World Model 2.0 Beyond Prediction: A Taxonomy and Review of Core Capabilities

YUJIE QIU, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, China

LEI FENG*, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, China

XUDONG WANG, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, China

JIACHENG WANG, College of Computing and Data Science, Nanyang Technological University, Singapore

ZHIXIANG YANG, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, China

ZEHUI XIONG, School of Electronics, Electrical Engineering and Computer Science (EEECS), Queen's University Belfast, United Kingdom

DUSIT NIYATO, College of Computing and Data Science, Nanyang Technological University, Singapore

World models have shown substantial potential for learning environmental dynamics and predicting future states. However, as environments become more complex and interactive, intelligent systems must make reliable decisions under uncertainty, anticipate alternative outcomes, and adapt as conditions change. This calls for world models that connect prediction with reasoning, action, and continual adaptation. Existing efforts remain fragmented and lack a systematic synthesis. To address this gap, this paper proposes world model 2.0 as a general cognitive architecture for complex interactive environments, develops a taxonomy of six core capabilities, and examines their integration within this architecture. It reviews representative methods and applications, analyzes their relationships, and summarizes key challenges and future research directions.

Additional Key Words and Phrases: World model 2.0, capability taxonomy, general cognitive architecture, challenges, future research directions.

*Corresponding author.

Authors' Contact Information: Yujie Qiu, qyj@bupt.edu.cn, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China; Lei Feng, fenglei@bupt.edu.cn, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China; Xudong Wang, xdwang@bupt.edu.cn, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China; Jiacheng Wang, jiacheng.wang@ntu.edu.sg, College of Computing and Data Science, Nanyang Technological University, Singapore, Singapore; Zhixiang Yang, yangzx@bupt.edu.cn, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China; Zehui Xiong, z.xiong@qub.ac.uk, School of Electronics, Electrical Engineering and Computer Science (EEECS), Queen's University Belfast, Belfast BT7 1NN, United Kingdom; Dusit Niyato, dniyato@ntu.edu.sg, College of Computing and Data Science, Nanyang Technological University, Singapore, Singapore.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

ACM Reference Format:
Yujie Qiu, Lei Feng, Xudong Wang, Jiacheng Wang, Zhixiang Yang, Zehui Xiong, and Dusit Niyato. 2026. World Model 2.0 Beyond Prediction: A Taxonomy and Review of Core Capabilities. *ACM Comput. Surv.* 1, 1 (November 2026), 35 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

1.1 Background

Artificial intelligence (AI) is driving a shift from data-driven automation toward systems capable of autonomous perception, understanding, and robust decision-making in complex real-world environments [54]. Its growing application across transportation, manufacturing, healthcare, and energy reflects the demand for systems that can handle complex tasks and collaborate with people and other intelligent systems [6].

Recent advances in foundation models, generative AI, and deep reinforcement learning have expanded AI capabilities in representation, generation, and sequential decision-making. Generative AI approaches include large language models (LLMs), vision-language models (VLMs), and diffusion models, with applications spanning communications, robotics, navigation, and resource allocation [96, 101, 146]. Foundation and generative models provide representation, prediction, and generation; model-free deep reinforcement learning learns policies through interaction, while model-based methods use learned dynamics for planning. However, integrating these complementary capabilities into systems that support causal reasoning, uncertainty-aware long-horizon prediction, physical consistency, and reliable decisions under changing or unseen conditions remains challenging.

World models learn internal representations of environmental dynamics and use action-conditioned transitions to predict future states, allowing agents to evaluate possible outcomes before acting. Their applications include robot control [99], autonomous driving [120], and embodied intelligence [155]. However, many existing models still rely primarily on statistical transition patterns, leaving causal mechanisms, calibrated multi-future prediction, and reliable adaptation under distribution shifts open challenges [13, 75, 79].

1.2 Motivation

Despite their applications in robot control, autonomous driving, and embodied intelligence, existing world models generally learn environmental transitions from historical interaction data and use them to predict future states. Their limitations become more evident in complex and changing settings: models may capture correlations without representing the mechanisms of state transitions, rely on a single or insufficiently calibrated set of future predictions, and separate prediction from decision-making. After deployment, limited adaptation further restricts their ability to respond to new environments, tasks, and multi-agent interactions [18, 60].

These limitations indicate that complex interactive environments require more from world models than accurate future-state prediction alone. A world model must also connect its predictions with decision-making and environmental interaction, so that predictions can inform actions and execution feedback can continually update the model. Motivated by this need, this paper introduces the concept of a next-generation world model, termed **world model 2.0**, and positions it as a general cognitive architecture for complex interactive environments. Its specific formulation, six capabilities, and the role of the general cognitive architecture in integrating them are developed in the following sections.

Before elaborating on this framework, it is important to position the present work within the existing survey literature on world models. The next section reviews related surveys, summarizes their scopes and emphases, and clarifies the positioning and contributions of this paper.

1.3 Related Surveys and Our Contributions

Several surveys examine the foundations and applications of world models. Ding et al. [18] identify state representation learning and future-state prediction as core functions, and review their applications in autonomous driving, robotics, and social simulation. Zhu et al. [153] review general world models and their development in video generation, autonomous driving, and autonomous agents. Xu et al. [116] survey applications across tasks and scenarios, with a focus on challenges in transferring learned representations and predictive capabilities to new domains. Domain-specific surveys further examine world models in particular tasks. For autonomous driving, Guan et al. [31] organize the literature around perception, memory, and control, covering environment representation, future-scene prediction, and their use in decision-making. Zhao et al. [142] review world models in cognitive agents and future networks, including their architectures, training, and applications in prediction, planning, and causal reasoning. Another survey considers world models alongside large language models and agentic AI in edge general intelligence [141]. In embodied intelligence, Wang et al. [105] and Shen et al. [90] survey World Action Models (WAMs), which combine environmental-state modeling with action generation. Gao et al. [25] review self-evolving agents and discuss how their world models can adapt over time, complementing the WAM literature with a focus on continual learning.

Overall, existing surveys cover the foundations, structures, and applications of world models, as well as emerging directions such as cross-domain generalization, action generation, and self-evolution. However, a systematic account of the core capabilities of next-generation world models and their relationships remains lacking. This paper addresses this gap by organizing the literature around six capabilities of world model 2.0 and the general cognitive architecture that integrates them. The specific contributions are as follows:

- We review the fundamental concepts, components, and typical workflows of world models, and synthesize relevant research to trace the field's development from environmental modeling and state prediction toward increasingly complex interaction tasks and integrated cognitive demands.
- Building on this trajectory, we introduce the concept of a next-generation world model, which we term world model 2.0. Centered on causal understanding, it connects environmental representation, dynamics prediction, and action decision-making within a unified architecture. It uses multiple possible future trajectories to inform action selection and incorporates execution feedback to recalibrate predictions, update memory, and adapt to environmental changes, forming a closed-loop system spanning perception, prediction, reasoning, and action.
- We present a taxonomy of six core capabilities and discuss the general cognitive architecture that integrates them. We organize the literature around the capabilities' definitions, representative technical realizations, and use cases, and synthesize how they are connected through the general cognitive architecture.
- Drawing on representative studies and use cases across capabilities, we cover settings including robot control, autonomous driving, embodied intelligence, language agents, wireless network control, and interactive simulation. We synthesize how world models are applied to different tasks and the findings reported in these studies. We also identify world models for further investigation.

1.4 Organization

The remainder of this paper is organized as follows. Section 2 reviews the components and workflow of traditional world models, defines world model 2.0, and presents the taxonomy and relationships among its six functional capabilities and integrative architecture. Section 3 examines each capability through its definition, representative technical realizations, and use cases. Section 4 discusses challenges and future directions for each dimension, and Section 5 concludes the paper.

2 Overview and Taxonomy

2.1 Traditional World Models: Components and Workflow

Traditional world models typically consist of three core components: an encoder, a dynamics model, and a decoder [18]. The encoder compresses high-dimensional observations into compact latent states; the dynamics model learns state-transition patterns conditioned on latent states and actions; and the decoder maps latent states back to observable outputs. Taking an agent operating in a Markov decision process (MDP) as an example, a typical world-model workflow is as follows [141].

In an MDP, let s_t denote the environment state at time t, $a_t \in \mathcal{A}$ the action taken by the agent, and r_t the immediate reward. After the agent executes a_t in state s_t, the environment transitions to the next state s_{t+1}:

$$s_{t+1} = T(s_t, a_t), \tag{1}$$

where $T(\cdot)$ is the true state-transition function, represented as $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, which maps a state-action pair to the corresponding next state. The interaction also produces an immediate reward:

$$r_t = R(s_t, a_t), \tag{2}$$

where $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the true reward function that evaluates the immediate outcome of taking action a_t in state s_t. The observation o_t is a partial observation of the environment state or a high-dimensional representation of it.

In real-world applications, environments involve many changing factors, and the corresponding state and action spaces may be high-dimensional or continuous. Explicitly representing or exhaustively estimating T and R is therefore often intractable. A world model learns parameterized approximations $\hat{T}_\theta$ and $\hat{R}_\theta$ from interaction data [19, 28]:

$$s_{t+1} \approx \hat{T}_\theta(s_t, a_t), \tag{3}$$

$$\hat{R}_\theta(s_t, a_t) \approx r_t, \tag{4}$$

where θ denotes the learnable model parameters.

During training, the agent interacts with the environment and records observations o_t, actions a_t, and rewards r_t. The encoder compresses observations into latent states [89]:

$$h_t = f_{\text{enc}}(h_{t-1}, o_t, a_{t-1}), \tag{5}$$

where h_t is the latent representation at time t and summarizes historical information. The dynamics model predicts the next latent state from the current latent state and action:

$$h_{t+1} = f_{\text{dyn}}(h_t, a_t). \tag{6}$$

The decoder reconstructs the corresponding observation from the latent state:

$$\hat{o}_t = f_{\text{dec}}(h_t). \tag{7}$$

Training can minimize the state-prediction error $\|s_{t+1} - \hat{T}_\theta(s_t, a_t)\|$ or maximize the likelihood $P_\theta(o_{t+1}, r_t \mid h_t, a_t)$ over experience data, enabling the model to learn the environment dynamics [143].

After training, the world model can serve as a proxy simulator of the environment. Starting from the current latent state h_t, it recursively predicts subsequent latent states and rewards for candidate action sequences, generating imagined trajectories:

$$\hat{\tau} = (h_t, a_t, \hat{r}_t;\ h_{t+1}, a_{t+1}, \hat{r}_{t+1};\ \ldots). \tag{8}$$

The agent evaluates the cumulative returns of these imagined trajectories and selects an action sequence accordingly. Its objective is to maximize the predicted discounted cumulative return [81, 92]:

$$\max_{\{a_{t+k}\}_{k=0}^{H-1}} \mathbb{E}_{\hat{T}_\theta}\left[\sum_{k=0}^{H-1} \gamma^k \hat{r}_{t+k}\right], \tag{9}$$

where H is the planning horizon and $\gamma \in [0, 1]$ is the discount factor.

In this typical workflow, the world model primarily simulates environmental dynamics and predicts future states, while action selection is commonly handled by a decision module outside the model. Thus, although a world model can support model-based planning, it generally remains a predictive component; decision-making and execution feedback do not yet form a continual closed loop within the same architecture.

2.2 Motivation and Definition of World Model 2.0

Many existing world models learn statistical correlations from historical observations and use them to predict future states [84]. Although this paradigm provides a useful proxy for complex real-world environments, several challenges remain: 1) Their predictions may reflect correlations in historical data without reliably capturing the causal relationships between actions, environmental factors, and future states [60]; 2) Uncertainty may be represented by a single predicted future or a limited set of samples, restricting consideration of plausible alternatives [47]; 3) Prediction is often separated from decision-making, and execution feedback is not consistently used to update the model [35, 56]; 4) Although recent studies have explored causal representations, multi-future prediction, continual adaptation, and multi-agent modeling, these capabilities are often studied separately and have not yet been reliably integrated. Deployment-time updates and task-specific representations may also limit adaptation to changing environments and transfer across domains [18].

To address these limitations, this paper proposes world model 2.0 as a general cognitive architecture for complex interactive environments. Rather than focusing only on future-state prediction, it uses causal understanding to connect environmental representation, dynamics prediction, and action decision-making. It reasons over multiple possible futures to inform action selection and uses execution feedback to recalibrate predictions, update memory, and adapt to environmental changes, forming a system that spans perception, prediction, reasoning, and action through continued interaction.

Fig. 1 compares a conventional predictive world model with world model 2.0. The former uses an encoder, a dynamics model, and a decoder for action-conditioned prediction, with action selection handled externally [23, 31]. The latter connects causal dynamics and multiple future predictions with action selection and feedback-based updates [60, 80, 102, 107].

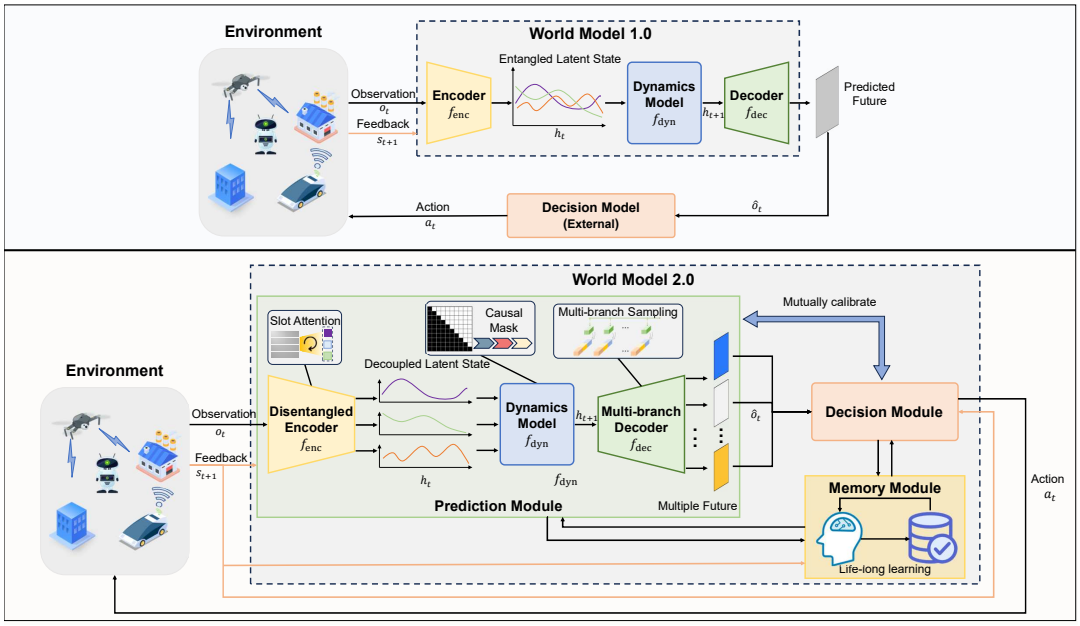


Fig. 1. Comparison between a conventional predictive world model and the proposed world model 2.0 architecture. The upper panel predicts an action-conditioned future from an entangled latent representation and relies on an external decision model for action selection. The lower panel combines disentangled and causally structured representations with multi-branch future prediction, mutual calibration between prediction and decision modules, execution feedback, knowledge memory, and lifelong learning to form a closed interaction loop. [20, 42, 60, 68, 78]

At time t, an attention-based encoder maps observation o_t to a structured representation [53]:

$$C_t = \mathrm{Enc}_{\phi}^{\mathrm{attn}}(o_t) = (Z_t, A_t), \tag{10}$$

where Z_t represents objects and other relevant factors, while A_t captures their learned dependencies. A causal mask guides information flow among these components [53, 60]. Using C_t to retrieve relevant experience from memory gives the context $\mathcal{M}_t^{\mathrm{ret}}$. The dynamics model then conditions on both the current representation and this retrieved context to generate multiple possible futures for each candidate action:

$$\hat{\mathbf{y}}^{(j,n)} \sim p_{\vartheta}\left(\cdot \mid C_t, \mathcal{M}_t^{\mathrm{ret}}, a_t^{(j)}\right), \tag{11}$$

where $\hat{\mathbf{y}}^{(j,n)}$ denotes the n-th predicted future branch for candidate action $a_t^{(j)}$. Multi-branch sampling and future-state decoding produce alternative outcomes [65, 148]. The decision module compares these outcomes in terms of task performance, costs, constraint violations, and risks, then selects an action for execution [100, 122].

After execution, observations are compared with predictions to recalibrate the model and assess the action outcome [72]. Prediction errors and execution outcomes inform model updates and are stored in memory. Relevant experience can then be retrieved alongside the current representation in subsequent predictions, closing the feedback loop [109, 112].

2.3 Capability Taxonomy

World model 2.0 can transform intelligent systems from passive predictors into proactive cognitive entities capable of long-horizon planning, autonomous decision-making, and lifelong self-evolution. The capabilities that world model 2.0 should possess include:

(1) **Causal understanding:** World model 2.0 uses causal understanding to represent the structures and mechanisms behind events across physical, semantic, and social dimensions [135]. Attention mechanisms can organize object- and factor-level representations [78], while autoregressive masks [60] or causal state-space models [87] can constrain temporal information flow. When appropriate identification assumptions are met, these mechanisms may support intervention or counterfactual reasoning and inform cross-domain generalization.
(2) **Multi-future reasoning:** World model 2.0 samples multiple branches over candidate actions [107] or latent states [12] to generate and evaluate plausible futures from a common initial state. This allows the system to consider alternatives before acting rather than relying only on a single prediction or trial and error [37].
(3) **Proactive decision-making:** A shared autoregressive architecture can connect prediction and decision-making: predicted future states inform the decision module, and execution feedback informs subsequent predictions [102, 105, 143]. Updating and calibrating both modules within a shared feature space forms a feedback loop from perception to action.
(4) **Open-ended and self-evolving learning:** World model 2.0 can update knowledge through accumulated and revised external memory [121, 137], and adapt model parameters online through methods such as parameter-efficient fine-tuning [77, 127]. Intervention and counterfactual reasoning may also help detect causal inconsistencies and revise internal structures as environments or tasks change [125].
(5) **Cross-domain generalization:** By representing relations between actions and state transitions, world model 2.0 can capture dynamics that transfer across environments [93, 110]. Depth regularization, dynamics-enhanced latent models, and cross-domain training can further help align transferable features such as geometry and motion while reducing dependence on scene-specific appearance [43, 52, 86].
(6) **Multi-agent coordination:** World model 2.0 extends modeling to interacting agents. Conditional denoising [140] or compositional reasoning [131] can structure joint states and actions, while identity embeddings and modular designs support cross-agent perspective alignment and knowledge reuse [108, 144].

The general cognitive architecture brings these six capabilities together through shared representations, memory, and interfaces [67, 73, 133, 153]. Compared with traditional world models, world model 2.0 connects prediction with causal understanding, proactive decision-making, continual adaptation, and cross-domain transfer.

2.4 Relationships among the Capabilities

Fig. 2 summarizes the relationships among the six capabilities. Causal understanding models action-conditioned state changes, multi-future reasoning compares alternative outcomes, and proactive decision-making uses these predictions to select actions [60, 102, 107]. Causal understanding also supports cross-domain transfer and agent-interaction modeling, while multi-future reasoning extends to joint actions and shared-state outcomes [43, 108]. Execution feedback updates dynamics and memory, supporting self-evolving learning and refining transferred models [121]. Shared representations, memory, and interfaces connect these capabilities within the general cognitive architecture [111, 152].

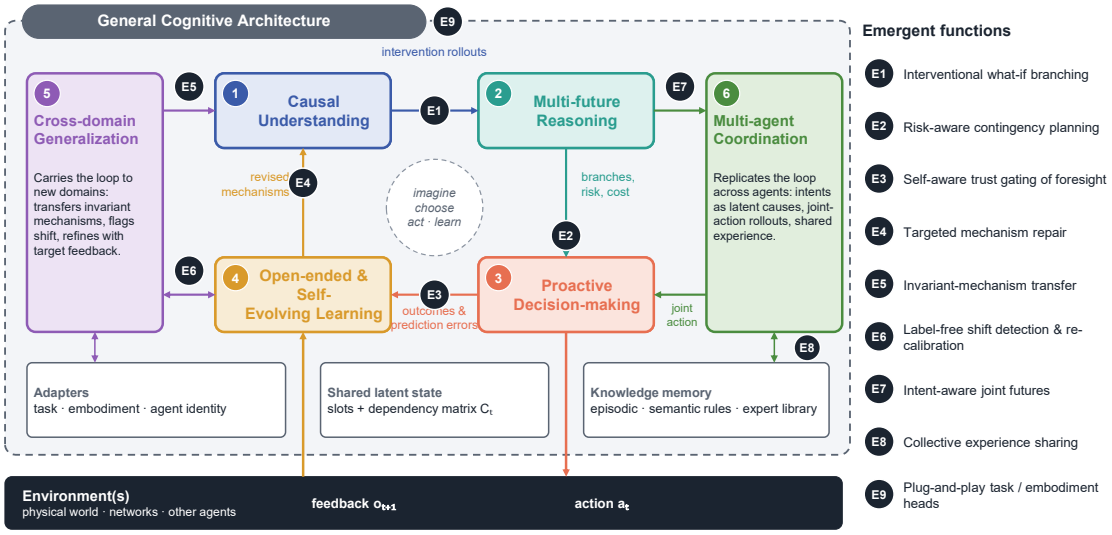


Fig. 2. An overview of the relationships among the six core capabilities of world model 2.0. Causal understanding supports multi-future reasoning through interventional what-if branching (E1), while multi-future reasoning supports proactive decision-making through risk-aware contingency planning (E2). Proactive decision-making and open-ended and self-evolving learning are coupled through execution outcomes and prediction errors, enabling self-aware confidence gating of foresight (E3), while the learning process performs targeted mechanism repair to revise causal mechanisms (E4). Cross-domain generalization transfers invariant mechanisms to causal understanding (E5) and connects with self-evolving learning through label-free shift detection and recalibration (E6). Multi-future reasoning extends to multi-agent coordination through intent-aware joint futures (E7), which support joint actions for proactive decision-making, while multi-agent coordination contributes collective experience through shared knowledge memory (E8). Shared latent states, knowledge memory, and adapters for tasks, embodiments, and agent identities connect these capabilities within the general cognitive architecture and enable plug-and-play task and embodiment heads (E9). [88, 93, 114, 144, 150]

3 Key Capabilities of World Model 2.0

Section 2 introduced six functional capabilities and the integrative architecture that connects them. This section examines these research dimensions through their definitions, representative technical realizations, and use cases, beginning with causal understanding.

3.1 Causal Understanding

3.1.1 *Foundational Concept.*

Traditional and general world models mainly learn statistical state transitions from observation sequences. They use past states and actions to predict the next state [18, 27, 102]. However, it does not explain why a state changes. Its predictions may also fail when the control policy, task distribution, environmental configuration, or interaction structure changes [4, 58]. Therefore, world model 2.0 no longer treats prediction accuracy as its sole objective. Instead, it uses causal understanding as a basic principle for organizing internal representations, modeling dynamics, and reasoning about decisions.

In complex interactive environments, causal reasoning aims to extend a model beyond describing what happened by supporting hypotheses about why it happened and predictions about how outcomes may change under alternative actions [3, 37].

Specifically, the model aims to identify candidate state variables with potential causal relevance from complex observations, estimate directed relations among these variables, and learn action-conditioned mechanisms that describe how the state changes [29, 63]. Claims about the causal interpretation of these variables or the stability of the learned mechanisms require suitable assumptions and supporting evidence. For example, when task performance degrades, the model should determine whether the cause is a change in physical conditions, an external disturbance, increased resource demand, unreliable sensing, or an inappropriate control decision[82]. Causal understanding capability enables the model to perform causal representation and attribution, intervention-effect analysis, and counterfactual inference [17, 29]. It provides structural support for multi-future reasoning, proactive decision-making, cross-domain generalization, and multi-agent coordination by representing how actions affect state transitions, environmental outcomes, and interactions among agents. Its causal knowledge can also be revised through intervention outcomes and feedback accumulated by open-ended and self-evolving learning.

3.1.2 Representative Technical Realization.

During continuous interaction with the environment, the model continues receiving new observations, each of which usually contains many coupled factors. In a complex real-world system, observations may jointly reflect physical conditions, external disturbances, resource demand, agent states, control policies, and historical actions. Traditional world models often compress this information into a unified latent state h_t and use it directly to predict the future. Although this holistic representation preserves the main statistical information in the observations, it mixes different objects and causal factors [115, 136]. The model therefore has difficulty identifying which entity caused a state change or which factor was first affected by a control action. Such entanglement at the representation level is a key reason why traditional world models struggle to develop causal understanding capability.

To address this issue, world model 2.0 first needs to separate observation features at the encoder [29, 63, 80]. This allows different objects or factors to have relatively independent representations, although the resulting slots are not necessarily identifiable causal variables. Slot-based self-attention provides one potential implementation [78]. Let o_t denote the raw observation at time t. The base encoder maps o_t to N element features $\{v_{t,i}\}_{i=1}^{N}$, where i is the element index. The model also defines K learnable slots $\{s_k\}_{k=1}^{K}$, where k is the slot index[1]. It then computes the assignment weights between the element features and the slots [48, 71]:

$$\alpha_{ik} = \mathrm{softmax}_k\left(\frac{(W_{\mathrm{Q}}s_k)^{\mathsf{T}}(W_{\mathrm{K}}v_{t,i})}{\sqrt{d_{\mathrm{att}}}}\right), \tag{12}$$

$$g_k = \sum_{i=1}^{N} \alpha_{ik} W_{\mathrm{V}} v_{t,i}, \tag{13}$$

$$z_t^k = \mathrm{GRU}(z_{t-1}^k, g_t^k), \tag{14}$$

where W_{Q}, W_{K}, and W_{V} are learnable query, key, and value projection matrices. d_{att} denotes the dimension of the query and key features. The weight α_{ik} measures how much element i is assigned to slot k, and $\mathrm{softmax}_k(\cdot)$ normalizes the weights over the slot index k. The vector g_k is the observation information collected by slot k. The operator $\mathrm{GRU}(\cdot)$ denotes a gated recurrent unit (GRU), and z_t^k is the resulting representation of object k. To preserve temporal causal consistency, each object representation is recurrently updated using only its previous state and the observation features available at the current time.

[1] N is selected according to the desired observation granularity and computational budget. The number of slots K is chosen using the expected number of objects or causal factors; unused slots can be masked when fewer objects are present.

To model pairwise dependencies among the object slots, the model treats them as structured latent variables and constructs a learned dependency matrix A_t. Its elements are estimated from pairwise slot interactions:

$$A_{t,jk} = \sigma\left(\frac{(W_Q^c z_t^k)^\top (W_K^c z_t^j)}{\sqrt{d_c}}\right), \tag{15}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, and $A_{t,jk}$ measures the learned dependency score from slot j on slot k under this parameterization, and W_Q^c and W_K^c are learnable projections. The structured latent representation is therefore defined as

$$C_t = (Z_t, A_t). \tag{16}$$

The dynamics model uses this matrix as a soft mask, allowing a target object to focus on slots that are potentially relevant to its dynamics. After obtaining the disentangled representations, causal understanding capability further requires the model to learn dynamics that describe how object states evolve as actions are applied [60, 126]. Regardless of the specific dynamics network, the model should combine the object representations, aggregated KPIs, and historical control variables into a state feature x_t at time t. It should also explicitly include task-relevant actions a_t, such as motion commands, manipulation operations, allocation decisions, or configuration updates, in the state transition. A general form is

$$\eta_{t+1} = f_\theta(\eta_t, a_t) + \omega_t, \tag{17}$$

$$x_t = h_\theta(\eta_t) + \varepsilon_t, \tag{18}$$

where η_t is the internal system state that cannot be fully observed. The functions f_θ and h_θ are the state transition function and the observation function, respectively, and θ denotes their model parameters. The terms ω_t and ε_t denote the process disturbance and observation noise. Because a_t directly enters the transition function, the model learns an action-conditioned state transition and predicts the next state given the supplied action. This formulation alone does not establish that the causal effect of an intervention on a_t has been identified; that interpretation requires suitable assumptions and supporting evidence. An alternative-action rollout is an action-conditioned prediction. It constitutes a counterfactual only when it refers to a particular observed case and specifies which factual evidence and latent disturbances are held fixed or inferred.

In this implementation, the encoder separates coupled observations into object-level representations, while the dynamics model explicitly learns action-conditioned state transitions. Therefore, they provide one way to implement causal capability in world model 2.0.

3.1.3 Use Cases.

The causal understanding capability of world model 2.0 models relationships among states, actions, and environmental changes, helping agents explain events and assess interventions. Representative studies are summarized in Table 1. Gkountouras et al. [29] learned causal variables from visual observations and text-described actions, using them to predict action effects and support long-horizon planning with a language agent. In robot control, Li et al. [60] used temporal causal attention to connect preceding actions with future visual states, improving long-horizon manipulation and adaptation to new configurations. Zhang et al. [139] used an action-conditioned transition graph to evaluate interventions under physical-parameter shifts, providing policy advice when predictions were reliable.

The same principle is being explored in self-evolving embodied intelligence and wireless network management. Yu et al. [125] proposed revising causal representations through interventions, counterfactual reasoning, and environmental

Table 1. Representative Studies on Causal Understanding in World Model 2.0.

Paper title	Domain	Main content	Main finding
Language Agents Meet Causality – Bridging LLMs and Causal World Models [29]	Embodied intelligence / language-agent planning	Learns disentangled causal variables from visual observations and language actions, and uses an action-conditioned transition model for planning.	Improves action-effect prediction and long-horizon planning, especially as the planning horizon increases.
Causal World Modeling for Robot Control [60]	Robot control	Jointly models video dynamics and robot actions with temporal causal conditioning and continuous real-world feedback.	Supports closed-loop control, improving long-horizon manipulation and adaptation to new configurations.
CausalNav: Reliability-Certified Causal World Models for Control under Physical-Parameter Shift [139]	Robot control	Uses an action-conditioned transition graph to evaluate interventions and gates causal advice according to its reliability.	Improves control safety by rejecting unreliable causal advice and falling back to the base controller.
Self-Evolving Cognitive Framework via Causal World Modeling for Embodied Scientific Intelligence [125]	Embodied intelligence	Proposes continual causal-model revision through interventions, counterfactual reasoning, and feedback.	Provides a conceptual basis for updating causal knowledge when environmental mechanisms change.
Agentic World Modeling for 6G: Near-Real-Time Generative State-Space Reasoning [87]	Wireless network control	Models physical resource blocks as causal control variables and generates KPI trajectories under alternative configurations.	Enables pre-deployment what-if analysis and policy screening with accurate and efficient near-real-time prediction.

feedback to adapt to changing mechanisms. Rezazadeh et al. [87] modeled how resource-block actions affect O-RAN performance indicators and generated KPI trajectories for alternative resource configurations. In these settings, causal understanding provides the mechanism-level basis for comparing alternative futures and selecting interventions, while feedback from executed interventions can further revise the learned causal structure. Together, these studies illustrate how causal models can help compare interventions and assess possible outcomes before deployment, particularly in dynamic settings with long-horizon objectives or safety constraints.

3.2 Multi-future Reasoning

3.2.1 Foundational Concept.

Traditional world models usually use historical states and actions to predict a single next state or the most likely future trajectory. Such point prediction is suitable for environments whose evolution is relatively deterministic [132]. However, when future states are jointly affected by random disturbances, unobserved factors, and external behavior, different outcomes are easily compressed into an average state [51, 114].

In world model 2.0, multi-future reasoning means that the model starts from the same historical conditions, builds a distributional representation of multiple possible subsequent states, and continuously rolls out each branch over time [42]. The model must generate distinct but plausible outcomes, preserve state continuity within each trajectory [5], and use the performance changes and risk information in the branches to support subsequent reasoning [147].

In complex interactive environments, even when the currently observed physical conditions, resource demand, and control configuration are identical, subsequent states may differ because of latent factors, stochastic disturbances, external behavior, or partial observability. Multi-future reasoning allows the model to retain normal, stressed, and

abnormal branches at the same time, providing a more complete view of the future for robust planning, resource management, and risk-constrained control [41].

In terms of reasoning depth, this capability has three progressive levels. The first generates multiple plausible futures, so the model no longer merges different outcomes into a single prediction. The second builds a coherent future distribution, so each branch follows stable temporal evolution and represents predictive uncertainty. The third evaluates each candidate action by aggregating performance, risk, and resource cost across its possible futures, thereby assessing how reliably it performs under the modeled uncertainty.

3.2.2 *Representative Technical Realization.*

In traditional world models, deterministic dynamics usually map the current latent state to a single next state. When trained to minimize average error, the model tends to output the mean of multiple future modes [113]. World model 2.0 instead needs to replace point mapping with a conditional distribution in latent-state space.

The model first organizes historical observations into a continuously updated belief state and introduces a stochastic latent variable into the state transition [22, 128]. Taking stochastic belief-state modeling as an example [12], and incorporating the action condition required for environment control, the core process can be written as

$$b_t = \mathcal{T}_\vartheta(b_{t-1}, \xi_{t-1}, a_{t-1}), \tag{19}$$

$$\xi_t \sim p_\vartheta(\xi_t \mid b_t), \tag{20}$$

$$y_t \sim p_\vartheta(y_t \mid b_t, \xi_t), \tag{21}$$

where b_t is the deterministic belief state that preserves historical information, $\mathcal{T}_\vartheta$ is the belief-state transition function, ξ_t is a stochastic latent variable sampled from the history-conditioned prior, and y_t is the predicted observable network state. The deterministic state maintains the continuity of system evolution, while different samples of ξ_t represent future changes that cannot be uniquely determined from the same history.

During training, the model uses encoded observation evidence e_t to construct the posterior $q_\varphi(\xi_t \mid b_t, e_t)$ and learn transition patterns that occur in the data. Reconstruction and Kullback-Leibler (KL) regularization constrain the posterior to remain close to the history-conditioned prior [9, 14, 97, 149]. During inference, the real future is unavailable, so the model samples only from the prior and can still generate plausible future branches.

After learning the distribution, the model can generate multiple futures from the same current belief state. When action selection is required, a planner first proposes candidate action sequences, and the world model samples several stochastic branches under each action. Let $\mathcal{A}_t = \{\mathbf{a}^{(j)}\}_{j=1}^{N_\mathrm{a}}$ be the candidate action set, where N_a is the number of candidate action sequences and $\mathbf{a}^{(j)} = (a_t^{(j)}, \ldots, a_{t+T_\mathrm{f}-1}^{(j)})$ is the j-th sequence. Let T_f be the future rollout length and N_s be the number of stochastic branches for each action. All branches share the current belief state $b_t^{(j,n)} = b_t$, and their initial latent variables are sampled from the history-conditioned prior, $\xi_t^{(j,n)} \sim p_\vartheta(\xi_t \mid b_t)$. The n-th future branch under action j can be written recursively as [37]

$$b_{t+d+1}^{(j,n)} = \mathcal{T}_\vartheta\left(b_{t+d}^{(j,n)}, \xi_{t+d}^{(j,n)}, a_{t+d}^{(j)}\right), \tag{22}$$

$$\xi_{t+d+1}^{(j,n)} \sim p_\vartheta\left(\xi_{t+d+1} \mid b_{t+d+1}^{(j,n)}\right), \tag{23}$$

$$\hat{y}_{t+d+1}^{(j,n)} \sim p_\vartheta\left(y_{t+d+1} \mid b_{t+d+1}^{(j,n)}, \xi_{t+d+1}^{(j,n)}\right), \tag{24}$$

$$\mathcal{Y}_t^{(j)} = \left\{\hat{\mathbf{y}}_{t+1:t+T_\mathrm{f}}^{(j,n)}\right\}_{n=1}^{N_\mathrm{s}}, \tag{25}$$

where j is the candidate-action index, n is the stochastic-branch index, and d is the rollout step relative to the current time. The variable $\hat{y}_{t+d+1}^{(j,n)}$ is the predicted state under action sequence j and latent sample n, while $\hat{\mathbf{y}}_{t+1:t+T_\mathrm{f}}^{(j,n)}$ is the corresponding complete trajectory. The set $\mathcal{Y}_t^{(j)}$ contains the future branches associated with $\mathbf{a}^{(j)}$. Thus, one candidate action can produce several distinct but temporally coherent futures.

After generating these action-conditioned futures, the model first evaluates each trajectory and then aggregates the scores across stochastic branches for each candidate action. Following the reward-based planning procedure in [107], the trajectory score and action selection can be written as

$$U^{(j,n)} = R_{\mathrm{per}}\left(\hat{\mathbf{y}}_{t+1:t+T_\mathrm{f}}^{(j,n)}\right) - \lambda_{\mathrm{con}} C_{\mathrm{con}}\left(\hat{\mathbf{y}}_{t+1:t+T_\mathrm{f}}^{(j,n)}\right), \tag{26}$$

$$\bar{U}^{(j)} = \frac{1}{N_\mathrm{s}} \sum_{n=1}^{N_\mathrm{s}} U^{(j,n)}, \tag{27}$$

$$j^\star = \arg\max_{1 \le j \le N_\mathrm{a}} \bar{U}^{(j)}, \tag{28}$$

$$a_t^\star = a_t^{(j^\star)}. \tag{29}$$

where $R_{\mathrm{per}}(\cdot)$ is the task-performance return, $C_{\mathrm{con}}(\cdot)$ is a nonnegative constraint-violation cost, λ_{con} controls the relative penalty assigned to constraint violations, and $U^{(j,n)}$ is the utility of the n-th future under action sequence j. The quantity $\bar{U}^{(j)}$ is the average utility over the stochastic branches generated for action sequence j. Since the branches are sampled from the predictive distribution, they are assigned equal weights in this Monte Carlo estimate. The planner selects the action sequence with the highest average utility; it does not select a particular stochastic outcome. Here, an additive penalty is used to distinguish task return from constraint-violation cost and to make their relative scale explicit through λ_{con}. This expected-utility criterion does not by itself guarantee safety if the predictive distribution is misspecified or fails to represent rare harmful outcomes.

3.2.3 Use Cases.

The multi-future reasoning capability of world model 2.0 generates plausible futures for candidate actions and compares their probability-weighted performance and risks, supporting forward-looking decisions under uncertainty. Representative studies are summarized in Table 2. Applications include autonomous driving, service robotics, crowd management, and surveillance, where behavior and interactions may be uncertain [41]. Drive-WM conditions multiview future videos on alternative driving maneuvers and uses image-based rewards to compare the resulting trajectories before execution. By evaluating the predicted scene under each maneuver, the model can support the selection of a suitable driving trajectory [107]. For mobile robots around people, Xu et al. [114] jointly model first-person observations and robot actions to generate multiple possible observation-action futures. These futures capture different possible robot motions and human responses, allowing observation-guided planning to account for interaction uncertainty.

In these applications, multi-future reasoning provides the possible consequences of candidate actions, while proactive decision-making uses the predicted performance and risks to select an action for execution. Multi-future reasoning has also been explored for uncertain behavior and unseen scenarios. Chaudhry [12] represents human motion with a stochastic belief state and anatomical constraints, supporting coherent long-horizon forecasts while limiting error accumulation. Dreamweaver learns object attributes and motion patterns, then recombines them to generate unseen scenes [5]. These studies show how multi-future reasoning can support decisions when behavior is diverse, environments are stochastic, or risks are substantial; in multi-agent settings, the same approach can compare joint actions and their effects on shared states.

Table 2. Representative Studies on Multi-Future Reasoning in World Model 2.0.

Paper title	Domain	Main content	Main finding
Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving [107]	Autonomous driving	Generates multiview future videos under alternative driving actions and evaluates candidate trajectories with image-based rewards.	Supports pre-execution maneuver comparison and improves planning robustness in simulated OOD driving conditions.
Action-Inclusive Multi-Future Prediction Using a Generative Model in Human-Related Scenes for Mobile Robots [114]	Robot control	Jointly generates multiple first-person observation-action futures using a temporally masked diffusion Transformer.	Demonstrates diverse, temporally plausible futures for visual-guided robot planning in human-populated scenes.
Semantic Belief-State World Model for 3D Human Motion Prediction [12]	Human motion prediction	Samples stochastic latent motion hypotheses within a recurrent belief state aligned with the SMPL-X body manifold.	Produces more stable and uncertainty-aware long-horizon motion forecasts with low inference cost.
Dreamweaver: Learning Compositional World Models from Pixels [5]	Embodied intelligence	Learns and recombines object attributes and motion concepts through block-slot representations and multi-future-frame prediction.	Supports compositional imagination of unseen scenes; it is more relevant to novel future composition than probabilistic branching.

3.3 Proactive Decision-Making

3.3.1 Foundational Concept.

The proactive decision-making capability of world model 2.0 means predicting future states and generating control actions within a unified representation and optimization process. The predicted environmental changes directly guide the current decision. The model must not only determine what may happen next, but also decide what action should be taken now in response [90, 105].

The key to proactive decision-making is collaboration between prediction and decision. Future-state prediction provides information about expected changes in environmental conditions, agent states, resource availability, and risk, allowing decisions to address adverse outcomes before they occur. In return, the objective of the action task shapes the shared representation and makes state prediction focus on dynamic factors that affect control outcomes. Environment prediction makes decisions more predictive, while decision supervision makes prediction more relevant to the task [38, 68]. In this process, causal understanding provides the mechanism-level basis for estimating how alternative actions may change future states and task outcomes, while multi-future reasoning provides multiple possible trajectories for comparing performance, risk, and resource costs before execution.

In multi-agent environments, proactive decision-making extends from selecting an individual action to selecting mutually consistent joint actions. The decision process must therefore consider inter-agent effects, shared resources, and collective objectives in addition to individual action outcomes.

In complex interactive environments, the model can reserve resources, adjust continuous controls, reconfigure task plans, or change agent assignments before rising demand or environmental disturbances degrade task performance. Control therefore moves from reactive response to proactive intervention.

3.3.2 Representative Technical Realization.

During operation, the model continuously receives environment observations and control actions. Given historical observations $\{o_{t-d+1}, \ldots, O_t\}$ and action chunks $\{a_{t-d}, \ldots, A_{t-1}\}$, it jointly generates future observations $\{O_{t+1}, \ldots, O_{t+h}\}$

and action chunks $\{A_t, \ldots, A_{t+h-1}\}$, where h is the prediction horizon. Observations can include multimodal physical states, agent states, resource conditions, and task feedback. Actions can represent controls such as motion commands, manipulation operations, scheduling decisions, configuration updates, and agent assignments.

World model 2.0 does not learn only an observation-to-action mapping. It also models the joint distribution of future states and actions through [105, 123]

$$\mathcal{L}_{\text{joint}} = \mathbb{E}_{(o_t, a_t, o_{t+1}) \sim \mathcal{D}} \left[-\log p_\omega \left(o_{t+1}, a_t \mid o_t \right) \right], \tag{30}$$

where $\mathcal{D}$ is the interaction dataset, and p_ω is the joint conditional distribution parameterized by ω. In a interactive system, $\mathcal{L}_{\text{joint}}$ is a negative log-likelihood loss that measures how well the model fits the observed joint outcome of the current observation, subsequent state, and control action.

Consider a unified representation mechanism with a shared Transformer. Historical observations, historical actions, and masked future observations are first encoded and concatenated. The shared backbone then fuses them into joint observation-action representations over the prediction horizon. This data flow can be written as [34, 150]

$$\begin{aligned} X_t = \text{Concat}(&E_o(o_{t-d+1:t}), E_a(a_{t-d:t-1}), \\ &E_o(\widetilde{o}_{t+1:t+d})), \end{aligned} \tag{31}$$

$$\{Z_{t+1}, \ldots, Z_{t+d}\} = \text{Transformer}_\psi(X_t), \tag{32}$$

where E_o and E_a are the observation and action encoders, $\widetilde{o}_{t+1:t+d}$ denotes the masked future observations, X_t is the concatenated input representation, and ψ denotes the parameters of the shared Transformer. Moreover, $Z_{t+j} = \{z_{t+j,1}, \ldots, z_{t+j,N}\}$ is the joint representation at the j-th future step. At inference, the joint model can generate candidate action sequences and their corresponding predicted state trajectories [34, 150]. A task-specific reward function or value estimator scores each candidate trajectory. For a candidate action sequence $A = (a_t, \ldots, a_{t+H-1})$, its predicted return can be written as

$$\hat{G}(A) = \sum_{j=1}^{H} \gamma^{j-1} \hat{r}_{t+j}, \tag{33}$$

where $\hat{r}_{t+j}$ is obtained from the predicted state and action using the task reward function or a learned reward estimator. The decision module selects $A^* = \arg\max_{A \in \mathcal{A}_{\text{cand}}} \hat{G}(A)$, executes the first action, and replans after receiving new feedback. Thus, joint state–action prediction supplies candidate outcomes, while reward maximization is carried out by the downstream scoring and selection step.

During training, the two output heads recover noisy actions and future observations. The action and observation losses are [34, 61]

$$\begin{aligned} \mathcal{L}_{\text{action}}(Z, a) &= \mathbb{E}_{\epsilon, r} \left[\left\| \epsilon - \epsilon_{\theta_a} \left(a^{(r)} \mid r, Z \right) \right\|_2^2 \right], \\ \mathcal{L}_{\text{state}}(Z, o) &= \mathbb{E}_{\epsilon, r} \left[\frac{1}{N} \sum_{i=1}^{N} \left\| \epsilon_i - \epsilon_{\theta_s} \left(o_i^{(r)} \mid r, z_i \right) \right\|_2^2 \right], \\ \mathcal{L}^{(t+j)} &= \mathcal{L}_{\text{action}}^{(t+j)} + \mathcal{L}_{\text{state}}^{(t+j)}, \\ \mathcal{L}_{\text{total}} &= \sum_{j=1}^{d} \mathcal{L}^{(t+j)}, \end{aligned} \tag{34}$$

where r is the diffusion step, $j = 1, \ldots, d$ is the future-step index, $a^{(r)}$ and $o_i^{(r)}$ are the noisy action and the i-th noisy observation feature, ϵ and ϵ_i are their noise terms, and ϵ_{θ_a} and ϵ_{θ_s} are the noise estimates produced by the two output heads. $\mathcal{L}_{\text{state}}$ is the prediction loss for future environment states. $\mathcal{L}_{\text{total}}$ is the total training loss of the diffusion model, representing the combined denoising error of future states and control actions over the prediction horizon. During inference, the action head can generate candidate action chunks directly from the joint representation without first reconstructing complete future observations. For each candidate action sequence, the decision module evaluates the predicted state trajectory using the task reward function and selects the sequence with the highest expected discounted return. It then executes the first action and replans after receiving new observations. Joint state–action prediction therefore provides candidate actions together with their predicted consequences, which can be used for reward-aware selection.

3.3.3 *Use Cases.*

The proactive decision-making capability of world model 2.0 jointly models future states and actions, allowing predictions to inform control and planning [90]. It has been explored in robotic manipulation, autonomous driving, and other embodied tasks requiring long-horizon planning, complex interaction, and feedback. Representative studies are summarized in Table 3.

In robotic manipulation, Cen et al. [11] proposed WorldVLA, which uses unified image, language, and action tokens to jointly predict future images and actions. This allows predicted visual changes to inform grasping and manipulation actions. Bi et al. [7] proposed Motus, which connects visual understanding, video generation, and action generation through shared representations, improving the use of predicted motion information in manipulation. He et al. [38] used language-guided semantic and spatial foresight to condition future-video and action generation, supporting instruction-following and spatially precise control. In these applications, multi-future reasoning provides alternative manipulation outcomes, while proactive decision-making selects an action sequence according to the task objective.

In autonomous driving, Zhao et al. [143] jointly model future visual states and vehicle trajectories, helping reduce error accumulation during long-horizon planning. The predicted scene dynamics can be used to evaluate candidate maneuvers before execution, while proactive decision-making selects a maneuver under driving objectives and constraints. Execution outcomes and prediction errors can then inform subsequent model updates, linking proactive decision-making with open-ended and self-evolving learning.

3.4 Open-Ended and Self-Evolving Learning

3.4.1 *Foundational Concept.*

Traditional world models are usually trained on a predefined dataset. Their internal knowledge and dynamics parameters remain fixed after deployment. Open-ended learning in world model 2.0 means that the environments, tasks, and dynamics observed during training do not form a closed set [24, 49]. New network states and interaction experience can continue to enter the model's knowledge throughout its lifecycle. Self-evolving learning means that the model uses discrepancies between predictions and real feedback to update its external memory and internal dynamics [20, 25]. This process is more than collecting online data or periodically retraining the model from scratch. It converts new experience into internal knowledge that improves later prediction and decision-making. In this sense, proactive decisions provide the interaction outcomes and prediction errors needed for self-evolution, while self-evolving learning updates the knowledge used by subsequent decisions.

Table 3. Representative Studies on Proactive Decision-Making in World Model 2.0.

Paper title	Domain	Main content	Main finding
WorldVLA: Towards Autoregressive Action World Model [11]	Robot control	Uses unified multimodal tokens and a shared autoregressive Transformer to jointly predict future images and actions.	Leveraging predicted visual dynamics improves scene understanding and grasping success.
Motus: A Unified Latent Action World Model [7]	Robot control	Integrates visual understanding, video generation, and action prediction through shared motion representations.	Shared physical and motion knowledge improves manipulation performance and cross-embodiment generalization.
SG-WAM: Text-Grounded and Spatial-Aware Semantic Guidance for World-Action Models [38]	Robot control	Uses language-grounded spatial foresight to guide future-video and action generation.	Improves instruction following, spatial precision, and robustness to environmental changes.
From Forecasting to Planning: Policy World Model for Collaborative State-Action Prediction [143]	Autonomous driving	Uses future-video forecasting as a predictive rationale for trajectory generation.	Anticipating future dynamics reduces collisions and enables safer autonomous-driving decisions.

When intervention outcomes or prediction errors reveal inconsistencies in the learned mechanisms, the model can revise its causal representations and dynamics accordingly.

In open real-world environments, new tasks, entities, tools, operating conditions, and interaction patterns continue to emerge. Physical conditions, demand distributions, and behavioral patterns also change over time. Open-ended and self-evolving learning allows the world model to correct its learned dynamics during operation and retain knowledge acquired in different environments. It therefore supports reliable long-term prediction, adaptive control, and system autonomy. The same capability can support cross-domain adaptation by incorporating target-domain feedback, and can support multi-agent coordination by retaining interaction histories, agent behavior patterns, and reusable coordination experience.

3.4.2 Representative Technical Realization.

The self-evolution of a world model begins with each real interaction. The model predicts the next state from the current latent state and control action. After the action is executed, the environment returns a new observation, from which the encoder obtains the real next latent state. By comparing the prediction with the real result, the model can determine whether its current dynamics still describe the environment. Let $\mathfrak{E}_t$ be the set of stored dynamics experts at time t. An expert in this set is denoted by E, with parameter set θ_{E}. Its prediction and feedback error are [69, 127]

$$\hat{h}_{t+1}(\mathsf{E}) = f_{\theta_{\mathsf{E}}}(h_t, a_t), \tag{35}$$

$$\mathcal{L}_t^{\mathrm{fb}}(\mathsf{E}) = \left\| h_{t+1} - \hat{h}_{t+1}(\mathsf{E}) \right\|_2^2, \tag{36}$$

where $\hat{h}_{t+1}(\mathsf{E})$ is the next latent state predicted by expert E, and $\mathcal{L}_t^{\mathrm{fb}}(\mathsf{E})$ is its feedback error. This error comes directly from the outcome observed after action execution.

After receiving real feedback, the model first writes the realized observation-action transition into episodic memory [121, 137]:

$$\mathcal{M}_{t+1}^{\mathrm{epi}} = \mathcal{M}_t^{\mathrm{epi}} \cup \{(o_t, a_t, o_{t+1})\}, \tag{37}$$

where $\mathcal{M}_t^{\text{epi}}$ is the episodic memory at time t. When the model encounters a similar state and action, it can retrieve the related transition to support the current prediction. Repeated prediction errors can also be summarized as reusable environmental rules [145]. This update makes recent experience immediately available to later predictions.

While updating the memory, the model also monitors the feedback error over a short time window. When the feedback error remains above the current expert's stable baseline over a short time window, the model determines that the environment has undergone a significant change, such as a change in physical conditions, task patterns, or interaction structure, and begins to revise its internal dynamics [124]. Once a dynamics update is triggered, the model first compares how its stored experts predict the current transition. If the current environment resembles a previously observed one, it directly reuses the expert with the lower feedback error. If no stored expert describes the transition accurately, the model adds a set of trainable parameters to the closest expert. Let these new parameters be θ_{new}. The update is [21, 77]

$$\theta_{\text{new}} \leftarrow \theta_{\text{new}} - \alpha_{\text{upd}} \nabla_{\theta_{\text{new}}} \left\| h_{t+1} - f_{\theta_{\text{E}},\theta_{\text{new}}}(h_t, a_t) \right\|_2^2, \tag{38}$$

where $f_{\theta_{\text{E}},\theta_{\text{new}}}$ is the dynamics model formed by the fixed expert parameters θ_{E} and the new parameters θ_{new}, while α_{upd} is the update step size. The pretrained backbone and the parameters of previous experts remain unchanged, while only θ_{new} is updated using the feedback from the current interaction. The interaction proceeds as follows: the model first predicts the next latent state with the selected expert; after the action is executed, it compares the prediction with the observed next state and stores the realized transition in episodic memory. If the feedback error remains within the normal range of an existing expert, that expert is retained or reused. If the error remains high, the model updates θ_{new} according to 38. Once the error becomes stable, the fixed parameters θ_E and the updated parameters θ_{new} are stored as a new expert, which can be selected when similar operating conditions occur again. In this way, new experience is first remembered, then used to update the dynamics only when existing experts are insufficient, and finally converted into reusable model knowledge.

3.4.3 Use Cases.

The multi-agent coordination capability of world model 2.0 models multiple agents' states, intentions, and joint actions under partial observations. It supports coordinated planning by representing how agents affect a shared environment, and has been studied in embodied cooperation, multi-agent reinforcement learning, multi-view simulation, autonomous driving, and wireless networks. Representative studies are summarized in Table 6. Zhang et al. [131] proposed COMBO, which infers other agents' intentions and simulates joint actions to support long-horizon embodied cooperation. Wu et al. [108] proposed MultiWorld, which assigns actions to specific agents and uses a global state representation to maintain consistency across agents' views, thereby supporting coordinated control in multi-agent environments. Zhao et al. [144] introduced a mixture-of-experts model that routes task-related dynamics and reward prediction to suitable experts, enabling knowledge reuse across similar multi-agent tasks while reducing interference between dissimilar tasks. Zhang et al. [140] used sequential action-conditioned diffusion modeling to capture inter-agent dependencies and joint effects on the shared state, supporting imagination-based policy learning and coordinated control.

Research also addresses cooperative driving and wireless networks. In autonomous driving, shared latent states combine observations from vehicles, infrastructure, and other traffic participants to support intention inference and coordinated planning [40]. In wireless communications, Meng et al. [76] predict network states under alternative resource configurations and exchange selected state information to coordinate distributed agents. Together, these studies examine coordination under partial information, interacting actions, and shared objectives.

Table 4. Representative Studies on Open-Ended and Self-Evolving Learning in World Model 2.0.

Paper title	Domain	Main content	Main finding
NavMorph: A Self-Evolving World Model for Vision-and-Language Navigation in Continuous Environments [121]	Vision-language navigation	Updates navigation representations through online interaction and evolution memory.	Improves adaptation and generalization in changing or unseen environments.
Self-Evolving World Models for LLM Agent Planning [137]	Language agents	Revises the world model at deployment time using episodic memory, reusable rules, and selective foresight.	Improves prediction reliability and long-horizon planning without parameter updates.
AdaWorldPolicy: World-Model-Driven Diffusion Policy with Online Adaptive Learning for Robotic Manipulation [127]	Robot control	Updates a world-model-driven manipulation policy from prediction-execution discrepancies and force feedback.	Enables adaptation to visual and physical changes with limited human intervention.
World Action Verifier: Self-Improving World Models via Forward-Inverse Asymmetry [69]	Embodied intelligence	Verifies predicted states through forward-inverse consistency and uses unreliable predictions for model improvement.	Demonstrates how forward–inverse consistency can be used to identify unreliable predictions and guide model updates.
EvoAgent: Self-Evolving Agent with Continual World Model for Long-Horizon Tasks [21]	Embodied intelligence	Combines world-model-guided control, self-verification, reflection, and continual experience updates.	Demonstrates a closed-loop process combining self-verification and continual experience updates for long-horizon tasks.
Self-Evolving Cognitive Framework via Causal World Modeling for Embodied Scientific Intelligence [125]	Embodied intelligence	Uses intervention, counterfactual reasoning, and feedback to continually refine causal knowledge.	Provides a theoretical basis for continual causal learning, but lacks empirical validation.

3.5 Cross-Domain Generalization

3.5.1 Foundational Concept.

Traditional world models are usually trained around one fixed environment or a group of closely related environments. Their representations and dynamics can become tied to specific physical conditions, embodiment configurations, interaction structures, and task distributions. Cross-domain generalization in world model 2.0 means that, after learning environment states and evolution rules in a source domain, the model can transfer their stable structure to an unseen target domain and still support effective state prediction and decision-making [26, 116]. Unlike conventional transfer learning, which typically transfers model parameters or feature representations and then fine-tunes them for a target task, cross-domain generalization in world model 2.0 aims to transfer domain-invariant environment dynamics, including how states evolve and respond to actions. It therefore reduces dependence on domain-specific observations and configurations, allowing the model to adapt across domains without building a separate world model from scratch. The transferred object is not a fixed set of observation values. It is the dynamic relation shared by different environments, such as changes in demand, resource competition, interaction effects, and action response. The model should reduce its dependence on regional appearance, sensor differences, and local configurations while retaining information that is useful for network dynamics.

In general interactive systems, this capability allows one world model to serve different environments, platforms, and operating conditions without building a complete model for every domain. It does not require identical prediction

accuracy in all domains. Its role is to preserve transferable representations and basic dynamics after a domain changes, so that the model can quickly interpret the target-domain state, predict action consequences, and provide a reliable starting point for planning, control, and resource management. After deployment in a target domain, open-ended and self-evolving learning can use new observations and prediction errors to identify transfer mismatches and refine the transferred representations and dynamics. Conversely, the knowledge accumulated across domains can provide a reusable basis for future transfer.

3.5.2 *Representative Technical Realization.*

Cross-domain generalization is not only how to move source-domain data to the target domain, but also what the model has actually learned from the source domain. World model 2.0 therefore first converts observations into latent representations of environment dynamics and then learns how an action changes the next state in that representation space. As an example of latent-state prediction, we reuse the encoder f_{enc} and dynamics model f_{dyn} defined earlier. The target branch is indicated by stop-gradient, and L_{hist} is the observation-history length [50, 129]:

$$\hat{h}_{t+1} = f_{\text{dyn}}\left(f_{\text{enc}}(o_{t-L_{\text{hist}}:t}), a_t\right), \tag{39}$$

$$h_{t+1} = \text{sg}[f_{\text{enc}}(o_{t+1})], \tag{40}$$

$$\mathcal{L}_{\text{pred}} = \left\|\hat{h}_{t+1} - h_{t+1}\right\|_2^2, \tag{41}$$

where $o_{t-L_{\text{hist}}:t}$ is the observation history, L_{hist} is its length, $\hat{h}_{t+1}$ is the predicted value of the next latent state h_{t+1}, $\text{sg}[\cdot]$ denotes stop-gradient, $\mathcal{L}_{\text{pred}}$ is the latent prediction loss, and $\|\cdot\|_2$ denotes the L_2 norm. After training, the model obtains a representation oriented toward future dynamics rather than a direct memory of source-domain appearance. This provides the basis for cross-domain transfer [16].

However, this step alone is not sufficient. Latent prediction may still preserve source-domain differences. To make the representation from the first step closer to cross-domain structural regularities, the model can add structural information during training [103]. It constrains the representation during training and need not be supplied continuously after deployment. Let c_t^{obs} be the observation feature, c_t^{str} the corresponding structural feature, and G_{str} a shared structural mapping [43, 66]:

$$\mathcal{L}_{\text{str}} = 1 - \cos\Big(G_{\text{str}}(c_t^{\text{obs}}), \text{sg}\big[G_{\text{str}}(c_t^{\text{str}})\big]\Big), \tag{42}$$

where$\mathcal{L}_{\text{str}}$ is the structural alignment loss and $\cos(\cdot,\cdot)$ denotes cosine similarity. Minimizing $\mathcal{L}_{\text{str}}$ maximizes the cosine similarity between the observation and structural branches for the same state, encouraging the observation representation to preserve domain-invariant information such as topology, spatial layout, and state changes. After source-domain training, the learned model parameters are transferred to the target domain instead of being trained again from scratch. Target-domain observations and actions are organized in the same state-action format, allowing the model to reuse the learned relations between environment states, actions, and state evolution. When the two domains are sufficiently similar, the transferred model can be used directly; otherwise, a small amount of target-domain trajectories is used to fine-tune the model before deployment [86]. The reliability of the transferred model is then evaluated in the target domain.

To determine whether the transferred model still supports state understanding, decision rollouts, and control, the system compares the model's rolled-out representation of the next state with the representation encoded from the

actual observation. Their mean squared error is used as an unlabeled score [43]:

$$\Delta_t^{\text{dom}} = \frac{1}{D_{\text{lat}}} \left\| \hat{h}_{t+1} - h_{t+1} \right\|_2^2, \tag{43}$$

where D_{lat} is the latent representation dimension and Δ_t^{dom} is the score for the current transition. A low score may indicate that the model's one-step latent prediction is close to the encoded observation for this transition, while a high or increasing score may indicate a possible model mismatch. Accordingly, 43 can be treated as a candidate indicator of model mismatch.

3.5.3 Use Cases.

The cross-domain generalization capability of world model 2.0 aims to transfer environmental dynamics across regions, sensors, robot platforms, and tasks rather than rely on scene-specific appearance. Representative studies are summarized in Table 5. Jaiswal et al. [43] used self-supervised JEPA-based representation learning with geographically diverse driving data to study zero-shot transfer to unseen cities and sensor configurations. Khan [50] introduced depth regularization to encourage RGB representations to retain stable 3D geometry, reducing visual-odometry error and improving out-of-domain scenario detection and long-horizon prediction.

In robot learning, Jang et al. [44] generated synthetic trajectories with pseudo-action labels to train visuomotor policies that transfer to new behaviors and environments. Here, cross-domain generalization improves the reuse of visual dynamics and action representations across robot platforms and task settings. Cross-domain generalization also supports the development of general-purpose simulation. PAN combines language-conditioned latent-state prediction and video diffusion to model actions, states, and visual changes across domains [110]. Vista uses geographically diverse driving data, structural constraints, and multi-level action conditioning to improve prediction and control across driving scenarios [26]. Together, these studies explore transfer across appearance, location, sensors, and embodiments, while target-domain feedback and causal representations can further refine the transferred dynamics.

3.6 Multi-Agent Coordination

3.6.1 Foundational Concept.

Early multi-agent world models often flatten all agent states and actions into one high-dimensional joint input and directly learn the overall transition [64]. As the number of agents grows, the joint action space expands rapidly, and the interactions among agents become harder to factorize. The multi-agent coordination capability of world model 2.0 means that the model can represent multiple agents in one shared environment state, distinguish their observations and actions, model their joint effects on the global state, and use this model to roll out and select mutually consistent behaviors [131].

Compared with the independent local models or directly flattened joint models used in traditional world models, world model 2.0 uses shared-state modeling and structured joint-action rollouts to preserve coupling among nodes without compressing all interactions into one difficult-to-scale high-dimensional transition function [108]. This capability supports joint task allocation, distributed control, shared-resource coordination, and cooperative planning. Its realization draws on causal understanding to model inter-agent effects, multi-future reasoning to evaluate alternative joint actions, and proactive decision-making to select mutually consistent behaviors. Open-ended and self-evolving learning can further support coordination by retaining interaction histories and updating reusable models of agent behavior and cooperation.

Table 5. Representative Studies on Cross-Domain Generalization in World Model 2.0.

Paper title	Domain	Main content	Main finding
Geographic Diversity Beats Data Volume for Cross-Domain Generalization in Zero-Label JEPA Driving World Models [43]	Autonomous driving	Learns transferable dynamics from geographically diverse driving data.	Reports transfer to unseen regions using geographically diverse driving data in the evaluated setting.
Depth-Regularized JEPA World Models Learn More Transferable Representations from Real Outdoor Robot Data [50]	Robot control	Uses depth supervision to preserve geometric information while retaining RGB-only inference.	Improves robustness in visual odometry, anomaly detection, and long-horizon prediction.
DreamGen: Unlocking Generalization in Robot Learning through Video World Models [44]	Robot control	Generates videos and pseudo-actions from images and language instructions.	Enables generalization to new robot behaviors and environments with sparse demonstrations.
PAN: A World Model for General, Interactable, and Long-Horizon World Simulation [110]	Embodied intelligence	Simulates future states under language-conditioned actions across video domains.	Demonstrates potential for open-domain, interactive, long-horizon simulation.
Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability [26]	Autonomous driving	Learns controllable future driving scenes from multi-region data.	Enables zero-shot cross-dataset control transfer and improves prediction fidelity.

3.6.2 *Representative Technical Realization.*

The multi-agent coordination of world model 2.0 consists of four connected stages. First, local observations from different agents are aligned into a shared environment state. The effects of joint actions on the next state are then decomposed into learnable sequential conditional transitions. During deployment, limited neighborhood information is used to support decentralized execution. Finally, the shared dynamics model rolls out different joint actions, and the imagined outcomes are fed back to the coordination policy.

The model first needs to align information that is observed locally by different nodes into a network state that can be jointly rolled out. Suppose that the system contains N_{ag} agents. The local observation and action of agent i at time t are denoted by o_t^i and a_t^i, and the joint action is $a_t^{1:N_{\mathrm{ag}}} = (a_t^1, \ldots, a_t^{N_{\mathrm{ag}}})$. In world model 2.0, the environment evolves according to the shared state s_t and the joint action [95, 119]:

$$s_{t+1} \sim P\left(\cdot \mid s_t, a_t^{1:N_{\mathrm{ag}}}\right), \tag{44}$$

where P is the stochastic global environment transition distribution, and s_t may include physical context, external disturbances, resource demand, entity states, spatial relations, and shared resource occupancy. The encoder uses each node's local observation together with permitted neighborhood information to estimate this shared state. Subsequent dynamics modeling can then operate on one common network state instead of isolated local states.

Directly learning $P(s_{t+1} \mid s_t, a_t^{1:N_{\mathrm{ag}}})$ requires all agent actions to enter one joint condition. To reduce this modeling difficulty, the model can use diffusion-inspired sequential conditioning. It starts from a highly noisy next state and

introduces action conditions according to an agent ordering. Let $s_{t+1}^{(0)} = s_{t+1}$ be the clean next state, $s_{t+1}^{(N_{\text{ag}})}$ the highest-noise state, and $\rho = (i_1, \ldots, i_{N_{\text{ag}}})$ an agent ordering. The joint transition can then be written as [140]

$$\begin{aligned} &P\left(s_{t+1}, s_{t+1}^{(1:N_{\text{ag}})} \mid s_t, a_t^{1:N_{\text{ag}}}\right) \\ &= p\left(s_{t+1}^{(N_{\text{ag}})}\right) \times \prod_{\ell=1}^{N_{\text{ag}}} p\left(s_{t+1}^{(\ell-1)} \mid s_{t+1}^{(\ell)}, a_t^{i_\ell}, s_t\right), \end{aligned} \tag{45}$$

where ℓ is the sequential denoising step, $s_{t+1}^{(1:N_{\text{ag}})}$ denotes the intermediate-state sequence, and p denotes the conditional transition density in the sequential denoising process. At the ℓ-th conditional transition, only the action of the ℓ-th agent in ρ is introduced, while every step is conditioned on the same current state s_t. The joint effect of multiple actions is therefore decomposed into a sequence of conditional transitions instead of one large joint condition.

During training, a denoising network D_θ recovers the clean next state from intermediate states at different noise levels. Let v_ℓ be the noise level at the ℓ-th conditioning step. Let $\mathcal{L}_{\text{ma}}(\theta)$ denote the multi-agent denoising loss [138]

$$\mathcal{L}_{\text{ma}}(\theta) = \mathbb{E}_{\rho, \{v_\ell\}}\left[\sum_{\ell=1}^{N_{\text{ag}}} \left\| D_\theta\left(s_{t+1}^{(\ell)}; v_\ell, s_t, a_t^{i_\ell}\right) - s_{t+1} \right\|_2^2\right]. \tag{46}$$

This objective makes every action condition contribute to next-state recovery and learns how different agents affect the shared state under different orderings. Once the joint dynamics are learned, the model can input different joint actions in imagination, compare their effects on task utility, latency, safety risk, and shared resource occupancy, and use the better joint outcomes for cooperative policy learning.

3.6.3 Use Cases.

The multi-agent coordination capability of world model 2.0 models multiple agents' states, intentions, and joint actions under partial observations. It supports coordinated planning by representing how agents affect a shared environment, and has been studied in embodied cooperation, multi-agent reinforcement learning, multi-view simulation, autonomous driving, and wireless networks. Representative studies are summarized in Table 6. Zhang et al. [131] proposed COMBO, which infers other agents' intentions and simulates joint actions to support long-horizon embodied cooperation. Wu et al. [108] proposed MultiWorld, which assigns actions to specific agents and uses a global state representation to maintain consistency across agents' views, thereby supporting coordinated control in multi-agent environments. Zhao et al. [144] introduced a mixture-of-experts model that routes task-related dynamics and reward prediction to suitable experts, enabling knowledge reuse across similar multi-agent tasks while reducing interference between dissimilar tasks. Zhang et al. [140] used sequential action-conditioned diffusion modeling to capture inter-agent dependencies and joint effects on the shared state, supporting imagination-based policy learning and coordinated control.

Research also addresses cooperative driving and wireless networks. In autonomous driving, shared latent states combine observations from vehicles, infrastructure, and other traffic participants to support intention inference and coordinated planning [40]. In wireless communications, Meng et al. [76] predict network states under alternative resource configurations and exchange selected state information to coordinate distributed agents. Together, these studies examine coordination under partial information, interacting actions, and shared objectives.

3.7 General Cognitive Architecture

3.7.1 Foundational Concept.

Table 6. Representative Studies on Multi-Agent Coordination in World Model 2.0.

Paper title	Domain	Main content	Main finding
COMBO: Compositional World Models for Embodied Multi-Agent Cooperation [131]	Embodied intelligence	Estimates a global state from partial observations, simulates joint actions with a compositional video diffusion model, and combines intent inference with tree-search planning.	Enables pre-execution evaluation of cooperative outcomes, improving success, planning efficiency, and scalability to different team sizes.
MultiWorld: Scalable Multi-Agent Multi-View Video World Models [108]	Embodied intelligence	Models multi-agent, multi-view environments with agent identity embeddings, adaptive action weighting, and a global state encoder.	Improves action following and view consistency, while supporting cooperative-failure simulation for control and risk analysis.
Learning and Planning Multi-Agent Tasks via an MoE-Based World Model [144]	Embodied intelligence	Uses MoE-based dynamics and reward models to route tasks to suitable experts and plans from predicted rollouts.	Reuses similar task dynamics and reduces interference, improving sample efficiency and multitask performance.
Revisiting Multi-Agent World Modeling from a Diffusion-Inspired Perspective [140]	Multi-agent cooperation	Models joint state transitions as sequentially action-conditioned diffusion denoising.	Reduces joint-action modeling complexity, captures inter-agent effects more accurately, and improves sample efficiency and control return.
Multi-Agent Embodied Autonomous Driving: From V2X Information Exchange to Shared World Models [40]	Autonomous driving	Reviews multi-source fusion, shared-state modeling, intent understanding, and cooperative planning.	Shared world models overcome isolated observations and improve joint decision-making in occluded and interactive traffic.
NetWorld: Communication-Based Diffusion World Model for Multi-Agent Reinforcement Learning in Wireless Networks [76]	Wireless network control	Combines a conditional diffusion world model, shared latent representations, inverse dynamics, and mean-field coordination.	Reduces online interaction in wireless resource management while improving decision performance, sample efficiency, and cross-task generalization.

Traditional world models typically consist of an encoder, a dynamics model, and a decoder. Their main function is to compress observations into latent states and predict subsequent environmental states, while perception, task understanding, planning, decision-making, and execution are often handled by separate external modules. Because these modules use independent representations and optimization objectives, traditional world models are generally designed for specific tasks and scenarios, making their learned environmental knowledge difficult to reuse.

The general cognitive architecture of world model 2.0 aims to provide a general cognitive substrate that can be reused across tasks, scenarios, and embodiments. As this substrate, the architecture encodes inputs from different tasks into a unified internal state and learns the corresponding environmental dynamics, thereby providing subsequent task modules with reusable state representations and environment predictions. Task-specific modules can connect to the substrate through feature alignment and interface adaptation, using the shared states to perform their respective planning, decision-making, or control functions. Consequently, adapting the system to a new task or scenario requires only the adjustment of the corresponding task module, rather than rebuilding the entire world model. The general cognitive architecture provides shared representations, memory, and interfaces through which the six core capabilities can interact. It connects causal understanding, multi-future reasoning, proactive decision-making, open-ended and self-evolving learning, cross-domain generalization, and multi-agent coordination within a reusable perception-to-action framework.

3.7.2 Representative Technical Realization.

In complex interactive systems, heterogeneous observations may reflect physical conditions, resource demand, agent states, task progress, and external disturbances. Their value formats, sampling intervals, and semantic meanings are different. World model 2.0 must therefore first convert these heterogeneous observations into a common state representation and establish a unified information interface. Each input type is encoded separately, and adapters then map the resulting features into a shared state space. In Pandora's architecture, for example, a state encoder extracts the environment state, a language encoder processes the task or action, a cognitive backbone integrates both representations, and a state generator reconstructs the next state [111].

After the shared state is formed, the key task is to make the action condition part of state evolution instead of separating prediction from control. The state-transition distribution is written as [130]:

$$P(s_t \mid s_{t-1}, a_{t-1}, \ldots, s_1, a_1), \tag{47}$$

where $P(\cdot)$ denotes the conditional state-transition distribution. The distribution shows that the next state depends jointly on historical states and actions, so actions directly enter the cognitive process.

After generating one state, the model appends it to the history and processes the next action, forming a continuous autoregressive rollout. The same cognitive backbone can therefore repeat state understanding and action-conditioned rollout across tasks without a separate predictor for each task.

For state representations and task instructions to interact in the same backbone, their representation spaces must also be aligned. Let $\mathbf{e}^{\mathrm{LM}}$ be the representation output by the cognitive backbone, $g_{\mathrm{adp}}(\cdot)$ the adapter connecting the backbone to the state generator, and $\mathbf{e}^{\mathrm{ref}}$ the generator's original conditioning representation. The alignment objective is [85]:

$$\mathcal{L}_{\mathrm{align}} = \left\| g_{\mathrm{adp}}\left(\mathbf{e}^{\mathrm{LM}}\right) - \mathbf{e}^{\mathrm{ref}} \right\|_2^2, \tag{48}$$

where $\mathcal{L}_{\mathrm{align}}$ is the representation-alignment loss. Minimizing it makes the backbone output readable by the state generator. Training then proceeds in two steps: large-scale state and text data first provide stable state-generation and task-understanding abilities; sequences with explicit actions and state changes then connect actions to their subsequent effects through instruction tuning.

After training, the model receives current environment observations and a task objective, and handles state understanding, state rollout, and action generation in one process. It first assesses how candidate actions may change the current state and then outputs a control action. The new observation after execution is fed back to update the next state. This allows the learned state representation and action understanding to be reused across tasks and embodiments and reduces information loss between rollout and control.

3.7.3 Use Cases.

The general cognitive architecture of world model 2.0 connects multimodal perception, task understanding, environmental dynamics, future rollouts, decision generation, and action execution through shared states and interfaces. Representative studies are summarized in Table 7. Its generality does not require a single network to implement every function; instead, transferable internal states allow environmental understanding, prediction, and control to be reused across tasks [153]. This approach has been explored in interactive simulation, robotic manipulation, autonomous driving, and autonomous agents [116].

Xiang et al. [111] proposed Pandora, which combines natural-language actions, video states, and future-scene generation in a unified framework, allowing a shared state-action interface to support controllable simulation across

Table 7. Representative Studies on General Cognitive Architectures in World Model 2.0.

Paper title	Domain	Main content	Main finding
Pandora: Towards General World Model with Natural Language Actions and Video States [111]	Interactive simulation / robotics / autonomous driving / game simulation	Jointly models language actions and video states to generate future scenes.	Enables real-time, language-controlled simulation across domains, demonstrating the feasibility of a general language-vision state interface.
Astra: General Interactive World Model with Autoregressive Denoising [152]	Environment exploration / robot control / autonomous driving	Unifies heterogeneous actions, including camera, navigation, and robot actions.	Maintains action responsiveness and long-horizon temporal consistency across exploration, robot control, and driving tasks.
ADriver-I: A General World Model for Autonomous Driving [46]	Autonomous driving	Recurrently predicts control signals and future scenes to form a vision-action loop.	Enables internal closed-loop driving simulation without requiring 3D bounding boxes or high-definition maps.
Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability [26]	Autonomous driving	Generates future driving scenes under multi-level actions and evaluates action reliability.	Provides high-fidelity, controllable prediction in unseen scenes and supports action evaluation without ground-truth action labels.

domains. Zhu et al. [152] proposed Astra, which uses action-aware adapters, memory, and action experts to map heterogeneous actions across exploration, robotics, and driving into a common predictive backbone. In autonomous driving, Jia et al. [46] introduced interleaved vision-action representations that connect scene prediction with control, so predicted future scenes can directly inform subsequent control signals. Gao et al. [26] combined future-state prediction, multi-level driving controls, and action evaluation through a unified conditioning interface, enabling the same predictive model to support planning and control across driving scenarios. These studies illustrate how shared states and explicit interfaces can connect perception, prediction, action, and control across tasks and domains.

4 Challenges And Future Directions

4.1 Causal Understanding

Although recent world models increasingly support action-conditioned simulation, it remains difficult to determine whether their latent dynamics capture stable causal mechanisms or mainly reflect correlations in the training data. Objects, attributes, actions, and scene context are often entangled in latent states, making it difficult to distinguish variables that merely predict an outcome from those that determine it [35]. Large-scale video data also contain relatively few controlled interventions, which limits the evidence available for identifying action effects [30]. As a result, visually plausible short-term predictions may still rely on invalid transition structures, with errors accumulating over long horizons or emerging when mechanisms, embodiments, or sensors change [2, 10, 39].

Future models should develop hierarchical, action-grounded representations that distinguish entities, attributes, relations, affordances, and action-conditioned transitions [80]. These representations should be learned from passive video together with targeted demonstrations and interventions, allowing models to compare alternative actions and their consequences [118]. Their reliability should be evaluated through intervention fidelity, counterfactual consistency, object persistence, long-horizon error, and robustness to changes in dynamics and embodiment. Uncertainty estimates should also indicate when a forecast is weakly supported [8].

4.2 Multi-Future Reasoning

Action-conditioned world models can generate multiple possible futures, which can help planners reason under uncertainty [107, 114]. However, it remains difficult to determine whether a set of futures is genuinely useful. Existing methods do not always assess diversity, coverage, calibration, physical plausibility, and downstream decision quality together [74, 147]. Individually plausible branches may also conflict with shared scene constraints, while rare but consequential events and unseen combinations of objects, attributes, and motions remain poorly covered [5, 14, 147, 154]. These limitations make it difficult to determine whether additional branches improve safety and planning or merely increase the number of possible outcomes.

Future benchmarks should therefore evaluate branch diversity, probability calibration, consistency, physical validity, and planning performance together. To improve the quality of generated futures, models can combine hierarchical temporal prediction with cross-branch constraints, risk-sensitive sampling, and explicit contingency branches. Physical or causal constraints can further help maintain feasible long-horizon and compositional predictions, while planning-oriented evaluation can test whether the generated branches improve actual decision quality.

4.3 Proactive Decision-Making

Action-centered world models increasingly connect future-state prediction with control generation [38, 68], but a prediction-guided controller must capture more than visually plausible observations. Visual fidelity does not guarantee that the model represents control-relevant factors such as contact, controllability, or task progress [15, 104]. When predictions are used recursively for planning, early errors can distort later states and decisions, particularly when the model must handle heterogeneous action interfaces [152]. Interactive settings further require decisions to account for resource limits, safety constraints, and the responses of other agents [1, 33].

Future systems should prioritize action-relevant states and connect them directly to action generation. Uncertainty-aware receding-horizon planning can update forecasts after each execution step and revise the remaining actions when the observed outcome differs from the prediction. Joint-action prediction, counterfactual rollouts, and feasibility checks can further help the planner account for constraints and other agents. Evaluation should therefore consider not only prediction quality but also control feasibility, task progress, and decision robustness.

4.4 Open-Ended and Self-Evolving Learning

Recent work has begun to adapt world models through deployment-time interaction [121, 127]. The remaining challenge is to ensure that these updates improve the model without reinforcing misleading feedback or erasing previously acquired capabilities. Reliable updating first requires distinguishing useful prediction errors from invalid feedback. For example, an action may lead to a visually plausible state while being physically unreachable, so predicted transitions need to be checked before they are stored [69, 70]. Even when new feedback is valid, adapting to new dynamics can cause forgetting [135]. Large interaction histories are also difficult to reuse unless the system can retrieve and combine relevant memories rather than treating all experience as a single replay buffer [94, 109].

Future systems should select updates using prediction error together with action reachability, novelty, uncertainty, recurrence, and downstream task impact. Stable knowledge should be separated from rapidly changing experience, while interaction histories should be converted into structured rules, causal relations, affordances, and skills whose provenance and confidence can be tracked. Evaluation should examine long-term adaptation, retention of previous capabilities, and recovery from harmful updates rather than focusing only on short-term task improvement [145].

4.5 Cross-Domain Generalization

World models increasingly use geometry-aware representations, shared motion spaces, and synthetic trajectories to transfer dynamics across domains [44, 50, 52]. However, transfer remains unreliable when changes in appearance, physical dynamics, action interfaces, and target-domain data occur together. Transfer can fail when domain-specific appearance or sensor conditions become entangled with the dynamics representation, distorting predictions even when the underlying physical rules remain stable [57, 83]. Even when the dynamics transfer, differences in embodiment and actuator limits may prevent the learned representation from producing executable actions [134, 151]. Sparse target-domain interaction further makes it difficult to distinguish a genuine dynamics shift from noise or an unusual event [45, 62].

A more transferable model should factorize domain-invariant dynamics, appearance-dependent observations, and embodiment-specific action interfaces. Transfer should also include reliability estimates that indicate whether the model can be used directly or requires calibration and further interaction. Future evaluation should measure long-horizon prediction, action feasibility, recovery from model mismatch, and data efficiency together. Such evaluation can clarify whether a model has learned reusable dynamics or only adapted to domain-specific regularities.

4.6 Multi-Agent Coordination

Multi-agent world models extend prediction beyond isolated agents by representing shared environments and the effects of joint actions [95, 108, 131]. Reliable coordination, however, depends on whether agents can maintain accurate and compatible beliefs as their interactions become more complex. Coordination begins with inferring a reliable shared state from partial and heterogeneous observations; errors in what agents know about the environment or one another can propagate into conflicting plans [55, 140]. As the team grows, the joint action space and the number of interaction dependencies make prediction and planning more costly [32]. Even when local predictions are individually plausible, differences in viewpoint, timing, or inferred scene structure can leave the resulting plans globally incompatible [92].

To support coordination, shared states should represent agent identity, spatial relations, uncertainty, and communication status. Joint dynamics can then be factorized into individual action effects, interactions, and global constraints, reducing planning cost while preserving important dependencies. Future benchmarks should test how coordination responds to changes in team size, observation coverage, communication delays, agent failures, and conflicting objectives. They should also evaluate whether agents can maintain compatible beliefs and recover when communication or observations are incomplete.

4.7 General Cognitive Architecture

Recent work aims to connect multimodal perception, language-conditioned actions, state prediction, and feedback through reusable architectures with shared representations and interfaces [110, 111]. The challenge is to make these components work together while preserving the information and constraints needed by each one. Language, vision, and action differ in temporal scale, semantic structure, and uncertainty, so a shared representation can create alignment errors or obscure task-relevant information [59, 98]. Even when abstract reasoning is coherent, its plans may rely on unreachable states or unexecutable actions [142]. These gaps can persist at the system level because perception, memory, prediction, planning, and control may operate at different timescales or optimize different objectives, leaving failures at the interfaces between components [36].

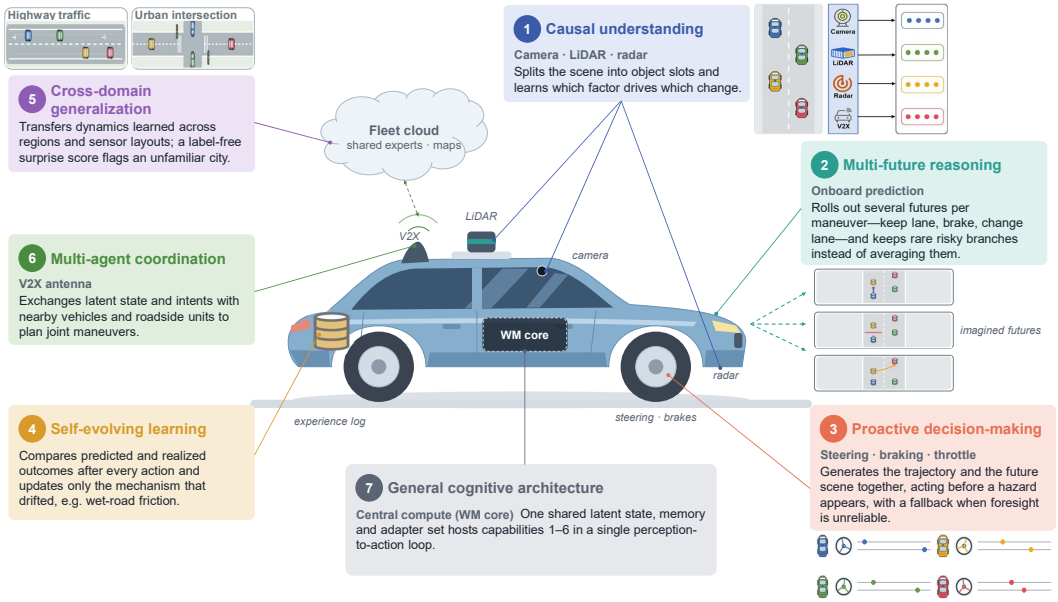


Fig. 3. Deployment of world model 2.0 in an autonomous driving system. Multimodal observations from cameras, LiDAR, radar, and vehicle-to-everything (V2X) communication are encoded into a shared environmental representation. Causal understanding and multi-future reasoning use this representation to analyze state changes and compare the outcomes of alternative maneuvers. Proactive decision-making then selects control actions, such as steering, acceleration, or braking, before execution. The observed consequences and prediction errors are fed back to update the model dynamics and memory, supporting self-evolving learning. Cross-domain generalization enables transfer across cities, road conditions, and sensor configurations, while multi-agent coordination models the behaviors and interactions of surrounding vehicles and roadside units. The general cognitive architecture provides the shared representations and interfaces that connect these capabilities into a closed-loop driving system.[86, 91, 106, 117]

Future architectures should use structured shared states and explicit interfaces between perception, memory, dynamics, planning, and control, while modality-specific encoders retain information needed by their respective tasks. Closed-loop interaction can help these components revise predictions and plans in response to executed actions and feedback. Evaluation should therefore examine state understanding, long-horizon prediction, intervention reasoning, action feasibility, planning, transfer, and reliability under distribution shift rather than relying on a single end-task score.

4.8 A Representative Application Scenario

Fig. 3 illustrates the closed-loop driving workflow. Multimodal observations support causal analysis and comparison of alternative maneuvers [107, 148]; the selected action is then evaluated against execution feedback, which updates the model and memory [102, 121]. Cross-domain generalization and multi-agent coordination extend this workflow across driving conditions and interacting road users [43, 108].

5 Conclusion

This paper introduces world model 2.0 as a general cognitive architecture for complex interactive environments. It develops a taxonomy of six core capabilities and examines how the architecture connects them, reviews representative

methods and applications, and identifies key challenges and future research directions. The reviewed work shows that these capabilities have been explored in different settings, while their integration and reliable evaluation within a unified architecture remain open problems.

References

[1] Hasaan Ahmad and Haixia Liu. 2026. The Cost of Dreaming: A Survey of Computational Constraints in Generative and Latent World Models. *IEEE Computational Intelligence Magazine* (2026).

[2] Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Manuel Wüthrich, and Stefan Bauer. 2020. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296* (2020).

[3] Feisal Alaswad, Batoul Aljaddouh, Maher Alrahhal, Talal Bonny, et al. 2026. Why We Need World Models for AGI: Where LLMs Fail and How World Models May Outperform. *arXiv preprint arXiv:2605.23972* (2026).

[4] Abi Aryan and Zac Liu. 2025. Causal Reflection with Language Models. *arXiv preprint arXiv:2508.04495* (2025).

[5] Junyeob Baek, Yi-Fu Wu, Gautam Singh, and Sungjin Ahn. 2025. Dreamweaver: Learning compositional world models from pixels. *arXiv preprint arXiv:2501.14174* (2025).

[6] Arti Bandhana and Jiří Vokřínek. 2025. AI-Driven Manufacturing: Surveying for Industry 4.0 and Beyond. In *Operations Research Forum*, Vol. 6. Springer, 145.

[7] Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. 2026. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 35101–35113.

[8] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. 2024. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*.

[9] Kenneth P Burnham and David R Anderson. 2001. Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife research* 28, 2 (2001), 111–119.

[10] Kunlin Cai, Rui Song, Jinghuai Zhang, Kaiyuan Zhang, Pranav Bodapati, Alicia Yu, Fnu Suya, Mohammad Rostami, Jiaqi Ma, and Yuan Tian. 2026. What-If World: A Causal Benchmark for General World Models in Embodied Scenarios. *arXiv preprint arXiv:2605.27589* (2026).

[11] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. 2025. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539* (2025).

[12] Sarim Chaudhry. 2026. Semantic Belief-State World Model for 3D Human Motion Prediction. *arXiv preprint arXiv:2601.03517* (2026).

[13] Zhikang Chen and Tingting Zhu. 2026. From Generative Engines to Actionable Simulators: The Imperative of Physical Grounding in World Models. *arXiv preprint arXiv:2601.15533* (2026).

[14] Lei Chu et al. 2026. Diverse Yet Consistent: Context-Guided Diffusion with Energy-Based Joint Refinement for Multi-Agent Motion Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1029–1039.

[15] Peterson Co, Sicheng Hu, Chunxuan Jiao, Hongyang Cheng, Yulin Luo, Yijie Xu, Sixiang Chen, Zhongxia Zhao, Zihao Wang, DaFeng Chi, et al. 2026. WorldSimProbe: Diagnosing Simulator Faithfulness in Action-Conditioned World Models for Embodied Manipulation. *arXiv preprint arXiv:2608.09298* (2026).

[16] Jingyi Cui, Qi Zhang, Hongwei Wen, and Yisen Wang. 2026. A Generalization Theory for JEPA-Based World Models. arXiv:2606.27014 [cs.LG] https://arxiv.org/abs/2606.27014

[17] Elisabeth Dillies, Quentin Delfosse, Jannis Blüml, Raban Emunds, Florian Peter Busch, and Kristian Kersting. 2025. Better decisions through the right causal world model. *arXiv preprint arXiv:2504.07257* (2025).

[18] Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. 2025. Understanding world or predicting future? a comprehensive survey of world models. *Comput. Surveys* 58, 3 (2025), 1–38.

[19] Fan Feng, Phillip Lippe, and Sara Magliacane. 2026. Learning interactive world model for object-centric reinforcement learning. *Advances in Neural Information Processing Systems* 38 (2026), 89827–89862.

[20] Tongtong Feng, Xin Wang, Zekai Zhou, Ren Wang, Yuwei Zhan, Guangyao Li, Qing Li, and Wenwu Zhu. 2025. EvolvingAgent: Curriculum Self-evolving Agent with Continual World Model for Long-Horizon Tasks. *arXiv preprint arXiv:2502.05907* (2025).

[21] Tongtong Feng, Xin Wang, Zekai Zhou, Ren Wang, Yu-Wei Zhan, Guangyao Li, Qing Li, and Wenwu Zhu. 2025. Evoagent: Self-evolving agent with continual world model for long-horizon tasks. (2025).

[22] Céline Finet, Stephane Da Silva Martins, Jean-Bernard Hayet, Ioannis Karamouzas, Javad Amirian, Sylvie Le Hégarat-Mascle, Julien Pettré, and Emanuel Aldea. 2025. Recent advances in multi-agent human trajectory prediction: A comprehensive review. *arXiv preprint arXiv:2506.14831* (2025).

[23] Karl Friston, Rosalyn J Moran, Yukie Nagai, Tadahiro Taniguchi, Hiroaki Gomi, and Josh Tenenbaum. 2021. World model learning and inference. *Neural Networks* 144 (2021), 573–590.

[24] Zhenfeng Gan, ZiTong Zeng, Jiajun Cheng, Yeke Song, Yongyi Tang, and Xueqian Wang. 2026. ActSWM: Action-Sensitive World Models for Long-Horizon Planning in Open-World Games. arXiv:2607.26712 [cs.RO] https://arxiv.org/abs/2607.26712

[25] Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, et al. 2025. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence. *arXiv preprint arXiv:2507.21046* (2025).

[26] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. 2024. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems* 37 (2024), 91560–91596.

[27] Yinfeng Gao, Qichao Zhang, Da-Wei Ding, and Dongbin Zhao. 2024. Dream to drive with predictive individual world model. *IEEE Transactions on Intelligent Vehicles* 9, 12 (2024), 8224–8238.

[28] Ignat Georgiev, Varun Giridhar, Nick Hansen, and Animesh Garg. 2025. PWM: Policy learning with multi-task world models. In *International Conference on Learning Representations*, Vol. 2025. 13737–13757.

[29] John Gkountouras, Matthias Lindemann, Phillip Lippe, Efstratios Gavves, and Ivan Titov. 2025. Language Agents Meet Causality–Bridging LLMs and Causal World Models. In *International Conference on Learning Representations*, Vol. 2025. 37311–37336.

[30] Jinwei Gu. 2025. Cosmos world foundation models for physical ai. In *Proceedings of the 3rd International Workshop on Rich Media With Generative AI*. 39–39.

[31] Yanchen Guan, Haicheng Liao, Zhenning Li, Jia Hu, Runze Yuan, Guohui Zhang, and Chengzhong Xu. 2024. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles* (2024).

[32] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* (2024).

[33] Weihang Guo, Zachary Kingston, and Lydia E Kavraki. 2025. Castl: Constraints as specifications through llm translation for long-horizon task and motion planning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 11957–11964.

[34] Yanjiang Guo, Yucheng Hu, Jianke Zhang, Yen-Jen Wang, Xiaoyu Chen, Chaochao Lu, and Jianyu Chen. 2024. Prediction with Action: Visual Policy Learning via Joint Denoising Process. arXiv:2411.18179 [cs.RO] https://arxiv.org/abs/2411.18179

[35] Tarun Gupta, Wenbo Gong, Chao Ma, Nick Pawlowski, Agrin Hilmkil, Meyer Scetbon, Marc Rigter, Ade Famoti, Ashley Juan Llorens, Jianfeng Gao, et al. 2024. The essential role of causality in foundation world models for embodied AI. *arXiv preprint arXiv:2402.06665* (2024).

[36] Jianhua Han, Meng Tian, Jiangtong Zhu, Fan He, Huixin Zhang, Sitong Guo, Dechang Zhu, Hao Tang, Pei Xu, Yuze Guo, et al. 2026. Percept-wam: Perception-enhanced world-awareness-action model for robust end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10642–10655.

[37] Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 conference on empirical methods in natural language processing*. 8154–8173.

[38] Junjie He, Junfeng Li, Zhide Zhong, Haodong Yan, Ruixin Li, Yangyang Zheng, Jiaguan Zhu, Tianran Zhang, Yuqiao Du, Wen Chen, et al. 2026. SG-WAM: Text-Grounded and Spatial-aware Semantic Guidance for World-Action Models. *arXiv preprint arXiv:2608.08839* (2026).

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

[40] Senkang Hu, Zhengru Fang, Yihang Tao, Zihan Fang, Yiqin Deng, and Yuguang Fang. 2026. Multi-Agent Embodied Autonomous Driving (MAEAD): From V2X Information Exchange to Shared World Models. arXiv:2606.13840 [cs.RO] https://arxiv.org/abs/2606.13840

[41] Renhao Huang, Hao Xue, Maurice Pagnucco, Flora D Salim, and Yang Song. 2025. Vision-based multi-future trajectory prediction: A survey. *IEEE Transactions on Neural Networks and Learning Systems* 36, 8 (2025), 13691–13708.

[42] Zhao Huang, Bowen Zhang, Zhongzhu Li, and Di Lin. 2026. Occdriver: Future occupancy guided dual-branch trajectory planner in autonomous driving. In *International Conference on Learning Representations*, Vol. 2026. 141940–141962.

[43] Santosh Jaiswal. 2026. Geographic Diversity Beats Data Volume for Cross-Domain Generalization in Zero-Label JEPA Driving World Models. arXiv:2607.04500 [cs.CV] https://arxiv.org/abs/2607.04500

[44] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. 2025. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705* (2025).

[45] Fanfan Ji, Yunpeng Chen, Luoqi Liu, and Xiao-Tong Yuan. 2023. Cross-domain few-shot classification via dense-sparse-dense regularization. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 3 (2023), 1352–1363.

[46] Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. 2023. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549* (2023).

[47] Bo Jiang, Shaoyu Chen, Hao Gao, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. 2026. VADv2: End-to-end vectorized autonomous driving via probabilistic planning. In *International Conference on Learning Representations*, Vol. 2026. 68886–68900.

[48] Jindong Jiang, Fei Deng, Gautam Singh, Minseung Lee, and Sungjin Ahn. 2024. Slot state space models. *Advances in Neural Information Processing Systems* 37 (2024), 11602–11633.

[49] Li Jin and Liu Jia. 2025. Embodied World Models Emerge from Navigational Task in Open-Ended Environments. arXiv:2504.11419 [cs.AI] https://arxiv.org/abs/2504.11419

[50] Usman M Khan. 2026. Depth-Regularized JEPA World Models Learn More Transferable Representations from Real Outdoor Robot Data. *arXiv preprint arXiv:2607.16314* (2026).

[51] Bernard Lange, Masha Itkina, Jiachen Li, and Mykel J Kochenderfer. 2024. Self-supervised multi-future occupancy forecasting for autonomous driving. *arXiv preprint arXiv:2407.21126* (2024).

[52] Seungjae Lee, Yoonkyo Jung, Inkook Chun, Yao-Chih Lee, Zikui Cai, Hongjia Huang, Aayush Talreja, Tan Dat Dao, Yongyuan Liang, Jia-Bin Huang, et al. 2025. TraceGen: World Modeling in 3D Trace Space Enables Learning from Cross-Embodiment Videos. *arXiv preprint arXiv:2511.21690* (2025).

[53] Anson Lei, Bernhard Schölkopf, and Ingmar Posner. 2026. Spartan: A sparse transformer world model attending to what matters. *Advances in Neural Information Processing Systems* 38 (2026), 154089–154114.

[54] Jiewu Leng, Jiajun Wang, Lianhong Zhou, Rongli Zhao, Chong Chen, Ding Zhang, Shuai Zheng, Qiang Liu, and Weiming Shen. 2026. A review of multi-modal deep learning towards agentic smart manufacturing. *Advanced Engineering Informatics* 70 (2026), 104117. doi:10.1016/j.aei.2025.104117

[55] Tomas Leroy-Stone. 2026. Dreaming Of Others: Latent Teammate Modeling In World Models For Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2605.31361* (2026).

[56] Baolu Li, Jingyu Qian, Rui Guo, Yilun Chen, Hanpeng Liu, Yuan Lin, Junhong Zhou, Ruixin Liu, Liu Yang, Yutong Zheng, et al. 2026. X-Foresight: A Joint Vision-Action Causal Forecasting Network via Predictive World Modeling. *arXiv preprint arXiv:2605.24892* (2026).

[57] Chenming Li, Daoan Zhang, Wenjian Huang, and Jianguo Zhang. 2023. Cross contrasting feature perturbation for domain generalization. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 1327–1337.

[58] Haiqing Li, Wenliang Zhong, Yinhao Wu, Hehuan Ma, Yuzhi Guo, Thao M Dang, and Junzhou Huang. 2026. Guidelines as Environments: A World Model Approach to Rule Following. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 16302–16318.

[59] Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. 2026. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. In *International Conference on Learning Representations*, Vol. 2026. 93136–93189.

[60] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. 2026. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998* (2026).

[61] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. 2025. Unified Video Action Model. arXiv:2503.00200 [cs.RO] https://arxiv.org/abs/2503.00200

[62] Shuangzhi Li, Junlong Shen, Lei Ma, and Xingyu Li. 2026. From dataset to real-world: general 3D object detection via generalized cross-domain few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 6415–6423.

[63] Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, and Animesh Garg. 2020. Causal discovery in physical systems from videos. *Advances in Neural Information Processing Systems* 33 (2020), 9180–9192.

[64] Zhuo Li, Weiran Wu, Yunlong Guo, Jian Sun, and Qing-Long Han. 2025. Embodied multi-agent systems: A review. *IEEE/CAA Journal of Automatica Sinica* 12, 6 (2025), 1095–1116.

[65] Alan Liang, Youquan Liu, Yu Yang, Dongyue Lu, Linfeng Li, Lingdong Kong, Huaici Zhao, and Wei Tsang Ooi. 2026. LiDARCrafter: Dynamic 4D world modeling from LiDAR sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 18406–18414.

[66] Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. 2024. How jepa avoids noisy features: The implicit bias of deep linear self distillation networks. *Advances in Neural Information Processing Systems* 37 (2024), 91300–91336.

[67] Fei Liu, Shichao Xie, Minghua Luo, Zedong Chu, Junjun Hu, Xiaolong Wu, and Mu Xu. 2026. Navforesee: A unified vision-language world model for hierarchical planning and dual-horizon navigation prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 32431–32440.

[68] Qiqi Liu, Huan Xu, Jingyu Li, Bin Sun, Zhihui Hao, Dangen She, Xiatian Zhu, and Li Zhang. 2026. Uni-world vla: Interleaved world modeling and planning for autonomous driving. *arXiv preprint arXiv:2603.27287* (2026).

[69] Yuejiang Liu, Fan Feng, Lingjing Kong, Weifeng Lu, Jinzhou Tang, Kun Zhang, Kevin Murphy, Chelsea Finn, and Yilun Du. 2026. World Action Verifier: Self-Improving World Models via Forward-Inverse Asymmetry. arXiv:2604.01985 [cs.LG] https://arxiv.org/abs/2604.01985

[70] Yushan Liu, Peibo Sun, Xintao Chao, Zhenyang Yang, Yifan Xie, Lingfeng Zhang, Shoujie Li, Chenyu Tang, Fang Chen, Xiao-Ping Zhang, et al. 2026. CheckVLA: Execution-Time Verification with Action-Conditioned World Model for Long-Horizon Mobile Manipulation. *arXiv preprint arXiv:2607.26789* (2026).

[71] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. 2020. Object-centric learning with slot attention. *Advances in neural information processing systems* 33 (2020), 11525–11538.

[72] Jiangran Lyu, Ziming Li, Xuesong Shi, Chaoyi Xu, Yizhou Wang, and He Wang. 2025. Dywa: Dynamics-adaptive world action model for generalizable non-prehensile manipulation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 11058–11068.

[73] Minghui Ma, Bin Guo, Han Wang, Mengqi Chen, Jingqi Liu, Yan Liu, and Zhiwen Yu. 2026. Cognitive World Models for Process-Level Social Influence Evaluation. *arXiv preprint arXiv:2606.29495* (2026).

[74] Ivan Malashin, Vadim Tynchenko, and Dmitry Martysyuk. 2026. From Pixels to Plans: A Review of Latent-Space Future Prediction. *Archives of Computational Methods in Engineering* (2026), 1–29.

[75] Yanghong Mei, Longteng Guo, Ming-Ming Yu, Guiyu Zhao, Xingjian He, and Jing Liu. 2026. NavWM: A Unified Navigation World Model for Foresight-Driven Planning. *arXiv preprint arXiv:2606.24101* (2026).

[76] Kechen Meng, Rongpeng Li, Yansha Deng, Zhifeng Zhao, and Honggang Zhang. 2026. NetWorld: Communication-Based Diffusion World Model for Multi-Agent Reinforcement Learning in Wireless Networks. arXiv:2602.00558 [cs.NI] https://arxiv.org/abs/2602.00558

[77] Jihun Moon and Seong-Woo Kim. 2026. SPREAD: Scalable Pre-Trained World Model for Adaptive Dynamics Model. *IEEE Robotics and Automation Letters* (2026).

[78] Yosuke Nishimoto and Takashi Matsubara. 2026. Object-centric world models for causality-aware reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 24585–24593.

[79] Qiuling Pan, Hong Zhou, and Zhouchen Lin. [n. d.]. Opinion: How Can Causal AI Benefit World Models?. In *NeurIPS 2025 Workshop on Embodied World Models for Decision Making*.

[80] Francesco Petri, Luigi Asprino, and Aldo Gangemi. 2025. Learning local causal world models with state space models and attention. *arXiv preprint arXiv:2505.02074* (2025).

[81] Kajal Pokharel, Hepeng Li, and Haibo He. 2026. World-Model-Based Deep Reinforcement Learning for Real-Time Microgrid Energy Management. *IEEE Transactions on Smart Grid* (2026).

[82] Roopesh Kumar Polaganga and Qilian Liang. 2024. Extending causal discovery to live 5G NR network with novel proportional fair scheduler enhancements. *IEEE Internet of Things Journal* 12, 1 (2024), 288–296.

[83] Ziteng Qiao, Dianxi Shi, Songchang Jin, Yanyan Shi, Luoxi Jing, and Chunping Qiu. 2025. Enhancing cross-domain generalization by fusing language-guided feature remapping. *Information Fusion* 119 (2025), 103029.

[84] Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. 2024. From causal to concept-based representation learning. *Advances in Neural Information Processing Systems* 37 (2024), 101250–101296.

[85] Baochang Ren, Yunzhi Yao, Rui Sun, Shuofei Qiao, Ningyu Zhang, and Huajun Chen. 2026. Aligning agentic world models via knowledgeable experience learning. *arXiv preprint arXiv:2601.13247* (2026).

[86] Zhongwei Ren, Yunchao Wei, Xiao Yu, Guixun Luo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. 2026. Videoworld 2: Learning transferable knowledge from real-world videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 40569–40580.

[87] Farhad Rezazadeh, Amir Ashtari Gargari, Hatim Chergui, Sandra Lagen, Merouane Debbah, Houbing Song, and Lingjia Liu. 2026. Agentic World Modeling for 6G: Near-Real-Time Generative State-Space Reasoning. arXiv:2511.02748 [cs.NI] https://arxiv.org/abs/2511.02748

[88] Jonathan Richens and Tom Everitt. 2024. Robust agents learn causal world models. In *International Conference on Learning Representations*, Vol. 2024. 15786–15817.

[89] Tankred Saanum, Peter Dayan, and Eric Schulz. 2024. Simplifying latent dynamics with softly state-invariant world models. *Advances in Neural Information Processing Systems* 37 (2024), 38355–38382.

[90] Qiuhong Shen, Shihua Zhang, Yue Liao, Qi Li, Zhenxiong Tan, Shizun Wang, Shuicheng Yan, and Xinchao Wang. 2026. World Action Models: A Survey. *arXiv preprint arXiv:2606.20781* (2026).

[91] Chen Shi, Shaoshuai Shi, Kehua Sheng, Bo Zhang, and Li Jiang. 2025. Drivex: Omni scene modeling for learning generalizable world knowledge in autonomous driving. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 28599–28609.

[92] Zifeng Shi, Meiqin Liu, Senlin Zhang, Ronghao Zheng, Shanling Dong, and Ping Wei. 2025. Gawm: Global-aware world model for multi-agent reinforcement learning. *arXiv preprint arXiv:2501.10116* (2025).

[93] Shayan Shokri. 2026. TERRA: Task-Embedded Reasoning and Representation Architecture for Cross-Domain Applications. *arXiv preprint arXiv:2606.01520* (2026).

[94] Sebastian Stapf, Pablo Acuaviva, Aram Davtyan, and Paolo Favaro. 2026. Composition of memory experts for diffusion world models. In *International Conference on Learning Representations*, Vol. 2026. 15377–15403.

[95] Huiqiang Sun, Zhan Peng, Size Wu, Kun Wang, Kang Liao, Dianyi Wang, Xingyu Zeng, Sheng Jin, Yangguang Li, Zhiguo Cao, et al. 2026. Prisma-World: Camera-Controllable Multi-Agent Video World Model. *arXiv preprint arXiv:2606.09507* (2026).

[96] Ruijin Sun, Nan Cheng, Changle Li, Wei Quan, Haibo Zhou, Ying Wang, Wei Zhang, and Xuemin Shen. 2025. A comprehensive survey of knowledge-driven deep learning for intelligent wireless network optimization in 6G. *IEEE Communications Surveys & Tutorials* (2025).

[97] Gianluca Tabella, Mohammed Ayalew Belay, Ismael Viejo, María Herrando, and Pierluigi Salvo Rossi. 2025. Failure Prediction in Manufacturing Processes via Kullback-Leibler Divergence. *IEEE Sensors Letters* (2025).

[98] Toru Takahashi. 2026. Toward AI That Understands Self and Others: A World-Model Theory of Cognitive Diversity and Alignment. *arXiv preprint arXiv:2605.29930* (2026).

[99] Yinzhou Tang, Yu Shang, Yinuo Chen, Bingwen Wei, Xin Zhang, Shu'ang Yu, Liangzhi Shi, Chao Yu, Chen Gao, Wei Wu, et al. 2026. Roboscape-r: Unified reward-observation world models for generalizable robotics training via rl. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1472–1481.

[100] MotuBrain Team, Chendong Xiang, Fan Bao, Haitian Liu, Hengkai Tan, Hongzhe Bi, James Li, Jiabao Liu, Jingrui Pang, Kiro Jing, et al. 2026. Motubrain: An advanced world action model for robot control. *arXiv preprint arXiv:2604.27792* (2026).

[101] Thai-Hoc Vu, Senthil Kumar Jagatheesaperumal, Minh-Duong Nguyen, Nguyen Van Huynh, Sunghwan Kim, and Quoc-Viet Pham. 2024. Applications of generative AI (GAI) for mobile and wireless networking: A survey. *IEEE Internet of Things Journal* 12, 2 (2024), 1266–1290.

[102] Chi Wan, Kangrui Wang, Yuan Si, Pingyue Zhang, and Manling Li. 2026. Worldagen: Unified state-action prediction with test-time world model training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 18584–18592.

[103] Qi Wang, Zhipeng Zhang, Baao Xie, Xin Jin, Yunbo Wang, Shiyu Wang, Liaomo Zheng, Xiaokang Yang, and Wenjun Zeng. 2025. Disentangled world models: Learning to transfer semantic knowledge from distracting videos for reinforcement learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2599–2608.

[104] Rui Wang, Yue Zhang, Jiehong Lin, Kuncheng Luo, Jianan Wang, Zhongrui Wang, and Xiaojuan Qi. 2026. When to trust imagination: Adaptive action execution for world action models. *arXiv preprint arXiv:2605.06222* (2026).

[105] Siyin Wang, Junhao Shi, Zhaoyang Fu, Xinzhe He, Feihong Liu, Chenchen Yang, Yikang Zhou, Zhaoye Fei, Jingjing Gong, Jinlan Fu, et al. 2026. World action models: The next frontier in embodied ai. *arXiv preprint arXiv:2605.12090* (2026).

[106] Yuqi Wang, Ke Cheng, Jiawei He, Qitai Wang, Hengchen Dai, Yuntao Chen, Fei Xia, and Zhaoxiang Zhang. 2024. Drivingdojo dataset: Advancing interactive and knowledge-enriched driving world model. *Advances in Neural Information Processing Systems* 37 (2024), 13020–13034.

[107] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. 2024. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14749–14759.

[108] Haoyu Wu, Jiwen Yu, Yingtian Zou, and Xihui Liu. 2026. Multiworld: Scalable multi-agent multi-view video world models. *arXiv preprint arXiv:2604.18564* (2026).

[109] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. 2026. Video world models with long-term spatial memory. *Advances in Neural Information Processing Systems* 38 (2026), 49371–49393.

[110] Jiannan Xiang, Yi Gu, Zihan Liu, Zeyu Feng, Qiyue Gao, Yiyan Hu, Benhao Huang, Guangyi Liu, Yichi Yang, Kun Zhou, et al. 2025. Pan: A world model for general, interactable, and long-horizon world simulation. *arXiv preprint arXiv:2511.09057* (2025).

[111] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. 2024. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455* (2024).

[112] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. 2026. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems* 38 (2026), 49632–49652.

[113] Eric Xing, Mingkai Deng, and Jinyu Hou. 2025. Critique of World Model. *arXiv preprint arXiv:2507.05169* (2025).

[114] Chenfei Xu, Huthaifa Ahmad, Yuya Okadome, Hiroshi Ishiguro, and Yutaka Nakamura. 2025. Action-Inclusive Multi-Future Prediction Using a Generative Model in Human-Related Scenes for Mobile Robots. *IEEE Access* (2025).

[115] Kaichen Xu, Yihang Du, Mianpeng Liu, Zimu Yu, and Xiaobo Sun. 2026. Causality-Induced Positional Encoding for Transformer-Based Representation Learning of Non-Sequential Features. *Advances in Neural Information Processing Systems* 38 (2026), 153284–153334.

[116] Kai Xu, Hang Zhao, Ruizhen Hu, Yuhang Huang, Ziqiao Zhou, Wancheng Feng, Li Yi, Sida Peng, Xing Liu, Zihao Liu, et al. 2026. From specialist to generalist: A comprehensive survey on world models. (2026).

[117] Tianyi Yan, Huan Zheng, Dubing Chen, Meizhi Qu, Yingying Shen, Lijun Zhou, Mingfei Tu, Bing Wang, Guang Chen, Hangjun Ye, et al. 2026. CausalDrive: Real-time causal world models for autonomous driving. In *European Conference on Computer Vision*. Springer, 134–151.

[118] Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. 2023. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114* (2023).

[119] Zhanjiang Yang, Yueming Li, Yang Shen, Meng Li, and Lijun Sun. 2026. MAPF-world: Action world model for multi-agent path finding. *IEEE Robotics and Automation Letters* (2026).

[120] Zhenya Yang, Zhe Liu, Yuxiang Lu, Liping Hou, Chenxuan Miao, Siyi Peng, Bailan Feng, Xiang Bai, and Hengshuang Zhao. 2026. Geniedrive: Towards physics-aware driving world model with 4d occupancy guided video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 35680–35690.

[121] Xuan Yao, Junyu Gao, and Changsheng Xu. 2025. Navmorph: A self-evolving world model for vision-and-language navigation in continuous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5536–5546.

[122] Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. 2026. GigaWorld-Policy: An Efficient Action-Centered World–Action Model. *arXiv preprint arXiv:2603.17240* (2026).

[123] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. 2026. World Action Models are Zero-shot Policies. arXiv:2602.15922 [cs.RO] https://arxiv.org/abs/2602.15922

[124] Dianzhi Yu, Vireo Zhang, Hongru Wang, Yanyu Chen, Minda Hu, Wanghan Xu, Siki Chen, Philip Torr, Zhenfei Yin, and Irwin King. 2026. Dynamic Mixture of Latent Memories for Self-Evolving Agents. *arXiv preprint arXiv:2605.21951* (2026).

[125] Yi Yu and Tetsunari Inamura. 2026. Self-Evolving Cognitive Framework via Causal World Modeling for Embodied Scientific Intelligence. *arXiv preprint arXiv:2606.22449* (2026).

[126] Zhongwei Yu, Jingqing Ruan, and Dengpeng Xing. 2024. Learning causal dynamics models in object-oriented environments. *arXiv preprint arXiv:2405.12615* (2024).

[127] Ge Yuan, Qiyuan Qiao, Jing Zhang, and Dong Xu. 2026. AdaWorldPolicy: World-Model-Driven Diffusion Policy with Online Adaptive Learning for Robotic Manipulation. *arXiv preprint arXiv:2602.20057* (2026).

[128] Gholamreza Zare and P Malekpour Alamdari. 2025. Diffusion-Based Latent Intent Evolution for Anticipatory and Goal-Transition-Aware Recommendation.

[129] Xi Zeng, Haojie Ren, and Ziying Song. 2026. PhyLatent: Learning Dynamics-Relevant Representations for JEPA World Models. *arXiv preprint arXiv:2608.05720* (2026).

[130] Chongyang Zhang, Rendong Wang, Hao Zheng, Hanwen Zhang, Yang Liu, Xiaolong Wei, and Bin Chong. 2026. Physiological World Models for Human State Transitions. *arXiv preprint arXiv:2608.15309* (2026).

[131] Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Behzad Dariush, Kwonjoon Lee, Yilun Du, and Chuang Gan. 2025. Combo: Compositional world models for embodied multi-agent cooperation. In *International Conference on Learning Representations*,

Vol. 2025. 49996–50019.

[132] Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal Patel, Paul Liang, et al. 2026. World-in-world: World models in a closed-loop world. In *International Conference on Learning Representations*, Vol. 2026. 55660–55699.

[133] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, et al. 2026. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *Advances in Neural Information Processing Systems* 38 (2026), 24195–24228.

[134] Xiangcheng Zhang and Yilun Du. 2026. World action planner: Generalizable decision-making with action-conditioned world models. *arXiv preprint arXiv:2607.27599* (2026).

[135] Xiaoyuan Zhang, Chengdong Ma, Yizhe Huang, Weidong Huang, Siyuan Qi, Song-Chun Zhu, Xue Feng, and Yaodong Yang. 2026. World models should prioritize the unification of physical and social dynamics. *Advances in Neural Information Processing Systems* 38 (2026).

[136] Xiangrui Zhang, Chunyue Song, Wei Dai, Zheng Zhang, Kaihua Gao, and Furong Gao. 2026. Causal graph spatial–temporal autoencoder for reliable and interpretable process monitoring. *IEEE transactions on neural networks and learning systems* (2026).

[137] Xuan Zhang, Wenxuan Zhang, See-Kiong Ng, and Yang Deng. 2026. Self-Evolving World Models for LLM Agent Planning. *arXiv preprint arXiv:2606.30639* (2026).

[138] Yang Zhang, Chenjia Bai, Bin Zhao, Junchi Yan, Xiu Li, and Xuelong Li. 2025. Decentralized Transformers with Centralized Aggregation are Sample-Efficient Multi-Agent World Models. arXiv:2406.15836 [cs.LG] https://arxiv.org/abs/2406.15836

[139] Yiyao Zhang, Diksha Goel, Hussain Ahmad, Shixun Huang, and Jun Shen. 2026. CausalNav: Reliability-Certified Causal World Models for Control under Physical-Parameter Shift. *arXiv preprint arXiv:2608.07809* (2026).

[140] Yang Zhang, Xinran Li, Jianing Ye, Shuang Qiu, Delin Qu, Xiu Li, Chongjie Zhang, and Chenjia Bai. 2026. Revisiting multi-agent world modeling from a diffusion-inspired perspective. *Advances in Neural Information Processing Systems* 38 (2026), 65439–65470.

[141] Changyuan Zhao, Guangyuan Liu, Ruichen Zhang, Yinqiu Liu, Jiacheng Wang, Jiawen Kang, Dusit Niyato, Zan Li, Xuemin Shen, Zhu Han, et al. 2026. Edge general intelligence through world models, large language models, and agentic AI: Fundamentals, solutions, and challenges. *IEEE Transactions on Cognitive Communications and Networking* (2026).

[142] Changyuan Zhao, Ruichen Zhang, Jiacheng Wang, Gaosheng Zhao, Dusit Niyato, Geng Sun, Shiwen Mao, and Dong In Kim. 2026. World models for cognitive agents: Transforming edge intelligence in future networks. *IEEE Communications Magazine* (2026).

[143] Zhida Zhao, Talas Fu, Yifan Wang, Lijun Wang, and Huchuan Lu. 2026. From forecasting to planning: Policy world model for collaborative state-action prediction. *Advances in Neural Information Processing Systems* 38 (2026), 134585–134611.

[144] Zijie Zhao, Zhongyue Zhao, Kaixuan Xu, Yuqian Fu, Jiajun Chai, Yuanheng Zhu, and Dongbin Zhao. 2026. Learning and planning multi-agent tasks via an moe-based world model. *Advances in Neural Information Processing Systems* 38 (2026), 40857–40890.

[145] Congjie Zheng, Chuanyi Xue, Bin Liang, Jun Yang, and Changshui Zhang. 2026. SEAGym: An Evaluation Environment for Self-Evolving LLM Agents. *arXiv preprint arXiv:2606.17546* (2026).

[146] Jie Zheng, Dusit Niyato, Ruichen Zhang, Jiacheng Wang, Jiangtian Nie, Hongyang Du, Jiawen Kang, Haijun Zhang, Abbas Jamalipour, and Dong In Kim. 2026. Advanced Deep Reinforcement Learning for Agentic AI and Their Applications in Wireless Network. *IEEE Transactions on Cognitive Communications and Networking* (2026).

[147] Lei Zheng, Luyao Zhang, Peiqi Yu, Yifan Sun, Sergio Grammatico, Jun Ma, and Changliu Liu. 2026. Contingency Planning for Safety-Critical Autonomous Vehicles: A Review and Perspectives. *arXiv preprint arXiv:2601.14880* (2026).

[148] Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, et al. 2025. World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 28632–28642.

[149] Chen Zhou, Ghassan AlRegib, Armin Parchami, and Kunjan Singh. 2024. TrajPRed: Trajectory prediction with region-based relation learning. *IEEE Transactions on Intelligent Transportation Systems* 25, 8 (2024), 9787–9796.

[150] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. 2025. Unified World Models: Coupling Video and Action Diffusion for Pretraining on Large Robotic Datasets. arXiv:2504.02792 [cs.RO] https://arxiv.org/abs/2504.02792

[151] Fangqi Zhu, Zhengyang Yan, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. 2026. Wmpo: World model-based policy optimization for vision-language-action models. In *International Conference on Learning Representations*, Vol. 2026. 62486–62502.

[152] Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jiwen Lu, and Jie Zhou. 2026. Astra: General interactive world model with autoregressive denoising. In *International Conference on Learning Representations*, Vol. 2026. 79167–79184.

[153] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Bohan Li, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, et al. 2024. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520* (2024).

[154] Ziyue Zhu, Zhanqian Wu, Zhenxin Zhu, Lijun Zhou, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, Jin Xie, et al. 2026. Worldsplat: Gaussian-centric feed-forward 4d scene generation for autonomous driving. In *International Conference on Learning Representations*, Vol. 2026. 68258–68277.

[155] Shilong Zou, Shilin Zhang, Yingji Zhang, Yuhang Huang, Yi Zhang, Zeyuan Ding, Han Dong, Junwei Liao, Yong Dai, Jian Tang, et al. 2026. Pelican-Sim 1.0: A General World Model Simulator for Embodied Intelligence. *arXiv preprint arXiv:2609.12036* (2026).