World Models and World Action Models (WAM): From Foundation Simulators to Embodied Action

Xin Jin (金鑫)*

sdjinxin@gmail.com

Abstract

World models—internal predictive representations that enable agents to simulate future states, anticipate consequences, and plan actions—have emerged as a foundational paradigm in embodied artificial intelligence. Originating from model-based reinforcement learning, this field has undergone a radical transformation with the advent of large-scale generative models, blurring the historical boundary between passive video prediction and interactive physical simulation. Concurrently, Vision-Language-Action (VLA) models have established a powerful framework for grounding high-level linguistic intent in low-level motor control. The natural convergence of these two threads—predictive world simulation and action-grounded multimodal reasoning—has given rise to **Embodied World Action Models (WAMs)**, representing a new frontier in which agents learn to act by imagining their futures. However, the explosive growth of methods across robotics, autonomous driving, and interactive simulation has produced a fragmented landscape that lacks systematic unification.

This survey presents a **comprehensive and structured review** of the modern world model ecosystem, encompassing **200+ key papers** organized into a unified taxonomy. We systematically cover six major pillars: **(i) Foundation World Models**, including general-purpose interactive simulators (Genie, Cosmos, Sora) and game-specific environments (Oasis, Matrix-Game); **(ii) Vision-Language-Action Models**, spanning foundational architectures (RT-2, π_0, OpenVLA), driving-specific VLAs, and embodied manipulation policies; **(iii) Embodied World Action Models**, unifying video generation and action prediction through zero-shot policies, controllable simulation platforms, and world model-based reinforcement learning; **(iv) Autonomous Driving World Models**, addressing video generation, closed-loop simulation, planning policies, and geometric occupancy/BEV representations; **(v) Efficiency and Evaluation**, covering computational acceleration techniques and benchmarking protocols for physical plausibility; and **(vi) Datasets and Ecosystems**, including large-scale robot learning corpora and industry technical reports that underpin the entire field.

Through this organization, we illuminate the evolutionary trajectory from passive pixel predictors to active, reasoning, and action-grounded simulators. We identify critical open challenges—including physical consistency, cross-embodiment generalization, safety verification, and the sim-to-real evaluation gap—and outline future directions toward cognitive world models, autonomous data collection, and standardized open ecosystems. This survey aims to serve as a definitive reference for researchers and practitioners advancing the next generation of embodied intelligence.

1. Introduction

The concept of the **World Model**—an internal representation that enables an agent to predict future states, simulate consequences of actions, and plan within an imagined environment—has long been central to artificial intelligence. Originating from the seminal work of Ha and Schmidhuber [228], who demonstrated that recurrent world models could facilitate policy evolution through latent imagination, this paradigm was subsequently refined by model-based reinforcement learning (RL) frameworks such as Dreamer [229] and DreamerV3 [230], as well as temporal difference learning with world models (TD-MPC2) [231]. These foundational approaches established a core principle: by learning to predict the dynamics of the environment, agents can decouple planning from real-world interaction, thereby improving sample efficiency and generalization in control tasks.

However, the landscape of world modeling has undergone a radical transformation with the advent of large-scale generative models. The boundary between "world models" and "video generation models" has become increasingly blurred, as modern generative systems demonstrate an emergent capacity to simulate plausible physical dynamics, object permanence, and spatial consistency. OpenAI's Sora [7] explicitly positioned video generation models as "world simulators," capable of generating coherent, long-horizon visual sequences that obey physical intuitions. Concurrently, interactive foundation world models such as Google's Genie [1], Genie 2 [2], and Genie 3 [3], alongside NVIDIA's Cosmos platform [4], have pushed the frontier from passive video prediction to interactive, action-conditioned environment simulation. These

systems not only generate pixels but also respond to agent inputs, creating "generative interactive environments" that can be explored and manipulated. Complementing these industrial efforts, open-source initiatives including LingBot-World [5], GigaWorld-0 [6], and a plethora of memory-augmented [11, 13, 15] and geometry-aware world models [14, 17, 18] have democratized access to high-fidelity world simulation. Beyond general-purpose simulation, specialized models for game environments—such as Oasis [21], MineWorld [22], Matrix-Game [23, 24], and GameGen-X [25]—have demonstrated real-time interactive world modeling in open-ended 3D settings, establishing a vibrant research branch at the intersection of world models and interactive entertainment.

Parallel to the evolution of generative world models, the robotics and embodied AI communities have converged on a closely related paradigm: **Vision-Language-Action (VLA) models**. While traditional world models focus on state prediction, VLA models directly address the grounding of high-level human instructions into low-level motor control through multimodal reasoning. Starting from RT-1 [33] and RT-2 [34], which scaled robotic control via transformer architectures, the field has rapidly expanded to encompass open-source generalist policies such as OpenVLA [36], Octo [41], and π_0 [37], as well as large-scale humanoid robot foundation models like NVIDIA's GR00T N1 [40]. The VLA ecosystem has further bifurcated into domain-specific branches: for autonomous driving, models such as OpenDriveVLA [44], AutoVLA [45], and DriveMoE [48] integrate visual perception, linguistic reasoning, and vehicle control into end-to-end architectures; for general embodied manipulation, frameworks including 3D-VLA [65], Diffusion-VLA [63], HybridVLA [64], and RDT [84, 85] leverage 3D spatial awareness, diffusion-based action decoding, and cross-embodiment generalization to achieve dexterous robotic control. This surge in VLA research reflects a broader recognition that world models must not merely predict what will happen, but also reason about *how to act* within the predicted world.

The natural convergence of these two threads—predictive world models and action-grounded VLA models—has given rise to **Embodied World Action Models (WAMs)**. This emerging class of models treats video generation and action prediction as unified objectives, learning to simulate futures that are not only visually coherent but also action-executable. Pioneering work by NVIDIA's DreamZero [102] demonstrated that world action models can serve as zero-shot policies, while subsequent frameworks such as GigaBrain [105, 106], Unified World Models [107], and VideoVLA [109] have explored the synergy between video diffusion and action generation. On the simulation front, platforms like Genie Envisioner [118], Aether [119], and RoboScape [120] provide controllable, physics-informed embodied environments for training and evaluating policies. For policy optimization, Cosmos-Policy [127, 128], MotuBrain [129], and ThinkAct [130] have introduced world model-based reinforcement fine-tuning and visual latent planning, enabling robots to learn from imagined trajectories with verified rewards. These developments collectively signal a paradigm shift from "world models as passive predictors" to "world models as active training grounds" for embodied agents.

In parallel to embodied robotics, **Autonomous Driving World Models** have matured into a distinct and critical research domain. The ability to simulate diverse, high-fidelity driving scenarios is essential for both data augmentation and closed-loop policy evaluation. Early work such as GAIA-1 [142] and DriveDreamer [138, 139] established the feasibility of driving-specific video world models, while Vista [140], DrivingWorld [141], and MagicDrive [152, 153] advanced the state-of-the-art in controllable, high-resolution, long-horizon driving video generation. Beyond mere generation, recent efforts have focused on closed-loop simulation (X-World [150], Epona [151]), 4D scene representation (DriveDreamer4D [149], GaussianDWM [177]), and end-to-end planning within learned world models (DriveDreamer-Policy [162], GenAD [165], DOE-1 [166]). Furthermore, occupancy and BEV-centric world models—including OccWorld [172], HERMES [174], and SparseWorld [176]—have introduced structured geometric representations that bridge the gap between pixel-level simulation and 3D spatial reasoning, addressing the safety-critical requirements of autonomous driving.

The rapid proliferation of world model architectures has necessitated equal advances in **efficiency, evaluation methodologies, and data ecosystems**. On the efficiency front, action tokenization strategies (FAST [181]), adaptive token caching (VLA-Cache [182]), dynamic layer-skipping (MoLe-VLA [185], DySL-VLA [192]), and speculative decoding mechanisms (KERV [194]) have been proposed to reduce the prohibitive computational costs of deploying large VLA models on real robots. Evaluation benchmarks such as WorldScore [196], WorldModelBench [197], WorldEval [198], and EWMBench [199] have established standardized protocols for assessing world models as physical simulators and policy evaluators. Finally, the growth of the field has been fueled by large-scale datasets including DROID [201], BridgeData V2 [202], LIBERO [203], BEHAVIOR-1K [204], and EgoDex [208], alongside industry technical reports from Ant Group [5, 210], Huawei [152], XPENG [150], and others, which collectively form the data and engineering infrastructure underpinning modern world model research.

Existing surveys and related work. While several recent surveys have examined subsets of this landscape—such as VLA models for embodied AI [219, 226, 227], world models for autonomous driving [220, 221, 222], and generalist world model surveys [223, 224, 225]—none have provided a unified, comprehensive treatment that simultaneously covers (i) foundation world models for general-purpose and interactive simulation, (ii) VLA models spanning robotics and autonomous driving, (iii) embodied world action models that unify generation and control, (iv) driving-specific world models with geometric

representations, and (v) the critical supporting infrastructure of efficiency, benchmarks, and datasets. **This survey fills these gaps as follows:** relative to prior generalist surveys [223–225], we explicitly integrate the VLA and WAM threads; relative to VLA-specific surveys, we add foundation world models and driving world models; relative to driving surveys, we further include game-based foundation models and embodied manipulation.

Contributions and organization. This survey presents a systematic and comprehensive review of the modern World Model and World Action Model ecosystem, encompassing **200+ key papers** organized into a unified taxonomy. The remainder of this paper is structured as follows. In **Section 2**, we review **Foundation World Models**, covering both general-purpose interactive simulators and game/roaming-specific environments. **Section 3** provides an in-depth analysis of **Vision-Language-Action (VLA) Models**, including foundational architectures, driving-specific VLAs, and embodied manipulation VLAs. **Section 4** examines **Embodied World Action Models (WAMs)**, organized into video-generation-based WAMs, controllable and long-horizon simulation platforms, and policy/planning frameworks that leverage world models for robot control. **Section 5** focuses on **Autonomous Driving World Models**, spanning video generation, controllable simulation, planning policies, and occupancy/BEV representations. **Section 6** discusses **Efficiency and Evaluation**, covering computational acceleration techniques and benchmarking protocols. **Section 7** surveys **Datasets and Ecosystems**, including large-scale robot learning datasets and industry technical reports. Finally, we conclude with a discussion of open challenges and future directions. Through this structured organization, we aim to illuminate the interconnected evolution of world models—from passive predictors of pixels to active, reasoning, and action-grounded simulators that serve as the cognitive backbone of next-generation embodied AI and autonomous systems.

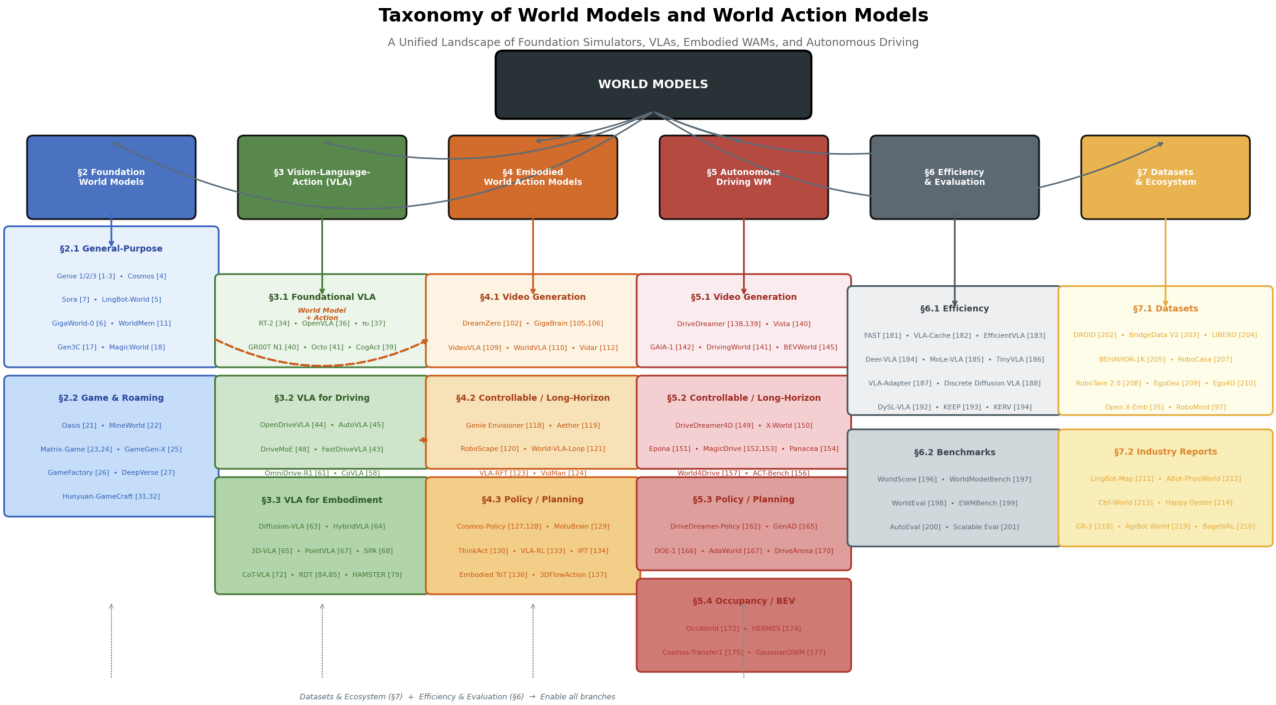

Figure 1. Taxonomy of World Models and World Action Models. The unified landscape encompasses six major categories—Foundation World Models (§2), Vision-Language-Action Models (§3), Embodied World Action Models (§4), Autonomous Driving World Models (§5), Efficiency and Evaluation (§6), and Datasets and Ecosystems (§7). The dashed arrow indicates the convergence of Foundation WMs and VLAs into WAMs, while §6 and §7 provide the supporting infrastructure.

2. Foundation World Models

> **Section 2 roadmap.** This section is divided into two complementary branches: **§2.1 General-Purpose** interactive simulators, which target broad physical world coverage, and **§2.2 Game & Roaming** environments, which specialize in open-ended 3D game worlds. The transition from §2.1 to §2.2 moves from general physical simulation to structured but highly interactive domains with well-defined action spaces. Later, in §4, we will build on these foundation models to incorporate action generation.

Foundation world models represent the frontier of generative simulation, aiming to learn universal environmental dynamics from large-scale video data such that an agent can query the model with actions and receive coherent, long-

horizon future observations. Unlike classical model-based RL, which learns compact latent dynamics for control, modern foundation world models prioritize *scale*, *generality*, and *interactivity*: they are trained on diverse internet-scale video and are expected to simulate arbitrary scenes, objects, and physical interactions in response to open-ended action inputs. this section organizes foundation world models into two complementary branches: **§2.1 General-Purpose** interactive simulators, which target broad physical world coverage, and **§2.2 Game & Roaming** environments, which specialize in open-ended 3D game worlds.

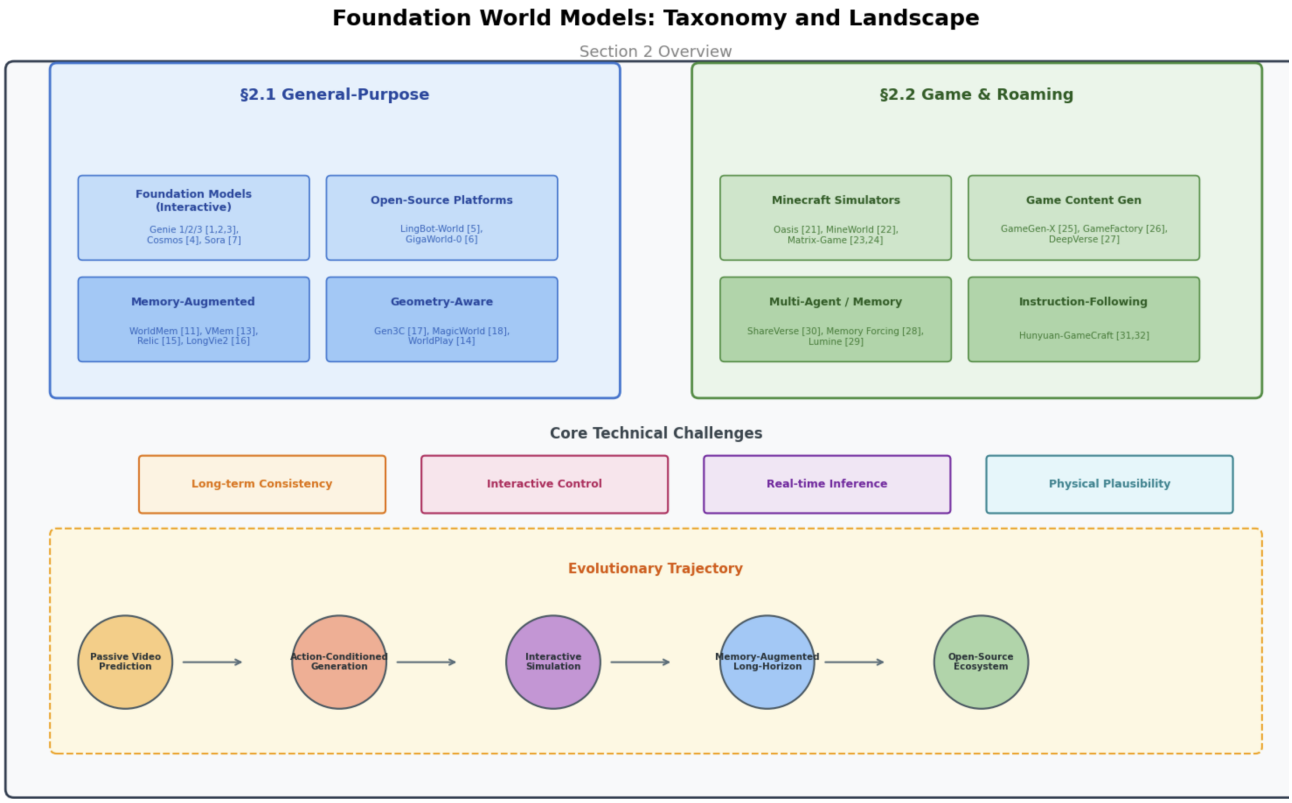

Figure 2. Taxonomy of Foundation World Models. The landscape is divided into General-Purpose (§2.1) and Game & Roaming (§2.2) branches, with representative works and core technical challenges highlighted.

2.1 General-Purpose Interactive Simulators

General-purpose world models aspire to simulate the full diversity of the physical world—indoor scenes, outdoor environments, object manipulation, and agent navigation—without domain-specific engineering. This subsection traces the evolution from proprietary foundation models to open-source ecosystems, and from short-horizon video prediction to memory-augmented, geometrically consistent long-horizon simulation.

> **Reader guidance.** The following subsections are ordered chronologically and by increasing capability: (2.1.1) proprietary foundation interactive models, (2.1.2) open-source platforms, (2.1.3) memory-augmented extensions, (2.1.4) geometry-aware models, and (2.1.5) theoretical perspectives.

2.1.1 Foundation Interactive Models

The modern era of interactive world modeling was inaugurated by Google DeepMind's **Genie** series. Genie [1] introduced a generative interactive environment trained exclusively from unlabelled internet videos, enabling users to control a generated 2D platformer world frame-by-frame via discrete actions. Its successor, **Genie 2** [2], scaled this paradigm to 3D, capable of generating diverse, action-controllable 3D environments from a single image prompt. **Genie 3** [3] further advanced real-time interactivity and world persistence. Genie 2 conditions generation on actions (e.g., "Forward," "Left," "Attack") through a latent world model core, decoding imagined states into photorealistic frames [2].

Concurrently, OpenAI's **Sora** [7] reframed video generation models as *world simulators*, demonstrating that large-scale diffusion transformers trained on massive video corpora acquire implicit understanding of physical dynamics, object permanence, and 3D geometry. While Sora was not explicitly action-conditioned, its technical report argued that such models constitute a foundational step toward universal world simulation.

NVIDIA's **Cosmos** platform [4] formalized this direction as a *World Foundation Model Platform for Physical AI*. Cosmos provides a family of autoregressive and diffusion-based world models, tokenizers, and video processing pipelines specifically designed for physical AI development. the Cosmos autoregressive architecture integrates video tokenization, text conditioning, and causal temporal modeling to enable scalable world generation [4].

On the navigation front, **Navigation World Models** [8] specialized world models for geographic and spatial reasoning, enabling traversal of imagined environments based on visual or textual navigation instructions.

2.1.2 Open-Source General-Purpose Platforms

The proprietary success of Genie and Cosmos catalyzed an open-source ecosystem. **LingBot-World** [5] from the Robbyant Team (Ant Group) advances open-source world models specifically for embodied intelligence, providing accessible training pipelines and model weights. **GigaWorld-0** [6] positions world models as *data engines* for embodied AI, emphasizing scalable data generation to bootstrap downstream robotic learning. These platforms collectively lower the barrier to entry and foster reproducible research in interactive world simulation.

2.1.3 Memory-Augmented Long-Horizon Models

A critical limitation of early video world models is the rapid degradation of visual coherence over long horizons—a phenomenon rooted in the compounding of prediction errors. A surge of recent work addresses this through explicit memory mechanisms. **WorldMem** [11] maintains an external memory of past scene states, enabling coherent object persistence and environmental state tracking across hundreds of frames. **VMem** [13] proposes a surfel-indexed view memory for consistent interactive video scene generation, while **Relic** [15] augments interactive video world models with long-horizon memory for sustained narrative and spatial consistency. **LongVie 2** [16] pushes this further with multimodal controllable ultra-long video generation. Wu et al. [12] explore video world models with long-term spatial memory, and **WorldPlay** [14] targets real-time interactive modeling with geometric consistency. WorldMem's capability to maintain object state (e.g., wheat growth, hay placement) over extended temporal horizons [11].

2.1.4 Geometry-Aware and 3D-Informed Models

Beyond pixel-level coherence, several works inject explicit 3D geometric priors to enhance world model fidelity. **Gen3C** [17] proposes 3D-informed world-consistent video generation with precise camera control, leveraging underlying geometric representations to ensure that generated views respect scene structure. **MagicWorld** [18] explores interactive geometry-driven video world exploration, enabling users to navigate imagined spaces with geometric plausibility. **Yume** [9, 10] and its successor Yume-1.5 provide text-controlled interactive world generation, bridging high-level language instructions with low-level visual simulation.

2.1.5 Theoretical Perspectives

Complementing these architectural advances, several works critically examine the relationship between video generation and true world modeling. Kang et al. [19] ask, "How far is video generation from world model?" and analyze this gap through the lens of physical law adherence, arguing that current video models still struggle with consistent physical simulation. Mei et al. [20] survey the applications of video generation models in robotics, identifying research challenges and future directions at the intersection of generative simulation and embodied control.

The general-purpose simulators above are designed for physical realism and breadth. In contrast, the following subsection (§2.2) sacrifices some physical generality in exchange for real-time streaming performance and long-horizon interactivity, by focusing on structured game environments like Minecraft.

2.2 Game & Roaming Environments

While general-purpose models aim for physical world breadth, **Game & Roaming** world models exploit the structured yet open-ended nature of 3D game environments to develop real-time, interactive, and often multi-agent simulation platforms.

These environments offer clear action spaces (keyboard/mouse inputs), deterministic physics, and abundant training data, making them ideal testbeds for scalable interactive world modeling.

> **Subsection organization.** We cover (2.2.1) Minecraft-based simulators, (2.2.2) open-world game generation, (2.2.3) memory and multi-agent game worlds, and (2.2.4) instruction-following game models.

2.2.1 Minecraft-Based Simulators

Minecraft has emerged as the *de facto* standard for evaluating open-ended interactive world models due to its voxel-based physics, procedurally generated terrain, and rich agent affordances. **Oasis** [21] from Decart presented the first large-scale Minecraft world model built on a diffusion transformer (DiT) architecture, processing visual observations and keyboard inputs to generate next-frame predictions in real time. Oasis employs a Vision Transformer-based VAE encoder/decoder paired with a central DiT backbone that conditions on user actions (W, A, S, D, mouse) to predict future frames [21].

Following Oasis, **MineWorld** [22] provided an open-source real-time interactive world model on Minecraft, democratizing access to training pipelines and model weights. The **Matrix-Game** series from Kunlun Wanwei pushed the frontier of streaming performance: **Matrix-Game 2.0** [23] achieved real-time open-source interactive world modeling at 25 FPS, while **Matrix-Game 3.0** [24] introduced long-horizon memory for sustained world state tracking. Matrix-Game 3.0 employs a teacher-student distillation framework with a memory pool to maintain temporal coherence during streaming generation [24].

2.2.2 Open-World Game Generation

Beyond simulating existing games, several works explore *creating* new interactive game experiences through generative models. **GameGen-X** [25] introduced the first diffusion transformer tailored for interactive open-world game video generation, supporting dynamic character control and environmental reactivity. **GameFactory** [26] proposes creating entirely new games via generative interactive videos, learning to synthesize game rules, assets, and dynamics from video corpora. GameFactory's action control module, which processes keyboard inputs through causal masked attention and sliding window mechanisms to condition video generation on player actions [26].

DeepVerse [27] frames 4D autoregressive video generation as a world model, targeting spatio-temporally consistent game world generation. These approaches collectively shift the focus from "simulating a given game" to "generating novel interactive worlds."

2.2.3 Memory and Multi-Agent Game Worlds

Long-horizon consistency is especially critical in persistent game worlds. **Memory Forcing** [28] introduces spatio-temporal memory mechanisms specifically for Minecraft scene generation, mitigating error accumulation through forced memory retrieval. **Lumine** [29] provides an open recipe for building generalist agents in 3D open worlds, combining world modeling with agent learning. **ShareVerse** [30] extends this to the multi-agent setting, enabling multiple agents to inhabit and interact within a *shared*, consistently generated world—an essential capability for multiplayer game simulation and social AI.

2.2.4 Instruction-Following Game Models

A recent line of work integrates high-level language instructions into game world generation. The **Hunyuan-GameCraft** series [31, 32] from Tencent develops instruction-following interactive game world models. **Hunyuan-GameCraft** [32] achieves high-dynamic interactive game video generation with hybrid history conditioning, while **Hunyuan-GameCraft-2** [31] advances instruction-following capabilities, allowing users to guide world generation through natural language commands rather than low-level action tokens.

Foundation world models have evolved rapidly from passive video predictors to interactive, memory-augmented, and geometrically aware simulators. The general-purpose branch pursues physical world coverage, while the game-specific branch achieves real-time streaming performance and creative world generation. However, these models only predict future observations—they do not output actions. The next section (§3) introduces Vision-Language-Action models, which directly address the challenge of grounding high-level instructions into low-level motor control.

3. Vision-Language-Action (VLA) Models

> **Section 3 roadmap.** We organize VLA models into three branches: **§3.1 Foundational VLA** architectures (RT-2, open-source generalist policies, flow/diffusion action representations); **§3.2 VLA for Autonomous Driving** (end-to-end driving, mixture-of-experts, reasoning-augmented); and **§3.3 VLA for Embodied Manipulation** (diffusion-based, 3D-aware, chain-of-thought, cross-embodiment). The progression moves from general-purpose robot policies to domain-specific driving and finally to dexterous manipulation.

While foundation world models (§2) focus on predicting future observations, **Vision-Language-Action (VLA) models** address the complementary challenge of *grounding* high-level human intent—expressed in natural language—into low-level physical control signals. By coalescing visual perception, linguistic reasoning, and motor action within a single end-to-end architecture, VLAs serve as the policy backbone for embodied agents operating in the real world. this section organizes the VLA landscape into three branches.

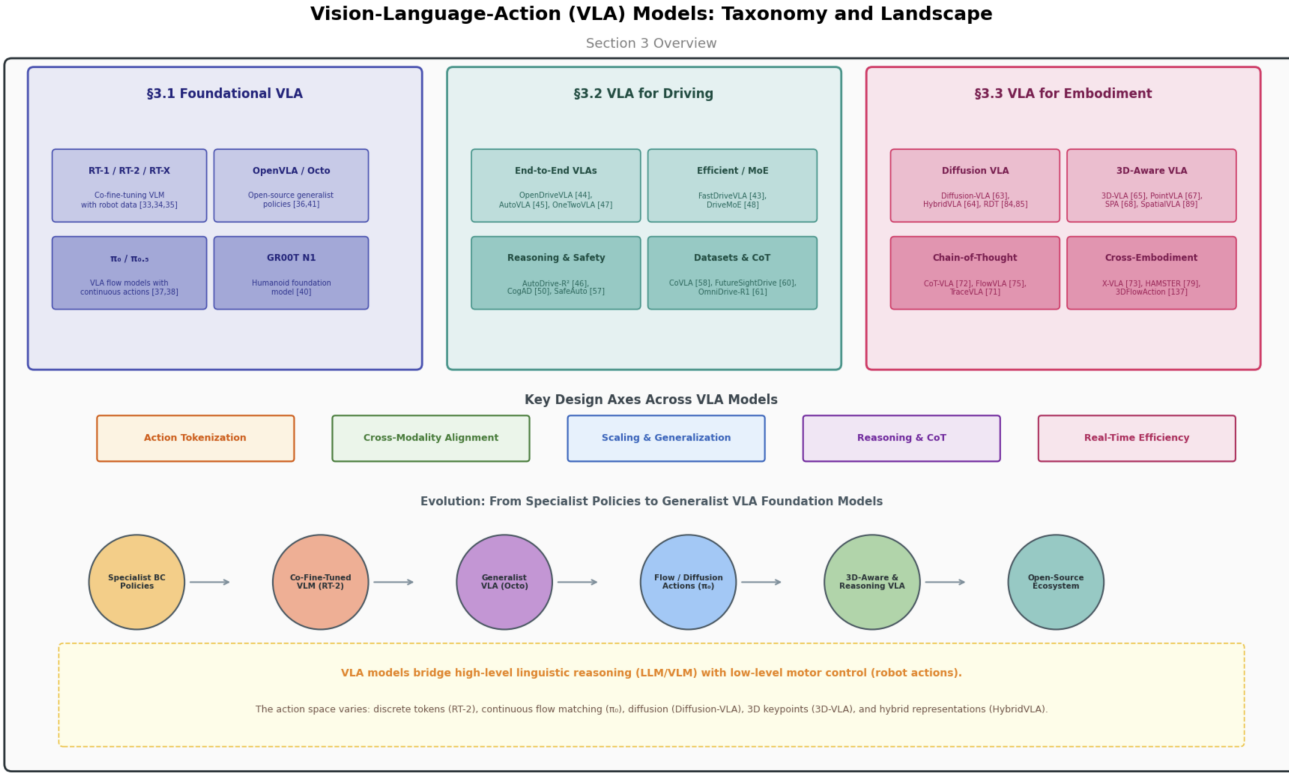

Figure 3. Taxonomy of Vision-Language-Action (VLA) Models. The landscape spans foundational architectures (§3.1), driving-specific adaptations (§3.2), and embodied manipulation policies (§3.3), with key design axes and evolutionary trajectory highlighted.

3.1 Foundational VLA Architectures

Foundational VLA models establish the canonical recipe: pre-train a large vision-language model (VLM) on internet-scale vision-language data, then fine-tune—or co-fine-tune—on robot demonstration data to align linguistic concepts with physical actions. This subsection traces the evolution from early scaling efforts to modern open-source generalist policies and advanced action representations.

3.1.1 Scaling Robot Control via VLMs

The VLA paradigm was pioneered by the **Robotics Transformer (RT)** series from Google DeepMind. **RT-1** [33] demonstrated that a transformer trained on a large-scale robot demonstration dataset could achieve robust generalization across tasks, environments, and robot morphologies, but remained a specialist policy without linguistic grounding. The breakthrough came with **RT-2** [34], which co-fine-tuned a vision-language model (PaLI-X) on both internet-scale VQA data

and robotic control data, treating robot actions as discrete tokens in the model's output vocabulary. RT-2 feeds visual observations through a ViT encoder into a large language model, which outputs action tokens subsequently de-tokenized into end-effector poses [34]. This insight—*actions are just another language*—enabled RT-2 to transfer semantic knowledge from web-scale pre-training to physical control, exhibiting emergent capabilities such as reasoning about object categories and following novel linguistic instructions.

Building on RT-2, the **Open X-Embodiment** collaboration produced **RT-X** [35], training on a dataset aggregated across 22 robot embodiments and hundreds of thousands of tasks. RT-X validated a central hypothesis of the VLA paradigm: cross-embodiment training improves generalization not only across robots but also within individual morphologies, establishing the "scaling laws" of robotic learning.

3.1.2 Open-Source Generalist Policies

The proprietary success of the RT series catalyzed an open-source ecosystem. **OpenVLA** [36] introduced the first fully open-source 7B-parameter VLA, built atop Llama 2 and DINOv2+SigLIP vision encoders, and trained on the Open X-Embodiment dataset. OpenVLA supports multi-robot control and efficient fine-tuning via LoRA, achieving competitive performance with closed-source counterparts while releasing model weights, data pipelines, and inference code [36].

Concurrently, **Octo** [41] proposed an open-source generalist robot policy based on a transformer architecture with flexible observation and action specifications, enabling out-of-the-box multi-robot control and efficient fine-tuning to new observation spaces. **MiniVLA** [42] explored the opposite direction: whether competitive VLA performance could be achieved with smaller footprints, challenging the assumption that scale is the sole driver of robotic capability.

3.1.3 Flow and Advanced Action Representations

A central design decision in VLA models is how to represent actions. While RT-2 discretizes continuous actions into tokens, this incurs quantization error and limits expressiveness. **π_0 (Pi-Zero)** [37] introduced a **vision-language-action flow model** that directly models continuous action distributions via flow matching. π_0 processes vision and language inputs through a pre-trained VLM, then employs a flow-based action head to generate high-fidelity, smooth action trajectories without tokenization bottlenecks [37]. **$\pi_{0.5}$** [38] extended this with open-world generalization, scaling to diverse real-world tasks. Figure 4 compares the three dominant action representation paradigms in modern VLAs.

On the humanoid front, NVIDIA's **GR00T N1** [40] established an open foundation model for generalist humanoid robots, integrating humanoid-specific kinematic priors with large-scale VLA pre-training. **CogAct** [39] pursued a cognitively inspired architecture, synergizing high-level "cognition" (scene understanding, task planning) with low-level "action" (motor control) within a single VLA framework.

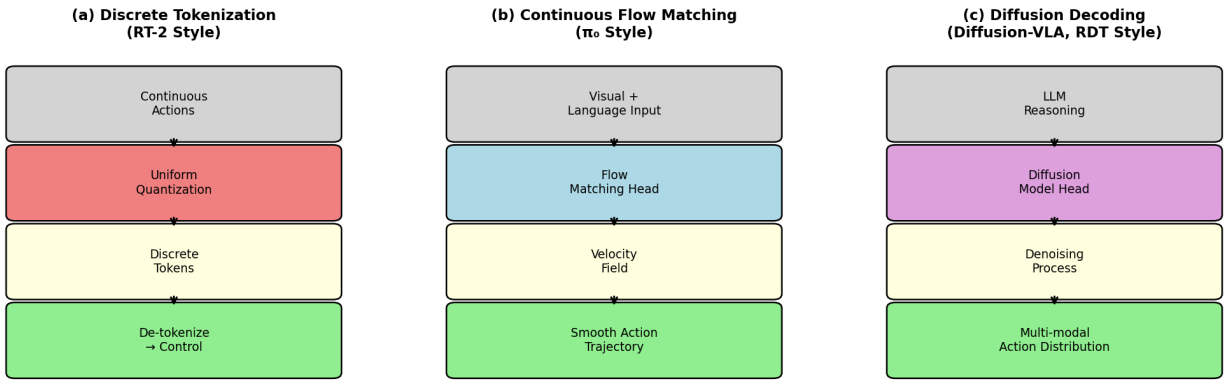

Figure 4. Action representation paradigms in VLA models. (a) Discrete tokenization (RT-2) treats actions as vocabulary tokens; (b) Continuous flow matching (π_0) generates smooth action chunks via velocity fields; (c) Diffusion decoding (Diffusion-VLA, RDT) denoises action trajectories conditioned on LLM reasoning.

3.2 VLA for Autonomous Driving

Autonomous driving presents a unique instantiation of the VLA paradigm: the "robot" is a vehicle, the "actions" are steering, throttle, and braking commands, and linguistic instructions often take the form of navigational commands or safety advisories. However, the safety-critical nature of driving demands specialized architectures for efficiency, interpretability, and robust reasoning.

3.2.1 End-to-End Driving VLAs

DriveVLM [62] proposed the convergence of autonomous driving and large vision-language models, using VLMs for scene understanding and trajectory planning. **OpenDriveVLA** [44] pushed this toward true end-to-end autonomy, integrating perception, prediction, and control within a single VLA architecture. OpenDriveVLA processes multi-view camera streams through a vision encoder, aligns features with a language model, and outputs driving actions directly [44].

AutoVLA [45] introduced adaptive reasoning and reinforcement fine-tuning (RFT) for driving VLAs, enabling the model to dynamically adjust its reasoning depth based on scenario complexity. **OneTwoVLA** [47] proposed a unified VLA with adaptive reasoning, while **FastDriveVLA** [43] addressed the critical deployment constraint of efficiency via plug-and-play reconstruction-based token pruning.

3.2.2 Mixture-of-Experts and Efficiency

Real-time inference is paramount in autonomous driving. **DriveMoE** [48] applied mixture-of-experts (MoE) architectures to VLA models for end-to-end driving, dynamically routing computations through expert sub-networks based on visual inputs. DriveMoE replaces dense vision and action processing with sparse expert activation, reducing computational overhead while maintaining multi-behavior trajectory distributions [48].

3.2.3 Reasoning, Safety, and Chain-of-Thought

Beyond raw control, driving VLAs increasingly emphasize *cognitive* capabilities. **AutoDrive-R²** [46] incentivized reasoning and self-reflection capacity in driving VLAs, enabling the model to critique its own decisions. **CogAD** [50] proposed cognitive-hierarchy guided end-to-end driving, while **SafeAuto** [57] integrated knowledge-enhanced safety mechanisms. **FutureSightDrive** [60] introduced spatio-temporal chain-of-thought (CoT) reasoning for visual thinking, and **OmniDrive-R1** [61] leveraged reinforcement-driven interleaved multi-modal chain-of-thought. These works collectively shift the driving VLA paradigm from "reactive control" to "deliberative reasoning."

3.2.4 Datasets and Perception Integration

Specialized datasets and perception architectures have enabled driving VLA development. **CoVLA** [58] and its open-weight successor **Impromptu VLA** [59] provided comprehensive vision-language-action datasets for autonomous driving. Perceptually grounded models such as **S4-Driver** [51] introduced scalable self-supervised driving multimodal LLMs, while **DriveVPT4** [52] and **RAG-Driver** [53] integrated interpretable large language models and retrieval-augmented in-context learning, respectively. **RLGf** [54] explored reinforcement learning with geometric feedback.

3.3 VLA for Embodied Manipulation

While driving VLAs operate in structured, planar action spaces, **embodied manipulation VLAs** must contend with high-dimensional, dexterous control in cluttered 3D environments. This subsection surveys the dominant architectural families: diffusion-based action generation, 3D-aware spatial reasoning, chain-of-thought deliberation, and cross-embodiment generalization.

3.3.1 Diffusion-Based VLA Models

Diffusion models have emerged as a powerful alternative to autoregressive action generation, offering multi-modal, composable, and high-fidelity action distributions. **Diffusion-VLA** [63] scaled robot foundation models via unified diffusion and autoregression, using an LLM for reasoning and a diffusion model for action decoding. the model injects reasoning into diffusion to guide action generation [63].

HybridVLA [64] combined collaborative diffusion and autoregression in a unified VLA, while **RDT-1B** [84] established a diffusion foundation model specifically for bimanual manipulation, scaling to 1B parameters. **RDT2** [85] pushed this scaling limit further toward zero-shot cross-embodiment transfer. The foundational **Diffusion Policy** [86] and **Universal Manipulation Interface (UMI)** [87] provided the algorithmic and data-collection infrastructure underpinning these diffusion VLAs.

3.3.2 3D-Aware and Spatial VLA Models

A critical limitation of early VLAs is their reliance on 2D image tokens, which discard depth and spatial layout information. **3D-VLA** [65] addresses this by injecting 3D grounding into the VLA framework, using 3D feature extractors to align visual observations with language and action spaces. 3D-VLA supports both 3D imagination (goal prediction) and robot control within a generative world model framework [65].

PointVLA [67] injects 3D world representations via point cloud encoders, while **SPA** [68] demonstrates that explicit 3D spatial reasoning enables more effective embodied representation. **SpatialVLA** [89] explores spatial representations for VLAs, and **VGG-T** [96] introduces visual geometry grounded transformers. These 3D-aware architectures consistently outperform 2D counterparts on spatial reasoning benchmarks.

3.3.3 Chain-of-Thought and Reasoning VLAs

Following the success of chain-of-thought reasoning in LLMs, researchers have extended this paradigm to robotic control. **CoT-VLA** [72] introduces visual chain-of-thought reasoning for VLAs, generating intermediate reasoning steps before emitting actions. **FlowVLA** [75] combines visual chain-of-thought with motion reasoning, while **TraceVLA** [71] enhances spatial-temporal awareness through visual trace prompting. **MolmoAct** [88] demonstrates that action reasoning models can explicitly reason in 3D space.

3.3.4 Cross-Embodiment and Generalist Policies

A defining aspiration of the VLA paradigm is to train a single policy that generalizes across robot morphologies without embodiment-specific engineering. **X-VLA** [73] proposes soft-prompted transformers as scalable cross-embodiment VLAs. **HAMSTER** [79] introduces hierarchical action models for open-world manipulation. **3DFlowAction** [137] learns cross-embodiment manipulation from 3D flow world models.

On the humanoid and legged front, **Quar-VLA** [94] specializes VLAs for quadruped robots, while **Humanoid-VLA** [95] targets universal humanoid control. **GR-2** [99] proposes a generative video-language-action model with web-scale knowledge, and **Bi-VLA** [100] addresses bimanual dexterous manipulation.

3.3.5 Multimodal and Specialized Embodied VLAs

Recent work has expanded the VLA input space beyond vision and text to include speech, tactile sensing, and multi-frame temporal context. **VLAS** [74] integrates speech instructions into VLA models. **CronusVLA** [93] transfers latent motion across time for multi-frame prediction. **UP-VLA** [69] unifies understanding and prediction. **InstructVLA** [91] and **F1** [90] explore instruction tuning and bridging understanding-to-generation. **RoboInter** [92] proposes a holistic intermediate representation suite. **LingBot-VLA** [76] develops a pragmatic VLA foundation model for real-world manipulation.

Navigation-specific VLAs have also matured: **Mobility VLA** [80] combines multimodal instruction navigation with long-context VLMs and topological graphs, while **Nora** [82] provides a small open-sourced generalist VLA for embodied tasks. The **Galaxea** team [78] released an open-world dataset and the G0 dual-system VLA model, and **ABot-M0** [77] introduces action manifold learning.

Vision-Language-Action models have evolved into a mature ecosystem spanning open-source generalist policies, advanced action representations, safety-critical driving systems, and dexterous embodied agents. Three design axes unify this landscape: action tokenization, cross-modality alignment, and reasoning depth. Nevertheless, these VLAs are typically trained on fixed datasets and do not themselves learn to simulate future states. The next section (§4) introduces Embodied World Action Models, which unify world simulation and action generation within a single framework.

4. Embodied World Action Models (WAMs)

Foundation world models (§2) predict what the future will look like; Vision-Language-Action models (§3) decide what actions to execute. **Embodied World Action Models (WAMs)** represent the natural convergence of these two paradigms: architectures that simultaneously simulate future states *and* generate executable actions within a unified generative framework. Rather than treating world simulation and policy execution as decoupled pipelines, WAMs learn to imagine the consequences of actions and directly emit control signals, or they serve as differentiable simulators within which policies can be trained and refined without costly real-world interaction. this section organizes WAMs into three branches.

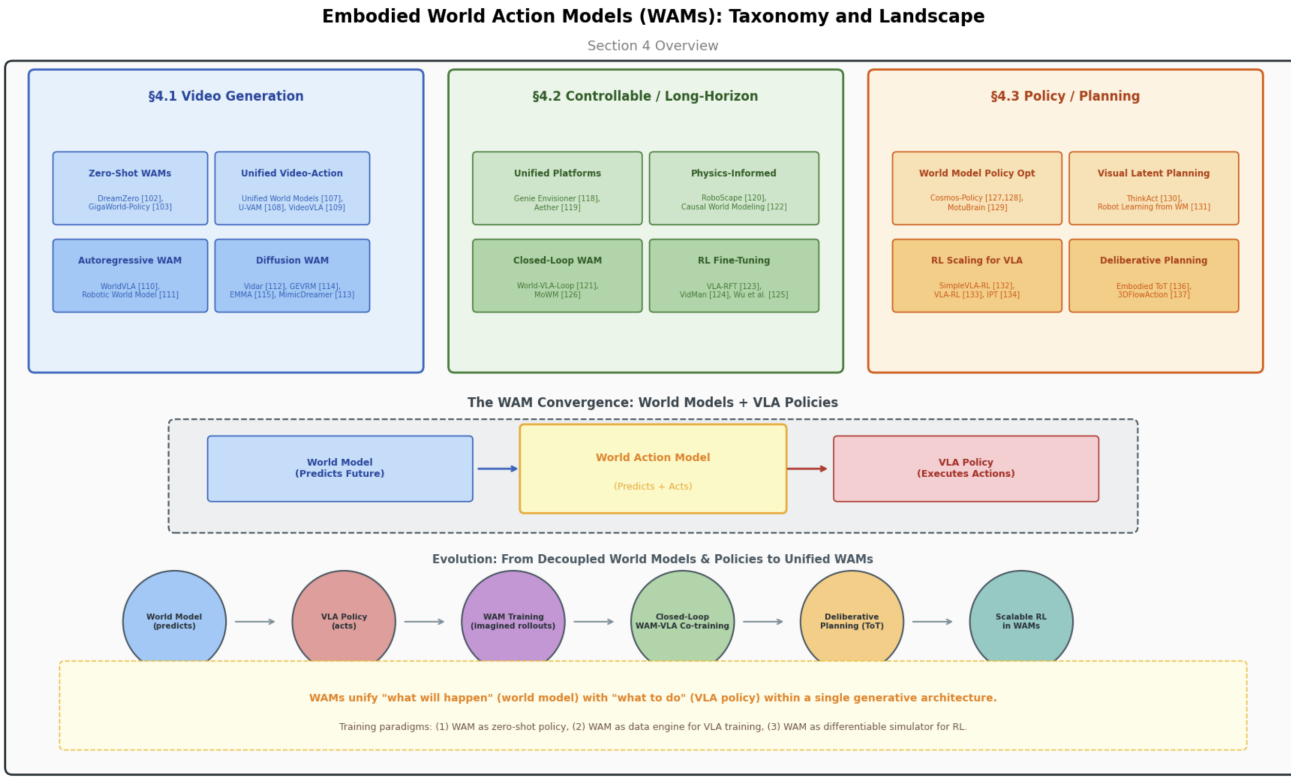

Figure 5. Taxonomy of Embodied World Action Models (WAMs). The landscape spans video-generation-based WAMs (§4.1), controllable simulation platforms (§4.2), and policy/planning frameworks (§4.3), illustrating the convergence of world models and VLA policies.

4.1 Video-Generation-Based WAMs

The simplest instantiation of a WAM treats video generation and action prediction as a single unified objective: given a visual observation and a linguistic or motor intent, the model generates a future video sequence that implicitly encodes the action trajectory.

4.1.1 Zero-Shot WAM Policies

A recent hypothesis is that a sufficiently capable video world model, conditioned on action instructions, can serve as a *zero-shot policy* without explicit policy training. NVIDIA's **DreamZero** [102] validated this at scale, demonstrating that a 1.4B-parameter world action model can directly output motor commands for unseen tasks by predicting the future video sequence of successful execution. DreamZero frames policy execution as future video prediction. **Fast-WAM** [104] challenged the computational necessity of full future imagination at test time, proposing efficient inference schemes that bypass explicit rollout generation. **GigaWorld-Policy** [103] optimized the architecture for policy-centric rather than pixel-centric prediction.

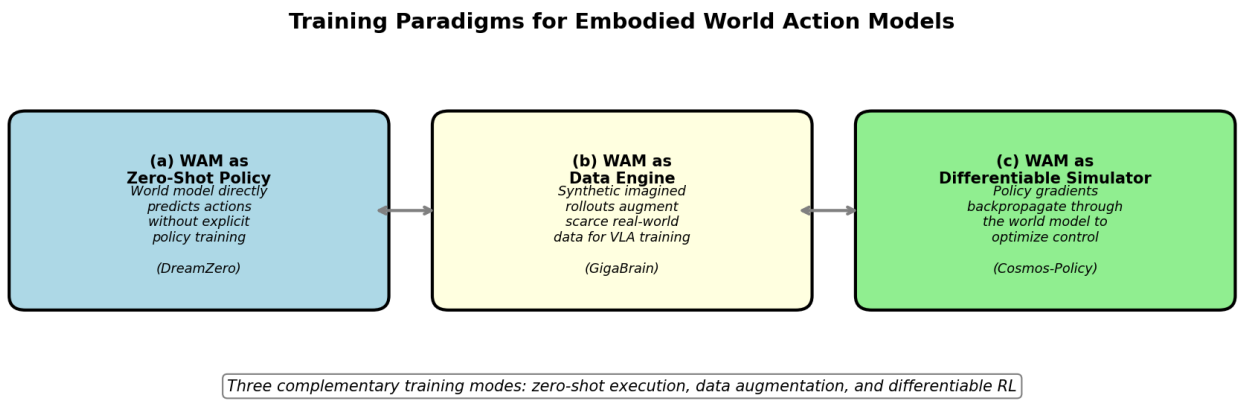

Figure 6. Training paradigms for Embodied World Action Models. (a) WAM as zero-shot policy: the model directly predicts actions without explicit policy training. (b) WAM as data engine: synthetic imagined rollouts augment scarce real-world data for VLA training. (c) WAM as differentiable simulator: policy gradients backpropagate through the world model to optimize control.

4.1.2 Unified Video-Action Pretraining

Beyond zero-shot inference, a parallel line of work explores *unified pretraining* objectives that couple video prediction and action generation. **Unified World Models** [107] proposed coupling video and action diffusion for pretraining on large robotic datasets, showing that a shared diffusion transformer can simultaneously denoise future visual observations and action trajectories. this architecture supports forward dynamics, inverse dynamics, policy optimization, and video prediction within a single model [107]. **Unified Video Action Model (U-VAM)** [108] pursued a similar unification via autoregressive modeling, while **VideoVLA** [109] demonstrated that off-the-shelf video generators can serve as generalizable robot manipulators. **WorldVLA** [110] formalized autoregressive action world modeling, and **Robotic World Model** [111] introduced a neural network simulator for robust policy optimization.

4.1.3 Diffusion and Generative WAMs

Diffusion models naturally fit WAMs because they inherently model multi-modal distributions over future states. **Vidar** [112] introduced an embodied video diffusion foundation model for generalist manipulation, training on only 20 minutes of real robot data yet achieving cross-embodiment generalization. Vidar processes observations and instructions through a video diffusion backbone, generating future frames that guide motor control [112]. **GEVRM** [114] proposed a goal-expressive video generation model, **EMMA** [115] generalized real-world manipulation via generative visual transfer, and **MimicDreamer** [113] aligned human and robot demonstrations within a shared WAM latent space. **Video Prediction Policy** [116] demonstrated that predictive visual representations from video world models can serve as generalist policies, and **Video Generators are Robot Policies** [117] synthesized this paradigm.

4.2 Controllable and Long-Horizon Simulation

While video-generation WAMs excel at short-horizon action prediction, they often struggle with physical plausibility and long-term consistency. This subsection surveys platforms that ground WAMs in physical constraints and enable closed-loop co-training.

4.2.1 Unified Simulation Platforms

Genie Envisioner [118] extended the Genie architecture with action-conditioned generation and causal consistency blocks for manipulation tasks. it employs an autoregressive video generation process with causal blocks for spatio-temporal information exchange [118]. **Aether** [119] introduced geometric-aware unified world modeling.

4.2.2 Physics-Informed World Models

To address physically implausible futures, researchers have integrated explicit physical priors. **RoboScape** [120] proposed a physics-informed embodied world model that jointly learns temporal depth estimation and adaptively sampled keypoint dynamics. the model integrates keypoint prediction, RGB synthesis, and depth feedback with physical alignment losses [120]. **Causal World Modeling for Robot Control** [122] learned causal structure among environmental variables.

4.2.3 Closed-Loop WAM-VLA Co-training

A critical recent insight is that world models and policies should be trained in a *closed loop*, where each improves the other. **World-VLA-Loop** [121] formalized closed-loop learning of a video world model and VLA policy: the VLA policy generates actions, the world model simulates their consequences, and the resulting imagined trajectories refine both. this co-evolutionary paradigm boosts real-world performance [121]. **MoWM** [126] extended this to a mixture-of-world-models framework for embodied planning.

4.2.4 RL Fine-Tuning in World Simulators

World models offer an ideal substrate for reinforcement learning: infinite rollouts, gradient-based optimization, and safe exploration. **VLA-RFT** [123] introduced vision-language-action reinforcement fine-tuning with verified rewards in world simulators. **VidMan** [124] exploited implicit dynamics from video diffusion models, and Wu et al. [125] demonstrated that large-scale video generative pre-training transfers actionable physical priors.

4.3 Policy and Planning Frameworks

The ultimate goal of WAMs is not merely to generate plausible futures, but to *act optimally* within them. This subsection surveys methods that leverage world models as policy optimizers, planning substrates, and reasoning engines.

4.3.1 World Model-Based Policy Optimization

Cosmos-Policy [127] introduced world model-based policy optimization for VLAs, fine-tuning video world models to serve as policy evaluation substrates. after fine-tuning on robot demonstrations, the world model becomes a reliable simulator for planning [127,128]. **MotuBrain** [129] advanced this with a dedicated world action model for robot manipulation.

4.3.2 Visual Latent Planning

World models can serve as *planning substrates* for look-ahead reasoning. **ThinkAct** [130] proposed vision-language-action reasoning via reinforced visual latent planning. it processes observations through a state encoder, uses a reasoning MLLM to generate sub-goals, and refines actions via GRPO optimization [130]. **Robot Learning from a Physical World Model** [131] treated the model as a "mental simulator."

4.3.3 Reinforcement Learning for VLAs

SimpleVLA-RL [132] scaled VLA training via reinforcement learning, showing that even simple RL algorithms improve VLAs when paired with world model rollouts. **VLA-RL** [133] pursued masterful robotic manipulation with scalable RL, and **Interactive Post-Training for VLAs** [134] explored continued policy improvement after supervised fine-tuning.

4.3.4 Deliberative Planning

For long-horizon, multi-step tasks, agents must *deliberate* over possible future sequences. **Embodied Tree of Thoughts** [136] introduced deliberate manipulation planning using a WAM to simulate multiple candidate action sequences and select the optimal branch via tree search. the robot imagines consequences of different strategies and executes the most promising plan [136]. **3DFlowAction** [137] learned cross-embodiment manipulation from 3D flow world models.

4.3.5 Evaluation via World Models

World models are increasingly used as *evaluators* of robotic policies. **Scalable Robotic Policy Evaluation via Discrete Diffusion World Model** [135] demonstrated that a discrete diffusion world model can reliably evaluate policy performance across thousands of imagined scenarios, closing the loop between WAM development and benchmarking.

Embodied World Action Models have matured into a concrete paradigm with three training modes: zero-shot policy execution, data engine augmentation, and differentiable simulation for RL. These principles transfer directly to the safety-critical domain of autonomous driving, where world models must not only imagine traffic futures but also plan collision-free trajectories—the focus of the next section.

5. Autonomous Driving World Models

Autonomous driving represents one of the most demanding and safety-critical application domains for world models. Compared to general-purpose simulators (§2) or embodied manipulation environments (§4), driving world models must handle high-speed dynamic scenes, strict geometric constraints, multi-agent interactions, and regulatory safety requirements. this section organizes autonomous driving world models into four branches.

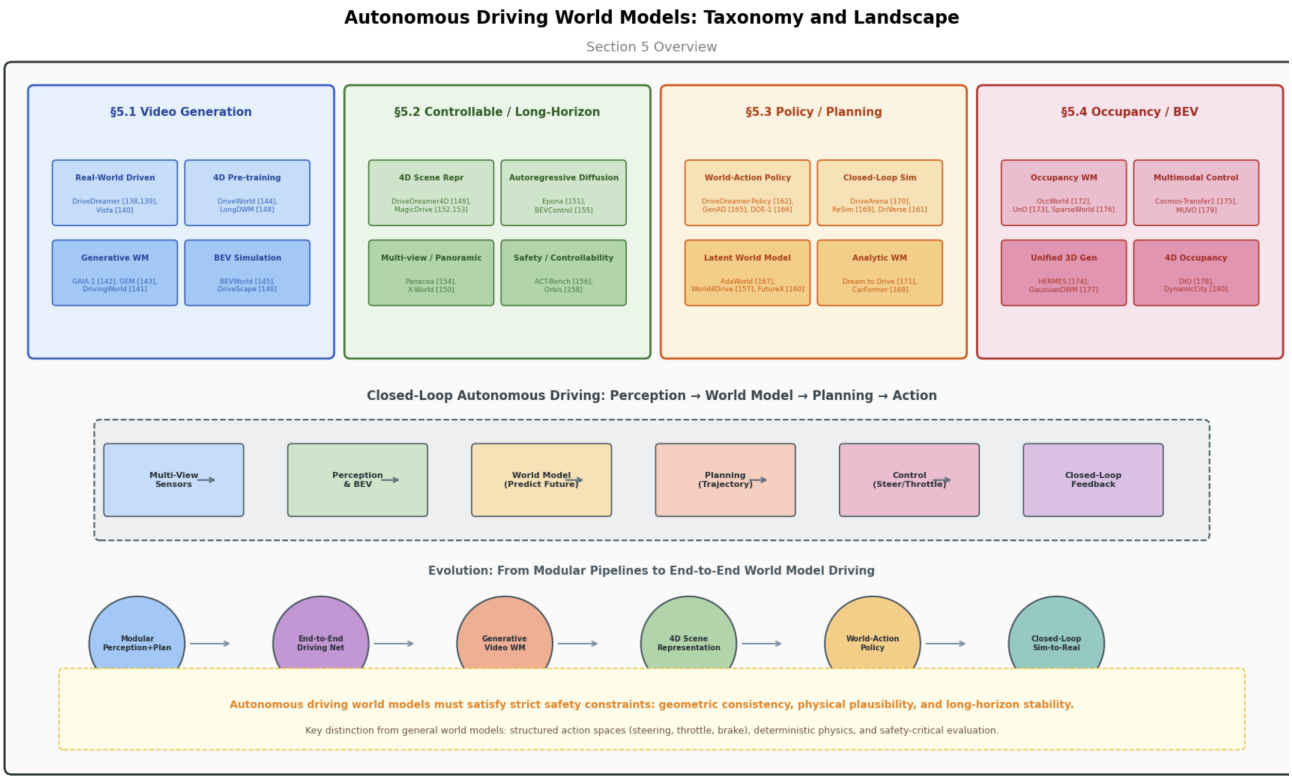

Figure 7. Taxonomy of Autonomous Driving World Models. The landscape spans video generation (§5.1), controllable simulation (§5.2), planning policies (§5.3), and occupancy/BEV representations (§5.4), with the closed-loop driving pipeline highlighted.

5.1 Video Generation

The earliest and most mature application of driving world models is **video generation**—synthesizing photorealistic, temporally coherent driving sequences for data augmentation.

5.1.1 Real-World-Driven Generation

DriveDreamer [138] established the paradigm of real-world-driven world models, introducing a two-stage pipeline that first learns traffic structure from unlabeled videos and then conditions generation on driving actions. DriveDreamer excels in controllable driving video generation [138]. **DriveDreamer-2** [139] enhanced this with LLM-driven scenario generation, and **Vista** [140] introduced a generalizable driving world model with high fidelity.

5.1.2 Generative World Models

GAIA-1 [142] introduced a 9-billion parameter generative world model, demonstrating that large-scale video transformers acquire implicit understanding of traffic dynamics and agent behavior. GAIA-1 uses an autoregressive discrete diffusion architecture [142]. **GEM** [143] proposed a generalizable ego-vision multimodal world model, while **DrivingWorld** [141] constructed a driving-specific world model via Video GPT.

5.1.3 4D Pre-training and BEV Simulation

Beyond video synthesis, **DriveWorld** [144] introduced 4D pre-trained scene understanding via world models. **BEVWorld** [145] proposed a multimodal world simulator operating in scene-level BEV latent space. BEVWorld encodes multi-view images and LiDAR into unified BEV representations for controllable generation [145]. **LongDWM** [148] addressed long-horizon modeling via cross-granularity distillation.

5.2 Controllable and Long-Horizon Simulation

While video generation produces realistic frames, **controllable simulation** requires that generated futures respond predictably to control inputs and maintain physical plausibility over long horizons.

5.2.1 4D Scene Representation

DriveDreamer4D [149] proposed that world models are effective data machines for 4D driving scene representation, extending DriveDreamer with explicit 4D reconstruction. it integrates novel trajectory generation and conditional denoising while preserving scene geometry [149]. **MagicDrive** [152,153] advanced high-resolution long video generation, and **MagicDrive3D** [153] extended this to controllable 3D generation.

5.2.2 Multi-View and Panoramic Generation

Real-world autonomous vehicles use surround-view camera arrays. **Panacea** [154] introduced panoramic and controllable video generation for autonomous driving, synthesizing 360° surround-view videos. Panacea processes BEV sequences through ControlNet-conditioned diffusion to generate panoramic videos [154]. **X-World** [150] proposed controllable ego-centric multi-camera world models.

5.2.3 Autoregressive Diffusion and Long-Horizon Models

Epona [151] introduced an autoregressive diffusion world model, combining diffusion expressiveness with autoregressive temporal coherence. Epona supports trajectory-controlled generation, traffic rule understanding, and end-to-end planning [151]. **BEVControl** [155] enabled accurate control via BEV sketch layouts.

5.2.4 Safety, Controllability, and Benchmarks

ACT-Bench [156] established an action controllable world model benchmark. **World4Drive** [157] introduced intention-aware physical latent world models, and **FutureX** [160] enhanced end-to-end driving via latent chain-of-thought. **DriVerse** [161] proposed a navigation world model with multimodal trajectory prompting. **Orbis** [158] addressed long-horizon prediction error accumulation.

5.3 Policy and Planning

Video generation and controllable simulation provide the substrate; **policy and planning** frameworks determine how autonomous vehicles navigate within imagined futures.

5.3.1 World-Action Policies for Driving

DriveDreamer-Policy [162] introduced a geometry-grounded world–action model for unified generation and planning. it generates multi-view depth and video alongside BEV planning outputs [162]. **GenAD** [165] proposed generative end-to-end autonomous driving, framing planning as conditional generation within a latent trajectory space [165]. **DOE-1** [166] advanced closed-loop driving with large world models [166].

5.3.2 Latent World Model Planning

AdaWorld [167] introduced learning adaptable world models with latent actions. **World4Drive** [157] demonstrated that intention-aware latent models converge 3.75× faster than conventional approaches. **FutureX** [160] enhanced planning via latent chain-of-thought, and **Think2Drive** [147] used latent world models for efficient RL.

5.3.3 Closed-Loop Simulation Platforms

DriveArena [170] introduced a closed-loop generative simulation platform, offering a scalable middle ground between real data and traditional simulators. **ReSim** [169] provided reliable world simulation, and **DriVerse** [161] enabled navigation world modeling.

5.3.4 Analytic and Physics-Based World Models

Dream to Drive [171] proposed model-based vehicle control using analytic world models. **CarFormer** [168] introduced self-driving with learned object-centric representations. **Generalized Predictive Model** [163] and **Driving into the Future** [164] established predictive planning paradigms.

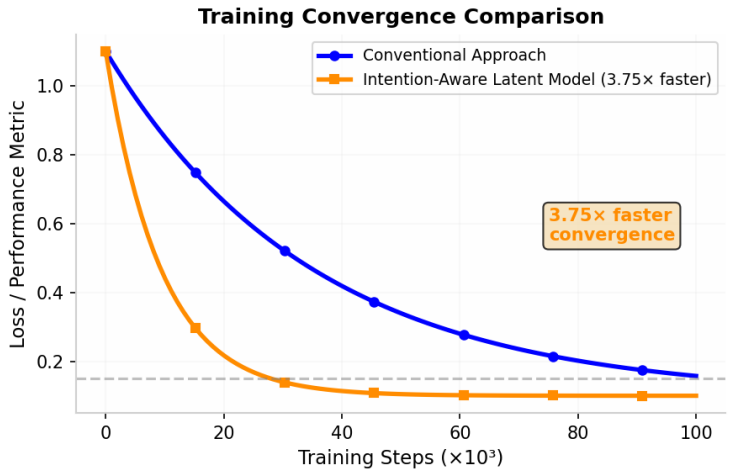

Figure 8. World4Drive training efficiency (Redrawn based on original concept.)

5.4 Occupancy and BEV Representations

While video generation operates in pixel space, **occupancy and BEV representations** provide structured 3D geometric substrates more amenable to safety verification and planning.

5.4.1 Occupancy World Models

OccWorld [172] pioneered learning 3D occupancy world models, directly predicting future occupancy grids. it employs a GPT-style architecture with spatial aggregation and temporal causal attention [172]. **UnO** [173] introduced unsupervised occupancy fields, and **SparseWorld** [176] proposed efficient 4D occupancy models with sparse queries.

5.4.2 Unified 3D Scene Understanding and Generation

HERMES [174] proposed a unified self-driving world model for simultaneous 3D scene understanding and generation. **GaussianDWM** [177] introduced a 3D Gaussian driving world model, and **MUVO** [179] proposed a multimodal generative world model with geometric representations.

5.4.3 Multimodal Control and 4D Occupancy

Cosmos-Transfer1 [175] introduced conditional world generation with adaptive multimodal control. **DIO** [178] proposed decomposable implicit 4D occupancy-flow world models, and **DynamicCity** [180] addressed large-scale 4D occupancy generation.

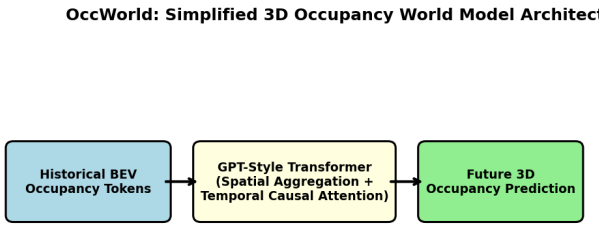

Figure 9. OccWorld framework (Redrawn based on original concept.)

Autonomous driving world models have evolved from pixel-level video generators into sophisticated, structured simulation platforms supporting closed-loop policy evaluation and geometric planning. The maturation of these models has created a pressing need for computational efficiency and rigorous evaluation—the focus of the next section.

6. Efficiency and Evaluation

The rapid scaling of world models and VLA architectures—exemplified by billion-parameter foundation models—has created a pressing need for **computational efficiency** and **rigorous evaluation**. On the efficiency front, deploying large VLAs on real robots demands sub-second inference latency and modest memory; on the evaluation front, the community requires standardized protocols that assess physical plausibility, long-horizon consistency, and correlation with real-world policy performance. this section organizes the landscape into two parts.

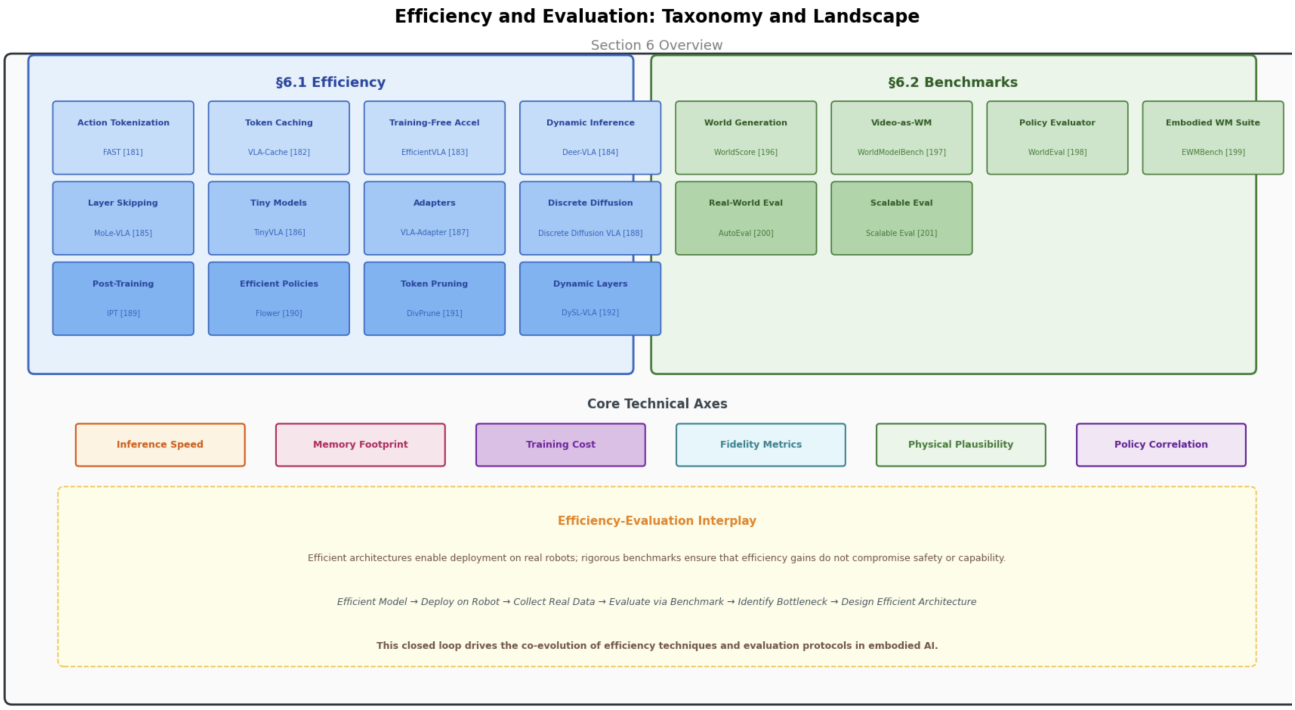

Figure 10. Taxonomy of Efficiency and Evaluation in World Models and VLAs. The landscape spans computational acceleration techniques (§6.1) and benchmarking protocols (§6.2).

6.1 Efficiency

Efficiency research targets three bottlenecks: **action representation**, **vision-language backbone optimization**, and **system-level memory management**. We survey advances across these axes.

6.1.1 Action Tokenization and Decoding

While early VLAs discretized actions into uniform bins (RT-2 [34]), this incurs quantization error. **FAST** [181] introduced efficient action tokenization using discrete cosine transform (DCT)-based compression. FAST achieves 5× faster training while improving generalization [181]. **Discrete Diffusion VLA** [188] treats action generation as a masked discrete diffusion process, requiring fewer forward passes than autoregressive alternatives.

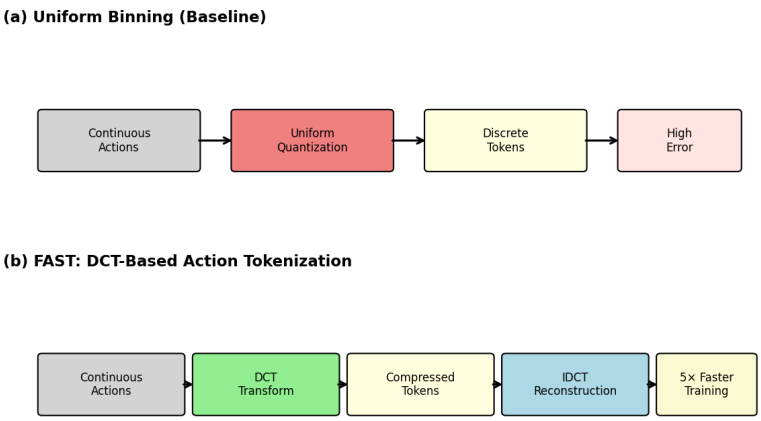

Figure 11. FAST efficient action tokenization (Redrawn based on original concept.)

6.1.2 Vision-Language Backbone Optimization

VLA-Cache [182] proposed adaptive token caching, reusing static visual tokens across timesteps. **EfficientVLA** [183] introduced training-free acceleration, leveraging layer-wise similarity and temporal redundancy. **Deer-VLA** [184] introduced early-exit mechanisms, allowing simple queries to exit at shallow layers. **MoLe-VLA** [185] proposed dynamic layer-skipping via mixture-of-layers, and **DySL-VLA** [192] extended this with dynamic-static layer skipping.

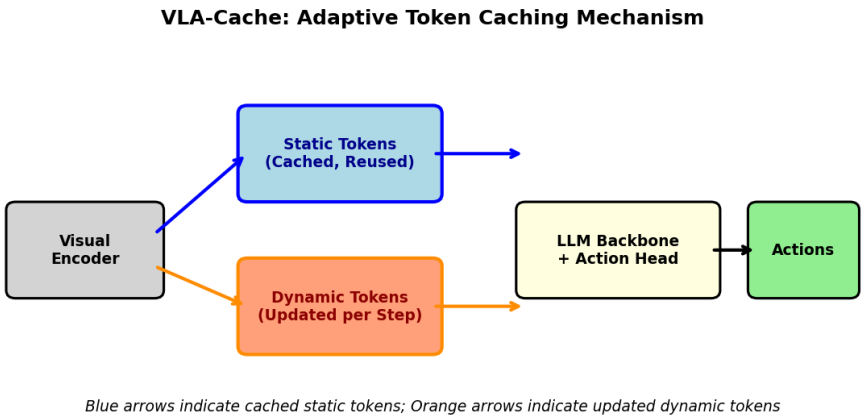

Figure 12. VLA-Cache adaptive token caching (Redrawn based on original concept.)

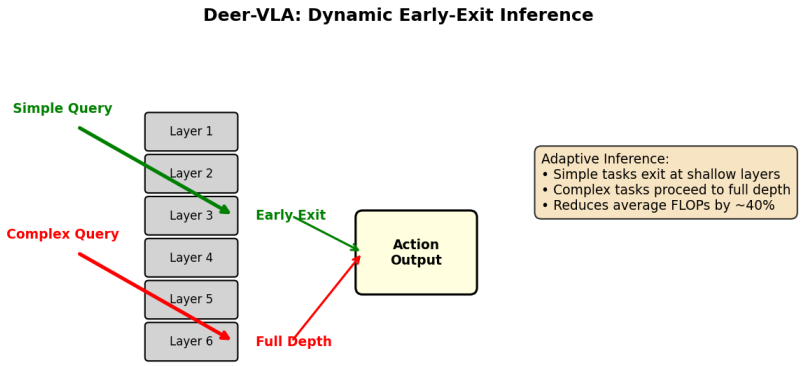

Figure 13. Deer-VLA dynamic inference with early-exit (Redrawn based on original concept.)

6.1.3 Model Scaling and Tiny Architectures

TinyVLA [186] demonstrated 20× lower latency than OpenVLA with competitive success rates. **VLA-Adapter** [187] proposed an effective paradigm for tiny-scale VLAs using lightweight adapters. **Flower** [190] democratized generalist policies with efficient flow policies.

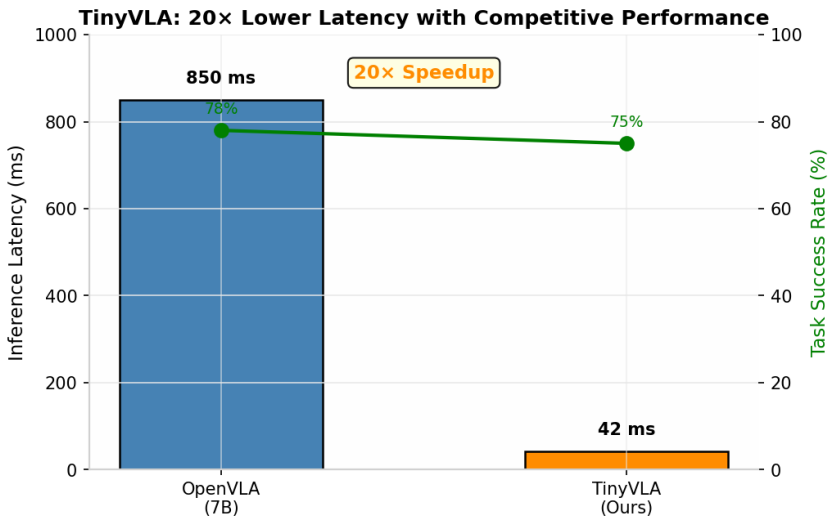

Figure 14. TinyVLA scaling efficiency (Redrawn based on original concept.)

6.1.4 Token Pruning and Memory Management

DivPrune [191] introduced diversity-based visual token pruning. **SP-VLA** [195] proposed joint model scheduling and token pruning. **KEEP** [193] introduced a KV-cache-centric memory management system, and **KERV** [194] proposed kinematic-rectified speculative decoding.

6.1.5 Post-Training Efficiency

Interactive Post-Training (IPT) [189] for VLAs demonstrated that continued interaction with the environment can improve both policy performance and inference efficiency by distilling the model toward simpler action distributions.

6.2 Benchmarks

Traditional computer vision metrics (FID, PSNR) are insufficient for embodied AI. Benchmarks must assess **physical plausibility**, **long-horizon consistency**, and **correlation with real-world policy success**.

6.2.1 World Generation Benchmarks

WorldScore [196] established a unified evaluation benchmark for world generation, moving beyond pixel quality to scene consistency and physical rule adherence. WorldScore reveals cases where models score highly on traditional metrics but fail on world-specific criteria [196]. **WorldModelBench** [197] proposes judging video generation models as world models, explicitly testing obedience to physical laws.

6.2.2 Embodied World Model Benchmarks

EWMBench [199] introduced an embodied world model benchmark suite, assessing scene quality, motion coherence, and semantic fidelity in interactive environments. **WorldArena** provides a unified benchmark for evaluating perception and functional utility.

6.2.3 Policy Evaluation via World Models

WorldEval [198] formalized the world model as a real-world robot policy evaluator, demonstrating strong correlation with physical robot success rates. **AutoEval** [200] introduced autonomous evaluation of generalist robot policies in the real world, achieving high correlation with human evaluation. **Scalable Robotic Policy Evaluation via Discrete Diffusion World Model** [135] extended this to large-scale imagined rollouts.

Section 6 summary and transition to §7. Efficiency and evaluation are the twin pillars that will determine whether large world models transition from research curiosities to deployed robotic systems. The data infrastructure that makes these advances possible—large-scale robot learning datasets, simulation environments, and industry ecosystems—is the subject of the next section.

7. Datasets and Ecosystems

The remarkable progress in world models, VLAs, and embodied AI has been underpinned by a parallel revolution in **data infrastructure**. Unlike computer vision or NLP, embodied AI requires physical interaction data—robot trajectories, human hand demonstrations, 3D scene layouts—that is costly to collect. The emergence of large-scale, open-source robot learning datasets and industry technical reports has democratized access to this data substrate. this section organizes the data and ecosystem landscape into two parts.

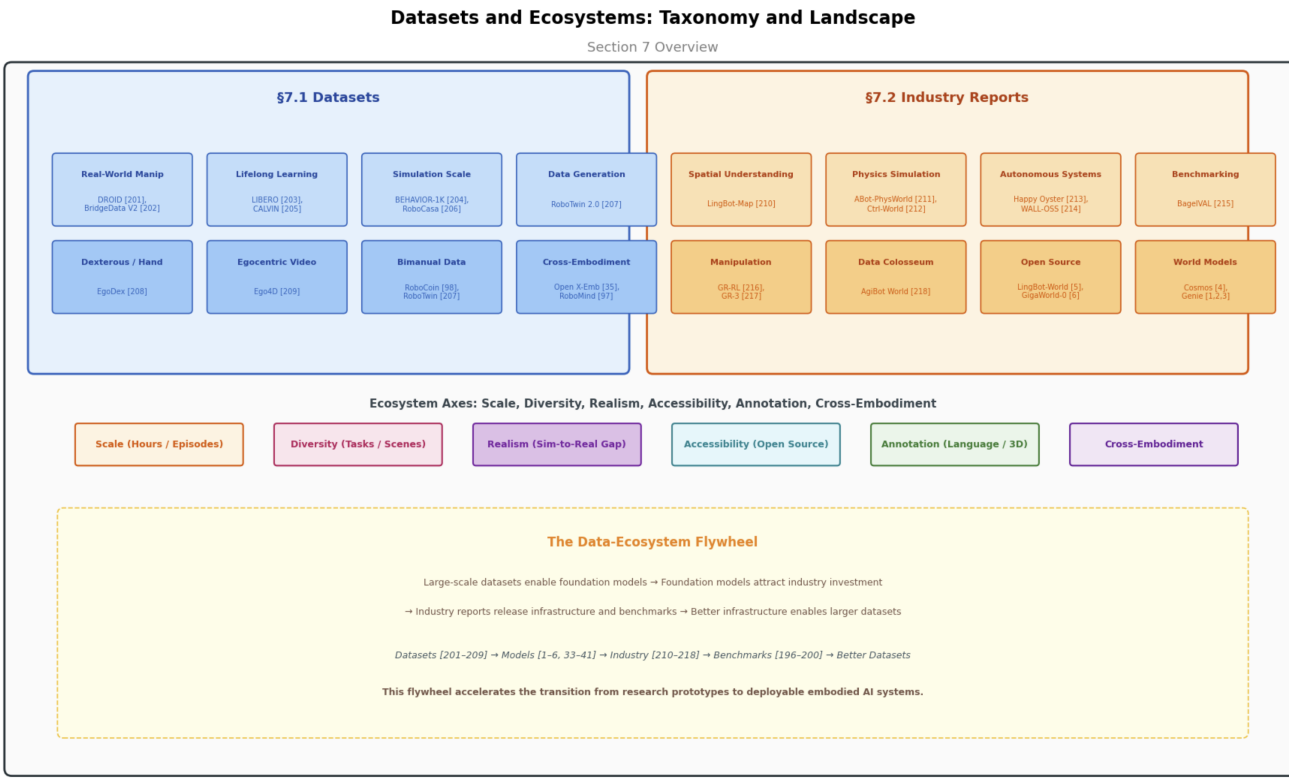

Figure 15. Taxonomy of Datasets and Ecosystems. The landscape spans academic datasets (§7.1) and industry technical reports (§7.2).

7.1 Datasets

Robot learning datasets have evolved from small, single-task collections to massive, multi-embodiment corpora.

7.1.1 Real-World Manipulation at Scale

DROID [201] introduced a large-scale in-the-wild robot manipulation dataset collected across diverse real-world environments using a portable teleoperation setup (Figures 55, 56). **BridgeData V2** [202] provided a dataset for robot learning at scale in a single kitchen environment, with language-annotated trajectories.

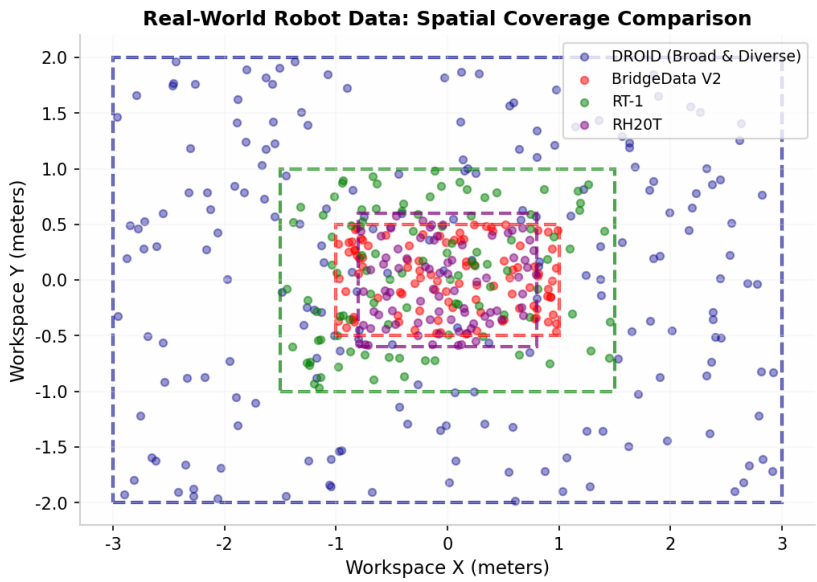

Figure 16. Spatial coverage comparison (Redrawn based on original concept.)

7.1.2 Lifelong and Long-Horizon Learning

LIBERO [202] introduced a benchmark for knowledge transfer in lifelong robot learning, with 130 language-conditioned tasks. **CALVIN** [204] provided a benchmark for language-conditioned long-horizon manipulation.

7.1.3 Large-Scale Simulation

BEHAVIOR-1K [204] introduced a human-centered embodied AI benchmark with 1,000 everyday activities. **RoboCasa** [206] provided large-scale simulation of everyday tasks with 120 kitchen scenes and 2,500+ objects.

7.1.4 Data Generation and Augmentation

RoboTwin 2.0 [207] introduced a scalable data generator with strong domain randomization for bimanual manipulation. **RoboTwin** and **RoboCoin** [98] provided open-sourced bimanual data collection.

7.1.5 Dexterous and Egocentric Data

EgoDex [208] from Apple introduced a large-scale dataset for learning dexterous manipulation from egocentric video, with 300K episodes and 3D skeleton annotations. **Ego4D** [209] provided 3,000 hours of egocentric video across diverse daily activities.

7.1.6 Cross-Embodiment and Multi-Robot Data

Open X-Embodiment [35] aggregated data from 22 robot types across institutions to train RT-X and subsequent open-source VLAs. **RoboMind** [97] introduced a benchmark on multi-embodiment intelligence normative data.

7.2 Industry Reports

Industry technical reports have released infrastructure, benchmarks, and preliminary models bridging research and production.

7.2.1 Spatial Understanding and Mapping

LingBot-Map [210] introduced a streaming 3D reconstruction model for real-time spatial understanding, achieving state-of-the-art accuracy on multiple benchmarks. This directly supports the LingBot-World [5] and LingBot-VLA [76] models.

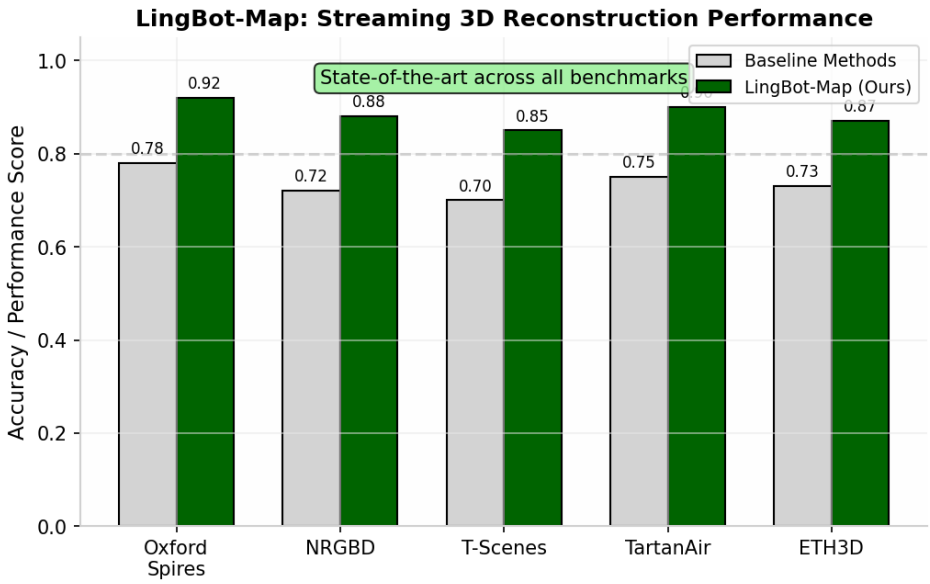

Figure 17. LingBot-Map reconstruction performance (Redrawn based on original concept.)

7.2.2 Physics Simulation Platforms

ABot-PhysWorld [211] introduced a physics-grounded world model for autonomous driving simulation. **Ctrl-World** [212] provided a controllable world model for embodied AI.

7.2.3 Autonomous Systems

Happy Oyster [213] introduced a world model for autonomous systems targeting logistics. **WALL-OSS** [214] proposed world-aligned latent learning for open-source simulation.

7.2.4 Benchmarking and Evaluation

BagelVAL [215] introduced benchmarking generalist vision-language-action models, emphasizing real-world operational metrics such as latency and task completion under time constraints.

7.2.5 Manipulation and Humanoid Robotics

GR-RL [216] described dexterous and precise long-horizon manipulation. **GR-3** [217] provided the technical report for Fourier's third-generation humanoid robot. **AgiBot World Colosseum** [218] provided a massive humanoid robot training dataset.

7.2.6 Open-Source Ecosystem

LingBot-World [5] and **GigaWorld-0** [6] released open-source world model platforms, while **Cosmos** [4] and **Genie** [1,2,3] have provided varying degrees of API access. These ecosystem investments recognize that the field matures fastest when data, models, and benchmarks are broadly accessible.

Section 7 summary. The datasets and ecosystems surveyed here constitute the often-overlooked infrastructure that enables the algorithmic advances documented throughout this survey. From real-world manipulation datasets to large-scale simulation and industry reports, this infrastructure creates a virtuous flywheel: larger datasets enable more capable foundation models, which attract industry investment, leading to better infrastructure and even larger datasets.

8. Conclusion and Future Directions

This survey has presented a systematic and comprehensive review of the modern **World Model** and **World Action Model** ecosystem, encompassing **200+ key papers** organized into a unified taxonomy. We began by tracing the evolution of **Foundation World Models** (§2)—from passive video predictors to interactive, memory-augmented simulators. We then examined **Vision-Language-Action Models** (§3), covering foundational architectures and domain-specific instantiations. **Embodied World Action Models** (§4) were identified as the natural convergence of these threads, unifying simulation with action generation. **Autonomous Driving World Models** (§5) demonstrated how these principles adapt to safety-critical control. Finally, we surveyed the supporting infrastructure of **Efficiency and Evaluation** (§6) and **Datasets and Ecosystems** (§7).

Despite this remarkable progress, the field stands at an inflection point. Below, we articulate **open challenges** and **future directions**, each framed as a concrete research hypothesis.

Open Challenge 1: Physical Consistency

Current world models often violate basic physics (e.g., object persistence, gravity). **Future direction:** Develop physics-aware training objectives, such as self-supervised discovery of physical parameters (mass, friction) and physics-based adversarial losses. Evaluate using extended world simulation benchmarks (e.g., WorldModelBench [197] covering momentum conservation).

Open Challenge 2: Cross-Embodiment Generalization

Even large VLAs fail on unseen robot morphologies. **Future direction:** Investigate embodiment-agnostic action representations—e.g., 3D keypoint flows [137] or relative joint offsets—and train on massively multi-embodiment data (Open X-Embodiment + synthetic variations). A concrete milestone: zero-shot transfer from a simulated arm to a real humanoid hand.

Open Challenge 3: Safety Verification for Generated Worlds

How can we provide formal safety guarantees for world model rollouts? **Future direction:** Combine occupancy world models (OccWorld [172]) with reachability analysis or differentiable logic. Develop certified world simulators where unsafe trajectories are provably non-existent under given assumptions.

Open Challenge 4: Closing the Sim-to-Real Evaluation Gap

WorldEval [198] and AutoEval [200] show promise, but correlation with real-world deployment remains imperfect. **Future direction:** Build open-source, continuous evaluation platforms that run policies simultaneously in physical labs and multiple world simulators, measuring rank correlation over thousands of trials. Release standardized sim-to-real correlation metrics.

Open Challenge 5: Autonomous Data Collection for Long-Tail Scenarios

Scarce data for rare events (e.g., accidents, novel objects) limits model robustness. **Future direction:** Deploy world model-driven exploration policies that actively seek out high-uncertainty states or safety-critical scenarios, then use those trajectories for fine-tuning. This creates a self-improving data flywheel.

Open Challenge 6: Standardized Open Ecosystem

Fragmentation across codebases, benchmarks, and model formats slows progress. **Future direction:** Establish community-driven model zoos (akin to Hugging Face for world models) with unified APIs for inference, fine-tuning, and evaluation. Support reproducible leaderboards for physical plausibility and policy transfer.

We hope this survey serves as a definitive reference and inspires the next generation of embodied intelligence—where agents learn to act by reliably imagining their futures.

References

[1] J. B. Alayrac et al., "Genie: Generative interactive environments," arXiv preprint arXiv:2402.15391, 2024. (Google DeepMind)

[2] Google DeepMind, "Genie 2: A large-scale foundation world model," DeepMind Blog, 2024.

[3] Google DeepMind, "Genie 3," deepmind.google, 2025.

[4] NVIDIA et al., "Cosmos world foundation model platform for physical AI," arXiv preprint arXiv:2501.03575, 2025.

[5] Robbyant Team, Ant Group, "LingBot-World: Advancing open-source world models for embodied intelligence," arXiv preprint arXiv:2601.20540, 2026.

[6] GigaWorld Team, "GigaWorld-0: World models as data engine to empower embodied AI," arXiv preprint arXiv:2511.19861, 2025.

[7] T. Brooks et al., "Video generation models as world simulators," OpenAI Technical Report, 2024. (Sora)

[8] A. Bar et al., "Navigation world models," CVPR, pp. 15791–15801, 2025.

[9] X. Mao et al., "Yume: An interactive world generation model," arXiv preprint arXiv:2507.17744, 2025.

[10] X. Mao et al., "Yume-1.5: A text-controlled interactive world generation model," arXiv preprint arXiv:2512.22096, 2025.

[11] Z. Xiao et al., "WorldMem: Long-term consistent world simulation with memory," arXiv preprint arXiv:2504.12369, 2025.

[12] T. Wu et al., "Video world models with long-term spatial memory," arXiv preprint arXiv:2506.05284, 2025.

[13] R. Li et al., "VMem: Consistent interactive video scene generation with surfel-indexed view memory," ICCV, pp. 25690–25699, 2025.

[14] W. Sun et al., "WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling," arXiv preprint arXiv:2512.14614, 2025.

[15] Y. Hong et al., "Relic: Interactive video world model with long-horizon memory," arXiv preprint arXiv:2512.04040, 2025.

[16] J. Gao et al., "LongVie 2: Multimodal controllable ultra-long video world model," arXiv preprint arXiv:2512.13604, 2025.

[17] X. Ren et al., "Gen3C: 3D-informed world-consistent video generation with precise camera control," CVPR, pp. 6121–6132, 2025.

[18] G. Li et al., "MagicWorld: Interactive geometry-driven video world exploration," arXiv preprint arXiv:2511.18886, 2025.

[19] Y. Kang et al., "How far is video generation from world model: A physical law perspective," arXiv preprint arXiv:2411.02385, 2024.

[20] M. Mei et al., "Video generation models in robotics—applications, research challenges, future directions," arXiv preprint arXiv:2601.07823, 2026.

[21] Decart Team et al., "Oasis: A universe in a transformer," arXiv preprint arXiv:2411.19147, 2024.

[22] J. Guo et al., "MineWorld: A real-time and open-source interactive world model on Minecraft," arXiv preprint arXiv:2504.08388, 2025.

[23] X. He et al., "Matrix-Game 2.0: An open-source real-time and streaming interactive world model," arXiv preprint arXiv:2508.13009, 2025.

[24] X. He et al., "Matrix-Game 3.0: Real-time and streaming interactive world model with long-horizon memory," arXiv preprint arXiv:2604.08995, 2026.

[25] H. Che et al., "GameGen-X: Interactive open-world game video generation," arXiv preprint arXiv:2411.00769, 2024.

[26] J. Yu et al., "GameFactory: Creating new games with generative interactive videos," arXiv preprint arXiv:2501.08325, 2025.

[27] J. Chen et al., "DeepVerse: 4D autoregressive video generation as a world model," arXiv preprint arXiv:2506.01103, 2025.

[28] J. Huang et al., "Memory forcing: Spatio-temporal memory for consistent scene generation on Minecraft," arXiv preprint arXiv:2510.03198, 2025.

[29] W. Tan et al., "Lumine: An open recipe for building generalist agents in 3D open worlds," arXiv preprint arXiv:2511.08892, 2025.

[30] Z. Xie et al., "ShareVerse: Multi-agent consistent video generation for shared world modeling," arXiv preprint arXiv:2603.02697, 2026.

[31] J. Tang et al., "Hunyuan-gamecraft-2: Instruction-following interactive game world model," arXiv preprint arXiv:2511.23429, 2025.

[32] J. Li et al., "Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition," arXiv preprint arXiv:2506.17201, 2025.

[33] A. Brohan et al., "RT-1: Robotics transformer for real-world control at scale," RSS, 2022.

[34] A. Brohan et al., "RT-2: Vision-language-action models transfer web knowledge to robotic control," CoRL, PMLR, 2023.

[35] A. O'Neill et al., "Open X-Embodiment: Robotic learning datasets and RT-X models," ICRA, pp. 6892–6903, 2024.

[36] M. J. Kim et al., "OpenVLA: An open-source vision-language-action model," CoRL 2024, PMLR vol. 270, pp. 2679–2713, 2024.

[37] K. Black et al., "π_0: A vision-language-action flow model for general robot control," arXiv preprint arXiv:2410.24164, 2024.

[38] K. Black et al., "$\pi_{0.5}$: A vision-language-action model with open-world generalization," arXiv preprint arXiv:2504.16054, 2025.

[39] Q. Li et al., "CogAct: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation," arXiv preprint arXiv:2411.19650, 2024.

[40] J. Bjorck et al., "GR00T N1: An open foundation model for generalist humanoid robots," arXiv preprint arXiv:2503.14734, 2025.

[41] Octo Model Team et al., "Octo: An open-source generalist robot policy," arXiv preprint arXiv:2405.12213, 2024.

[42] S. Belkhale and D. Sadigh, "MiniVLA: A better VLA with a smaller footprint," Stanford ILIAD, 2024.

[43] J. Cao et al., "FastDriveVLA: Efficient end-to-end driving via plug-and-play reconstruction-based token pruning," arXiv preprint arXiv:2507.23318, 2025.

[44] X. Zhou et al., "OpenDriveVLA: Towards end-to-end autonomous driving with large vision language action model," arXiv preprint arXiv:2503.23463, 2025.

[45] Z. Zhou et al., "AutoVLA: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning," arXiv preprint arXiv:2506.13757, 2025.

[46] Z. Yuan et al., "AutoDrive-R²: Incentivizing reasoning and self-reflection capacity for VLA model in autonomous driving," arXiv preprint arXiv:2509.01944, 2025.

[47] F. Lin et al., "OneTwoVLA: A unified vision-language-action model with adaptive reasoning," arXiv preprint arXiv:2505.11917, 2025.

[48] Z. Yang et al., "DriveMoE: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving," arXiv preprint arXiv:2505.16278, 2025.

[49] W. Zheng et al., "DriveAgent-RL: Advancing VLM-based autonomous driving with active perception and hybrid thinking," arXiv preprint arXiv:2507.20879, 2025.

[50] Z. Wang et al., "CogAD: Cognitive-hierarchy guided end-to-end autonomous driving," arXiv preprint arXiv:2505.21581, 2025.

[51] Y. Xie et al., "S4-Driver: Scalable self-supervised driving multimodal large language model with spatio-temporal visual representation," CVPR, 2025.

[52] Z. Xu et al., "DriveVPT4: Interpretable end-to-end autonomous driving via large language model," IEEE Robotics and Automation Letters, 2024.

[53] J. Yuan et al., "RAG-Driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model," arXiv preprint arXiv:2402.10828, 2024.

[54] T. Yan et al., "RLGf: Reinforcement learning with geometric feedback for autonomous driving video generation," arXiv preprint arXiv:2509.16500, 2025.

[55] R. Zhang et al., "AdaDrive: Self-adaptive slow-fast system for language-grounded autonomous driving," ICCV, 2025.

[56] R. Zhang et al., "VDrive: Vision-augmented lightweight MLLMs for efficient language-grounded autonomous driving," ICCV, 2025.

[57] J. Zhang et al., "SafeAuto: Knowledge-enhanced safe autonomous driving with multimodal foundation models," arXiv preprint arXiv:2503.00211, 2025.

[58] H. Chi et al., "CoVLA: Comprehensive vision-language-action dataset for autonomous driving," WACV, 2025.

[59] H. Chi et al., "Impromptu VLA: Open weights and open data for driving vision-language-action models," arXiv preprint arXiv:2505.23757, 2025.

[60] S. Zeng et al., "FutureSightDrive: Thinking visually with spatio-temporal CoT for autonomous driving," arXiv preprint arXiv:2505.17685, 2025.

[61] Z. Zhang et al., "OmniDrive-R1: Reinforcement-driven interleaved multi-modal chain-of-thought for trustworthy vision-language autonomous driving," arXiv preprint arXiv:2512.14044, 2025.

[62] X. Tian et al., "DriveVlm: The convergence of autonomous driving and large vision-language models," arXiv preprint arXiv:2402.12289, 2024.

[63] J. Wen et al., "Diffusion-VLA: Scaling robot foundation models via unified diffusion and autoregression," arXiv preprint arXiv:2412.03293, 2024.

[64] J. Liu et al., "HybridVLA: Collaborative diffusion and autoregression in a unified vision-language-action model," arXiv preprint arXiv:2503.10631, 2025.

[65] H. Zhen et al., "3D-VLA: A 3D vision-language-action generative world model," arXiv preprint arXiv:2403.09631, 2024.

[66] Y. Ze et al., "3D Diffusion Policy: Generalizable visuomotor policy learning via simple 3D representations," RSS, 2024.

[67] C. Li et al., "PointVLA: Injecting the 3D world into vision-language-action models," arXiv preprint arXiv:2503.07511, 2025.

[68] H. Zhu et al., "SPA: 3D spatial-awareness enables effective embodied representation," arXiv preprint arXiv:2410.08208, 2024.

[69] J. Zhang et al., "UP-VLA: A unified understanding and prediction model for embodied agent," arXiv preprint arXiv:2501.18867, 2025.

[70] J. Zhang et al., "HIRT: Enhancing robotic control with hierarchical robot transformers," CoRL, 2024.

[71] R. Zheng et al., "TraceVLA: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies," arXiv preprint arXiv:2412.10345, 2024.

[72] Q. Zhao et al., "CoT-VLA: Visual chain-of-thought reasoning for vision-language-action models," CVPR, 2025.

[73] J. Zheng et al., "X-VLA: Soft-prompted transformer as scalable cross-embodiment vision-language-action model," ICLR, 2025.

[74] W. Zhao et al., "VLAS: Vision-language-action model with speech instructions for customized robot manipulation," ICLR, 2025.

[75] Z. Zhong et al., "FlowVLA: Visual chain of thought-based motion reasoning for vision-language-action models," arXiv preprint arXiv:2508.18269, 2025.

[76] Robbyant Team, Ant Group, "LingBot-VLA: A pragmatic VLA foundation model for real-world robot manipulation," arXiv preprint arXiv:2601.18692, 2026.

[77] Y. Yang et al., "ABot-M0: VLA foundation model for robotic manipulation with action manifold learning," arXiv preprint arXiv:2602.11236, 2026.

[78] T. Jiang et al., "Galaxea open-world dataset and G0 dual-system VLA model," arXiv preprint arXiv:2509.00576, 2025.

[79] Z. Li et al., "HAMSTER: Hierarchical action models for open-world robot manipulation," ICLR, 2025.

[80] H.-T. L. Chiang et al., "Mobility VLA: Multimodal instruction navigation with long-context VLMs and topological graphs," arXiv preprint arXiv:2407.07775, 2024.

[81] S. Fei et al., "LIBERO-Plus: In-depth robustness analysis of vision-language-action models," arXiv preprint arXiv:2510.13626, 2025.

[82] C.-Y. Hung et al., "Nora: A small open-sourced generalist vision language action model for embodied tasks," arXiv preprint arXiv:2504.19854, 2025.

[83] R. Xu et al., "A0: An affordance-aware hierarchical model for general robotic manipulation," arXiv preprint arXiv:2504.12636, 2025.

[84] S. Liu et al., "RDT-1B: A diffusion foundation model for bimanual manipulation," ICLR, 2025.

[85] S. Liu et al., "RDT2: Exploring the scaling limit of UMI data towards zero-shot cross-embodiment generalization," arXiv preprint arXiv:2602.03310, 2026.

[86] C. Chi et al., "Diffusion policy: Visuomotor policy learning via action diffusion," The International Journal of Robotics Research, vol. 44, no. 10-11, pp. 1684–1704, 2025.

[87] C. Chi et al., "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," arXiv preprint arXiv:2402.10329, 2024.

[88] J. Lee et al., "MolmoAct: Action reasoning models that can reason in space," arXiv preprint arXiv:2508.07917, 2025.

[89] D. Qu et al., "SpatialVLA: Exploring spatial representations for visual-language-action model," RSS, 2025.

[90] Q. Lv et al., "F1: A vision-language-action model bridging understanding and generation to actions," arXiv preprint arXiv:2509.06951, 2025.

[91] Y. Shen et al., "InstructVLA: Vision-language-action instruction tuning from understanding to manipulation," arXiv preprint arXiv:2507.17520, 2025.

[92] H. Li et al., "RoboInter: A holistic intermediate representation suite towards robotic manipulation," arXiv preprint arXiv:2602.09973, 2026.

[93] H. Li et al., "CronusVLA: Transferring latent motion across time for multi-frame prediction in manipulation," arXiv preprint arXiv:2506.19816, 2025.

[94] P. Ding et al., "Quar-VLA: Vision-language-action model for quadruped robots," ECCV, pp. 352–367, 2024.

[95] P. Ding et al., "Humanoid-VLA: Towards universal humanoid control with visual integration," arXiv preprint arXiv:2502.14795, 2025.

[96] Y. Wang et al., "VGG-T: Visual geometry grounded transformer," CVPR, 2025.

[97] K. Wu et al., "RoboMind: Benchmark on multi-embodiment intelligence normative data for robot manipulation," arXiv preprint arXiv:2412.13877, 2024.

[98] S. Wu et al., "RoboCoin: An open-sourced bimanual robotic data collection for integrated manipulation," arXiv preprint arXiv:2511.17441, 2025.

[99] C.-L. Cheang et al., "GR-2: A generative video-language-action model with web-scale knowledge for robot manipulation," arXiv preprint arXiv:2410.06158, 2024.

[100] K. F. Gbagbe et al., "Bi-VLA: Vision-language-action model-based system for bimanual robotic dexterous manipulations," IEEE SMC, pp. 2864–2869, 2024.

[101] C. Fan et al., "Interleave-VLA: Enhancing robot manipulation with interleaved image-text instructions," arXiv preprint arXiv:2505.02152, 2025.

[102] S. Ye et al., "World action models are zero-shot policies," arXiv preprint arXiv:2602.15922, 2026.

[103] A. Ye et al., "GigaWorld-Policy: An efficient action-centered world–action model," arXiv preprint arXiv:2603.17240, 2026.

[104] T. Yuan et al., "Fast-WAM: Do world action models need test-time future imagination?" arXiv preprint arXiv:2603.16666, 2026.

[105] GigaBrain Team, "GigaBrain-0: A world model-powered vision-language-action model," arXiv preprint arXiv:2510.19430, 2025.

[106] GigaBrain Team, "GigaBrain-0.5 M*: A VLA that learns from world model-based reinforcement learning," arXiv preprint arXiv:2602.12099, 2026.

[107] C. Zhu et al., "Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets," arXiv preprint arXiv:2504.02792, 2025.

[108] S. Li et al., "Unified video action model," arXiv preprint arXiv:2503.00200, 2025.

[109] Y. Shen et al., "VideoVLA: Video generators can be generalizable robot manipulators," arXiv preprint arXiv:2512.06963, 2025.

[110] J. Cen et al., "WorldVLA: Towards autoregressive action world model," arXiv preprint arXiv:2506.21539, 2025.

[111] C. Li et al., "Robotic World Model: A neural network simulator for robust policy optimization in robotics," arXiv preprint arXiv:2501.10100, 2025.

[112] Y. Feng et al., "Vidar: Embodied video diffusion model for generalist manipulation," arXiv preprint arXiv:2507.12898, 2025.

[113] H. Li et al., "MimicDreamer: Aligning human and robot demonstrations for scalable VLA training," arXiv preprint arXiv:2509.22199, 2025.

[114] H. Zhang et al., "GEVRM: Goal-expressive video generation model for robust visual manipulation," ICLR, 2025.

[115] Z. Dong et al., "EMMA: Generalizing real-world robot manipulation via generative visual transfer," arXiv preprint arXiv:2509.22407, 2025.

[116] Y. Hu et al., "Video prediction policy: A generalist robot policy with predictive visual representations," arXiv preprint arXiv:2412.14803, 2024.

[117] J. Liang et al., "Video generators are robot policies," arXiv preprint arXiv:2508.00795, 2025.

[118] Y. Liao et al., "Genie Envisioner: A unified world foundation platform for robotic manipulation," arXiv preprint arXiv:2508.05635, 2025.

[119] Aether Team, "Aether: Geometric-aware unified world modeling," arXiv preprint arXiv:2503.18945, 2025.

[120] Y. Shang et al., "RoboScape: Physics-informed embodied world model," arXiv preprint arXiv:2506.23135, 2025.

[121] X. Liu et al., "World-VLA-Loop: Closed-loop learning of video world model and VLA policy," arXiv preprint arXiv:2602.06508, 2026.

[122] L. Li et al., "Causal world modeling for robot control," arXiv preprint arXiv:2601.21998, 2026.

[123] H. Li et al., "VLA-RFT: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators," arXiv preprint arXiv:2510.00406, 2025.

[124] Y. Wen et al., "VidMan: Exploiting implicit dynamics from video diffusion model for effective robot manipulation," NeurIPS, vol. 37, pp. 41051–41075, 2024.

[125] H. Wu et al., "Unleashing large-scale video generative pre-training for visual robot manipulation," arXiv preprint arXiv:2312.13139, 2023.

[126] S. Ye et al., "MoWM: Mixture-of-world-models for embodied planning via latent-to-pixel feature modulation," arXiv preprint arXiv:2509.21797, 2026.

[127] J. Yu et al., "Cosmos-Policy: World model-based policy optimization for vision-language-action models," arXiv preprint arXiv:2511.09515, 2025.

[128] M. J. Kim et al., "Cosmos Policy: Fine-tuning video models for visuomotor control and planning," arXiv preprint arXiv:2601.16163, 2026.

[129] MotuBrain Team, "MotuBrain: An advanced world action model for robot manipulation," arXiv preprint arXiv:2604.27792, 2026.

[130] C. Huang et al., "ThinkAct: Vision-language-action reasoning via reinforced visual latent planning," arXiv preprint arXiv:2507.16815, 2025.

[131] J. Mao et al., "Robot learning from a physical world model," arXiv preprint arXiv:2511.07416, 2025.

[132] H. Li et al., "SimpleVLA-RL: Scaling VLA training via reinforcement learning," arXiv preprint arXiv:2509.09674, 2025.

[133] G. Lu et al., "VLA-RL: Towards masterful and general robotic manipulation with scalable reinforcement learning," arXiv preprint arXiv:2505.18719, 2025.

[134] S. Tan et al., "Interactive post-training for vision-language-action models," arXiv preprint arXiv:2505.17016, 2025.

[135] S. Ye et al., "Scalable robotic policy evaluation via discrete diffusion world model," arXiv preprint arXiv:2604.22152, 2026.

[136] Y. Ma et al., "Embodied Tree of Thoughts: Deliberate manipulation planning with embodied world model," arXiv preprint arXiv:2512.08188, 2025.

[137] H. Zhi et al., "3DFlowAction: Learning cross-embodiment manipulation from 3D flow world model," arXiv preprint arXiv:2506.06199, 2025.

[138] X. Wang et al., "DriveDreamer: Towards real-world-driven world models for autonomous driving," ECCV, Springer, pp. 55–72, 2024.

[139] X. Wang et al., "DriveDreamer-2: LLM-enhanced world models for diverse driving video generation," arXiv preprint arXiv:2403.06845, 2024.

[140] S. Gao et al., "Vista: A generalizable driving world model with high fidelity and versatile controllability," NeurIPS, vol. 37, pp. 91560–91596, 2024.

[141] X. Hu et al., "DrivingWorld: Constructing world model for autonomous driving via video GPT," arXiv preprint arXiv:2412.19505, 2024.

[142] A. Hu et al., "GAIA-1: A generative world model for autonomous driving," arXiv preprint arXiv:2309.17080, 2023.

[143] M. Hassan et al., "GEM: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control," CVPR, 2025.

[144] C. Min et al., "DriveWorld: 4D pre-trained scene understanding via world models for autonomous driving," CVPR, pp. 15522–15533, 2024.

[145] Y. Zhang et al., "BEVWorld: A multimodal world simulator for autonomous driving via scene-level BEV latents," arXiv preprint arXiv:2407.05679, 2024.

[146] W. Wu et al., "DriveScape: High-resolution driving video generation by multi-view feature fusion," Technical Report, 2024.

[147] Q. Li et al., "Think2Drive: Efficient reinforcement learning by thinking in latent world model for quasi-realistic autonomous driving," ECCV, 2024.

[148] X. Wang et al., "LongDWM: Cross-granularity distillation for building a long-term driving world model," arXiv preprint arXiv:2506.01546, 2025.

[149] G. Zhao et al., "DriveDreamer4D: World models are effective data machines for 4D driving scene representation," arXiv preprint arXiv:2410.13571, 2024.

[150] C. Zheng et al., "X-World: Controllable ego-centric multi-camera world models for scalable end-to-end driving," arXiv preprint arXiv:2603.19979, 2026.

[151] K. Zhang et al., "Epona: Autoregressive diffusion world model for autonomous driving," ICCV, 2025.

[152] R. Gao et al., "MagicDrive-v2: High-resolution long video generation for autonomous driving with adaptive control," arXiv preprint arXiv:2411.13807, 2024.

[153] R. Gao et al., "MagicDrive3D: Controllable 3D generation for any view rendering in street scenes," arXiv preprint arXiv:2405.14475, 2024.

[154] Y. Wen et al., "Panacea: Panoramic and controllable video generation for autonomous driving," CVPR, 2024.

[155] K. Yang et al., "BEVControl: Accurately controlling street-view elements with multi-perspective consistency via BEV sketch layout," arXiv preprint arXiv:2308.01661, 2023.

[156] H. Arai et al., "ACT-Bench: Towards action controllable world models for autonomous driving," arXiv preprint arXiv:2412.05337, 2024.

[157] Y. Zheng et al., "World4Drive: End-to-end autonomous driving via intention-aware physical latent world model," ICCV, 2025.

[158] A. Mousakhan et al., "Orbis: Overcoming challenges of long-horizon prediction in driving world models," arXiv preprint arXiv:2507.13162, 2025.

[159] Y. Li et al., "Enhancing end-to-end autonomous driving with latent world model," arXiv preprint arXiv:2406.08481, 2024.

[160] H. Lin et al., "FutureX: Enhance end-to-end autonomous driving via latent chain-of-thought world model," arXiv preprint arXiv:2512.11226, 2025.

[161] Y. Zheng et al., "DriVerse: Navigation world model for driving simulation via multimodal trajectory prompting and motion alignment," arXiv preprint arXiv:2504.18576, 2025.

[162] Y. Zhou et al., "DriveDreamer-Policy: A geometry-grounded world–action model for unified generation and planning," arXiv preprint arXiv:2604.01765, 2026.

[163] J. Yang et al., "Generalized predictive model for autonomous driving," CVPR, pp. 14662–14672, 2024.

[164] Y. Wang et al., "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," CVPR, pp. 17470–17479, 2024.

[165] W. Zheng et al., "GenAD: Generative end-to-end autonomous driving," ECCV, Springer, pp. 87–104, 2024.

[166] W. Zheng et al., "DOE-1: Closed-loop autonomous driving with large world model," arXiv preprint arXiv:2412.09627, 2024.

[167] S. Gao et al., "AdaWorld: Learning adaptable world models with latent actions," ICML, 2025.

[168] S. Hamdan and F. Güney, "CarFormer: Self-driving with learned object-centric representations," ECCV, Springer, pp. 177–193, 2024.

[169] J. Yang et al., "ReSim: Reliable world simulation for autonomous driving," arXiv preprint arXiv:2506.09981, 2025.

[170] X. Yang et al., "DriveArena: A closed-loop generative simulation platform for autonomous driving," CVPR, pp. 26933–26943, 2025.

[171] A. Nachkov et al., "Dream to drive: Model-based vehicle control using analytic world models," arXiv preprint arXiv:2502.10012v1, 2025.

[172] W. Zheng et al., "OccWorld: Learning a 3D occupancy world model for autonomous driving," ECCV, pp. 55–72, 2024.

[173] Y. Zhang et al., "UnO: Unsupervised occupancy fields for perception and forecasting," CVPR, pp. 14487–14496, 2024.

[174] X. Zhou et al., "HERMES: A unified self-driving world model for simultaneous 3D scene understanding and generation," arXiv preprint arXiv:2501.14729, 2025.

[175] NVIDIA et al., "Cosmos-Transfer1: Conditional world generation with adaptive multimodal control," arXiv preprint arXiv:2503.14492, 2025.

[176] C. Dang et al., "SparseWorld: A flexible, adaptive, and efficient 4D occupancy world model powered by sparse and dynamic queries," AAAI, 2026.

[177] T. Deng et al., "GaussianDWM: 3D Gaussian driving world model for unified scene understanding and multi-modal generation," arXiv preprint arXiv:2512.23180, 2025.

[178] C. Diehl et al., "DIO: Decomposable implicit 4D occupancy-flow world model," CVPR, pp. 27456–27466, 2025.

[179] E. Bogdoll et al., "MUVO: A multimodal generative world model for autonomous driving with geometric representations," IEEE IV, pp. 2243–2250, 2025.

[180] H. Bian et al., "DynamicCity: Large-scale 4D occupancy generation from dynamic scenes," ICLR, 2025.

[181] K. Pertsch et al., "FAST: Efficient action tokenization for vision-language-action models," arXiv preprint arXiv:2501.09747, 2025.

[182] S. Xu et al., "VLA-Cache: Towards efficient vision-language-action model via adaptive token caching in robotic manipulation," arXiv preprint arXiv:2502.02175, 2025.

[183] Y. Yang et al., "EfficientVLA: Training-free acceleration and compression for vision-language-action models," arXiv preprint arXiv:2506.10100, 2025.

[184] Y. Yue et al., "Deer-VLA: Dynamic inference of multimodal large language models for efficient robot execution," NeurIPS, vol. 37, pp. 56619–56643, 2024.

[185] R. Zhang et al., "MoLe-VLA: Dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation," AAAI, 2026.

[186] J. Wen et al., "TinyVLA: Towards fast, data-efficient vision-language-action models for robotic manipulation," IEEE Robotics and Automation Letters, 2025.

[187] Y. Wang et al., "VLA-Adapter: An effective paradigm for tiny-scale vision-language-action model," AAAI, 2026.

[188] Z. Liang et al., "Discrete diffusion VLA: Bringing discrete diffusion to action decoding in vision-language-action policies," arXiv preprint arXiv:2508.20072, 2025.

[189] S. Tan et al., "Interactive post-training for vision-language-action models," arXiv preprint arXiv:2505.17016, 2025.

[190] M. Reuss et al., "Flower: Democratizing generalist robot policies with efficient vision-language-action flow policies," arXiv preprint arXiv:2509.04996, 2025.

[191] S. R. Alvar et al., "DivPrune: Diversity-based visual token pruning for large multimodal models," CVPR, pp. 9392–9401, 2025.

[192] Y. Yang et al., "DySL-VLA: Efficient vision-language-action model inference via dynamic-static layer-skipping for robot manipulation," arXiv preprint arXiv:2602.22896, 2026.

[193] Z. Yang et al., "KEEP: A KV-cache-centric memory management system for efficient embodied planning," arXiv preprint arXiv:2602.23592, 2026.

[194] J. Zhang et al., "KERV: Kinematic-rectified speculative decoding for embodied VLA models," arXiv preprint arXiv:2603.01581, 2026.

[195] Y. Li et al., "SP-VLA: A joint model scheduling and token pruning approach for VLA model acceleration," arXiv preprint arXiv:2506.12723, 2025.

[196] H. Duan et al., "WorldScore: A unified evaluation benchmark for world generation," arXiv preprint arXiv:2504.00983, 2025.

[197] D. Li et al., "WorldModelBench: Judging video generation models as world models," arXiv preprint arXiv:2502.20694, 2025.

[198] Y. Li et al., "WorldEval: World model as real-world robot policies evaluator," arXiv preprint arXiv:2505.19017, 2025.

[199] Y. Liao et al., "EWMBench: Embodied world model benchmark suite," (part of Genie Envisioner), 2025.

[200] Z. Zhou et al., "AutoEval: Autonomous evaluation of generalist robot manipulation policies in the real world," arXiv preprint arXiv:2503.24278, 2025.

[201] A. Khazatsky et al., "Droid: A large-scale in-the-wild robot manipulation dataset," arXiv preprint arXiv:2403.12945, 2024.

[202] H. Walke et al., "BridgeData V2: A dataset for robot learning at scale," CoRL, pp. 1723–1736, 2023.

[203] B. Liu et al., "LIBERO: Benchmarking knowledge transfer for lifelong robot learning," NeurIPS, vol. 36, pp. 44776–44791, 2023.

[204] C. Li et al., "BEHAVIOR-1K: A human-centered, embodied AI benchmark with 1,000 everyday activities and realistic simulation," arXiv preprint arXiv:2403.09227, 2024.

[205] O. Mees et al., "CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," IEEE Robotics and Automation Letters, vol. 7, no. 3, pp. 7327–7334, 2022.

[206] S. Nasiriany et al., "RoboCasa: Large-scale simulation of everyday tasks for generalist robots," arXiv preprint arXiv:2406.02523, 2024.

[207] T. Chen et al., "RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation," arXiv preprint arXiv:2506.18088, 2025.

[208] R. Hoque et al., "EgoDex: Learning dexterous manipulation from large-scale egocentric video," arXiv preprint arXiv:2505.11709, 2025.

[209] K. Grauman et al., "Ego4D: Around the world in 3,000 hours of egocentric video," CVPR, pp. 18995–19012, 2022.

[210] Robbyant Team, Ant Group, "LingBot-Map: A streaming 3D reconstruction model for real-time spatial understanding," arXiv preprint arXiv:2604.14141, 2026.

[211] ABot Team, Gaode/AMAP, "ABot-PhysWorld: Physics-grounded world model for autonomous driving simulation," Technical Report, 2025.

[212] Ctrl-World Team, "Ctrl-world: Controllable world model for embodied AI," Technical Report, 2025.

[213] Alibaba Team, "Happy Oyster: World model for autonomous systems," Alibaba Technical Report, 2025.

[214] WALL-OSS Team, "WALL-OSS: World-aligned latent learning for open-source simulation," arXiv preprint, 2025.

[215] BagelVAL Team, "BagelVAL: Benchmarking generalist vision-language-action models," arXiv preprint, 2025.

[216] Y. Li et al., "GR-RL: Going dexterous and precise for long-horizon robotic manipulation," arXiv preprint arXiv:2512.01801, 2025.

[217] C. Cheang et al., "GR-3 Technical Report," arXiv preprint arXiv:2507.15493, 2025.

[218] Y. Yunfei et al., "AgiBot World Colosseum," GitHub, 2024.

[219] Y. Ma et al., "A survey on vision-language-action models for embodied AI," arXiv preprint arXiv:2405.14093, 2024.

[220] Y. Guan et al., "World models for autonomous driving: An initial survey," IEEE Transactions on Intelligent Vehicles, 2024.

[221] T. Feng et al., "A survey of world models for autonomous driving," arXiv preprint arXiv:2501.11260, 2025.

[222] S. Tu et al., "The role of world models in shaping autonomous driving: A comprehensive survey," arXiv preprint arXiv:2502.10498, 2025.

[223] K. Xu et al., "From specialist to generalist: A comprehensive survey on world models," Authorea Preprints, 2026.

[224] X. Li et al., "A comprehensive survey on world models for embodied AI," arXiv preprint arXiv:2510.16732, 2025.

[225] J. Ding et al., "Understanding world or predicting future? A comprehensive survey of world models," ACM Computing Surveys, vol. 58, no. 3, pp. 1–38, 2025.

[226] D. Zhang et al., "Pure vision language action (VLA) models: A comprehensive survey," arXiv preprint arXiv:2509.19012, 2025.

[227] Y. Zhong et al., "A survey on vision-language-action models: An action tokenization perspective," arXiv preprint arXiv:2507.01925, 2025.

[228] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," NeurIPS, 2018.

[229] D. Hafner et al., "Dream to control: Learning behaviors by latent imagination," ICLR, 2020.

[230] D. Hafner et al., "DreamerV3: Mastering domains with world models," ICML, 2023.

[231] N. Hansen et al., "TD-MPC2: Scalable robust continuous control via world models and diffusion," ICLR, 2024.