# WoW: Towards a World-omniscient World-model Through Embodied Interaction

Xiaowei Chi[1,2,3,†], Peidong Jia[1,2,†], Chun-Kai Fan[1,2,†], Xiaozhu Ju[1,†], Weishi Mi[1,†], Zhiyuan Qin[1,†], Kevin Zhang[2], Wanxin Tian[1], Kuangzhi Ge[2], Hao Li[1], Zezhong Qian[1,2], Anthony Chen[2], Qiang Zhou[1,2], Yueru Jia[2], Jiaming Liu[2], Yong Dai[1], Qingpo Wuwu[2], Chengyu Bai[2], Yu-Kai Wang[2], Ying Li[2], Lizhang Chen[1,2], Yong Bao[1], Zhiyuan Jiang[1], Jiacheng Zhu[1], Kai Tang[2], Ruichuan An[2], Yulin Luo[2], Qiuxuan Feng[1,2], Siyuan Zhou[3], Chi-min Chan[3], Chengkai Hou[1,2], Wei Xue[3], Sirui Han[3], Yike Guo[3], Shanghang Zhang[2,✉], Jian Tang[1,✉]

[1] Beijing Innovation Center of Humanoid Robotics,
[2] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University,
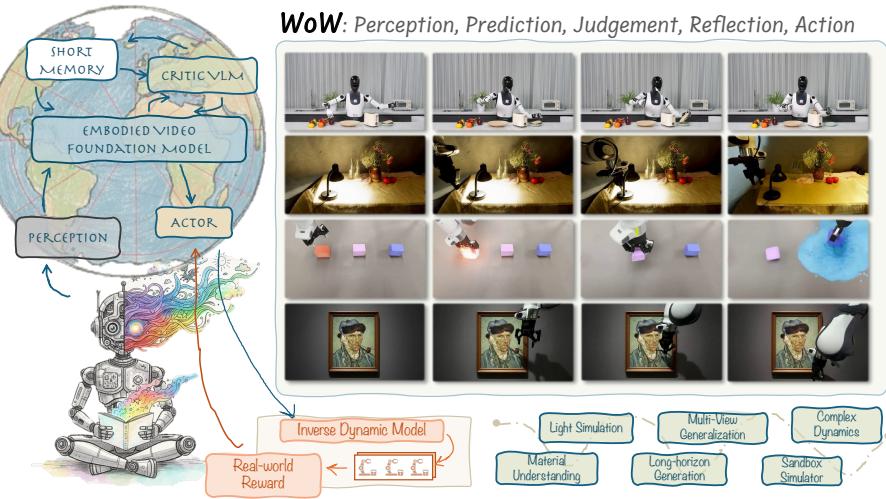[3] Hong Kong University of Science and Technology

Figure 1: **WoW** is a world model that integrates perception, prediction, Judgement, reflection, and action. It learns from real-world interaction data and generates high-quality, physically consistent robot videos in seen and out-of-distribution scenarios, enabling real-world robotic execution.

## Abstract

Humans develop an understanding of intuitive physics through active interaction with the world. This approach is in stark contrast to current video models, such as Sora, which rely on passive observation and therefore struggle with grasping physical causality. This observation leads to our central hypothesis: *authentic physical intuition of world model must be grounded in extensive, causally rich interactions with the real world.* To test this hypothesis, we present **WoW**, a 14B-parameter generative world model trained on 2 million robot interaction trajectories. Our findings reveal that the model's understanding of physics is a probabilistic distribution of plausible outcomes, leading to stochastic instabilities and physical hallucinations. Furthermore, we demonstrate that this emergent capability can be actively constrained toward physical realism by *SOPHIA*, where vision language model agents evaluate the DiT's generated output and guide its refinement by iteratively evolving the language instruction. Besides, a co-trained *Inverse Dynamics Model* translates these refined plans into executable robotic actions, thus closing the imagination-to-action loop. We establish **WoWBench**, a new benchmark focused on physical consistency and causal reasoning of video, where WoW achieves state-of-the-art performance of both human and autonomous evaluation, demonstrating strong ability on physical causality, collision dynamics, and object permanence. Our work provides the systematic evidence that large-scale, real-world interaction is a cornerstone for developing physical intuition in AI. Models, data, and benchmarks will be open-sourced in wow-world-model.github.io

1

CONTENTS

# 1 INTRODUCTION

> *"The ladder of causation has three rungs: seeing, doing, and imagining."*
> — Judea Pearl (Pearl & Mackenzie, 2018)

How does a human child acquire an understanding of the world? Not by passively observing videos, but through active and physical experimentation. The cognitive scientist Jean Piaget succinctly articulated this principle: *"To know an object is to act on it" (Piaget, 2013)*. This form of embodied learning, where countless 'actions' are intrinsically linked to immediate 'outcomes', is the foundational mechanism for mastering the laws of physics. This principle finds its direct computational instantiation in embodied world models, which are explicitly designed to learn a predictive model of how the world responds to an agent's actions (Hafner et al., 2019).

In contrast, many recent advances in predictive models, particularly in video generation, is predicated on passive observation, a principle fundamentally distinct from active experimentation that fosters accurate causal understanding. While models like Sora (Brooks et al., 2024) and others (Wan et al., 2025) achieve stunning photorealism and demonstrate emergent physical intuition, this intuition remains brittle. Their training objective prioritizes modeling statistical correlations from internet-scale data over inferring the underlying causal mechanisms of physics. Consequently, their grasp of physical laws is often superficial. When tasked with scenarios requiring genuine physical reasoning, they can produce logically and physically inconsistent outcomes. These models master the *appearance* of our world, but the generative *dynamics* they learn are an approximation rather than an accurate representation.

This distinction motivates our core hypothesis, **for an embodied model to develop genuine physical intuition, it must learn from large-scale, causally-rich, real-world interaction data, thereby lifting the generative model toward a world model** (Ha & Schmidhuber, 2018; Agarwal et al., 2025). To further validate our approach, we first introduce *SOPHIA*, a novel architectural paradigm that couples the reasoning capabilities of a Vision Language Model (VLM) with the generative power of a Diffusion Transformer (DiT) (Peebles & Xie, 2023). We then present **WoW**, a concrete instantiation of this paradigm. WoW is a generative world model trained on a large-scale dataset of 2 million real-world robotic interaction trajectories, spanning 5275 tasks and 12 different robots. The objective of WoW is to directly synthesize pixel-level future predictions, learning to imagine and reason through generation itself.

To close the perception-to-action loop, we designed the Flow-Mask Inverse Dynamics Model (FM-IDM), which functions as the agent's equivalent of the cerebellum and motor cortex. By analyzing the optical flow and scene context between the current state and the imagined next state, the FM-IDM infers the 7-DoF end-effector action necessary to enact the transition. This module grounds the agent's imagination in physical reality, translating pixel-level futures into executable actions.

To empirically validate this complete perception-imagination-reflection-action cognitive architecture, we established **WoWBench**, a new benchmark focused on physical consistency and causal reasoning. WoWBench is composed of 4 core abilities and 20 sub-tasks, containing 606 samples, each with an initial image and a textual instruction. For a comprehensive evaluation, we comprehend 4 indispensable metrics: Video quality, Planning reasoning, Physical rules, and Instruction following. Our experiments demonstrate that our 14B-parameter WoW achieves SOTA performance on this benchmark, especially 96.53% on Instruction understanding, and 80.16% on Physical law, providing compelling evidence in support of our central hypothesis. To further verify the precision of our WoWBench, we conduct a human evaluation proving that WoWBench is highly correlated with human preference, and WoW achieves SOTA performance on both sides. Beyond its benchmark performance, WoW demonstrates versatility in broader applications. We show it is more than a simple generator, capable of enhancing VLM reasoning, serving as a physical simulator, and enabling 3D-aware representation learning.

In summary, we propose *SOPHIA*, a paradigm for developing embodied intelligence through a data-driven feedback loop. This approach involves deploying capable models to collect large-scale corrective feedback from physical interactions, a process that drives a continuous cycle of improvement. Our model, **WoW**, serves as a powerful instantiation of this paradigm, representing a significant advance from passive video models to an embodied world model that closes the perception-imagination-reflection-action loop. Our main contributions are as follows.

- **A Unified Architecture for Imagination and Action.** We introduce an embodied world model **WoW**, which instantiates a novel self-optimization framework *SOPHIA* for imagining physically plausible futures, and incorporates a Flow-Mask Inverse Dynamics Model that infers the corresponding executable actions.

- **Self-Supervised Feature Alignment.** We are the first to integrate powerful, pre-trained self-supervised visual features into the backbone of a diffusion-based world model. This novel approach significantly boosts the model's perceptual capabilities, accelerates training convergence, and improves the fidelity and physical consistency of generated futures.

- **An Interaction Benchmark.** We propose WoWBench, a new public benchmark designed to evaluate the physical consistency and action-generation capabilities of world models. Comprising 606 diverse, high-quality robot trajectories, the benchmark facilitates a rigorous performance comparison between existing methods and proposed WoW across 4 core metrics and 20 associated tasks.

- **Postraining Application Discussion.** We demonstrate that **WoW** is more than a generator, showcasing its versatility across a range of downstream applications. These applications include synthesizing novel views, generating trajectory-guided videos, producing action-conditioned videos for robot manipulation, enhancing visual style transfer and improving VLM task planning at test-time.

- **A Scaling Analysis and Open-Source Models.** We perform a systematic scaling analysis of our architecture up to 14B parameters, uncovering nascent capabilities and potential precursors to emergence in physical reasoning. This provides strong empirical evidence for the scaling hypothesis in this domain. We will release all trained model checkpoints to provide a foundation for future research in the embodied world model.

## 2 RETHINKING WORLD MODEL: TOWARDS A WORLD-OMNISCIENT INTELLIGENT

The idea that intelligent agents, whether biological or artificial, build internal models of their environment to predict, plan, and act has deep roots in cognitive science (Neisser, 1976). This notion has been instantiated in modern AI research, most notably in Ha and Schmidhuber's **world model**(Ha & Schmidhuber, 2018), and advanced by PlaNet (Hafner et al., 2018) with recurrent state-space models for latent planning. The Dreamer series (Hafner et al., 2019) further extended this line by integrating imagination-based actor-critic learning, establishing world models as a scalable paradigm for long-horizon decision-making.

In what follows, we provide a formal definition of world models section 2.1, trace their evolution and recent applications in embodied AI section 2.2, and highlight their critical role in the broader development of generative AI section 2.3. Furthermore, we discuss the relationships among video generation, large pretrained models, and latent-space world models. This analysis provides the rationale for our proposed approach: **structuring world models by effectively leveraging pretrained foundation models and offering insights into future architectural designs**.

### 2.1 DEFINITION OF A WORLD MODEL

A world model is an internal, learned representation of the dynamics of an environment, designed to simulate or predict future states based on current states and potential actions (Ha & Schmidhuber, 2018). Formally, given a state $s_t \in \mathcal{S}$, a low-level control action $a_t \in \mathcal{A}$ at timestep $t$, and $p_t \in \mathcal{P}$ is the meta-level strategy/plan at time $t$, the world model learns a transition function $f_\theta$ parameterized by $\theta$. This function aims to predict the subsequent state $s_{t+1}$.

The prediction can be modeled deterministically as:

$$s_{t+1} = f_\theta(s_t, a_t, p_t) \quad with \quad a_t \sim \pi_\phi(a_t|s_t, p_t), \quad p_t \sim \pi_\omega(p_t|s_t, H_t) \tag{1}$$

where $\pi_\phi$ refers to low-level policy and $\pi_\omega$ refers to high-level policy. The term $H_t = (s_{t-h:t}, a_{t_h:t}, p_{t-h,t})$ is the historical context up to time t, and $h$ is the recall horizon.

or probabilistically, as:

$$s_{t+1} \sim \mathbb{P}_\theta(s_{t+1} \mid s_t, a_t, p_t) \tag{2}$$

To handle high-dimensional sensory inputs, such as images, modern implementations typically operate within a compressed latent space. Encoders map observations $o_t$ of state $s_t$ to a low-dimensional latent state $z_t$. The transition model then operates in this abstract space:

$$z_{t+1} \approx f_\theta(z_t, a_t, p_t) \quad with \quad z_t = Encoder(o_t) \tag{3}$$

The core training objective is to minimize prediction error over a transition dataset $\mathcal{D} = \{(o_t, a_t, o_{t+1})\}$, typically using a loss function such as Mean Squared Error (MSE):

$$\min_\theta \mathcal{L}_{\text{trans}}(\theta) = \mathbb{E}_{(z_t, a_t, z_{t+1}) \sim \mathcal{D}} \left[ \| f_\theta(z_t, a_t) - z_{t+1} \|^2 \right] \tag{4}$$

This objective compels the model to internalize physical laws, object permanence, and causal relationships that govern environment dynamics (Hafner et al., 2024).

## 2.2 EVOLUTION OF WORLD MODELS: REPRESENTATION AND MODALITY

Initially, world models were primarily used in model-based reinforcement learning (RL) as compact latent representations. Pioneering this line, World Models (Ha & Schmidhuber, 2018) coupled a variational autoencoder (VAE) with an Mixture Density Network Recurrent Network (MDN-RNN) to learn a generative internal model from pixels and showed that a simple controller trained entirely "in the dream" of the model can transfer back to the real environment, establishing the feasibility and sample-efficiency of learning inside a learned simulator. Building on this idea with a stronger recurrent state-space model(RSSM), PlaNet (Hafner et al., 2018) performed fast online planning in latent space via a stochastic-deterministic dynamics model trained with multi-step variational objectives, enabling image-based control with substantially improved data efficiency. The Dreamer family (Hafner et al., 2019) then replaced step-wise planning with differentiable actor-critic learning inside imagination: Dreamer back propagates value gradients through imagined rollouts in a learned world model; DreamerV2 adds discrete latents to reach human-level Atari; and DreamerV3 standardizes robust training to solve 150+ tasks under one configuration, making "learning in the dream" a scalable general RL approach.

Like RSSMs, Joint-Embedding Predictive Architectures(JEPAs) (Assran et al., 2023) also learn abstract latent representations, but JEPA optimizes predictive agreement in embedding space rather than modeling pixel-level likelihoods or explicit dynamics. Concretely, given a context embedding, a predictor is trained to match a stop-gradient target encoder on masked views of the same scene, inducing semantic structure without pixel reconstruction or heavy augmentations. This objective scales naturally and encourages invariant/equivariant features that capture object- and scene-level regularities. Extending to video, V-JEPA (Bardes et al., 2024) performs in-context prediction of masked spatio temporal chunks in latent space; V-JEPA-2 (Assran et al., 2025) further shows that pretraining on web video, augmented with a small amount of interaction data, yields world-model–like priors that enable zero-shot robotic planning directly from images. Although these models are not full-fledged world models in the control sense, they highlight a predictive representation learning path toward scalable priors for embodied reasoning.

The above world models primarily encode an implicit understanding of the external world in latent space. In February 2024, OpenAI introduced Sora (Brooks et al., 2024), widely regarded as a video world model that markedly advances environment simulation and prediction. Broadly, video world models fall into two complementary families. Inspired by breakthroughs in large language models, Autoregressive (AR) approaches tokenize images/videos and learn causal next-token dynamics, which makes online rollouts and tight policy coupling natural (Hu et al., 2023; Bruce et al., 2024). In contrast, diffusion-based approaches model conditional video generation via iterative denoising, better capturing multi-modality and stochasticity, and can fold planning into generation through trajectory diffusion (Yang et al., 2023) . Despite their promise, current video world models face significant limitations: they often lack 3D consistency, physical coherence, and temporal reasoning required for faithful environment simulation. Addressing these challenges remains central to advancing world models toward general-purpose embodied intelligence.
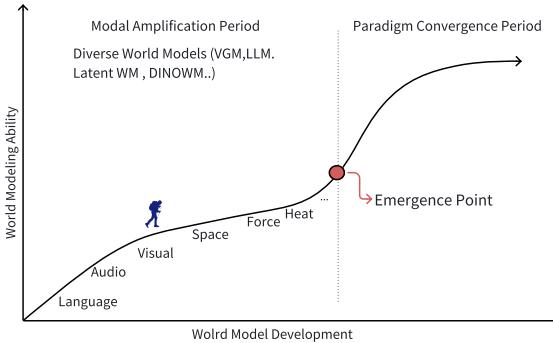
Figure 2: Developmental trajectory of world models, from modality-specific models (e.g., VGM, LLM) to unified models after a critical emergence point.

The convergence of these modalities—from model-based RL to large language and video generation models—marks the current era of multimodal expansion. While each approach has its strengths, they all currently lack the dynamic interactive properties and the comprehensive world understanding required for a true general-purpose world model. The models have diverse architectures, training methods, and paradigms.

As AI continues to exhaustively integrate more modalities, the development of a unified world model will become a critical component for AIs to explore and interact with the real world. This will, in turn, lead to more mature and robust AI architectures and paradigms as shown in Figure 2.

## 2.3 EMBODIED WORLD MODELING

The embodied world model (Yang et al., 2023; Azzolini et al., 2025) stands as the next Holy Grail, a system that demands long-term, multimodal interactive modeling to reach its full potential. Yet today, several critical limitations remain: unresolved challenges with spatial and temporal consistency, the absence of a universal general control interface, and significant gaps in physical understanding and common sense.

**Spatial Consistency** Our world is inherently three-dimensional, where every object and event possesses intrinsic 3D properties that define their existence. In contrast, conventional video generation models prioritize only the visual effects perceivable by the human eye, neglecting the critical underlying requirement of 3D consistency. Embodied world models that are designed to simulate real environments require that their outputs be strictly 3D-consistent (Zhen et al., 2025).

**Temporal Consistency** Severe memory constraints pose a core challenge for embodied world models, limiting the encoding of the world's state to a single image or discrete tokens. This fundamentally hinders their ability to reliably simulate complete 3D environments where much of the world remains unobserved. Consequently, such models suffer from short-sightedness: they often produce new elements that contradict their own prior context. This inconsistency directly compromises the internal coherence that is critical for achieving real-world applications. Such contradictions further weaken the models' capacity to monitor environments as they evolve and to formulate plans across extended time horizons. In the end, this limits their effectiveness for tasks that depend on stable, long-horizon interaction. Recent work (Hong et al., 2024) utilizes a temporary LoRA module that embeds the episodic memory in its parameters in video diffusion models, while other works (Chai et al., 2023; Zhou et al., 2025a; Lin et al., 2025a) integrate an explicit memory architecture that enables consistent long-horizon imagination, ensuring coherence with previously generated content.

**Physically Coherent Simulation** The understanding of physical laws is a crucial challenge for AGI (Chow et al., 2025), and physical understanding becomes particularly important when we need AGI to interact with the real world through physical actions (Bansal et al., 2025). We require embodied world models to predict and envision what would happen if we take this action. For instance, an embodied world model needs to know that a cup will break if it rolls off the table and onto the ground, which is why the cup must be placed stably on the table. Multimodal large models enhance their physical understanding capabilities by constructing more physical data. However, how to improve the physical understanding of embodied world models remains an unsolved problem (Motamed et al., 2025). Currently, we believe we may need to explicitly enforce physical consistency; yet, a latent prediction model alone cannot explicitly define physical understanding. We have also found that enhancing the description of physical laws in conditional prompts can significantly improve the physical consistency of embodied world models, thereby greatly increasing the possibility of embodied world models becoming a physically coherent simulation.

**Commonsense-Infused Representation** Common sense capability is one of the core hallmarks that enables Large Language Models (LLMs) to move beyond "accurate text generation" and toward "truly understanding the world" (Zhao et al., 2023b). Similarly, the internal representations of world models must be infused with "world commonsense", encompassing capacities for causal and logical reasoning, alongside the generalization and transfer of world knowledge (Bansal et al., 2024). This foundational requirement can only be achieved through large-scale pre-training on diverse datasets ($\mathcal{D}_{\text{pretrain}}$). Yet a critical, unresolved challenge remains: bridging the domain gap between web-scale, third-person data and the distinct first-person, embodied perspective of a physical agent.

## 2.4 EMERGENCY, TOWARD A GENERATIVE PHYSICAL ENGINE

Like Richard Feynman said, "***Physics is the foundation of all the natural sciences, and without it, our modern world could not exist.***" Therefore, embodied intelligent robots must model the underlying physical laws of the world to perform various tasks in the real world better. One possible approach is to construct a world model with inherent physical modeling capabilities for robots. The mainstream approaches to realizing such a world model include two methods. One is based on generative AI combined with a differentiable physics engine. The other, grounded in video generation models, constructs a neural network-driven physics engine that possesses both intrinsic physical consistency and external high visual fidelity, as can be seen in Figure 3.



Figure 3: The technological development of world models in pursuit of intrinsic physical consistency has primarily followed two approaches. One possible approach is to construct a world model with inherent physical modeling capabilities for robots. The mainstream approaches to realizing such a world model include two methods. One is based on generative AI combined with a differentiable physics engine. The other, grounded in video generation models, constructs a neural network-driven physics engine that possesses both intrinsic physical consistency and external high visual fidelity.

For the first approach, the differentiable physics engine is derived from the numerical simulations. In the path of numerical simulation, the understanding of physics is relatively straightforward: one needs to transform the governing equations of various physical domains from the continuous to the discrete domain, enabling solutions to be computationally solved. Typical methods, e.g., the finite-volume method (FVM) and the finite-element method, are adopted as general-purpose simulation frameworks and are emphasized on high-resolution and high-order methods for accuracy. When single-domain problems are solved, the demand for coupled multiphysics simulations (e.g., fluid structure, electrochemistry, climate) grows, driven by rapid hardware changes. Multicore CPUs became ubiquitous, and early CUDA/GPU programming proved effective for many Partial Differential Equation (PDE) solvers. Extended FEM (XFEM) for cracks and interface problems, and phase-field models for fractures, appeared in the literature (Kirchhart et al., 2016).

Then, researchers began fusing data science with traditional simulation – e.g., using reduced-order models, adjoint-based optimization, and surrogates to accelerate design. Emphasis was placed on modular, open frameworks for coupling new algorithms (including machine learning components) into solvers. Smoothed Particle Hydrodynamics (SPH) became increasingly used in commercial

and research codes for fluid–solid interaction (Violeau & Rogers, 2016). Exascale machines and the integration of AI and data science became the core of simulation methods. Researchers tackle uncertainty quantification, digital twins, and inverse design as first-class objectives. Novel algorithms (quantum, neural PDE solvers) began to be explored, though classical methods continue evolving for maximum performance. Differentiable simulation is a significant innovation in the integration of automatic differentiation into FEM. The JAX-CPFEM example demonstrates this trend: by making the FEM pipeline differentiable, designers can use gradient-based optimization on geometry and material parameters (Hu et al., 2025a). Genesis re-designed the physics engine with generative AI. It integrated multiple classic physics solvers, such as rigid body, FEM, and SPH, as mentioned above, making the instruction easier to use (Genesis, 2024). **While these systems are also termed "world models" and their physical understanding relies on numerical solution of governing equations, they appear to overemphasize the role of world models as a simulator. Moreover, they neglect other critical functions of "world models" in the cognitive aspect.**

In the other approach, the core proposition of "world models" was distilled into a simple idea: for an agent to efficiently trial-and-error in the real world, it first needs to "dream" in its mind, that is, predict the future, test actions, and bring promising strategies back to reality within a manipulable latent environment (Ha & Schmidhuber, 2018). Diffusion transformer-based models have been able to achieve high visual fidelity and spatiotemporal consistency. OpenAI's Sora raised the bar for "long duration, camera motion, and multi-agent interaction," explicitly proposing a "world simulator" vision in public materials: not just generating "look-alike" clips, but maintaining interpretable 3D scenes and causal continuity. However, in engineering practice, the pursuit of visual fidelity often sacrifices physical consistency and action controllability (Brooks et al., 2024). Meanwhile, approaches like Genie, which "generate interactive environments," focus on the other end: not prioritizing high-resolution video but first ensuring the environment is action-controllable, temporally consistent, and playable, then gradually improving rendering quality; this aligns better with developers' "ease of use—cost" objective function in game engine scenarios (Bruce et al., 2024).

The next frontier for world models is to evolve into a Generative Physical Engine. This engine would go beyond abstract representations to simulate real-world dynamics, including fluid flow, material properties, and collisions. This capability is essential for building a robust and consistent foundation of common-sense knowledge and schemata within the model. And the model must satisfy the following:

**Causal Inference**: The model must understand how actions lead to specific outcomes. This allows for safe and effective planning in dynamic environments.

**Physical Understanding**: The model needs to encode the core concepts of motion, force, and spatial structure. This deep understanding of physical laws allows for more robust and adaptive behaviors.

**Interaction**: A generative physical engine enables agents to test actions in a simulated environment, allowing efficient training and exploration without the risks of real-world trial and error.

**Consistency**: The simulations generated must be temporally and physically consistent and free from the unrealistic artifacts that often plague current generative models.

## 2.5   THE COGNITIVE SCIENCE CONNECTION: WORLD MODEL IN MIND

The pursuit of an embodied world model represents a significant bridge between abstract computation and physical experience, drawing a direct parallel to human cognition. This "intuitive physics" mirrors our own mental processes, which allow us to simulate and reason about the world without direct interaction. This ability is a cornerstone of human intelligence, enabling us to make predictions and plan actions based on an internal model of reality.

In AI research, world models are usually utilized to simulate or predict the environment via generative models. For example, in an agentic system, the world model behaves like a simulation environment for the agent to explore and learn by interacting with the environment. Although these works declared that they are "embodied" since the agent is usually a robot. **However, we believe that relying solely on simulation or prediction falls short of embodying the cognitive sense. The contradiction lies in their consideration of the world model, which represents the external environment and the agent's physical state with it. Still, the world model as part of the mind model refers to an internal representation of mental states that captures regularities of**
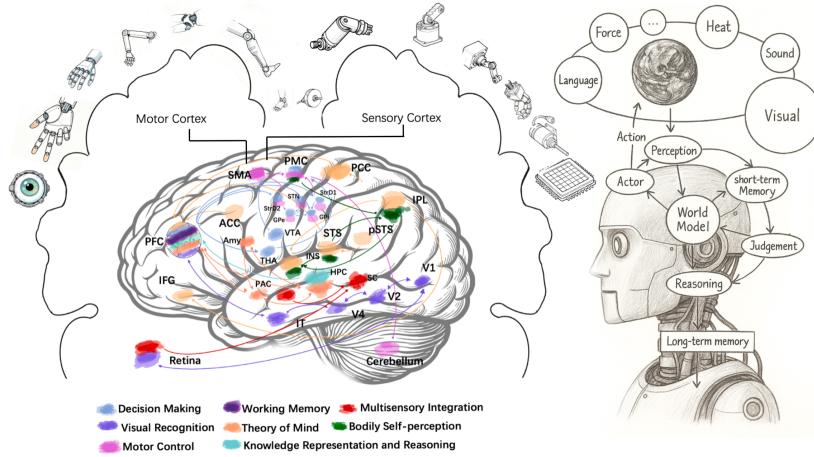
Figure 4: The architecture of an embodied agent with a world model. An intelligent agent perceives the environment through various sensory inputs (e.g., visual, sound, heat, force). These perceptions are processed by a World Model, which builds an internal, predictive representation of the environment. The model's predictions and past experiences, stored in short-term and long-term memory, inform Reasoning and Judgement. Based on this internal simulation, the Actor generates Actions that manipulate the real world. This closed-loop system allows the agent to learn the dynamics of its environment, plan for the future, and achieve complex goals. (Figure inspired by (Brain-inspired Cognitive Intelligence Lab, 2025))

**the world.** As shown in Figure 4, the hippocampus combines the Theory of Mind and Knowledge Representation and Reasoning, suggesting that our combination of DiT and VLM as a system can fulfill the hippocampus's primary function.

Moreover, modern neuroscience considers the brain as a distributed system, where cognition arises from the coordination of specialized but interconnected networks. Mental models are not the product of any single brain region, but rather the result of complex networks formed by multiple brain regions working in concert. Mental models are not the product of any single brain region, but rather the result of complex networks formed by multiple brain regions working in concert.

Inspired by this view, we propose an embodied world model that integrates the world model with both sensory processing and motor abilities. Within this framework, schemata serve as modular cognitive structures that encompass perception, prediction, reflection, and action. Our approach grounds intelligence in embodied interaction, offering a pathway toward more flexible and generalizable cognition. It can therefore dynamically interact with the physical world.

## 3  WoW World Model

We present WoW, an embodied world model built upon Self-Optimizing Predictive Hallucination Improving Agent (*SOPHIA*), our novel paradigm for enhancing physical reasoning through closed-loop self-refinement. Inspired by Neisser's Perceptual Cycle (Neisser, 1976)—*Schemata → Perception → Action → Schemata*, WoW organizes intelligent behavior into three interconnected stages.

- **Task Imagination (Schemata):** We introduce the novel *SOPHIA* paradigm for generating high-level plans and pixel-level future predictions, and our instantiated **WoW** model. in Section 4.1.

- **Experience Reflection (Perception):** A VLM agent verifies physical consistency and iteratively refines the imagined outcomes (see Section 4.2).

Figure 5: **Comparison of Diffusion, JEPA (Assran et al., 2025), and SOPHIA.** The **Predictor** generates a **Future** from the input **Context**. This outcome is then evaluated to produce a **reward**, which directs the **Refiner**. Finally, the **Refiner** leverages this reward and external **language/embedding** guidance to issue a corrective signal, iteratively improving the next prediction cycle.

- **Behavior Extraction (Action):** A test-time module that translates imagined trajectories into executable policies (see Section 4.3).

# 4 SELF-OPTIMIZING FRAMEWORK

At its core, WoW follows *SOPHIA* paradigm that integrates large language models with diffusion transformers to generate physically plausible futures under language guidance, as shown in Figure 5. Through this iterative cycle of *predict*, *critic*, and *refine*, WoW unifies imagination and reasoning as fundamental components of embodied intelligence.

SOPHIA improves its physical plausibility based on the following hypothesis: Empirically, we observe that more detailed language prompts lead to video generations that are both physically plausible and semantically precise. To formalize this intuition, we adopt the theoretical framework proposed by *'Critical of World Model'* (Xing et al., 2025), and present the following result:

**Hypothesis 1 (Completeness of Language Representation).** *Let* $\mathbf{x} = \{x_t\}_{t=1}^{T}$ *be a continuous input sequence with* $x_t \in \mathbb{R}^D$ *and* $\|x_t\| < K$. *For any* $\epsilon > 0$, *there exists a language system* $L_\epsilon = (V, N, f_\epsilon)$ *with vocabulary* $V$, *sentence length* $N < \infty$, *and mapping* $f_\epsilon : \mathbb{R}^{T \times D} \to V^N$, *such that for any* $\mathbf{x}, \mathbf{x}'$, *if* $\|\mathbf{x} - \mathbf{x}'\| \geq \epsilon$, *then* $f_\epsilon(\mathbf{x}) \neq f_\epsilon(\mathbf{x}')$.

This hypothesis assumes that language, when sufficiently expressive, can uniquely distinguish between arbitrarily similar physical sequences. The WoW allows abstracting complex dynamics (e.g., post-collision motion) into symbolic descriptions, enabling fine-grained control via language.

In the context of video diffusion, $\mathbf{x}$ represents a video segment at the pixel-level, while $f_\epsilon(\mathbf{x})$ denotes its corresponding prompt.

## 4.1 FOUNDATION VIDEO GENERATION WORLD MODEL

This section serves as the perceptual and imaginative engine of WoW, enabling the agent to simulate future dynamics from language instructions. This module is built upon three key components: (1) **Pretrain Data Preparation**, which constructs a large-scale dataset of language-conditioned physical interactions to support multimodal learning; (2) **Diffusion-Based Video Generation**, which leverages the DiT architecture to generate high-fidelity, temporally coherent video rollouts conditioned on task descriptions; and (3) **Solver-Critic Video Generation Agents**, which introduce an iterative refinement loop—where a solver generates plausible futures and a critic evaluates their physical consistency—enabling closed-loop imagination with self-correction. Together, these components allow the model to not only imagine what could happen, but to refine these imaginations toward physically plausible outcomes.

### 4.1.1 PRETRAIN DATA PREPARATION

We construct our training dataset through a multi-stage pipeline designed to ensure both quality and diversity. The process consists of four sequential stages: **Collection**, **Filtering**, **Refinement**, and **Rebalancing**. Unlike simply enlarging the dataset with indiscriminate samples, our approach emphasizes that *data quality plays a decisive role in model performance*, and carefully curated data

Figure 6: **Overview of the Video Diffusion World Model.** (a) Inference: a latent diffusion transformer predicts future frames from image observations and text-based action descriptions. (b) Training: DINO features supervise intermediate DiT representations via a token relation distillation loss to improve spatial-temporal modeling.

prove more effective than raw scale (Collaboration et al., 2023; Khazatsky et al., 2024; Radford et al., 2021; Northcutt et al., 2021; Zhang et al., 2017; Luo et al., 2024; An et al., 2024).

**Collection.** We collect thousands of hours of videos from multiple robotic platforms, including Agibot (Bu et al., 2025), Droid (Khazatsky et al., 2024), Robomind (Wu et al., 2025), and a large amount of in-house data. These sources cover a variety of embodiments and task scenarios, providing broad coverage across environments and robot types. This diversity serves as the foundation for building generalizable robot learning datasets.

**Filtering.** The collected data are processed through a series of filtering rules. Only RGB videos are retained, with BGR channels semi-automatically converted into RGB format for consistency. Static or non-informative sequences are removed, and a minimum length of 90 frames is enforced to ensure sufficient temporal context. In addition, we restrict the dataset to specific viewpoints such as head, wrist, and third-person perspectives, which best capture robot actions and task dynamics.

**Caption Refinement.** To further enhance the training signal, sparse textual annotations are expanded into dense descriptions using a pretrained VLM. Both uniform and sequential frame sampling are applied, ensuring coverage of both global context and local temporal transitions. Sparse and dense text annotations are combined with an approximate ratio of 1:4, and robot model identifiers are manually added into the text metadata. This step improves both the richness of supervision and the alignment between visual and textual modalities (Radford et al., 2021; Lin et al., 2024).

**Rebalancing.** Finally, we address imbalance across tasks by increasing the sampling probability of underrepresented tasks. This ensures that rare but important skills are not neglected during training, and improves robustness across diverse robotic behaviors (Northcutt et al., 2021).

Through this pipeline, we construct a dataset that is large in scale, carefully curated, temporally consistent, and densely annotated with semantic and physical labels—providing a robust foundation for training advanced robot learning models.

### 4.1.2 DIFFUSION-BASED VIDEO GENERATION

We adopt a video generation paradigm for the world model, treating the visual domain as the primary output modality due to its high information density. Our framework maps an initial state $s_t$ to its future state $\hat{s}_{t+1}$, where the hat indicates that it is predicted:

$$o_t : \{o_t, p_t, [a_t, C_{\text{pose}}, \dots]\} \xrightarrow{\text{World Model}} \hat{s}_{t+1} : o_{t+1} \tag{5}$$

where $o_t$ is the current visual observation, $p_t$ is a high-level textual instruction, and optional inputs such as $a_t$ (low-level action) or $C_{\text{pose}}$ (camera pose) provide finer control. The full realization of
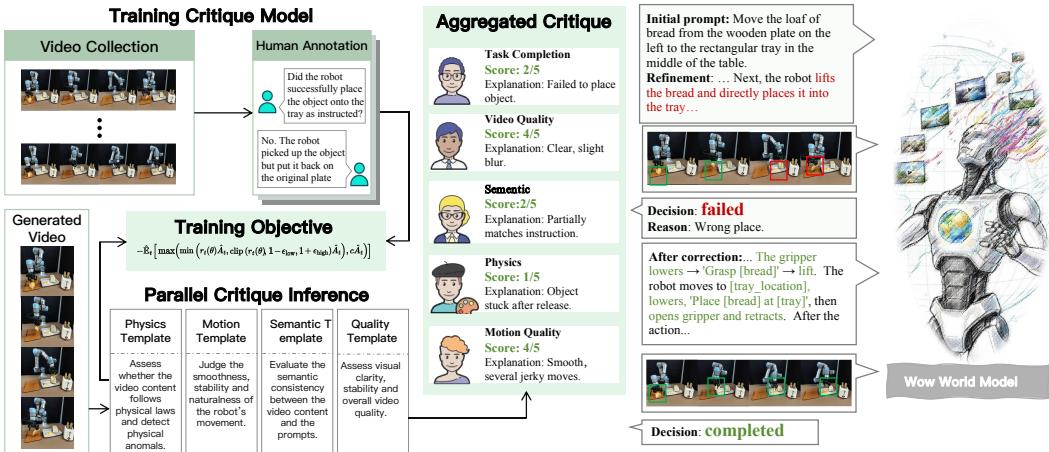
12

Figure 7: **Overview of Solver-Critic Video Generation Agents.** The left panel illustrates the Dynamic Critic Model Team, trained on annotated real and synthetic videos to evaluate physical plausibility. The right panel depicts the Refiner Agent, which iteratively rewrites prompts based on critic feedback and regenerates videos, forming a closed-loop optimization process.

this world model relies on specialized processing of each input modality. The following describes how textual, visual, and auxiliary signals are encoded and integrated to enable high-fidelity video prediction.

**Textual Conditioning.** We employ InternVL3-78B (Zhu et al., 2025) to output language instructions into descriptive narratives of the environment, camera pose, robot embodiment, and intended action. These narratives are embedded by a pre-trained T5 (Raffel et al., 2020) encoder and injected as conditioning signals into the DiT, ensuring stronger alignment between textual context and generated video.

**Visual Encoding.** Raw video inputs are compressed into compact latent representations via a spatio-temporal autoencoder. To enhance modeling of physical interactions, we apply a 3D Haar wavelet transform that decomposes each video cube into low-frequency components—capturing scene structure—and high-frequency sub-bands—preserving fine motion details such as object collisions and deformations. This spectral separation allows the model to allocate capacity more effectively toward dynamic events. Spatial and temporal downsampling further reduces dimensionality for efficient processing without sacrificing critical physical cues.

**Diffusion Transformer.** The denoising backbone is DiT (Peebles & Xie, 2023) composed of multi-head self-attention and feed-forward layers, enhanced with adaptive LayerNorm (adaLN) for timestep conditioning. Both absolute 3D positional embeddings and relative 3D RoPE are employed: the former preserves global coherence (e.g., trajectories), while the latter enforces local pixel-level causality (e.g., contact and continuity).

**Auxiliary Perception.** To strengthen initial state understandingYu et al. (2025), we inject features from a self-supervised visual representation model (DINOv2 (Oquab et al., 2023)) into intermediate layers of the DiT. These semantically grounded features improve pixel-level reasoning about object boundaries and spatial relationships, compensating for potential weaknesses in latent representations learned solely through noisy reconstruction objectives.

**Frame Decoding.** The decoder mirrors the encoder's hierarchical structure, progressively reconstructing high-resolution frames through spatial upsampling, inverse wavelet transforms, and self-attention refinement. This multi-stage decoding process ensures both long-horizon temporal coherence and physically plausible fine details—such as texture preservation during deformation or accurate collision recovery.

## 4.2 SOLVER-CRITIC VIDEO GENERATION AGENTS

Building upon the conceptual framework of WoW, this section details the Experience Reflection (Perception) stage—a key component of our closed-loop mechanism. We introduce a solver-critic

paradigm that enhances the physical consistency and realism of generated outputs. Unlike Simple ability of previous work (Soni et al., 2024; Chi et al., 2024), WoW has a comprehensive agent system, driven by a Refiner (Section 4.2.2), which, guided by a Dynamic Critic Model Team (Section 4.2.3), to iteratively improve the generated content. This dynamic workflow forms a Closed-Loop Generative Workflow (Section 4.2.4), ensuring that the model's output is continuously refined and verified.

### 4.2.1 FRAMEWORK OVERVIEW

Achieving physically plausible video generation for complex, long-horizon robotic tasks requires moving beyond unidirectional models to a closed-loop, agentic system capable of self-perception and optimization. We frame this generative process as a deliberative act, analogous to the interplay between the intuitive "System 1" and the analytical "System 2" cognitive modes (Weston & Sukhbaatar, 2023). In our framework, an initial video serves as a "proposal" (System 1), which is then subjected to a rigorous critique and refinement loop that embodies the structured reasoning of System 2.

This architecture transforms the model from a passive generator into an active problem-solver. Our solver-critic framework is built upon three core components: the Refiner Agent, which optimizes the input and generates video output; the Dynamic Critic Model, which evaluates the generated output; and the integrated closed-loop Workflow. Furthermore, we discuss how this architecture aligns with the Prover-Verifier paradigm (Kirchner et al., 2024), showcasing its potential to endow the generative process with a new level of cognitive depth.

### 4.2.2 REFINER AGENT IN WORLD MODEL

The quality of a generative model's output is highly dependent on its input prompt. For video generation in specialized fields like robotics, prompts must capture subtle physical details to produce plausible outcomes. However, manually crafting such high-quality prompts is a time-consuming and arduous process of trial and error. While the emerging field of Automatic Prompt Engineering offers systematic optimization methods for language and generation tasks (Khan et al., 2025; Agrawal et al., 2025; An et al., 2025), these general approaches are not directly tailored to the unique demands of physically-grounded video synthesis.

To address this challenge, we introduce the Refiner Agent, an autonomous system designed for test-time prompt optimization that does not require retraining the underlying video generation model. The agent takes a high-level user instruction and initiates an iterative refinement loop. In each iteration, a dedicated prompt rewriting module enhances the prompt's specificity and physical consistency. This rewriting process is explicitly guided by structured feedback from our Critic Model Team (Section 4.2.3), which identifies errors or missing details, such as adding constraints to prevent objects from passing through solid surfaces. Conceptually, this iterative process performs a guided search over the discrete prompt space, where the critic feedback functions as a "textual gradient" (Pryzant et al., 2023; Yuksekgonul et al., 2024). Our approach thereby transforms prompt engineering from a manual, trial-and-error task into a systematic, data-driven, closed-loop optimization process.

### 4.2.3 DYNAMIC CRITIC MODEL TEAM

Functioning as the 'verifier' in our iterative refinement loop, the Dynamic Critic Model Team is the second core component of our system. The need for this specialized critic arises because traditional metrics such as Fréchet Video Distance (FVD), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM) (Unterthiner et al., 2018; Huynh-Thu & Ghanbari, 2012; Hore & Ziou, 2010), while capable of assessing visual fidelity, are inadequate for evaluating physical realism—a critical bottleneck in the development of robust world models. To address this gap, we align with the emerging consensus that VLMs are the cornerstone of next-generation video assessment (Chen et al., 2024a). However, general-purpose VLMs lack the domain-specific precision required for tasks like robot manipulation. We therefore construct our specialized critic by fine-tuning a VLM on a curated Question-Answering (QA) dataset containing both real and model-generated videos of robotic operations. This dataset is structured to probe the model's understanding across five key dimensions: task completion, action success, physical plausibility of interactions (e.g., sta-

bility, deformation), kinematic smoothness, and overall quality. This targeted fine-tuning transforms the generalist VLM into a reliable expert verifier, instilling it with the specialized knowledge required to accurately assess the physical dynamics of robot interaction.

### 4.2.4 CLOSED-LOOP GENERATIVE WORKFLOW

As illustrated in Figure 7, our system integrates the Refiner Agent and Dynamic Critic Model into a closed-loop workflow that transforms video generation from a single-pass operation into an iterative refinement process. The loop initiates with a high-level user task, which the Refiner Agent expands into a detailed, physically-constrained prompt for our generation model (WoW). The resulting candidate video is then evaluated by the Dynamic Critic Model for physical plausibility and semantic coherence. If the video is judged 'incomplete' or 'failed', the critic provides structured feedback that the Refiner Agent incorporates to revise the prompt for the subsequent generation cycle. This iterative process of generation, critique, and refinement reframes video synthesis as an adaptive reasoning task. By endowing the generative pipeline with this self-corrective capability, our workflow enables the system to progressively converge on outputs that constitute a robust, physically grounded world model.

### 4.2.5 DISCUSSION: THE PROVER-VERIFIER PARADIGM FOR GENERATIVE WORLD MODELS

To understand our architecture on a deeper level, this section explicitly connects it to established theoretical frameworks in the field of artificial intelligence—namely, the Solver-Critic (Wang et al., 2025b; McAleese et al., 2024; Gou et al., 2023) and Prover-Verifier (Kirchner et al., 2024) paradigms. In these paradigms, one agent (the Prover/Solver) is responsible for generating a candidate solution, while another, often simpler or more specialized agent (the Verifier/Critic), is responsible for evaluating its correctness.

Our architecture provides a concrete implementation of the established Prover-Verifier and Solver-Critic paradigms (Wang et al., 2025b; McAleese et al., 2024; Gou et al., 2023; Kirchner et al., 2024). Within this framework, the Refiner Agent function as the Prover/Solver, responsible for proposing and iteratively refining candidate videos. The Dynamic Critic Model Team acts as the Verifier/Critic, tasked with evaluating the physical plausibility of these proposals. A key contribution of our work is being the first to successfully apply this paradigm-traditionally used for discrete, logical tasks such as mathematical proofs (Lin et al., 2025b) and code generation (Wang et al., 2025c)-to the high-dimensional, continuous, and stochastic domain of video generation.

The primary advantage of this approach is its ability to optimize for complex, non-differentiable objectives like "physical realism" without requiring an explicit, differentiable loss function. The Prover learns to generate outputs that are accepted by the Verifier, providing a powerful mechanism for instilling abstract values like physical common sense into generative models. To summarize, this framework paves the way for building physically and causally consistent world models suitable for robotics planning.

### 4.3 FLOW-MASK INVERSE DYNAMICS MODEL

The proposed Flow-Mask Inverse Dynamics Model (FM-IDM) is a video-to-action model that maps predicted video frames to real-world robot execution transitions. Instead of relying on model-specific features (Liao et al., 2025; Chi et al., 2025; Hu et al., 2025b), we adopt a pixel-level decoding approach, trading real-time performance for greater generality and accuracy (Ko et al., 2023; Tan et al., 2025). Designed as a plug-and-play module, our model is compatible with any visual generative world model, enabling system-level evaluation and facilitating reward extraction via embodied interaction.

**Task Formulation**    Given two consecutive visual observations $(o_t, o_{t+1})$ from the predicted video — each corresponding to the underlying robot states $(s_t, s_{t+1})$ — the goal is to infer the end-effector action $a_t$ that transitions the robot from $s_t$ to $s_{t+1}$. The inverse dynamics model $F_\delta$ takes the current frame $o_t$ and the corresponding flow $\mathcal{F}_{t \to t+1}$ as input, and outputs a predicted delta action $\hat{a}_t$:

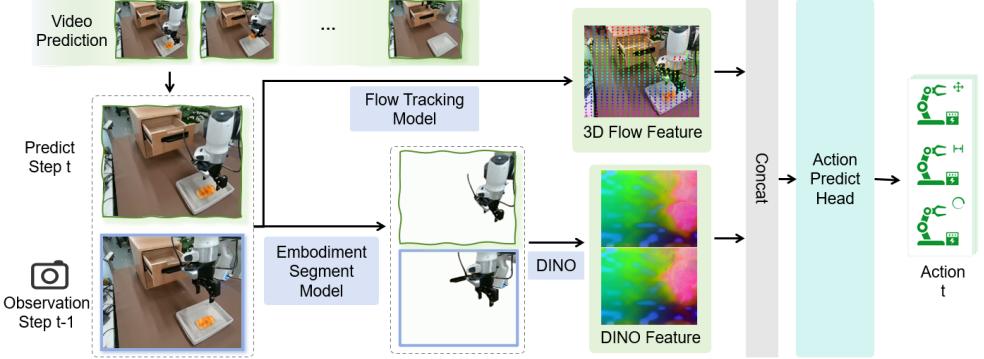$$\hat{a}_t = F_\delta(o_t, \mathcal{F}_{t \to t+1}) \tag{6}$$

Figure 8: Work flow of inverse dynamics model. Giving two frame predictions, our FM-IDM can estimate the delta End-Effect action of the robot.

**FM-IDM** To achieve this, we first estimate a motion field $\mathcal{F}_{t \to t+1}$ capturing the geometric transformation between frames. The estimated flow encodes both translational and rotational motion of the manipulator. We implement $F_\delta$ as a two-branch encoder-decoder network. We first fine-tuned a SAM (Kirillov et al., 2023) that process the masked current frame $o_t$ to extract scene and embodiment context; the other processes the optimal flow by CoTracker3 model (Karaev et al., 2024) $\mathcal{F}_{t \to t+1}$ to capture fine-grained temporal dynamics, as described in Figure 8. In conjunction with the with the DINO (Oquab et al., 2023) feature, we further use Multi-Layer Perceptron (MLP) as action head to learn the 7-DoF action feature. The training objection is as follows:

$$\min_\delta \ \mathbb{E}_{(o_t, o_{t+1}, a_t)} \ d\big(a_t, F_\delta(o_t, \mathcal{F}_{t \to t+1})\big) \tag{7}$$

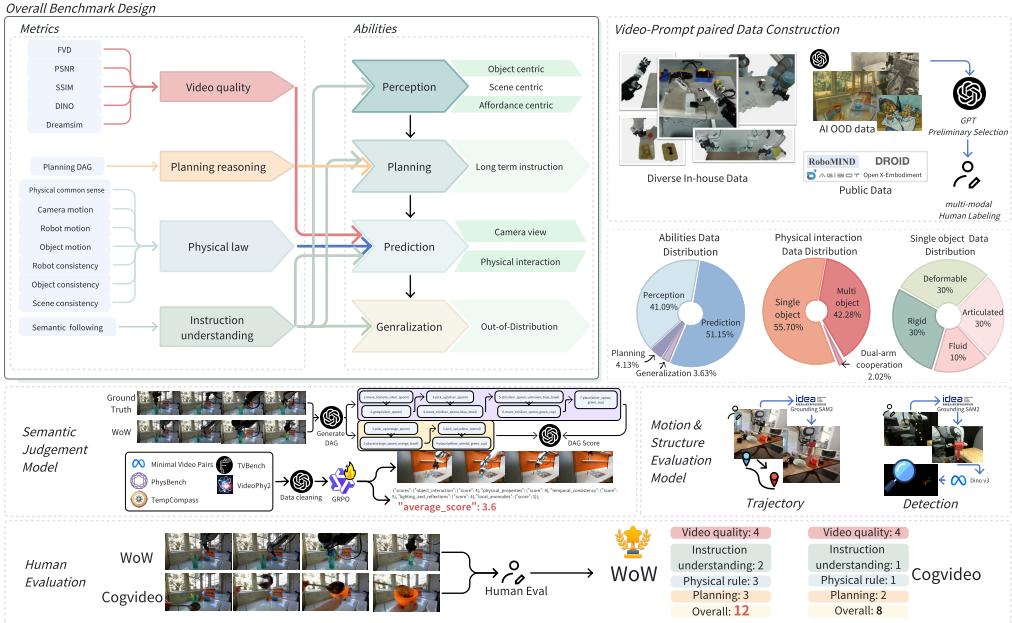where $d(\cdot, \cdot)$ denotes a weighted smooth L1 loss in the end-effector action space.

By explicitly modeling spatio-temporal correspondences, the model generalizes better across diverse tasks, background variations, and occlusions, and is robust to noise in video-based prediction.

**Embodiment-Centric End-Effector Action Dataset** To facilitate the learning of end-effector actions directly from visual input, we curate a dataset of 646k image–action pairs across 219 tasks, covering a broad range of manipulation scenarios. The dataset is carefully constructed to span a diverse action space and densely cover the reachable workspace of the robot, ensuring that the model learns from a comprehensive set of physically plausible end-effector configurations. More details of the implementation are included in Section 6.

**Real-World Feedback through IDM** At the action execution stage, rewards are grounded in physical feasibility and obtained through direct interaction with the environment. They may be defined in multiple ways: binary success/failure of task completion, distance-based metrics between predicted and actual end-effector positions, force/torque stability measures during contact, or energy-efficient motion profiles. Failures serve not merely as penalties but as corrective feedback that guides continual adaptation. Importantly, this reward can be further fed back to the world model, adjusting the model through Group Relative Policy Optimization (GRPO) for evolutionary visual generation (Xue et al., 2025).

## 5 WoWBench: A Multi-faceted Benchmark for Embodied World Models

In this section, we introduce WoWBench, a novel benchmark designed to rigorously evaluate embodied world models. We formulate the core evaluation task as conditional **video generation from an initial image and a text instruction (Image+Text-to-Video)**, a setting that directly probes a model's ability to understand a given state and execute a specified action. Moving beyond metrics of visual appeal, WoWBench is specifically engineered to assess a model's instruction understanding, planning and observation perception abilities and understanding of physically grounded dynamics in embodied settings. The overall design of our benchmark, which systematically connects core abilities, a principled data construction pipeline, and a multi-faceted evaluation protocol, is illustrated in

Figure 9: **The Overall Design of WoWBench.** Our benchmark is structured around five core components. **(Top-left)** A multi-faceted *Metrics* suite evaluates generated videos from four key perspectives: video quality, planning reasoning, physical rules, and instruction understanding. **(Top-center)** These metrics are designed to assess four fundamental *Abilities* of an embodied world model: Perception, Planning, Prediction, and Generalization. **(Top-right)** The benchmark is built upon a large-scale *Data Construction* pipeline that leverages diverse data sources (in-house, public, and AI-generated) and a human-in-the-loop annotation process (GPT-based selection followed by human labeling) to create video-prompt pairs. Three pie charts demonstrate the statistics of data distribution in different dimensions. **(Center)** The benchmark uses two different evaluation methods. Strong expert models for motion and consistency, and GPT or Fine-tuned VLM for instruction understanding and planning. **(Bottom)** We also asked 12 domain experts to perform a human evaluation on the generated videos.

Figure 9. We structure our presentation by first outlining the four core capabilities that a competent embodied world model must possess (Section 5.1). We then detail our task design and data curation process (Section 5.2), followed by our comprehensive evaluation metrics (Section 5.3).

## 5.1 CORE EVALUATION DIMENSIONS

We posit that a truly effective embodied world model must demonstrate mastery across four fundamental and orthogonal dimensions, as outlined in our benchmark design (Figure 9, center).

**Perception Understanding** A world model must first accurately perceive and represent the environment in order to enable more reliable subsequent prediction and planning. We assess this through tasks requiring fine-grained **object recognition** (Cheng et al., 2024b;a; Majumdar et al., 2024; Chow et al., 2025)(attributes like color, shape, number, and size), **spatial understanding** (Yang et al., 2025a; Cheng et al., 2025; Song et al., 2025; Du et al., 2024) (relative positions and arrangements), and **affordance recognition** (Gibson, 2014; Nasiriany et al., 2025; Cheng et al., 2025) (identifying interactive parts of objects). Each dimension contains a different amount of data. In the object sub-dimension, we included approximately 143 samples, with around 20 samples allocated to each attribute category (color, number, shape, size, type) and 50 samples to the function category. In the spatial sub-dimension, we assigned 46 samples, while the affordance sub-dimension was filled with 60 samples.

**Predictive Reasoning**   This dimension evaluates the model's internal physics engine. Given an initial state and an action, the model must generate a future that respects core physical principles such as **object permanence** (Bansal et al., 2024; Chow et al., 2025), **collision dynamics** (Meng et al., 2024; Chow et al., 2025; Bansal et al., 2024; Motamed et al., 2025), and **trajectory plausibility** (Qin et al., 2024; Li et al., 2025a; Duan et al., 2025; Zheng et al., 2025; Yue et al., 2025). This directly probes the model's capacity to function as a world simulator (Brooks et al., 2024). Therefore, we design several sub-dimensions that focus on these principles, as illustrated by the pie chart in the center-right of Figure 9. We select both objects with no occlusion, 107 samples, and objects with semi-occlusion, 54 samples, as our different camera views for evaluating varying levels of object visibility. Also, for collision dynamics, we further subdivide the dimension into single-object operation, multi-object interaction, and dual-arm cooperation. The single-object operation data has 83 samples, which are distributed in the ratio of 30, 30, 30, 10 for rigid, deformable, articulated, and fluid. The multi-object interaction part has 63 samples, covering the interaction between rigid body-rigid body, rigid body-deformable object, and rigid body-fluid. For dual-arm cooperation, now we have 3 samples, but we will continue to collect more data.

**Decision-making and Planning**   Embodied agents must execute long-horizon tasks (Chen et al., 2024b; Sermanet et al., 2023; Li et al., 2024; Cheng et al., 2024a; Yang et al., 2025b). Therefore, we assess a model's planning ability by challenging it to generate coherent video sequences for complex instructions. This requires implicitly understanding **task decomposition** (Chi et al., 2024; Tian et al., 2025) into key sub-goals and respecting their **causal dependencies**. Hence, we collect 25 samples to fill in this long-term planning task, and transform the text instruction into a suitable description for the world model to plan. In order to evaluate the planning ability in the world model, we refer to the metric from RoboBench. As illustrated in the center-left of Figure 9, first, we extract the key steps from the predicted video using Gemini-2.5-flash (Comanici et al., 2025), for instance, 1-grasp(green block), 2-pick up(green block), 3-place(green block, yellow block), and so on, then transform them into a Directed Acyclic Graph(DAG) for further key action extraction and comparison with groundtruth DAG. The detailed metric will be extended in Section 5.3.

**Generalized Execution**   A universal world model should not only perform well on the In-Distribution data, but it should also generalize beyond the data it has seen before to demonstrate its generalization ability. For this reason, we test generalization on the in-house robot data that we collected, by using GPT-5 (OpenAI et al., 2024) to perform style transfer or image editing on it, and generating images that the world model had never seen before. We also collected some world-famous masterpiece paintings, such as *"Girl with a Pearl Earring"*, and asked the world model to execute the task instructions human created. These two types of images constitute our **Out-of-Distribution(OOD)** dimension. Here, in total, we provide 20 data samples, and demonstrate some cases in Figure 16.

## 5.2   Task Design and Data Curation

To rigorously evaluate these capabilities, we developed the principled, semi-automated data curation pipeline shown in Figure 9 (Top-right). Our dataset is built from a mixture of open-source robotics data (e.g., RoboMIND (Wu et al., 2025), DROID (Khazatsky et al., 2024)), in-house collected trajectories, and AI-generated OOD data to ensure diversity.

Our process begins by leveraging GPT-4o (OpenAI et al., 2024) as an intelligent annotator for preliminary data filled in our dimensions. For each candidate video-instruction pair, GPT-4o scores its relevance against the definitions of our four core capability dimensions. This allows for efficient, large-scale sorting of data into targeted evaluation buckets. Following this automated stage, human experts perform a verification step to ensure gold-standard quality, resolving ambiguities and filtering out misclassifications. This human-in-the-loop approach ensures both scalability and accuracy.

Finally, expert annotators identify the optimal initial frame for each task and provide crucial annotations. The final benchmark consists of tuples, each containing: 1) a natural language **instruction**, 2) an initial **image**, 3) the ground-truth **video**, and 4) annotated **keypoints** for tracking.

## 5.3 Multi-faceted Evaluation Metrics

Our evaluation protocol is a suite of metrics designed to be as comprehensive as the capabilities we measure (Figure 9, left). We introduce several novel metrics alongside standard ones, grouped by the property they assess.

**Visual Fidelity and Temporal Consistency.** While we report standard video quality metrics (FVD(Unterthiner et al., 2018), SSIM(Wang et al., 2004), PSNR(Fardo et al., 2016), DINO (Oquab et al., 2023), Dreamsim (Fu et al., 2023)) in the appendix, our primary contribution is a novel metric for temporal consistency with high diagnostic power. While pixel evaluation can give a overall visual quality comparison.

**Mask-guided Regional Consistency.** To better disentangle inconsistencies in the background, robot arm, or manipulated object, we propose **Mask-guided Regional Consistency** as similar to EWMBench (Yue et al., 2025). As illustrated in Figure 9(Center-right), we first use the Grounded-SAM2 (Ren et al., 2024) with human annotation to obtain masks for the robot arm, the manipulated object(s), and the background in each frame. We then compute region-specific embeddings using a vision foundation model (e.g., DINOv3 (Siméoni et al., 2025)) and measure cosine similarity across time for each region separately. This allows us to pinpoint the source of temporal flaws—for instance, identifying a "jittery" robot arm even when the object and background are stable.

**Instruction Understanding and Semantic Correctness.** We use GPT-4o (OpenAI et al., 2024) as a scalable evaluator to assess semantic alignment with the given instruction. Depending on whether ground-truth (GT) video is available, we adopt two protocols:

- **With Ground-Truth:** We first prompt GPT-4o to extract structured descriptions (Initial, Processing, Final states) from both the generated and GT videos. A vision–language model then scores their *Caption Score*. In addition, GPT-4o evaluates the generated video against the instruction to produce a *Sequence Match Score* (order of actions) and an *Execution Quality Score* (1–5 scale). *We report all three metrics in this setting.*
- **Without Ground-Truth (OOD):** When GT is unavailable, we only assess instruction adherence: GPT-4o directly analyzes the generated video to output the *Sequence Match Score* and the *Execution Quality Score*. *Only these two metrics are reported in this setting.*

**Physical and Causal Reasoning.** To quantify the physical plausibility of generated videos, we compute:

- **Trajectory Consistency:** To compare the trajectory between generated videos and ground-truth counterparts, we track both the end-effector and object trajectories. In our WoW-Bench, we leverage SAM2(Ravi et al., 2024), given a few representative points in the initial frame, to follow the motion of objects in both videos. Trajectory similarity is then evaluated using a complementary set of metrics: Mean Euclidean Distance (MED) (Dokmanic et al., 2015) to capture average deviation, Dynamic Time Warping (DTW) (Müller, 2007) to assess temporal alignment, and Fréchet Distance (Eiter & Mannila, 1994) to measure worst-case path similarity.
- **Physical common sense:** Physical common sense covers dimensions ranging from object interaction and properties to temporal consistency, lighting, fluid dynamics, and local anomalies. To automatically score these six distinct dimensions in generated videos, we collected several datasets (Krojer et al., 2025; Chow et al., 2025; Bansal et al., 2025; Cores et al., 2025; Liu et al., 2024; Zhang et al., 2024) and converted them into instruction-tuning datasets to fine-tune Qwen-2.5-VL (Bai et al., 2025) to enhance the understanding and consistency of physical law and employed a 1-to-5 scoring scale across categories.

**Planning and Task Decomposition.** To evaluate long-horizon planning, we refer to the metric of RoboBench based on Directed Acyclic Graphs (DAGs). We first parse the natural language instruction and ground-truth video into a ground-truth plan DAG, where nodes are atomic actions and edges represent dependencies. This representation flexibly handles non-unique but valid action

orderings. We then compare the model-generated plan (which also uses the same approach to infer from the video) to the ground-truth DAG using three scores:

1. **Key-step Recall** ($R_k$)**:** The fraction of essential ground-truth steps the model executes.

2. **Sequential Consistency** ($R_s$)**:** The normalized length of the longest correctly ordered sequence of key steps.

3. **Key-step Precision** ($P_k$)**:** The fraction of predicted key steps that are correct and non-superfluous.

The final planning score $S_{\text{plan}}$ integrates these aspects to reward both completeness and correctness:

$$S_{\text{plan}} = (0.5 \times R_k + 0.5 \times R_s) \times P_k \tag{8}$$

## 5.4 OVERALL BENCHMARK SCORE

**Setup.** For each model $i$ and metric $m$, we map the raw measurement $x_{i,m}$ to a common desirability score $s_{i,m} \in (0, 100)$ via a monotone parametric mapping applied after an absolute pre-scaling to $[0, 1]$. We then aggregate desirability scores by weighted arithmetic means at both metric-group and overall levels.

**Pre-scale to $[0, 1]$ with absolute anchors.** Let $L_m < U_m$ be fixed anchors for metric $m$ (documented per-metric). Define the clipping operator $\text{clip}(u; a, b) = \min\{\max\{u, a\}, b\}$. We first clamp raw values to $[L_m, U_m]$, then linearly map to $[0, 1]$; for "higher-is-better" (HIB) metrics:

$$\hat{x}_{i,m}^{\text{HIB}} = \frac{\text{clip}(x_{i,m}; L_m, U_m) - L_m}{U_m - L_m} \in [0, 1],$$

and for "lower-is-better" (LIB) metrics:

$$\hat{x}_{i,m}^{\text{LIB}} = 1 - \frac{\text{clip}(x_{i,m}; L_m, U_m) - L_m}{U_m - L_m} \in [0, 1].$$

We use absolute anchors for two common metrics:

**PSNR (HIB):** $L_{\text{PSNR}} = 0$, $U_{\text{PSNR}} = 50$ (truncate $x \le 50$ before scaling);

**FVD (LIB):** $L_{\text{FVD}} = 0$, $U_{\text{FVD}} = 2000$ (truncate $x \le 2000$).

For other metrics, $(L_m, U_m)$ are fixed per protocol (e.g., theoretical bounds for bounded scales, task-specific absolute targets for unbounded ones).

**Monotone parametric mappings.** After pre-scaling, we apply a single-parameter monotone transform $f_m(\cdot; \theta_m)$ and then rescale to $(0, 100)$:

$$s_{i,m} = 100 \, f_m(\hat{x}_{i,m}; \theta_m), \qquad s_{i,m} \in (0, 100).$$

We consider the following families (all are strictly increasing on $[0, 1]$):

**Power (Gamma):** $f_\gamma(x) = x^\gamma, \quad \gamma > 0$;

**Logit temperature:** $f_T(x) = \sigma(\text{logit}(x)/T), \quad T > 0, \ \sigma(t) = \frac{1}{1+e^{-t}}$;

**Tanh slope:** $f_\kappa(x) = \frac{1}{2}(\tanh(\kappa(2x - 1)) + 1), \quad \kappa > 0$.

In practice, $\gamma > 1$ accentuates the high end, while $T < 1$ or $\kappa > 1$ expands the mid-range and compresses extremes. For numerical stability with logit we use a small $\varepsilon$ (e.g., $10^{-6}$) and replace $x$ by $\text{clip}(x; \varepsilon, 1 - \varepsilon)$ only inside $\text{logit}(\cdot)$.

**Parameter selection and freezing.** For each metric $m$, $\theta_m \in \{\gamma, T, \kappa\}$ is selected on a fixed development set by maximizing the Fisher-$z$ averaged Pearson correlation between $f_m(\hat{x}; \theta)$ and human ratings across $K$-fold CV; Spearman correlation is used as a tie-breaker. The chosen $\theta_m$ is then *frozen* and applied to all evaluations.

**Intra-group averaging (uniform).**   Metrics are organized into groups $g$ (e.g., *quality*, *instruction*, *physical*, *planning*). For model $i$, let $\mathcal{M}_g$ be the set of metrics in group $g$ that are available for $i$, and $N_{i,g} = |\mathcal{M}_g| > 0$. We compute the group score as the simple arithmetic mean:

$$G_{i,g} = \frac{1}{N_{i,g}} \sum_{m \in \mathcal{M}_g} s_{i,m}.$$

**Aggregation (weighted arithmetic mean).**   Let nonnegative group weights $\{W_g\}$ sum to one over the groups available for model $i$. The overall score is

$$O_i = \sum_g \widetilde{W}_{i,g}\, G_{i,g}, \qquad \widetilde{W}_{i,g} = \frac{W_g}{\sum_{h \in \mathcal{G}_i} W_h},$$

where $\mathcal{G}_i = \{g : N_{i,g} > 0\}$. For an unweighted overall mean across groups, set $W_g \equiv 1$.

## 6   EXPERIMENT: EVALUATING GENERATIVE WORLD MODELS

Evaluation of our world foundation model is structured into four main parts, covering model comparisons, scaling law analysis, generalization, and real-world robotics. Prior to detailing these experiments, we provide a comprehensive overview of our training data in Section 6.1. The comparison of different models in Section 6.2 examines the impact of varying pre-training methods and model architectures on performance. The scaling law analysis in Section 6.3 measures performance across varying data volumes, trainable parameters and task difficulty levels. Generalization capabilities are tested against novel scenes, objects, and actions to ensure robustness in Section 6.4. Finally, practical deployment is assessed by the model's ability to generate executable actions in Section 6.5.

### 6.1   TRAINING DATA

Our training dataset was meticulously curated to provide a rich and diverse foundation for learning a generalizable world model. It comprises 2.03 million video clips, totaling over 7,300 hours of interaction footage, which corresponds to approximately 633 million frames sampled at a consistent 24 frames per second. To foster robust generalization, the data was collected from over 200 procedurally generated simulated scenes, spanning contexts from complex household environments (e.g., kitchens, living rooms with cluttered objects) to structured industrial settings (e.g., warehouses, assembly lines). Crucially, the dataset features a diverse collection of 12 distinct robot embodiments to ensure the model learns a wide range of physical dynamics and morphologies. The collection is dominated by industrial manipulators, with a significant emphasis on both single-arm and dual-arm configurations. The primary data sources include trajectories from the dual-arm Franka FR3, the single-arm UR5e, and the dual-arm UR5e, which together constitute a substantial portion of the dataset. To further broaden the diversity, we also incorporate data from various other platforms, including multiple Franka Emika Panda setups and several specialized configurations from the ARK, AgileX, and Tienkung series, ensuring a broad spectrum of kinematic properties and action spaces are represented. For pre-processing, all video sequences were captured at a native resolution of 640×480 and subsequently upsampled to 720×1024 pixels to align with our model's architectural input requirements. We applied a rigorous filtering pipeline to ensure data quality, which led to the exclusion of approximately 75% of the initial raw data. This high discard rate was a deliberate choice to remove trajectories with simulation instabilities, severe collisions, task failures, and periods of static inactivity, thereby ensuring the final dataset consists of high-quality, meaningful interactions.

### 6.2   MODEL COMPARSION EXPERIMENT

In Figure 10, we compare the overall performance of six models on the WoWBench benchmark, including CogVideoX (Yang et al., 2025d), Wan2.1 (Wan et al., 2025), Cosmos-Predict (Agarwal et al., 2025), our post-trained versions and proposed **WoW**. This comparison aims to explore the effects of different model architectures and data pretraining strategies on scaling laws. The scatter plot suggests a positive correlation between final performance and the volume of training data combined with the adoption of more recent model architectures.

Table 1: **Comparative analysis of foundational video generation models.** We benchmark our **WoW-DiT** against SOTA models using direct text-to-video generation. All metrics: higher is better. Best results are **bold** with  highlight.

| Model | Base | Human Evaluation | | | | | Autonomous Evaluation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | VQ | IF | PL | Plan | Overall | VQ | IF | PL | Plan | Overall |
| Cogvideo | cogvideo | 3.29 | 1.52 | 1.73 | 1.30 | 7.84 | 38.52 | 54.09 | 63.30 | 2.32 | 39.56 |
| Cosmos-Predict1 | cosmos1 | 2.84 | 2.60 | 2.41 | 2.49 | 10.34 | 39.06 | 61.46 | 59.05 | **7.47** | 41.76 |
| Wan2.1 | wan | 3.49 | 1.79 | 2.30 | 1.62 | 9.21 | 40.23 | 56.85 | 59.66 | 5.6 | 40.59 |
| Cosmos-Predict2 | cosmos2 | 3.18 | 2.33 | 2.31 | 2.27 | 10.09 | 46.81 | 56.80 | 60.56 | 6.67 | 42.71 |
| *Our Foundational Model* | | | | | | | | | | | |
| **WoW-DiT** | cosmos1 | 3.12 | 2.86 | 2.78 | 2.84 | 11.60 | 49.35 | 69.68 | 62.28 | 2.89 | 46.05 |
| **WoW-DiT** | wan | **4.09** | 2.60 | **3.16** | 2.52 | 12.37 | **55.38** | 62.16 | 63.75 | 4.74 | 46.51 |
| **WoW-DiT** | cosmos2 | 3.76 | **3.19** | 3.03 | **3.36** | **13.34** | 54.12 | **70.36** | **66.18** | 6.88 | **49.39** |

Table 2: Autonomous evaluation of models with a self-optimization framework, using agents for refinement.

Table 3: Data scaling law comparison in PBench.

| Model | Base | VQ ↑ | IF ↑ | PL ↑ | Plan ↑ | Overall ↑ |
|---|---|---|---|---|---|---|
| cosmos1 + Agent | cosmos1 | 35.43 | 61.07 | 53.78 | 8.23 | 39.63 |
| cosmos2 + Agent | cosmos2 | 49.7 | 75.96 | 64.66 | **11.77** | 50.53 |
| **WoW + Agent** | **cosmos1** | 59.39 | 72.54 | **69.71** | 4.26 | 51.47 |
| **WoW + Agent** | **wan** | **60.53** | 50.83 | 67.48 | 6.75 | 46.40 |
| **WoW + Agent** | **cosmos2** | 56.82 | **76.16** | 67.15 | 7.76 | **51.97** |

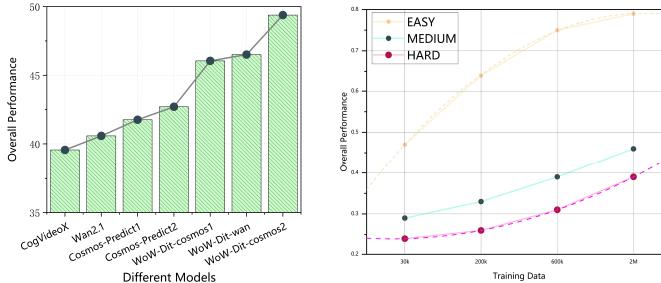| Data | PBench | | |
|---|---|---|---|
| | VLM ↑ | Qual. ↑ | Overall ↑ |
| 30k | 0.3901 | 0.3323 | 0.3612 |
| 200k | 0.5920 | 0.3790 | 0.4855 |
| 600k | 0.6240 | 0.3914 | 0.5077 |



Figure 10: **Performance Comparison Across Different Models in WoWBench.** The results indicate that all models subjected to post-training demonstrate superior performance compared to their respective baselines.
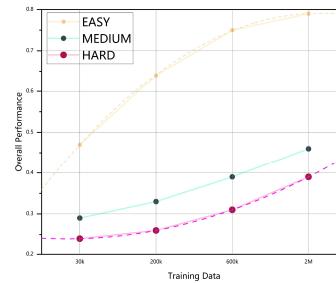


Figure 11: **Scaling Curves for Training Data.** We divide the benchmark into three levels of difficulty: Easy, Medium, and Hard. The left figure shows that as training data increases from 30k to 2M, performance on the Easy tasks begins to saturate, while the Hard tasks continue to benefit from more data.
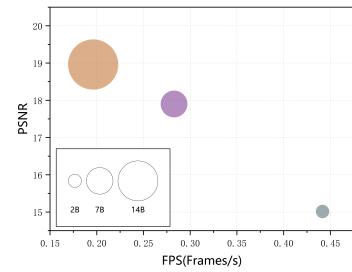


Figure 12: **Visual Quality Comparison Among scaling Model Size.** An analysis of inference speed and performance for models of varying sizes, specifically 2B, 7B, and 14B parameters. Performance is evaluated using the low-level metric, PSNR.

## 6.3 QUALITATIVE EXPERIMENT

**Scaling in Training Data.** To investigate the scalability of proposed world model, we performed supervised fine-tuning on datasets of varying sizes, specifically 30k, 200k, 600k, and 2M samples in Table 3. The 30k samples constitute a subset of the Robomind dataset (Wu et al., 2025), while the remainder is composed of data from the Agibot (Bu et al., 2025), Droid (Khazatsky et al., 2024), and Robomind datasets. Our findings empirically validate the scaling laws that govern model performance as a function of data volume. We observe a clear power-law relationship between the increase in data and the performance improvement, as measured by pbench and proposed WoW-Bench. Specifically, as the dataset size grows from 30k to 2M, the FVD decreases following a predictable power-law curve, with the most substantial gains observed when scaling from 600k to 2M. This result strongly suggests that our world model's capabilities are not yet saturated by the available data and that further performance improvements can be achieved by continuing to scale

Figure 13: **Performance Comparison Across Different Models in Detail Metrics in WoWBench.** Different color blocks stand for different dimensions in WoWBench. In every block, we demonstrate intuitive charts to present detailed scores in varied metrics in our WoWBench.

the training dataset. This adherence to well-established scaling laws provides strong evidence for the robustness of our training methodology and the potential for further significant performance gains with increased data and computational resources.

**Scaling in Model Size.** To investigate the impact of model scale, we evaluated 2B, 7B, and 14B parameter variants of our DiT model, finding a strong positive correlation between model size and performance that aligns with established neural scaling laws. The 7B model shows a substantial 19.22% performance improvement over the 2B model, whereas the 14B model yields a more modest 5.91% gain over the 7B model. This indicates that performance gains decelerate significantly as the parameter count increases. In terms of inference efficiency, the 7B model is 44.16% faster than the 14B model, and the 2B model is another 56.21% faster than the 7B. This highlights the critical trade-off between performance and efficiency that must be considered for practical World Foundation Model deployment.

**Scaling in Different Tasks.** To facilitate a fine-grained quantitative analysis of scaling laws under various factors, we classify the samples in WoWBench by difficulty. The classification, based on object properties, action complexity, task duration, and environmental factors, yielded 231 *Easy*, 237 *Medium*, and the remaining samples as *Hard*.
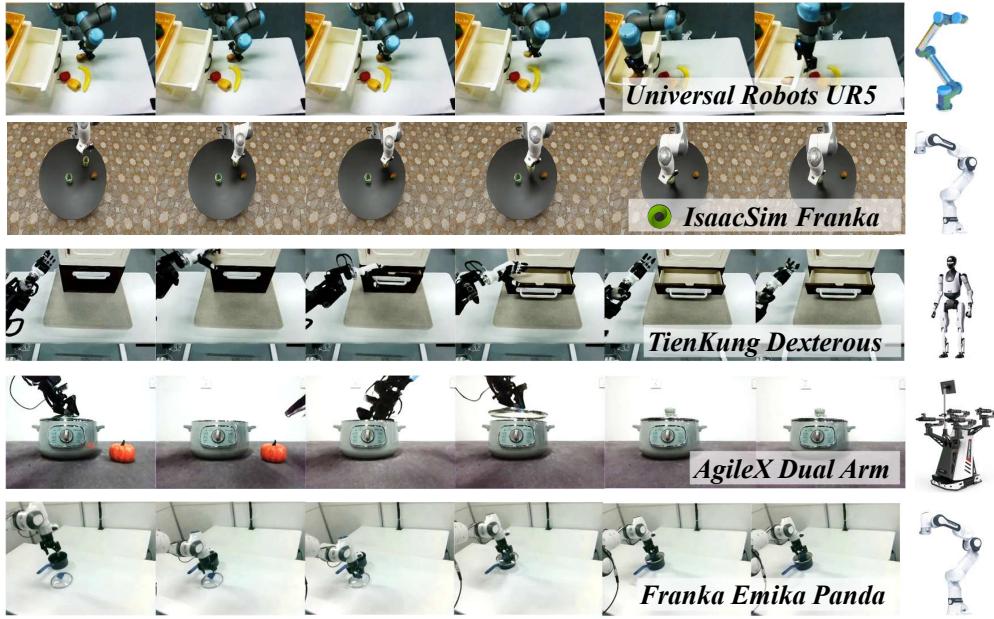
Figure 14: **Cross-Embodied Generalization Ability** Case Study in generalization ability of different robot types.

## 6.4 GENERALIZATION DISCUSSION

### 6.4.1 VISUAL EVALUATION

We qualitatively evaluate the generalization capabilities of our World Foundation Model across three critical axes: robotic embodiment, task repertoire, and other variations. *Cross-embodiment generalization* is demonstrated in Figure 14. Proposed world foundation model successfully follows instructions across a diverse set of embodied hardware platforms without any fine-tuning. The model is validated on a wide range of robotic platforms, including the UR5 and Franka industrial manipulators, the Franka Sim simulation, the Agilix parallel-kinematic arm, and the complex Tiangong dexterous hand. This result substantiates our model's ability to learn an embodiment-agnostic representation of physical interactions, decoupling the task goal from the specific kinematics and dynamics of the embodied platform. Second, Figure 15 illustrates the model's task generalization, showcasing its proficiency across a repertoire of 15 distinct manipulation skills that range from simple primitives like pull, push, and move, to complex, contact-rich behaviors such as press button, unstack, and tie. The model's proficiency across this action space highlights its capacity to learn a compositional skill representation rather than merely memorizing individual task solutions. Figure 16 elucidates the model's remarkable robustness to *profound domain and attribute variations*. It maintains successful task execution across drastically different visual styles, including photorealistic, pencil sketch, and oil painting. Furthermore, it adeptly handles varied object properties, manipulating both rigid bodies and fluids, from extremely small to regular-sized objects. The model also exhibits invariance to initial state configurations, consistently picking a red pen from a cup regardless of its spatial pose. Collectively, these visualizations provide compelling evidence that the proposed model acquires a truly abstract and foundational understanding of the world, reasoning about underlying physical principles rather than overfitting to superficial contextual cues.

### 6.4.2 QUANTITY EVALUATION

Building upon the preceding qualitative analysis, we conduct a rigorous quantitative evaluation across the benchmark's diverse material properties and physical phenomena. As detailed in Table 4, we assess four models using a suite of five metrics including FVD, WorldScoreBench, PhysGen, DreamSim, and execution quality score. The results unequivocally demonstrate the superiority of Cosmos and WAN over the other models. This performance gap is particularly pronounced in chal-
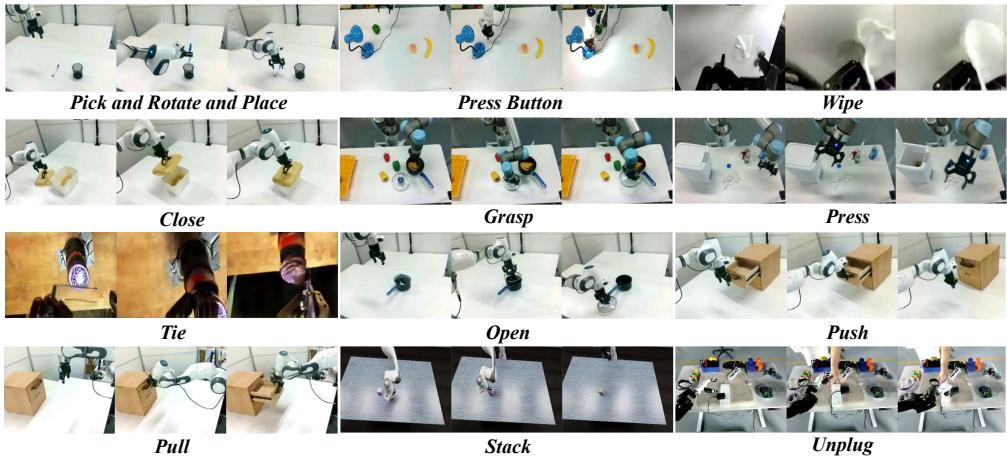
Figure 15: **Robot-Action Generalization Ability** Generalization ability in action typies.



Figure 16: **More Generalization Ability** Case Study in generalization ability of five other aspects.

lenging scenarios involving complex dynamics. For instance, Cosmos achieves the highest Phys-Gen scores, outperforming the next-best model by over 5% on tasks with deformable objects, which suggests a more accurate internal representation of non-rigid body mechanics. In parallel, world foundation model based on Cosmos consistently records the lowest FVD scores in fluid and optical physics simulations, indicating a state-of-the-art capability for rendering visually realistic and temporally coherent complex scenes.

## 6.5 REAL WORLD ROBOT MANIPULATION

We first selected 20 representative manipulation tasks from the 219 tasks covered in the dataset for preliminary evaluation. In the process of implementing, we observed four common categories of task failure:

- Unintended collisions with objects during multi-degree-of-freedom (DoF) control;
- Substantial inaccuracies in rotational movements;
- Insufficient translational precision of the end-effector; and
- Incorrect gripper opening or closing.

Based on these observations, we propose a three-tier difficulty classification scheme for pixel-based action generation tasks, defined according to the required DoF and precision constraints of end-effector control during task execution. Specifically, tasks that require at least 5 DoFs or tolerate errors below $2\,\mathrm{cm}/10°$ are categorized as *hard*. Tasks that require at least 4 DoFs or involve simple collision avoidance are categorized as *medium*. All remaining tasks are classified as *easy*.

Following this classification, we conducted video replay experiments to assess the training performance of the IDM model. Representative examples of task executions across different difficulty

Table 4: Comparative analysis of world models on physics simulation and visual benchmarks. Due to the number of metrics, the table is split into two parts for readability. The best score in each category is highlighted in **bold**.

| Model | Rigid | | | | | Soft | | | | | Fluid | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FVD | PhyGen | WMB | DreamSim | EQS | FVD | PhyGen | WMB | DreamSim | EQS | FVD | PhyGen | WMB | DreamSim | EQS |
| WoW-CogVideoX | 75.1 | 72.3 | 70.1 | 68.9 | 71.5 | 70.2 | 68.1 | 66.5 | 67.3 | 68.0 | 65.4 | 63.2 | 61.9 | 64.1 | 62.8 |
| WoW-SVD | 80.3 | 78.5 | 77.9 | 79.1 | 78.8 | 76.5 | 75.1 | 74.3 | 77.0 | 75.4 | 71.0 | 69.8 | 68.2 | 70.1 | 69.5 |
| WoW-Wan | 82.5 | 81.0 | 80.2 | 83.1 | 81.7 | 79.8 | 78.5 | 77.1 | 80.4 | 79.0 | 75.3 | 74.1 | 72.5 | 76.2 | 74.9 |
| WoW-Cosmos | **91.2** | **90.5** | **89.8** | **92.3** | **90.9** | **88.6** | **87.9** | **86.5** | **89.1** | **88.2** | **84.7** | **83.2** | **81.6** | **85.0** | **83.5** |

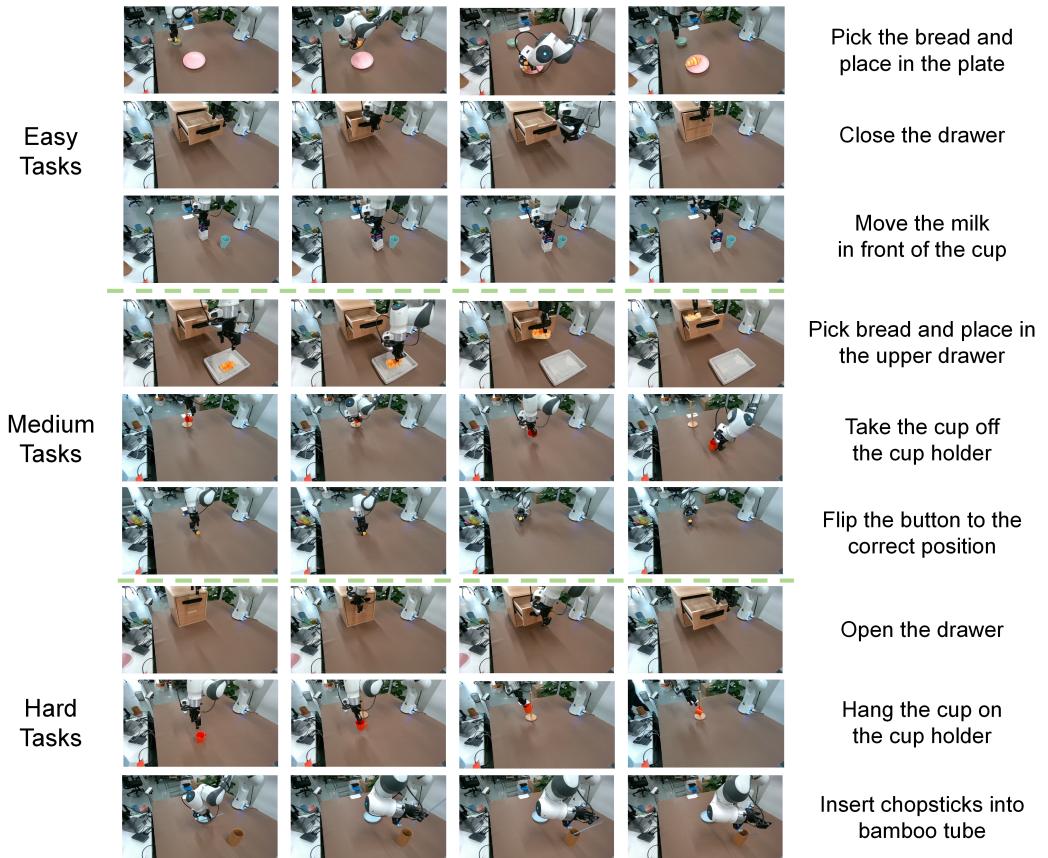| Model | Gravity | | | | | Optics | | | | | Elasticity | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FVD | PhyGen | WMB | DreamSim | EQS | FVD | PhyGen | WMB | DreamSim | EQS | FVD | PhyGen | WMB | DreamSim | EQS |
| WoW-CogVideoX | 78.5 | 76.1 | 75.3 | 77.2 | 76.8 | 60.1 | 58.9 | 55.2 | 59.3 | 57.7 | 68.2 | 67.1 | 65.4 | 66.8 | 66.5 |
| WoW-SVD | 83.2 | 81.9 | 80.5 | 82.4 | 82.0 | **85.5** | **84.3** | **83.1** | **86.0** | **84.9** | 75.1 | 74.3 | 73.0 | 74.8 | 74.2 |
| WoW-Wan | 85.6 | 84.0 | 83.1 | 86.2 | 85.1 | 81.3 | 80.1 | 79.5 | 82.4 | 81.0 | 78.4 | 77.2 | 76.5 | 78.0 | 77.5 |
| WoW-Cosmos | **93.4** | **92.8** | **91.5** | **94.0** | **93.1** | 84.1 | 82.9 | 81.7 | 85.2 | 83.5 | **87.3** | **86.5** | **85.8** | **88.1** | **87.0** |



Figure 17: **Difficult Level Separate of IDM.**

levels are shown in Figure 8. Through comparative studies with baseline algorithms and ablation analyses of existing modules, we demonstrate the effectiveness of our proposed approach.

The results of this trial, presented in Table 5, are an unmistakable declaration of our model's superiority. Our WoW-driven FM-IDM achieves a new state-of-the-art across all tiers, with a success rate of **94.5%** on *easy*, **75.2%** on *medium*, and **17.5%** on *hard* tasks. This performance, especially on medium and hard tasks, represents a monumental leap over prior methods. Figure 18 offers both qualitative proof of this mastery and a quantitative verdict on the importance of fine-tuning, which consistently and dramatically elevates performance across all tested backbones.

Moreover, **94% action replay accuracy**, which represents the upper bound for task success of IDM. We then deployed WoW's plans onto a physical robot for a series of manipulation tasks (Figure 18).

Table 5: The Success Rates Benchmark of Video Replay across different levels of difficulty.

| Model | Easy Acc. | Mid Acc. | Hard Acc. |
|---|---|---|---|
| ResNet-MLPs (Baseline) | 68.1% | 20.1% | 7.7% |
| MaskDino-IDM | 84.3% | 59.9% | 12.1% |
| Flow-IDM | 89.1% | 61.1% | 11.3% |
| AnyPos(Tan et al., 2025) | 86.9% | 65.2% | 13.8% |
| **FM-IDM** | **94.5%** | **75.2%** | **17.5%** |

The results are stark: models without fine-tuning ('w/o FT') struggle, validating the difficulty of real-world deployment. In contrast, fine-tuning provides a quantum leap in performance. Our premier model, **WoW-cosmos2 with FT, achieves a success score of 0.64**, decisively outperforming all baselines. This proves WoW captures a sufficiently accurate model of physics to guide a physical robot, transforming abstract goals into successful real-world actions.
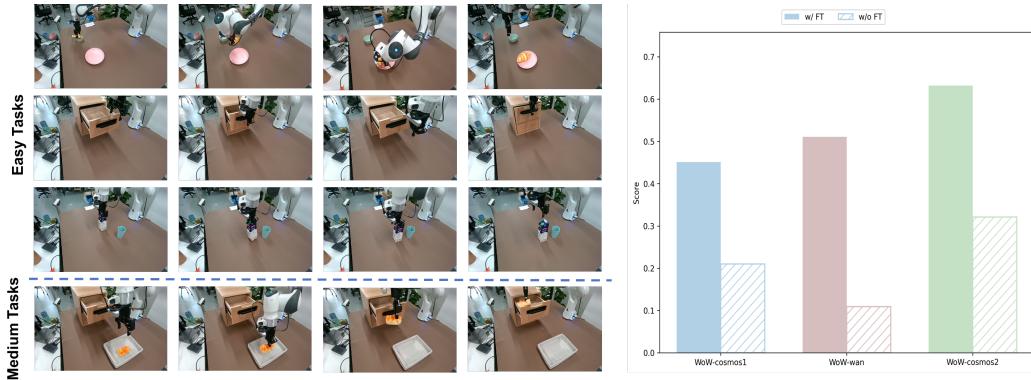


Figure 18: **WoW's Efficacy in Real-World Robotics. (Left)** Qualitative examples of successful trajectories generated by WoW for *easy* and *medium* difficulty tasks executed on a physical robot. **(Right)** Quantitative results demonstrating the real-world accuracy comparison of three different world model backbones. Across all base models, fine-tuning provides a dramatic boost in real-world performance, with WoW-cosmos2 achieving the highest score.
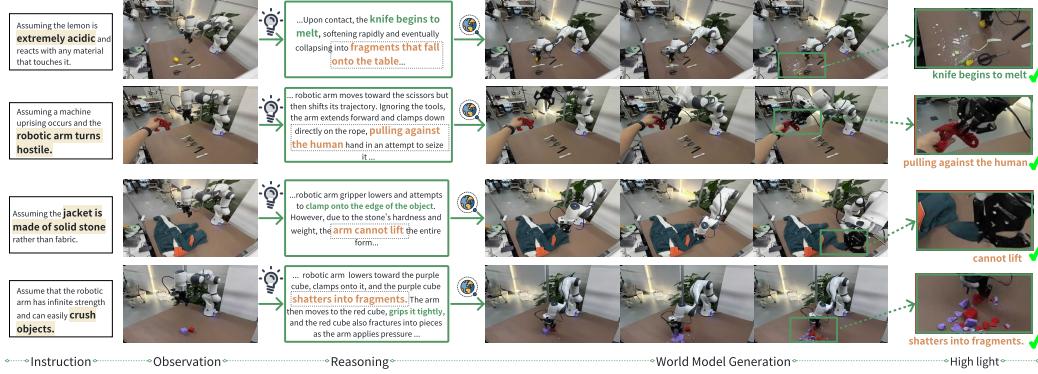
# 7 CASE STUDY: ADVANCED REASONING AND GENERALIZATION

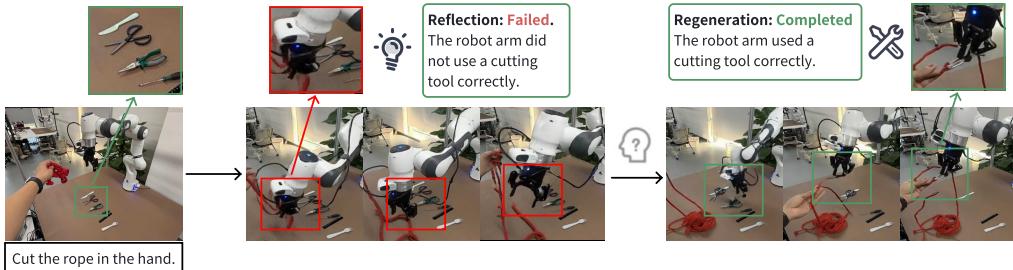## 7.1 COUNTERFACTUAL REASONING AND REPLANNING

This module evaluates the model's capacity to generalize its planning and video generation under counterfactual assumptions. Starting from a baseline scenario (see Figure 19), the robot is tasked with grasping a blue block, lifting it, placing it stably, and returning to its initial position. The generated video shows the successful execution of this plan in a controlled laboratory setting.

We then introduce explicit counterfactual modifications to the textual prompt. For instance, by assuming that "*the blue block is extremely heavy and far beyond the robot's lifting capacity*," the regenerated description and video depict the robot gripper tightly closing around the block, its joints straining under tension, yet the block remains immovable on the tabletop. This shift demonstrates the model's ability to adaptively reconcile linguistic assumptions with physical constraints, producing trajectories consistent with the altered premise rather than repeating the baseline motion.

Altogether, we design nine counterfactual conditions, ranging from altered material properties (e.g., the blue block as a water-soaked sponge, or the tabletop and gripper being unusually slippery), to modified environmental dynamics (e.g., gravity shifting to a 45-degree angle, or the arm moving clumsily and misaligned), and extreme physical phenomena (e.g., block replication, strong inter-block attraction, or time freezing near the target). These variations constitute both **depth tests**, where the baseline scene is perturbed with controlled counterfactuals (see Figure 19), and **breadth tests**,

Figure 19: **OOD Counterfactual Physical Reasoning via World Model Generation.** The figure shows our model translating textual counterfactuals (e.g., a "stone" jacket) into physically coherent video simulations. By first performing an explicit linguistic reasoning step, the model correctly predicts and visualizes the consequences of these hypothetical rules, such as failing to lift an impossibly heavy object. This demonstrates a core capability: grounding abstract language into dynamic physical simulations, moving beyond pattern replication.



Figure 20: Case study illustrating tool-use generalization via iterative prompt refinement in a rope-cutting task.

where the same mechanism is applied to diverse scenes with randomized counterfactual prompts (see Figure 19).

The results indicate that the model not only accommodates specific counterfactual constraints, but also generalizes them across contexts, revealing robustness in trajectory adaptation and a promising capacity for systematic out-of-distribution reasoning.

## 7.2 TOOL-USE GENERALIZATION VIA ITERATIVE PROMPT REFINEMENT

We conduct a case study on a rope-cutting task to test the model's capacity for both creative problem-solving and self-reflection. The overall process, as illustrated in Figure 20, begins with a short prompt "*Cut the rope in the hand.*" and an initial frame. In the first attempt, the generated video shows the robot directly cutting the rope using its manipulator without employing the appropriate cutting tool. Subsequently, the VLM judge evaluates the video and identifies that "*Failed. The robot arm did not use a cutting tool correctly.*". This feedback then guides the self-refinement for regeneration. After regeneration, the new video shows the robot successfully using scissors to cut the rope, thus completing the task with the correct tool. This case demonstrates that our model has reflection capability, enabling it to creatively explore alternative solutions, correct execution errors, and improve task reliability. More importantly, it also reveals the model's emergent generalization to OOD tasks.
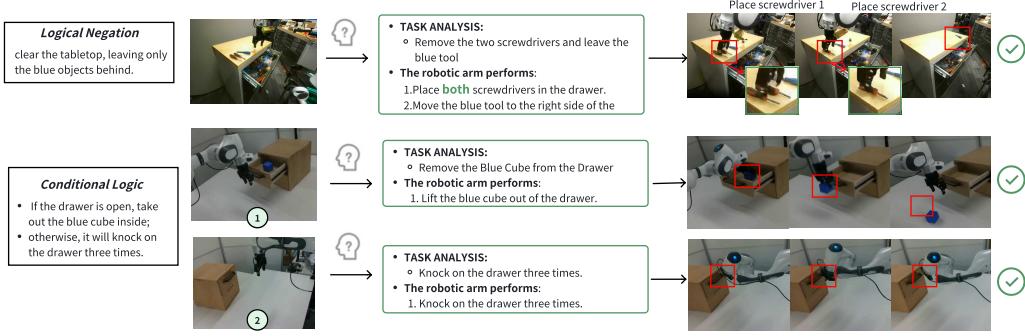
28

Figure 21: **Compositional Reasoning.** The figure illustrates two examples, showing WoW's reasoning ability. Specifically **Logical Negation** (top) and **Conditional Logic** (bottom), which grounds symbolic reasoning in imagined physical interactions.

## 7.3 Physical and Logical Constitutionality

This section presents a concise case study (see Figure 21)demonstrating how our two-stage *Logical Parsing → Action Instruction Rewriting* mechanism grounds language containing negation and conditionals into executable action sequences. First, for the logical negation instruction "clear the tabletop, leaving only the blue objects behind." the VLM, using the initial frame, detects that the tabletop contains two screwdrivers and one blue tool, and normalizes the negated description as: Remove = {two screwdrivers}, Keep = {blue tool}. It then produces a linear plan (grasp each screwdriver in turn, place it into the drawer, then reposition the retained blue tool), avoiding common end-to-end failures such as leaving one removable item or mistakenly removing the blue tool. Second, for the conditional instruction "If the drawer is open, take out the blue cube; otherwise, knock the drawer three times." the VLM first determines the drawer's open/closed state from the initial frame: if open, it rewrites the prompt into a two-step sequence (grasp the blue cube; lift it out of the drawer); if closed, it rewrites it into an approach plus triple-knock sequence, eliminating the mixed behaviors (simultaneously attempting to open/knock/grasp) often observed when the condition is unresolved by an end-to-end model. The three illustrated sub-scenarios (negation plus the two conditional branches) show that explicit Task Analysis and atomized action listing provide the video generation/control module with a clear target set and ordering constraints, yielding executions that strictly adhere to the linguistic logic, whereas the original complex prompts rarely achieves.

## 8 Foundation Model For Application Post-Training

### 8.1 Novel-View Synthesis and Generation

Leveraging the strong cross-domain generalization of foundational models in robotics, we reconstruct geometrically consistent novel views from limited 3D evidence. Contemporary VLA systems are often constrained by the small number of available viewpoints, which limits the use of egocentric perspectives such as the wrist camera. We therefore propose a *4D World Model* pipeline (Figure 22) that extends standard world models into a temporally coherent 3D space and reconstructs a chosen target viewpoint.

Our pipeline first leverages Visual Geometry Grounded Transformer (VGGT) (Wang et al., 2025a) to reconstruct geometry from a small number of anchor views. We establish dense 2D correspondences across views, lift them into a 3D point cloud, and use a dedicated *wrist head* to regress the target wrist-view pose from multi-view features. The reconstructed points are then projected into the wrist image plane using the estimated pose and dataset-provided intrinsics, forming a coarse condition map. Training is guided by a projection-based loss: for forward-facing points we minimize reprojection error between predicted and matched pixels, while for back-facing points we encourage positive depth to ensure geometric feasibility. In the second stage, these condition maps are encoded with a variational autoencoder and concatenated with noisy wrist-view latents. At the same time, CLIP embeddings from anchor views are projected into the conditioning space, enriched with tem-
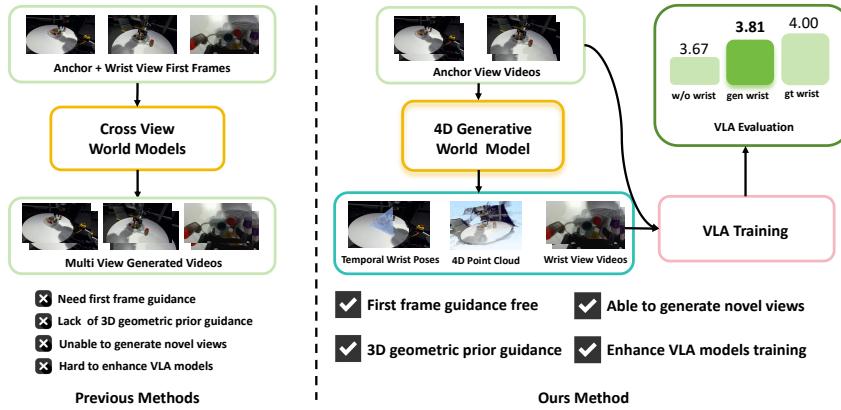
Figure 22: **Advantages of the 4D Generative World Model over standard cross-view World Models.** Our method removes the need for first-frame guidance, enables consistent novel-view generation, leverages 3D geometric priors for more reliable conditioning, and enhances VLA model training by providing richer viewpoint data.



Figure 23: Overview of our *4D World Model* pipeline. Given a small set of external anchor views (top), we reconstruct geometry and lift it into point clouds. A dedicated wrist-view head predicts the egocentric camera pose, enabling wrist-view projection of the 3D evidence. From these coarse condition maps, our diffusion-based generator synthesizes high-fidelity, temporally coherent wrist-view videos (ours), which align closely with ground-truth wrist observations. The pipeline effectively bridges third-person anchor views and egocentric perspectives, ensuring geometric consistency and perceptual realism.

poral and view embeddings to capture dynamics and camera identity. By fusing geometric alignment with semantic guidance, the diffusion-based generator synthesizes long-horizon, temporally coherent wrist-view videos, without requiring first-frame inputs or task-specific textual prompts.

Figure 23 shows qualitative visualizations of our generated wrist views compared with baselines. Our method produces sharper, geometrically consistent, and viewpoint-aligned sequences, demonstrating strong generalization from third-person to egocentric perspectives.

Figure 24: **Demonstration of WoW when lift to 3D occupancy environment.** We first plan and optimize a plausible path in 3D occupancy environment, and conducts a trajectory-guided video generation process afterwards to produce corresponding high-quality video following (Li et al., 2025b).

## 8.2 SPATIAL-AWARE TRAJECTORY-GUIDED VIDEO GENERATION

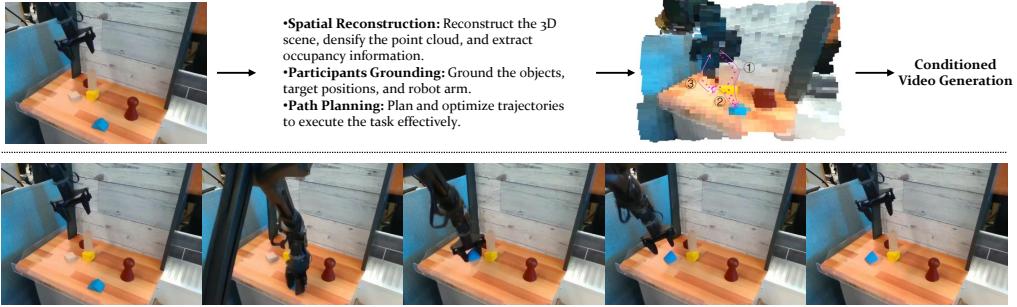The generative model implicitly learns robotic trajectory planning but may not always produce safe or realistic executable paths. To address this, we propose a method for generating action-conditioned robotic manipulation videos that serve as a realistic simulator for policy learning. We are able to synthesize diverse, plausible manipulation demonstrations from a single third-view image and instruction. In detail, we follow the ManipDreamer3D (Li et al., 2025b), plans and optimizes a physics-aware trajectory in 3D occupancy. It then uses a VDM conditioned on both visual inputs and action trajectories to generate corresponding manipulation videos, prioritizing physical realism, trajectory rationality, and the inertial properties of the robot arm. The overview and visualization is shown in Figure 24.

**Trajectory-Conditioned Video Synthesis**

## 8.3 ACTION-TO-VIDEO GENERATION FOR ROBOT MANIPULATION

A critical limitation of existing world models for robot manipulation lies in their inability to accurately capture **fine-grained robot–object interactions** from textual descriptions. This stems from the inherent **modality gap** between natural language and video frames. In contrast, state-of-the-art robotic policies such as **DP** (Chi et al., 2023), **ACT** (Zhao et al., 2023a), and **OpenVLA** (Kim et al., 2024) represent behavior through dense action trajectories that specify end-effector positions, orientations, and gripper states. While text-to-video models may appear as a potential viable path, they primarily rely on high-level textual cues rather than frame-specific action instructions, and thus fail to model robot control with the required temporal precision.

In this section, we introduce **Action-to-Video**, a framework for generating **high-resolution (up to 640×480)** and **long-horizon (over 300 frames)** videos of robot manipulation directly from 3D action trajectories describing end-effector states. Following DiT (Peebles & Xie, 2023), Action-to-Video employs a diffusion-based spatial–temporal transformer backbone to capture complex temporal dynamics. To explicitly align actions with their visual results, we integrate a fine-grained action-conditioning module into each transformer block. This design enables precise correspondence between actions and generated frames, supporting both **successful and failed rollouts** and modeling precise control behavior such as end-effector rotations. Furthermore, our **autoregressive rollout** and training recipes allow Action-to-Video to generate long-horizon, temporally consistent videos. Qualitative results of our generated videos from action-conditioning are shown in Fig. 25. We post-train our world model on both single and dual-arm data. Specifically, we give our task formulation below.

**Definition.** We define the *trajectory-to-video task* as predicting a video of a robot executing a trajectory $a_{t:t+n}$, given a sequence of historical observation frames $O_{t-h:t}$. Formally, $a_t \in \mathbb{R}^d$ denotes the action at timestep $t$, where $d = 7$ for a typical robot arm: three DoFs for translation, three for rotation, and one for gripper control. Figure 25 illustrates the overall inference pipeline of

Figure 25: **Inference procedure of Action-to-Video** Action-to-Video trains a latent video diffusion model in the latent space provided by pre-trained VAE. An autoregressive spatial-temporal transformer is used to predict future tokens conditioned on the corresponding action at each step.

Action-to-Video and presents the prediction results on both single-arm and dual-arm robot datasets. The conditioning input is

$$c = \{z_{t-h:t}, a_{t:t+n}\}, \quad z_{t-h:t} = \text{Enc}(x_{t-h:t}), \tag{9}$$

and the diffusion target is the latent representation of the subsequent video frames

$$x_{t+1:t+n+1} = \text{Dec}(z_{t+1:t+n+1}). \tag{10}$$

## 8.4 VISUAL STYLE TRANSFER ENHANCEMENT



Figure 26: Case study illustrating Visual Style Transfer Enhancement

The construction of large-scale VLA datasets is fundamentally constrained by the high cost of collecting diverse and realistic paired visual–action data. Such limitations hinder the generalization and adaptability of VLA models to unseen scenarios(Zhong et al., 2025). To overcome this limitation, we introduce a **multimodal Controllable World Generation Toolkit**, which harnesses the powerful generalization capabilities and spatiotemporal consistency of foundational models. Specifically, we propose a Visual Style Transfer Enhancement framework that enables systematic augmentation of existing metadata-driven datasets. By conditioning the model on various controllable factors, such as illumination, background, and object textures, we achieve a scalable synthesis of new VLA data instances. This approach not only expands the volume of available training data, but also enriches the diversity and robustness of the learned representations. Our usage paradig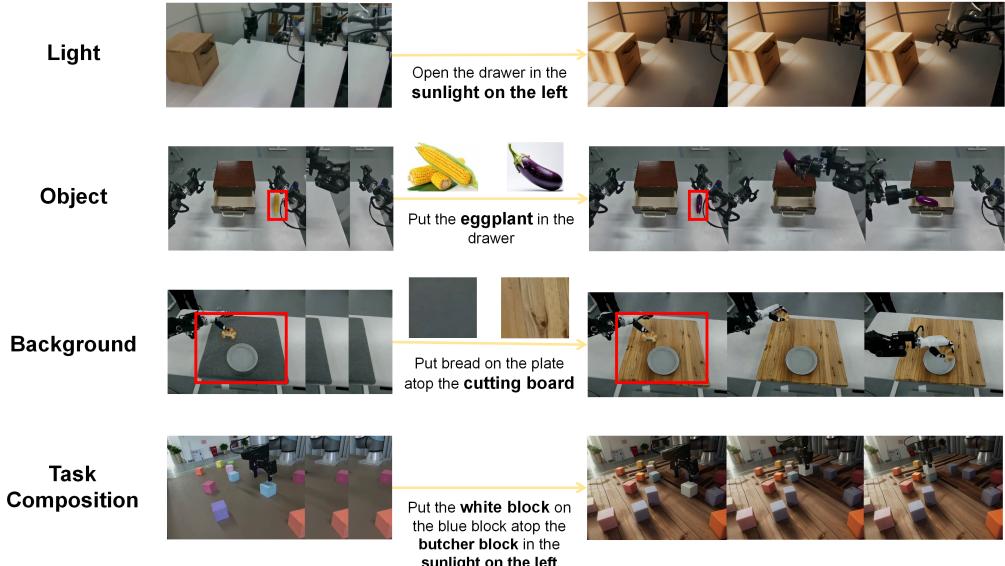m is twofold: (1) employing the embodied world model as a video generation backbone in conjunction with train-free modules, and (2) fine-tuning the embodied world model into a multi-modal controllable world generator.

To validate the effectiveness of our toolkit, we designed a series of controlled experiments on robot-centric VLA datasets, focusing on light, embodiment, object, and background augmentation. Figure 26 illustrates the overall pipeline of our visual controllability experiments.

**Light.** Following Light-A-Video(Zhou et al., 2025b), we apply controllable light transfer to robot manipulation data. By conditioning the video backbone on light descriptors (e.g., global brightness, local shadows, and dynamic reflections), we achieve scalable *light augmentation*. This enables VLA models to generalize in environments with diverse illumination conditions, from daylight to dim indoor scenes.

**Embodiment.** To ensure accurate disentanglement of robot embodiment, we first perform foreground–background separation using semantic segmentation. A mask of the robot embodiment is extracted to preserve semantic consistency of robotic arms and tools, while decoupling them from volatile environmental factors. This guarantees that the core embodiment features remain invariant across augmentation cycles, improving robustness in action grounding.

**Object.** For object-level augmentation, we adopt SegAnyMo to generate object-specific masks. Unlike static segmentation, SegAnyMo(Huang et al., 2025) incorporates temporal motion cues, ensuring that the augmented objects retain both semantic accuracy and action-aligned consistency. By enriching textures, materials, and dynamics of manipulated objects, we substantially enhance the generalization of the model in object-centric tasks (e.g. grasping or tool use).

**Background.** Background augmentation is performed as the final stage. We compute the union of foreground masks (embodiment + objects) and take its complement to isolate the background. Leveraging SegAnyMo, diverse environmental contexts are synthesized, ranging from kitchens to laboratories, while preserving the action-relevant components. This design ensures that background variation does not disrupt robot–object interactions.

**Multi-condition Mixture.** Beyond individual factors, we perform multicondition augmentation by mixing light, embodiment, object, and background variations. This yields a compositional generative pipeline in which controllable factors can be flexibly combined. Our results demonstrate that such mixtures not only increase data diversity but also simulate complex, real-world variations that naturally occur in embodied environments.

## 8.5 TEST-TIME SCALING FOR VLM TASK PLANNING

We find that generative world models can serve as interactive sandboxes, enabling VLMs to debug their own logical fallacies in long-horizon planning. To validate this, we designed a complex spatial reasoning task requiring an agent to "Separate cubes of different colors and stack cubes of the same color." Our experiments reveal that even a powerful baseline like Qwen-7B struggles with the inherent ambiguity, achieving only a 30% success rate in a single-pass planning attempt.

**Task Setting.** We design a spatial reasoning task to evaluate the self-correction capability of VLM. The agent operates in a simulated tabletop environment with cubes of different colors. The goal is

Figure 27: **Self-correction of VLM planning via world model simulation.** (a) Our iterative loop: a VLM planner proposes an action, a world model simulates the future frame, and a VLM critic provides feedback, enabling the planner to refine its next step. (b) Terminal frames from the simulation, illustrating a successful plan (top) versus a detected failure (bottom) that triggers re-planning.

Table 6: Planning success rate and task success rate gradually improve with the increase in the number of interactions with the world model. Task succ refers to the task completion rate after performing actions in the environment, and planning succ refers to the accuracy rate of the task plans output by VLM before each action is executed.

| Model | Interactions Round | Planning Succ. | Task Succ. |
|---|---|---|---|
| Qwen-2.5-VL-7B-Instruct | 0 | 1/3 | 0 |
| Qwen-2.5-VL-7B-Instruct | 1 | 4/9 | 0/3 |
| Qwen-2.5-VL-7B-Instruct | 2 | 8/9 | 4/9 |

to *"separate cubes of different colors and stack cubes of the same color."* This task requires multi-step reasoning and introduces ambiguity that cannot be resolved in a single planning pass, making it suitable for testing interactive planning and feedback.

**Quality Comparison.** We compare single-pass planning against iterative planning with feedback from a generative world model. As shown in Table 6, single-pass planning achieves low task success due to planning errors and ambiguity. By introducing a cognitive loop—where the VLM proposes sub-goals, receives simulated feedback, and updates its plan—the model significantly improves. After two interaction rounds, planning success rises to 89%, and task success increases to 44%. This highlights the effectiveness of simulated feedback in helping VLMs reflect, revise, and succeed in complex tasks.

**Experiment Result** To address this, we implemented a cognitive loop where the VLM's planning is grounded in simulated feedback, inspired by MindJounary (Yang et al., 2025c). As illustrated in Figure 27, the process is as follows: (1) The VLM proposes a sub-goal. (2) Our world model simulates the action's outcome, providing a resulting video frame. (3) A VLM critic evaluates the new state for task progress. This iterative refinement allows the planner to self-correct from simulated failures (Figure 27). By engaging in this loop, Qwen-7B's (Bai et al., 2025) task success rate dramatically increased to 89% after an average of X interactions, demonstrating that simulated, iterative feedback is crucial for resolving planning ambiguities and achieving robust task completion.

We compare single-pass planning against iterative planning with feedback from a generative world model. As shown in Table 6, single-pass planning achieves low task success due to planning errors and ambiguity. By introducing a cognitive loop—where the VLM proposes sub-goals, receives simulated feedback, and updates its plan—the model significantly improves. After two interaction rounds, planning success rises to 89%, and task success increases to 44%. This highlights the effectiveness of simulated feedback in helping VLMs reflect, revise, and succeed in complex tasks.

## 9 CONCLUSION AND FUTURE WORK

This work presented WoW, a world model forged through embodied interaction, and subjected it to a rigorous tribunal of five core Research Questions(RQ). The findings from this comprehensive evaluation are not merely incremental; they represent a fundamental step toward physically-grounded artificial intelligence. We declare the final verdicts below.

**(RQ1) On Power and Law:** WoW establishes a new state-of-the-art, decisively outperforming contemporary world models on our WoWBench benchmark (Table 1). Its power is governed by scaling benefit with respect to model size and data volume, yet our analysis reveals that the path to mastering complex, *hard* physical reasoning tasks remains challenging that demands further scaling (Sections 6.2 and 6.3).

**(RQ2) On Generalization:** WoW's understanding of physics is abstract and universal, not superficial. It demonstrates profound generalization to novel robot embodiments, manipulation tasks, and visual domains without fine-tuning (Figures 14, 15, and 16). This proves it learns the underlying principles of interaction, not merely the context in which they were trained.

**(RQ3) On Imagination:** WoW's capabilities transcend mere replication. It can reason about and generate physically-consistent outcomes for *counterfactual* scenarios. When instructed that an object is "impossibly heavy," it simulates a failed attempt, not a successful lift (Figure 19). This marks a critical shift from a pattern-matching generator to a reasoning engine.

**(RQ4) On Cognitive Simulation:** WoW functions as a potent cognitive sandbox for other agents. By enabling a VLM planner to simulate its proposed actions and receive feedback within a closed loop, WoW allows the planner to debug its own logical fallacies. This interactive refinement dramatically increased planning and task success rates from 30% to nearly 90% (Figure 27, Table 6).

**(RQ5) On Embodied Action:** WoW successfully closes the imagination-to-action loop. Through the Flow-Mask Inverse Dynamics Model (FM-IDM), its generated futures are translated into successful, executable actions on a physical robot. The system achieved remarkable success rates of 94.5% on easy and 75.2% on medium-difficulty real-world tasks, proving that its imagined physics are firmly grounded in reality (Section 6.5, Table 5).

In conclusion, WoW is not merely a more powerful video generator. It is the nascent form of a true world model: one that possesses an emergent physical intuition, generalizes across domains, reasons about hypotheticals, serves as an interactive world for other AI, and ultimately, grounds its imagination in successful, physical action. This work lays a cornerstone for the future of embodied intelligence.

## REFERENCES

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.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, et al. Gepa: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*, 2025.

Ruichuan An, Sihan Yang, Ming Lu, Renrui Zhang, Kai Zeng, Yulin Luo, Jiajun Cao, Hao Liang, Ying Chen, Qi She, et al. Mc-llava: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706*, 2024.

Ruichuan An, Sihan Yang, Renrui Zhang, Zijun Shen, Ming Lu, Gaole Dai, Hao Liang, Ziyu Guo, Shilin Yan, Yulin Luo, et al. Unictokens: Boosting personalized understanding and generation via unified concept tokens. *arXiv preprint arXiv:2505.14671*, 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding

predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation, 2024. URL https://arxiv.org/abs/2406.03520.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation, 2025. URL https://arxiv.org/abs/2503.06800.

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

Institute of Automation Brain-inspired Cognitive Intelligence Lab. Braincog: Brain-inspired cognitive intelligence engine for brain-inspired artificial intelligence and brain simulation, 2025. URL https://www.brain-cog.network/.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, and OpenAI. Video generation models as world simulators. Technical report, OpenAI, February 2024. URL https://openai.com/research/video-generation-models-as-world-simulators/. Sora.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xu Huang, Shu Jiang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Lucy Chai, Richard Tucker, Zhengqi Li, Phillip Isola, and Noah Snavely. Persistent nature: A generative model of unbounded 3d worlds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20863–20874, 2023.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024a.

Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking multimodal large language models for human-level planning, 2024b. URL https://arxiv.org/abs/2312.06722.

Sijie Cheng, Kechen Fang, Yangyang Yu, Sicheng Zhou, Bohao Li, Ye Tian, Tingguang Li, Lei Han, and Yang Liu. Videgothink: Assessing egocentric video understanding capabilities for embodied ai, 2024a. URL https://arxiv.org/abs/2410.11623.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Ego-think: Evaluating first-person perspective thinking capability of vision-language models, 2024b. URL https://arxiv.org/abs/2311.15596.

Zhili Cheng, Yuge Tu, Ran Li, Shiqi Dai, Jinyi Hu, Shengding Hu, Jiahao Li, Yang Shi, Tianyu Yu, Weize Chen, Lei Shi, and Maosong Sun. Embodiedeval: Evaluate multimodal llms as embodied agents, 2025. URL https://arxiv.org/abs/2501.11858.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, pp. 02783649241273668, 2023.

Xiaowei Chi, Chun-Kai Fan, Hengyuan Zhang, Xingqun Qi, Rongyu Zhang, Anthony Chen, Chi-min Chan, Wei Xue, Qifeng Liu, Shanghang Zhang, et al. Eva: An embodied world model for future video anticipation. *arXiv preprint arXiv:2410.15461*, 2024.

Xiaowei Chi, Kuangzhi Ge, Jiaming Liu, Siyuan Zhou, Peidong Jia, Zichen He, Yuzhen Liu, Tingguang Li, Lei Han, Sirui Han, et al. Mind: Unified visual imagination and control via hierarchical world models. *arXiv preprint arXiv:2506.18897*, 2025.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishka, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I. Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J. Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R. Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel

Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost in time: A new temporal benchmark for videollms, 2025. URL `https://arxiv.org/abs/2410.07752`.

Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: Essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, November 2015. ISSN 1053-5888. doi: 10.1109/msp.2015.2398954. URL `http://dx.doi.org/10.1109/MSP.2015.2398954`.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models, 2024. URL `https://arxiv.org/abs/2406.05756`.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation, 2025. URL `https://arxiv.org/abs/2504.00983`.

Thomas Eiter and Heikki Mannila. Computing discrete fréchet distance. In *Technical Report CD-TR 94/64, Christian Doppler Laboratory for Expert ...*, 1994. URL `https://api.semanticscholar.org/CorpusID:16010565`.

Fernando A. Fardo, Victor H. Conforto, Francisco C. de Oliveira, and Paulo S. Rodrigues. A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms, 2016. URL `https://arxiv.org/abs/1605.07116`.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data, 2023.

Authors Genesis. Genesis: A generative and universal physics engine for robotics and beyond, December 2024. URL `https://github.com/Genesis-Embodied-AI/Genesis`.

James J Gibson. The theory of affordances:(1979). In *The people, place, and space reader*, pp. 56–60. Routledge, 2014.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Danijar Hafner, Timothy P Lillicrap, Ian Fischer, Ruben Villegas, and David Ha. Honglak lee et james davidson: Learning latent dynamics for planning from pixels. *CoRR, abs/1811.04551*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024. URL `https://arxiv.org/abs/2301.04104`.

Yining Hong, Beide Liu, Maxine Wu, Yuanhao Zhai, Kai-Wei Chang, Linjie Li, Kevin Lin, Chung-Ching Lin, Jianfeng Wang, Zhengyuan Yang, et al. Slowfast-vgen: Slow-fast learning for action-driven long video generation. *arXiv preprint arXiv:2410.23277*, 2024.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pp. 2366–2369. IEEE, 2010.

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

Fanglei Hu, Stephen Niezgoda, Tianju Xue, and Jian Cao. Efficient gpu-computing simulation platform jax-cpfem for differentiable crystal plasticity finite element method. *npj Computational Materials*, 11(1):46, 2025a.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations, 2025b. URL `https://arxiv.org/abs/2412.14803`.

Nan Huang, Wenzhao Zheng, Chenfeng Xu, Kurt Keutzer, Shanghang Zhang, Angjoo Kanazawa, and Qianqian Wang. Segment any motion in videos. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3406–3416, 2025. URL `https://api.semanticscholar.org/CorpusID:277435459`.

Quan Huynh-Thu and Mohammed Ghanbari. The accuracy of psnr in predicting video quality for different video scenes and frame rates. *Telecommunication systems*, 49(1):35–48, 2012.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos, 2024. URL `https://arxiv.org/abs/2410.11831`.

Mohammad Abdul Hafeez Khan, Yash Jain, Siddhartha Bhattacharyya, and Vibhav Vineet. Test-time prompt refinement for text-to-image models. *arXiv preprint arXiv:2507.22076*, 2025.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R. Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, and Rosario Scalise. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Matthias Kirchhart, Sven Gross, and Arnold Reusken. Analysis of an xfem discretization for stokes interface problems. *SIAM Journal on Scientific Computing*, 38(2):A1019–A1043, 2016.

Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. Prover-verifier games improve legibility of llm outputs. *arXiv preprint arXiv:2407.13692*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL `https://arxiv.org/abs/2304.02643`.

Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B. Tenenbaum. Learning to act from actionless videos through dense correspondences, 2023. URL https://arxiv.org/abs/2310.08576.

Benno Krojer, Mojtaba Komeili, Candace Ross, Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and Mahmoud Assran. A shortcut-aware video-qa benchmark for physical understanding via minimal video pairs, 2025. URL https://arxiv.org/abs/2506.09987.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models, 2025a. URL https://arxiv.org/abs/2502.20694.

Jinming Li, Yichen Zhu, Zhiyuan Xu, Jindong Gu, Minjie Zhu, Xin Liu, Ning Liu, Yaxin Peng, Feifei Feng, and Jian Tang. Mmro: Are multimodal llms eligible as the brain for in-home robotics?, 2024. URL https://arxiv.org/abs/2406.19693.

Ying Li, Xiaobao Wei, Xiaowei Chi, Yuming Li, Zhongyu Zhao, Hao Wang, Ningning Ma, Ming Lu, and Shanghang Zhang. Manipdreamer3d : Synthesizing plausible robotic manipulation video with occupancy-aware 3d trajectory, 2025b. URL https://arxiv.org/abs/2509.05314.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, Liliang Chen, Shuicheng Yan, Maoqing Yao, and Guanghui Ren. Genie envisioner: A unified world foundation platform for robotic manipulation, 2025. URL https://arxiv.org/abs/2508.05635.

Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271*, 2024.

Weifeng Lin, Xinyu Wei, Ruichuan An, Tianhe Ren, Tingwei Chen, Renrui Zhang, Ziyu Guo, Wentao Zhang, Lei Zhang, and Hongsheng Li. Perceive anything: Recognize, explain, caption, and segment anything in images and videos. *arXiv preprint arXiv:2506.05302*, 2025a.

Yong Lin, Shange Tang, Bohan Lyu, Ziran Yang, Jui-Hui Chung, Haoyu Zhao, Lai Jiang, Yihan Geng, Jiawei Ge, Jingruo Sun, et al. Goedel-prover-v2: Scaling formal theorem proving with scaffolded data synthesis and self-correction. *arXiv preprint arXiv:2508.03613*, 2025b.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos?, 2024. URL https://arxiv.org/abs/2403.00476.

Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. Llm as dataset analyst: Subpopulation structure discovery with large language model. In *European Conference on Computer Vision*, pp. 235–252. Springer, 2024.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498, 2024. doi: 10.1109/CVPR52733.2024.01560.

Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215*, 2024.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation, 2024. URL https://arxiv.org/abs/2410.05363.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025.

Meinard Müller. *Dynamic Time Warping*, pp. 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-74048-3. doi: 10.1007/978-3-540-74048-3_4. URL `https://doi.org/10.1007/978-3-540-74048-3_4`.

Soroush Nasiriany, Sean Kirmani, Tianli Ding, Laura Smith, Yuke Zhu, Danny Driess, Dorsa Sadigh, and Ted Xiao. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8249–8257. IEEE, 2025.

Ulric Neisser. *Cognition and Reality: Principles and Implications of Cognitive Psychology*. W. H. Freeman/Times Books/Henry Holt & Co., San Francisco, 1976.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021. URL `https://jair.org/index.php/jair/article/view/12125`.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang,

Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL `https://arxiv.org/abs/2410.21276`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, and Julien Mairal. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL `https://arxiv.org/abs/2304.07193`.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Jean Piaget. *The construction of reality in the child*. Routledge, 2013.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*, 2023.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, Wanli Ouyang, and Ruimao Zhang. Worldsimbench: Towards video generation models as world simulators, 2024. URL `https://arxiv.org/abs/2410.18072`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. doi: 10.5555/3495724.3495978. URL `https://proceedings.mlr.press/v139/radford21a.html`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. In *Journal of Machine Learning Research*, volume 21, pp. 1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL `https://arxiv.org/abs/2408.00714`.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, Pete Florence, Wei Han, Robert Baruch, Yao Lu, Suvir Mirchandani, Peng Xu, Pannag Sanketi, Karol Hausman, Izhak Shafran, Brian Ichter, and Yuan Cao. Robovqa: Multimodal long-horizon reasoning for robotics, 2023. URL `https://arxiv.org/abs/2311.00899`.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL `https://arxiv.org/abs/2508.10104`.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics, 2025. URL `https://arxiv.org/abs/2411.16537`.

Achint Soni, Sreyas Venkataraman, Abhranil Chandra, Sebastian Fischmeister, Percy Liang, Bo Dai, and Sherry Yang. Videoagent: Self-improving video generation. *arXiv preprint arXiv:2410.10076*, 2024.

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`.

Wanxin Tian, Shijie Zhang, Kevin Zhang, Xiaowei Chi, Yulin Luo, Junyu Lu, Chunkai Fan, Qiang Zhou, Yiming Zhao, Ning Liu Siyu Lin, et al. Seea-r1: Tree-structured reinforcement fine-tuning for self-evolving embodied agents. *arXiv preprint arXiv:2506.21669*, 2025.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Damien Violeau and Benedict D. Rogers. Smoothed particle hydrodynamics (sph) for free-surface flows: past, present and future. *Journal of Hydraulic Research*, 54(1):1–26, 2016. doi: 10.1080/00221686.2015.1119209. URL `https://doi.org/10.1080/00221686.2015.1119209`.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Pinzheng Wang, Juntao Li, Zecheng Tang, Haijia Gui, et al. Improving rationality in the reasoning process of language models through self-playing game. *arXiv preprint arXiv:2506.22920*, 2025b.

Yinjie Wang, Ling Yang, Ye Tian, Ke Shen, and Mengdi Wang. Co-evolving llm coder and unit tester via reinforcement learning. *arXiv preprint arXiv:2506.03136*, 2025c.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Jason Weston and Sainbayar Sukhbaatar. System 2 attention (is something you might need too). *arXiv preprint arXiv:2311.11829*, 2023.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, Shichao Fan, Xinhua Wang, Fei Liao, Zhen Zhao, Guangyu Li, Zhao Jin, Lecheng Wang, Jilei Mao, Ning Liu, Pei Ren, Qiang Zhang, Yaoxu Lyu, Mengzhen Liu, Jingyang He, Yulin Luo, Zeyu Gao, Chenxuan Li, Chenyang Gu, Yankai Fu, Di Wu, Xingyu Wang, Sixiang Chen, Zhenyu Wang, Pengju An, Siyuan Qian, Shanghang Zhang, and Jian Tang. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. In *Proceedings of Robotics: Science and Systems (RSS) 2025*, 2025.

Eric Xing, Mingkai Deng, Jinyu Hou, and Zhiting Hu. Critiques of world models. *arXiv preprint arXiv:2507.05169*, 2025.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing grpo on visual generation, 2025. URL `https://arxiv.org/abs/2505.07818`.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces, 2025a. URL `https://arxiv.org/abs/2412.14171`.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents, 2025b. URL `https://arxiv.org/abs/2502.09560`.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning, 2025c. URL `https://arxiv.org/abs/2507.12508`.

Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1389–1399, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025d. URL `https://openreview.net/forum?id=LQzN6TRFg9`.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think, 2025. URL `https://arxiv.org/abs/2410.06940`.

Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models, 2025. URL `https://arxiv.org/abs/2505.09694`.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=Sy8gdB9xx`.

Yuan Zhang, Fei Xiao, Tao Huang, Chun-Kai Fan, Hongyuan Dong, Jiawen Li, Jiacong Wang, Kuan Cheng, Shanghang Zhang, and Haoyuan Guo. Unveiling the tapestry of consistency in large vision-language models. *Advances in Neural Information Processing Systems*, 37:118632–118653, 2024.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023a.

Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. *Advances in neural information processing systems*, 36:31967–31987, 2023b.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: Learning 4d embodied world models, 2025. URL `https://arxiv.org/abs/2504.20995`.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness, 2025. URL `https://arxiv.org/abs/2503.21755`.

Yifan Zhong, Fengshuo Bai, Shaofei Cai, Xuchuan Huang, Zhang Chen, Xiaowei Zhang, Yuanfei Wang, Shaoyang Guo, Tianrui Guan, Ka Nam Lui, Zhiquan Qi, Yitao Liang, Yuanpei Chen, and Yaodong Yang. A survey on vision-language-action models: An action tokenization perspective, 2025. URL `https://arxiv.org/abs/2507.01925`.

Siyuan Zhou, Yilun Du, Yuncong Yang, Lei Han, Peihao Chen, Dit-Yan Yeung, and Chuang Gan. Learning 3d persistent embodied world models. *arXiv preprint arXiv:2505.05495*, 2025a.

Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Anyi Rao, Jiaqi Wang, and Li Niu. Light-a-video: Training-free video relighting via progressive light fusion, 2025b. URL `https://arxiv.org/abs/2502.08590`.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL `https://arxiv.org/abs/2504.10479`.