Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Visual World-Action Models for Embodied AI and Autonomous Driving: A Systematic Survey of Spatial Representations, Generative Dynamics, and Benchmarks

Binay Kumar Singh

binay.singh@ucf.edu

University of Central Florida

Niels Da Vitoria Lobo

University of Central Florida

Systematic Review

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11864906/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Visual World-Action Models for Embodied AI and Autonomous Driving: A Systematic Survey of Spatial Representations, Generative Dynamics, and Benchmarks

Binay Kumar Singh* and Niels Da Vitoria Lobo

University of Central Florida, Orlando, FL, USA.

*Corresponding author(s). E-mail(s): binay.singh@ucf.edu;
Contributing authors: Niels.DaVitoriaLobo@ucf.edu;

Abstract

Rapid progress in generative video modeling, 3D scene representation, and large-scale embodied datasets has catalyzed the emergence of World-Action Models (WAMs)—unified predictive frameworks that jointly model environment dynamics and agent actions over extended temporal horizons. Unlike reactive Vision-Language-Action (VLA) policies that map current observations directly to motor commands, WAMs internalize physical causality, spatial geometry, and counterfactual outcomes to support closed-loop planning, simulation, and policy optimization. Recent surveys organize WAMs mainly by functional role and architectural coupling; fewer analyze them from a visual-computing perspective that contrasts how WAMs represent 2D, 3D, and 4D world states alongside physical accuracy and deployment latency. We establish a formal mathematical distinction between VLAs and WAMs, introduce a three-axis taxonomy spanning Latent Dynamics, Pixel/Video-Generative, and 3D/Occupancy-Centric paradigms, and compare 28 representative architectures across robotics and autonomous driving. We treat driving WAMs in depth, catalog 18 core benchmarks and simulators, and compile verified leaderboards of reported results on NAVSIM, nuScenes, Occ3D, LIBERO, and CALVIN. We further analyze the divergence between visual generation quality (e.g., FVD) and physical task success, formalize three physics-grounded diagnostics with an open-source reference implementation, and outline five open challenges with engineering remedies. Code is available at https://github.com/binayksingh/awesome-visual-world-action-models.

Keywords: World-Action Models, Autonomous Driving, Embodied AI, Generative World Models, 3D Scene Representation, Robot Learning

1 Introduction

Building embodied agents capable of perceiving, reasoning, and acting reliably in unstructured 3D environments remains a central goal of computer vision and robotics [1, 2]. Over the past two years, Vision-Language-Action (VLA) models have demonstrated strong zero-shot generalization by adapting vision-language backbones to output low-level control tokens [3, 4]. However, purely reactive VLAs lack an explicit internal mechanism to simulate physical consequences before execution, leading to compounding errors in long-horizon manipulation, occluded multi-view traffic, and out-of-distribution contacts [5, 6].

To overcome these limitations, World-Action Models (WAMs) unify predictive world modeling with action conditioning. By learning the joint transition dynamics of visual observations and agent actions, a WAM acts simultaneously as a neural physics simulator, a counterfactual planner, and a dense representation learner [7–9].

1.1 Formal Distinction: VLAs vs. WAMs

Let $o_t \in \mathcal{O}$ denote multi-modal visual observations (e.g., monocular/multi-view RGB frames, depth maps, or 3D occupancy grids) at timestep t, $a_t \in \mathcal{A}$ denote the agent's continuous or discrete action, and $l \in \mathcal{L}$ denote a high-level language or goal specification.

- **Reactive VLA policy:** maps current perceptual history and task instructions directly to an immediate motor command without explicitly modeling environmental state transitions,
$$a_t = \pi_\theta(o_{1:t}, l). \tag{1}$$
- **World-Action Model (WAM):** learns a generative or predictive transition model over latent, pixel, or 3D spatial states $z_{t+1:t+H}$ coupled with action generation or trajectory scoring over prediction horizon H,
$$p_\phi\big(z_{t+1:t+H},\, a_{t:t+H-1} \mid o_{1:t},\, a_{1:t-1},\, l\big), \tag{2}$$
typically trained with a composite objective
$$\mathcal{L}_{\text{WAM}} = \lambda_1 \mathcal{L}_{\text{fwd}}(\hat{z}, z) + \lambda_2 \mathcal{L}_{\text{act}}(\hat{a}, a) + \lambda_3 \mathcal{L}_{\text{geo}}, \tag{3}$$
where $\mathcal{L}_{\text{geo}}$ is an optional geometric regularizer (Eq. (7), Section 3.5).

Definition 1 (Visual World-Action Model) A *Visual World-Action Model* is a model that, given visual observations $o_{1:t}$, past actions $a_{1:t-1}$, and an optional instruction l, (i) predicts future world states $z_{t+1:t+H}$ in a latent, 2D pixel, or 3D/4D spatial parameterization, and (ii) uses that prediction to generate, select, or train actions $a_{t:t+H-1}$, such that its forecasts change with the action ($\Delta_{\text{act}} > 0$, Eq. (8)). A model that predicts futures without conditioning on or informing actions is a world model; a model that outputs actions without predicting future states is a reactive VLA policy.

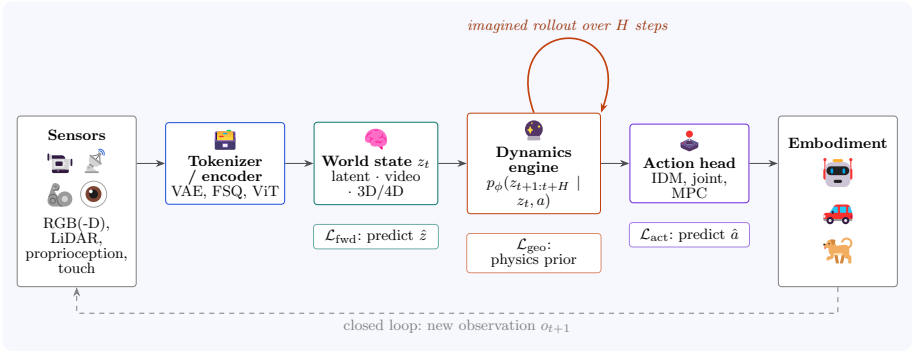


Fig. 1 Anatomy of a Visual World-Action Model. Multimodal observations are tokenized into a world state z_t; a dynamics engine imagines future states over a horizon H; an action head turns the imagined future into commands for a robot arm, vehicle, or legged robot, and the loop closes on new observations. Training combines forward-prediction, action, and optional geometric losses (Eq. (3)).

Figure 1 summarizes this generic pipeline, which every architecture reviewed in this survey instantiates differently.

1.2 Relation to Existing Surveys

Several recent surveys cover adjacent ground. Li et al. [10] organize world models for embodied AI along functionality, temporal modeling, and spatial representation; Kong et al. [11] focus on native 3D and 4D world modeling through video, occupancy, and LiDAR generation; and Shang et al. [12] classify embodied world models by generation modality, control signal, and generation view. Surveys dedicated specifically to WAMs have also appeared: Wang et al. [6] define WAMs as jointly modeling future states and actions and distinguish cascaded from joint designs; Shen et al. [13] organize WAMs by what they must generate (rendered futures, latent futures, or generation-free action reasoning); and Lu et al. [14] provide a robotics-oriented taxonomy of representations, transition models, and action interfaces. Singh and Lobo [15] place predictive world modeling within a broader pixels-to-actions hierarchy for spatiotemporal understanding, survey shared urban state representations for autonomous mobility [16], and test empirically whether driving world models depend on their conditioning action [17].

Driving-specific surveys form a second line of work: Guan et al. [18] give an early overview of world models for autonomous driving; Feng et al. [19] organize them into future-world generation, behaviour planning, and prediction–planning interaction, and maintain a public benchmark repository; and Tu et al. [20] categorize driving world models by the modality of the predicted scene (video, point cloud, occupancy, latent feature, traffic map) and report the performance of representative methods. These remain the most complete references for driving world models as scene generators. Table 1 summarizes how all of these surveys are organized.

The present survey is complementary to these works. Rather than re-deriving a general taxonomy, it adopts a visual-computing perspective, contrasting how WAMs parameterize 2D, 3D, and 4D world states, and it contributes physics-grounded evaluation diagnostics, a hardware failure-mode taxonomy, and a deployment-oriented

Table 1 Recent surveys related to world(-action) models and how each is organized. Descriptions follow each survey's own stated scope.

Survey	Primary organizing axis	Domain focus
Guan et al. [18] (2024)	Foundations, applications, and open problems of world models	Autonomous driving
Feng et al. [19] (2025)	Future-world generation, behaviour planning, prediction–planning interaction	Autonomous driving
Tu et al. [20] (2025)	Modality of the predicted scene (video, point cloud, occupancy, latent, map)	Autonomous driving
Li et al. [10] (2025)	Functionality, temporal modeling, spatial representation	Embodied AI
Kong et al. [11] (2025)	Native 3D/4D generation via video, occupancy, and LiDAR	3D/4D world modeling
Shang et al. [12] (2026)	Generation modality, control signal, generation view	Embodied AI
Wang et al. [6] (2026)	Cascaded vs. joint modeling of future states and actions	Embodied AI (WAMs)
Shen et al. [13] (2026)	What must be generated: rendered, latent, or no future	Embodied AI (WAMs)
Lu et al. [14] (2026)	Representations, transition models, action interfaces	Robot learning and control
Zhang et al. [21] (2026)	Tutorial: imagine-then-execute, feature-conditioned, joint video–action, auxiliary prediction	Robotics
This survey	Spatial-state parameterization (2D/3D/4D) × coupling × temporal engine; physics-grounded metrics with code; verified cross-domain leaderboards	Robotics *and* autonomous driving

analysis of latency and safety that the works above treat only briefly. It also places driving and robotics under one action-conditioned formalism, so that results and failure modes can be compared across the two domains.

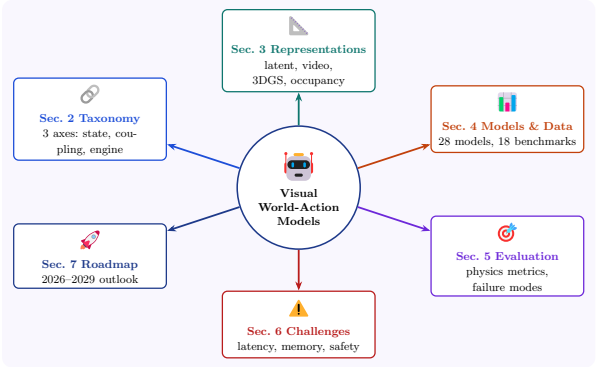


Fig. 2 Organization of this survey. Starting from a formal definition (Section 1.1), the survey proceeds through taxonomy, visual and spatial representations, models and datasets, physics-grounded evaluation, open challenges, and a research roadmap.

1.3 Contributions and Survey Scope

Building on the definitions and broad design space established in the surveys above, this survey targets the visual computing, 3D scene representation, and benchmarking dimensions essential for computer vision researchers:

1. **Canonical three-axis taxonomy:** we organize WAMs by spatial state representation, architectural coupling, and temporal generation engine (Section 2, Figs. 4–5).
2. **Visual and spatial representation analysis:** we contrast 2D video diffusion tokens against latent predictive spaces, 3D Gaussian Splatting (3DGS), and 4D occupancy grids (Section 3, Fig. 7).
3. **Master lookup tables and benchmark synthesis:** we provide standardized comparison tables covering representative WAMs and 18 datasets (Section 4), alongside an analysis of visual fidelity vs. physical control success (Section 5, Fig. 10).
4. **Driving world-action models and verified leaderboards:** we review driving WAMs by family (Section 4.4) and compile reported results on NAVSIM, nuScenes, Occ3D, LIBERO, and CALVIN from the original papers (Section 5.2).
5. **Physics-grounded metrics with code:** we formalize IVR, E_{dn}, and Δ_{act} (Section 5.4) and release a reference implementation.
6. **Actionable open challenges and roadmap:** five research questions (RQ1–RQ5) with measurable targets, a named reporting checklist (WAM-RC), and concrete technical remedies for five core bottlenecks and a multi-year evolution timeline (Section 6, Fig. 14).

Figure 2 gives a visual overview of the organization of the survey.

1.4 Historical Evolution

The conceptual roots of modern World-Action Models trace back to early recurrent state-space planners such as World Models [2] and PlaNet [22], which demonstrated that planning in a compact stochastic latent space could solve continuous control benchmarks from raw pixels without simulating full RGB frames at inference time. DreamerV2 [23] and DreamerV3 [24] stabilized actor-critic learning inside imagined latent rollouts via discrete categorical latents and symlog transformations, achieving broad mastery across 150+ tasks. In parallel, pixel-level video prediction frameworks such as FitVid [25] exposed how naive L_2 reconstruction blurs multi-modal stochastic futures, motivating the shift toward latent diffusion transformers (DiTs) and masked joint-embedding architectures (V-JEPA [9]). Between 2023 and 2026, two inflection points merged these threads into unified WAMs: (1) internet-scale generative video pretraining (UniSim [5], Cosmos-WFM [26]), which endowed predictors with rich zero-shot physical priors, and (2) 3D geometric lifting (OccWorld [27], ManiGaussian [28]), which grounded predicted futures in metric space. Figure 3 places these milestones on a common timeline.

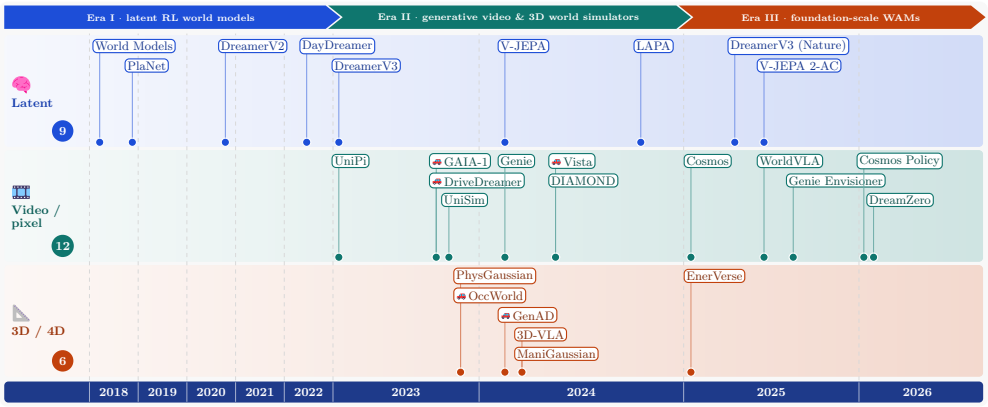


Fig. 3 Milestones of visual world(-action) models placed by year of first public release (arXiv or project release) and coloured by spatial-state paradigm (Axis I). Driving models are marked with a car icon; the circled number in each lane gives the count of milestones shown for that paradigm, and the top ribbon marks three broad eras. References: World Models [2], PlaNet [22], DreamerV2/V3 [23, 24], DayDreamer [29], V-JEPA/V-JEPA 2 [9, 30], LAPA [31], UniPi [32], GAIA-1 [7], DriveDreamer [33], UniSim [5], Genie [8], Vista [34], DIAMOND [35], Cosmos [26], WorldVLA [36], Genie Envisioner [37], DreamZero [38], Cosmos Policy [39], OccWorld [27], 3D-VLA [40], ManiGaussian [28], GenAD [41], PhysGaussian [42], EnerVerse [43].

2 Three-Axis Taxonomy of WAMs

As illustrated in Fig. 4, we categorize existing Visual World-Action Models along three orthogonal design axes governing computational cost, visual fidelity, and control precision.

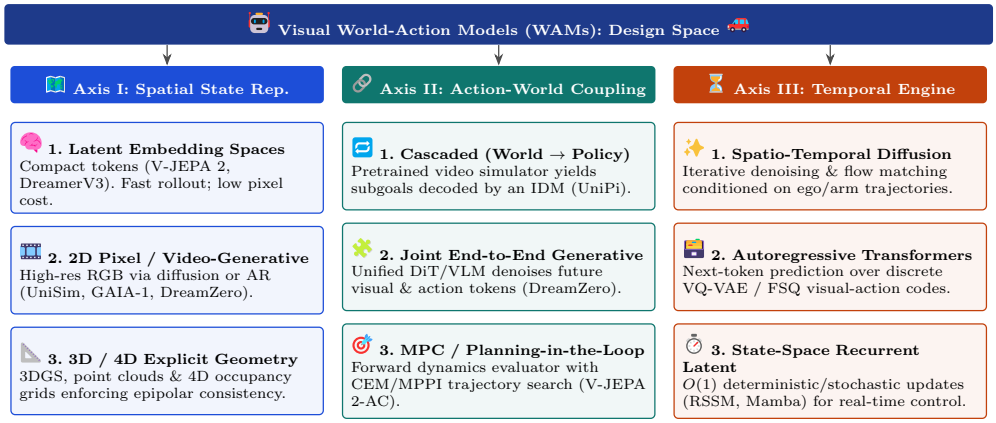


Fig. 4 Proposed three-axis taxonomy categorizing Visual World-Action Models across spatial state representations (Axis I), action-world architectural coupling (Axis II), and temporal transition engines (Axis III).

2.1 Axis I: Spatial State Representation

The choice of state space z_t governs the trade-off between photorealism, 3D geometric accuracy, and rollout speed:

1. **Latent embedding spaces.** Methods such as DreamerV3 [24] and V-JEPA 2 [30] compress observations into compact feature tokens via variational bottlenecks or joint-embedding predictive architectures (JEPA). By discarding unpredictable pixel textures (e.g., foliage motion or specular glare), latent WAMs achieve low memory overhead and rapid rollouts.
2. **2D pixel / video-generative spaces.** Video-based WAMs (e.g., UniSim [5], GAIA-1 [7], Cosmos [26]) represent states in pixel or VAE latent space. While compatible with internet-scale video pretraining, 2D predictions lack explicit 3D projective constraints.
3. **3D / 4D explicit geometry.** Geometry-aware WAMs lift 2D image features into metric 3D representations—such as 4D voxel occupancy grids (OccWorld [27]), point clouds (3D-VLA [40]), or dynamic 3D Gaussian Splatting (ManiGaussian [28])—to preserve multi-view epipolar consistency and collision boundaries.

2.2 Axis II: Action-World Coupling

Figure 5 contrasts reactive VLAs against the three primary structural paradigms for coupling future prediction with control:

1. **Cascaded (world $\rightarrow$ policy).** A pretrained video simulator generates future visual subgoals $\hat{o}_{t+1:t+H}$ conditioned on language l, which are decoded into motor commands by an inverse dynamics model (IDM) (e.g., UniPi [32]). Modular pretraining on actionless internet video is straightforward, though synthesized subgoal artifacts can propagate into the IDM.
2. **Joint end-to-end generative.** A unified backbone—typically a Diffusion Transformer (DiT) or autoregressive VLM—simultaneously models future visual/spatial tokens and action tokens within a shared attention space (e.g., DreamZero [38], WorldVLA [36]). Joint modeling enables action gradients to refine physical world representations directly.
3. **MPC / planning-in-the-loop.** The world model acts purely as a forward dynamics evaluator $z_{t+1} = f_\phi(z_t, a_t)$. At test time, optimizers (e.g., CEM, MPPI) sample candidate action sequences and select trajectories minimizing goal distance (e.g., V-JEPA 2-AC [30]).

2.3 Axis III: Temporal Dynamics Engine

Temporal transition backbones dictate both multi-step physical coherence and closed-loop inference latency:

1. **Spatio-temporal diffusion and flow matching.** Performs iterative denoising across space-time volumes conditioned on ego-motion or end-effector trajectories [26, 33]. Flow-matching variants reduce sampling steps while preserving sharp contact dynamics.

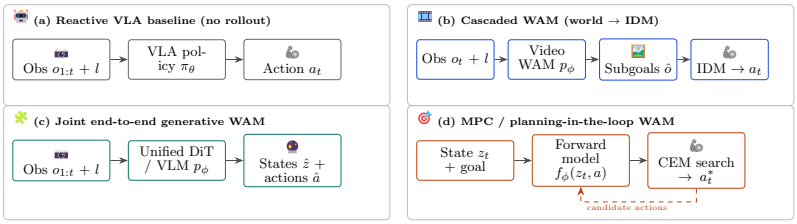


Fig. 5 Comparison of control architectures: (a) purely reactive VLAs lack future state simulation, whereas WAMs incorporate predictive dynamics via (b) cascaded video subgoal synthesis, (c) joint end-to-end visual-action generation, or (d) test-time MPC trajectory optimization.

2. **Autoregressive token transformers.** Discretizes multi-view images and actions via VQ-VAE or FSQ tokenizers, casting future prediction as causal next-token sequence modeling [7, 8].
3. **State-space recurrent latent engines.** Employs deterministic-stochastic recurrent state-space models (RSSM) or selective state-space layers (Mamba) for $O(1)$ per-step inference complexity during high-frequency closed-loop control [24].

2.4 Action-Conditioning Mechanisms and Pretraining Curricula

A central architectural question in WAM design is how low-dimensional motor commands $a_t \in \mathcal{A}$ (e.g., 7-DoF end-effector deltas or ego-vehicle steering angles) steer high-capacity spatio-temporal backbones without washing out visual priors. Existing architectures converge on three conditioning interfaces: (i) *Adaptive Layer Normalization* (AdaLN-Zero), where projected action embeddings regress scale, shift, and residual gate parameters $(\gamma_t, \beta_t, \alpha_t)$ inside each DiT block (common in DiT-based world models such as Cosmos [26]); (ii) *interleaved token conditioning*, where discrete VQ/FSQ action codes are embedded directly into the causal attention sequence alongside visual patches (GAIA-1 [7], WorldVLA [36]); and (iii) *unsupervised latent action quantization*, where an inverse-dynamics encoder infers discrete pseudo-actions $\tilde{a}_t = q_\psi(o_t, o_{t+1})$ from unlabeled video before grounding onto physical robot trajectories via a lightweight linear or MLP adapter (Genie [8], LAPA [31]).

To amplify action controllability during sampling without sacrificing visual coherence, diffusion-based WAMs increasingly employ Classifier-Free Action Guidance (CFAG). By randomly dropping action conditioning ($a_t \to \varnothing$) during training, inference blends unconditional physical priors with action-steered velocity fields using a scalar guidance weight $s_a \geq 1$, ensuring predicted trajectories respond faithfully to counterfactual steering or gripper commands:

$$\tilde{\epsilon}_\phi(z_t, a_t, l) = \epsilon_\phi(z_t, \varnothing, l) + s_a\big(\epsilon_\phi(z_t, a_t, l) - \epsilon_\phi(z_t, \varnothing, l)\big). \tag{4}$$

- **Two-stage curriculum (video → action):** large-scale action-free video pretraining followed by comparatively small action post-training preserves broad intuitive physics while adapting to embodiment-specific kinematics; for example, V-JEPA 2 is pretrained on over one million hours of internet video and then post-trained into V-JEPA 2-AC with less than 62 hours of unlabelled DROID robot video [30].

- **Latent action alignment:** VQ-based latent actions bridge human demonstration videos and multi-DoF robot actuators when paired teleoperation trajectories are scarce.

2.5 Visual-Action Tokenization Bottlenecks

Across both autoregressive and diffusion WAMs, tokenization forms the primary information bottleneck linking visual perception and motor control. Video-generative WAMs typically compress spatio-temporal clips through causal 3D Variational Autoencoders (3D VAEs) with typical 8×8 spatial and $4\times$–$8\times$ temporal downsampling [26], minimizing a composite reconstruction, perceptual, and KL-divergence objective:

$$\mathcal{L}_{\text{tok}} = \left\| o_{1:T} - D(E(o_{1:T})) \right\|_2^2 + \beta_{\text{LPIPS}}\, \mathcal{L}_{\text{perc}} + \beta_{\text{KL}}\, D_{\text{KL}}\big(q(z \mid o) \,\|\, p(z)\big). \tag{5}$$

Autoregressive token transformers (GAIA-1 [7], Genie [8], MAGVIT-v2 [44]) replace continuous Gaussian latents with discrete vocabularies via VQ-VAE or Finite Scalar Quantization (FSQ) [45]. Unlike standard vector quantization, whose codebooks are prone to under-utilization and collapse as they grow, FSQ projects each latent channel onto a fixed grid of scalar levels whose product forms an implicit codebook, avoiding codebook collapse without auxiliary commitment losses [45].

On the control side, naive per-dimension, per-timestep discretization of continuous 7-DoF or 14-DoF robot actions into 256 bins (as in RT-2 and OpenVLA [3, 4]) performs poorly on dexterous, high-frequency control data and inflates sequence length for multi-step action chunks [46]. To eliminate this redundancy, frequency-domain action tokenizers such as FAST [46] apply the Discrete Cosine Transform (DCT) followed by byte-pair encoding across action chunks $a_{t:t+H-1}$, compressing correlated action sequences before tokenization; FAST reduces VLA training time by up to $5\times$ and enables autoregressive VLAs on high-frequency dexterous tasks where binning fails [46]. Harmonizing visual tokenizers and frequency-space action tokens within a shared transformer vocabulary remains an active design frontier (Fig. 6):

- **Causal 3D spatio-temporal VAEs:** enforce temporal causality during video encoding so frame t never leaks future frames $t + \tau$, allowing streaming frame-by-frame closed-loop rollouts.
- **FSQ vs. VQ-VAE discrete bottlenecks:** FSQ eliminates exponential moving average (EMA) codebook updates and scales to large implicit vocabularies (e.g., the 64,000-entry FSQ vocabulary of the Cosmos discrete tokenizers [26]).
- **DCT / FAST action compression:** transforms smooth multi-step joint trajectories into compact spectral tokens, shortening autoregressive action sequences [46].

Key takeaways

- WAMs differ from VLAs by predicting future states $z_{t+1:t+H}$ in addition to actions (Eqs. (1)–(2)).

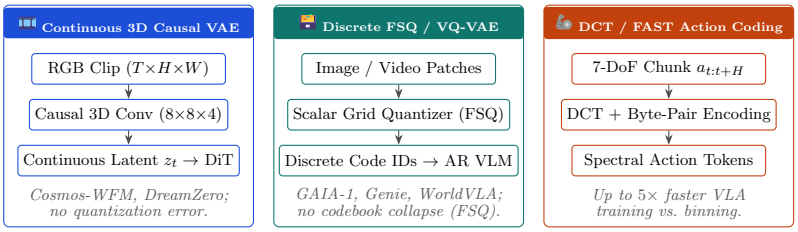


Fig. 6 Visual and action tokenization pipelines (Section 2.5): contrasting continuous 3D causal VAE latents for DiTs, discrete FSQ/VQ-VAE visual vocabularies for autoregressive VLMs, and DCT-based FAST frequency-space action compression.

- The design space is spatial state × action–world coupling × temporal engine; the coupling decides whether the world model acts as a simulator, a planner, or a training signal.

---

3 Visual and Spatial Representations

A defining challenge in visual computing is selecting a state representation that balances photorealistic appearance with geometric and physical fidelity (Fig. 7).

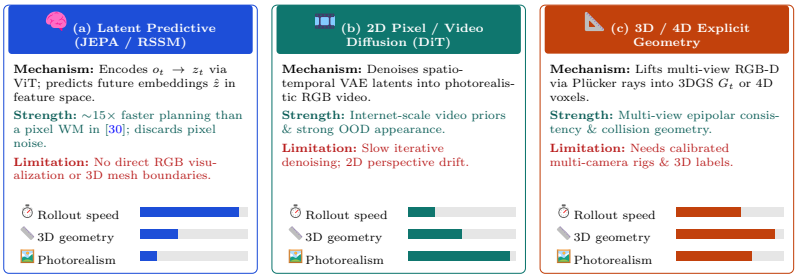


Fig. 7 Visual-computing trade-offs across the three primary spatial state representations used in World-Action Models, contrasting inference speed, photorealism, and 3D geometric consistency. Bar lengths are qualitative.

3.1 2D Video Diffusion vs. Latent Predictive Models

Video-generative WAMs (such as UniSim [5], DriveDreamer [33], and Cosmos [26]) synthesize future RGB frames directly using latent diffusion transformers (DiTs). While visually compelling and naturally compatible with internet-scale video pretraining, pixel-space generation incurs heavy iterative-denoising latency and frequently suffers from object permanence drift when objects become temporarily occluded. Conversely, Joint-Embedding Predictive Architectures (e.g., V-JEPA 2 [30]) predict abstract feature representations rather than raw pixels, discarding unpredictable high-frequency textures; in the V-JEPA 2 study, latent planning took about 16 s per action versus about 4 min for a Cosmos-based video-generation planner on the same robot tasks (∼15×) [30].

3.2 3D-Aware and 4D Occupancy World Models

To eliminate 2D perspective distortion in multi-camera robotics and autonomous driving, recent WAMs lift visual features into explicit 3D or 4D representations:

- **4D voxel occupancy forecasting:** models such as OccWorld [27] predict future 3D semantic occupancy grids conditioned on ego-vehicle trajectories, directly encoding free space and collision boundaries without rendering RGB textures.
- **3D Gaussian Splatting (3DGS) dynamics:** combining 3DGS [47] with learned particle or rigid-body deformation fields (e.g., ManiGaussian [28]) enables real-time novel-view rendering and physically grounded manipulation rollouts.

3.3 Multi-View Geometric Consistency

In multi-camera robotic manipulation (e.g., wrist + third-person cameras) and autonomous driving (e.g., 6-camera surround rigs), independent 2D frame prediction frequently violates cross-view epipolar geometry. Modern visual WAMs enforce spatial coherence through three mechanisms:

- **Plücker ray embeddings:** camera intrinsics K and extrinsics $[R \mid t]$ are encoded per pixel as 6D Plücker ray coordinates $(d, m = o \times d)$ and injected into spatial attention layers for view-consistent synthesis.
- **Cross-view epipolar attention:** restricts spatial attention across simultaneous camera views along projected epipolar lines, preventing object duplication across overlapping fields of view.
- **Deformable 3D Gaussian fields:** updates anisotropic 3D Gaussians $G = \{(\mu_i, \Sigma_i, \alpha_i, c_i)\}_{i=1}^{N}$ via an SE(3) neural deformation field conditioned on robot joint velocities, inheriting the real-time (≥ 30 FPS at 1080p) differentiable rasterization of 3DGS [47].

3.4 Explicit Lifting Equations and Visuo-Tactile World States

Formally, for any calibrated camera with intrinsics K and extrinsics $[R \mid t]$, each pixel coordinate $\tilde{u} = [u, v, 1]^\top$ maps to a 6D Plücker ray $r_u = (d_u, m_u)$ encoding unit ray direction and moment about the world origin, while deformable 3DGS world models propagate scene primitives G_t through an SE(3) neural deformation field T_ψ conditioned on action commands and joint velocities $\dot{q}_t$:

$$r_u = (d_u, m_u) = \left(\frac{RK^{-1}\tilde{u}}{\|RK^{-1}\tilde{u}\|_2},\ t \times d_u \right), \qquad G_{t+1} = T_\psi(G_t, a_t, \dot{q}_t). \tag{6}$$

However, even geometrically consistent 3D visual representations suffer from severe perceptual aliasing during contact-rich manipulation (e.g., peg-in-hole insertion, cable routing, or deformable cloth folding), where the robot end-effector occludes the exact contact patch. To resolve unobservable contact forces, emerging multimodal WAMs augment o_t with high-resolution optical tactile skins (e.g., GelSight/DIGIT deformation maps) and 6-axis wrist force-torque histories. Predicting joint visuo-tactile future

states enables models to distinguish a rigid slip event from compliant deformation purely within latent rollout space.

- **Occlusion-resilient contact grounding:** jointly forecasting RGB-D and tactile shear fields prevents generative WAMs from hallucinating object penetration when fingers block camera sightlines.
- **Cross-modal latent synchronization:** asymmetric temporal rates (30 Hz RGB vs. 1 kHz force-torque) are aligned via temporal window pooling inside multi-modal tokenizers.

3.5 Object-Centric Particles and Neural SDFs

Beyond voxel occupancy grids [27] and rigid-body 3D Gaussian Splatting [28], manipulation of fluids, granular media, ropes, and cloth requires continuous continuum-mechanics representations. PhysGaussian [42] couples anisotropic 3D Gaussians directly with Material Point Method (MPM) continuum solvers, enabling Newtonian stress-strain simulation directly on reconstructed visual primitives without intermediate tetrahedral meshing. In parallel, object-centric world models such as SlotFormer [48] factorize scenes into discrete entity slots $z_t = \{s_t^{(1)}, \ldots, s_t^{(M)}\}$ that interact via graph-structured transformer attention, preserving object permanence even under temporary occlusion. For rigid obstacle avoidance, Neural Signed Distance Fields (SDFs) Φ_ϕ complement generative rendering with smooth distance gradients regularized via the Eikonal constraint, yielding differentiable minimum-distance collision margins. Together with an epipolar term over fundamental matrix F_{uv}, this defines the geometric regularizer used in Eq. (3):

$$\mathcal{L}_{\text{geo}} = \lambda_{\text{epi}} \left\| p_v^\top F_{uv}\, p_u \right\|_2^2 + \lambda_{\text{eik}}\, \mathbb{E}_x \Big[\big(\|\nabla_x \Phi_\phi(x, z_t)\|_2 - 1 \big)^2 \Big]. \tag{7}$$

3.6 Viewpoint and Egocentric–Exocentric Transfer

A pervasive visual-computing challenge in robotic WAMs is camera-viewpoint non-stationarity. Unlike autonomous driving rigs with fixed roof-mounted extrinsics, bimanual and mobile manipulators rely heavily on wrist-mounted cameras whose extrinsics $[R_t \mid t_t]$ change at every control step as a direct function of action a_t. Furthermore, scaling WAMs via abundant human egocentric datasets (e.g., Ego4D's 3,670 hours of first-person video [49]) introduces an embodiment-viewpoint gap between human eye-level streams and robot third-person or wrist perspectives. Recent approaches bridge this gap using dense 3D point-track conditioning (Track2Act [50], CoTracker3 [51]) and goal-conditioned human video synthesis (Gen2Act [52]), decoupling invariant scene motion from camera ego-motion via relative Plücker ray embeddings.

Key takeaways

- Latent states are fast but hard to inspect; video states are photorealistic but slow and physically weak; 3D/4D states give metric collision checks at higher data and annotation cost.

Table 2 Comparative taxonomy of 3D and 4D scene parameterizations in Visual World-Action Models (qualitative comparison).

Representation	Representative models	Rendering / query speed	Collision & contact fidelity	Deformable object support
2D latent / pixel	UniSim [5], DreamZero [38]	Slow (iterative denoising)	Weak (2D perspective drift)	Appearance only (morphs)
Latent JEPA tokens	V-JEPA 2-AC [30], DreamerV3 [24]	Fast (no pixel decoding)	Implicit (via goal/cost)	Implicit feature tracking
3D point clouds	3D-VLA [40]	Moderate	Moderate (sparse holes)	Limited (rigid / semi-rigid)
4D semantic voxels	OccWorld [27], Occ3D [53]	Moderate (voxel grid)	High (metric free-space)	Coarse voxel discretization
Dynamic 3DGS + MPM	ManiGaussian [28], PhysGaussian [42]	Real-time raster ($\geq$30 FPS [47])	High (epipolar + density)	Strong (MPM / SE(3) fields)
Object slots / SDFs	SlotFormer [48], neural SDFs	Fast (compact state)	Exact (analytic distance)	Entity-level dynamics

- Recent systems mix representations, e.g. video priors with 3D Gaussians or occupancy, to trade realism against geometric control.

4 Master Comparison and Datasets

To serve as a rapid reference for empirical comparisons, Tables 3 and 4 summarize 28 representative Visual World-Action Models (16 core systems and 12 extended entries, six of them driving WAMs) across architecture, spatial modality, action interface, and primary application domain. Table 5 catalogs 18 core training datasets, simulation suites, and physics evaluation benchmarks, and Fig. 8 summarizes the paradigm trade-offs and the composition of the surveyed models.

4.1 Extended Architecture Profiles: Action Interfaces, Horizons and Latency

Beyond the 16 core systems cataloged in Table 3, rapid diversification across interactive simulation and unsupervised video pretraining has introduced specialized paradigms—notably pixel-level EDM diffusion for crisp high-frequency control (DIAMOND [35]), multi-modal controllable driving simulators (Vista [34], GenAD [41]), interactive autoregressive tokenizers (iVideoGPT [56]), unsupervised latent-action VLMs (LAPA [31]), and sparse-memory 4DGS manipulation models (EnerVerse [43]). Table 6 synthesizes the practical deployment envelope across representative WAM families by contrasting their action parameterization, the control rate or horizon reported by the original authors, and open-weight availability.

Late-2026 developments.

Work released between April and October 2026 sharpens the representation question at the core of this survey. Fast-WAM [83] finds that removing video co-training hurts far more than removing test-time future generation, and skips imagination at inference

Table 3 Representative Visual World-Action Models (2023–2026): 16 core systems. Extended entries are listed in Table 4.

Model	Year	Spatial rep.	Dynamics engine	Action coupling	Domain & key contribution
DreamerV3 [24]	2023	Latent (RSSM)	Recurrent SSM	Joint (actor-critic)	General RL; one configuration across 150+ tasks (Nature 2025).
GAIA-1 [7]	2023	2D video tokens	Autoregressive trans.	Joint (tokenized)	Driving; multimodal video-text-action synthesis.
UniPi [32]	2023	2D pixel frames	Video diffusion	Cascaded (IDM)	Robotics; text-guided video plans decoded to actions.
UniSim [5]	2024	2D pixel/latent	Video diffusion	Cascaded (simulator)	Universal simulator unifying manipulation & navigation.
Genie / Genie 2 [8, 54]	2024	2D video tokens	Spatio-temp. trans.	Latent action space	Playable worlds from unlabeled video (11B params).
DriveDreamer [33]	2024	2D + HD map/3D box	Structured diffusion	Joint trajectory	Driving; controllable multi-view traffic synthesis.
V-JEPA [9]	2024	Latent features	Masked ViT predictor	Cascaded / probing	Self-supervised visual dynamics without pixel loss.
3D-VLA [40]	2024	3D point / RGB-D	3D generative LLM	Joint embodied	3D perception, goal point-cloud generation & control.
OccWorld [27]	2024	4D semantic occ.	Autoregressive trans.	Joint ego-motion	Driving; joint 3D occupancy & trajectory rollout.
ManiGaussian [28]	2024	Dynamic 3DGS	SE(3) deformation	Joint policy	Manipulation; semantic Gaussian splatting world model.
Cosmos-WFM [26]	2025	2D/3D-aware tokens	Diffusion & AR	Joint / post-trained	WFMs curated from 20M h of video (∼100M clips).
V-JEPA 2-AC [30]	2025	Latent features	Masked ViT predictor	MPC planning	Zero-shot robot planning via action-conditioned JEPA.
WorldVLA [36]	2025	2D visual tokens	Autoregressive VLM	Joint action-image	Unified action and world model autoregression.
UnifoLM-WMA-0 [55]	2025	2D video	Video diffusion	Dual (sim + policy)	Open-source WMA framework deployed on Unitree robots.
Genie Envisioner [37]	2025	Multi-view video latent	Video diffusion	Joint (GE-Act decoder)	GE-Base WM + flow-matching GE-Act + GE-Sim neural simulator.
DreamZero [38]	2026	2D video latent	Wan-2.1 DiT (14B)	Joint video-action	Joint video-action prediction; real-time closed loop at 7 Hz.

Table 4 Extended Visual World-Action Model entries (2024–2026): general models (Section 4.1) followed by driving WAMs (Section 4.4).

Model	Year	Spatial rep.	Dynamics engine	Action coupling	Domain & key contribution
DIAMOND [35]	2024	2D pixel (EDM)	Pixel diffusion	Joint (RL policy)	High-frequency diffusion world model preserving crisp visual details.
Vista [34]	2024	2D high-res latent	Video diffusion	Multi-modal action	Generalizable driving WM controllable via steering, trajectory & text.
GenAD [41]	2024	Instance-centric BEV	Structured VAE + GRU	Joint planning	Generative end-to-end driving via instance-centric trajectory priors.
iVideoGPT [56]	2024	2D VQ tokens	Autoregressive GPT	Tokenized obs-act	Interactive autoregressive world model over observation-action streams.
LAPA [31]	2025	2D + latent action	VLM + VQ latent act.	Latent pretraining	Unsupervised latent action quantization from unlabeled human videos.
EnerVerse [43]	2025	4D Gaussian + video	Chunkwise diffusion	Sparse memory + policy	Embodied foundation WM pairing sparse context memory with 4DGS.
Drive-WM [57]	2024	Multi-view video	Video diffusion	Maneuver rollouts + reward	Driving; multiview forecasting used to select trajectories by image-based rewards.
LAW [58]	2025	Latent scene features	Latent predictor	Self-supervised aux. task	Driving; predicts future features from current features and ego trajectory.
WoTE [59]	2025	BEV semantic maps	BEV world model	Trajectory evaluation	Driving; scores candidate trajectories by imagined BEV futures.
World4Drive [60]	2025	Latent + VFM priors	Intention-aware latent WM	Generate & select	Driving; perception-annotation-free planning via latent world model.
Epona [61]	2025	2D video latent	Autoregressive diffusion	Joint trajectory + video	Driving; minutes-long generation; WM doubles as real-time planner.
DriveVLA-W0 [62]	2026	Image tokens / features	AR or diffusion WM	Joint VLA + world modeling	Driving; future-image prediction as dense supervision for VLAs.

(190 ms latency, over 4× faster than imagine-then-execute WAMs). A controlled study of latent versus explicit WAMs [84] reports that latent models match explicit ones in distribution but lose their generalization advantage, with most of the gap arising in the first denoising step. Representation-centric designs follow: ReWAM [85] builds compact world states from DINO features and reaches 93.6% on RoboTwin 2.0 without generative video pretraining, Box-WAM [86] replaces dense future frames with oriented box trajectories of task-relevant objects, and X-WAM [87] lifts video priors to unified

Table 5 Core datasets, closed-loop simulators, and physics evaluation benchmarks (18 suites).

Category	Benchmark / dataset	Modalities	Scale & primary evaluation role
Manipulation	Open X-Embodiment [63]	RGB, depth, joint	1M+ trajectories, 22 embodiments, 527 skills.
Manipulation	DROID [64]	Multi-view stereo RGB	76k trajectories (350 h), 564 scenes, 84 tasks.
Manipulation	BridgeData V2 [65]	RGB, proprioception	60,096 trajectories across 24 environments.
Manipulation	LIBERO / LIBERO-Plus [66, 67]	Multi-view RGB, prop.	130 tasks in 4 suites; 7 perturbation axes (Plus).
Manipulation	CALVIN (ABC→D) [68]	RGB-D, tactile, prop.	Zero-shot long-horizon language-conditioned chaining.
Manipulation	RoboTwin 2.0 [69]	Multi-view RGB-D (sim)	50 bimanual tasks with strong domain randomization.
Manipulation	RoboCasa [70]	Photorealistic RGB-D	100 household (kitchen) tasks for generalist policies.
Driving	nuScenes [71]	6-cam RGB, LiDAR, radar	1,000 scenes; standard occupancy & L2 planning benchmark.
Driving	Waymo Open Dataset [72]	5-cam RGB, LiDAR	1,150 scenes of 20 s; perception & sensor simulation.
Driving	OpenScene / Occ3D [53]	6-cam RGB, 3D voxels	Dense 3D semantic occupancy ground truth for 4D WAMs.
Driving	NAVSIM [73]	Multi-cam, BEV	Non-reactive simulation-based scoring (PDMS) of planners.
Driving	DrivingGen [74]	Video + ego trajectory	Visual, trajectory & controllability metrics for driving WMs.
Navigation	Habitat 3.0 [75]	Photorealistic 3D	Human-robot social navigation and rearrangement.
Navigation	HM3D / MP3D [76, 77]	3D mesh, RGB-D	1,000 (HM3D) and 90 (MP3D) building-scale scans.
Physics & WAM	WorldModelBench [78]	Video + action	Evaluates action-following & physical consistency in video WMs.
Physics & WAM	WorldArena [79]	Video + embodied tasks	16 perceptual metrics + data-engine, policy-evaluator & planner utility.
Physics & WAM	Physics-IQ / IntPhys 2 [80, 81]	Video	Physical understanding (mechanics, fluids, optics; permanence, solidity).
Physics & WAM	VBench-2.0 [82]	Generated video	18 dimensions incl. physics & commonsense faithfulness.

4D world-action modeling (79.2% on RoboCasa, 90.7% on RoboTwin 2.0). AW-0 [88] adds a third generative stream of code-based skill programs alongside future video and actions.

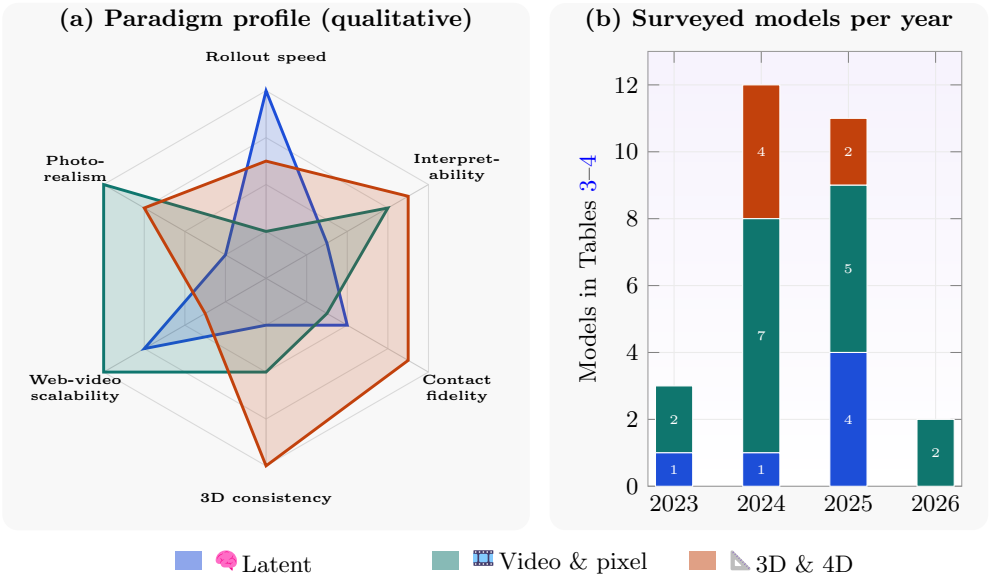


Fig. 8 (a) Qualitative profile of the three spatial-state paradigms, summarizing the trade-offs discussed in Sections 3–5 (axes are ordinal scores assigned by the authors, not measurements). (b) Distribution of the 28 models in Tables 3–4 by listed year and spatial representation: video/pixel-space WAMs dominate, 3D/4D models concentrate in 2024, and latent WAMs grow in 2025 with driving planners.

Table 6 Extended WAMs and deployment-relevant properties as reported by the original papers (n.r. = not reported in the cited source).

Model	Action interface	Reported rate / horizon	Open weights
DIAMOND [35]	Discrete game actions (Atari)	n.r.	Yes
Vista [34]	Steering & speed, trajectory, command, goal point	n.r.	Yes
GenAD [41]	Instance-centric BEV waypoints	3 s planning horizon (nuScenes)	Yes
iVideoGPT [56]	Tokenized robot actions	n.r.	Yes
LAPA [31]	VQ latent actions → robot actions	n.r.	Yes
EnerVerse [43]	Action chunks + multi-view future space	n.r.	Partial
V-JEPA 2-AC [30]	End-effector deltas via CEM planning	∼16 s per planned action	Yes
DreamZero [38]	Joint video-action chunks	7 Hz closed loop	Yes

4.2 Data Scale, Action-Free Video Mining and Synthetic Sim-to-Real Pipelines

A defining lesson of 2024–2026 WAM scaling is the acute asymmetry between video availability and action-labeled embodied data. Whereas web-scale video corpora supply tens of millions of hours of passive physical interactions (gravity, pouring, tool use, traffic flow), real-world robot teleoperation corpora such as Open X-Embodiment [63] (1M+ trajectories across 22 embodiments) and DROID [64] (76k trajectories, 350 h)

remain orders of magnitude smaller. Consequently, modern pipelines adopt a three-pillar data curriculum: (1) massive action-free video pretraining (e.g., the Cosmos WFMs, curated from a 20M-hour video collection into ∼100M clips [26]; Ego4D [49] supplying 3,670 hours of first-person footage), (2) unsupervised inverse-dynamics or latent-motion pseudo-labeling (LAPA [31], Moto [89]), and (3) GPU-accelerated synthetic rollout generation in photorealistic simulators such as ManiSkill3 [90], RoboCasa [70], and Cosmos-Drive-Dreams [91].

Synthetic-to-real transfer remains double-edged: while procedural domain randomization improves robustness to unseen lighting and distractor clutter, simulated rigid-body contacts frequently exhibit simplified Coulomb friction and infinite stiffness that diverge from compliant real-world grippers. To calibrate simulated dynamics against physical hardware, SimplerEnv [92] aligns simulator camera intrinsics, robot control gains, and green-screen real backgrounds, achieving strong correlation between simulated and real-world policy performance (summarized across WAM data recipes in Table 7).

Table 7 Pretraining data mix, parameter scale, and action labeling strategies across representative WAMs.

Model	Pretraining corpus / scale	Param. scale	Action grounding mechanism	Primary eval. setting
DreamerV3 [24]	Online RL interaction (150+ tasks)	12M–400M	Direct env. action + reward	Atari, DMC, Minecraft
GAIA-1 [7]	∼4,700 h proprietary UK driving	∼9B total	Discretized speed & curvature	Urban driving rollout
Cosmos-WFM [26]	∼100M clips from 20M h video	4B–14B (diffusion & AR)	Post-trained action adapters	Driving & manipulation
V-JEPA 2-AC [30]	1M+ h internet video + DROID	∼1B ViT-g encoder	<62 h robot video (CEM-MPC)	Zero-shot Franka pick-place
Genie Envisioner [37]	∼3,000 h, 1M+ AgiBot World episodes [93]	Video DiT	Flow-matching GE-Act decoder	Dual-arm manipulation
DreamZero [38]	Pretrained Wan video model + heterogeneous robot data	14B	Joint video-action prediction	Unseen real-robot tasks
LAPA [31]	Unlabeled human video + OXE	7B VLM	VQ latent action pseudo-labels	SimplerEnv & real WidowX
EnerVerse [43]	Manipulation video (+ 4DGS data engine)	Chunk-wise video diffusion	Action chunks with sparse memory	Long-horizon manipulation

4.3 Legged Locomotion, Humanoid Whole-Body Control and Cross-Domain Envelopes

While most early visual WAMs targeted planar autonomous driving or static-base tabletop arms, extending predictive world models to quadrupeds and bipedal

humanoids introduces coupled floating-base dynamics, discontinuous ground impacts, and 20–40+ actuated degrees of freedom. DayDreamer [29] demonstrated that a latent RSSM world model can train a physical A1 quadruped to walk from scratch in roughly one hour of real-world experience by learning compact proprioceptive-visual transitions without risking destructive policy exploration. Building on this paradigm, World Model-based Perception (WMP) [94] trains an RSSM-style world model over depth images entirely in simulation and uses its latent state for visual locomotion; the simulated world model still predicts real-world trajectories, and the policy transfers to a Unitree A1 robot.

For full-scale humanoid loco-manipulation, recent frameworks such as UnifoLM-WMA-0 [55] unify multi-view egocentric video generation with dual policy heads that couple whole-body balance stabilization and bimanual object interaction. As illustrated in Fig. 9, embodied domains occupy markedly different regions of the spatio-temporal design space: driving requires metric lookahead over tens of metres with moderate occlusion, dexterous manipulation requires fine contact fidelity under severe hand occlusion, and legged locomotion requires fast proprioceptive-visual feedback to prevent falls (the ranges in Fig. 9 are indicative engineering values, not measurements).

- **Floating-base proprioceptive fusion:** legged WAMs fuse low-rate exteroceptive inputs (e.g., depth) with high-rate proprioception; WMP, for instance, updates its world-model state at a lower rate than the policy to fit onboard compute [94].
- **Whole-body loco-manipulation:** requires predicting how upper-body payload dynamics shift center-of-mass stability margins during concurrent bipedal walking.

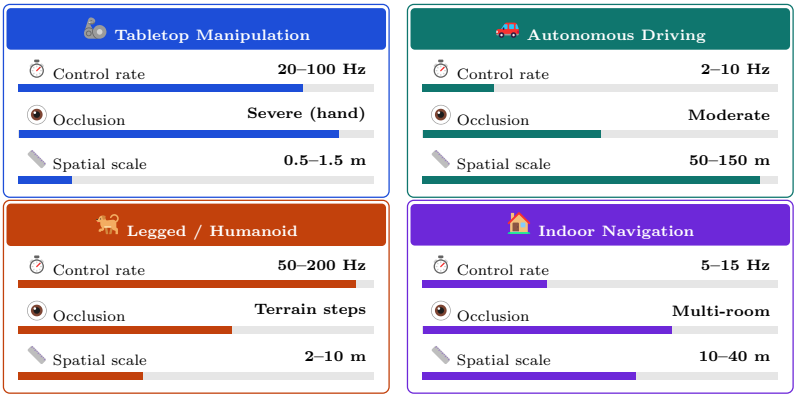


Fig. 9 Cross-domain operational envelopes (Section 4.3): indicative spatial horizon, control rate, and occlusion severity across tabletop manipulation, autonomous driving, legged locomotion, and indoor navigation. Values are typical engineering ranges, not measured results.

4.4 World-Action Models for Autonomous Driving

Driving is where WAMs meet safety-critical deployment most directly: the ego vehicle must predict how traffic evolves under its own candidate actions, at a few hertz, over tens to hundreds of metres. Four families of driving WAMs can be distinguished by what the world model predicts and how its prediction reaches the planner (Fig. 11):

- **Video-generative driving simulators** predict future (multi-view) camera video conditioned on ego trajectory, commands, or layout: GAIA-1 [7], DriveDreamer [33], Drive-WM [57], Vista [34], and Epona [61]. Besides synthesizing training data (Cosmos-Drive-Dreams [91]), Drive-WM and Epona use their rollouts for planning, selecting trajectories by image-based rewards or predicting them jointly with the video.
- **Occupancy and BEV world models** forecast 3D semantic occupancy or BEV semantics, giving metric free space in which collisions can be checked directly: OccWorld [27], DOME [95], and WoTE [59], which scores candidate trajectories by their imagined BEV futures.
- **Latent world models for end-to-end planning** predict future scene features as a self-supervised signal: LAW [58] conditions the prediction on the ego trajectory, and World4Drive [60] generates and selects intention-specific trajectories without perception annotations.
- **World-model-augmented driving VLAs** add future-image prediction to VLA training so that large models receive denser supervision than sparse actions alone (DriveVLA-W0 [62]). In 2026, joint video–action driving WAMs such as SimWAM [96] pair a pretrained video backbone with an action expert, hierarchical slow–fast world models appear in Drive-HWM [97], and WALT [98] aligns planner trajectories with a world-model latent space.

Across these families the world model is increasingly used as a *trajectory evaluator* rather than only as a renderer, which shifts the relevant question from "does the video look real?" to "does the prediction change correctly with the action?". Evaluation has moved accordingly: open-loop nuScenes L2 and collision rates [71] are sensitive to protocol (GenAD reports an average L2 of 0.91 m or 0.52 m depending on how horizons are averaged [41]), so recent work reports the simulation-based PDM score of NAVSIM [73], 4D occupancy forecasting on Occ3D [53], and generation metrics on DrivingGen [74]. Section 5.2 compiles the reported numbers, and Section 5.4 adds tests of whether forecasts actually depend on the action.

Key takeaways

- Video WAMs, foundation-scale data, and scarce action labels define the 2024–2026 landscape (Tables 3–7).
- Driving WAMs fall into four families; most now use the world model to score candidate trajectories rather than only to render video.
- Reported driving results depend strongly on protocol, so compare within one benchmark and one protocol only.

5 Quantitative Benchmarks and Analysis

Evaluating WAMs requires measuring both perceptual realism and downstream control utility. A recurring observation across recent benchmarks is the weak correlation between visual generation metrics and physical task success (Fig. 10).

5.1 Perceptual vs. Closed-Loop Metrics

Standard video generation evaluations diverge sharply from embodied control reliability when contact dynamics or kinematic feasibility are violated.

- **Perceptual vs. physical metrics:** Fréchet Video Distance (FVD), PSNR, and LPIPS measure visual sharpness and temporal smoothness, but fail to penalize physically impossible events (e.g., robotic grippers passing through rigid objects or vehicles violating kinematic bicycle constraints).
- **Closed-loop policy evaluation:** state-of-the-art WAMs increasingly report closed-loop task success rates on simulation suites (e.g., LIBERO-Long, CALVIN ABC→D) and open-loop/closed-loop L_2 trajectory error and collision rates on nuScenes.

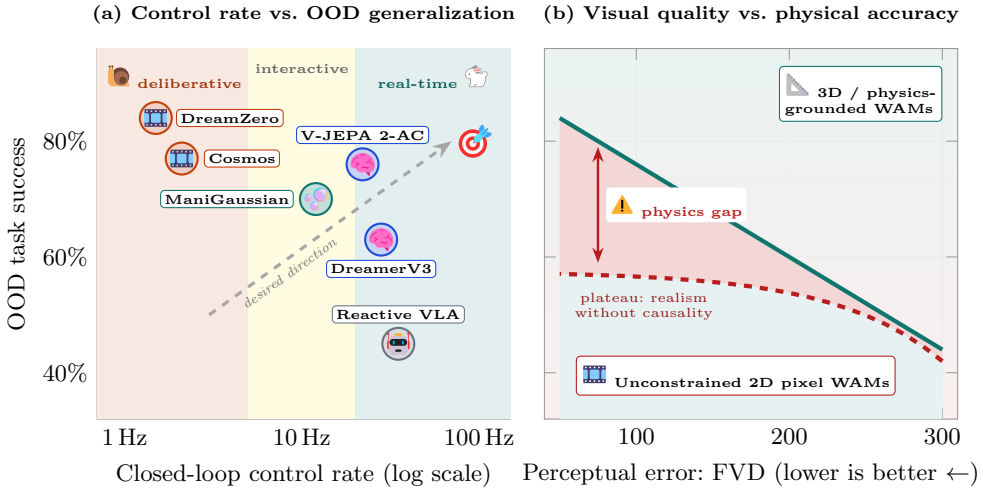


Fig. 10 Schematic illustration of trade-offs reported across 2024–2026 studies (positions are qualitative, not measured values): (a) video-generative WAMs attain strong OOD task success but run at low control rates, whereas latent and 3DGS WAMs reach higher rates; (b) improving FVD alone saturates physical task success when 2D models hallucinate contact dynamics.

5.2 Reported Results on Driving and Manipulation Benchmarks

Tables 8–10 compile results *as reported by the original papers* (or, where marked, as compiled by a cited paper); we did not re-run any model. Protocols differ across papers, so entries should be compared only within the same table block. Four observations stand out:

- **World models as trajectory scorers help on NAVSIM.** Scoring candidates with a BEV world model (WoTE, PDMS 88.3) outperforms both a single-trajectory baseline (TransFuser, 84.0) and model-free scoring (Hydra-MDP, 86.5), and every camera-only world-model method matches or exceeds the camera–LiDAR TransFuser baseline; the joint video–action SimWAM reaches 91.9 with a single camera.

All entries remain below the human reference (94.8), mostly through lower ego progress.

- **Occupancy forecasts decay quickly with horizon.** DOME-O drops from 35.11 mIoU at 1 s to 20.29 at 3 s, and camera-input forecasting (DOME-F, 18.25 average) lags occupancy-input forecasting (27.10) [95].
- **Generation quality has improved sharply**, from an FVD of 452.0 (DriveDreamer) to 82.8 (Epona) on nuScenes [61], but Section 5.4 explains why FVD alone does not certify physical accuracy.
- **Manipulation benchmarks are near saturation.** LIBERO averages reach 98.5% (Cosmos Policy), yet LIBERO-Plus perturbations reduce strong models from 95% to below 30% [67]. On CALVIN ABC→D, policies that predict future visual states lead (Seer 4.28, VPP 4.33, DreamVLA 4.44 average tasks completed).

Table 8 NAVSIM v1 (navtest) results as reported by the original papers. WM = how a world model is used. C = camera, L = LiDAR. NC: no at-fault collision, DAC: drivable-area compliance, TTC: time-to-collision, EP: ego progress, PDMS: PDM score (all ↑).

Method	WM usage	Input	NC	DAC	TTC	EP	PDMS
Human reference [73]	–	–	100	100	100	87.5	94.8
TransFuser [73, 99]	None (single trajectory)	C & L	97.7	92.8	92.8	79.2	84.0
Hydra-MDP [100]	None (model-free scoring)[a]	C & L	98.3	96.0	94.6	78.7	86.5
LAW [58]	Latent WM, self-supervised	C	96.4	95.4	88.7	81.7	84.6
World4Drive [60]	Latent WM, generate & select	C	97.4	94.3	92.8	79.9	85.1
Epona [61]	Video WM as planner	C	97.9	95.1	93.8	80.4	86.2
WoTE [59]	BEV WM scores candidates	C & L	98.5	96.8	94.9	81.9	88.3
DriveVLA-W0 [62]	Future-image prediction in VLA[b]	C	98.7	96.2	95.5	82.2	88.4
SimWAM [96]	Joint video–action WAM[c]	C	98.5	99.1	96.1	86.5	91.9

[a] As reported in [59]. [b] Single front camera, query-based action expert; a best-of-N variant reports 93.0. [c] Single front camera, RL-tuned, one trajectory per scene (no test-time selection).

5.3 Compute, Latency, and Robustness Trade-offs

Results reported across 2025–2026 studies point to a trade-off between visual-physical grounding and real-time control latency:

- **Generalization vs. speed:** video-generative WAMs report large generalization gains—DreamZero reports more than 2× better generalization than state-of-the-art VLAs on unseen real-robot tasks while running closed-loop at 7 Hz [38], and Cosmos

Table 9 Driving world models on nuScenes and Occ3D as reported by the original papers. (a) Open-loop planning with the averaged (VAD-style) protocol; L2 in metres, collision in %, both averaged over 1–3 s (↓). (b) 4D occupancy forecasting, average over 1–3 s (↑); "-O" = occupancy input, "-D/-F" = camera input. (c) Video generation on the nuScenes validation set (↓), as compiled in [61].

(a) Planning

Method	L2	Col.
VAD-Base [101][c]	0.72	0.22
OccWorld-O [27]	0.64	0.24
LAW[e] [58]	0.61	0.30
GenAD [41]	0.52	0.19
World4Drive [60]	0.50	0.16

(b) Occ3D forecasting

Method	mIoU	IoU
Copy & paste	11.33	20.52
OccWorld-D [27]	8.62	16.53
OccWorld-O [27]	17.14	26.63
DOME-F [95]	18.25	28.84
DOME-O [95]	27.10	36.36

(c) Generation

Method	FID	FVD
DriveDreamer [33]	52.6	452.0
Drive-WM [57]	15.8	122.7
Vista [34]	6.9	89.4
Epona [61]	7.5	82.8

[c] As reported in [41] under the averaged protocol. [e] Perception-free variant. Copy & paste repeats the last observed occupancy (baseline from [95]).

Table 10 Manipulation results as reported in the cited papers. (a) LIBERO success rate (%, ↑) per suite; values for π_0 and UniVLA as compiled in [39]. (b) CALVIN ABC→D average number of consecutive tasks completed out of 5 (↑).

(a) LIBERO

Method	Spatial	Object	Goal	Long	Avg.
OpenVLA [4]	84.7	88.4	79.2	53.7	76.5
WorldVLA [36]	87.6	96.2	83.4	60.0	81.8
π_0 [102]	96.8	98.8	95.8	85.2	94.2
UniVLA [103]	96.5	96.8	95.6	92.0	95.2
OpenVLA-OFT [104]	97.6	98.4	97.9	94.5	97.1
Cosmos Policy [39]	98.1	100.0	98.2	97.6	98.5

(b) CALVIN ABC→D

Method	Avg. len.
UniPi [32][d]	0.92
SuSIE [105]	2.69
GR-1 [106]	3.06
Seer [107]	4.28
VPP [108]	4.33
DreamVLA [109]	4.44

[d] As reported in [108].

Policy reaches 98.5% average success on LIBERO and 67.1% on RoboCasa [39]—but multi-step denoising keeps their control rates below those of lightweight reactive policies.

- **Latent rollout efficiency:** non-reconstructive latent WAMs bypass pixel decoding entirely; V-JEPA 2-AC plans in about 16 s per action versus about 4 min for a Cosmos-based video planner on the same tasks [30].
- **Training compute scaling:** pretraining video world foundation models is extremely compute-intensive—the Cosmos WFMs were trained on 20M hours of video using 10,000 NVIDIA H100 GPUs over about three months [26]—making parameter-efficient adapters and consistency distillation vital for downstream deployment.

5.4 Formalizing Physics-Grounded Metrics

To close the perceptual-versus-physical evaluation gap identified in Fig. 10b, quantitative WAM benchmarks should supplement FVD with three verifiable geometric and causal diagnostics: (1) the *Interpenetration Volume Ratio* (IVR), measuring the

fraction of predicted 3D occupancy or mesh volume overlapping rigid obstacles across horizon H; (2) *Depth-Normal Temporal Consistency* (E_{dn}), penalizing non-rigid surface morphing across consecutive synthesized views; and (3) *Action-Counterfactual Sensitivity* (Δ_{act}), which measures whether distinct candidate actions $a_t \neq a'_t$ induce proportionally distinct visual/spatial rollouts $\hat{z}_{t+1:t+H}$ rather than collapsing to an unconditional video average:

$$\mathrm{IVR} = \frac{1}{H}\sum_{\tau=1}^{H} \frac{\mathrm{Vol}\big(M_{\mathrm{agent}}(t{+}\tau) \cap M_{\mathrm{env}}(t{+}\tau)\big)}{\mathrm{Vol}\big(M_{\mathrm{agent}}(t{+}\tau)\big)},$$

$$E_{\mathrm{dn}} = \frac{1}{H-1}\sum_{\tau=1}^{H-1} \mathbb{E}_{p\in\Omega_s}\left[\frac{|\hat{D}_{\tau+1}(p) - \tilde{D}_{\tau}(p)|}{\tilde{D}_{\tau}(p)} + \lambda_n\big(1 - \hat{\mathbf{n}}_{\tau+1}(p)^{\top}\tilde{\mathbf{n}}_{\tau}(p)\big)\right], \tag{8}$$

$$\Delta_{\mathrm{act}} = \mathbb{E}_{a_t\neq a'_t}\left[\frac{\|\hat{z}_{t+1:t+H}(a_t) - \hat{z}_{t+1:t+H}(a'_t)\|_2}{\|a_t - a'_t\|_2}\right],$$

where $\hat{D}_\tau$ and $\hat{\mathbf{n}}_\tau$ are the depth and surface normals of generated frame τ, $\tilde{D}_\tau$ and $\tilde{\mathbf{n}}_\tau$ are frame τ reprojected into frame $\tau{+}1$ with the commanded camera motion, and Ω_s is the set of static pixels. A reference implementation of all three metrics accompanies this survey (see Code availability). In driving semantic forecasting, the executed action raises mid-term mIoU by at most 2.5 points while a wrong one lowers it by up to 23.2 [17]: accuracy gains show that an action is used, not how strongly forecasts depend on it.

5.5 Cross-Domain Divergence

Comparing results across robotics and driving reveals a sharp architectural bifurcation driven by physical scale and latency budgets. Autonomous driving WAMs typically plan at a few hertz over coarse spatial scales (tens to hundreds of metres) with standardized surround-camera rigs (six cameras in nuScenes [71]), favoring 4D semantic occupancy grids (OccWorld [27]) and BEV-conditioned multi-view video diffusion (DriveDreamer [33], Vista [34]). Figure 11 illustrates this driving-specific instantiation of the WAM loop. Conversely, tabletop dexterous manipulation requires fine contact fidelity under heavy hand-object occlusion at higher control rates, driving convergence toward wrist-mounted epipolar 3DGS (ManiGaussian [28]) and non-reconstructive latent MPC (V-JEPA 2-AC [30]).

5.6 Closed-Loop Benchmark Synthesis: Manipulation, Driving and Compounding Horizon Drift

Evaluating WAMs against reactive VLAs (such as RT-2 [3], OpenVLA [4], Octo [110], and flow-matching π_0 [102]) requires separating in-distribution imitation from out-of-distribution (OOD) visual and physical generalization. Across standard long-horizon tabletop benchmarks such as LIBERO-Long [66] and CALVIN ABC→D [68] (which chains five consecutive language-conditioned subtasks in an unseen environment D),

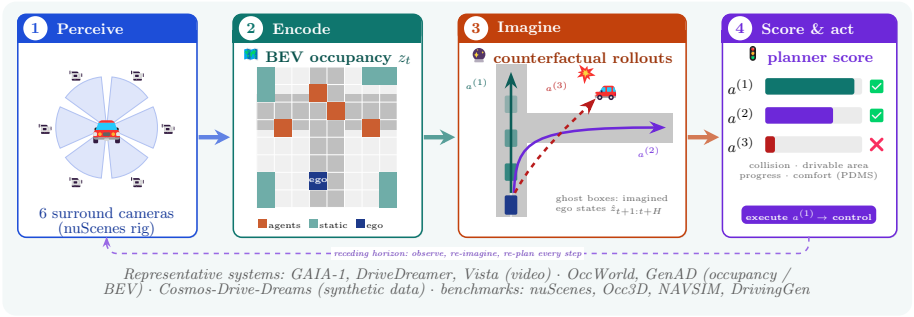


Fig. 11 World-action modeling for autonomous driving. (1) Surround-view images are (2) encoded into a bird's-eye-view or 4D occupancy world state; (3) the world model rolls out counterfactual futures for candidate ego actions; and (4) a planner scores them on safety and progress criteria such as the PDM score used by NAVSIM [73]. Representative systems: [7, 27, 33, 34, 41, 91]; benchmarks: [53, 71, 73, 74].

reactive single-step policies suffer sharp compounding drift when subtle object displacements accumulate across subtasks. In contrast, action-conditioned world models that verify subgoals or optimize trajectories over horizon H maintain higher multi-stage completion rates, although perturbation suites such as LIBERO-Plus show that success rates can collapse from about 95% to below 30% under camera and robot-state shifts [67], and distributed real-world protocols such as RoboArena [111] are emerging to rank generalist policies more reliably.

In autonomous driving, closed-loop benchmarks such as NAVSIM [73] and nuScenes [71] reveal an equally important methodological pitfall: open-loop L_2 trajectory error alone correlates poorly with real collision avoidance because ego-conditioned datasets allow policies to extrapolate smooth curves from current velocity without reacting to other agents. Consequently, 2025–2026 driving evaluations pair open-loop L_2 and 4D occupancy mIoU (Occ3D [53]) with closed-loop Predictive Driver Model Scores (PDMS) and multi-agent collision rates. Figure 12 further contrasts how prediction error scales with rollout horizon H: open-loop autoregressive pixel rollouts diverge super-linearly beyond $H > 16$ steps due to error feedback, whereas chunked flow-matching and closed-loop MPC replanning anchor predictions to fresh observations every Δt steps.

5.7 Qualitative Failure Taxonomy and Cross-Embodiment Hardware Case Studies

Beyond aggregate success rates, deploying visual WAMs on physical actuators exposes four recurring qualitative failure modes that seldom appear in static video benchmarks (summarized in Table 12 and Fig. 13). First, *contact ghosting and rigid-body interpenetration* arise when 2D latent diffusion predictors lack explicit signed-distance or occupancy constraints: predicted frames depict a gripper cleanly grasping a mug even as its fingers intersect a table edge or pass through a handle. Second, *occlusion identity swapping* occurs when two visually similar objects cross paths behind an articulated robot arm; without persistent slot memory [48] or 3D Gaussian particle tracking [28], video backbones merge object identities upon disocclusion. Third, *action-conditioning*

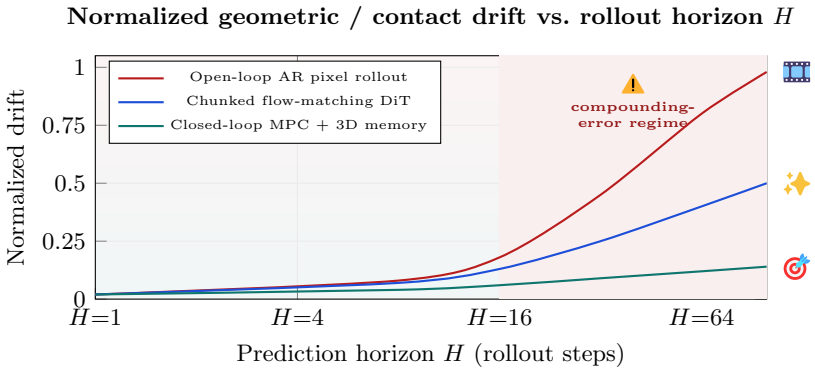


Fig. 12 Schematic illustration of compounding rollout drift vs. prediction horizon H (Section 5.6; curves are qualitative, not measured values): open-loop autoregressive pixel rollouts diverge super-linearly beyond $H > 16$ steps, whereas chunked flow matching and closed-loop MPC replanning anchor geometric drift.

Table 11 Standardized evaluation protocols and primary failure modes across core embodied AI benchmarks (numbers are quoted from the cited sources).

Benchmark suite	Domain & task horizon	Primary evaluation metrics	Reactive VLA limitation	WAM advantage & bottleneck
LIBERO / LIBERO-Plus [66, 67]	Manipulation (long-horizon)	Task success rate (%) under perturbations	Success can fall from ∼95% to below 30% under camera / robot-state perturbations [67]	Video-model policies reach 98.5% average on LIBERO [39]; robustness still open
CALVIN (ABC→D) [68]	Manipulation (5-task chain)	Avg. completed chain length (0–5)	Error accumulation across task transitions	Subgoal rehearsal improves chaining reliability
SimplerEnv [92]	Real-to-sim (WidowX / Google)	Sim-real Pearson correlation & success	Sensitive to visual texture mismatch	Enables rapid policy screening before hardware
nuScenes / Occ3D [53, 71]	Driving (3.0 s horizon)	L_2 error (m), collision (%), Occ mIoU	Open-loop ego-velocity shortcutting	4D occupancy captures free space & collisions
NAVSIM [73]	Driving (closed-loop)	PDMS (no-collision, drivable, comfort)	Reactive steering drift in dense traffic	Counterfactual rollouts score multi-agent risk
WorldScore / WorldModel-Bench [78, 112]	Physics & controllability	Action alignment, 3D consistency, FVD	N/A (policy does not generate world)	Exposes contact hallucination & action ignoring

collapse (or trajectory shortcutting) arises when video backbones are trained predominantly on passive human/driving footage where future motion is highly predictable from momentum alone; during inference, the model ignores sharp corrective action

inputs a_t and hallucinates a smooth nominal trajectory regardless of steering or joint commands [78, 112]. Fourth, *temporal flicker at chunk boundaries* degrades closed-loop precision whenever open-loop H-step action chunks are re-sampled without prefix conditioning [38, 43], inducing high-frequency torque spikes on physical joints.

Hardware case studies clarify how architectural design resolves these trade-offs in practice. On a 7-DoF Franka Panda tabletop platform, V-JEPA 2-AC [30] achieves zero-shot pick-and-place after post-training on less than 62 hours of DROID video and test-time CEM planning in latent space, trading pixel reconstruction for much cheaper planning (∼16 s per action); conversely, DreamZero [38] (14B) and Genie Envisioner [37] leverage video-diffusion backbones with joint or decoded action heads for diverse manipulation where fine visual cues matter, with DreamZero reaching 7 Hz closed-loop control. In quadrupedal locomotion, DayDreamer [29] learns to walk on a real Unitree A1 from scratch in about one hour, and WMP [94] transfers a simulation-trained latent world model with depth perception to the A1 over gaps, stairs and slopes—settings where multi-second diffusion samplers would be too slow for balance control. Finally, in urban driving evaluations (nuScenes [71], NAVSIM [73]), explicit 4D occupancy predictors such as OccWorld [27] and GenAD [41] verify drivable free space directly in metric ego-coordinates, preventing the perspective drift observed in purely 2D video generators.

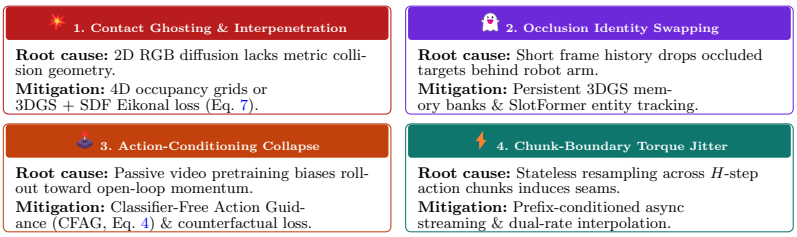


Fig. 13 Real-world WAM failure modes and targeted architectural mitigations (Section 5.7): linking observable hardware errors to their underlying modeling root causes and concrete engineering remedies.

Key takeaways

- Better FVD does not imply better control; report closed-loop success together with physics diagnostics (IVR, E_{dn}, Δ_{act}).
- Saturated benchmarks such as LIBERO hide brittleness that perturbation suites expose.
- Test whether a world model's forecasts depend on the action, not only whether accuracy improves.

6 Open Challenges and Future Directions

To advance WAMs from offline video generators to reliable embodied controllers, five core bottlenecks must be addressed (Fig. 14).

Table 12 Cross-embodiment physical failure modes, observable symptoms, and architectural mitigations.

Failure mode	Observable hardware symptom	Primary root cause	Diagnostic metric	Targeted mitigation
Contact ghosting	Gripper/bumper penetrates rigid obstacle	2D RGB lacks metric collision surface	3D penetration depth (mm); IVR (Eq. 8)	4D occ. / 3DGS + SDF Eikonal loss (Eq. 7)
Occlusion ID swap	Mis-grasped target after arm occlusion	Short context window; no object slots	3D track / slot consistency	Persistent 3DGS memory + SlotFormer [48]
Action collapse	Rollout ignores sharp corrective command	Momentum bias from passive video pretraining	Action sensitivity Δ_{act} (Eq. 8)	CFAG guidance (Eq. 4) + counterfactual loss
Chunk-edge jitter	Torque spike at H-step replanning seam	Stateless chunk resampling across steps	Jerk (m/s^3) & torque RMS	Prefix-conditioned async streaming [38, 43]
OOD friction drift	Object slips despite visually sound grasp	Visual appearance decouples from mass/μ	Slip event rate (%)	Visuo-tactile latent fusion (Section 3.4)

6.1 Five Core Bottlenecks

Bridging offline predictive modeling and high-frequency physical deployment requires resolving foundational failure modes across causality, latency, memory, embodiment alignment, and metrics. We state each as a research question (RQ) with a measurable target so that future work can report progress against it.

1. **RQ1 – Causal hallucination and intuitive physics:** generative WAMs frequently confuse visual correlation with physical causality, hallucinating object deformations or disappearing obstacles under contact-rich interactions. *Target:* near-zero IVR and stable E_{dn} during contact, at FVD comparable to unconstrained 2D models.
2. **RQ2 – Closed-loop inference latency:** iterative diffusion denoising limits control rates (e.g., 7 Hz for the 14B DreamZero [38]), whereas dexterous manipulation and legged balance typically demand much faster feedback. *Target:* reactive-VLA control rates (tens of Hz) on a named edge GPU with no loss in closed-loop success.
3. **RQ3 – Long-horizon spatial memory:** standard context windows forget object states once they leave the camera frustum; integrating persistent 3D neural maps with generative dynamics remains an open frontier. *Target:* correct object identity and pose after occlusions lasting minutes, reported as re-identification accuracy versus occlusion length.
4. **RQ4 – Cross-embodiment action alignment:** harmonizing heterogeneous action spaces (e.g., 7-DoF end-effector deltas, dexterous hands, and vehicle steering) into unified latent action tokens. *Target:* one action interface whose zero-shot transfer to a new embodiment matches per-embodiment training.

5. **RQ5 – Physics-grounded evaluation metrics:** developing automated evaluators combining 3D contact detection, depth consistency, and kinematic feasibility alongside FVD. *Target:* automated metrics whose rank correlation with closed-loop success is reported across at least two benchmarks.

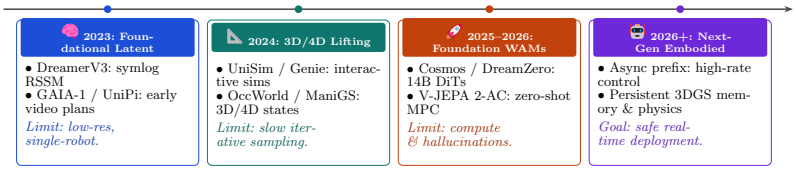


Fig. 14 Chronological evolution of Visual World-Action Models (2023–2026+) and key engineering directions for real-time embodied deployment.

6.2 Concrete Engineering Remedies

Recent systems point toward three actionable architectural remedies for closing the latency and physics gap:

- **Asynchronous and prefix-conditioned execution:** overlapping model inference with action chunk execution using prefix-conditioned generation eliminates chunk-boundary jerky motion without sacrificing precision.
- **Persistent 3D neural memory banks:** pairing short-context video transformers with an incrementally updated 3DGS or voxel feature map allows agents to retrieve occluded object states via camera-frustum queries over hundreds of steps.
- **Differentiable physics contact losses:** supervising generative rollouts with auxiliary 3D bounding-box penetration penalties, optical-flow cycle consistency, and depth-normal alignment reduces non-physical object morphing.

6.3 Epistemic Rollout Uncertainty and Hybrid WAM–VLA Fail-Safe Switching

A sixth critical bottleneck for real-world deployment is epistemic uncertainty calibration during rollout. Unlike deterministic physics engines, generative WAMs can produce visually sharp yet physically catastrophic rollouts when encountering novel object geometries or unseen friction regimes—an effect masked by single-sample trajectory evaluation. Confidence-gated hybrid WAM–VLA execution (Fig. 15) offers a practical engineering remedy: by evaluating rollout dispersion $\sigma^2_{\text{rollout}}(z_t, a_{t:t+H}) = \text{Var}_{k=1\ldots K}\left[f_\phi^{(k)}(z_t, a)\right]$ across lightweight latent ensembles or multi-seed flow samples, an embodied controller can dynamically fall back to a high-frequency reactive VLA safety policy π_θ (or trigger conservative braking) whenever predictive disagreement exceeds a calibrated conformal threshold τ_{safe}:

$$a_t^{\text{exec}} = \begin{cases} a_t^{\text{WAM-MPC}}, & \text{if } \sigma^2_{\text{rollout}}(z_t, a_{t:t+H}) \le \tau_{\text{safe}}, \\ \pi_\theta^{\text{VLA}}(o_{1:t}, l), & \text{otherwise (safe reactive fallback).} \end{cases} \tag{9}$$

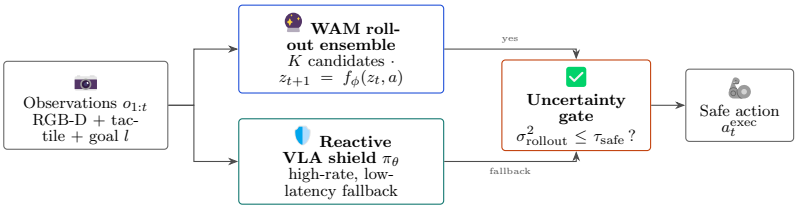


Fig. 15 Confidence-gated hybrid WAM–VLA safety architecture (Section 6.3, Eq. (9)): multi-seed rollout dispersion $\sigma^2_{\text{rollout}}$ dynamically gates execution between deliberative WAM-MPC and a high-frequency reactive VLA shield.

6.4 Edge Deployment and Inference Acceleration

Closing the gap between low-rate video-diffusion WAMs and high-rate dexterous hardware requires co-designing temporal distillation with edge systems engineering. Standard DDPM/DDIM samplers require tens of sequential network evaluations per generated chunk, which saturates embedded GPUs. Consistency Models [113] and Rectified Flow Distillation [114] straighten curved probability ODE trajectories so high-fidelity rollouts converge in 1–4 sampling steps:

$$\mathcal{L}_{\text{RF-distill}} = \mathbb{E}_{t,z_0,z_1}\left[\left\|v_\phi(z_t,t,a_t) - (z_1 - z_0)\right\|_2^2\right], \qquad z_t = (1-t)\,z_0 + t\,z_1. \tag{10}$$

Combined with temporal KV-cache reuse across overlapping prefix frames, FP8 tensor-core quantization, and asynchronous dual-rate execution—where a slower world-model planner runs on a host GPU while a lightweight action-chunk interpolator executes at the robot's control rate (Fig. 16)—embodied systems can approach smooth real-time control without sacrificing predictive foresight.

- **1–4 step flow and consistency distillation:** replaces tens of denoising steps with one to a few [113, 114], at some cost in sample fidelity.
- **Dual-rate async host-edge stack:** decouples low-rate deliberative world rollouts from high-rate local action-chunk streaming and CBF safety filtering.
- **Standardized hardware and latency reporting:** wall-clock Hz, VRAM footprint, and physical collision rates should be reported alongside FVD and success rate.

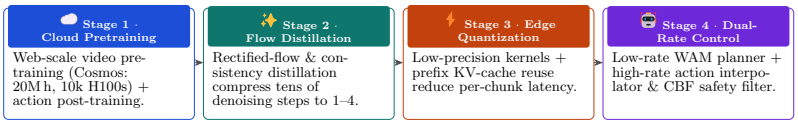


Fig. 16 End-to-end edge deployment and acceleration stack (Section 6.4): bridging large-scale foundation pretraining with few-step flow distillation, low-precision inference, and dual-rate robot control.

6.5 Safety, Reproducibility and Physical Verification

Transitioning WAMs from academic benchmarks to human-shared homes, factories, and public roadways demands formal physical verification and reproducible reporting.

Because neural world models offer probabilistic rather than formal guarantees, safety-critical deployments pair WAM trajectory proposals with Control Barrier Functions (CBFs) [115] and hardware torque/velocity limits that veto commands violating kinematic envelopes regardless of visual confidence. For scientific reproducibility across the community, future surveys and benchmark releases should report closed-loop wall-clock latency (ms/step on standardized hardware), physical collision rates alongside perceptual FVD, and open-weight action-conditioned checkpoints.

Key takeaways

- Latency, long-horizon memory, and calibrated uncertainty are the main barriers to deploying WAMs.
- Few-step distillation, dual-rate execution, and uncertainty-gated fallback to a reactive policy are the most practical remedies today.

7 Conclusion

Visual World-Action Models represent a paradigm shift from reactive perception-to-action mappings toward predictive, physically grounded embodied intelligence. By unifying visual representation learning, 3D spatial geometry, and action-conditioned dynamics, WAMs provide a scalable foundation for general-purpose robotics and autonomous driving. To translate these findings into concrete engineering practice, Fig. 17 provides a practitioner decision guide mapping target embodiment constraints—control rate, occlusion severity, contact compliance, and onboard compute budget—to the recommended spatial representation, dynamics backbone, and inference acceleration stack. For unified reference across all equations, Table A1 (Appendix A) consolidates the mathematical symbols used throughout this survey. The leaderboards, paper list, figures, and the reference implementation of IVR, E_{dn}, and Δ_{act} are maintained at https://github.com/binayksingh/awesome-visual-world-action-models; we invite the community to report against RQ1–RQ5 (Section 6.1) and the WAM-RC checklist (Table 13).

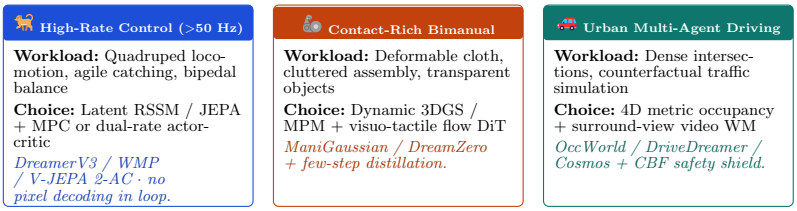


Fig. 17 Practitioner architecture selection guide (Section 7): mapping deployment requirements across control frequency, 3D contact fidelity, and compute constraints to recommended WAM paradigms and acceleration stacks.

7.1 Multi-Year Research Roadmap (2026–2029)

Synthesizing the architectural trends, empirical scaling observations, and physical failure modes reviewed across Sections 2–6, we outline a structured three-horizon roadmap for Visual World-Action Models. Near-term efforts (2026–2027) should prioritize standardized closed-loop evaluation protocols that pair perceptual metrics with 3D collision checking, action sensitivity auditing (Eq. (8)), and wall-clock latency reporting on fixed edge GPUs, eliminating the false confidence induced by FVD-only leaderboards. Equally urgent is the release of open-weight, action-conditioned 3D/4D foundation checkpoints trained on harmonized human-egocentric and multi-embodiment teleoperation mixtures.

Medium-term breakthroughs (2027–2028) hinge on unifying persistent 3D Gaussian and continuum-mechanics memory with sub-20 ms distilled flow transformers, enabling robots to maintain metric world states across multi-minute, multi-room mobile manipulation while streaming tactile and force feedback into shared latent tokens. Over the longer horizon (2028+), closing the loop between internet-scale passive video pretraining, physics-verified synthetic rollouts, and formal Control Barrier Function (CBF) safety certificates will transform WAMs from experimental trajectory predictors into certifiable general-purpose physical collaborators.

- **Phase I (2026–2027): physics-audited benchmarks and fast flow.** Standardize 3D penetration, PDMS, and action-sensitivity reporting alongside 1–4 step consistency distillation for 50+ Hz edge execution.
- **Phase II (2027–2028): persistent 4D memory and visuo-tactile grounding.** Couple dynamic 3DGS/MPM world state banks with high-rate optical tactile encoders and cross-embodiment spectral action vocabularies.
- **Phase III (2028+): certifiable physical foundation models.** Integrate conformal uncertainty quantification and differentiable CBF safety shields into general-purpose humanoid and autonomous vehicle planners.

7.2 Community Reproducibility Standard

As generative world-action modeling matures across foundation models such as Genie 2 [54], WorldDreamer [116], and GR-2 [117], inconsistent evaluation protocols remain a major barrier to fair cross-paper comparison. Many recent reports evaluate open-loop visual fidelity on selected clips without disclosing closed-loop replanning rates, failed grasp collisions, or physical commonsense scores on dedicated suites such as PhysBench [118]. Adopting a standardized reproducibility checklist will accelerate scientific progress across both computer vision and robot learning.

We therefore propose the *WAM Reporting Checklist* (WAM-RC, Table 13) for future submissions benchmarking Visual World-Action Models.

Acknowledgements. The authors used Claude (Anthropic), a large language model, to assist with literature verification, drafting and editing of text, preparation of figures and tables, and implementation of the accompanying code. The authors

Table 13 The WAM Reporting Checklist (WAM-RC): items that papers benchmarking Visual World-Action Models should report.

#	Item	What to report
1	Closed-loop latency	Wall-clock ms per action chunk and effective control rate (Hz) on a named GPU, including decoding, denoising, and MPC; not FLOPs alone.
2	Perceptual and physical metrics together	FVD/LPIPS alongside physical contact fidelity (IVR or 3D penetration depth, collision rate, E_{dn}) and action dependence (Δ_{act} or an executed/wrong/no-action intervention test).
3	Data provenance	Web-video, simulation, and teleoperation hours reported separately; pretraining video hours separated from embodiment-specific action-labeled hours (e.g., RT-1 [119], OXE [63], DROID [64]).
4	Failure breakdown	Success split by occlusion, lighting, and long-horizon chaining, plus results on a perturbation suite.
5	Benchmark protocol	Benchmark version and split (e.g., NAVSIM v1 navtest), metric definition, number of seeds or trials, and whether any best-of-N or test-time selection is used.
6	Ethical and safety compliance	Torque/velocity limits, emergency stops, and privacy safeguards for egocentric footage.

reviewed, verified, and edited all content and take full responsibility for the article. Figure icons are from Microsoft Fluent Emoji (https://github.com/microsoft/fluentui-emoji), released under the MIT License.

Declarations

- **Funding:** The authors received no funding for this work.
- **Conflict of interest/Competing interests:** The authors declare no competing interests.
- **Ethics approval and consent to participate:** Not applicable.
- **Consent for publication:** Not applicable.
- **Data availability:** Not applicable; this article reviews published work and does not generate new data. All benchmark results reported in the tables are taken from the cited publications.
- **Materials availability:** Not applicable.
- **Code availability:** A reference implementation of the IVR, E_{dn}, and Δ_{act} metrics, the curated paper list, the leaderboards, and the figures are available at https://github.com/binayksingh/awesome-visual-world-action-models.
- **Author contribution:** B.K.S. conceived the survey, performed the literature search and analysis, prepared the figures, tables, and code, and wrote the manuscript. N.D.V.L. supervised the work and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Appendix A Unified Mathematical Notation

Table A1: Unified mathematical notation and symbol glossary.

Symbol	Domain / space	Definition & role in survey	Primary ref.
$o_t,\ o_{1:t}$	$\mathcal{O}$ (RGB-D / tactile)	Multi-view visual and sensory observation at timestep t	Sec. 1.1, Eq. (1)
$a_t,\ a_{t:t+H-1}$	$\mathcal{A}$ ($\mathbb{R}^d$ / SE(3))	Motor command or H-step action chunk executed by agent	Sec. 1.1, Eq. (2)
l	$\mathcal{L}$ (text / goal)	Natural-language instruction or target goal specification	Sec. 1.1, Eq. (1)
$z_t,\ \hat{z}_{t+1}$	$\mathcal{Z}$ (latent / 3DGS)	Compact latent, voxel, or Gaussian world state at step t	Sec. 2.1, Eq. (3)
$\pi_\theta,\ p_\phi$	Policy / dynamics	Reactive VLA policy π_θ vs. generative WAM transition p_ϕ	Sec. 1.1, Eqs. (1)–(2)
H	$\mathbb{Z}_+$ (steps)	Temporal prediction / planning horizon for closed-loop rollout	Sec. 1.1, Fig. 12
$[R_t \mid t_t],\ F_{uv}$	SE(3) / $\mathbb{R}^{3\times3}$	Camera extrinsics & fundamental matrix linking views u, v	Secs. 3.3–3.6, Eq. (7)
G_t	$\{(\mu_i, \Sigma_i, \alpha_i, c_i)\}$	Dynamic 3D Gaussian Splatting scene representation at step t	Sec. 3.4, Eq. (6)
s_a	$\mathbb{R}_{\geq1}$	Classifier-Free Action Guidance (CFAG) weight	Sec. 2.4, Eq. (4)
IVR, Δ_{act}	$\mathbb{R}$ (metrics)	Interpenetration Volume Ratio & Action-Counterfactual Sensitivity	Sec. 5.4, Eq. (8)
$\Phi_\phi(x, z_t)$	$\mathbb{R}$	Neural Signed Distance Field (SDF)	Sec. 3.5, Eq. (7)
$\sigma^2_{\text{rollout}},\ \tau_{\text{safe}}$	$\mathbb{R}_+$	Multi-seed rollout dispersion & conformal safety threshold	Sec. 6.3, Eq. (9)

References

[1] LeCun, Y.: A Path Towards Autonomous Machine Intelligence, Version 0.9.2. OpenReview position paper, https://openreview.net/forum?id=BZ5a1r-kVsf (2022)

[2] Ha, D., Schmidhuber, J.: Recurrent world models facilitate policy evolution. In: Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 31 (2018)

[3] Brohan, A., Brown, N., Carbajal, J., *et al.*: RT-2: Vision-language-action models transfer web knowledge to robotic control. In: Proc. 7th Conf. Robot Learn. (CoRL), PMLR, vol. 229 (2023)

[4] Kim, M.J., Pertsch, K., Karamcheti, S., *et al.*: OpenVLA: An open-source vision-language-action model. In: Proc. 8th Conf. Robot Learn. (CoRL), PMLR, vol. 270 (2024)

[5] Yang, S., Du, Y., Ghasemipour, K., *et al.*: Learning interactive real-world simulators. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2024)

[6] Wang, S., Shi, J., Fu, Z., et al.: World Action Models: The Next Frontier in Embodied AI. arXiv:2605.12090 (2026)

[7] Hu, A., Russell, L., Yeo, H., et al.: GAIA-1: A Generative World Model for Autonomous Driving. arXiv:2309.17080 (2023)

[8] Bruce, J., Dennis, M., Edwards, A., *et al.*: Genie: Generative interactive environments. In: Proc. 41st Int. Conf. Mach. Learn. (ICML), PMLR, vol. 235 (2024)

[9] Bardes, A., Garrido, Q., Ponce, J., et al.: Revisiting feature prediction for learning visual representations from video. Trans. Mach. Learn. Res. (2024)

[10] Li, X., He, X., Zhang, L., et al.: A Comprehensive Survey on World Models for Embodied AI. arXiv:2510.16732 (2025)

[11] Kong, L., Yang, Y., Mei, J., et al.: 3D and 4D World Modeling: A Survey. arXiv:2509.07996 (2025)

[12] Shang, Y., Tang, Y., Zhang, X., et al.: A Survey of Embodied World Models. Preprints.org (2026). https://doi.org/10.20944/preprints202604.0928.v1

[13] Shen, Q., Zhang, S., Liao, Y., et al.: World Action Models: A Survey. arXiv:2606.20781 (2026)

[14] Lu, Z., Zhai, H., Wang, G., et al.: World-Action Models for Robot Learning and Control: A Survey. arXiv:2609.16074 (2026)
[15] Singh, B.K., Lobo, N.D.V.: From Pixels to Actions: A Comprehensive Review of Spatiotemporal Understanding from Visual Perception to Vision-Language-Action Models. Preprints 2026, 2026112452 (2026). https://doi.org/10.20944/preprints202611.2452.v1
[16] Singh, B.K., Lobo, N.D.V.: Urban state representation for autonomous mobility: A survey of multimodal perception, architectures, datasets, and benchmarks. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW), UrbanState Workshop, to appear (2027). https://doi.org/10.48550/arXiv.2611.28129
[17] Singh, B.K., Lobo, N.D.V.: Does the action matter? Probing action-conditioned semantic forecasting as a world model for autonomous driving. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW), WAM-AD Workshop, to appear (2027). https://doi.org/10.48550/arXiv.2611.24840
[18] Guan, Y., Liao, H., Li, Z., et al.: World Models for Autonomous Driving: An Initial Survey. arXiv:2403.02622 (2024)
[19] Feng, T., Wang, W., Yang, Y.: A Survey of World Models for Autonomous Driving. arXiv:2501.11260 (2025)
[20] Tu, S., Zhou, X., Liang, D., et al.: The Role of World Models in Shaping Autonomous Driving: A Comprehensive Survey. arXiv:2502.10498 (2025)
[21] Zhang, X., Zeng, X., Zhang, W.: From World Models to World Action Models: A Concise Tutorial for Robotics. arXiv:2607.00836 (2026)
[22] Hafner, D., Lillicrap, T., Fischer, I., *et al.*: Learning latent dynamics for planning from pixels. In: Proc. 36th Int. Conf. Mach. Learn. (ICML), PMLR, vol. 97, pp. 2555–2565 (2019)
[23] Hafner, D., Lillicrap, T., Norouzi, M., Ba, J.: Mastering Atari with discrete world models. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2021)
[24] Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering diverse control tasks through world models. Nature **640**(8059), 647–653 (2025) https://doi.org/10.1038/s41586-025-08744-2
[25] Babaeizadeh, M., Saffar, M.T., Nair, S., et al.: FitVid: Overfitting in Pixel-Level Video Prediction. arXiv:2106.13195 (2021)
[26] NVIDIA, Agarwal, N., Ali, A., et al.: Cosmos World Foundation Model Platform for Physical AI. arXiv:2501.03575 (2025)
[27] Zheng, W., Chen, W., Huang, Y., *et al.*: OccWorld: Learning a 3D occupancy world model for autonomous driving. In: Proc. Eur. Conf. Comput. Vis. (ECCV), LNCS, pp. 55–72. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72624-8_4
[28] Lu, G., Zhang, S., Wang, Z., *et al.*: ManiGaussian: Dynamic Gaussian splatting for multi-task robotic manipulation. In: Proc. Eur. Conf. Comput. Vis. (ECCV), LNCS, pp. 349–366. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72761-0_20
[29] Wu, P., Escontrela, A., Hafner, D., *et al.*: DayDreamer: World models for physical robot learning. In: Proc. 6th Conf. Robot Learn. (CoRL), PMLR, vol. 205 (2022)
[30] Assran, M., Bardes, A., Fan, D., et al.: V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. arXiv:2506.09985 (2025)
[31] Ye, S., Jang, J., Jeon, B., *et al.*: Latent action pretraining from videos. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2025)
[32] Du, Y., Yang, M., Dai, B., *et al.*: Learning universal policies via text-guided video generation. In: Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 36 (2023)
[33] Wang, X., Zhu, Z., Huang, G., *et al.*: DriveDreamer: Towards real-world-drive world models for autonomous driving. In: Proc. Eur. Conf. Comput. Vis. (ECCV), LNCS, pp. 55–72. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-73195-2_4
[34] Gao, S., Yang, J., Chen, L., *et al.*: Vista: A generalizable driving world model with high fidelity and versatile controllability. In: Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 37 (2024)
[35] Alonso, E., Jelley, A., Micheli, V., Kanervisto, A., Storkey, A., Pearce, T., Fleuret, F.: Diffusion for world modeling: Visual details matter in Atari. In: Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 37 (2024)
[36] Cen, J., Yu, C., Yuan, H., et al.: WorldVLA: Towards Autoregressive Action World Model. arXiv:2506.21539 (2025)
[37] Liao, Y., Zhou, P., Huang, S., *et al.*: Genie envisioner: A unified world foundation platform for robotic manipulation. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2026)
[38] Ye, S., Ge, Y., Zheng, K., et al.: World Action Models are Zero-shot Policies. arXiv:2602.15922 (2026)
[39] Kim, M.J., Gao, Y., Lin, T.-Y., *et al.*: Cosmos policy: Fine-tuning video models for visuomotor control and planning. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2026)
[40] Zhen, H., Qiu, X., Chen, P., *et al.*: 3D-VLA: A 3D vision-language-action generative world model. In: Proc. 41st Int. Conf. Mach. Learn. (ICML), PMLR, vol. 235 (2024)
[41] Zheng, W., Song, R., Guo, X., *et al.*: GenAD: Generative end-to-end autonomous driving. In: Proc. Eur. Conf. Comput. Vis. (ECCV), LNCS, pp. 87–104. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-73650-6_6

[42] Xie, T., Zong, Z., Qiu, Y., *et al.*: PhysGaussian: Physics-integrated 3D Gaussians for generative dynamics. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4389–4398 (2024). https://doi.org/10.1109/CVPR52733.2024.00420

[43] Huang, S., Chen, L., Zhou, P., et al.: EnerVerse: Envisioning Embodied Future Space for Robotics Manipulation. arXiv:2501.01895 (2025)

[44] Yu, L., Lezama, J., Gundavarapu, N.B., *et al.*: Language model beats diffusion – tokenizer is key to visual generation. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2024)

[45] Mentzer, F., Minnen, D., Agustsson, E., Tschannen, M.: Finite scalar quantization: VQ-VAE made simple. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2024)

[46] Pertsch, K., Stachowicz, K., Ichter, B., *et al.*: FAST: Efficient action tokenization for vision-language-action models. In: Robot.: Sci. Syst. XXI (2025)

[47] Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D Gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 1–14 (2023) https://doi.org/10.1145/3592433

[48] Wu, Z., Dvornik, N., Greff, K., *et al.*: SlotFormer: Unsupervised visual dynamics simulation with object-centric models. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2023)

[49] Grauman, K., Westbury, A., Byrne, E., *et al.*: Ego4D: Around the world in 3,000 hours of egocentric video. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 18973–18990 (2022). https://doi.org/10.1109/CVPR52688.2022.01842

[50] Bharadhwaj, H., Mottaghi, R., Gupta, A., Tulsiani, S.: Track2Act: Predicting point tracks from internet videos enables generalizable robot manipulation. In: Proc. Eur. Conf. Comput. Vis. (ECCV), LNCS, pp. 306–324. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-73116-7_18

[51] Karaev, N., Makarov, I., Wang, J., *et al.*: CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 6013–6022 (2025)

[52] Bharadhwaj, H., Dwibedi, D., Gupta, A., et al.: Gen2Act: Human Video Generation in Novel Scenarios enables Generalizable Robot Manipulation. arXiv:2409.16283 (2024)

[53] Tian, X., Jiang, T., Yun, L., *et al.*: Occ3D: A large-scale 3D occupancy prediction benchmark for autonomous driving. In: Adv. Neural Inf. Process. Syst. (NeurIPS), Datasets Benchmarks Track, vol. 36 (2023)

[54] Parker-Holder, J., Ball, P., Bruce, J., et al.: Genie 2: A Large-Scale Foundation World Model. Google DeepMind blog, https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/ (2024)

[55] Unitree Robotics: UnifoLM-WMA-0: A World-Model-Action (WMA) Framework under UnifoLM Family. Open-source release, https://github.com/unitreerobotics/unifolm-world-model-action (2025)

[56] Wu, J., Yin, S., Feng, N., *et al.*: iVideoGPT: Interactive VideoGPTs are scalable world models. In: Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 37 (2024)

[57] Wang, Y., He, J., Fan, L., *et al.*: Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) (2024)

[58] Li, Y., Fan, L., He, J., *et al.*: Enhancing end-to-end autonomous driving with latent world model. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2025)

[59] Li, Y., Wang, Y., Liu, Y., *et al.*: End-to-end driving with online trajectory evaluation via BEV world model. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV) (2025)

[60] Zheng, Y., Yang, P., Xing, Z., *et al.*: World4Drive: End-to-end autonomous driving via intention-aware physical latent world model. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 28632–28642 (2025)

[61] Zhang, K., Tang, Z., Hu, X., *et al.*: Epona: Autoregressive diffusion world model for autonomous driving. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 27220–27230 (2025)

[62] Li, Y., Shang, S., Liu, W., *et al.*: DriveVLA-W0: World models amplify data scaling law in autonomous driving. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2026)

[63] Open X-Embodiment Collaboration, O'Neill, A., Rehman, A., *et al.*: Open X-embodiment: Robotic learning datasets and RT-X models. In: Proc. IEEE Int. Conf. Robot. Autom. (ICRA), pp. 6892–6903 (2024). https://doi.org/10.1109/ICRA57147.2024.10611477

[64] Khazatsky, A., Pertsch, K., Nair, S., *et al.*: DROID: A large-scale in-the-wild robot manipulation dataset. In: Robot.: Sci. Syst. XX (2024). https://doi.org/10.15607/RSS.2024.XX.120

[65] Walke, H., Black, K., Lee, A., *et al.*: BridgeData V2: A dataset for robot learning at scale. In: Proc. 7th Conf. Robot Learn. (CoRL), PMLR, vol. 229 (2023)

[66] Liu, B., Zhu, Y., Gao, C., *et al.*: LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In: Adv. Neural Inf. Process. Syst. (NeurIPS), Datasets Benchmarks Track, vol. 36 (2023)

[67] Fei, S., Wang, S., Shi, J., et al.: LIBERO-Plus: In-depth Robustness Analysis of Vision-Language-Action Models. arXiv:2510.13626 (2025)

[68] Mees, O., Hermann, L., Rosete-Beas, E., Burgard, W.: CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robot. Autom. Lett. **7**(3), 7327–7334 (2022) https://doi.org/10.1109/LRA.2022.3180108

[69] Chen, T., Chen, Z., Chen, B., *et al.*: RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. In: Proc. 43rd Int. Conf. Mach. Learn. (ICML), PMLR, vol. 306 (2026)
[70] Nasiriany, S., Maddukuri, A., Zhang, L., *et al.*: RoboCasa: Large-scale simulation of household tasks for generalist robots. In: Robot.: Sci. Syst. XX (2024). https://doi.org/10.15607/RSS.2024.XX.050
[71] Caesar, H., Bankiti, V., Lang, A.H., *et al.*: nuScenes: A multimodal dataset for autonomous driving. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 11618–11628 (2020). https://doi.org/10.1109/CVPR42600.2020.01164
[72] Sun, P., Kretzschmar, H., Dotiwalla, X., *et al.*: Scalability in perception for autonomous driving: Waymo open dataset. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2443–2451 (2020). https://doi.org/10.1109/CVPR42600.2020.00252
[73] Dauner, D., Hallgarten, M., Li, T., *et al.*: NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In: Adv. Neural Inf. Process. Syst. (NeurIPS), Datasets Benchmarks Track, vol. 37 (2024)
[74] Zhou, Y., Shao, H., Wang, L., *et al.*: DrivingGen: A comprehensive benchmark for generative video world models in autonomous driving. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2026)
[75] Puig, X., Undersander, E., Szot, A., *et al.*: Habitat 3.0: A co-habitat for humans, avatars and robots. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2024)
[76] Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., *et al.*: Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI. In: Adv. Neural Inf. Process. Syst. (NeurIPS), Datasets Benchmarks Track (2021)
[77] Chang, A., Dai, A., Funkhouser, T., *et al.*: Matterport3D: Learning from RGB-D data in indoor environments. In: Proc. Int. Conf. 3D Vis. (3DV), pp. 667–676 (2017). https://doi.org/10.1109/3DV.2017.00081
[78] Li, D., Fang, Y., Chen, Y., et al.: WorldModelBench: Judging Video Generation Models As World Models. arXiv:2502.20694 (2025)
[79] Shang, Y., Li, Z., Ma, Y., et al.: WorldArena: A Unified Benchmark for Evaluating Perception and Functional Utility of Embodied World Models. arXiv:2602.08971 (2026)
[80] Motamed, S., Culp, L., Swersky, K., *et al.*: Do generative video models understand physical principles? In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), pp. 948–958 (2026). https://doi.org/10.1109/WACV61042.2026.00099
[81] Bordes, F., Garrido, Q., Kao, J.T., et al.: IntPhys 2: Benchmarking Intuitive Physics Understanding In Complex Synthetic Environments. arXiv:2506.09849 (2025)
[82] Zheng, D., Huang, Z., Liu, H., et al.: VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness. arXiv:2503.21755 (2025)
[83] Yuan, T., Dong, Z., Liu, Y., Zhao, H.: Fast-WAM: Do World Action Models Need Test-time Future Imagination? arXiv:2603.16666 (2026)
[84] Zhou, R., Ni, Z., Fan, Z., et al.: What Makes World Action Models Generalize? An Empirical Study of Test-Time Future Modeling. arXiv:2609.34981 (2026)
[85] Jiang, H., Liu, L., Wang, X., et al.: Rethinking Representations for World-Action Modeling. arXiv:2609.38163 (2026)
[86] Yu, Q., Ding, W., Wang, G.X., et al.: Box-WAM: Box Trajectories as Compact State Representations for World Action Models. arXiv:2610.17526 (2026)
[87] Guo, J., Li, Q., Li, P., et al.: Unified 4D World Action Modeling from Video Priors with Asynchronous Denoising. arXiv:2604.26694 (2026)
[88] Liu, Z., Niu, D., Zhang, B., et al.: Agentic World Action Models. arXiv:2610.30136 (2026)
[89] Chen, Y., Ge, Y., Tang, W., et al.: Moto: Latent Motion Token as the Bridging Language for Learning Robot Manipulation from Videos. arXiv:2412.04445 (2024)
[90] Tao, S., Xiang, F., Shukla, A., *et al.*: ManiSkill3: GPU parallelized robotics simulation and rendering for generalizable embodied AI. In: Robot.: Sci. Syst. XXI (2025)
[91] Ren, X., Lu, Y., Cao, T., et al.: Cosmos-Drive-Dreams: Scalable Synthetic Driving Data Generation with World Foundation Models. arXiv:2506.09042 (2025)
[92] Li, X., Hsu, K., Gu, J., *et al.*: Evaluating real-world robot manipulation policies in simulation. In: Proc. 8th Conf. Robot Learn. (CoRL), PMLR, vol. 270, pp. 3705–3728 (2025)
[93] AgiBot World Contributors, Bu, Q., Cai, J., et al.: AgiBot World Colosseo: A Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems. arXiv:2503.06669 (2025)
[94] Lai, H., Cao, J., Xu, J., *et al.*: World model-based perception for visual legged locomotion. In: Proc. IEEE Int. Conf. Robot. Autom. (ICRA), pp. 11531–11537 (2025). https://doi.org/10.1109/ICRA55743.2025.11128762
[95] Gu, S., Yin, W., Jin, B., et al.: DOME: Taming Diffusion Model into High-Fidelity Controllable Occupancy World Model. arXiv:2410.10429 (2024)
[96] Zhao, Z., Zhou, X., Xu, T., et al.: SimWAM: A Simple World Action Model for End-to-End Autonomous Driving. arXiv:2608.07468 (2026)
[97] Fan, Z., Zhang, T., Wu, W., et al.: Drive-HWM: Hierarchical World Models for Dynamic-Latent Guided Autonomous Driving. arXiv:2609.03572 (2026)

[98] Jia, M., Guo, J., Shu, Z., et al.: WALT: Learning World-Model-Aligned Latent Trajectories for Autonomous Driving. arXiv:2609.30436 (2026)
[99] Chitta, K., Prakash, A., Jaeger, B., *et al.*: TransFuser: Imitation with transformer-based sensor fusion for autonomous driving. IEEE Trans. Pattern Anal. Mach. Intell. **45**(11), 12878–12895 (2023) https://doi.org/10.1109/TPAMI.2022.3200245
[100] Li, Z., Li, K., Wang, S., et al.: Hydra-MDP: End-to-End Multimodal Planning with Multi-Target Hydra-Distillation. arXiv:2406.06978 (2024)
[101] Jiang, B., Chen, S., Xu, Q., *et al.*: VAD: Vectorized scene representation for efficient autonomous driving. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 8340–8350 (2023). https://doi.org/10.1109/ICCV51070.2023.00766
[102] Black, K., Brown, N., Driess, D., *et al.*: π_0: A vision-language-action flow model for general robot control. In: Robot.: Sci. Syst. XXI (2025). https://doi.org/10.15607/RSS.2025.XXI.010
[103] Bu, Q., Yang, Y., Cai, J., *et al.*: UniVLA: Learning to act anywhere with task-centric latent actions. In: Robot.: Sci. Syst. (RSS) (2025)
[104] Kim, M.J., Finn, C., Liang, P.: Fine-tuning vision-language-action models: Optimizing speed and success. In: Robot.: Sci. Syst. (RSS) (2025)
[105] Black, K., Nakamoto, M., Atreya, P., *et al.*: Zero-shot robotic manipulation with pretrained image-editing diffusion models. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2024)
[106] Wu, H., Jing, Y., Cheang, C., *et al.*: Unleashing large-scale video generative pre-training for visual robot manipulation. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2024)
[107] Tian, Y., Yang, S., Zeng, J., *et al.*: Predictive inverse dynamics models are scalable learners for robotic manipulation. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2025)
[108] Hu, Y., Guo, Y., Wang, P., *et al.*: Video prediction policy: A generalist robot policy with predictive visual representations. In: Proc. 42nd Int. Conf. Mach. Learn. (ICML). PMLR, vol. 267, pp. 24328–24346 (2025)
[109] Zhang, W., Liu, H., Qi, Z., *et al.*: DreamVLA: A vision-language-action model dreamed with comprehensive world knowledge. In: Adv. Neural Inf. Process. Syst. (NeurIPS) (2025)
[110] Octo Model Team, Ghosh, D., Walke, H., *et al.*: Octo: An open-source generalist robot policy. In: Robot.: Sci. Syst. XX (2024). https://doi.org/10.15607/RSS.2024.XX.090
[111] Atreya, P., Pertsch, K., Lee, T., *et al.*: RoboArena: Distributed real-world evaluation of generalist robot policies. In: Proc. 9th Conf. Robot Learn. (CoRL), PMLR, vol. 305, pp. 336–364 (2025)
[112] Duan, H., Yu, H.-X., Chen, S., *et al.*: WorldScore: A unified evaluation benchmark for world generation. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 27713–27724 (2025)
[113] Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: Proc. 40th Int. Conf. Mach. Learn. (ICML), PMLR, vol. 202 (2023)
[114] Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2023)
[115] Ames, A.D., Coogan, S., Egerstedt, M., *et al.*: Control barrier functions: Theory and applications. In: Proc. 18th Eur. Control Conf. (ECC), pp. 3420–3431 (2019). https://doi.org/10.23919/ECC.2019.8796030
[116] Wang, X., Zhu, Z., Huang, G., et al.: WorldDreamer: Towards General World Models for Video Generation via Predicting Masked Tokens. arXiv:2401.09985 (2024)
[117] Cheang, C.-L., Chen, G., Jing, Y., et al.: GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation. arXiv:2410.06158 (2024)
[118] Chow, W., Mao, J., Li, B., *et al.*: PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2025)
[119] Brohan, A., Brown, N., Carbajal, J., *et al.*: RT-1: Robotics transformer for real-world control at scale. In: Robot.: Sci. Syst. XIX (2023). https://doi.org/10.15607/RSS.2023.XIX.025