WORLD MODELS IN ROBOTICS AND AUTONOMOUS DRIVING: A SURVEY OF THE LATENCY-FIDELITY TRADE-OFF AND THE CASE FOR UNIFIED SCENE-MOTION DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

World models – learned predictive models that compress an agent's observations into a latent representation and forecast how that representation evolves under action – have become a unifying substrate for both simulated control and autonomous driving. This paper synthesizes thirty such systems (2018–2026) along four axes: the five training objectives in current use (variational latent dynamics, discrete autoregressive token prediction, denoising score matching, joint-embedding feature prediction, and value-equivalent dynamics); four architecture families (recurrent/variational, Transformer/diffusion, Joint-Embedding Predictive Architectures, and driving/robotics-specific systems); the choice of 3D scene representation, from raw pixels to bird's-eye-view latents, occupancy grids, and point clouds; and five largely incommensurable evaluation protocols, under which only about a quarter of latency-relevant papers report any throughput figure at all. A persistent tension between representational richness and inference latency runs through every family: the most expressive generative world models are, by their own authors' account, not yet real-time, while real-time-capable models sacrifice visual fidelity, multi-agent modeling, or explicit 3D grounding. We argue that unifying scene and motion representation within a single, low-latency model – rather than treating world modeling and action planning as separate stages – is the central open challenge for physical, real-time robotic and autonomous-driving deployment, and examine three candidate directions toward it – continuous-time (Neural ODE) dynamics, cross-embodiment generalization, and structured state-space backbones – each grounded in evidence from the surveyed corpus or its margins, though none yet demonstrated at the scale the gap requires.

Keywords: world models, model-based reinforcement learning, autonomous driving, scene representation, Joint-Embedding Predictive Architectures, diffusion models, vision-language models, partially observable Markov decision processes, evaluation protocols, latency-fidelity trade-off, structured state-space models

1 INTRODUCTION

Learning a *world model* – a network that compresses raw sensory input into a latent representation and predicts how that representation evolves under action – has become the dominant strategy for giving an agent a way to plan or act without touching the real environment at every step. Ha and Schmidhuber's *World Models* (Ha & Schmidhuber, 2018b;a) frame this directly as an analogy to human cognition, and show that a controller trained entirely inside a model's hallucinated rollouts transfers back to the real environment. This "learn a model, then act inside it" pattern connects nearly every architecture surveyed here, from the recurrent, variational world models of 2018–2022, through the Transformer- and diffusion-based generative models of 2023–2024, to the Joint-Embedding Predictive Architecture (JEPA) family, and finally to driving- and robotics-specific

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Chronological timeline of the thirty surveyed world models, 2018–2026, colored by architecture family and marked by application domain.

systems that couple these dynamics cores with structured 3D scene representations. Figure 1 situates the thirty papers surveyed here on a single chronological timeline, grouped by architecture family.

This paper synthesizes thirty such systems, published 2018–2026, along four axes: architecture family and scene representation (Section 3), downstream deployment role (Section 3.C), and evaluation protocol (Section 4). The domain-specific transition-kernel taxonomy, the full architecture-comparison table, three candidate future backbones (continuous-time dynamics, cross-embodiment generalization, structured state-space models), and a condensed statement of the five training objectives are integrated directly into the sections above rather than deferred to an appendix (full derivations in the Appendix). The Appendix reproduces the remaining per-paper analysis, extended challenges discussion, and additional comparison tables this main text condenses; states how this survey positions itself against three adjacent literatures – model-based reinforcement learning, end-to-end autonomous driving, generative video-diffusion surveys; and gives the citation-chaining search process and inclusion criteria used to assemble the corpus (Appendix B).

A central theme threading through all four axes is a tension between representational richness and inference latency: the most visually expressive generative world models are, by their own authors' account, not yet real-time, while the models compact enough for real-time operation typically sacrifice visual fidelity, multi-agent modeling, or explicit 3D grounding. Section 5 argues that unifying scene and motion representation within a single, low-latency model – rather than treating world modeling and action planning as separate stages evaluated in isolation – is the field's central open challenge for physical, real-time deployment.

2 Background: World Models as POMDPs

Every world model surveyed here approximates the same control-theoretic object: a partially observable Markov decision process (POMDP)

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, R, \Omega, O, \gamma)\,, \tag{1}$$

following PlaNet and Dreamer's own problem statement (Hafner et al., 2019; 2020), where the true state $s \in \mathcal{S}$ is never observed directly, only a high-dimensional observation $o \in \Omega$ (an image, a LiDAR sweep, a fused multi-sensor frame) generated by the observation function $O(o_t \mid s_t)$.

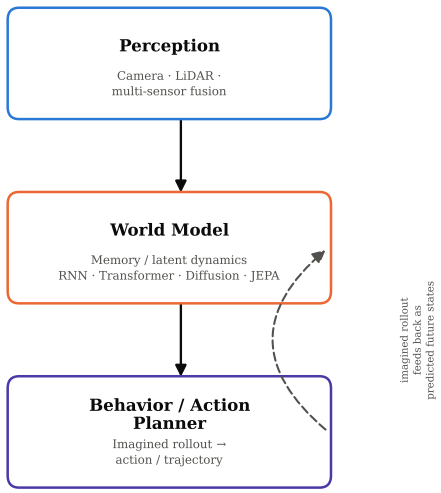

Figure 2: The perception / world-model / behavior pipeline every surveyed system instantiates a version of.

A world model is a set of learned functions $\hat{T}_\theta, \hat{O}_\theta, \hat{R}_\theta$ approximating $\mathcal{M}$'s true transition, observation, and reward mappings from past experience, via a compact latent state s_t standing in for the true, unobserved state. Concretely, this reduces to a *perception–world-model–behavior* pipeline (Figure 2): perception encodes raw sensor input into a scene representation, the world model predicts how that representation evolves ($\hat{T}_\theta$ above), and behavior or planning consumes the predicted representation to select an action – the pattern every architecture family surveyed in Section D instantiates.

The thirty surveyed systems instantiate five distinct training objectives for fitting θ (full derivations, with each paper's own notation, in Appendix C). *Variational latent dynamics*: the PlaNet/Dreamer Recurrent State-Space Model (RSSM) line (Hafner et al., 2019; 2020; 2021; 2023) splits the latent into a deterministic component h_t and a stochastic component s_t,

$$\begin{aligned} h_t &= f(h_{t-1}, s_{t-1}, a_{t-1}), \\ s_t &\sim p_\theta(s_t \mid h_t), \quad o_t \sim p_\theta(o_t \mid h_t, s_t), \end{aligned} \tag{2}$$

and fits an approximate posterior $q_\theta(s_t \mid h_t, o_t)$ by maximizing a variational evidence lower bound that trades off reconstruction of o_t against a KL penalty toward the prior $p_\theta(s_t \mid h_t)$ used for imagined rollouts (full ELBO in Appendix C). *Discrete autoregressive token prediction*: IRIS and Genie (Micheli et al., 2023; Bruce et al., 2024) tokenize each frame and model dynamics with a GPT-style Transformer over interleaved frame and action tokens. *Denoising score matching*: DriveDreamer and Copilot4D (Wang et al., 2023; Zhang et al., 2024) cast dynamics prediction as conditional diffusion, training a denoising network ϵ_ϕ to recover injected Gaussian noise from a scene condition c,

$$\min_\phi \; \mathbb{E}_{\mathcal{Z}_0,\, \epsilon,\, t,\, c} \left[\left\| \epsilon - \epsilon_\phi(\mathcal{Z}_t, t, c) \right\|_2^2 \right], \tag{3}$$

recovering the future observation at sampling time by iteratively reversing this process from noise, conditioned on c (Copilot4D applies the same principle to a discrete, tokenized point-cloud space instead). *Joint-embedding feature prediction*: I-JEPA and V-JEPA (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025), instantiating LeCun's JEPA proposal (LeCun, 2022), predict masked-patch representations from visible context using a context encoder, an EMA-updated target encoder, and a predictor, with no pixel reconstruction at all:

$$\min_\theta \; \left\| \hat{s}_\theta(z_x, m) - \mathrm{sg}(s_{\bar{\theta}}(y)) \right\|_2^2, \qquad \bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\theta, \tag{4}$$

where $\mathrm{sg}(\cdot)$ is a stop-gradient and the target encoder $\bar{\theta}$ is an exponential moving average of the online encoder θ, preventing the collapse a naive Siamese predictor would otherwise reach trivially. *Value-equivalent dynamics*: MuZero (Schrittwieser et al., 2020) trains a latent state to be sufficient

for predicting policy, value, and reward – with no requirement that it resemble the true environment state –

$$\mathcal{L}_{\text{MuZero}} = \sum_{k=0}^{K} \left[\ell^r\left(u_{t+k},\, r_t^k\right) + \ell^v\left(z_{t+k},\, v_t^k\right) + \ell^p\left(\pi_{t+k},\, p_t^k\right) \right], \tag{5}$$

where r_t^k, v_t^k, p_t^k are the model's k-step-unrolled reward, value, and policy predictions and $u_{t+k}, z_{t+k}, \pi_{t+k}$ their respective training targets, and plans via Monte Carlo tree search over the learned model at decision time.

A trained model becomes behavior via one of three mechanisms (Appendix C): *gradient-based imagination*, backpropagating policy gradients through differentiable latent rollouts (Dreamer's actor-critic, DayDreamer (Wu et al., 2022)); *search-based model-predictive control* over the latent rollout (PlaNet's cross-entropy-method planner, MuZero's tree search); or *conditional generative decoding*, sampling the next observation directly given an action (Genie, GAIA-1, GameNGen).

3 Architecture Taxonomy, Scene Representation, and Deployment Roles

3.1 Four Architecture Families

Recurrent and variational (2018–2022). Ha and Schmidhuber's V/M/C (Ha & Schmidhuber, 2018b;a) trains a controller entirely inside imagined rollouts. PlaNet (Hafner et al., 2019): RSSM (Section C) plus online cross-entropy-method planning. Dreamer (Hafner et al., 2020): differentiable actor-critic via backpropagated value gradients through imagined trajectories. DreamerV2 (Hafner et al., 2021): discrete categorical latents. DreamerV3 (Hafner et al., 2023): normalization/balancing lets one fixed configuration solve 150+ tasks across eight domains, including Minecraft diamond collection from raw pixels and sparse reward. DayDreamer (Wu et al., 2022): same algorithm, online, on physical robots, no simulator. MuZero (Schrittwieser et al., 2020): a value-equivalent model plus Monte Carlo tree search, superhuman on Atari/Go/chess/shogi without knowledge of the dynamics; TD-MPC/TD-MPC2 (Hansen et al., 2022; 2024) extend the same principle to continuous-action robotics via a receding-horizon optimizer, 104 tasks under one hyperparameter set. In driving, MILE (Hu et al., 2022) adapts the recurrent template to a bird's-eye-view latent, +31% CARLA driving score over prior state of the art from purely offline imitation.

Transformer and diffusion (2023–2024). IRIS (Micheli et al., 2023): discrete autoencoder plus autoregressive Transformer, 1.046 mean human-normalized Atari 100k score from roughly two hours of gameplay per game. Genie (Bruce et al., 2024): 11B-parameter, action-free, latent action interface from internet video alone. GameNGen (Valevski et al., 2025): diffusion-based neural game engine running DOOM in real time at 20 FPS. Cosmos (NVIDIA et al., 2025): open-weight world foundation model platform for physical AI (diffusion and autoregressive variants over tokenized video), demonstrated on manipulation, driving, and camera control – like GAIA-1/GAIA-2, not real-time, built for offline policy training. In driving: GAIA-1/GAIA-2 (Hu et al., 2023; Russell et al., 2025) generate multi-camera video conditioned on ego-vehicle, agent, and road-semantic signals (explicitly not real-time); DriveDreamer/DriveDreamer-2 (Wang et al., 2023; Zhao et al., 2025) condition diffusion on HD maps and, in the second version, LLM-generated scene descriptions; Vista (Gao et al., 2024) improves FID 55%/FVD 27% over the prior best driving world model and doubles as a reward function; WorldDreamer (Wang et al., 2024) proposes a masked-token objective on nuScenes; ViDAR (Yang et al., 2024b), Copilot4D (Zhang et al., 2024), and OccWorld (Zheng et al., 2024) move prediction to 3D point clouds/occupancy grids; ADriver-I (Jia et al., 2023) finetunes a multimodal LLM for direct control-signal prediction, paired with a video-diffusion decoder.

Joint-Embedding Predictive Architectures. LeCun's JEPA proposal (LeCun, 2022) predicts a target representation from context via an EMA target encoder, avoiding collapse without pixel reconstruction. I-JEPA (Assran et al., 2023)/V-JEPA (Bardes et al., 2024): frozen-backbone pretraining for images/video; V-JEPA 2 (Assran et al., 2025): over a million hours of video plus robot data, 4–16 FPS action-conditioned inference for zero-shot pick-and-place; AD-L-JEPA (Zhu et al., 2025): adapts the objective to LiDAR, BEV-embedding prediction for automotive 3D detection at a fraction of prior pretraining GPU cost.

Convergence toward unified scene-motion representation. WorldDrive (Gui et al., 2026), the most recent system surveyed, argues all four families above leave scene and motion representation as separate, non-planner-shared stages, and proposes a trajectory-aware diffusion Transformer at 53 ms end-to-end versus baselines the authors cite at 570–850 ms. MILE and OccWorld partially anticipate this with earlier, single-family results; Section 5 returns to this convergence as the survey's central open challenge.

Table 1: Architecture comparison across the surveyed world models. "Latency / real-time viability" reports only figures the paper itself states; "Not reported" means no such figure was found in the extracted text.

Model	Year	Architecture Type	Scene Representation	Latency / Real-Time Viability
World Models (Ha & Schmidhuber, 2018b)	2018	RNN (MDN-RNN + VAE)	Image (raw pixels)	Not reported
World Models, NeurIPS ed. (Ha & Schmidhuber, 2018a)	2018	RNN (MDN-RNN + VAE)	Image (raw pixels)	Not reported
PlaNet (Hafner et al., 2019)	2019	RNN (RSSM)	Image (raw pixels)	Not reported
Dreamer (Hafner et al., 2020)	2020	RNN (RSSM)	Image (raw pixels)	Not reported
MuZero (Schrittwieser et al., 2020)	2020	Other (ResNet + MCTS, value-equivalent)	Image / board state	Real-time-compatible (100 ms per-move search budget)
DreamerV2 (Hafner et al., 2021)	2021	RNN (RSSM, discrete latents)	Image (raw pixels)	Not reported
MILE (Hu et al., 2022)	2022	RNN (recurrent BEV world model)	BEV	Real-time (43.0 Hz fully-recurrent inference, vs. 6.2 Hz windowed)
LeCun position paper (LeCun, 2022)	2022	JEPA (conceptual proposal)	N/A (conceptual)	N/A
DayDreamer (Wu et al., 2022)	2022	RNN (RSSM)	Image (onboard camera)	Real-time (trained/deployed online on physical robots)
I-JEPA (Assran et al., 2023)	2023	JEPA (Vision Transformer)	Image (static images)	N/A (representation pretraining, not closed-loop)
DreamerV3 (Hafner et al., 2023)	2023	RNN (RSSM)	Image (raw pixels)	Not reported
GAIA-1 (Hu et al., 2023)	2023	Transformer + Diffusion (hybrid)	Image (multi-camera video)	Not real-time (explicitly stated by authors)
ADriver-I (Jia et al., 2023)	2023	Transformer (MLLM) + Diffusion	Image (vision-action pairs)	Not reported
IRIS (Micheli et al., 2023)	2023	Transformer (autoregressive, VQ tokens)	Image (raw pixels)	Not reported
DriveDreamer (Wang et al., 2023)	2023	Diffusion	Image + BEV/HD-map conditioning	Not reported
UniSim (Yang et al., 2024a)	2024	Diffusion	Image (raw pixels)	Not reported
V-JEPA (Bardes et al., 2024)	2024	JEPA	Image (video)	N/A (representation pretraining)
Genie (Bruce et al., 2024)	2024	Transformer (ST-Transformer + VQ latent actions)	Image (raw pixels)	Not real-time ($\sim$1 FPS as reported)
Vista (Gao et al., 2024)	2024	Diffusion	Image (camera video)	Not reported
GameNGen (Valevski et al., 2025)	2024	Diffusion (augmented Stable Diffusion)	Image (raw pixels)	Real-time (20 FPS interactive play reported)
WorldDreamer (Wang et al., 2024)	2024	Transformer (masked visual-token prediction)	Image (raw pixels)	Not reported
ViDAR (Yang et al., 2024b)	2024	Transformer (latent rendering)	Image $\rightarrow$ 3D Point Cloud (forecast)	Not reported
Copilot4D (Zhang et al., 2024)	2024	Diffusion (discrete, VQ tokens)	3D Point Cloud (LiDAR)	Not reported
DriveDreamer-2 (Zhao et al., 2025)	2024	Diffusion + LLM front-end	Image + BEV/HD-map conditioning	Not reported
OccWorld (Zheng et al., 2024)	2024	Transformer (GPT-style, occupancy tokens)	3D Occupancy (voxel grid)	Not reported
Cosmos (NVIDIA et al., 2025)	2025	Diffusion + Transformer (autoregressive, dual variants)	Image (tokenized video)	Not real-time (offline policy training; no latency figure reported)
V-JEPA 2 (Assran et al., 2025)	2025	JEPA	Image (video) + robot proprioception	Near real-time (4–16 FPS reported for robot action inference)
GAIA-2 (Russell et al., 2025)	2025	Diffusion (DiT-based)	Image (multi-camera video, 5–6 cams)	Not yet real-time (explicit limitation stated)
AD-L-JEPA (Zhu et al., 2025)	2025	JEPA	BEV Embedding (LiDAR input)	Not reported
WorldDrive (Gui et al., 2026)	2026	Diffusion Transformer (TA-DiT)	Image (multi-view) + motion fusion	Real-time (53 ms end-to-end, vs. 570–850 ms baselines)

3.2 Scene Representation and Domain-Specific Dynamics

The surveyed systems make four distinct geometric choices (full treatment in Appendix F): raw *image-space* pixels (Dreamer line, Genie, GameNGen), *bird's-eye-view* latents (MILE, DriveDreamer/DriveDreamer-2), explicit *3D occupancy or point clouds* (OccWorld, ViDAR, Copilot4D), and *representation-space* embeddings with no reconstruction (AD-L-JEPA) – no single choice dominates the most recent papers. Transition-function entropy also differs fundamentally by domain: gaming is close to a deterministic, hand-authored program (Valevski et al., 2025; Schrittwieser et al., 2020); AI-agent (GUI) transitions follow a human-authored reachability graph rather than physical law; robotics/autonomous-vehicle transitions are irreducibly stochastic, continuous, and multi-agent, reflecting genuine environmental variation rather than a gap the model could close (Hu et al., 2022; Russell et al., 2025). The next subsection formalizes this taxonomy; Appendix E gives the full domain-by-domain discussion.

3.3 A Mathematical Taxonomy of Domain-Specific Transition Functions

The three domains above imply structurally different transition kernels $T(s_{t+1} \mid s_t, a_t)$, summarized in Fig. 3 and derived in full in Appendix E. *Gaming* is close to a Dirac point mass at a hand-authored engine's deterministic output over a small discrete action set (Valevski et al., 2025; Schrittwieser et al., 2020):

$$T_{\text{game}}(s_{t+1} \mid s_t, a_t) = \delta(s_{t+1} - f_{\text{engine}}(s_t, a_t)), \quad a_t \in \mathcal{A}_{\text{discrete}}. \tag{6}$$

AI agents are likewise effectively deterministic, but constrained by a human-authored interaction graph $\mathcal{R}(r_t, m_t)$ rather than physical law (Xu et al., 2026):

$$T_{\text{agent}}(s_{t+1} \mid s_t, a_t) = \begin{cases} p(s_{t+1}) & s_{t+1} \in \mathcal{R}(r_t, m_t), \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Robotics and autonomous vehicles are irreducibly stochastic and continuous, reflecting genuine environmental and multi-agent variation that no engine or reachability graph can capture (Hu et al., 2022; Russell et al., 2025; Rao et al., 2026):

$$T_{\text{robot}}(s_{t+1} \mid s_t, a_t) = \mathcal{N}(f_{\text{robot}}(s_t, a_t),\, \Sigma(s_t, a_t)), \quad a_t \in \mathcal{A}_{\text{continuous}}. \tag{8}$$

Gaming and AI agents are both low-entropy, for different reasons – a known program versus a human-authored graph – while robotics is irreducibly stochastic because f_{robot} is only ever an approximation to physical law, never an exact program the way f_{engine} is for gaming.

3.4 Downstream Deployment Roles

Across the corpus, a trained world model plays one of four recurring roles (full treatment in Appendix G): a *simulator or data engine* generating training data or rare safety-critical scenarios (UniSim (Yang et al., 2024a), GAIA-1/2); a *planner*, consumed via gradient-based imagination or search (Dreamer, DayDreamer, MILE, Vista); a *representation-pretraining backbone* evaluated on downstream transfer rather than closed-loop control (I-JEPA, V-JEPA, ViDAR, AD-L-JEPA); or the *dynamics core inside a unified perception-action stack* that also performs perception and control (ADriver-I, V-JEPA 2-AC, WorldDrive). These roles correlate with architecture: gradient-based imagination favors RSSM, pretraining favors JEPA, and unified stacks are the newest systems in the corpus.

4 Evaluation Protocols and the Latency Question

The thirty surveyed papers report results on five largely incommensurable metric families, detailed in full in Appendix H. *Game and control benchmarks*: human-normalized score on Atari (MuZero's full 57-game suite, DreamerV2's 55 games, IRIS's data-efficient Atari 100k subset) and DreamerV3's 150+ tasks across eight domains, evaluated under training budgets that differ by up to two orders of magnitude even when nominally reporting the same metric. *Generative video-quality metrics*: Fréchet Video/Inception Distance (FVD/FID) and, for GAIA-2, Fréchet DINO Distance (FDD) – computed against different held-out video sets and, in GAIA-2's case, a different feature space than

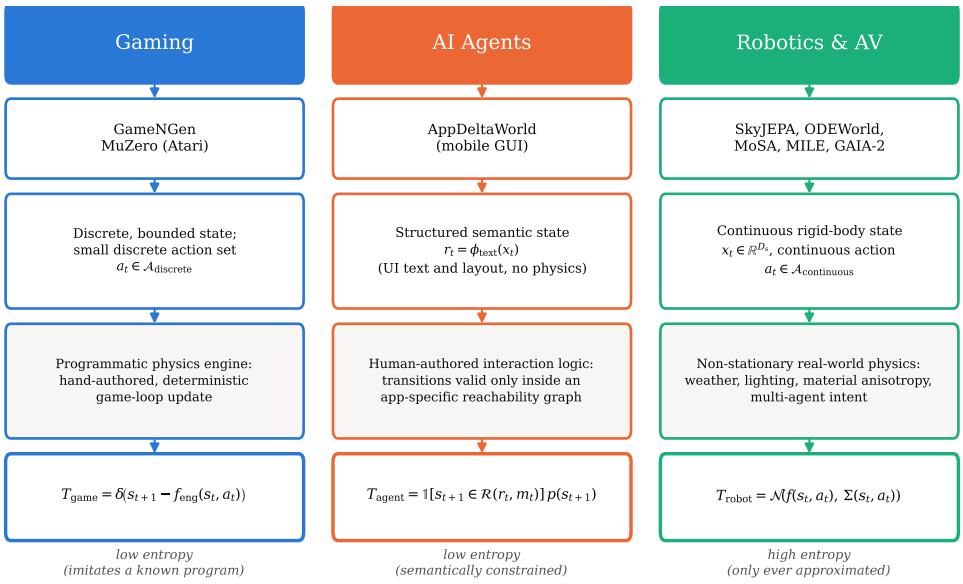

Figure 3: A visual taxonomy of the three domain-specific transition kernels derived above. Each column traces, top to bottom, the representative surveyed systems, the character of the state and action space, the governing constraint on plausible transitions, and the resulting formal kernel. Full per-domain derivation, with direct quotations from each source paper, is in Appendix E.

FVD, so figures are not directly comparable even within the same architecture family (WorldDrive, in the same family as GAIA-2, reports no such metric at all). *Driving perception and planning metrics*: nuScenes Detection Score, mIoU occupancy forecasting, Chamfer Distance, CARLA Driving Score, and the composite NAVSIM Predictive Driver Model Score. *Representation-transfer metrics*: ImageNet-1K and Something-Something-v2 linear-probe accuracy (I-JEPA, V-JEPA 2), unrelated to closed-loop control. *Latency/throughput*: standard mobile-robot closed-loop control calls for 10–30 Hz update rates (Siegwart et al., 2011); excluding the three papers for which latency is inapplicable (LeCun's position paper, I-JEPA, V-JEPA), only 6 of the remaining 26 empirically-deployed papers (∼23%) report any inference-time or throughput figure at all, and those six are measured on different hardware under different batch and resolution settings.

These five families answer five genuinely different questions – game-playing competence, generative fidelity, driving quality, forecasting accuracy, and representation transferability – so no single number could summarize all five without losing what makes each meaningful. The more diagnostic finding is that comparability breaks down even *within* a family, and that latency, specifically, is reported by fewer than a quarter of the papers for which it is a meaningful question at all – leaving the field without a shared basis for certifying real-time viability, the evaluation-side counterpart of the architectural latency bottleneck of Section 3.

5 Discussion: Open Challenges and the Central Tension

Latency–fidelity trade-off. The most expressive generative world models – GAIA-1 and GAIA-2 – are, by their own authors' statement, not yet real-time, while systems that do report a real-time figure (MILE at 43.0 Hz, GameNGen at 20 FPS, WorldDrive at 53 ms) are each less representationally rich than the GAIA line. This trade-off runs through every family and reflects current engineering, not an architectural law. Toward closing it, GAIA-2 proposes distillation and inference-time acceleration (Russell et al., 2025), while V-JEPA 2-AC's 16-second per-action planning time against diffusion baseline Cosmos's four minutes (Assran et al., 2025) shows representation-space prediction is already viable.

Hallucination and long-horizon error accumulation. Autoregressive latent rollouts compound small one-step errors: a model can score well on FVD/FID at short horizons yet drift into physically

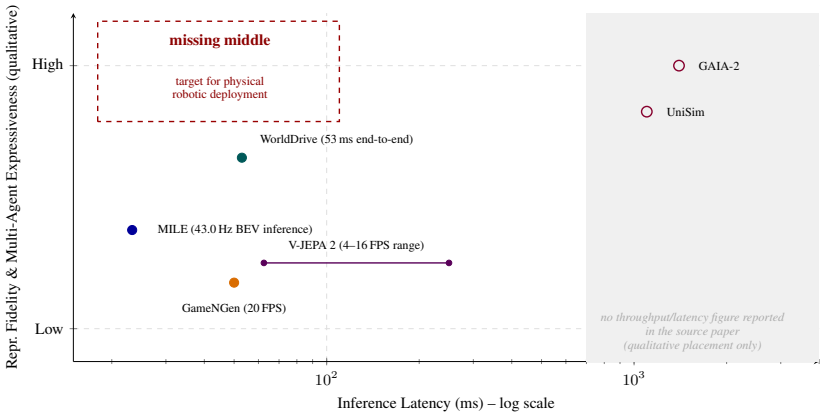

Figure 4: The latency–fidelity Pareto frontier this survey argues is the field's central open problem. *Filled markers* report an actual latency or throughput figure, converted to milliseconds where the source gives Hz or FPS (MILE (Hu et al., 2022), GameNGen (Valevski et al., 2025), WorldDrive (Gui et al., 2026)); V-JEPA 2's 4–16 FPS range (Assran et al., 2025) is drawn as a range, since that is what its authors report. *Hollow markers inside the shaded band* – GAIA-2 (Russell et al., 2025), UniSim (Yang et al., 2024a) – report no latency or throughput figure at all (Section I.D), so their horizontal placement is qualitative. The vertical axis is likewise a qualitative synthesis of representational and multi-agent richness (Sections D–G), not a benchmarked score, since no such shared metric exists across this literature (Section H). The dashed "missing middle" region is empty by construction: the low-latency, high-fidelity combination no system here has yet reported achieving.

implausible states over long ones. Vista, GAIA-2, and UniSim converge on a learned per-step confidence signal as the concrete near-term fix (Gao et al., 2024; Russell et al., 2025; Yang et al., 2024a).

Sim-to-real and out-of-domain generalization. Systems trained in CARLA or on offline driving logs must generalize to weather, lighting, and geographic conditions unseen in training (MILE, GAIA-2); DayDreamer's online, simulator-free training is one direct response, unavailable to the generative and JEPA-pretraining systems evaluated offline. ViDAR and AD-L-JEPA both propose scaling pretraining data toward richer, action-conditioned objectives (Yang et al., 2024b; Zhu et al., 2025).

Absence of a shared evaluation protocol. Section 4's five-family fragmentation is itself an open problem: without a shared closed-loop, real-time benchmark, claims of progress across papers in different families are not directly comparable. WorldDrive's composite NAVSIM score and MILE's supplementary reward metric (Gui et al., 2026; Hu et al., 2022) are two steps toward such a protocol.

Multi-agent and safety-critical deployment. GAIA-2's multi-agent conditioning and WorldDrive's planner-shared motion representation target this directly, though staged, certifiable architectures and the unified, end-to-end ones (ADriver-I, V-JEPA 2-AC, WorldDrive) this survey favors for latency remain in tension. GAIA-2's roadmap toward "richer agent behavior modeling" (Russell et al., 2025) is the most concrete step.

Synthesis. These five challenges are facets of one underlying tension (Section 6 draws out the full argument). WorldDrive is the most complete evidence to date, but a single 2026 preprint, not peer-reviewed – so we treat it, and the three candidate backbones below, as promising, not field-wide; none of the thirty core papers adopts any of them. The third, structured state-space models, is developed alongside the other two below.

5.1 Continuous-Time Dynamics and Neural ODEs

This subsection and the two below argue this survey's own position, not a corpus synthesis: only continuous-time dynamics has a corpus representative, ODEWorld (Liu et al., 2026) (Appendix B);

cross-embodiment generalization and structured state-space backbones draw on evidence in adjacent literatures, since no surveyed world model has attempted either.

Every backbone family surveyed assumes a uniform, discrete time grid – fixed RSSM timesteps, GameNGen's game loop, evenly spaced Transformer/diffusion tokens – with no native notion of time *between* observations, though real sensors sample irregularly across camera, IMU, and LiDAR. Chen et al.'s Neural ODE parameterizes "the derivative of the hidden state using a neural network," queryable at any time value (Chen et al., 2018), later extended to "naturally handl[ing] arbitrary time gaps between observations" (Rubanova et al., 2019). This survey's corpus already contains a world model built this way: ODEWorld reframes latent dynamics as a continuous velocity field ($z_\tau = z_0 + \int_0^\tau v_\theta(z_t, t; z_0, c)\, dt$), reporting "arbitrary temporal resolution and even backward prediction" (Liu et al., 2026) – evaluated so far only on video reconstruction quality, not a real irregularly-sampled sensor stream or continuous-time controller.

5.2 Cross-Embodiment Generalization

Every robotics/driving system surveyed here is trained on one embodiment – MILE and the GAIA/DriveDreamer/Vista line around a single chassis, SkyJEPA written for a quadrotor with "four motor forces" (Rao et al., 2026) – so retraining for a new chassis is a separate run, not a transfer the architecture addresses. A cross-embodiment path already exists at the *policy-learning* level: Open X-Embodiment's "dataset from 22 different robots... demonstrating 527 skills" yields an RT-X policy that "exhibits positive transfer... by leveraging experience from other platforms" (Open X-Embodiment Collaboration, 2024). Open X-Embodiment's scope, though, is "robotic manipulation," not the wheeled, aerial, and underwater diversity a genuinely embodiment-general *world model* needs – single-morphology dynamics cores must give way to a cross-embodiment *core* built on embodiment-agnostic physical quantities ($SE(3)$ ego-pose, depth, contact geometry) rather than raw pixels.

5.3 A Candidate Future Backbone: Structured State-Space Models

A structured state-space model (SSM) models sequences as a linear recurrence with input-dependent parameters: "5× higher throughput than Transformers and linear scaling in sequence length" (Gu & Dao, 2023). Outside this corpus, S4WM reports 500-step imagination MSE of 44.0 against 219.4 (RSSM) and 224.4 (Transformer) baselines (Deng et al., 2023) – evidence no surveyed family combines.

6 Conclusion

This paper has synthesized thirty world models (2018–2026) along four axes – architecture family (Section 3), 3D scene representation, downstream deployment role, and evaluation protocol (Section 4) – converging on one argument: representational richness trades off against inference latency, compounded by open problems in long-horizon hallucination, sim-to-real generalization, and multi-agent safety-critical reasoning; no system surveyed is simultaneously real-time, safe, generalizable, and adequately evaluated. Unifying scene and motion representation in one low-latency model is the central challenge; continuous-time dynamics, cross-embodiment generalization, and structured state-space backbones (Section 5) are promising, though not field-wide, paths toward it.

AI use statement

We used a generative AI system (Claude, Anthropic) substantially in producing this manuscript: for literature synthesis and drafting of the survey prose in every section; for technical formalization of the mathematical notation, equations, and derivations presented in Sections 2–5 and the Appendix; and for iterative restructuring of the manuscript, including the main-text/appendix split, to fit the target page budget. We did not use generative AI tools to conduct experiments or to generate the underlying research findings this survey reports on – there are none; this is a literature synthesis, and every claim, quotation, and figure attributed to a surveyed paper is drawn from that paper's own published text, not generated. We reviewed all AI-assisted content, checking each formalized equation for mathematical correctness and each direct quotation and citation against its source paper.

We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

This paper synthesizes and analyzes previously published research. It introduces no new datasets, human-subjects data, or trained models, and we are not aware of ethical concerns specific to this survey beyond those already addressed by the authors of the individual papers surveyed. All quotations from surveyed papers are reproduced under fair-use scholarly citation norms and attributed to their original source.

Reproducibility statement

This is a survey paper: there are no new experiments to reproduce. Every quantitative claim restated from the surveyed literature is cited to its original source, and every direct quotation is drawn verbatim from the corresponding paper. Appendix B states, in full, the corpus search process and inclusion/exclusion criteria used to assemble the 30-paper corpus, so that the corpus itself can be independently audited or extended.

Author Contributions

Sole-authored. The author conceived the survey's scope, four-axis organizational structure, and corpus inclusion criteria; directed and reviewed the AI-assisted literature synthesis, drafting, and technical formalization described in the AI Use Statement above; and verified the final content, including all cited claims, against the original papers.

Acknowledgments

This work received no external funding. We thank the maintainers of arXiv and the publishers making the surveyed papers openly accessible, without which the source-grounded verification this survey relies on would not have been possible.

References

Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Józefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020. arXiv:1808.00177.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. arXiv:2301.08243.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair,

Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.15391.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *arXiv preprint arXiv:2306.16927*, 2024.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. arXiv:1806.07366.

Fei Deng, Junyeong Park, and Sungjin Ahn. Facing off world model backbones: RNNs, Transformers, and S4. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2307.02064.

Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.17398.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Xingtai Gui, Meijie Zhang, Tianyi Yan, Wencheng Han, Jiahao Gong, Feiyang Tan, Cheng-zhong Xu, and Jianbing Shen. Bridging scene generation and planning: Driving with world model via unifying vision and motion representation. *arXiv preprint arXiv:2603.14948*, 2026.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018a. arXiv:1809.01999.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018b.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning (ICML)*, 2019. arXiv:1811.04551.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1912.01603.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with discrete world models. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.02193.

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

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. *International Conference on Machine Learning (ICML)*, 2022. arXiv:2203.04955.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.16828.

Bohan Hou, Gen Li, Jindou Jia, Tuo An, Xinying Guo, Sicong Leng, Haoran Geng, Yanjie Ze, Tatsuya Harada, Philip Torr, Oier Mees, Marc Pollefeys, Zhuang Liu, Jiajun Wu, Pieter Abbeel, Jitendra Malik, Yilun Du, and Jianfei Yang. World model for robot learning: A comprehensive survey. *arXiv preprint arXiv:2605.00080*, 2026.

Anthony Hu, Gianluca Corrado, Nicolas Griffiths, Zak Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, and Jamie Shotton. Model-based imitation learning for urban driving. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2210.07729.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. ADriver-I: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5580–5590, 2017. arXiv:1703.04977.

Philip Koopman and Michael Wagner. Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine*, 9(1):90–96, 2017. doi: 10.1109/MITS.2016.2583491.

Yann LeCun. A path towards autonomous machine intelligence. Position paper, version 0.9.2, 2022. URL `https://openreview.net/pdf/315d43ba26f55357a84cec9a7ed15a6610094f79.pdf`.

Dongxiu Liu, Haoyi Niu, Peng Cheng, Yuan Gao, Xirui Kang, Sangli Teng, Koushil Sreenath, and Xianyuan Zhan. ODEWorld: A continuous predictive architecture via physical-time flow. *arXiv preprint arXiv:2607.27924*, 2026.

Javier Alejandro Lopetegui Gonzalez, Paul Pacaud, and Cordelia Schmid. Spatially aware world action model via geometric latent diffusion. *arXiv preprint arXiv:2609.02531*, 2026.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2209.00588.

Thomas M. Moerland, Joost Broekens, Aske Plaat, and Catholijn M. Jonker. Model-based reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, 16(1):1–118, 2023. arXiv:2006.16712.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. *arXiv preprint arXiv:2310.08864*, 2024.

Pratyaksh Rao, Wancong Zhang, Randall Balestriero, Yann LeCun, and Giuseppe Loianno. Sky-JEPA: Learning long-horizon world models for zero-shot sim-to-real control of quadrotors. *arXiv preprint arXiv:2606.23444*, 2026.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15, pp. 627–635, 2011. arXiv:1011.0686.

Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent ODEs for irregularly-sampled time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1907.03907.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. GAIA-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588:604–609, 2020. arXiv:1911.08265.

Roland Siegwart, Illah Reza Nourbakhsh, and Davide Scaramuzza. *Introduction to Autonomous Mobile Robots*. MIT Press, 2nd edition, 2011.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017. arXiv:1703.06907.

Sifan Tu, Xin Zhou, Dingkang Liang, Xingyu Jiang, Yumeng Zhang, Xiaofan Li, and Xiang Bai. The role of world models in shaping autonomous driving: A comprehensive survey. *arXiv preprint arXiv:2502.10498*, 2025.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2408.14837.

Jiaxu Wang, Junhao He, Jingkai Sun, Yi Gu, Yunyang Mo, Jiahang Cao, Qiang Zhang, and Renjing Xu. MoSA: Motion-constrained stress adaptation for mitigating real-to-sim gap in continuum dynamics via learning residual anisotropy. *International Conference on Machine Learning (ICML)*, 2026. arXiv:2605.22597.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drive-Dreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. World-Dreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985*, 2024.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and Pieter Abbeel. DayDreamer: World models for physical robot learning. In *Conference on Robot Learning (CoRL)*, 2022. arXiv:2206.14176.

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647*, 2023.

Weikai Xu, Yunren Feng, Haoxiang Lei, Kun Huang, Yuxuan Liu, Kang Zhao, Xiaolin Hu, Shuo Shang, and Bo An. AppDeltaWorld: Transition-grounded delta code world model for mobile GUI agents. *arXiv preprint arXiv:2608.05891*, 2026.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations (ICLR)*, 2024a. arXiv:2310.06114.

Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b. arXiv:2312.17655.

Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. Copilot4D: Learning unsupervised world models for autonomous driving via discrete diffusion. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2311.01017.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. DriveDreamer-2: LLM-enhanced world models for diverse driving video generation. In *AAAI Conference on Artificial Intelligence*, 2025. arXiv:2403.06845.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. OccWorld: Learning a 3D occupancy world model for autonomous driving. In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2311.16038.

Haoran Zhu, Zhenyuan Dong, Kristi Topollai, Beiyao Sha, and Anna Choromanska. Self-supervised representation learning with joint embedding predictive architecture for automotive LiDAR object detection. *arXiv preprint arXiv:2501.04969*, 2025.

Appendix

Contents

A Additional Comparison Tables 16

B Related Work, Positioning, and Corpus Methodology 19

B.1 Positioning Against Prior Surveys 19

B.2 Literature Search and Selection Methodology 20

B.3 Survey Scope 21

C Background and Problem Formulation 21

C.1 The World Model as a POMDP 22

C.2 RSSM Latent Dynamics and the Variational ELBO 22

C.3 Discrete Autoregressive Token Prediction 23

C.4 Denoising Score Matching for Diffusion World Models 23

C.5 JEPA Feature Prediction with EMA Target Encoders 23

C.6 Value-Equivalent Dynamics 24

C.7 The Imagination Loop: Latent Rollouts for Policy Gradients and Model-Predictive Control 24

D Architectural Evolution: From Recurrent Latent Dynamics to Joint-Embedding Prediction 25

D.1 Recurrent and Variational World Models (2018–2022) 26

D.2 Transformer and Diffusion World Models (2023–2024) 26

D.3 Joint-Embedding Predictive Architectures (JEPA) 27

D.4 Vision-Language Models and Spatial Grounding 27

D.5 Multi-Modal Scene Representation: Sensor Fusion 27

D.6 Convergence: Unifying Scene and Motion Representation 27

E Domain-Specific Dynamics: Agents, Gaming, and Robotics 28

E.1 Gaming: Discrete, Bounded, and Programmatically Deterministic 28

E.2 AI Agents: Semantic State and Interaction Logic, Not Physics 29

E.3 Robotics and Autonomous Vehicles: Continuous, Stochastic, and Multi-Agent . . . 29

E.4 A Mathematical Taxonomy of Domain-Specific Transition Functions: Supporting Evidence 30

F Scene Understanding and Representation for Robotics and Autonomous Vehicles 31

F.1 Geometric Representation Trade-offs 31

F.2 Geometric Representation and Physical Grounding Cues 31

F.3 Semantic Cues and Interpretability 33

F.4 Temporal Dynamics and Uncertainty 33

F.5 Action-Conditioning and Controllability 33

G World Models for Downstream Robotics and Autonomous-Driving Tasks 34

G.1 World Models as Simulators and Data Engines 34

G.2 World Models as Planners 34

G.3 World Models as Representation-Pretraining Backbones 34
G.4 World Models Inside Unified Perception-Action Stacks 34
H Benchmarks and Evaluation Protocols 35
H.1 General-Domain Control and Game Benchmarks 35
H.2 Generative Video-Quality Metrics 36
H.3 Driving-Specific Perception and Planning Metrics 36
H.4 Representation-Transfer Evaluation 36
H.5 The Absence of a Shared Evaluation Protocol 36
I Open Challenges and Future Directions 37
I.1 The Latency–Fidelity Trade-off 37
I.2 Hallucination and Long-Horizon Error Accumulation 37
I.3 Sim-to-Real and Out-of-Domain Generalization 38
I.4 The Absence of a Shared Evaluation Protocol 38
I.5 Multi-Agent Interaction and Safety-Critical Deployment 38
I.6 Synthesis: A Single Underlying Tension 39

APPENDIX OVERVIEW

This appendix reproduces, in full, the full training-objective derivations, per-paper architectural analysis, scene-representation treatment, domain-dynamics derivation, downstream-role analysis, extended evaluation discussion, open-challenges discussion, and additional comparison tables that the main text condenses for space, plus the corpus methodology and related-work positioning omitted from the main text entirely. Appendix A's two additional comparison tables are placed first, immediately below, as the appendix's most directly reusable reference material; the remaining sections follow in the order they are cited from the main text. Per a reviewer-lens page-budget pass, two promotions from an earlier restructuring attempt are kept in the main text because they are cheap and high-value: the complete 30-row architecture-comparison table (Section 3) and the Latency-Fidelity Pareto Frontier figure (Section 5). Two others were reverted here because, under a hard 9-page limit, re-deriving already-published training objectives and airing speculative future-backbone essays are lower priority than a reviewer would place on core survey rigor: the full training-objective derivations (kept condensed in Section 2, full version here in Appendix C). One candidate-future-backbone subsection, structured state-space models, is kept appendix-only (Appendix I) as a full essay, since a reviewer would weight it below the other two under a hard page limit; the other two, continuous-time dynamics and cross-embodiment generalization, were promoted directly into Section 5 at the user's later, explicit request to elevate this survey's forward-looking argument (compressed to fit the main-text budget), so Appendix I no longer carries a separate copy of either – it points back to Section 5 instead. The domain-specific transition-kernel taxonomy is a middle case: its figure and three defining equations are kept in Section 3 (a genuine, citable formalization worth the main text's attention), and this appendix's own copy of that subsection (Appendix E) does not repeat that figure or those equations a second time – it keeps only the extended per-domain evidentiary prose that substantiates them and refers back to Section 3's numbering. Similarly, the corpus-timeline figure (Section 1, right below the abstract) and the perception/world-model/behavior pipeline diagram (Section 2, beside the training-objectives paragraph) each live only in the main text, not in this appendix's Appendix B and Appendix D respectively, which point back to them by number. Other than these promotions, every subsection below is verbatim content also used by `main_iclr2027.tex` (the full-length ICLR-styled variant of this survey) and, other than citation style, by `main.tex` (the IEEEtran conference variant) – nothing here is written specifically for this appendix, so the three main files cannot drift apart on facts.

A ADDITIONAL COMPARISON TABLES

Table 2: Benchmark datasets and key reported metric(s) as stated in each paper, restricted to papers evaluated on a driving benchmark; the metrics tabulated span closed-loop planning outcomes, open-loop perception/geometry scores, and video-generation distributional distances (Section H.C partitions these by protocol), not a single metric family. Metrics are quoted as stated in each paper; "proprietary" means the paper explicitly reports training/evaluating on an internal, non-public fleet dataset rather than nuScenes or Waymo.

Model	Year	Benchmark Dataset(s)	Key Reported Metric(s)
MILE (Hu et al., 2022)	2022	CARLA simulator (not nuScenes/Waymo)	+31% Driving Score over prior state of the art
GAIA-1 (Hu et al., 2023)	2023	Proprietary (4,700 h London driving fleet data; nuScenes/Waymo cited only as related work, not used)	Qualitative / expert evaluation of scenario realism; no public benchmark number reported
ADriver-I (Jia et al., 2023)	2023	nuScenes + private fleet dataset	Multi-threshold control-signal prediction accuracy, cross-dataset (nuScenes vs. private) comparison
DriveDreamer (Wang et al., 2023)	2023	nuScenes	FID and FVD (video generation quality)
Vista (Gao et al., 2024)	2024	nuScenes	FID improved 55% and FVD improved 27% over the best prior driving world model; also used as a reward function to score candidate actions
WorldDreamer (Wang et al., 2024)	2024	nuScenes (700 training scenes) + self-collected data	Reports high-fidelity image-to-video generation; explicit FID/FVD values not located in the extracted text
ViDAR (Yang et al., 2024b)	2024	nuScenes	As a self-supervised pretraining signal: +3.1% NDS (3D detection), +4.3% mAP, +4.6% mIoU, $\sim$33% reduction in Chamfer Distance (point-cloud forecasting)
Copilot4D (Zhang et al., 2024)	2024	nuScenes, KITTI Odometry, Argoverse2	Chamfer Distance reduced $>$65% vs. prior SOTA at 1 s prediction horizon; also reports L1 depth / AbsRel error
DriveDreamer-2 (Zhao et al., 2025)	2024	nuScenes	FID 11.2 and FVD 55.7 (first-frame multi-view image condition), a $\sim$30%/$\sim$50% relative improvement over the previous best-performing methods on the nuScenes validation set
OccWorld (Zheng et al., 2024)	2024	nuScenes	mIoU 17.13, IoU 26.63 for 3 s future occupancy forecast
GAIA-2 (Russell et al., 2025)	2025	Proprietary internal fleet dataset ($\sim$25M video clips; not nuScenes/Waymo)	FID and FDD (Fréchet DINO Distance) visual-fidelity metrics
AD-L-JEPA (Zhu et al., 2025)	2025	Waymo, KITTI3D, ONCE	3D detection mAP gains of +1.61 and +2.98 from JEPA-based pretraining (dataset-size dependent)
WorldDrive (Gui et al., 2026)	2026	nuScenes, NAVSIM, NAVSIM-v2	L2 trajectory error (m) and Collision Rate (%), plus NAVSIM composite sub-metrics (NC, DAC, TTC, Comfort, EP, PDMS); 53 ms end-to-end latency

Table 3: Full coverage matrix (appendix): the 30 core architecturally-surveyed papers, with publication year, author affiliation(s), scene representation, architecture backbone, application domain, and the dataset(s) the paper itself trains or evaluates on, plus a clearly demarcated supplement of the four further 2026 papers (AppDeltaWorld, SkyJEPA, MoSA, SA-WAM) drawn on in Section E and Section F for domain-taxonomy and geometric-grounding material this core matrix does not otherwise cover. ODEWorld, the fifth paper contributing to this survey's 35-paper corpus (Section 1), is discussed only in Section 5.1 as a supporting citation for an argued position rather than as a system with dataset/benchmark results comparable to the rows below, and is accordingly omitted from this reproducibility-oriented matrix.

Model	Yr.	Affiliation(s)	Representation	Backbone	Domain	Dataset(s)
World Models (Ha & Schmidhuber, 2018b)	2018	Google Brain; NNAISENSE; IDSIA (USI/SUPSI)	Image (raw pixels)	RNN	General	CarRacing-v0; VizDoom (Take Cover)
World Models, NeurIPS ed. (Ha & Schmidhuber, 2018a)	2018	Google Brain; NNAISENSE; IDSIA (USI/SUPSI)	Image (raw pixels)	RNN	General	CarRacing-v0; VizDoom (Take Cover)
PlaNet (Hafner et al., 2019)	2019	Google Brain; U. Toronto; DeepMind; Google Research; U. Michigan	Image (raw pixels)	RNN	General	DeepMind Control Suite
Dreamer (Hafner et al., 2020)	2020	U. Toronto; DeepMind; Google Brain	Image (raw pixels)	RNN	General	DeepMind Control Suite
MuZero (Schrittwieser et al., 2020)	2020	DeepMind; UCL	Image / board state	Other	General	Atari-57; Go; Chess; Shogi
DreamerV2 (Hafner et al., 2021)	2021	Google Research; DeepMind; U. Toronto	Image (raw pixels)	RNN	General	Atari (55 games)
MILE (Hu et al., 2022)	2022	Wayve; U. Cambridge	BEV	RNN	Driving	CARLA simulator
LeCun Position Paper (LeCun, 2022)	2022	NYU; Meta FAIR	N/A (conceptual)	JEPA	General	N/A (position paper)
DayDreamer (Wu et al., 2022)	2022	UC Berkeley	Image (onboard camera)	RNN	General	Physical robot tasks, 4 robots (no public dataset)
I-JEPA (Assran et al., 2023)	2023	Meta FAIR; McGill; Mila; NYU	Image (static images)	JEPA	General	ImageNet-1K
DreamerV3 (Hafner et al., 2023)	2023	Google DeepMind; U. Toronto	Image (raw pixels)	RNN	General	Control Suite; Atari; ProcGen; DMLab; Minecraft; BSuite
GAIA-1 (Hu et al., 2023)	2023	Wayve	Image (multi-camera video)	Transformer	Driving	Proprietary (4,700 h London fleet data)
ADriver-I (Jia et al., 2023)	2023	MEGVII; Waseda U.; USTC; Mach Drive	Image (vision-action pairs)	Transformer	Driving	nuScenes; private fleet dataset
IRIS (Micheli et al., 2023)	2023	U. Geneva	Image (raw pixels)	Transformer	General	Atari 100k
DriveDreamer (Wang et al., 2023)	2023	GigaAI; Tsinghua U.	Image + BEV/HD-map	Diffusion	Driving	nuScenes
UniSim (Yang et al., 2024a)	2024	UC Berkeley; Google DeepMind; MIT; U. Alberta	Image (raw pixels)	Diffusion	General	Habitat/HM3D; Language Table; Bridge; RT-1/RT-2; Ego4D; EPIC-KITCHENS; SSv2
V-JEPA (Bardes et al., 2024)	2024	Meta FAIR; Inria; ENS; NYU	Image (video)	JEPA	General	Kinetics-400; Something-Something-v2; ImageNet-1K
Genie (Bruce et al., 2024)	2024	Google DeepMind	Image (raw pixels)	Transformer	General	30,000 h Internet gameplay video (2D platformer games)
Vista (Gao et al., 2024)	2024	HKUST; OpenDriveLab (Shanghai AI Lab); U. Tübingen; HKU	Image (camera video)	Diffusion	Driving	nuScenes
GameNGen (Valevski et al., 2025)	2024	Google Research; Tel Aviv U.; Google DeepMind	Image (raw pixels)	Diffusion	General	DOOM gameplay data
WorldDreamer (Wang et al., 2024)	2024	GigaAI; Tsinghua U.	Image (raw pixels)	Transformer	Driving	nuScenes (700 scenes) + self-collected data
ViDAR (Yang et al., 2024b)	2024	OpenDriveLab (Shanghai AI Lab)	Image $\rightarrow$ 3D Point Cloud	Transformer	Driving	nuScenes
Copilot4D (Zhang et al., 2024)	2024	Waabi; U. Toronto	3D Point Cloud (LiDAR)	Diffusion	Driving	nuScenes; KITTI Odometry; Argoverse2
DriveDreamer-2 (Zhao et al., 2025)	2024	CASIA; GigaAI	Image + BEV/HD-map	Diffusion	Driving	nuScenes
OccWorld (Zheng et al., 2024)	2024	Tsinghua U.	3D Occupancy (voxel)	Transformer	Driving	nuScenes
Cosmos (NVIDIA et al., 2025)	2025	NVIDIA	Image (tokenized video)	Diffusion	General	Large curated pretraining video corpus (platform release); demonstrated on robotic manipulation, autonomous driving, and camera-control tasks
V-JEPA 2 (Assran et al., 2025)	2025	Meta FAIR; Mila	Image (video) + proprioception	JEPA	General	1M+ h internet video; Droid (robot); SSv2; Epic-Kitchens-100
GAIA-2 (Russell et al., 2025)	2025	Wayve	Image (multi-camera, 5–6 cams)	Diffusion	Driving	Proprietary ($\sim$25M video clips)
AD-L-JEPA (Zhu et al., 2025)	2025	NYU	BEV Embedding (LiDAR input)	JEPA	Driving	Waymo; KITTI3D; ONCE
WorldDrive (Gui et al., 2026)	2026	U. Macau; Afari Intelligent Drive	Image (multi-view) + motion fusion	Diffusion	Driving	nuScenes; NAVSIM; NAVSIM-v2
Domain-taxonomy / geometric-grounding supplement (Sections E, F) – not part of the 30-paper architectural core above						
AppDeltaWorld (Xu et al., 2026)	2026	Nanyang Technological U.; UESTC; Renmin U. of China; Xiamen U.	Structured HTML/code (GUI state)	Other (retrieval + code generation)	General (mobile GUI agents)	CMGUIBench-500; AndroidLens; MobileGym; MobileWorld
SkyJEPA (Rao et al., 2026)	2026	UC Berkeley; NYU; Brown U.	Latent embedding (quadrotor rigid-body state)	JEPA	General (aerial robotics)	Automated domain-randomized simulation pipeline; real-world outdoor closed-loop flight
MoSA (Wang et al., 2026)	2026	HKUST (Guangzhou); CUHK; U. of Hong Kong	3D Gaussian splat / continuum deformation field	Other (physics-informed residual network)	General (robot manipulation, continuum dynamics)	PAC-NeRF; Spring-GS; self-collected real-world multi-view dataset
SA-WAM (Lopetegui Gonzalez et al., 2026)	2026	Inria; PSL Research U.	RGB + depth (geometric latent diffusion)	Diffusion	General (robot manipulation)	RoboCasa; LIBERO-Plus; real-world UR5 arm

B Related Work, Positioning, and Corpus Methodology

This appendix states in full how this survey's 30-paper corpus was assembled and how it positions itself against three adjacent survey literatures and three more directly competing world-model surveys – material condensed to a single citation in Section 1 of the main text for space. Figure 1 appears in the main text (Section 1) rather than being reproduced here a second time.

B.1 Positioning Against Prior Surveys

Twenty-one of the thirty papers analyzed here (70%), and twenty-four (80%) including 2022, were published in 2023 or later, concentrated in the period covered by the JEPA-based, vision-language-grounded, and multi-camera driving systems of Sections D.C–F. A synthesis of this specific intersection – world-model architecture, multi-sensor 3D scene representation, and closed-loop deployment latency – was not possible before this recent concentration of publications existed to synthesize, and, as argued below, does not yet exist in the adjacent survey literatures that predate or sit outside it. Table 1 and Figure 1 make this concentration directly visible.

World models sit at the intersection of at least three literatures large enough to have produced their own surveys, each worth distinguishing from the present one rather than treating the thirty papers here as if no organizing precedent existed.

Moerland et al.'s *Model-based Reinforcement Learning: A Survey* (Moerland et al., 2023) represents surveys of *model-based reinforcement learning*: its stated scope is "a broad conceptual overview of the combination of planning and learning for MDP optimization," from which the RSSM, discrete-token, and value-equivalent objectives of Sections C.2, C.3, and C.6 are drawn. It names no standard RL benchmark beyond Go, chess, and shogi (via AlphaGo Zero/MuZero), and no LiDAR, multi-sensor fusion, BEV representation, occupancy grids, or autonomous driving at all – Table 1's "Scene Representation" column and Section F's BEV/occupancy/point-cloud treatment sit outside its scope because the driving- and robotics-specific papers surveyed here pose a design axis a conceptual, planning-focused treatment of model-based RL does not need to.

Chen et al.'s *End-to-end Autonomous Driving: Challenges and Frontiers* (Chen et al., 2024), reviewing over 270 papers, is a more directly adjacent survey of *end-to-end autonomous driving*. It discusses world models under one subsection, framing them in a single, undifferentiated role – sample-efficiency tools "particularly helpful in driving, as simulators like CARLA are relatively slow." It names none of GAIA-1, DriveDreamer, or the other diffusion-based driving world models of Section D, and collapses the finer deployment roles Sections G.A–D separate – simulator/data-engine, planner, pretraining backbone, unified perception-action stack – into one observation that "world model learning for end-to-end autonomous driving is an emerging and promising direction." Formalizing the imagination loop's two mechanisms (Section C.7) and separating them from the simulator/pretraining roles is this survey's point of emphasis added to that literature, not a correction of it.

Xing et al.'s *A Survey on Video Diffusion Models* (Xing et al., 2023) represents surveys of *generative video and diffusion models*, covering "video generation, video editing, and other video understanding tasks." Its metrics – FID, SSIM, CLIPSIM, FVD, Video Inception Score – are exactly the visual/distributional-fidelity family Section H identifies as one of several incommensurable evaluation families in this literature, but it treats action-conditioning, closed-loop control, and real-time latency not at all (action labels there are dataset categories, e.g. UCF-101's "101 action categories," not a control concern). It is a natural home for the diffusion-based driving world models surveyed here (GAIA-1, GAIA-2, DriveDreamer, DriveDreamer-2, Vista, WorldDreamer), but the representational-richness/real-time tension Section I argues is the field's central open problem falls outside its evaluation axes.

Three further surveys, published closer in time to the present submission and covering more directly overlapping ground, are addressed separately from the three adjacent-literature surveys above, since each is itself a world-model survey rather than a survey of a neighboring topic.

Feng et al.'s *A Survey of World Models for Autonomous Driving* (Feng et al., 2025) organizes the driving-world-model literature into three tiers – generation of the future physical world (image, BEV, occupancy, and point-cloud prediction), behavior planning, and prediction-planning interaction –

and states its contribution as "a technical roadmap" for the field rather than a claim of novelty over prior work. Its scope is autonomous driving alone: gaming (Section D.B), GUI/agent world models (Section E), and JEPA-style joint-embedding prediction (Section D.C) fall outside its taxonomy, which organizes by pipeline stage rather than by architecture family or training objective.

Tu et al.'s *The Role of World Models in Shaping Autonomous Driving* (Tu et al., 2025) likewise scopes to "Driving World Models," organizing by predicted output modality (video, point cloud, occupancy, latent features, traffic maps) and pairing this with a systematic review of the surrounding simulator, dataset, and metric ecosystem – a depth of driving-specific infrastructure coverage this survey does not attempt to match. Its taxonomy is modality-of-output rather than the training-objective and latency-fidelity axes Sections C and I use, and, like Feng et al., its corpus is driving-only.

Hou et al.'s *World Model for Robot Learning: A Comprehensive Survey* (Hou et al., 2026) is the closest in stated ambition to the present survey – explicitly a cross-cutting synthesis of "disparate architectures, functional roles, and embodied application domains," organized around how a world model couples to a policy (simulator, planning, and foundation-model roles), with autonomous driving and manipulation both treated as application domains under one robot-learning perspective. It differs in axis rather than in ambition: its organizing dimension is policy-integration role, not the latency-fidelity trade-off that Table 1 and Figure 4 make this survey's central empirical throughline, and its stated scope does not itemize gaming or the JEPA family as first-class categories alongside driving and manipulation, whereas Sections D.C and G treat JEPA's representation-pretraining role as one of four deployment roles on equal footing with simulation, planning, and unified perception-action stacks.

This survey's contribution, relative to all six surveys discussed above, is integration along a specific, empirically-grounded axis rather than depth in any single dimension or domain. Relative to the three adjacent-literature surveys, it applies a conceptual model-based-RL formalism (Section C) to the multi-sensor 3D scene representations an autonomous-driving survey covers but does not functionally differentiate (Sections F and G), while holding the visual-fidelity metrics a generative-video survey catalogs accountable to the closed-loop, real-time deployment question a physical system must answer (Sections H and I). Relative to the three world-model-specific surveys just discussed – each scoped to driving alone, or organized by output modality or policy-integration role – its distinguishing move is to hold four domains (gaming, GUI agents, robotics, and autonomous driving) and five training objectives to one shared POMDP formalism and one shared empirical throughline: the latency-fidelity trade-off, quantified directly from each surveyed system's own reported throughput and fidelity figures (Section H), which none of the six surveys discussed here use as their organizing argument.

B.2 Literature Search and Selection Methodology

This subsection states plainly how the thirty papers analyzed here were identified and selected. This survey is a narrative and architectural synthesis organized around four axes – architecture family, scene representation, downstream functional role, and evaluation protocol, synthesized together in Section 6 – not a systematic review in the PRISMA sense; it makes no claim to exhaustive, database-logged coverage of every qualifying paper.

Search process. The corpus was assembled through expert-guided, citation-informed curation rather than a single systematic Boolean query. An initial candidate set was seeded from the field's foundational touchstones – Ha and Schmidhuber's *World Models* (Ha & Schmidhuber, 2018b;a), the PlaNet/Dreamer RSSM line (Hafner et al., 2019; 2020), MuZero (Schrittwieser et al., 2020), and LeCun's JEPA position paper (LeCun, 2022) – and expanded via forward/backward citation-chaining and by tracking arXiv and the NeurIPS/ICML/ICLR/CVPR proceedings. This was not logged as a set of dated, reproducible query strings against IEEE Xplore, Scopus, or a similar database, which is a methodological limitation relative to a fully systematic review, stated here rather than implied to be more rigorous than it was.

Inclusion criteria. A candidate was included if it satisfied all of: (i) it presents or substantively extends a world model in the sense of Section C – a learned component predicting how a latent or explicit environment representation evolves; (ii) it reports an empirical evaluation, with the sole exception of LeCun's JEPA position paper (LeCun, 2022), retained despite being non-empirical as

the explicit conceptual proposal several later empirical papers (I-JEPA, V-JEPA, V-JEPA 2, AD-L-JEPA) directly instantiate; (iii) a full technical paper is available, enabling the same quote-level grounding applied to every claim in this survey, rather than a product announcement or abstract alone; and (iv) publication falls within 2018–2026, chosen because 2018 marks Ha and Schmidhuber's original World Models architecture, this survey's starting point.

Exclusion criteria. Candidates were excluded for failing any criterion above – most commonly for lacking a full technical paper at the depth needed for grounded extraction. OpenAI's Sora, DeepMind's Genie 2, and DeepMind's Genie 3 (announced August 2025) were excluded on this basis: as of this writing, DeepMind has published a blog announcement and a citable BibTeX record for Genie 3, but no accompanying technical report or arXiv paper, so the quote-level grounding this survey applies to every other system is not obtainable for it – the same criterion, not a double standard, that excluded Sora and Genie 2 earlier. Beyond these three named cases, no dated log of every candidate considered and excluded was kept, part of the systematicity gap the next paragraph discloses. Prior survey and review articles were deliberately excluded from this primary corpus – not out of scope, but to avoid conflating primary contributions with meta-analyses of them – and are instead discussed on their own terms in Section B.1 above.

Acknowledged bias. Because the search relied on citation-chaining from a small seed set and on tracking a handful of major venues rather than an exhaustive database query, the resulting corpus is visibly skewed toward highly-cited, frontier-industrial-lab-authored work (Google DeepMind, Meta FAIR, Wayve, and similar organizations account for a disproportionate share relative to smaller academic labs or non-English-language venues) – a real limitation on representativeness, not merely a formality to disclose: a comparably capable world model from a venue or region this search did not reach would not appear here regardless of merit. Extending this into a fully systematic review – dated, logged search strings against arXiv, IEEE Xplore, and the ACM Digital Library, with a PRISMA-style screened/excluded/included record – is left as future work.

B.3 Survey Scope

This survey covers thirty papers in four families tracing the architectural trajectory of world models from 2018 to 2026. The first comprises the foundational recurrent and variational world models establishing the learn-a-latent-dynamics-model-then-plan paradigm: the original World Models architecture (Ha & Schmidhuber, 2018b;a), the RSSM Dreamer line (Hafner et al., 2019; 2020; 2021; 2023), DayDreamer's extension to physical robots (Wu et al., 2022), and MuZero's value-equivalent planning model (Schrittwieser et al., 2020). The second covers the shift to Transformer- and diffusion-based architectures for sequential decision-making and open-ended video/driving-scene generation (Micheli et al., 2023; Bruce et al., 2024; Valevski et al., 2025; Yang et al., 2024a; NVIDIA et al., 2025). The third is the JEPA line, from LeCun's original proposal through its instantiation across images, video, and automotive LiDAR perception (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025; Zhu et al., 2025). The fourth comprises world models purpose-built for autonomous driving and robotics perception, spanning bird's-eye-view, point-cloud, and occupancy-grid scene representations (Hu et al., 2022; 2023; Russell et al., 2025; Wang et al., 2023; Zhao et al., 2025; Wang et al., 2024; Gao et al., 2024; Zheng et al., 2024; Zhang et al., 2024; Yang et al., 2024b; Jia et al., 2023; Gui et al., 2026). The first three families are defined by architecture; the fourth by application domain, and *deliberately* overlaps with them – Copilot4D and AD-L-JEPA, diffusion- and JEPA-architected respectively (Table 1), are cross-listed for their driving/robotics focus rather than silently duplicated by oversight.

Table 1 and Table 2 summarize this corpus along the axes above; the coverage matrix (Table 3) extends the comparison with each paper's affiliation and dataset for full reproducibility. Beyond this 30-paper core, five further 2026 papers are drawn on later: three (AppDeltaWorld, Sky-JEPA, MoSA) instantiate the domain-specific transition kernels of Section E, a fourth (SA-WAM) grounds the geometric-representation argument of Section F, and a fifth (ODEWorld) is cited exclusively in Section I's argued-position subsection on continuous-time dynamics. The first four are listed, clearly demarcated, in a supplementary block of Table 3; ODEWorld is omitted from that reproducibility-oriented matrix since it supports an argued position rather than a system with comparable dataset/benchmark results. Together these 35 papers are the full set this survey draws on; Section I's occasional references to a "35-paper corpus" mean this larger set, not the 30-paper architectural core tabulated above.

C Background and Problem Formulation

Section 1 introduced "learn a model, then plan or act inside it" as the throughline connecting the thirty papers surveyed here. This section makes that throughline precise: it states the shared control-theoretic problem every surveyed world model approximately solves, formalizes the five distinct training objectives used to fit that approximation – variational latent dynamics, discrete autoregressive token prediction, denoising score matching, joint-embedding feature prediction, and value-equivalent dynamics – each with the notation and equations the corresponding papers report themselves, and closes by formalizing the *imagination loop*: the mechanism by which a trained world model's latent rollouts are converted into a behavior, either by backpropagating value gradients through imagined trajectories or by searching over them with model-predictive control.

C.1 The World Model as a POMDP

PlaNet and Dreamer both state their problem setup as a *partially observable Markov decision process* (POMDP) (Hafner et al., 2019; 2020): an agent does not observe the true state of its environment directly, only a high-dimensional observation (an image, a LiDAR sweep, a multi-camera frame) generated from that state. Formally, a POMDP is the tuple

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, R, \Omega, O, \gamma)\,, \tag{9}$$

where $\mathcal{S}$ is the space of true environment states $s \in \mathcal{S}$, never directly observed; $\mathcal{A}$ the action space; $T(s_{t+1} \mid s_t, a_t)$ the transition kernel; $R(s_t, a_t)$ the reward function; Ω the observation space $o \in \Omega$ (pixels, point clouds, or a fused multi-sensor frame, depending on the paper); $O(o_t \mid s_t)$ the observation function; and $\gamma \in [0, 1)$ a discount factor. Following Dreamer's notation, the agent's task reduces to learning a policy $a_t \sim p(a_t \mid o_{\le t}, a_{<t})$ maximizing the expected discounted return $\mathbb{E}_p\left[\sum_{t=1}^{T} \gamma^t r_t\right]$ under the unknown dynamics $o_t, r_t \sim p(o_t, r_t \mid o_{<t}, a_{<t})$.

A *world model*, in the sense used throughout this survey, is a set of learned functions $\hat{T}_\theta$, $\hat{O}_\theta$, $\hat{R}_\theta$ approximating the true, unknown transition, observation, and reward mappings of Eq. (9) from a dataset of past experience $\mathcal{D} = \{(o_t, a_t, r_t)\}_{t=1}^{T}$, typically via a compact latent state s_t (or, in the token- and representation-space families, a discrete token sequence or embedding) standing in for the true, unobserved $s \in \mathcal{S}$. Every family in Section D is a different parametric form for $\hat{T}_\theta$, $\hat{O}_\theta$, $\hat{R}_\theta$ and a different training objective for fitting θ to $\mathcal{D}$, though not every family instantiates all three: $\hat{R}_\theta$ matters only to planning-oriented families that consume it to select actions (Section G), while the generative simulator and representation-pretraining architectures – GAIA-1, GAIA-2, DriveDreamer, Genie, GameNGen, Vista, WorldDreamer – define no reward model and approximate only $\hat{T}_\theta$ and, where they decode observations, $\hat{O}_\theta$. The JEPA family is a further limiting case: I-JEPA and V-JEPA are the asymptotic case of LeCun's proposal in which the optional latent variable z is absent, so the reward and action elements of Eq. (9) are simply unpopulated at pretraining time, rather than architectures that themselves define a transition kernel over actions and rewards. The next five subsections formalize each choice in turn.

C.2 RSSM Latent Dynamics and the Variational ELBO

PlaNet and the Dreamer line (Hafner et al., 2019; 2020; 2021; 2023) instantiate the world model as a *recurrent state-space model* (RSSM) splitting the latent state into a deterministic component h_t and a stochastic component s_t,

$$\begin{aligned}
\text{Deterministic state:} \quad & h_t = f(h_{t-1}, s_{t-1}, a_{t-1})\,, \\
\text{Stochastic state:} \quad & s_t \sim p_\theta(s_t \mid h_t)\,, \\
\text{Observation model:} \quad & o_t \sim p_\theta(o_t \mid h_t, s_t)\,, \\
\text{Reward model:} \quad & r_t \sim p_\theta(r_t \mid h_t, s_t)\,,
\end{aligned} \tag{10}$$

where f is a recurrent network deterministically propagating $s_{t-1}, h_{t-1}, a_{t-1}$ forward, and $p_\theta(s_t \mid h_t)$ is the learned *prior* over the stochastic state given only the deterministic history – the distribution used for imagining forward without new observations. Because $o_{1:T}$ is only indirectly related to $s_{1:T}$, PlaNet introduces an approximate posterior (encoder) $q_\theta(s_t \mid h_t, o_t)$ additionally conditioned

on the current observation, and trains via a variational evidence lower bound (ELBO) obtained by Jensen's inequality,

$$\ln p_\theta(o_{1:T} \mid a_{1:T}) \;\geq\; \sum_{t=1}^{T} \underbrace{\mathbb{E}_{q_\theta(s_t \mid h_t, o_t)}[\ln p_\theta(o_t \mid h_t, s_t)]}_{\text{reconstruction}} - \sum_{t=1}^{T} \underbrace{\mathbb{E}_{q_{t-1}}[\mathrm{KL}[q_\theta(s_t \mid h_t, o_t) \| p_\theta(s_t \mid h_t)]]}_{\text{complexity}}, \tag{11}$$

where $q_{t-1} \triangleq q_\theta(s_{t-1} \mid h_{t-1}, o_{t-1})$, with an analogous reward-reconstruction term added by the same argument. The reconstruction term rewards the posterior for encoding enough information about o_t to reconstruct it; the KL term regularizes the posterior toward the prior $p_\theta(s_t \mid h_t)$ – exactly the distribution used for prediction without observations, the mechanism letting a model trained this way later imagine forward in latent space alone.

Dreamer's own presentation of the identical model reverses this notation: it writes a representation model $p_\theta(s_t \mid s_{t-1}, a_{t-1}, o_t)$, a transition model $q_\theta(s_t \mid s_{t-1}, a_{t-1})$, and a reward model $q_\theta(r_t \mid s_t)$, reserving p for the distribution that "generates samples in the real environment" and q for the prior/transition model that "enable[s] latent imagination" (Hafner et al., 2020) – tracking each distribution's role in imagination rather than its role as posterior or prior. This is reproduced here exactly as Dreamer defines it, but is a single, self-contained exception confined to this paragraph: everywhere else in this survey, including Eq. 11 above and every other p_θ/q_θ pair below, q_θ denotes the observation-conditioned approximate posterior (encoder) and p_θ the prior/generative model, matching the standard VAE convention of Kingma and Welling.

C.3 Discrete Autoregressive Token Prediction

IRIS and Genie (Micheli et al., 2023; Bruce et al., 2024) instead represent the scene as a sequence of discrete tokens and model dynamics with an autoregressive Transformer. IRIS first learns a discrete autoencoder (E, D): the encoder $E : \mathbb{R}^{h \times w \times 3} \to \{1, \ldots, N\}^K$ maps an input frame x_t to K tokens from a vocabulary of size N, via nearest-neighbor lookup $z_t^k = \arg\min_i \|y_t^k - e_i\|_2$ against a learned embedding table $\mathcal{E} = \{e_i\}_{i=1}^N$; the decoder D maps the K tokens back to a reconstructed frame. A GPT-style autoregressive Transformer G then models dynamics entirely at the token level, factorizing each future frame's tokens autoregressively as

$$\hat{z}_{t+1}^k \sim p_G\big(\hat{z}_{t+1}^k \;\big|\; z_{\leq t}, a_{\leq t}, z_{t+1}^{<k}\big), \qquad k = 1, \ldots, K, \tag{12}$$

with reward and termination heads $\hat{r}_t \sim p_G(\hat{r}_t \mid z_{\leq t}, a_{\leq t})$ and $\hat{d}_t \sim p_G(\hat{d}_t \mid z_{\leq t}, a_{\leq t})$, where $z_{t+1}^{<k}$ denotes the tokens of frame $t{+}1$ already predicted earlier in the autoregressive order. G is trained with a cross-entropy loss between predicted and true next-frame tokens (cross-entropy or MSE for the reward head, depending on its type). Genie applies the same recipe at larger scale and without ground-truth action labels, learning a latent action model jointly with the token dynamics so a user can control the rollout through an inferred action interface.

C.4 Denoising Score Matching for Diffusion World Models

DriveDreamer and Copilot4D (Wang et al., 2023; Zhang et al., 2024) instead cast dynamics prediction as conditional generation with a diffusion model. DriveDreamer's Auto-DM component defines a forward process gradually corrupting a clean latent feature $\mathcal{Z}_0$ with Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to produce a noised latent $\mathcal{Z}_t$, and trains a denoising network ϵ_ϕ to recover the added noise from $\mathcal{Z}_t$, the diffusion timestep t, and a structured driving-scene condition c (an HD map, 3D boxes, or, in DriveDreamer-2, an LLM-generated scene description). The training objective is

$$\min_\phi \; \mathcal{L} = \mathbb{E}_{\mathcal{Z}_0,\, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\, t,\, c}\left[\|\epsilon - \epsilon_\phi(\mathcal{Z}_t, t, c)\|_2^2\right], \tag{13}$$

with t sampled uniformly from $[1, T]$ per training step. At sampling time, the future scene is generated by iteratively applying the learned reverse (denoising) process from pure noise, conditioned on c – for a world model, the mechanism turning an action or scene-layout condition into a predicted

future observation. Copilot4D applies the same denoising principle to a *discrete* token space: it tokenizes point-cloud observations with a VQVAE, then recasts MaskGIT's masked-token-prediction as a discrete diffusion process, decoding and denoising many tokens in parallel rather than one at a time as in Eq. (12), while retaining the same denoising character as Eq. (13) applied to categorical rather than Gaussian corruption.

C.5 JEPA Feature Prediction with EMA Target Encoders

I-JEPA and V-JEPA (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025) depart from all three objectives above by predicting in representation space rather than reconstructing pixels, tokens, or noise. The architecture has three networks: a context encoder f_θ, a target encoder $f_{\bar{\theta}}$, and a predictor g_ϕ. Given an input signal y (an image or video clip), I-JEPA splits it into N non-overlapping patches and computes target representations $s_y = f_{\bar{\theta}}(y)$ by passing the *full* signal through the target encoder. A context block x (a subset, with target-overlapping patches removed) is passed through the context encoder f_θ, and the predictor g_ϕ – conditioned on a mask token carrying the target block's positional embedding, playing the role of the latent variable z in LeCun's general JEPA formulation (LeCun, 2022) – produces a prediction $\hat{s}_y(i)$ for each of M sampled target blocks B_i. The training loss is the average L_2 distance between predicted and target patch representations,

$$\begin{aligned} \mathcal{L}(\theta, \phi) &= \frac{1}{M} \sum_{i=1}^{M} D(\hat{s}_y(i), s_y(i)) \\ &= \frac{1}{M} \sum_{i=1}^{M} \sum_{j \in B_i} \left\| \hat{s}_{y_j} - s_{y_j} \right\|_2^2 . \end{aligned} \tag{14}$$

Only ϕ and θ are updated by gradient descent on Eq. (14); the target-encoder parameters $\bar{\theta}$ receive no gradient and are instead updated after each step as an exponential moving average (EMA) of θ,

$$\bar{\theta} \leftarrow \lambda \bar{\theta} + (1 - \lambda)\theta, \qquad \lambda \in [0, 1), \tag{15}$$

a design the I-JEPA authors identify as essential for avoiding representation collapse – the degenerate solution in which the encoders output a constant vector, trivially minimizing Eq. (14). V-JEPA and V-JEPA 2 apply the same encoder/predictor/EMA structure to video clips, predicting masked spatiotemporal tubes rather than masked spatial patches, and V-JEPA 2-AC further conditions the predictor on an action, so g_ϕ becomes an action-conditioned latent dynamics model directly comparable to the RSSM transition model of Eq. (10), but operating entirely in representation space with no observation or reward reconstruction term.

C.6 Value-Equivalent Dynamics

MuZero (Schrittwieser et al., 2020) departs from all four objectives above in a different way: rather than training its latent state to reconstruct observations, tokens, denoised features, or held-out representations, it trains the latent state to be *value-equivalent* – sufficient to predict future policy, value, and reward targets, with no requirement that it resemble the true environment state at all. The model has three jointly trained functions: a representation function h_θ encoding past observations into an initial hidden state, $s^0 = h_\theta(o_1, \ldots, o_t)$; a dynamics function g_θ producing an immediate reward and next hidden state, $r^k, s^k = g_\theta(s^{k-1}, a^k)$; and a prediction function f_θ computing a policy and value estimate, $p^k, v^k = f_\theta(s^k)$. The hidden state s^k carries no semantics of the true environment state – its only role is supporting accurate prediction of p^k, v^k, r^k. Given a trajectory sampled from a replay buffer, the model is unrolled for K hypothetical steps and trained by minimizing, at every real time step t,

$$\begin{aligned} l_t(\theta) = \sum_{k=0}^{K} \Big[& l_r\big(u_{t+k}, r_t^k\big) + l_v\big(z_{t+k}, v_t^k\big) \\ & + l_p\big(\pi_{t+k}, p_t^k\big) \Big] + c\|\theta\|^2, \end{aligned} \tag{16}$$

where u_{t+k} is the true observed reward; π_{t+k} the improved policy target from Monte Carlo tree search (MCTS) conditioned on the model's own p^k, v^k, r^k; z_{t+k} a bootstrapped n-step return, $z_t = u_{t+1} + \gamma u_{t+2} + \cdots + \gamma^{n-1} u_{t+n} + \gamma^n \nu_{t+n}$, with ν the MCTS-estimated value; l_r, l_v, l_p the reward,

value, and policy loss terms; and $c\|\theta\|^2$ an L_2 regularization term. At action-selection time, MuZero performs MCTS over the *learned* model $(h_\theta, g_\theta, f_\theta)$ rather than the real environment, selecting a_{t+1} by sampling from the resulting search policy π_t – an instance of the search-based planning branch of the imagination loop formalized next.

C.7 The Imagination Loop: Latent Rollouts for Policy Gradients and Model-Predictive Control

Every objective above yields a learned, latent-space approximation of the environment's dynamics; none, on its own, produces a behavior. Across the papers surveyed here, a trained world model is converted into a behavior by one of three primary mechanisms, the first two operating by rolling the model forward in latent space rather than interacting with the real environment.

Gradient-based imagination. Dreamer's actor-critic defines an *imagination environment*: a fully observed MDP induced by the latent dynamics of Eq. (10), in which imagined trajectories $\{(s_\tau, a_\tau)\}_{\tau=t}^{t+H}$ of horizon H are unrolled from a real model state s_t, following $s_\tau \sim q_\theta(s_\tau \mid s_{\tau-1}, a_{\tau-1})$, $r_\tau \sim q_\theta(r_\tau \mid s_\tau)$, and a stochastic policy $a_\tau \sim q_\phi(a_\tau \mid s_\tau)$. A value model $v_\psi(s_\tau)$ is trained to regress a λ-return $V_\lambda(s_\tau)$ over the imagined trajectory, while the policy parameters ϕ are updated by propagating the *analytic gradient* of the summed value estimate directly back through the differentiable latent dynamics,

$$\begin{aligned} \phi &\leftarrow \phi + \alpha \nabla_\phi \sum_{\tau=t}^{t+H} V_\lambda(s_\tau), \\ \psi &\leftarrow \psi - \alpha \nabla_\psi \sum_{\tau=t}^{t+H} \tfrac{1}{2}\big(v_\psi(s_\tau) - V_\lambda(s_\tau)\big)^2. \end{aligned} \tag{17}$$

Because the reparameterized transition $q_\theta(s_\tau \mid s_{\tau-1}, a_{\tau-1})$ is differentiable with respect to $a_{\tau-1}$, the gradient of a multi-step imagined return flows all the way back through every imagined transition to the policy, without a single additional interaction with the real environment – exactly the property Ha and Schmidhuber's original controller exploited when trained "entirely inside of its own hallucinated dream" (Section G), and the property DayDreamer (Wu et al., 2022) exploits to train and act concurrently on physical robot hardware.

Search-based model-predictive control. PlaNet and MuZero instead treat the same latent rollout as a simulator to search over rather than a differentiable path to backpropagate through. PlaNet uses the cross-entropy method (CEM) to optimize an action sequence $a_{t:t+H}$ against predicted cumulative reward in latent space via Eq. (10), executes only the first action, and replans at the next step – the standard receding-horizon structure of MPC. MuZero replaces CEM with Monte Carlo tree search guided by the learned policy and value heads $p^k, v^k = f_\theta(s^k)$ of Eq. (16), searching over the value-equivalent latent space induced by g_θ rather than a space trained to reconstruct observations. In both mechanisms, the property making the search tractable is the same one making gradient-based imagination efficient: because the world model predicts directly in a compact latent space, thousands of candidate trajectories can be evaluated or differentiated through without ever rendering an image or querying the real environment.

Conditional generative decoding and direct policy heads. A third mechanism covers systems whose inference pathway is neither gradient-based imagination nor search-based MPC. The Transformer- and diffusion-based generative simulators surveyed in Section D – Genie, GAIA-1, GameNGen, UniSim (Bruce et al., 2024; Hu et al., 2023; Valevski et al., 2025; Yang et al., 2024a) – convert a trained model into behavior by conditional autoregressive decoding or iterative denoising: queried with an action or control signal, the model directly samples the next observation, with no imagined-value estimate or explicit search interposed at all. ADriver-I (Jia et al., 2023) is a further variant, outputting control tokens directly from a fine-tuned multimodal LLM. This taxonomy still does not exhaust every inference pathway this survey covers: I-JEPA and V-JEPA (Assran et al., 2023; Bardes et al., 2024) serve, at pretraining time, as frozen representation backbones with no behavior-selection mechanism intrinsic to the world model itself, and require a downstream, task-specific inference head trained separately to produce any action, control token, or planning output at all. Section D traces how each subsequent architecture family inherits, and in the Transformer-, diffusion-, and JEPA-based cases substantially modifies, this same underlying imagination loop.

D Architectural Evolution: From Recurrent Latent Dynamics to Joint-Embedding Prediction

Across the architecture families traced in this section, nearly every surveyed system instantiates the same three-stage pipeline: a *perception* stage that encodes raw sensor input into a scene representation, a *world model* stage that predicts how that representation evolves over time, and a *behavior or planning* stage that consumes the predicted representation to select an action – depicted in Figure 2 (Section 2, alongside the POMDP formalism this pipeline instantiates).

D.1 Recurrent and Variational World Models (2018–2022)

The earliest models represent a scene as a low-dimensional latent vector from a variational encoder and model temporal dynamics with a recurrent network. Ha and Schmidhuber's *World Models* (Ha & Schmidhuber, 2018b;a) decompose the agent into a VAE (V), a mixture-density recurrent network (M) modeling $P(z_{t+1} \mid z_t, a_t, h_t)$, and a small linear controller (C); the key result is that C can be trained *entirely inside rollouts generated by M* – "inside of its own hallucinated dream" – and still transfer to the real environment, establishing the train-in-imagination pattern every later recurrent world model here reuses. PlaNet (Hafner et al., 2019) introduces the Recurrent State-Space Model (RSSM), pairing both a deterministic and a stochastic latent component – reported as "crucial for high planning performance" – with online cross-entropy-method planning. Dreamer (Hafner et al., 2020) keeps the RSSM but learns an actor-critic by "propagating analytic gradients of learned state values back through trajectories imagined in the compact state space." DreamerV2 (Hafner et al., 2021) switches to discrete categorical latents, an ablated design choice the authors report as central to accuracy. DreamerV3 (Hafner et al., 2023) adds normalization and balancing techniques letting one fixed configuration solve "over 150 diverse tasks," including collecting a diamond in Minecraft from raw pixels and sparse reward. DayDreamer (Wu et al., 2022) runs the same algorithm online on physical robots, with no simulator in the loop, directly addressing the authors' concern that simulator-only training "fails to capture the complexity of the real world."

A parallel line replaces the RSSM with tree search over model-predicted quantities. MuZero (Schrittwieser et al., 2020) trains a model to predict only reward, policy, and value – never reconstructing observations – and plans via Monte Carlo tree search, reaching superhuman Atari, Go, chess, and shogi performance "without any knowledge of their underlying dynamics." MuZero is evaluated only on discrete-action games, leaving open whether value-equivalent planning transfers to continuous-action robotics; TD-MPC and TD-MPC2 (Hansen et al., 2022; 2024) close that gap, pairing the same value-equivalent principle with a receding-horizon optimizer rather than MCTS. TD-MPC2 reports "104 online RL tasks spanning" four domains "with a single set of hyperparameters," and separately trains "a single 317M parameter agent to perform 80 tasks across multiple task domains, embodiments, and action spaces" – direct evidence that value-equivalent planning scales to the continuous-control regime MuZero does not test. Like MuZero, TD-MPC2 is an "implicit (decoder-free) world model," never reconstructing observations; whether this yields a latency advantage over the reconstruction-based families below is evidence this survey has not found directly reported, and Section I treats it as evidenced by architecture rather than measured comparison.

In driving, MILE (Hu et al., 2022) adapts the recurrent template to a bird's-eye-view scene representation, learning a world model and an imitation-learning policy jointly and offline, using "3D geometry as an inductive bias." Trained with no online interaction, MILE improves CARLA driving score by 31% over the prior state of the art in a held-out town and weather condition, and decodes predicted states directly into interpretable BEV segmentation maps.

D.2 Transformer and Diffusion World Models (2023–2024)

Around 2023 the recurrent core is replaced by an autoregressive Transformer or diffusion process, and application broadens to open-ended video and driving-scene generation. IRIS (Micheli et al., 2023) replaces the RSSM with "a discrete autoencoder and an autoregressive Transformer," reaching a mean human-normalized Atari 100k score of 1.046 from about two hours of gameplay per game. Genie (Bruce et al., 2024) scales this to an 11B-parameter, action-*free* model that learns a latent action interface purely from internet video. GameNGen (Valevski et al., 2025) builds an interactive diffusion-based (augmented Stable Diffusion) neural game engine, the first reported to run in real

time – 20 FPS, next-frame PSNR of 29.4, comparable to lossy JPEG – while remaining playable over multi-minute DOOM sessions.

In driving, GAIA-1 (Hu et al., 2023) models discrete video/text/action tokens with a Transformer plus diffusion decoder, explicitly stated as not real-time. DriveDreamer and DriveDreamer-2 (Wang et al., 2023; Zhao et al., 2025) condition diffusion on HD maps/BEV layouts and, in the second version, an LLM-generated scene description. Vista (Gao et al., 2024) improves FID by 55% and FVD by 27% over the prior best driving world model and doubles as a learned reward function. WorldDreamer (Wang et al., 2024) proposes a masked-token objective framed as general-purpose but evaluated only on the driving-domain nuScenes dataset. GAIA-2 (Russell et al., 2025) moves to a latent diffusion model unifying multi-agent interaction and multi-camera consistency across five to six cameras and geographically diverse environments (UK, US, Germany); like GAIA-1, not yet real-time. ViDAR (Yang et al., 2024b) forecasts future 3D point clouds as a self-supervised pretraining signal that transfers to detection, occupancy, and planning. Copilot4D (Zhang et al., 2024) tokenizes LiDAR point clouds and applies discrete diffusion in token space. OccWorld (Zheng et al., 2024) models a 3D semantic occupancy grid, reporting competitive planning *without* instance-level or map supervision. ADriver-I (Jia et al., 2023) fine-tunes a multimodal LLM to predict control signals directly, paired with a video-diffusion decoder, evaluated on both public nuScenes and a private fleet dataset.

D.3 Joint-Embedding Predictive Architectures (JEPA)

LeCun's position paper (LeCun, 2022) proposes the Joint-Embedding Predictive Architecture (JEPA): two encoders map inputs x, y to representations s_x, s_y, a predictor predicts s_y from s_x, and training minimizes prediction error *in representation space* rather than reconstructing y – JEPA "is not generative... it merely captures the dependencies between x and y."

I-JEPA (Assran et al., 2023) instantiates this for static images, predicting masked target-block representations from a visible context block. V-JEPA (Bardes et al., 2024) extends the same objective to video, "solely using a feature prediction objective, without... pretrained image encoders, text, negative examples, reconstruction, or other sources of supervision." V-JEPA 2 (Assran et al., 2025) scales to over a million hours of video plus robot interaction data, reports state-of-the-art action anticipation on Epic-Kitchens-100, and – one of the few JEPA-family papers to report a latency figure at all – 4–16 FPS for downstream robot action inference. AD-L-JEPA (Zhu et al., 2025) adapts the objective to LiDAR, predicting bird's-eye-view embeddings ("neither generative nor contrastive") for a 3D detection mAP gain of 1.61–2.98 points on ONCE at substantially lower GPU cost than the prior pretraining state of the art.

D.4 Vision-Language Models and Spatial Grounding

A further shift, concentrated in the most recent driving papers, couples the dynamics core to a pretrained vision-language backbone. ADriver-I (Jia et al., 2023) fine-tunes a multimodal LLM (built on Vicuna) as the world model's core, predicting steering and speed directly from interleaved vision/action tokens, paired with a video diffusion model that renders the resulting scene. V-JEPA 2 (Assran et al., 2025) is evaluated against vision-language-action baselines and general-purpose VLMs (InternVL 2.5, Qwen2VL, Qwen2.5VL) on video-language reasoning benchmarks, and despite no language supervision is competitive with or exceeds several of them. UniSim (Yang et al., 2024a) uses a learned video world model to generate training experience for both vision-language and low-level control policies. Together these point to a convergence: VLM backbones fine-tuned directly for control (ADriver-I), used as baselines/consumers for representation-learning world models (V-JEPA 2, UniSim), and evaluated on driving benchmarks such as nuScenes rather than only general-purpose corpora.

D.5 Multi-Modal Scene Representation: Sensor Fusion

Scene representation correlates with sensing modality. Camera-only systems – MILE, the GAIA line, the DriveDreamer line (Hu et al., 2022; 2023; Russell et al., 2025; Wang et al., 2023; Zhao et al., 2025) – decode into BEV or multi-camera video without a second modality. LiDAR-only systems take the complementary path: Copilot4D and ViDAR (Zhang et al., 2024; Yang et al., 2024b)

forecast future point clouds, and AD-L-JEPA (Zhu et al., 2025) predicts LiDAR BEV embeddings in representation space. DayDreamer (Wu et al., 2022) fuses "proprioceptive and camera inputs" within a single RSSM encoder for online, simulator-free deployment.

None of the thirty papers adopts an explicit LiDAR-camera fusion or panoptic-segmentation front end, even though several cite this class of technique as related work (MILE, DriveDreamer, ViDAR). This gap is notable given the stochastic-latent design PlaNet and MILE already share, motivated explicitly to avoid "overconfident predictions" and "capture multiple possible futures"; a fused, panoptic-aware front end combining LiDAR geometric precision, camera semantic density, and instance-level uncertainty would extend that same design to a strictly richer observation space than any camera-only or LiDAR-only system here currently uses – an open direction, not a solved component of the current literature.

D.6 Convergence: Unifying Scene and Motion Representation

WorldDrive (Gui et al., 2026), the most recent driving system surveyed, argues that existing driving world models "primarily focus on visual scene representation," while motion representation is "not explicitly designed to be planner-shared and inheritable." Its trajectory-aware diffusion Transformer unifies vision and motion representation in one model, reporting 53 ms end-to-end latency against baselines the authors cite at 570–850 ms, evaluated on nuScenes and NAVSIM/NAVSIM-v2 with composite planning metrics.

WorldDrive is not the only system moving toward this unification, though it is the most complete to date: MILE already reports "complex driving manoeuvres" executed "from plans entirely predicted in imagination" (Hu et al., 2022), and OccWorld reports competitive planning from a 3D occupancy world model without instance-level or map supervision (Zheng et al., 2024). WorldDrive's contribution is best read as the most complete instance of a direction this corpus had already partially anticipated, not as this survey's sole evidence for it – and, as a 2026 preprint rather than a peer-reviewed publication, its results should be read as promising and single-source pending independent verification.

This convergence is constrained by a latency bottleneck running through every family surveyed: only a minority of the driving/robotics systems in Table 1 report a real-time or near-real-time figure at all – MuZero's 100 ms search budget (Schrittwieser et al., 2020), MILE's 43.0 Hz BEV inference (Hu et al., 2022), GameNGen's 20 FPS (Valevski et al., 2025), V-JEPA 2's 4–16 FPS (Assran et al., 2025), and WorldDrive's 53 ms (Gui et al., 2026) – while the most expressive generative models, GAIA-1 and GAIA-2, are by their own authors' statement not yet real-time (Hu et al., 2023; Russell et al., 2025), and several further systems report no latency figure at all. This representational-richness/real-time-viability trade-off, recurring across every family from recurrent latents through diffusion to JEPA, is the bottleneck the remainder of this survey confronts. Section E next asks a more basic question: whether the architectures surveyed above, together with a further set of physics-informed and geometry-focused 2026 systems, are even trying to learn the *same kind* of transition function across the gaming, AI-agent, and robotics domains they target.

E Domain-Specific Dynamics: Agents, Gaming, and Robotics

Sections D and F treat the world-model backbone and its scene representation as if a single, generic transition function $T(s_{t+1} \mid s_t, a_t)$ were being learned throughout. In practice, the thirty surveyed papers – together with five additional 2026 papers on physics-informed dynamics, 3D geometric grounding, and the simulation-to-reality domain gap identified for this section – span three qualitatively different application domains whose transition functions differ enormously in state-space topology, action semantics, and entropy. The five supplemental papers were each selected *after* the three-domain structure below had already emerged from the thirty core papers' own descriptions of their target systems, specifically because each supplies a concrete mathematical formulation – a state vector, an action space, a governing constraint – where the core papers state the relevant property only in prose. This section makes that difference explicit, domain by domain, before formalizing it mathematically in Section E.4.

E.1 Gaming: Discrete, Bounded, and Programmatically Deterministic

Video games are, at the implementation level, ordinary software: the state a world model must predict is produced by a fixed, hand-authored program rather than physical law. GameNGen's authors make this point directly, observing that "while vastly different game engines exist, the game state updates and rendering logic in all are composed of a set of manual rules, programmed or configured by hand" (Valevski et al., 2025) – a game engine already *is* a transition function, and the world model's job is to imitate an existing deterministic mapping rather than discover an unknown physical one. GameNGen frames its own neural simulator the same way, running the "game-loop (update state based on input and render it to the screen)" (Valevski et al., 2025) at 20 frames per second while "maintaining a level of visual quality comparable to that of the original game" (Valevski et al., 2025).

This programmatic origin gives gaming transitions two properties not shared by the other domains here. First, the action space is typically small and discrete: MuZero encodes an Atari action "as a one hot vector which is tiled appropriately into planes" (Schrittwieser et al., 2020), noting that "in Atari due to the much smaller action space 50 simulations per move are sufficient" (Schrittwieser et al., 2020) for its search. Second, the transitions themselves are low-entropy, often exactly deterministic: MuZero's dynamics function is "represented deterministically" in its base formulation, crediting "the deterministic nature of the Atari simulator" (Schrittwieser et al., 2020) in its evaluation protocol. Bounded, discrete state spaces and programmatic, low-entropy dynamics are thus not simplifying assumptions for tractability – they are, per the surveyed authors' own descriptions, a reasonably accurate model of what a game engine actually is.

E.2 AI Agents: Semantic State and Interaction Logic, Not Physics

A second domain – world models for web and mobile GUI agents – has no counterpart among the original thirty architectures, motivating the addition of AppDeltaWorld (Xu et al., 2026) for this section. AppDeltaWorld targets mobile GUI agents predicting how an app's interface changes in response to a tap or swipe, and its authors are explicit that the governing rules are semantic and interactive rather than physical: predicting the next screen requires following "the interaction logic in real apps" (Xu et al., 2026), formalized as a *transition-grounded* constraint. Concretely, AppDeltaWorld converts a screenshot x_t together with an app identifier and action m_t into a structured textual state

$$r_t = \phi_{\text{text}}(x_t) = [f_t, d_t, \ell_t, q_t, p_t], \tag{18}$$

where $f_t, d_t, \ell_t, q_t, p_t$ are the functional template, screen description, region layout, slot schema, and page type (Xu et al., 2026) – built entirely from UI structure and text, with no notion of position, velocity, or force. Because the space of valid next screens for a given app/action is finite and enumerable from the app's navigation graph, AppDeltaWorld maintains an explicit "action-transition index" per app: "unsupported actions are treated as invalid or low-confidence transitions rather than unconstrained hallucinated states" (Xu et al., 2026). There is no Newtonian constraint here; plausibility is enforced by membership in a discrete, app-specific reachability graph, making this domain's transition function closer in character to gaming (Section E.1) than to robotics below, but governed by human-authored interface semantics rather than a physics or game engine.

E.3 Robotics and Autonomous Vehicles: Continuous, Stochastic, and Multi-Agent

The robotics and autonomous-driving systems already surveyed (MILE, GAIA-1/2, DriveDreamer, OccWorld, Copilot4D, AD-L-JEPA) occupy the opposite end of this spectrum, and the newly added papers make the contrast concrete. SkyJEPA models a quadrotor as a physical rigid-body vector rather than a rendered frame or UI schema:

$$x_t = \begin{bmatrix} p_t^\top & v_t^\top & r_{x,t}^\top & r_{y,t}^\top & r_{z,t}^\top & \omega_t^\top \end{bmatrix}^\top, \tag{19}$$

with position $p_t \in \mathbb{R}^3$, velocity $v_t \in \mathbb{R}^3$, attitude as the columns $r_{x,t}, r_{y,t}, r_{z,t} \in \mathbb{R}^3$ of a rotation matrix $R_t \in SO(3)$, and angular velocity $\omega_t \in \mathbb{R}^3$, driven by a four-dimensional motor-force action (Rao et al., 2026). Unlike gaming and agents, this state and its dynamics $x_{t+1} = f(x_t, a_t)$ (Rao et al., 2026) are continuous, and the underlying physical process can only be approximated, not

exactly imitated: SkyJEPA's long-horizon prober exists precisely because autoregressive rollouts accumulate "small one-step errors arising from uncertainty, approximation, or inaccuracies" that "accumulate over time, leading to drift, instability, and physically implausible trajectories" (Rao et al., 2026) – compounding error with no analogue against a deterministic game engine or finite UI graph.

This compounding rollout error is an *epistemic* source of uncertainty in the sense of Kendall and Gal (Kendall & Gal, 2017): reducible in principle with a better model, and formally located in f_{robot}'s approximation error in Eq. (8) (Section 3) rather than the covariance term Σ. MoSA documents a second epistemic instance at the level of material physics: even objects a simulator treats as ideally isotropic "typically exhibit small but systematic anisotropy and heterogeneity" (Wang et al., 2026), a gap in the physics *model* – one a purely programmatic engine (Section E.1) systematically fails to capture – rather than genuine randomness, and reducible by a more detailed material model, as MoSA's own residual-stress correction demonstrates. Both belong, formally, in f_{robot}.

Two further sources of entropy are *aleatoric* instead – irreducible, reflecting genuine environmental randomness – and populate the covariance term $\Sigma(s_t, a_t)$ of Eq. (8). The first is physical stochasticity independent of modeling error: MILE evaluates under "a completely new town and new weather conditions" (Hu et al., 2022), across "four different weather conditions (ClearNoon, WetNoon, HardRainNoon, ClearSunset)" (Hu et al., 2022), since a driving world model must remain valid under conditions a game engine would never need to model; GAIA-2 likewise conditions on "environmental factors" among "a diverse set of domain-specific control signals" (Russell et al., 2025). The second is other agents: GAIA-2 lists "multi-agent interactions" among the requirements a generative driving world model must satisfy that a single-agent game or GUI transition does not (Russell et al., 2025), modeling "the locations and movements of other agents (e.g., vehicles, pedestrians, cyclists)" (Russell et al., 2025) and distinguishing "situations... caused by the ego agent, and those caused by other agents" (Russell et al., 2025) – intent that is neither hard-coded nor constrained to a finite graph, but must itself be inferred as an irreducibly external random variable.

E.4 A Mathematical Taxonomy of Domain-Specific Transition Functions: Supporting Evidence

Section 3 states this taxonomy formally: every world model here learns some approximation to $T(s_{t+1} \mid s_t, a_t)$, but the surveyed authors' own descriptions imply three structurally different kernels, summarized in Figure 3 and stated as Eqs. (6)–(8). This subsection does not restate that figure or those equations, and instead substantiates each domain's kernel with the quotation-level evidence the main text's one-paragraph summary had no space for.

Gaming. Because the ground-truth process is, in the common case, an existing deterministic program (Valevski et al., 2025) operating over a small, discrete action set (Schrittwieser et al., 2020), the ideal gaming kernel is, to first approximation, a point mass at the engine's own output – Eq. (6) – where f_{engine} is the hand-authored game-loop update function (Valevski et al., 2025) and $\mathcal{A}_{\text{discrete}}$ a finite action alphabet (Schrittwieser et al., 2020). The learning problem is predominantly imitation of a known functional form, and for a fixed, seed-independent engine the residual entropy reflects only the world model's own approximation error. This Dirac kernel is a simplifying idealization rather than exact for every gaming system here, however: DreamerV3 is evaluated in part on ProcGen, a 16-game benchmark that "features randomized levels and visual distractions" (Hafner et al., 2023) – randomness intrinsic to the ground-truth transition function itself, not an artifact of the world model's approximation. The kernel is most accurate for a fixed, non-procedural engine of the kind MuZero's Atari-57 and IRIS's Atari 100k target, and should be read as a low-entropy idealization rather than a zero-entropy one.

AI agents. The GUI-agent kernel is likewise effectively deterministic given the current structured state $r_t = \phi_{\text{text}}(x_t)$ of Eq. (18), but its support is restricted not by physical law but by an app-specific, human-authored interaction graph $\mathcal{R}(r_t, m_t)$ of reachable next screens (Xu et al., 2026) – the constraint Eq. (7) formalizes. Specializing the generic POMDP state/action of Eq. (9) to this domain, $s_t := r_t$ and $a_t := m_t$: this directly reflects AppDeltaWorld's design choice to treat an action whose target lies outside $\mathcal{R}(r_t, m_t)$ as invalid, rather than merely improbable (Xu et al., 2026). There is no Newtonian term in Eq. (7) at all; the constraint is entirely semantic.

Robotics and autonomous vehicles. The robotics kernel has no closed-form deterministic core to fall back on. Its state is continuous, as in SkyJEPA's rigid-body vector $x_t \in \mathbb{R}^{D_s}$ of Eq. (19) (Rao et al., 2026); its action set $\mathcal{A}_{\text{continuous}} \subset \mathbb{R}^{D_a}$ is likewise continuous (four continuous motor forces in SkyJEPA's case (Rao et al., 2026)); and it is irreducibly stochastic, reflecting environmental variation MILE and GAIA-2 condition on or evaluate against (Hu et al., 2022; Russell et al., 2025), unmodeled material effects MoSA's residual-anisotropy correction targets (Wang et al., 2026), and other-agent behavior GAIA-2's multi-agent conditioning captures (Russell et al., 2025). Specializing Eq. (9) to $s_t := x_t$ gives Eq. (8). Its continuous action set is illustrated with SkyJEPA's four motor-force actions, the cleanest purely-continuous, low-level robotics action space in this corpus, but it does not by itself cover every driving action interface surveyed: Vista additionally conditions on discrete "command" and continuous "goal point" inputs alongside low-level trajectory actions (Section F.E), while GAIA-1, GAIA-2, DriveDreamer-2, and ADriver-I accept structured or language-specified scene conditions rather than a real-valued action vector alone – so the action space most relevant to driving systems here is the hybrid set $\mathcal{A}_{\text{AV}} \subseteq \mathbb{R}^{D_a} \cup \mathcal{A}_{\text{discrete}}$, combining Eq. (8)'s continuous term with the discrete alphabet $\mathcal{A}_{\text{discrete}}$ already introduced for gaming; Eq. (8)'s purely-continuous form should be read as the low-level, single-modality limiting case, most exact for SkyJEPA-style platforms. $\Sigma(s_t, a_t)$ is a state- and action-dependent covariance absorbing environmental and multi-agent stochasticity, and f_{robot} is itself only ever an approximation to the true physical dynamics – never an exact program, as f_{engine} is for gaming – which is precisely why SkyJEPA reports autoregressive rollout error that "accumulate[s] over time" (Rao et al., 2026) where a game engine or GUI transition graph would not drift at all.

Equations (6)–(8) make explicit what Sections E.1–E.3 argued in prose: the three domains surveyed are not merely different applications of a single world-modeling recipe, but require transition functions of fundamentally different *degree* of entropy and constraint structure – gaming and AI agents both low-entropy for different reasons, robotics irreducibly stochastic because f_{robot} is only ever an approximation to physical law – rather than a strict deterministic/stochastic split, a distinction Section F revisits at the level of *representation* rather than dynamics, and one Section I argues the field has not yet fully reckoned with when comparing systems across domains on a single benchmark axis.

This taxonomy's illustrative examples should be read accordingly: because AppDeltaWorld, SkyJEPA, and MoSA were selected for their fit to a taxonomic cell already identified from the core papers' own system descriptions, Eq. (18) and Eq. (19) together with Eqs. (6)–(8) demonstrate what each domain's kernel looks like in closed form rather than constituting independent evidence that the three-way split is itself correct. That independent support instead comes from the core corpus's own prose, quoted throughout Sections E.1–E.3 above; the supplemental papers' role is to make those prose descriptions mathematically precise, not to re-establish them.

F Scene Understanding and Representation for Robotics and Autonomous Vehicles

This section looks at the thirty papers along the axis introduced in Table 1's "Scene Representation" column: how each model encodes a scene's 3D structure, expresses uncertainty about how that scene will evolve, and exposes control over the futures it predicts.

F.1 Geometric Representation Trade-offs

The surveyed papers make four distinct choices about where, geometrically, prediction happens. The largest group predicts directly in *image space*: the Dreamer line (Hafner et al., 2019; 2020; 2021; 2023), Genie (Bruce et al., 2024), and GameNGen (Valevski et al., 2025) all reconstruct or roll out raw pixels, keeping the model domain-agnostic but leaving 3D structure implicit. A second group moves prediction into *bird's-eye-view* (BEV) space: MILE (Hu et al., 2022) decodes its recurrent latent into a BEV semantic-segmentation grid, and DriveDreamer/DriveDreamer-2 (Wang et al., 2023; Zhao et al., 2025) condition diffusion on BEV layouts from HD maps. A third commits to explicit *3D geometry*: OccWorld (Zheng et al., 2024) represents the scene as a 3D semantic occupancy grid, motivated as "more fine-grained 3D structure" while remaining "economical" from sparse LiDAR; ViDAR (Yang et al., 2024b) and Copilot4D (Zhang et al., 2024) predict raw 3D point clouds directly, with Copilot4D tokenizing the cloud and applying discrete diffusion in token

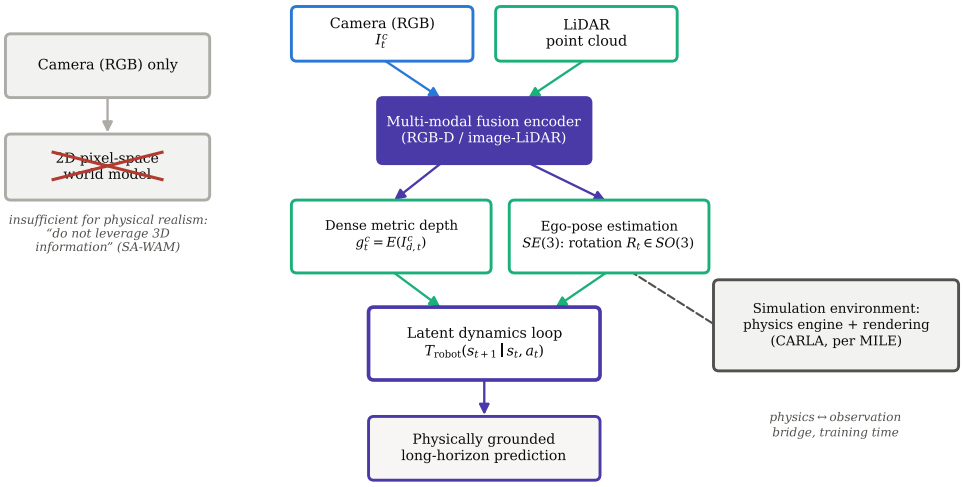

Figure 5: The physical-constraints and geometric-grounding pipeline detailed in this subsection. *Left:* the insufficient RGB-only path, which SA-WAM's authors identify as failing to "leverage 3D information" (Lopetegui Gonzalez et al., 2026). *Center:* the grounded path – camera and LiDAR observations fused into dense metric depth (Eq. 1 of (Lopetegui Gonzalez et al., 2026)) and $SE(3)/SO(3)$ ego-pose ((Zhang et al., 2024; Rao et al., 2026)), both folded directly into the latent dynamics loop $T_{\text{robot}}(s_{t+1} \mid s_t, a_t)$ rather than consumed only at the input. *Right:* the simulation-environment bridge – CARLA, as used by MILE (Hu et al., 2022) – that couples a programmatic physics engine to visual rendering during training, shown dashed to indicate it supplies training-time data rather than participating in the inference-time loop.

space. AD-L-JEPA (Zhu et al., 2025) occupies a fourth position: it operates on LiDAR input like Copilot4D and ViDAR, but predicts a BEV *embedding* rather than reconstructing geometry, placing it in both the point-cloud and representation-space camps. No single geometric choice dominates – image-space, BEV, occupancy, and point-cloud representations are all still active choices as of the most recent papers surveyed here.

F.2 Geometric Representation and Physical Grounding Cues

The four camps above differ not only in where prediction happens, but in whether that representation can support physically grounded reasoning at all. This subsection argues that pure image-space prediction is insufficient for physical realism, and details the geometric machinery – $SE(3)/SO(3)$ pose/attitude structure, dense depth, and multi-modal sensor fusion – that improves the model's geometric *observability* of the scene (Fig. 5). This representation-side grounding – richer, metrically-scaled observations fed into the same learned transition function – is distinct from *dynamics-side* grounding, in which $T(s_{t+1} \mid s_t, a_t)$ itself is constrained to respect momentum, collision, or kinematic feasibility (e.g. via physics-informed neural ODEs or Hamiltonian/Lagrangian networks); the machinery below improves what the model can observe, not what it is constrained to predict, and the two should not be conflated when assessing what an architectural choice guarantees about physical realism.

The insufficiency of 2D pixels. SA-WAM motivates its architecture by identifying this limitation directly: "the prevailing models operate exclusively on RGB observations and do not leverage 3D information" (Lopetegui Gonzalez et al., 2026), a specific gap for "manipulation-intensive tasks where object geometry plays a critical role" (Lopetegui Gonzalez et al., 2026). The consequence is a representation that cannot, even in principle, expose metric quantities – distances, contact points, clearances – a downstream controller needs. OccWorld makes the analogous argument for driving: a 3D occupancy grid captures "more fine-grained 3D structure" than the 2D maps it replaces (Zheng et al., 2024), since occupancy represents free space and obstacle geometry in the same physical units the vehicle must reason about.

$SE(3)$/$SO(3)$ structure in actions and state. A world model that predicts only appearance has no notion of where its own agent is in physical space. Two systems here build rigid-body structure directly into the transition being learned – as a *provided* input or state, not something the model estimates from observation. Copilot4D treats ego-vehicle motion as a structured action: "the observations $\{o^{(t)}\}_t$ are point clouds from the Lidar sensor, and the actions $\{a^{(t)}\}_t$ are $SE(3)$ poses of the ego vehicle" (Zhang et al., 2024), conditioning next-cloud prediction on a known rigid-body transformation supplied as input rather than inferred as localization or odometry would. SkyJEPA uses the same structure at the level of state, representing quadrotor attitude as the columns of $R_t \in SO(3)$ inside its state vector (Eq. (19) of Section E) (Rao et al., 2026) – again a given, not estimated, quantity. Parameterizing a known ego-motion or attitude as manifold-valued rather than an unconstrained Euclidean vector plausibly supports better generalization across viewpoints, but it is not pose estimation, nor does it force the learned kernel to respect $SE(3)$/$SO(3)$ structure internally – that would require the kernel's own parameterization to be equivariant, which neither paper claims to verify.

Dense depth in the latent dynamics loop. A physically grounded world model also needs dense, metric depth rather than sparse or implicit RGB-only cues. SA-WAM injects depth directly into the same latent sequence used for RGB and action prediction: for camera c, RGB and geometric latent frames are $v_t^c = E(I_t^c)$, $g_t^c = E(I_{d,t}^c)$, with E a frozen video-VAE encoder and $I_{d,t}^c$ a tokenizer-compatible encoding of metric depth d_t^c (Lopetegui Gonzalez et al., 2026). Because metric depth is positive and unbounded while the tokenizer expects a bounded range, SA-WAM compares linear, inverse-depth, and log-scale normalizations (Lopetegui Gonzalez et al., 2026), finding log-scale – which allocates "greater resolution to the near field while retaining useful precision at larger distances" (Lopetegui Gonzalez et al., 2026) – best for downstream policy success. The structural point is that dense depth can be a first-class citizen of the same latent-dynamics loop predicting RGB and action, rather than a separate perception module.

Multi-modal sensor fusion: LiDAR and camera. Several driving-specific systems fuse LiDAR with camera observations rather than estimating depth from RGB alone. ViDAR constructs exactly this pairing, "using Image-LiDAR sequences to pre-train visual encoders" (Yang et al., 2024b), jointly modeling "multi-view geometry and temporal dynamics" (Yang et al., 2024b). Copilot4D and AD-L-JEPA operate on LiDAR point clouds directly rather than as an auxiliary signal (Zhu et al., 2025; Zhang et al., 2024). The common thread is that camera observations alone under-determine 3D structure, while LiDAR provides direct, if sparse, metric range measurements – and that combining both, rather than using either alone, is what these systems treat as necessary for accurate physical grounding.

Simulation environments as a physics–observation bridge. Grounding a world model in real physical law requires a training environment coupling a genuine physics engine to policy-relevant visual observation. MILE trains and evaluates inside CARLA, "an Open Urban Driving Simulator" (Hu et al., 2022), evaluating across "new town and new weather conditions" unseen during training (Hu et al., 2022) – an explicit test of whether dynamics learned from simulated physics-plus-rendering generalize rather than memorize. While AirSim and ROS-based stacks are widely used elsewhere in robotics for this bridging role, neither is named in any paper surveyed here; CARLA (Hu et al., 2022) is the only concretely evidenced example in this corpus.

F.3 Semantic Cues and Interpretability

Several driving-specific models pair their geometric representation with an explicit semantic decoding step. MILE decodes its latent state into an eight-class BEV semantic map (background, road, lane marking, vehicles, pedestrians, traffic-light states), so "diverse and plausible future states and actions" can be inspected visually rather than taken on faith; OccWorld's 3D occupancy grid serves the same purpose in three dimensions. By contrast, the purely generative video models – GAIA-1, GAIA-2, Vista, WorldDreamer (Hu et al., 2023; Russell et al., 2025; Gao et al., 2024; Wang et al., 2024) – decode no explicit semantic layer; GAIA-1 instead reports semantic understanding as *emergent*, evidenced by the model's ability to "understand and disentangle important concepts such as static and dynamic elements" without semantic-segmentation supervision – a real split between interpretable-by-construction decoding (MILE, OccWorld) and interpretability as an emergent property of scale (GAIA-1).

F.4 Temporal Dynamics and Uncertainty

A recurring concern across recurrent-latent papers is that a model collapsing multiple plausible futures into one blurry average is of little use for planning. PlaNet's RSSM is deliberately split into deterministic and stochastic components specifically to avoid "overconfident predictions" and failure to "capture multiple possible futures"; MILE inherits this design and reports the consequence: its BEV decoder visualizes "multi-modal futures." Generative video models handle the same problem by sampling instead – GAIA-1 and DriveDreamer-family models can be resampled for different plausible continuations. The JEPA family sidesteps the question in its base form – I-JEPA and V-JEPA predict a single point in representation space – though LeCun's original proposal frames JEPA's predictor as *optionally* conditioned on a latent variable z to capture "information about y that is not deducible from x," representing multiple plausible futures without committing to one.

F.5 Action-Conditioning and Controllability

Finally, the surveyed papers differ in how much control a user or downstream planner has over which future is generated. At the low-control end, I-JEPA and V-JEPA are unconditional representation learners with no action input. In the middle, the Dreamer line and MILE condition on a low-level action vector supplied by the policy being trained. At the high-control end, several recent driving world models expose rich, structured conditioning: GAIA-1 offers "fine-grained control over ego-vehicle behavior and scene features through action and language conditioning"; Vista incorporates "a versatile set of controls from high-level intentions (command, goal point) to low-level maneuvers (trajectory, ...)"; GAIA-2 conditions on "a diverse set of domain-specific control signals" spanning ego-vehicle dynamics, agent configurations, environmental factors, and road semantics; and DriveDreamer-2 inserts an LLM front end so free-form *text prompts* can specify the traffic scenario to generate, positioned as the first world model producing customized, uncommon scenarios "in a user-friendly manner." This progression – no conditioning, to low-level action conditioning, to structured multi-attribute conditioning, to free-form language conditioning – runs roughly parallel to the architectural progression of Section D, setting up Section G's discussion of what these increasingly controllable world models are used *for*.

G World Models for Downstream Robotics and Autonomous-Driving Tasks

Sections D and F examined world models as architectures. This section examines them as *components*: once trained, what role does a world model actually play in the surrounding system? Across the thirty papers, that role falls into four recurring, not mutually exclusive, patterns.

G.1 World Models as Simulators and Data Engines

The most direct downstream use is as a source of training data or evaluation scenarios that would otherwise require real-world collection. UniSim (Yang et al., 2024a) states this role explicitly, positioning itself as "a universal simulator of real-world interaction" training both high-level vision-language and low-level RL policies entirely in simulation, deployed to real robots "in zero shot" by making the simulator "nearly visually indistinguishable from the real world." The same paper notes a second use: generating rare events (e.g. self-driving crashes) that are "expensive or dangerous" to collect in reality, for training rare-event detectors. GAIA-1 and GAIA-2 (Hu et al., 2023; Russell et al., 2025) occupy the driving-specific version of this role, framed as "advancing the use of generative world models as a core tool in the development of autonomous systems," with explicit conditioning over geography, weather, road configuration, and "rare but critical edge-case scenarios" that real-world data collection under-samples. DriveDreamer-2 (Zhao et al., 2025) reports a concrete consequence: video generated by the model, used as additional training data, measurably improves 3D detection and tracking on held-out driving-perception benchmarks – a world model's synthetic output feeding back into a conventional perception pipeline.

G.2 World Models as Planners

A second role treats the world model as the substrate the agent plans or trains *inside*, via two mechanisms. Gradient-based imagination: Dreamer's actor-critic is trained by "propagating analytic gradients of learned state values back through trajectories imagined in the compact state space," and Ha and Schmidhuber's original controller is trained "entirely inside of its own hallucinated dream" with no gradient through the real environment. Search-based planning: PlaNet plans online with the cross-entropy method directly in RSSM latent space, and MuZero (Schrittwieser et al., 2020) combines its value-equivalent model with Monte Carlo tree search rather than a differentiable actor. DayDreamer (Wu et al., 2022) shows the gradient-based version is not confined to simulation: the same Dreamer model trains and plans online, on a physical robot, with "a learner thread continuously train[ing] the world model and actor-critic behavior, while an actor thread in parallel computes actions." MILE and Vista (Hu et al., 2022; Gao et al., 2024) extend the planner role to driving: MILE executes "complex driving manoeuvres" "from plans entirely predicted in imagination," and Vista uses the trained world model directly as a *reward function* to score and rank candidate driving actions, rather than only as a source of imagined rollouts for a separately trained critic.

G.3 World Models as Representation-Pretraining Backbones

A third role, concentrated in the JEPA family and the point-cloud driving models, treats world modeling as a self-supervised pretraining objective whose payoff is a reusable encoder. I-JEPA and V-JEPA (Assran et al., 2023; Bardes et al., 2024) are evaluated almost entirely on downstream transfer – frozen-backbone linear probing and fine-tuning – rather than closed-loop control. ViDAR (Yang et al., 2024b) makes the same move for driving perception: its future-point-cloud forecasting objective is "a scalable, label-free pretraining signal" that transfers to measurable gains in downstream 3D object detection, occupancy prediction, and motion planning (Table 2). AD-L-JEPA (Zhu et al., 2025) is the most explicit instance: described by its authors as "the first JEPA-based pretraining method for autonomous driving," evaluated purely on how much its pretrained LiDAR representations improve downstream 3D detection mAP, at a fraction of the GPU cost of prior pretraining methods, rather than any closed-loop metric.

G.4 World Models Inside Unified Perception-Action Stacks

A fourth, more recent role folds the world model into a single model that also performs perception and action selection, rather than keeping three separate stages. ADriver-I (Jia et al., 2023) is the clearest example: one fine-tuned multimodal LLM both predicts low-level control signals and, paired with a video diffusion module, predicts the resulting future scene, so "the predicted current action directly affects the scene in the future" within one autoregressive loop. V-JEPA 2's robotic extension, V-JEPA 2-AC (Assran et al., 2025), reports a similar unification from the representation-learning side: after post-training on fewer than 62 hours of unlabeled robot video, the resulting action-conditioned world model is deployed *zero-shot* on physical robot arms in two different labs for pick-and-place, "without collecting any data from the robots in these environments, and without any task-specific training or reward." WorldDrive (Gui et al., 2026) motivates this unification as a response to keeping scene generation and motion planning as separate stages, which leaves the motion representation "not explicitly designed to be planner-shared and inheritable"; its trajectory-aware architecture closes that gap.

Read together, these four roles show "world model" names a family of deployment patterns, not a single downstream use, correlating closely with the architectural choices of Sections D and F: gradient-based imagination favors the differentiable RSSM line, pretraining favors the JEPA line, and unified perception-action stacks are, so far, the province of the newest Transformer-, diffusion-, and JEPA-based systems in this survey.

H Benchmarks and Evaluation Protocols

This section examines a fourth axis along which the surveyed papers differ sharply: what each paper actually measures, and against what benchmark. Table 2 restricts to the thirteen papers evaluated on a driving-relevant benchmark; this section widens the view to the evaluation protocols used across

all thirty papers and asks what these disparate metrics do and do not tell a reader about closed-loop, real-time viability. Because "real-time" and "near real-time" recur throughout this survey as labels applied across figures measured on different hardware and task definitions, this survey adopts an explicit threshold rather than treating the label as self-evident: standard closed-loop control practice calls for update rates of 10–30 Hz for mobile-robot perception and planning (Siegwart et al., 2011), and this survey treats a reported figure at or above that range as meeting a domain-appropriate real-time threshold, a figure clearly below it (such as V-JEPA 2's 4 FPS at the low end of its reported 4–16 FPS range) as *near* real-time at best, and no reported number at all as simply not comparable on this axis. Table 1's labels for individual systems are reproduced as their own authors describe them, and should be read against this 10–30 Hz reference rather than as a claim that every so-labeled system meets an identical bar.

H.1 General-Domain Control and Game Benchmarks

The recurrent and Transformer-based world models here are evaluated almost exclusively on established RL and game suites, though *which* suite varies considerably and is not always Atari. The Atari Learning Environment is used explicitly by four of the thirty papers (Table 3): MuZero (the full Atari-57 suite), DreamerV2 (55 games), IRIS (the Atari 100k subset), and DreamerV3 (one of its eight evaluation domains). Others report on adjacent but distinct game benchmarks: Ha and Schmidhuber's original World Models targets CarRacing-v0 and VizDoom's Take Cover, and GameNGen targets DOOM gameplay data. Two Atari protocols recur among the four papers that use it: the full-data protocol (MuZero, DreamerV2) trains for up to 200 million frames per game across all 57 games and reports human-normalized score, with MuZero reporting "a new state of the art for both mean and median normalized score across the 57 Atari games"; the data-efficient Atari 100k protocol restricts training to roughly two hours of gameplay per game across 26 games, with IRIS reporting a mean human-normalized score of 1.046, "outperform[ing] humans on 10 out of 26 games," using the same score definition despite the two-orders-of-magnitude difference in training budget. DreamerV3 is evaluated most broadly: eight distinct domains and more than 150 tasks under one fixed hyperparameter configuration, spanning Atari, ProcGen (16 procedurally generated games), DeepMind Control Suite (proprioceptive and visual variants), BSuite (23 environments, 468 configurations), and Minecraft diamond collection from raw pixels and sparse reward. Genie and GameNGen, by contrast, are evaluated on visual and interactive quality rather than task reward (below). MuZero additionally reports Go, chess, and shogi results by Elo rating rather than any score-based metric – human-normalized score is itself only one of several incompatible scales in use across this literature, not a universal one.

H.2 Generative Video-Quality Metrics

The Transformer- and diffusion-based generative models – Genie, GameNGen, and the entire GAIA/DriveDreamer/Vista/WorldDreamer/WorldDrive line – are evaluated on metrics borrowed from video generation rather than RL. Fréchet Video Distance (FVD) and Fréchet Inception Distance (FID) are most widely reported: DriveDreamer-2 reports "FID of 11.2 and FVD of 55.7," a relative improvement of "$\sim$30% and $\sim$50% compared to previous best-performing" methods, and Vista reports FID improved by 55% and FVD by 27% over the best prior driving world model on nuScenes. GAIA-2 reports FID alongside FDD (Fréchet DINO Distance), computed over DINO self-supervised embeddings rather than the I3D features FVD uses (Russell et al., 2025), so its figure is not directly interchangeable with DriveDreamer-2's or Vista's. WorldDrive, despite the same diffusion-based generative family, reports no FVD/FID/FDD-style metric at all – its evaluation is entirely planning-focused (Section H.C), a further instance of metric fragmentation within the same architectural family and domain. Genie evaluates its Platformer environments with FVD of 82.7 alongside a frame-prediction metric it calls Δt PSNR, and GameNGen reports next-frame PSNR of 29.4 – comparable, the authors note, "to lossy JPEG compression" – as the basis for its real-time-playability claim. These metrics measure distributional closeness between generated and real video, not whether an agent conditioning its actions on the generated rollout would act correctly, and among the papers reporting one of these three fidelity metrics, none also reports a closed-loop control outcome for the same model.

H.3 Driving-Specific Perception and Planning Metrics

The thirteen driving-domain papers in Table 2 split into perception-style and planning-style protocols. Perception-style evaluation borrows metrics from 3D object detection and occupancy-forecasting: ViDAR reports "3.1% NDS on 3D detection" and a "∼33% Chamfer Distance reduction on point cloud" forecasting, using nuScenes Detection Score (NDS) alongside mAP and mIoU; OccWorld reports "an average IoU of 26.63 and mIoU of 17.13 for 3s future" occupancy forecasting; and Copilot4D reports Chamfer Distance reduced by more than 65% versus prior state of the art at a 1s horizon, alongside L1 depth error and AbsRel via raycasting against ground-truth LiDAR. Planning-style evaluation instead scores the predicted or executed trajectory directly: MILE reports a $+31\%$ improvement in CARLA's aggregate Driving Score, and WorldDrive reports L2 trajectory error and collision rate alongside the composite NAVSIM Predictive Driver Model Score (PDMS), aggregating no-collision rate, drivable-area compliance, time-to-collision, comfort, and ego progress into one number. AD-L-JEPA sits outside both groups: evaluated purely on downstream 3D detection mAP gained from pretrained features, isolating representation quality from either video fidelity or closed-loop outcome.

H.4 Representation-Transfer Evaluation

The JEPA family and UniSim are evaluated on a fourth protocol: downstream transfer of a frozen, pretrained representation rather than any property of the model's own output. I-JEPA reports ImageNet-1K linear-probe accuracy, including a low-label regime with only 1% of available labels, plus transfer to CIFAR-100 and Place205. V-JEPA 2 extends this to video and action: 77.3 top-1 accuracy on Something-Something v2 with an attentive probe, and state-of-the-art on Epic-Kitchens-100 human action anticipation, with "39.7 recall-at-5, which is a 44% relative improvement over the previous best model." None of these transfer metrics involve closed-loop control; V-JEPA 2's own robot deployment (V-JEPA 2-AC, Section G) is reported as a separate, qualitative zero-shot pick-and-place demonstration rather than a benchmark score in the same table.

H.5 The Absence of a Shared Evaluation Protocol

Two distinct claims should be kept separate here. The first is unsurprising and not itself a failure: a human-normalized Atari score, an FVD/FID/FDD figure, a driving-score percentage, a Chamfer Distance, and a linear-probe accuracy are five different scales answering five different sub-task questions, and no single number could summarize all five without discarding what makes each meaningful. The narrower, genuinely diagnostic claim is that comparability breaks down even *within* a single sub-task. IRIS and MuZero both report "human-normalized score" on Atari under training budgets differing by roughly two orders of magnitude; FID/FVD/FDD figures are reported by eight different papers against eight different held-out video sets rather than one shared split, and not even every system in the same architecture family and domain reports the same fidelity metric, or any at all. The latency/real-time-viability column compounds this: excluding the three papers for which closed-loop latency is definitionally inapplicable (LeCun's position paper, I-JEPA, V-JEPA – conceptual or representation-pretraining work), only 6 of the remaining 26 empirically-deployed papers (∼23%) report any quantitative inference-time or throughput figure at all – a near-universal omission, not merely an inconsistency – and of those six, MuZero's 100 ms search budget, MILE's 43.0 Hz, Genie's ∼1 FPS, GameNGen's 20 FPS, V-JEPA 2's 4–16 FPS, and WorldDrive's 53 ms end-to-end latency are measured on different hardware, batch and resolution settings, and definitions of one inference step. A model's video-quality, task-reward, or transfer-probe score therefore says little on its own about closed-loop, real-time viability, and a model's reported latency says little about the representational richness sacrificed to achieve it. This within-sub-task gap between what the field measures and what physical robotic deployment actually requires is the evaluation-side counterpart of the architectural latency bottleneck identified in Section D, and motivates the open challenges discussed next.

I Open Challenges and Future Directions

The preceding sections traced four axes of variation across the thirty surveyed papers – architecture, scene representation, downstream role, and evaluation protocol – and each repeatedly surfaced the

same handful of open problems, stated in the surveyed authors' own limitations and future-work sections rather than inferred externally. This section collects five such problems and argues that they are related facets of a single underlying tension: representational and behavioral richness scales against real-time viability, data efficiency, and evaluability, and no system surveyed here satisfies all of these simultaneously.

I.1 The Latency–Fidelity Trade-off

Section D already identified this pattern, though the throughput figures behind it are self-reported numbers measured on different hardware, batch sizes, and resolution settings (Section H.E) rather than a matched benchmark: the most expressive generative world models – GAIA-1 and GAIA-2 – are, by their own authors' statement, not yet real-time, and GAIA-1's authors attribute this specifically to sampling procedure and hardware maturity ("parallel sampling [...], reducing the number of latent variables [...], and improvements in hardware accelerators"), not to architectural expressiveness – implying the gap could narrow under equal engineering investment rather than a change of architecture. The systems that do report real-time or near-real-time figures – MILE's 43.0 Hz BEV inference, GameNGen's 20 FPS, V-JEPA 2's 4–16 FPS, WorldDrive's 53 ms – achieve it by constraining representation (BEV rather than multi-camera video, discrete actions rather than free-form generation, representation-space prediction rather than pixel reconstruction), not by closing the gap while keeping full generative fidelity. No paper surveyed here reports both diffusion/Transformer-level fidelity and real-time deployment on the same system; TD-MPC and TD-MPC2 (Hansen et al., 2022; 2024) bear on this from the constrained-representation side, planning over a compact, decoder-free state like MuZero, but report no directly comparable latency figure, so this survey treats decoder-free planning as architecturally consistent with, rather than measured evidence for, the low-latency end of the trade-off.

Figure 4 lays this out visually: the fast, constrained systems (MILE, GameNGen, WorldDrive) cluster at low latency but moderate richness; GAIA-2 and UniSim sit at the opposite extreme, reporting no throughput figure at all. The upper-left quadrant – low latency *and* high fidelity – is conspicuously empty across this survey's 35-paper corpus: the "missing middle" this survey's synthesis identifies as the actual deployment target for physical robotics.

I.2 Hallucination and Long-Horizon Error Accumulation

Predicted futures degrade the further a world model predicts without new grounding observations, but this survey's review of failure reports finds at least two mechanistically distinct failure modes. The first is *generative hallucination under weak conditioning*: UniSim reports that when an action is unrealistic given the scene, "we observe hallucinations (e.g., the table turns into a sink or the view turns away from the tabletop robot and a sink shows up)," and states a preference, not yet realized, for a simulator that "detect[s] actions that are not possible to simulate as opposed to hallucinating unrealistic outcomes." The second, distinct failure is *autoregressive error accumulation under covariate shift*: V-JEPA 2-AC reports that "autoregressive prediction suffers from error accumulation" as prediction horizon grows, independent of pixel- or latent-space prediction – the classical sequential-prediction problem Ross et al. (Ross et al., 2011) formalize for imitation learning, where per-step errors compound under the model's own induced distribution rather than the training distribution. UniSim separately notes a related limited-context failure: a simulator "conditioned on a few frames of the recent history cannot capture long-term memory," so an out-of-view object can vanish once memory of it falls outside the conditioning window. These call for different fixes: hallucination under weak conditioning calls for an explicit action-feasibility or confidence signal, while covariate-shift error accumulation calls for uncertainty-aware replanning, periodic observation re-grounding, or explicit state-space confinement of the kind DAgger-style imitation learning uses. Whether the fix is explicit uncertainty modeling (the stochastic-latent approach PlaNet and MILE already use), longer or externalized memory, or a feasibility signal, remains unresolved across the papers surveyed here.

I.3 Sim-to-Real and Out-of-Domain Generalization

A world model's competence is bounded by what its training distribution covers, and every paper reporting this limitation reports it as still open. UniSim frames its contribution as "bridging the

sim-to-real gap in embodied learning," yet reports that generalization to an unseen robot is "limited," since the simulator trains on only four robot morphologies – attributed to data scale, with further scaling proposed as the fix rather than reported as achieved. DayDreamer takes the opposite approach, training and planning directly on physical robots online, and reports its own cost: "learning on hardware over many hours creates wear on robots that may require human intervention or repair." Domain randomization – systematically varying a simulator's lighting, texture, object mass/friction, and sensor noise during training – sits between these poles and is empirically validated for robotic manipulation and locomotion (Tobin et al., 2017; Andrychowicz et al., 2020); this survey's own corpus corroborates it independently: SkyJEPA trains with "a domain-randomized simulation pipeline for automated dataset generation" for "zero-shot sim-to-real transfer" (Rao et al., 2026), and SA-WAM reports "strong gains in randomized environments" in real-world robot-arm evaluation (Lopetegui Gonzalez et al., 2026). The three strategies – simulate at scale and transfer, train directly on hardware and absorb the cost, or randomize the simulator – are different responses to the same unresolved data-availability/real-world-grounding trade-off, not a problem any paper surveyed here claims to have fully closed.

I.4 The Absence of a Shared Evaluation Protocol

Section H developed this point at the level of metrics: a human-normalized Atari score, an FVD/FID pair, a driving score, a Chamfer Distance, and a linear-probe accuracy cannot be placed on one comparable axis, and only about 6 of the 26 papers for which closed-loop latency is a meaningful question report any quantitative latency or throughput figure at all. MILE's authors make an unusually explicit version of this critique from inside the driving literature itself: CARLA's own aggregate Driving Score, obtained by "multiplying the route completion with the infraction penalty," lets a single severe infraction disproportionately collapse the score regardless of how much of the route was driven safely – which is why MILE reports a supplementary timestep-level reward metric to correct for it. If the field's own most careful driving benchmark is judged flawed by one of its principal users, an evaluation protocol capable of jointly certifying representational quality, closed-loop safety, and real-time viability remains an open construction problem.

I.5 Multi-Agent Interaction and Safety-Critical Deployment

GAIA-2 is the clearest system designed to reason about other independent agents rather than a passively-reactive environment, unifying "multi-agent interaction, fine-grained control, and multi-camera consistency" so that safety-relevant scenarios can be generated and stress-tested rather than only encountered by chance. Across the rest of the survey, explicit multi-agent modeling is the exception: the recurrent, JEPA, and pretraining-backbone papers are overwhelmingly evaluated in single-agent or ego-only settings, and none of the thirty papers reports a quantitative safety metric for hallucinated or slowly-updating predictions.

This survey's own closing argument that unifying scene and motion representation within a single model is promising sits in tension with a requirement from outside this survey's corpus: functional-safety standards for automotive systems generally require a fault be traceable to a specific, independently verifiable component so a certifying body can audit each stage in isolation (Koopman & Wagner, 2017). A single, unified model that jointly produces a scene prediction and a motion plan is, by construction, harder to fault-isolate than an explicitly staged pipeline, and none of the unification-oriented systems surveyed here, WorldDrive included, report a fault-attribution mechanism. A future architecture would need to unify representational stages for efficiency *and* retain interpretable decomposability sufficient for independent, stage-level safety auditing, rather than treating the two as mutually exclusive.

I.6 Synthesis: A Single Underlying Tension

These five problems are not independent: a model that predicts richer, more multi-agent, more semantically detailed futures tends to be slower and harder to certify; a model fast enough for closed-loop deployment tends to achieve that speed by constraining representation, inheriting hallucination and error-accumulation risk when pushed beyond its training distribution or prediction horizon; and the field currently lacks a shared protocol capable of exposing this trade-off on one comparable scale. None of the thirty papers surveyed reports a system simultaneously real-time, safe under long-

horizon and multi-agent prediction, generalizable outside its training distribution, and evaluated against a benchmark the field agrees is sufficient. Closing that combined gap, rather than improving any one axis in isolation, is the open challenge this survey's Introduction and Section D's latency-bottleneck discussion both point toward.