生数 GensPI

General World Models from First-Principles

Jun Zhu, Hengkai Tan, Jintao Zhang, Min Zhao, Fan Bao, Bo Zhang

GensPI Technology, Tsinghua University

Abstract

World models now span latent simulators, video generators, interactive environments, and robot policies, but the field lacks a shared definition of what makes such a model *general*. This article develops a first-principles account of the general world model (GWM) as a single foundation that integrates three inseparable faculties at scale: **Understanding** the present world, **Imagining** possible and action-conditioned futures, and **Acting** to realize goals while incorporating feedback. From this closed loop, we derive a five-level roadmap of increasing agency: World Generation, Interactive World, Actionable World, Autonomous World Agent, and World Orchestrator. To construct a GWM, we organize around three requirements. First, a five-layer data pyramid combines the breadth of web, curated, and egocentric video with action-recording human demonstrations and real-robot experience. Synthetic data augments all five layers, while imperfect action-aligned trajectories provide evidence about intervention effects, failure, and recovery. Second, as one current architectural instantiation, a Mixture-of-Transformers design specializes parameters for language, video, and action while sharing joint attention, enabling understanding, future prediction, and action to exchange information within one foundation. Third, scalable training and accelerated inference determine whether this foundation can progress from offline generation to real-time interaction and physical control. We further present current systems to illustrate the first three roadmap levels: Vidu implements large-scale world generation; Vidu S1 adds continuous, voice-conditioned interaction; Motus unifies understanding, future-state prediction, and action learning; and Motubrain scales this loop toward deployable cross-embodiment control. These examples support a common development path across digital and physical domains, showing promise to ultimately achieve general embodied intelligence, but do not yet establish the autonomy or orchestration required by L4 and L5. Finally, we discuss directions toward these upper levels, including joint evaluation, causal and physical grounding, persistent memory, online self-improvement, efficient closed-loop deployment, and safe, controllable autonomy.

Keywords: world models, embodied intelligence, video generation, foundation models, multimodal architectures, robot learning, first-principles reasoning.

1 Introduction

World models have become central to artificial intelligence because prediction changes the nature of behavior [1]. A system that only recognizes the present can select a familiar response; while a system that models how the world evolves can compare possible futures, anticipate the consequences of intervention, and revise its behavior when observation differs from expectation. This idea connects predictive coding and active inference in neuroscience [2–4], internal forward models of motor control [5], latent model-based agents [6–8], and proposals for autonomous machine intelligence [9]. Across these traditions, prediction is not an auxiliary task: it is the mechanism that connects perception to planning and action.

Yet "world model" now names systems with very different capabilities. Latent world models learn compact dynamics for control, often within defined environments [6–8]. Large video generators produce high-fidelity visual futures but need not expose causal structure or physical actions [10]. Interactive generative environments and generated simulators produce world trajectories that respond to user or agent inputs, yet their control interfaces are often tied to a particular environment and need not correspond to actions

executable by a physical embodiment [11–13]. Vision-Language-Action (VLA) policies map observations and instructions directly to robot commands, while explicit future-state prediction is optional [14–16]. Each approach captures part of the problem. What remains unclear is what must be combined for a single foundation to operate generally across digital and physical worlds.

We address this question from first principles, taking the human world model as a functional inspiration. Humans do not understand, predict, and act through isolated faculties: accumulated experience supports an internal representation of the present, imagination anticipates how the world may evolve under possible actions, and interaction tests and continually refines these expectations [1, 5]. Following this organization, a *general world model* (GWM) must integrate three inseparable faculties in one single foundation. It must **Understand** the present by constructing a grounded, task-relevant state from multimodal observations; **Imagine** possible futures, including the consequences of candidate actions; and **Act** to realize goals while using feedback to update its predictions and decisions. Imagination is operationalized through prediction, but it is broader than next-frame forecasting: it includes uncertain, counterfactual, and action-conditioned futures. Action, likewise, is more than an output head; it grounds the model's internal state in causal interaction. A GWM is therefore defined by a closed Understanding–Imagination–Action loop rather than by any single representation or generative objective.

To develop GWMs, we propose a five-level roadmap of increasing agency. **L1: World Generation** learns to produce coherent world trajectories. **L2: Interactive World** makes those trajectories respond continuously to external input. **L3: Actionable World** connects predicted change to actions that intervene in a physical environment. **L4: Autonomous World Agent** adds self-directed exploration, intermediate goal formation, and continual improvement. **L5: World Orchestrator** extends agency from one model to the coordination of multiple robots, digital agents, tools, resources, and people. The progression is thus **generation** $\rightarrow$ **interaction** $\rightarrow$ **action** $\rightarrow$ **autonomy** $\rightarrow$ **organization**. It separates capabilities that are sometimes conflated and makes clear which stages current systems have and have not reached.

The roadmap identifies three requirements for construction. The first is *data*. We organize real-world experience as a five-layer pyramid, moving from web-scale and curated videos through egocentric observations and action-recording human demonstrations to real-robot trajectories. Synthetic data acts across these layers rather than occupying a fixed level, while imperfect action-aligned trajectories [17, 18] complement successful demonstrations with failure and recovery evidence. The lower layers provide breadth, while the upper layers provide progressively stronger intervention and embodiment grounding. Inverse dynamics [19], latent predictive learning [20, 21], and generative modeling [22] offer complementary ways to recover structures from the massive scale of action-free videos to support large-scale pre-training. This realizes a machine form of *learning by seeing* [23, 24]: abundant observation builds broad world knowledge, and scarce interaction data grounds that knowledge in a particular body and action space. The second requirement is *architecture*. Understanding, future prediction, and action depend on shared knowledge of objects, dynamics, goals, and history, yet language, video, and continuous actions have different statistical and computational requirements. We therefore study a Mixture-of-Transformers (MoT) design [25] that specializes parameters by modality while sharing joint attention. This creates a common computational context without forcing every token through identical transformations, enabling the same foundation to decode pixels, actions, or both. The third requirement is *compute*. Foundation-scale video training demands distributed and memory-efficient optimization, while interactive generation and robot control require few-step sampling, efficient attention, streaming state, and latency-aware action execution. Compute therefore determines not only model scale, but whether a learned world model can remain inside a real-time feedback loop.

Finally, we present our current practice to illustrate the first three roadmap levels. **Vidu** implements L1 through large-scale video generation [26]. **Vidu S1** advances to L2 through continuous, voice-conditioned real-time generation [27]. At L3, **Motus** unifies understanding, world prediction, and action learning [28], while **Motubrain** scales the same world-action loop toward long-horizon, multiview, and cross-embodiment robot control [29]. These systems are supporting examples, not evidence that the roadmap is complete. In particular, the L3 systems operate under externally specified goals and bounded task distributions; none of the examples demonstrates the open-ended exploration and continual self-improvement required by L4 or the multi-agent organization required by L5. L1–L3 themselves also remain active development targets in quality,

efficiency, and controllability.

In summary, this article makes four contributions. First, it defines a GWM through the Understanding–Imagination–Action triad and formalizes the corresponding closed loop. Second, it proposes a five-level roadmap that distinguishes world generation, interaction, physical action, autonomy, and orchestration. Third, it develops a construction program organized around the data pyramid, a current MoT-based unification design, and scalable training and inference. Fourth, it connects this program to current systems, clarifies the boundary between today's L1–L3 examples and the still-open L4–L5 capabilities, and situates the framework among latent world models, predictive representation learning, generative simulators, VLAs, world-action models, and unified multimodal foundations, shedding light on the development of dependable autonomous world agents at the upper levels of the roadmap.

2 General World Model and its Roadmap

We provide a formal definition of general world model and envision its development roadmap as well.

2.1 GWM as a Foundation of Understanding, Imagination and Action

The most reliable starting point is the world model humans already possess. Consider learning to drive a car or ride a bicycle. Before we acquire the skill, our movements are hesitant and unprepared; we overcorrect, we stall, we fall. But once the skill is internalized, something remarkable happens: the brain begins to *predict*. For every possible action, it anticipates the consequence and the state that will follow, and it feeds that anticipation back into the muscles. The result is motion that is safe, reliable, and smooth. What we call "skill" is, in large part, a well-trained internal model of how the world responds to us [1].

This everyday intuition is mirrored by what neuroscience has learned about the brain. A leading account holds that the brain is fundamentally a *prediction machine*: rather than passively registering sensory input, it continuously generates top-down predictions of incoming signals and transmits only the *prediction error*—the mismatch between expectation and reality—up the cortical hierarchy. This is the theory of predictive coding, first formalized in the visual cortex [2] and later generalized into the free-energy principle and active inference [3], under which perception and action are two sides of a single imperative: minimize surprise about the world. Andy Clark's synthesis [4] frames the brain, in exactly these terms, as an organ that models its environment in order to anticipate it. In motor control specifically, the cerebellum is understood to learn *internal forward models* [5] that predict the sensory consequences of our own movements—the computational substance of the "skill" described above.

Crucially, this model is not innate in full; it is *developed*. Developmental psychology shows that infants arrive equipped with a scaffolding of "core knowledge"—primitive expectations about objects, space, number, and agents—onto which experience is built [30]. From this scaffold, and through relentless sensorimotor interaction with their surroundings, children construct an increasingly accurate model of how the world behaves: that unsupported objects fall, that hidden objects persist, that actions produce predictable effects [31]. In other words, the human world model is *bootstrapped from broad perceptual experience first, then refined through embodied action*—a developmental order that, as later sections argue, maps directly onto the strategy of pretraining on vast video before grounding in scarce robot interaction.

One paradigm that deserves special emphasis is *learning by seeing*. Long before we act, we watch. A child who has never held a cup learns an astonishing amount about pouring simply by observing others do it—where the hand goes, how the liquid falls, what counts as success or spillage. Bandura's social-learning theory established that much of human competence is acquired through *observation* rather than trial-and-error reward: we build a model of an action by watching it performed, and only later execute it ourselves [23]. The neural machinery for this appears to run deep—mirror-neuron systems fire both when an agent performs an action and when it merely observes the same action in another, as if perception and the plan for action shared one representation [24]. Learning by seeing is thus not a shortcut but a primary channel of world-model acquisition: it lets an agent absorb the dynamics and consequences of countless actions it has never physically taken.

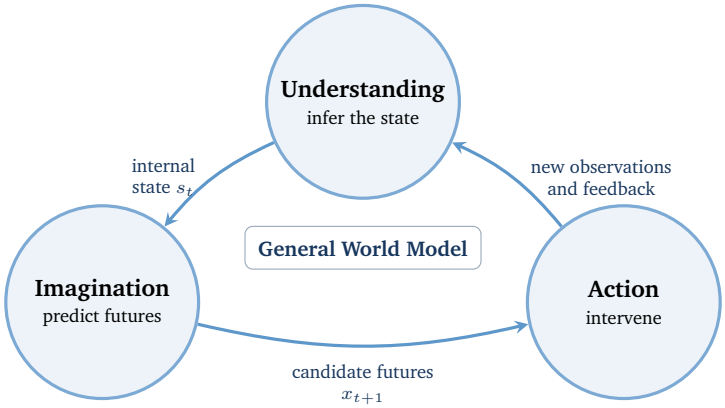

Figure 1 The three cores of a general world model form a closed loop. Understanding infers the current state, imagination predicts possible futures, and action changes the world, producing new evidence for understanding.

If these functions characterize the human world model, then first-principles reasoning suggests that a machine world model must integrate three corresponding capabilities:

First, it must understand the world. Understanding requires more than recognizing objects or describing a scene. The model must integrate visual, linguistic, auditory, and embodied observations into an internal state that captures what entities are present, how they are related, what has changed over time, and which aspects of the situation matter for the current goal. It must also preserve uncertainty when the evidence is incomplete rather than committing to a single unsupported interpretation. This internal state provides the grounded, task-relevant account of the present on which prediction and action depend.

Second, it must predict and imagine possible futures. From its internal state, the model must infer how the world may evolve both without intervention and under candidate actions. Because real environments are uncertain, imagination should represent multiple plausible outcomes rather than extrapolate a single deterministic trajectory. By comparing these counterfactual futures, the model can anticipate risks, evaluate whether an action advances a goal, and construct plans before committing to them in reality. Imagination thus turns an understanding of what *is* into a model of what *could be*.

Third, it must act in the world and learn from the consequences. Understanding and imagination become operational only when the model can translate a goal and its predicted futures into effective interventions. It must represent how candidate actions change the world, select the actions whose anticipated consequences advance the goal, execute them through a particular embodiment, and use the resulting feedback to correct subsequent predictions and decisions. Action is therefore not merely an output modality appended to a simulator; it grounds the model's knowledge in cause and effect and closes the loop between internal imagination and external reality. A system that can describe or generate the world but cannot deliberately change it remains incomplete as a general world model.

Taken together, **Understanding – Imagination – Action** form the essential triad of a general world model. Formally, a general world model can be described with a few variables. At time t, x_t is the model's observation of the world, a_t is an input that can affect the world, g is a condition or goal, and s_t is the model's internal summary of what it currently believes about the world. Here, an observation may contain pixels, audio, language, or robot sensors; an input may be a user's real-time control at the interactive stage or a robot action at the embodied stage. Both a_t and g are optional: a video generator may use g but no action, whereas an interactive stream may respond to a_t without an explicit goal.

Symbol	Meaning
x_t	World observation (*e.g.* vision, tactile, audio, and force) at time t
$x_{\leq t}$	Observation history up to time t
a_t	Optional interaction input or physical action at time t
g or g_t	Optional fixed or dynamically updated goal
s_t	Internal world state inferred from the history
m_t	Persistent memory accumulated up to time t
U	Understanding operator that infers the internal state
p	Conditional distribution over future observations
π	Policy, or action distribution, used to select an intervention

With this notation, the three cores form one simple closed loop:

$$\begin{aligned} \textbf{Understanding:} \quad & s_t = U(x_{\leq t}, [a_{<t}]), \\ \textbf{Imagination:} \quad & x_{t+1} \sim p(x_{t+1} \mid s_t, [a_t], [g]) \,, \\ \textbf{Action:} \quad & a_t \sim \pi(a_t \mid s_t, [g]). \end{aligned} \tag{1}$$

Here, $\sim$ means that an observation or action is sampled from the corresponding conditional distribution to account for intrinsic uncertainty, rather than being assumed to be deterministic. Square brackets $[\cdot]$ denote optional inputs. Thus, prediction may be conditioned on a goal, an interaction, both, or neither; the action branch is activated only when the model must intervene in the world. Understanding turns past experience into a useful internal state; prediction or imagination rolls that state forward into possible futures; and action chooses how to intervene so that a desired future becomes real. Concretely, the action model can propose candidate actions, while imagination predicts their consequences; imagined states then guide subsequent actions, which condition further prediction, forming recurrent rollouts toward a goal-aligned outcome. Once an action is executed, the resulting observation x_{t+1} updates s_{t+1} and corrects subsequent imagination and action, closing the loop summarized in Figure 1. As we shall see, the roadmap in Section 2.2 progressively strengthens this loop: from generating observations, to conditioning prediction on interaction, to acting, setting goals autonomously, and finally coordinating many actors and resources.

2.2 Roadmap of GWM

We envision the development of general world models as five levels of increasing agency: **World Generation** → **Interactive World** → **Actionable World** → **Autonomous World Agent** → **World Orchestrator**. The progression begins with models that generate coherent pixels and digital worlds, then closes a real-time interaction loop with users. Through World Action Models (WAMs), it next crosses from the digital domain into the physical world, by jointly predicting how the world will evolve and producing actions that change the environment. At the higher levels, the system shifts from executing human-specified goals to proactively perceiving, exploring, planning, and learning; it ultimately coordinates robots, digital agents, tools, and people to organize and shape both digital and physical worlds. In short, the central axis of the roadmap is **Pixel** → **Interaction** → **Action** → **Autonomy** → **Organization**.

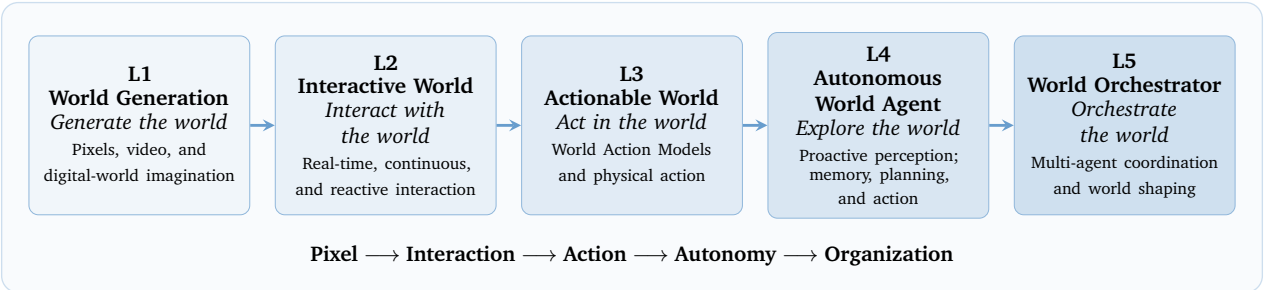

Table 1 Formal modeling objectives and learned capabilities for the five levels of the GWM roadmap.

Level	Modeling objective	Capability learned
L1	$p(x_{1:t} \mid g)$	Generate a coherent world trajectory from a condition or goal.
L2	$p(x_{t+1} \mid x_{\leq t}, a_{\leq t})$	Predict the world's next response under continuous interaction.
L3	$(a_t, x_{t+1}) \sim \pi(a_t \mid x_{\leq t}, g)p(x_{t+1} \mid x_{\leq t}, a_t, g)$	Produce task-driven actions that change the physical world.
L4	$g_t = f(g_{\text{long-term}}, s_t, m_t, h),\ a_t \sim \pi(a_t \mid x_{\leq t}, g_t, m_t)$	Form goals from memory m_t and human preferences h, then explore, plan, and learn proactively.
L5	$\pi : (s_t, g_t) \mapsto \{\text{plans}, \text{roles}, \text{actions}, \text{resources}\}$	Organize multiple robots, agents, people, tools, and resources.

The corresponding modeling objectives make the increase in agency explicit. **L1** primarily learns to imagine a world; **L2** makes the imagined world react continuously to external input; **L3** adds the capacity to act, turning the world model into a world action model that intervenes in the environment, observes the resulting feedback, and closes the understanding–prediction–action loop. **L4** then moves from reactive execution of human-specified tasks to proactive behavior: the agent forms immediate goals from its world state, decides what to observe or learn, and may seek help or recruit another robot. For example, it could notice an unreported puddle and clean it before someone slips, or investigate an experimental result that contradicts its prediction. **L5** further expands this autonomy from a single agent to system-level organization. Rather than selecting only its own actions, it allocates roles, tools, information, and resources across people, robots, and digital agents—for example, dynamically reorganizing a factory as equipment, supplies, and priorities change.

Table 1 provides a formal definition of each level. In the L4 objective, $g_{\text{long-term}}$ denotes the externally specified long-term objective, and h denotes human preferences or constraints. The function f is the agent's goal-formation mechanism, which converts these inputs and the current internal state s_t into an immediate goal g_t. In L5, π is generalized from an individual action policy to an orchestration policy whose outputs may include plans, roles, actions, and resource allocations.

Language models across the roadmap. We would like to remark that language is an integral component of a GWM rather than a separate development axis. At L1, a text encoder or large language model (LLM) can provide semantic conditions for world generation; at L2, language becomes a natural channel for continuously instructing and interacting with the generated world; and at L3, a vision–language or understanding expert interprets objects, relations, instructions, and goals and connects them to future prediction and physical action. Its role expands further at L4, where autonomous agents require higher-level reasoning, planning, memory, and tool use, and at L5, where world orchestration additionally requires communication, task decomposition, and coordination among robots, digital agents, and people. The long-term direction is therefore not for a language model to replace the GWM or GWM to replace LLMs, but for language understanding and reasoning to become increasingly fused with the GWM's perceptual state, imagination, action, and feedback loop as the roadmap advances toward agentic and organizational intelligence.

3 Building General World Models: Data, Architecture, and Compute

Turning the preceding definition and roadmap into a working system requires three elements to scale together: *data*, *architecture*, and *compute*. Data determines the breadth of world knowledge the model can acquire; architecture determines whether understanding, imagination, and action can share and transfer that knowledge; and compute determines whether the resulting system can be trained at foundation-model scale and deployed within the latency constraints of interaction and control. This section develops these

requirements and then traces their realization from video generation and real-time interaction to embodied action.

3.1 Data: Scaling Experience from Web Video to Robot Interaction

For a GWM, the central data problem is an asymmetry between scale and grounding. Web videos offer extraordinary breadth but rarely record the actions or forces that produced an observed transition; robot interactions provide those signals but are expensive, narrow, and tied to particular hardware. Understanding, imagination, and action therefore cannot be learned by collecting more samples of a single kind. They require a composition of data sources that vary in abundance, task relevance, action supervision, embodiment alignment, and proximity to deployment. A viable strategy combines broad observation at the base with progressively stronger intervention grounding toward the top, while retaining the ability to exploit increasing data and compute [32–34].

3.1.1 The Data Pyramid

We organize real-world training experience into a five-layer **data pyramid**, as highlighted in Figure 2. Moving upward generally strengthens task, action, and embodiment alignment, while reducing the amount and diversity of data available. Each layer contributes a different form of evidence to the understanding–imagination–action loop:

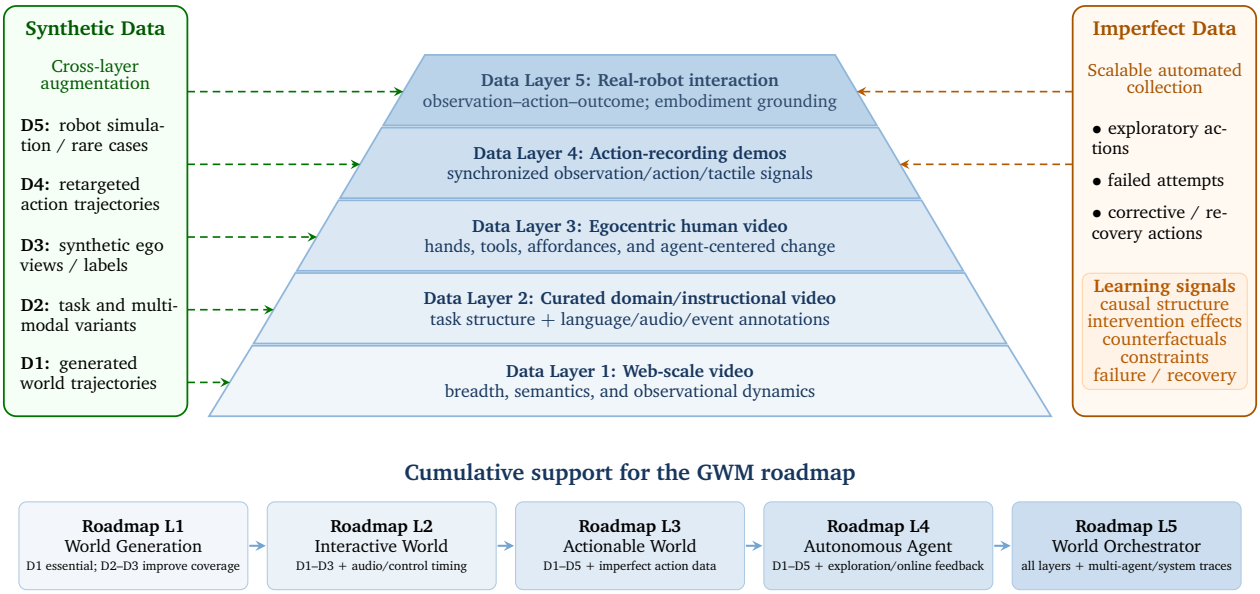

Figure 2 The GWM data pyramid and its relationship to the five-level capability roadmap. The center contains cumulative real-world experience. Synthetic data and imperfect data are parallel axes rather than additional tiers: synthetic data can augment every layer, whereas imperfect action-aligned trajectories are most directly associated with Data Layers 4–5. The bottom row summarizes the cumulative data support required as the roadmap advances.

- **Data Layer 1 — Web-scale video.** Broad, diverse footage supplies visual semantics, scene structure, motion patterns, human behavior, and recurring regularities of the physical and social world. Its primary contribution is broad *understanding* and an observational prior for *imagination*: what the world contains and how it commonly changes. It has the greatest scale but the weakest task and action supervision.
- **Data Layer 2 — Curated domain and instructional video.** Videos selected around activities such as cooking, assembly, tool use, navigation, and driving concentrate the events most relevant to downstream reasoning and control while retaining substantial diversity. This layer contributes task structure, temporal stages, object affordances, and examples of goal-directed behavior, sharpening *understanding* and making *imagination* more task relevant even when executable actions remain unobserved.

- **Data Layer 3 — Egocentric human video (Ego data).** First-person recordings of everyday and industrial work capture dense hand–object interaction, tool use, task structure, and physical dynamics, but generally lack robot-executable action labels. Ego data is a particularly scalable source of action-relevant experience because its agent-centered view approximates the human perception–action loop. It connects visual change to the behavior of an acting body, improving action-relevant *understanding* and first-person *imagination*. Examples include Ego4D, EgoDex, Egocentric-10K, and Egocentric-100K [35–38].
- **Data Layer 4 — Action-recording human demonstrations (Universal Manipulation Interface, or UMI, data).** This layer augments egocentric video with instrumented handheld or wearable devices that record actions synchronized with observations, including richer manipulation signals such as hand or gripper motion and tactile feedback. By better aligning the sensing and action representations used in data collection and robot deployment, UMI improves train–deployment consistency and bridges embodiment-free human demonstrations and physical execution. It thereby supports action-conditioned prediction, inverse dynamics, and action learning after embodiment adaptation. Examples include UMI, FastUMI, and the multi-finger DexUMI [39–41].
- **Data Layer 5 — Real-robot interaction.** Cross-embodiment corpora and target-specific experience provide synchronized observations, robot states, actions, and outcomes. Successful, instruction-aligned demonstrations directly supervise goal-conditioned *action*; imperfect or exploratory trajectories may not complete a specified task, but their aligned observation–action streams still reveal embodiment dynamics, action consequences, failure states, and recovery opportunities. This layer therefore grounds both action-conditioned *imagination* and executable *action*, while calibrating the model to real sensors, actuators, and control frequencies. Open X-Embodiment and AgiBot World illustrate large-scale aggregation of real-robot experience [42, 43]; such data is the most directly grounded but the most costly to collect.

Note that the pyramid is not a ranking in which one layer replaces another. The lower layers provide breadth: what objects, agents, environments, and events look like and how they commonly evolve. The middle layers increasingly expose task structure and the motions that produce change. The apex grounds these priors in the dynamics and action space of real machines. Importantly, task success is not a prerequisite for every action-aligned trajectory to be useful. Successful demonstrations teach which actions achieve an instruction, whereas imperfect interaction teaches what actions actually do, including ineffective motion, failure, and recovery. A GWM should therefore learn broad world knowledge from the base and use the scarce apex both to acquire goal-directed behavior and to correct its model of physical intervention.

Synthetic data complements all five layers rather than occupying a fixed level. Generative models and simulation can add controlled variations, rare or unsafe cases, privileged state, and aligned instruction–observation–action trajectories; systems such as MimicGen, for example, expand robot demonstrations across task configurations [44]. Its main contribution is scalable and controllable coverage, not real-world grounding: synthetic dynamics and action effects may deviate from deployment and should therefore be calibrated or filtered with real interaction.

3.1.2 Learning from Action-Free Videos

Action-free videos deserve special emphasis because they are available at massive scale and are extremely valuable: temporal observation contains substantially more than appearance. Across frames, a model can learn object persistence, three-dimensional and temporal structure, motion, agent behavior, and statistical regularities of interaction. However, these signals do not by themselves identify the true action, force, intention, or causal mechanism behind every transition; those quantities may be hidden or ambiguous. The key technical question is therefore how to extract the predictive structures that videos do contain and connect them to the action-grounded layers above.

Three complementary learning routes aim to address this question. First, *inverse dynamics* asks what motion or action could explain an observed transition. Video PreTraining (VPT) showed that an inverse-dynamics model learned from a relatively small labeled set can pseudo-label a much larger collection of unlabeled video [19]. In a general video collection, however, such an inferred action may describe only a latent mode of change; it is

not automatically a command that a particular robot can execute. Second, *latent predictive learning* trains a model to predict missing or future representations rather than every pixel. The Video Joint-Embedding Predictive Architecture (V-JEPA) demonstrates that masked spatiotemporal prediction can yield transferable representations of motion and interaction without action labels [20]. Third, *generative modeling* asks the model to reproduce or forecast the observable world in detail. Diffusion objectives [45] and scalable transformer backbones such as the U-shaped Vision Transformer (U-ViT) [22] make this objective practical at large scale, forcing the model to capture appearance, temporal consistency, and multimodal future uncertainty.

We emphasize that these routes should not be treated as mutually exclusive. Inverse dynamics adds approximate action semantics, latent prediction emphasizes compact and predictable structure, and pixel generation preserves detailed futures that can be inspected or used for interaction. Together they implement the machine analogue of *learning by seeing* [23, 24]. The higher layers of the pyramid then resolve what passive observation cannot. Videos can show *what changes are possible or likely*, but they cannot by themselves determine *which changes a particular body can cause*. Action-aligned human and robot data bind latent changes and generated futures to actual motion, commands, outcomes, and feedback. The strategy is therefore not to replace robot experience with videos, but to let scarce robot experience ground a broad world model rather than building one from scratch.

3.1.3 From Observed Change to Grounded Action

Moving from prediction to action requires answering three progressively stronger questions: What change followed an action? Which changes can this body reliably cause? Which action should it choose for a goal? Different forms of action-aligned data answer different parts of this progression.

Synchronized observation–action data remains useful even without a reliable instruction or successful task completion. A robot may simply explore, repeat an ineffective motion, or move without reaching a recognizable goal. If its observations and executed actions are aligned in time, the trajectory still teaches forward dynamics (i.e., what follows an action) and inverse dynamics (i.e., what action could have produced an observed change). Across many such transitions, the model also learns *controllability*: which parts of the future depend on the robot and which arise from the rest of the world.

Instruction-conditioned failures add another kind of evidence. The instruction states what was intended, while the unsuccessful trajectory reveals actions or states that do not achieve it. Such data can teach failure prediction, physical or task constraints, and recovery behavior rather than only successful imitation. Finally, **instruction-aligned successful demonstrations** connect a goal to actions that realize it and provide the strongest supervision for a goal-conditioned policy.

Imperfect trajectories are therefore weak supervision for copying successful behavior, but strong supervision for learning action consequences, controllability, failure, and recovery. Together, the upper layers of the pyramid turn observational knowledge into action-conditioned imagination and then into grounded, goal-directed action.

3.2 Architecture: Unifying Understanding, Imagination, and Action

The data pyramid supplies complementary forms of experience, but an architecture is needed to make their knowledge transferable. A pipeline of independently trained perception, prediction, and action modules can be effective, yet its interfaces determine in advance what information may pass between faculties. A caption, discrete plan, or fixed latent state may omit uncertainty, geometry, or temporal detail that later becomes important for prediction or action. Separate modules may also relearn overlapping structure and make it difficult for abundant video data to influence an action policy trained from comparatively few trajectories. The architectural objective is therefore not simply to place three modules next to one another, but to give the three faculties a shared computational context. At the same time, unification should not imply that every modality uses identical parameters. Language tokens, video features, and continuous action vectors differ in dimensionality, statistics, and training objective. Complete parameter sharing can create competition among modalities, whereas complete separation prevents useful transfer. The required design principle is therefore *specialize where representations differ and share where reasoning must interact*.

3.2.1 Bridging Language and Heterogeneous World Representations

The preceding principle becomes concrete when language is combined with the heterogeneous signals of an embodied world. Dual-coding theory provides a useful cognitive analogy: human cognition includes an interacting *verbal system* for linguistic information and a *nonverbal system* for perceptual objects, events, and imagery [46]. Correspondingly, a GWM can be viewed as coupling a *language-centric stream* with an *observation–action stream*. The language-centric stream represents concepts, instructions, goals, rules, and abstract reasoning through discrete tokens, commonly trained with token-level objectives. The observation–action stream encompasses vision, audio, force, touch, temperature, proprioception, and continuous actions, using predictive, diffusion, or flow-based objectives to represent world states, dynamics, affordances, and interventions. Their different statistics and objectives motivate specialized computation rather than indiscriminate parameter sharing.

Specialization, however, cannot become separation. The three GWM faculties require the two streams to refer to the same evolving world. Language helps Understanding identify task-relevant entities and relations, conditions Imagination on goals and counterfactual queries, and specifies the intent that Action should realize. Conversely, observations ground linguistic concepts in the current state; predicted futures give semantic plans physical consequences; and action feedback reveals whether the model's interpretation and plan were correct. Thus, language primarily contributes semantic abstraction and long-horizon reasoning, while observation and action primarily contribute grounding, dynamics, and control, but neither stream owns a faculty exclusively. Their interaction is what closes the Understanding–Imagination–Action loop.

This coupling must also accommodate different time scales. Semantic reasoning and long-horizon planning may evolve relatively slowly, whereas state estimation, future prediction, and physical control must track the environment continuously. Section 3.2's general objective can therefore be stated more precisely: preserve transformations suited to each representation and time scale, but provide a shared computational space through which all streams jointly update the same world model. With this criterion established, the next subsection examines MoT as one current architectural instantiation of the required balance between specialization and interaction.

3.2.2 Mixture-of-Transformers as a Current Design

Against the preceding criterion, Mixture of Transformers (MoT) [25] provides a direct architectural instantiation: it assigns modality-specific transformer parameters while retaining joint attention across modalities. Text, visual, and action tokens use their own projections, normalization, and feed-forward transformations, but participate in a common attention operation. Each token is routed deterministically to the expert associated with its modality; unlike learned-gating Mixture-of-Experts architectures [47, 48], MoT does not need to discover the routing assignment or balance traffic among interchangeable experts. The expert roles are fixed and interpretable, while shared attention allows information to cross modality boundaries.

For a GWM, these streams can be organized around three functionally aligned experts:

- A language-centered **understanding expert** integrates semantic knowledge with observations to represent the current state, task, and history.
- An observation-centered **imagination expert** models missing, future, or counterfactual world states and can decode detailed predictions into pixel space.
- An **action expert** conditions on the shared context to generate continuous control trajectories, for example through a flow-matching objective [49].

These experts are not independent modules connected only through final outputs. Joint attention allows an action token to condition directly on visual and linguistic context, a future-state token to condition on a candidate action, and an understanding token to incorporate predicted outcomes. This creates a single differentiable path through the Understanding–Imagination–Action loop while retaining parameters adapted to each modality.

The architecture supports the heterogeneous supervision supplied by the data pyramid. Web and egocentric

video can train visual understanding and generative prediction without robot actions; language–video pairs can ground instructions and semantics; synthetic or inferred actions can add weak intervention supervision; and robot trajectories can train action generation together with action-conditioned future prediction. Joint training allows these objectives to update the shared cross-modal computation, so that action learning begins from representations already informed by large-scale observation.

3.2.3 Relation to Alternative Architectures

MoT occupies a middle ground among several established designs:

- **Dense early fusion**, as represented by Chameleon [50], maximizes parameter sharing. MoT instead reserves modality-specific capacity while preserving direct token-level interaction.
- **Learned-gating Mixture-of-Experts** dynamically selects experts and can discover useful computational specialization. MoT uses the known modality identity for deterministic routing, simplifying optimization but providing less adaptive routing within a modality.
- **Hybrid shared-backbone models**, such as Transfusion [51], combine autoregressive and diffusion objectives within one backbone. MoT further separates modality-specific transformations and extends the shared context to continuous actions.
- **Specialist pipelines** permit independent optimization, replacement, and safety isolation of components. MoT trades some of that modularity for richer cross-modal transfer and end-to-end coordination.

This unification introduces its own challenges. Joint attention can become expensive for long video contexts; imbalanced datasets can cause large visual objectives to dominate scarce action signals; and shared optimization does not by itself guarantee physical correctness, causal grounding, or safe control. These issues require efficient attention, balanced sampling and losses, staged training, and action-grounded post-training. MoT should therefore be understood not as a complete solution, but as the architectural substrate on which the three faculties can be learned together and scaled without forcing them into identical representations.

3.3 Compute: Scaling Training and Meeting Real-Time Budgets

Compute is the third component of a GWM because the model must solve two different systems problems. During training, it must process heterogeneous datasets and long spatiotemporal contexts at foundation-model scale. During deployment, it must produce useful predictions or actions before the world changes. A system can therefore be compute-efficient in one sense and impractical in the other: high-throughput training does not guarantee low-latency inference, and a fast policy obtained from narrow data does not provide the breadth sought by a general world model.

3.3.1 Scaling Multimodal Training

Videos dominate the training cost because their token count grows across both space and time. Higher resolution, longer duration, and denser frame sampling improve the information available to the model but increase memory, attention cost, and communication. The scaling behavior observed in language models [32, 33] and diffusion transformers [52] suggests that useful capability emerges when model capacity, data, and compute are expanded together; increasing only one component eventually yields diminishing returns.

Training a GWM consequently requires parallelism along several axes. Data parallelism distributes examples, tensor and pipeline parallelism divide model parameters and layers, and sequence or context parallelism partitions long video streams. Mixed-precision arithmetic, activation checkpointing, and memory-efficient attention increase the context or batch size that fits on each device. The MoT architecture adds another form of structure: modality-specific parameters need only process their corresponding tokens, although the shared attention operation remains a major cost for long multimodal sequences.

For GWM, compute allocation must also reflect the data pyramid. If training samples are drawn only in proportion to dataset size, web videos can overwhelm the much smaller action datasets; if robot data are oversampled too aggressively, the model can lose breadth or overfit particular embodiments. A practical

curriculum therefore begins with broad visual and multimodal pretraining, introduces action-conditioned prediction and control objectives progressively, and reserves high-cost physical interaction data for grounding and post-training. Balanced sampling, objective weighting, and expert-specific learning schedules are as important as aggregate floating-point operations: they determine which faculty benefits from the available compute.

3.3.2 Inference Acceleration

The deployment budget depends on the GWM level. Offline world generation may tolerate seconds or minutes of computation for a high-quality sequence. An interactive world must respond quickly enough to preserve conversational and visual continuity. A robot policy operates under a harder closed-loop deadline: observations must be encoded, futures evaluated, and actions issued before latency makes the internal state stale. Efficient inference must therefore be treated as part of the model design rather than as a final implementation detail.

Inference acceleration can be organized into two complementary classes:

- **Increase computation speed.** This class executes the required operations more efficiently through low-bit quantization, hardware-aware kernels, operator fusion, compilation, and optimized serving. For example, the SageAttention [53–58] family quantizes attention computation to provide plug-and-play acceleration for language, image, and video models; related gains can be obtained from low-precision linear layers and efficient memory movement. At the serving layer, TurboServe [59] uses migration-aware placement and load-driven GPU autoscaling to improve the efficiency of long-lived streaming generation sessions.
- **Reduce computational complexity.** This class decreases the number of operations required to generate each output. Sparse attention avoids unnecessary token-pair computation [60–63], while improved samplers and distillation reduce the number of network evaluations. DPM-Solver obtains high-order updates from the diffusion probability-flow ODE and can generate samples in roughly 10–20 function evaluations without retraining [64]. With additional training, consistency distillation can compress generation into one or a few steps [65]. Distribution Matching Distillation (DMD) instead trains a one-step generator to match the teacher model's output distribution, avoiding the need to follow each multi-step sampling trajectory [66]. rCM [67] further scales continuous-time consistency distillation to application-scale image and video models while preserving generation quality and diversity. Based on these distillation methods, recent advances have introduced autoregressive diffusion distillation approaches [68–70] that compress diffusion models into a few-step autoregressive generation paradigm. By reducing the generation unit size, they substantially lower inference latency and improve interactivity. These techniques establish the technological foundation for real-time interactive video generation.

Note that the two classes can be composed to achieve higher efficiency. For example, TurboDiffusion [71] combines low-bit and sparse attention, rCM step distillation, W8A8 linear layers [72], and system-level optimizations in a full-stack framework, reporting a 100–200$\times$ end-to-end acceleration for large video diffusion models at comparable generation quality. More generally, a few-step generator can combine quantized FFN and attention operations with efficient cluster-scale serving.

Ultimately, compute links the GWM roadmap to deployable systems. Scaling training makes broad world knowledge possible, while inference acceleration determines whether that knowledge can support continuous interaction and physical action. The following sections examine these regimes in turn: high-fidelity video generation in Vidu, real-time interactive generation in Vidu S1, and prediction-coupled robot control in Motubrain.

4 Current Practice: Implementing the GWM Roadmap

The roadmap in Section 2.2 is aspirational, but its first three levels already have concrete implementations. As highlighted in Table 2, we report the progression from **Vidu** to **Vidu S1**, **Motus**, and **Motubrain** to illustrate how the same broad strategy—video pretraining, unified multimodal modeling, and inference acceleration—can move from offline world generation to real-time interaction and then to physical action. We emphasize

that these systems should not be read as proof that a fully general world model has been achieved. Rather, they are current engineering instances of successive capabilities required by the roadmap.

Table 2 Current example systems implementing L1–L3.

Models	Roadmap level	Implemented capability	Remaining boundary
Vidu	L1: World Generation	Conditional, coherent video generation	Offline generation without a continuous interaction loop
Vidu S1	L2: Interactive World	Streaming generation controlled by real-time voice input	Interaction remains within a generated digital world
Motus/Motubrain	L3: Actionable World	Unified understanding, future prediction, action modeling, and deployable cross-embodiment control	Goals remain externally specified; open-ended exploration and self-improvement remain future work

4.1 L1 — Vidu: Learning to Generate a World

The first roadmap level requires a model to generate coherent world trajectories from a condition. Vidu provides a direct implementation of this level. Introduced in May 2024 soon after Sora [10], Vidu uses a U-ViT diffusion backbone and can generate high-performance videos at up to 1080p and 16 seconds in a single generation [26]. Building on this foundation, the subsequent Vidu Q family pioneered the *Reference-to-Video* (R2V) paradigm, extending beyond conventional image-to-video generation in which an image primarily anchors the initial visual state. R2V instead accepts multiple reference images for subjects, characters, objects, or scenes, together with a text prompt specifying the desired dynamics, and synthesizes a new video while preserving the referenced visual identities and characteristics. This turns visual references into a flexible compositional control interface and supports stronger multi-entity consistency across changing motions, viewpoints, and scenes [73]. R2V is now becoming a standard feature for industrial video generators.

Beyond the Vidu family, other systems further broadened the practice of L1 world generation. For instance, Wan introduced an open suite of scalable video foundation models covering text-to-video, image-to-video, editing, and personalization [74]; Kling advanced high-resolution, long-duration generation with complex motion [75]; Google's Veo family emphasized high-fidelity, controllable generation and, with Veo 3, native audio synthesis [76]; while Seedance 2.0 unified text, image, audio, and video conditioning for joint audio-video generation [77]. The relevance of these systems to a GWM is not limited to visual quality. Maintaining subjects, motion, camera behavior, scene structure, and multimodal alignment across time requires models to learn reusable regularities from large-scale video, making generation a practical training objective for the Understanding and Imagination faculties.

This capability should be interpreted carefully. A visually plausible rollout demonstrates useful knowledge of appearance and common dynamics, but it does not by itself establish causal correctness or reliable action consequences. Vidu and similar systems therefore implement L1 rather than the full GWM: they generate a world for a viewer, but do not yet maintain a continuous feedback loop with an external actor. Their principal contribution to the roadmap is the scalable visual foundation on which later interaction and action capabilities can be built.

4.2 L2 — Vidu S1: Real-Time Interactive Imagination

A conventional generator (e.g., Vidu) samples an entire world trajectory from a fixed condition, $p(x_{1:t} \mid g)$: the user states an intention once, the model renders a clip, and the result is consumed only after generation is complete. A general world model must eventually operate in a different regime. It must preserve a state of the unfolding world, accept new inputs while that world is evolving, and revise its imagined future quickly enough for the user or agent to remain inside the loop. Real-time streaming is therefore not merely faster content creation. It is the transition from *offline imagination* to *interactive imagination*—from L1 World Generation to the L2 Interactive World in the roadmap above.

This transition exposes limitations that video quality alone cannot solve. Most large video models still follow an offline, one-shot diffusion paradigm: a user submits a prompt, waits for the full denoising process, and receives the completed video only at the end. Changing the generation schedule to an autoregressive one does not by itself create interaction; many streaming methods still cannot accept an intervention once generation has begun. Natural control is another obstacle: speech is often used only to drive lip motion rather than as an explicit instruction about what should happen next. Finally, small prediction errors accumulate over long rollouts, while the compute and serving cost of diffusion can prevent even a capable model from responding at interactive speed. These are the four barriers targeted by **Vidu S1** [27]: intervention, control, temporal stability, and deployable real-time inference.

Vidu S1 changes the role of speech from a soundtrack into a control channel. During generation, a user can issue a new spoken instruction at any moment—for example, asking the character to wave, make a gesture, stand, or sit—and the instruction guides the *future* video rather than merely animating the mouth. The model exposes a unified conditioning interface for speech, text, and reference images, and it jointly generates synchronized video and audio. Users can provide images of real people, animated characters, or pets and select different voice tones, turning the same streaming foundation into many personalized interactive characters. For a general world model, the important advance is not the avatar application itself, but the shift from producing a fixed clip to continuously revising the imagined future as new instructions arrive [27].

Technically, S1 constructs this capability in three stages. It first trains a bidirectional video–audio diffusion model to establish a high-quality generative prior. It then adapts that model to causal generation, combining teacher forcing with Diffusion Forcing so that each new state is predicted from the conditions available so far and from a possibly imperfect generated history. Finally, Distribution Matching Distillation with Phased Consistency Model regularization compresses the autoregressive denoising process into a few sampling steps while suppressing the drift and mode collapse that otherwise emerge during long rollouts. At inference time, a fixed-size sliding window keeps the per-step cost independent of elapsed video length; a persistent reference context anchors character identity, while RoPE repositioning and the proposed TwinCache reuse complementary noisy and clean historical states to improve both efficiency and temporal stability. Combined with the TurboDiffusion [71] inference engine and the TurboServe [59] serving system, these optimizations make causal generation fast enough for live interaction, reaching 540p at up to 42 frames per second on consumer GPUs.

The connection to the GWM formulation in Eq. (1) is direct. Let $x_{\leq t}$ denote the generated audiovisual history, and let a_t denote the user's current spoken intervention. Vidu S1 realizes an online conditional imagination process of the form $p(x_{t+1} \mid x_{\leq t}, a_{\leq t})$: the past supplies the current world state, while a newly arriving instruction redirects what the model imagines next. In this sense, streaming generation is an indispensable component of a GWM, because imagination must be causal, persistent, and responsive rather than an offline artifact. Vidu S1 does not yet complete the full GWM loop. It neither emits physical actions through a policy π nor autonomously forms goals. Instead, it establishes the bridge from generation to interaction. The next step is to carry this same online loop from pixels into actions, and from a responsive digital world into the physical world.

4.3 L3 — Motus and Motubrain: From Unified Understanding, Prediction, and Action to Deployable Control

Physical action requires the model to connect visual change to motor intervention. Motus and Motubrain are successive implementations of this same L3 objective: Motus establishes a unified understanding–prediction–action foundation, while Motubrain extends that foundation toward broader, faster, and more deployable robot control.

Motus: unifying understanding, prediction, and action. Introduced and open-sourced in December 2025, Motus implements L3 through a unified latent action world model [28]. Its Mixture-of-Transformers architecture coordinates understanding, video-generation, and action experts, while its UniDiffuser-style formulation *jointly models video and action through their conditional, marginal, and joint distributions*. World modeling, VLA prediction, inverse dynamics, video generation, and joint video–action prediction thus become different query modes of the same multimodal distribution. VLA policies, video generation models, world

models, and inverse-dynamics models can accordingly be viewed as special cases of this unified foundation, which may further incorporate continuous physical modalities.

Motus also connects the three construction requirements in Section 3. Its system-specific six-layer data organization is a finer-grained implementation than the five-layer conceptual pyramid used in this article; both combine broad video with increasingly action-grounded experience and use synthetic data as cross-layer augmentation. Its unified objective *absorbs heterogeneous supervision from every pyramid layer*, from action-free Internet video to human demonstrations and robot trajectories, without requiring every sample to contain every modality. Optical-flow-derived latent actions provide transferable motion supervision, while staged training adapts the representation to robot control. Meanwhile, MoT *composes mature pretrained experts*: video-generation models contribute motion, 4D spatiotemporal, and physical-interaction priors, while vision-language models contribute perception, semantics, and reasoning. Shared attention connects these priors while retaining modality-specialized transformations. In roadmap terms, Motus realizes L3 by making prediction operational: imagined or predicted state changes are grounded in multimodal representations of the present, inform the action distribution, and receive feedback from executed actions about whether both the inferred state and predicted consequences were useful.

Motubrain: scaling and deploying the L3 loop. Motubrain extends Motus into a scalable and deployable world action model [29]. Relative to Motus, it adds *unified arbitrary-view modeling* for different camera configurations, an independent text stream that directly couples language understanding to action generation, and a *unified relative-EEF action representation* that shares transferable action structure across robot embodiments.

Motubrain further scales up deployment through caching, low-precision computation, CUDA graphs, and real-time action chunking. The full model operates at about 5 Hz, approximately $10\times$ faster than Motus, while an IDM-style *Video-to-Action mode* skips video decoding and reaches up to 11 Hz [29], keeping a large embodied foundation model inside the real-time control loop. The evaluation results provide evidence for both task breadth and robustness. For example, Motubrain achieves average success rates of 95.8% in clean settings and 96.1% under randomized settings on RoboTwin 2.0 [29, 78], and can adapt to new humanoid embodiments with 50–100 trajectories. These results support the data and architecture theses: broad video and heterogeneous robot experience can be combined on one predictive-action foundation, then specialized to a new body with comparatively limited embodiment-specific data.

Together, Motus and Motubrain demonstrate the progression available *within* L3: first unify understanding, world prediction, and action learning, then expand the observation, embodiment, task, and deployment scope of the resulting controller. Their remaining limitation is the scope of agency. Both primarily execute externally supplied goals within learned task distributions; neither yet constitutes an L4 autonomous world agent that forms intermediate goals, explores open-endedly, and improves continually through online interaction.

4.4 What Current Practice Establishes—and What It Does Not

Taken together, the four systems implement a concrete path through the first three roadmap levels:

Vidu: generate $\longrightarrow$ Vidu S1: interact $\longrightarrow$ Motus/Motubrain: unify understanding, prediction, and action

The progression also makes the roles of data, architecture, and compute visible. Video pretraining supplies broad visual knowledge; MoT-style unification transfers that knowledge across understanding, future prediction, and action; and inference acceleration turns an otherwise offline generative model into an interactive or closed-loop system. Current practice therefore supports the central claim of this article: world generation, interaction, and action can be developed as stages of one foundation-model program rather than as unrelated applications.

The upper roadmap remains open. None of these systems yet demonstrates persistent self-directed goal formation, autonomous exploration, continual online learning in unrestricted environments, or orchestration across many agents and people. Those capabilities distinguish an L3 world action model from the L4 Autonomous World Agent and L5 World Orchestrator. The current systems establish a credible implementation path to those levels; they do not remove the need for the memory, causality, evaluation, efficiency, and safety advances discussed later.

5 Related Work

World modeling spans several research traditions that differ in what they predict, how they represent the world, and whether prediction is connected to action. We organize the literature along these distinctions.

Latent World Models for Model-Based Control Early neural world models learned compact states in which an agent could predict and plan. The Vision–Memory–Controller decomposition of Ha and Schmidhuber [6] separated visual encoding, latent dynamics, and control, while Dreamer learned policies from imagined latent trajectories [7]. DreamerV3 subsequently demonstrated that one configuration could solve more than 150 tasks across heterogeneous domains [8]. This line establishes the value of closing prediction and control in a learned representation. Its emphasis is sample-efficient interaction within defined environments; scaling the representation to open-ended video, language, and many physical embodiments is a different objective.

Predictive Representation Learning Joint-Embedding Predictive Architectures (JEPAs) predict target representations from context rather than reconstructing all observable detail [9]. I-JEPA applies the principle to masked image regions [79], and V-JEPA learns spatiotemporal visual representations from videos [20]. V-JEPA 2 extends the approach with an action-conditioned world model that supports image-goal planning on robots [21]. These methods show that predictable latent structures can support efficient understanding and planning without pixel generation. In the framework of this article, latent and pixel prediction are complementary: latent targets can discard nuisance detail, whereas decoded video provides explicit futures for interaction, inspection, and joint training with generative objectives.

Generative and Interactive World Models Diffusion models [45], latent diffusion [80], and video diffusion [81] made high-fidelity temporal generation scalable. Transformer backbones such as U-ViT [22] and the Diffusion Transformer (DiT) [52] further connected generation quality to model and compute scale. Sora framed large video generators as world simulators [10]; Cosmos developed world foundation models as reusable infrastructure for physical-AI simulation, synthetic data, and downstream customization [82]. Such systems learn broad visual regularities, although perceptual realism alone does not guarantee causal accuracy or effective control. Interactive generative models add a sequence of control inputs to the rollout. Genie infers a latent action space from video and enables frame-by-frame intervention without ground-truth action labels [11]. GameNGen demonstrates real-time neural simulation of a game [12], GAIA-1 studies action-conditioned generation for autonomous driving [83], and minWM [84] enables real-time interactive, camera-controllable world exploration. These models are closely related to L2 of our roadmap: they turn a conditioned video generator into a responsive environment, but their controls may be latent, domain-specific, or disconnected from a physical embodiment.

Robot Policies and World Action Models Vision-Language-Action (VLA) models learn direct mappings from observations and instructions to continuous robot actions. OpenVLA [16], the π family [15], and diffusion-based foundation policies such as RDT-1B [14] demonstrate strong language-conditioned manipulation across tasks and robots. Generalist AI's GEN-0 instead emphasizes native multimodal pretraining on large-scale, high-fidelity physical-interaction data and jointly streaming sensing and action, while GEN-1 further scales this foundation and combines pretraining, post-training, reinforcement learning, and inference-time techniques to improve adaptation and real-world task performance [85, 86]. RDT-1B scales a diffusion transformer through multi-robot pretraining and a unified action representation, while RDT2 extends this direction with large-scale, embodiment-agnostic UMI data and targets zero-shot deployment on previously unseen robot embodiments [87]. Their principal objective is successful action prediction; an explicit decoded future is optional. This directness can simplify control and reduce inference cost, while the extent to which large-scale action-free video contributes depends on the pretraining and representation used by each system.

A neighboring line makes predicted world evolution central to the policy. UniPi formulates decision-making as text-conditioned video generation followed by action inference [88], and GR-1 transfers large-scale video generative pretraining to manipulation [89]. Vidar adapts a video diffusion foundation for bimanual control [90]. DreamZero jointly predicts video and action with a pretrained video backbone and demonstrates real-time closed-loop execution and cross-embodiment transfer [91]. Motus and Motubrain, discussed as

current practice in Section 4, belong to this world-action-model family. The common hypothesis is that predicting observable consequences provides a useful intermediate structure for action learning, particularly when robot trajectories are scarce relative to videos.

Positioning of This Article This article does not propose that one of these traditions should replace all others. It synthesizes them into a definition and development program for a GWM. Relative to latent model-based control, it emphasizes foundation-scale observational pretraining and cross-embodiment transfer. Relative to video world simulators, it requires an explicit path from imagination to action. Relative to VLAs, it treats future-state prediction as a first-class output rather than only an implicit internal feature. Relative to existing multimodal foundations, it organizes data, architecture, compute, and deployment around a five-level progression from world generation to world orchestration. The distinctive contribution is therefore the joint framing of the Understanding–Imagination–Action triad, the data pyramid, the unified architecture, and the roadmap—not a claim that any current system has already achieved general embodied intelligence.

6 Conclusion and Outlook

This article has developed a first-principles account of the general world model. The central claim is that a GWM should not be identified with a video generator, a latent simulator, or a robot policy alone. It is a foundation that integrates three inseparable faculties: **Understanding** the present world, **Imagining** possible futures and interventions, and **Acting** to realize goals while incorporating feedback. This triad leads naturally to a five-level roadmap from World Generation and Interactive World to Actionable World, Autonomous World Agent, and World Orchestrator.

To build such a foundation, we analyze the required data, architecture, and compute. The five-layer data pyramid moves from web-scale video through curated domain and instructional video, egocentric human video, action-recording human demonstrations, and real-robot interaction. These layers trade scale for progressively stronger task, action, and embodiment grounding. Synthetic data augments all five layers with controllable variations and trajectories, but remains subject to real-world calibration. Imperfect action-aligned trajectories complement successful demonstrations by exposing intervention effects, failure states, and recovery behavior. The MoT architecture specializes transformations by modality while sharing joint attention, allowing language, video, and action to exchange information without forcing them into identical representations. Scalable training and accelerated inference determine whether the foundation remains an offline model or becomes a responsive system capable of interaction and control.

Finally, we present the current implementations to make the roadmap concrete, with Vidu, Vidu S1, and Motus/Motubrain Specifically corresponding to L1, L2, and L3. Together, these systems support the feasibility of a shared foundation-model program across generation, interaction, and control. They do not yet demonstrate L4 or L5: the L3 systems operate under externally specified goals and bounded task and embodiment distributions, and none of the examples autonomously explores, improves continually, or coordinates open-ended multi-agent systems. Closing this gap motivates six directions:

First, evaluation must measure generality rather than isolated task success. Existing benchmarks such as RoboTwin 2.0 [78] and RoboCasa365 [92] evaluate important aspects of manipulation, but a GWM also requires joint measurement of semantic understanding, long-horizon predictive consistency, the agreement between imagined and realized futures, robustness under intervention, and transfer across tasks, environments, and embodiments. Evaluation should expose tradeoffs among the three faculties rather than allowing strength in one to conceal weakness in another.

Second, prediction must become causal and physically grounded. A plausible video is not necessarily a correct model of dynamics. Reliable planning requires sensitivity to interventions, contact, friction, force, object permanence, and failure modes outside the training distribution. Counterfactual evaluation and action-grounded data are therefore necessary to distinguish a model that reproduces visual correlations from one that predicts the consequences of control.

Third, a GWM needs persistent and revisable memory. Streaming generation can extend a sequence indefinitely [93], yet duration alone does not guarantee coherence. An autonomous agent must remember

objects outside the current view, maintain commitments from earlier instructions, update beliefs when observations contradict predictions, and retrieve relevant experience over minutes, days, or longer. The required memory is not merely a larger fixed context, but an explicit world state that can be written, queried, and corrected.

Fourth, pretraining must be connected to online learning and self-improvement. Video teaches how the world commonly evolves, while demonstrations teach how other agents acted; neither alone determines which behavior best achieves a new goal. Model-based reinforcement learning [7, 8] provides a starting point: imagined experience can propose and evaluate behavior, and real interaction can correct the learned dynamics. Crucially, this closed loop allows a world model to use real-world ground truth from its interactions with the physical world to revise itself continually and thereby support self-improvement. Scaling this cycle to open-ended video foundations will require uncertainty-aware exploration, safeguards against model exploitation, and careful integration of simulated and physical feedback.

Fifth, capability must fit the latency and resource budget of deployment. Few-step generation, low-bit attention, sparse attention, caching, streaming, and action chunking make real-time operation increasingly practical, but acceleration methods often introduce a tradeoff between capability and efficiency. Efficiency must be evaluated through closed-loop outcomes rather than throughput alone.

Sixth, safety and controllability must grow with agency. A model that can imagine and act must represent constraints, communicate uncertainty, expose its proposed futures for inspection, and keep executed actions within safe operating envelopes. Persistent memory and online learning add further requirements: the system must preserve human intent, resist harmful distribution shifts, and remain auditable as its internal state changes. Interpretability of predicted states and plans is therefore not separate from performance; it is part of what makes an increasingly autonomous world model deployable among people.

These problems are tightly coupled. Better evaluation supplies objectives for online learning; physical grounding restricts which imagined futures are trusted; persistent memory supports long-horizon planning; efficient inference keeps that planning inside the control loop; and safety determines which actions may be explored at all. The near-term task is to strengthen the first three roadmap levels while building the capabilities required by the fourth. The longer-term goal is a foundation that can maintain an accurate world state, imagine consequential alternatives, act under explicit constraints, and learn continuously across digital and physical environments. Achieving that goal would turn the general world model from a promising architectural program into dependable embodied intelligence.

References

[1] Kenneth J. W. Craik. *The Nature of Explanation*. Cambridge University Press, Cambridge, 1943.

[2] Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

[3] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[4] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

[5] Daniel M. Wolpert, Zoubin Ghahramani, and Michael I. Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, 1995.

[6] David Ha and Jürgen Schmidhuber. World models. In *Advances in Neural Information Processing Systems*, 2018.

[7] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

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

[9] Yann LeCun. A path towards autonomous machine intelligence. Open Review, 2022.

[10] Tim Brooks, Bill Peebles, et al. Video generation models as world simulators. Technical report, OpenAI, 2024.

[11] Jake Bruce, Michael Dennis, Ashley Edwards, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.

[12] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

[13] Xinjie Wang, Liu Liu, Yu Cao, Ruiqi Wu, Wenkang Qin, Dehui Wang, Wei Sui, and Zhizhong Su. EmbodiedGen: Towards a generative 3d world engine for embodied intelligence. *arXiv preprint arXiv:2506.10600*, 2025.

[14] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.07864.

[15] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems*, 2025. URL https://arxiv.org/abs/2410.24164.

[16] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[17] Physical Intelligence, Bo Ai, Ali Amin, et al. $\pi_{0.7}$: A steerable generalist robotic foundation model with emergent capabilities, 2026. URL https://arxiv.org/abs/2604.15483.

[18] Hengkai Tan, Yao Feng, Xinyi Mao, Shuhe Huang, Guodong Liu, Zhongkai Hao, Hang Su, and Jun Zhu. Anypos: Automated task-agnostic actions for bimanual manipulation, 2025. URL https://arxiv.org/abs/2507.12768.

[19] Bowen Baker, Ilge Akkaya, Peter Zhokhov, et al. Video pretraining (VPT): Learning to act by watching unlabeled online videos. In *Advances in Neural Information Processing Systems*, 2022.

[20] Adrien Bardes, Quentin Garrido, Jean Ponce, et al. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

[21] Mahmoud Assran et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[22] Fan Bao, Shen Nie, Kaiwen Xue, et al. All are worth words: A ViT backbone for diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[23] Albert Bandura. *Social Learning Theory*. Prentice-Hall, 1977.

[24] Giacomo Rizzolatti and Laila Craighero. The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192, 2004.

[25] Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2411.04996.

[26] Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models, 2024. URL https://arxiv.org/abs/2405.04233.

[27] Jintao Zhang, Kai Jiang, Jintao Chen, Xu Wang, Yang Luo, Yuji Wang, Dechuang Chen, Jungang Li, Chengyang Ye, Marco Chen, et al. Vidu s1: A real-time interactive video generation model. *arXiv preprint arXiv:2607.03118*, 2026.

[28] Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 35101–35113, 2026.

[29] Motubrain Team, Chendong Xiang, Fan Bao, Haitian Liu, Hengkai Tan, Hongzhe Bi, James Li, Jiabao Liu, Jingrui Pang, Kiro Jing, Louis Liu, Mengchen Cai, Rongxu Cui, Ruowen Zhao, Runqing Wang, Shuhe Huang, Yao Feng, Yinze Rong, Zeyuan Wang, and Jun Zhu. Motubrain: An advanced world action model for robot control. *arXiv preprint arXiv:2604.27792*, 2026.

[30] Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007.

[31] Renée Baillargeon. Infants' physical world. *Current Directions in Psychological Science*, 13(3):89–94, 2004.

[32] Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[33] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, 2022.

[34] Richard S. Sutton. The bitter lesson. Incomplete Ideas blog, 2019.

[35] Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[36] Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

[37] Build AI. Egocentric-10k. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/builddotai/Egocentric-10K.

[38] Build AI. Egocentric-100k. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/builddotai/Egocentric-100K.

[39] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Robotics: Science and Systems*, 2024.

[40] Zhaxizhuoma, Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Pengan Chen, Pingrui Zhang, Haoming Song, Delin Qu, Dong Wang, Zhigang Wang, Nieqing Cao, Yan Ding, Bin Zhao, and Xuelong Li. FastUMI: A scalable and hardware-independent universal manipulation interface with dataset. *arXiv preprint arXiv:2409.19499*, 2024.

[41] Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. DexUMI: Using human hand as the universal manipulation interface for dexterous manipulation. *arXiv preprint arXiv:2505.21864*, 2025.

[42] Open X-Embodiment Collaboration. Open X-embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation*, 2024.

[43] AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, et al. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[44] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. MimicGen: A data generation system for scalable robot learning using human demonstrations. In *Conference on Robot Learning*, 2023.

[45] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[46] Allan Paivio. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45(3):255–287, 1991. doi: 10.1037/h0084295.

[47] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

[48] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23, 2022.

[49] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

[50] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[51] Chunting Zhou, Lili Yu, Arun Babu, et al. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

[52] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

[53] Jintao Zhang, Jia Wei, Pengle Zhang, Jun Zhu, and Jianfei Chen. Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. In *International Conference on Learning Representations (ICLR)*, 2025.

[54] Jintao Zhang, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, and Jianfei Chen. Sageattention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization. In *International Conference on Machine Learning (ICML)*, 2025.

[55] Jintao Zhang, Xiaoming Xu, Jia Wei, Haofeng Huang, Pengle Zhang, Chendong Xiang, Jun Zhu, and Jianfei Chen. Sageattention2++: A more efficient implementation of sageattention2. *arXiv preprint arXiv:2505.21136*, 2025.

[56] Jintao Zhang, Jia Wei, Pengle Zhang, Xiaoming Xu, Haofeng Huang, Haoxu Wang, Kai Jiang, Jun Zhu, and Jianfei Chen. Sageattention3: Microscaling fp4 attention for inference and an exploration of 8-bit training. *arXiv preprint arXiv:2505.11594*, 2025.

[57] Jintao Zhang, Marco Chen, Haoxu Wang, Kai Jiang, Ion Stoica, Joseph E Gonzalez, Jianfei Chen, and Jun Zhu. Sagebwd: A trainable low-bit attention. *arXiv preprint arXiv:2603.02170*, 2026.

[58] Jintao Zhang, Rundong Su, Chunyu Liu, Jia Wei, Ziteng Wang, Haoxu Wang, Pengle Zhang, Huiqiang Jiang, Haofeng Huang, Chendong Xiang, et al. Efficient attention methods: Hardware-efficient, sparse, compact, and linear attention. *arXiv preprint*, 2025.

[59] Youhe Jiang, Haoxu Wang, Haotong Bao, Kai Jiang, Jianfei Chen, Jun Zhu, Fangcheng Fu, and Jintao Zhang. Turboserve: Serving streaming video generation efficiently and economically. *arXiv preprint arXiv:2606.19271*, 2026.

[60] Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. Spargeattention: Accurate and training-free sparse attention accelerating any model inference. *arXiv preprint arXiv:2502.18137*, 2025.

[61] Jintao Zhang, Haoxu Wang, Kai Jiang, Shuo Yang, Kaiwen Zheng, Haocheng Xi, Ziteng Wang, Hongzhou Zhu, Min Zhao, Ion Stoica, Joseph E. Gonzalez, Jun Zhu, and Jianfei Chen. Sla: Beyond sparsity in diffusion transformers via fine-tunable sparse-linear attention. *arXiv preprint arXiv:2509.24006*, 2025.

[62] Jintao Zhang, Haoxu Wang, Kai Jiang, Kaiwen Zheng, Youhe Jiang, Ion Stoica, Jianfei Chen, Jun Zhu, and Joseph E Gonzalez. Sla2: Sparse-linear attention with learnable routing and qat. *arXiv preprint arXiv:2602.12675*, 2026.

[63] Jintao Zhang, Kai Jiang, Chendong Xiang, Weiqi Feng, Yuezhou Hu, Haocheng Xi, Jianfei Chen, and Jun Zhu. Spargeattention2: Trainable sparse attention via hybrid top-k+ top-p masking and distillation fine-tuning. *arXiv preprint arXiv:2602.13515*, 2026.

[64] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022.

[65] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

[66] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[67] Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. *arXiv preprint arXiv:2510.08431*, 2025.

[68] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

[69] Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

[70] Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026.

[71] Jintao Zhang, Kaiwen Zheng, Kai Jiang, Haoxu Wang, Ion Stoica, Joseph E Gonzalez, Jianfei Chen, and Jun Zhu. Turbodiffusion: Accelerating video diffusion models by 100-200 times. *arXiv preprint arXiv:2512.16093*, 2025.

[72] Pengle Zhang, Jia Wei, Jintao Zhang, Jun Zhu, and Jianfei Chen. Accurate int8 training through dynamic block-level fallback. *arXiv preprint arXiv:2503.08040*, 2025.

[73] Vidu Team. Vidu Q2: Reference-to-video. Vidu API Documentation, 2025. URL https://platform.vidu.com/docs/reference-to-video. Reference-to-Video capability added to the Vidu Q2 family on October 21, 2025.

[74] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[75] Kling Team. Kling-omni technical report. ArXiv, 2512.16776v1, 2025.

[76] Google. Veo 3 announcement, 2025. Accessed: Aug 31, 2026. URL https://blog.google/technology/ai/generative-media-models-io-2025/.

[77] Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Zetao Fang, Chengjian Feng, Yu Gao, Diandian Gu, Dong Guo, Hanzhong Guo, Qiushan Guo, Boyang Hao, Hongxiang Hao, Haoxun He, Jiaao He, Qian He, Tuyen Hoang, Heng Hu, Ruoqing Hu, Yuxiang Hu, Jiancheng Huang, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Jishuo Jin, Ming Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bingchuan Li, Gen Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, J. H. Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Jihao Liu, Jingtuo Liu, Shu Liu, Sichao Liu, Wei Liu, Xue Liu, Zuxi Liu, Ruijie Lu, Lecheng Lyu, Jingting Ma, Tianxiang Ma, Xiaonan Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Ronggui Peng, Xueqiong Qu, Yuxi Ren, Yuchen Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zirui Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang, Shulei Wang, Weichen Wang, Xuanda Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Guoqiang Wei, Meng Wei, Di Wu, Guohong Wu, Hanjie Wu, Huachao Wu, Jian Wu, Jie Wu, Ruolan Wu, Shaojin Wu, Xiaohu Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Xin Xia, Xuefeng Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yihang Yang, Zhixian Yang, Ziyan Yang, Fulong Ye, Bingqian Yi, Xing Yin, Yongbin You, Linxiao Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chenlin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyuan Zhang, Shuo Zhang, Xiaohe Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Huating Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng Zheng, Zerong Zheng, Kuan Zhu, and Feilong Zuo. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

[78] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan-ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

[79] Mahmoud Assran, Quentin Duval, Ishan Misra, et al. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[80] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[81] Jonathan Ho, Tim Salimans, Alexey Gritsenko, et al. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022.

[82] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

[83] Anthony Hu, Lloyd Russell, Hudson Yeo, et al. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[84] Min Zhao, Hongzhou Zhu, Bokai Yan, Zihan Zhou, Yimin Chen, Wenqiang Sun, Kaiwen Zheng, Guande He, Xiao Yang, Chongxuan Li, et al. minwm: A full-stack open-source framework for real-time interactive video world models. *arXiv preprint arXiv:2605.30263*, 2026.

[85] Generalist Team. Gen-0: Embodied foundation models that scale with physical interaction. *Generalist AI Blog*, 2025. https://generalistai.com/blog/gen-0.

[86] Generalist Team. Gen-1: Scaling embodied foundation models to mastery. *Generalist AI Blog*, 2026. https://generalistai.com/blog/gen-1.

[87] Songming Liu, Bangguo Li, Kai Ma, Lingxuan Wu, Hengkai Tan, Xiao Ouyang, Hang Su, and Jun Zhu. Rdt2: Exploring the scaling limit of umi data towards zero-shot cross-embodiment generalization, 2026. URL https://arxiv.org/abs/2602.03310.

[88] Yilun Du, Sherry Yang, Bo Dai, et al. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, 2023.

[89] Hongtao Wu, Ya Jing, Chi Cheang, et al. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, 2024.

[90] Yao Feng, Hengkai Tan, Xinyi Mao, Guodong Liu, Shuhe Huang, Chendong Xiang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist bimanual manipulation. *arXiv preprint arXiv:2507.12898*, 2025.

[91] 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. World action models are zero-shot policies. In *2nd ICLR Workshop on World Models: Understanding, Modelling and Scaling*, 2026. URL https://arxiv.org/abs/2602.15922.

[92] Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. RoboCasa365: A large-scale simulation framework for training and benchmarking generalist robots. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2603.04356.

[93] Boyuan Chen, Diego Marti Monso, et al. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, 2024.