Do Vision-Language Models Have Internal World Models? Towards an Atomic Evaluation

Qiyue Gao*, **Xinyu Pi***, **Kevin Liu, Junrong Chen, Ruolan Yang**
Xinqi Huang, Xinyu Fang, Lu Sun, Gautham Kishore, Bo Ai, Stone Tao, Mengyang Liu
Jiaxi Yang, Chao-Jung Lai, Chuanyang Jin, Jiannan Xiang, Benhao Huang, Zeming Chen
David Danks, Hao Su, Tianmin Shu, Ziqiao Ma, Lianhui Qin, Zhiting Hu
Maitrix.org UC San Diego JHU Cornell Tech EPFL UMich
https://wm-abench.maitrix.org/ **equal contribution*

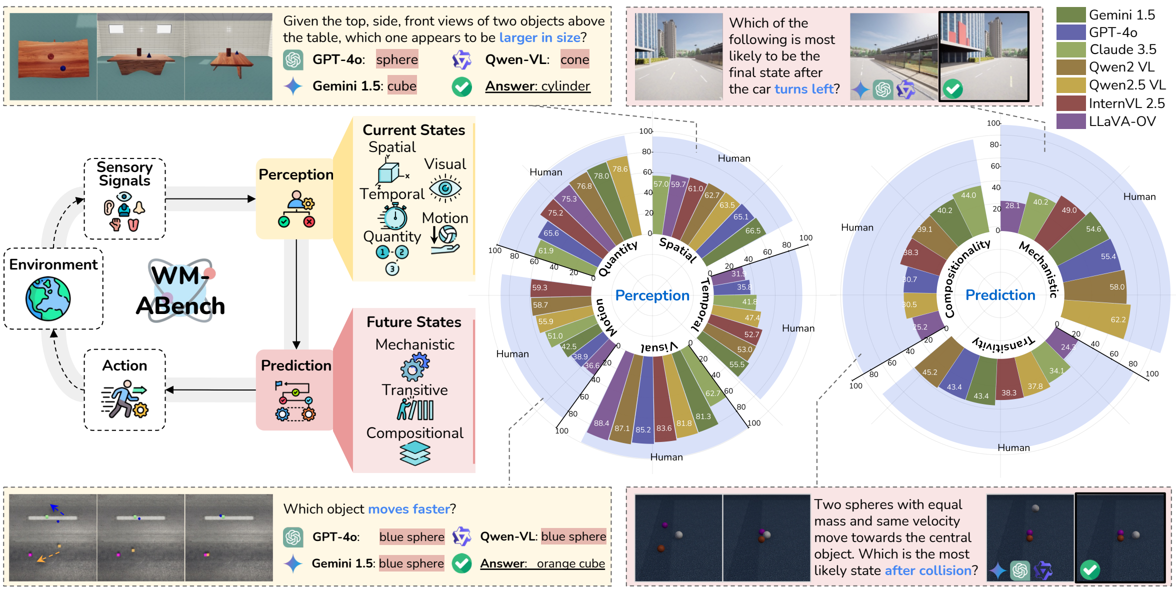


Figure 1: WM-ABench overview. Grounded in comparative psychology and cognitive science, our framework delineates the functioning of world models into two distinct stages: (1) *perception* stage, encompassing visual, spatial, temporal, quantitative, and motion perceptions, and (2) *prediction* stage, involving mechanistic simulation, transitive inference, and compositional inference. Large-scale evaluation reveals VLMs' limitations as WMs.

Abstract

Internal world models (WMs) enable agents to understand the world's state and predict transitions, serving as the basis for advanced deliberative reasoning. Recent large Vision-Language Models (VLMs), such as GPT-4o and Gemini, exhibit potential as general-purpose WMs. While the latest studies have evaluated and shown limitations in specific capabilities such as visual understanding, a systematic evaluation of VLMs' fundamental WM abilities remains absent. Drawing on comparative psychology and cognitive science, we propose a two-stage framework that assesses *perception* (visual, spatial, temporal, quantitative, and motion) and *prediction* (mechanistic simulation, transitive inference, compositional inference) to provide an atomic evaluation of VLMs as WMs. Guided by this framework, we introduce WM-ABench, a large-scale benchmark comprising 23 fine-grained evaluation dimensions across 6 diverse simulated environments with controlled counterfactual simulations. Through 660 experiments on 15 latest commercial and open-source VLMs, we find that these models exhibit striking limitations in basic world modeling abilities. For instance, all models perform at near-random accuracy when distinguishing motion trajectories. Additionally, they lack disentangled understanding—e.g., they tend to believe blue objects move faster than green ones. More rich results and analyses reveal significant gaps between VLMs and human-level world modeling.

1 Introduction

World models (WMs) in agents provide internal representations of the external world (Johnson-Laird, 1983; Wooldridge and Jennings, 1995). They simulate how the current state transforms to the next (Kappes and Morewedge, 2016; Russell

Findings of the Association for Computational Linguistics: ACL 2025, pages 26170–26195
July 27 - August 1, 2025

and Norvig, 2016), enabling agents to perform deliberative planning by predicting probable future states and choosing the most favorable one (Ha and Schmidhuber, 2018b; LeCun, 2022; Hu and Shu, 2023; Ai et al., 2024; Tian et al., 2025). Different environments operate under distinct mechanisms and dynamics—for example, the transition mechanics of autonomous vehicles, surgical robots, and rover-based spacecraft vary significantly. Prior work handled this diversity by building environment-specific world models, but they failed to adapt across various real applications.

Recent large-scale Vision-Language Models (VLMs; Google, 2023; OpenAI et al., 2024; Anthropic, 2024; Li et al., 2024b, *inter alia*) enhance generalist LLMs with visual semantics and encapsulate extensive knowledge of world dynamics, making them promising potential world models for general domains. Unlike video generative world models which directly generate the next states, VLMs can reason in their latent representation space and forecast via language. However, their language grounding and world simulation capabilities may still be insufficient in various aspects. For example, existing benchmarks have revealed their vulnerability in visual or spatiotemporal perception (Goyal et al., 2020; Shangguan et al., 2024; Fu et al., 2024b; Zhang et al., 2025) and future state prediction driven by intuitive physics (Bakhtin et al., 2019; Yi et al., 2020a; Bear et al., 2022). These limitations underscore the need for a more systematic evaluation. A robust world model must integrate multiple fundamental abilities in perception and prediction, yet previous studies that assess these aspects in isolation provide only a partial view. To address this, we propose an *atomic* evaluation framework, systematically testing the essential aspects (and their interactions) of a VLM's internal world model from first principles.

With theories and evidence from comparative psychology and cognitive science (Spelke, 2000; Knill and Pouget, 2004; Olmstead and Kuhlmeier, 2015), we first present a systematic conceptual framework to formalize the functioning of world models (Figure 1). We decompose the process into two stages: (1) *perception* stage, which involves *visual*, *spatial*, *temporal*, *quantitative*, and *motion* perceptions (Baillargeon et al., 1985; Merleau-Ponty, 2004; Coren et al., 2004; Hoffmann et al., 2011); and (2) *prediction* stage, which involves *mechanistic simulation*, *transitive inference*, and *compositional inference* (Hegarty, 2004; Barsalou, 2008; Prystawski et al., 2023).

Following this framework, we create `WM-ABench`, the World Model Atomic Benchmark that covers 23 fine-grained dimensions (Figure 2) of world modeling and over 100,000 instances curated from 6 different simulators (Dosovitskiy et al., 2017a; Gan et al., 2021; Szot et al., 2021a; Tao et al., 2024; Bear et al., 2022; Gu et al., 2023). By systematically manipulating environmental factors and simulating counterfactual actions, we generate incorrect states for models to differentiate from the correct ones to ensure *controlled* studies. We also measure human performance to verify the fairness and solvability of our problems.

We conduct 660 experiments on 15 state-of-the-art VLMs and find that, while they excel in certain aspects of visual and quantitative perception, they are surprisingly limited in many other dimensions (Figure 1).

Specifically, VLMs exhibit (1) weak perception of space, time, and motion; (2) insufficient knowledge of physical causality in intuitive physics and agentic actions; (3) limited transitive and compositional reasoning capabilities, showing near-random accuracy. Our further analyses reveal a lack of independent and robust world representations in VLMs, e.g., mistakenly associating color with speed. These findings reveal significant gaps between current VLMs and human-level world modeling, shedding light on the need for deeper understanding, grounding, and reasoning over the perceptual world and its transition mechanisms before VLMs can truly serve as generalist world models.

2 The Dual-Stage Conceptual Framework

We view world modeling as a two-stage process of *perception* and *prediction* (Knill and Pouget, 2004; Ha and Schmidhuber, 2018b). In the first stage, agents form internal representations of the current state by sensing and encoding environmental stimuli. In the second stage, agents use these internal representations to extrapolate future states, refining their model whenever new ground-truth observations arrive. This dual-stage framework explains (1) how raw sensory signals are converted into compact world representations; and (2) how these representations then guide forward simulations. The two stages of WMs are core functions in agents' advanced planning and decision-making. We brief our formulation here and leave the in-depth discussion in Appendix A.

2.1 Perception Stage

Perception involves extracting and organizing essential information from multi-sensory cues (Merleau-Ponty, 2004; Coren et al., 2004). It is not just bottom-up signal processing but also top-down inference grounded in prior knowledge. For instance, many intelligent animals (e.g., crows, dolphins, chimpanzees) grasp *object permanence* by maintaining accurate representations of hidden or partially occluded objects (Baillargeon et al., 1985; Hoffmann et al., 2011). While real-world perception spans multitudinous modalities (time, vision, temperature, humidity, audition, proprioception, etc.), we only focus on a limited number of perceptual dimensions. From an epistemic perspective, we design these dimensions to maximize the information captured about the external world while minimizing representational dimensions, ensuring they remain mutually orthogonal. From the practical perspective, we only cover the dimensions that modern VLMs can access. In this work, we consider 5 perceptual dimensions in our framework: space, time, motion, quantity, and vision. To make models' perceptual competency empirically testable along each dimension, we further break the major 5 perceptual dimensions into sub-dimensions:

Space and Time. All entities in spacetime must be located at some position and occupy some room or period, i.e., *extension*. Any pair of entities to exist in spacetime necessarily have spatiotemporal relations (e.g., front, left, before, after). Thus, we break spacetime down into *position*, *extension*, and *relations* (Reichenbach, 2012).

Motion. At every moment, motion can be described as a vector with speed and direction. The integration of motion along time forms a trajectory. Thus we consider *direction*, *speed*, and *trajectory* (Johansson, 1975).

Quantity. Axiomatically, quantities can either be discrete or continuous. For any pair of quantities, there could be quantitative relations (e.g., more, less, n-times) (Kaufman et al., 1949; Hurst and Piantadosi, 2024). We consider *discrete*, *continuous*, and *relations* (Kadosh and Dowker, 2015) quantities in this study.

Vision. In contrast with the previous four dimensions, the vision channel is extremely broad and hard to enumerate, including *orientation*, *density*, *edge*, *color*, *shape*, to name a few (Marr, 2010). We consider salient features like *color*, *shape*, and *material* (i.e., texture and reflexivity).

2.2 Prediction Stage

Once current-state representations are established, the agent must predict how future states evolve in response to both natural dynamics and possible actions. We distinguish 3 primary sub-dimensions:

Mechanistic Simulation. Agents should understand the causality of intuitive physical dynamics (e.g., motion, collisions) and intentional actions to simulate the next state (Hegarty, 2004; Barsalou, 2008). For example, predicting how one ball bounces off a wall or another ball depends on basic principles like momentum or elasticity.

Transitive Inference. Multi-step forecasts are often needed for tasks requiring long-horizon planning (Prystawski et al., 2023). Rather than only predicting the immediate next state, robust world models should extrapolate further into the future by chaining intermediate predictions.

Compositional Inference. Real-world scenarios usually involve multiple interacting objects and agents (e.g., two incoming balls hitting a third one from different directions). Agents must merge known mechanisms to predict novel outcomes, even if that specific combination has not been observed (Xu and Denison, 2009; Eckert et al., 2021; Gweon et al., 2010). This requires compositional reasoning, where partial pre-conditions (e.g., "hit from the left" plus "hit from the right") merge into an overall post-condition (e.g., "moves straight up").

3 The WM-ABench Benchmark

Following the above taxonomy, we design and implement corresponding tasks for benchmarking VLMs as WMs. Figure 2 presents a visual overview of our complete benchmark tasks, offering readers a high-level understanding of the framework.

Controlled Experiment and Causal Analysis. To ensure controlled evaluation, we exhaustively iterate over all dimensions, keep all other dimensions fixed as independent variables, and allow only one to vary at each data point. This methodology, which holds all variables constant except the independent ones across data points, allows us to draw causal conclusions (e.g., changing color causes the model to misperceive size). Rather than claiming generality or complete coverage (Raji et al., 2021; Saxon et al., 2024),

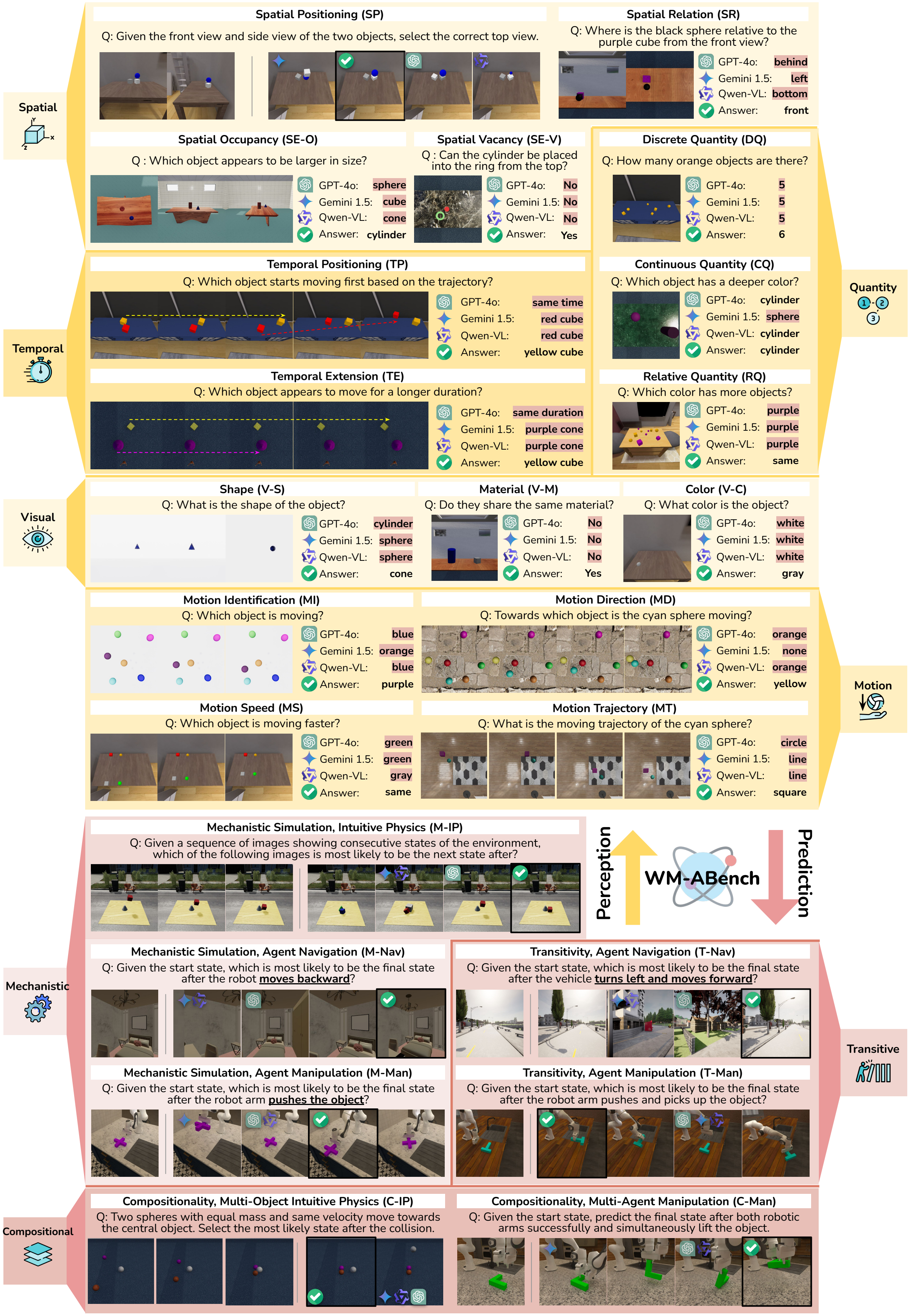
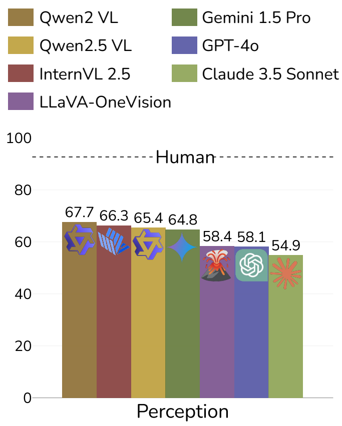


Figure 2: Overview of WM-ABench tasks. The Perception stage (top) covers Spatial, Temporal, Visual, Quantity, and Motion dimensions, each shown with example questions and outputs. The Prediction stage (bottom) includes Mechanistic Simulation, which covers Intuitive Physics (e.g., drop), Agent Navigation (e.g., turn left), and Agent Manipulation (e.g., push), plus Transitivity and Compositionality tasks that build on these transitions.



Model	Space								Time			
	SR		SE-V		SE-O		SP		TP		TE	
	Mani.	TDW	Mani.	TDW	Mani.	TDW	Mani.	TDW	Mani.	TDW	Mani.	TDW
Open-source Models												
NVILA	53.7	40.4	55.0	74.4	84.9	92.9	35.4	26.9	35.5	59.1	32.7	9.0
QWen2-VL-72b	42.2	72.6	42.6	57.0	87.6	93.6	58.9	47.1	46.5	75.5	36.4	53.4
QWen2.5-VL-72b	57.8	70.6	49.5	57.7	97.5	93.5	36.1	45.5	48.4	52.9	35.0	53.5
InternVL2.5-78b	44.2	65.8	51.3	73.5	98.4	93.0	31.3	30.9	43.5	63.6	43.9	59.7
Llama 3.2 vision-90b	50.7	47.5	50.0	43.8	78.6	25.2	25.0	24.6	33.4	16.8	38.0	37.2
LLaVA-OneVision	71.0	61.8	56.6	53.1	96.5	90.9	24.8	22.7	38.5	31.1	32.2	25.8
Closed-source Models												
GPT-4o-mini	37.6	37.7	61.3	69.1	98.2	35.3	37.0	38.1	31.7	47.1	29.0	40.6
GPT-4o	55.1	58.7	66.1	70.4	96.6	78.9	47.4	47.5	37.5	45.7	35.0	25.0
Claude 3.5 Sonnet	54.6	59.9	64.2	58.0	73.2	72.6	41.7	31.9	42.0	49.1	32.2	43.7
Gemini-1.5-flash	54.5	59.1	50.0	50.0	80.6	79.6	34.0	21.8	37.5	50.9	33.4	56.1
Gemini-1.5-pro	49.1	65.0	64.0	68.5	83.6	92.9	44.7	43.4	42.8	60.1	77.0	42.2
GPT-4.5*	72.0	79.0	55.0	52.0	95.0	92.0	57.0	38.0	47.0	66.0	55.0	67.0
OpenAI o3*	88.0	88.0	88.0	99.0	100.0	97.0	62.0	56.0	60.0	72.0	53.0	60.0
Claude 3.7 Sonnet*	56.0	72.0	60.0	76.0	82.0	63.0	70.0	60.0	35.0	38.4	22.0	49.0
Gemini-2.5-pro*	97.0	90.0	84.0	78.0	100.0	99.0	48.0	62.0	54.0	65.0	35.0	37.0
Random	25.0	25.0	50.0	50.0	50.0	50.0	25.0	25.0	33.3	33.3	33.3	33.3
Human	90.0	100.0	98.0	100.0	100.0	86.0	92.0	92.0	80.0	82.0	86.0	90.0

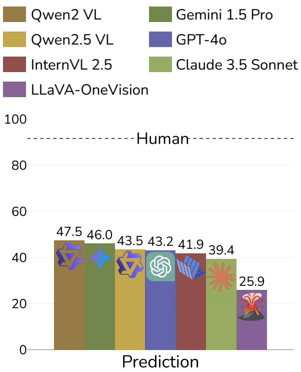
Model	Vision				Motion								Quantity				
	V-C		V-S	V-M	MS		MD		MI		MT		DQ		CQ	RQ	
	Mani.	TDW	TDW	TDW	Mani.	TDW	Mani.	TDW	Mani.	TDW	Mani.	TDW	Mani.	TDW	TDW	Mani.	TDW
Open-source Models																	
NVILA	94.6	88.0	98.2	56.3	34.2	45.1	52.6	53.5	33.3	35.3	24.4	14.8	57.8	81.4	63.9	73.5	58.0
QWen2-VL-72b	97.2	88.6	99.0	63.7	33.3	58.8	85.4	85.4	71.1	84.4	29.5	21.5	75.7	83.1	75.5	75.5	74.2
QWen2.5-VL-72b	94.7	89.4	91.5	51.4	47.6	74.7	69.8	71.9	59.7	66.8	28.9	27.7	73.1	84.7	88.0	75.9	71.3
InternVL2.5-78b	95.4	89.3	95.7	53.9	23.5	55.3	88.5	92.1	72.2	85.6	31.3	25.6	76.4	81.4	76.7	72.1	69.2
Llama 3.2 vision-90b	95.2	87.0	99.6	50.0	27.4	42.4	51.4	36.1	23.7	23.8	23.8	25.9	66.5	70.8	72.0	70.6	67.4
LLaVA-OneVision	95.2	88.8	98.4	71.2	45.2	40.5	54.1	53.1	33.7	25.5	26.1	14.5	67.5	76.9	81.1	74.0	76.8
Closed-source Models																	
GPT-4o-mini	98.1	87.8	99.7	50.0	24.1	36.7	32.5	37.8	53.9	51.4	26.6	17.1	31.3	60.1	73.9	60.1	54.5
GPT-4o	98.6	88.0	98.7	55.6	24.2	45.3	39.6	69.4	34.7	38.8	29.9	28.9	57.0	61.3	87.3	64.6	57.8
Claude 3.5 Sonnet	42.1	62.4	96.2	50.0	35.8	69.2	55.7	83.8	43.9	52.6	24.8	41.7	76.3	73.9	43.3	65.0	49.6
Gemini-1.5-flash	98.3	88.0	98.7	50.0	36.2	43.3	49.4	76.2	25.2	34.4	24.3	14.5	76.1	81.8	80.6	69.2	66.2
Gemini-1.5-pro	99.1	86.8	89.2	50.0	36.1	58.8	56.0	66.1	30.0	35.2	24.2	33.8	84.9	83.1	79.8	73.0	69.1
GPT-4.5*	100.0	87.0	99.0	51.0	25.0	57.0	75.0	92.0	79.0	88.0	28.0	61.0	61.0	87.0	89.0	76.0	70.0
OpenAI o3*	100.0	87.0	99.0	54.0	27.0	63.0	89.0	100.0	83.0	89.0	48.0	70.0	85.0	90.0	96.0	93.0	92.0
Claude 3.7 Sonnet*	27.3	76.0	98.0	53.0	32.0	61.0	43.8	85.7	34.0	46.0	22.0	25.0	72.0	69.0	49.0	69.0	49.0
Gemini-2.5-pro*	100.0	87.0	99.0	55.0	42.0	68.0	51.0	98.0	46.0	60.0	36.0	51.0	81.0	92.0	89.0	92.0	93.0
Random	25.0	25.0	25.0	50.0	33.3	33.3	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	50.0	33.3	33.3
Human	100.0	88.0	100.0	84.0	84.0	90.0	100.0	98.0	96.0	98.0	76.0	100.0	98.0	100.0	98.0	98.0	100.0

Table 1: Results on the Perception tasks in our WM-ABench, reported as accuracies (%). Models are evaluated in two simulators ManiSkill and ThreeDWorld). Cell shades indicate different performance levels (dark red indicates proficient performance; dark blue indicates performance close to random, and the lighter intermediate shades represent levels in between). **Random** and **Human** provide reference baselines. * denotes models evaluated on the subset (100 instances) of each dataset.

this benchmark aims to provide a precise, atomic diagnosis of models' perception and prediction, establishing a clear checklist for VLMs as world models.

Fighting Shortcuts and Spurious Correlations. Pre-trained models such as VLMs are known to rely on shortcuts and spurious correlations (Ye et al., 2024; Steinmann et al., 2024). To test whether VLMs can truly simulate and extrapolate into the next states, rather than relying on some spurious correlations, we generate hard negative options in our benchmark. We consider 2 methodologies to generate counterfactual states: counterfactual action and counterfactual previous states. Consider a ground truth transition triplet (S_t^*, a^*, S_{t+1}^*). For counterfactual action-based option generation, we fix the ground-truth previous state and perturb the action, and the transition becomes (S_t^*, a', S_{t+1}'). For counterfactual state-based option generation, we keep the ground-truth action, and perturb the previous state, so that the transition becomes (S_t', a^*, S_{t+1}''). False options generated under these methodologies often exhibit high visual similarity to the ground truth state, requiring models to possess a genuine understanding of world dynamics to distinguish and eliminate counterfactual states.

Data Collection. To rigorously evaluate models, a large and diverse dataset of test cases is essential. While manual data collection is possible, it is costly and often impractical for obtaining images where only a single factor varies for controlled studies. Therefore, we utilize compute-scalable simulation frameworks to synthetically generate a substantial number of test cases, minimizing human labor while ensuring precise control over variations. To



Model	Mechanistic Simulation								Transitivity				Compositionality		
	Car. nav	Hab. nav	Phys. coll	Phys. slide	Phys. drop	Mani. drop	Mani. lift	Mani. push	Car. nav	Hab. nav	Mani. pu-pi	Mani. pi-ro	TDW coll	Mani. push	Mani. lift
Open-source Models															
NVILA	31.3	16.7	28.4	44.6	44.7	48.9	33.8	26.4	32.4	4.2	29.1	26.4	26.6	25.5	31.8
Qwen2-VL-72b	62.0	65.3	30.4	40.3	59.2	95.3	91.9	20.0	50.1	43.8	55.7	31.2	24.3	51.3	41.8
Qwen2.5-VL-72b	76.6	70.9	36.5	51.9	55.0	88.9	85.5	31.9	61.1	34.7	24.4	30.9	21.6	34.8	35.2
InternVL2.5-78b	57.9	54.6	24.9	41.3	51.3	73.7	60.3	28.1	51.1	31.9	35.6	34.7	40.1	47.0	27.8
Llama-3.2-90b	24.7	33.3	24.0	25.4	24.7	23.7	26.3	23.1	23.9	33.6	25.6	20.8	25.9	24.1	25.3
LLaVA-OneVision	19.0	57.9	25.8	20.5	19.4	26.1	30.8	25.1	22.6	26.7	25.4	22.4	22.4	25.0	28.2
Closed-source Models															
GPT-4o	61.6	69.6	39.2	37.6	49.4	83.9	71.2	30.3	42.6	43.2	42.4	45.5	22.8	33.8	35.6
GPT-4o-mini	45.8	34.0	25.8	32.9	49.9	77.7	68.2	30.9	41.6	24.4	32.9	39.4	37.2	40.8	30.8
Claude 3.5 Sonnet	36.3	39.6	15.0	22.1	36.7	86.3	57.7	28.0	35.3	27.5	36.3	37.5	39.0	41.9	51.0
Gemini-1.5-flash	36.8	41.1	27.3	30.2	42.2	75.7	65.8	18.4	40.7	39.0	42.7	37.0	40.2	30.8	46.2
Gemini-1.5-pro	44.3	57.6	47.5	37.6	60.1	79.4	76.8	33.5	48.1	43.3	46.2	35.9	34.1	36.8	49.6
GPT-4.5*	71.0	34.0	55.0	72.0	56.0	86.0	59.0	40.0	47.0	34.0	44.0	35.0	32.0	41.0	39.0
OpenAI o3*	85.0	87.0	60.0	78.0	79.0	80.0	83.0	50.0	61.0	51.0	42.0	48.0	33.0	37.0	37.0
Claude 3.7 Sonnet*	59.0	51.0	38.5	49.0	67.7	96.0	56.4	39.4	44.0	26.0	34.3	48.4	41.5	29.0	49.5
Gemini-2.5-pro*	81.0	77.0	53.0	68.0	73.0	87.0	76.0	44.0	66.0	67.0	42.0	54.0	38.0	47.0	57.0
Random	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Human	98.0	98.0	100.0	98.0	86.0	100.0	100.0	100.0	78.0	90.0	80.0	82.0	84.0	88.0	100.0

Table 2: Results of the Prediction tasks in WM Benchmark (%). Our mechanistic simulation tasks cover three categories: intuitive physics (e.g., drop, slide collision), agent manipulation (e.g., lift, push), and navigation (e.g., turn left, move forward). * denotes models evaluated on the subset (100 instances) of each dataset.

avoid bias towards one single environment, we use a wide variety of simulation frameworks to create data, including ThreeDWorld (TDW; Gan et al., 2021), ManiSkill (Tao et al., 2024; Gu et al., 2023), Habitat 2.0 (Szot et al., 2021a), Physion (Bear et al., 2022), and Carla (Dosovitskiy et al., 2017a). Our diverse set of simulators allows the benchmark to incorporate various world dynamics that align with human intuition while leveraging a broad range of assets to address diverse questions.

4 Experiments

4.1 Experimental Setup

Evaluated Models. We evaluate a range of state-of-the-art VLMs on `WM-ABench`, including **closed-source models**: Gemini-1.5-pro, Gemini-1.5-flash (Google et al., 2024), GPT-4o (`gpt-4o-2024-08-06`), GPT-4o-mini (OpenAI et al., 2024), Claude 3.5 Sonnet (Anthropic, 2024), and **open-source models**: Qwen2-VL (72B) (Bai et al., 2023), Qwen2.5-VL (72B) (Team, 2024), InternVL2.5 (78B) (Chen et al., 2024), LLaVA-OneVision (72B) (Li et al., 2024a), NVILA (15B) (Liu et al., 2024c), and Llama 3.2-Vision (90B) (Meta AI, 2024).

We used the same system prompt (see Appendix E.2) and greedy decoding for all questions to maintain consistent output formatting. We evaluate model performance by comparing the parsed labels from model outputs to the ground-truth labels, and skip the instances where model outputs failed to be parsed by our template.

Human Evaluation. For each subtask, we randomly sample 50 questions and employ Amazon Mechanic Turk for human evaluations. We asked 3 evaluators for each question and finalized the labels via majority voting. The inter-rater agreement is measured by Fleiss' kappa on 50 samples for each task. All tasks are above moderate agreement (Fleiss' $k > 0.4$). Detailed instructions are available in Appendix E.1).

4.2 Main Results on Perception Tasks

Table 1 presents the results for perception tasks. In fact, closed-source VLMs do not exhibit an overwhelming advantage over open-source models in terms of perceptual capabilities. Across the 5 dimensions, Qwen2-VL achieves the highest overall performance with an average accuracy of 67.7%. All models still fall behind human-level perception, which is near-perfect or substantially more accurate across all the perception tasks.

Recommendation 1: Improve 3D Perception. Spatial tasks require grounding and reasoning over 3D semantics, including constructing robust scene representations from limited views. Even the most advanced models achieve less than 60% accuracy in spatial positioning tasks. These results suggest that current VLMs struggle to form robust internal 3D representations, in line with previous findings (El Banani et al., 2024; Zhang et al., 2025). We recommend future VLMs transition from relying solely on 2D semantics to incorporating 3D priors or explicit 3D representations.

Recommendation 2: Improve Temporal and Motion Understanding. We found that models struggle with coherent temporal representations across consecutive frames, as reflected in their low performance on temporal extension (TE). In contrast, they perform much better on tasks relying on a subset of frames, such as temporal positioning (TP). Similarly, while models perform relatively well on motion detection (MD), they exhibit *near-random performance* on motion trajectory (MT), which demands a thorough understanding of consecutive states. These results suggest that current VLMs still struggle to perceive and form dynamic scene representations. We recommend future VLMs incorporate temporal and motion priors by leveraging the rich visual dynamics in videos (Song et al., 2024; Ko et al., 2024).

4.3 Main Results on Prediction Tasks

Table 2 presents the results for prediction tasks, which generally pose greater challenges for VLMs than perception tasks. Qwen2-VL achieves the highest average accuracy of 47.5%. Similar to perception tasks, open-source VLMs perform on par with closed-source ones, and all VLMs fall behind human performance by a large margin.

Recommendation 3: Improve Cause-and-Effect Understanding. Our prediction tasks cover intuitive physics (e.g., drop, collision) and agent-initiated actions like navigation (e.g., turn left, move forward), and manipulation (e.g., lift, push). We find that VLMs struggle with predicting the post-condition of physical transitions and manipulation actions. For instance, on the ManiSkill simulator, Qwen2-VL achieves 95.3% accuracy in predicting the outcomes of dropping and 91.4% in lifting. However, its performance on pushing objects in the same environment is close to random, and its accuracy in predicting dropping outcomes on Physion decreases to 59.2%. These results suggest that current VLMs still struggle to reliably predict the cause and effect of physical processes and actions (Gao et al., 2018).

Recommendation 4: Improve Transitive and Compositional World Modeling. Our results also reveal that VLMs do poorly in transitive and compositional inference when reasoning over sequential or concurrent actions. In transitive inference tasks, models consistently underperform, with even the best achieving only 43.8% on the multi-step navigation task in Habitat, which is far below the human accuracy of 90.0%. Similarly, in compositional inference tasks, the gap remains substantial: the best models reach only 40.2% on the collision prediction task in TDW and 51.3% on the manipulation task in ManiSkill, compared to the human performance of 84.0% and 88.0%.

4.4 Performance of Frontier VLMs

We extend our evaluation to include the most recent frontier VLMs (Kavukcuoglu, 2025; OpenAI, 2025b,a; Anthropic, 2025), including visual reasoning models such as OpenAI o3. Due to computational constraints and API limitations, we evaluate these frontier models on a subset of 100 instances per task.

Frontier Models Reach Human Parity in Static Perception. Frontier VLMs (e.g., Gemini-2.5-Pro, o3, GPT-4.5) demonstrate clear improvements over previous state-of-the-art models across static perception tasks. On the Size Comparison (SE-O) task in ThreeDWorld, frontier models achieve 92-100% accuracy, up from GPT-4o's 88%, even exceeding the 93% human baseline. Similarly, Relative Position (SR) performance improves from 71-73% (LlaVA-OneVision, QWen2-VL) to 90-97% (Gemini-2.5-Pro).

Notable Gains in Mechanistic Simulation but Gaps Remain. We also observe substantial improvements from these models in mechanistic simulation tasks. In Carla Navigation task, accuracy rises from 60% (GPT-4o) to 81-85% for Gemini-2.5-Pro and OpenAI o3, though still below 98% human performance. Intuitive physics tasks (Physion Collision, Slide) advance from 47-60% to 60-79%.

Persistent Challenges in Spatial, Temporal, and Compositional Inference. Depite these performance gains, certain spatial and temporal perception tasks remain challenging. Multi-view spatial reasoning (SP) for these frontier models does not exceed 70%, well under the 90%+ human level. Temporal Extension (TE) shows minimal improvement, with only 7.3% gains from GPT-4.5. Advanced prediction tasks also shows little progress, with the highest ManiSkill Lift performance (57%, Gemini-2.5-Pro) significantly lagging human capabilities.

5 Further Analyses and Discussions

5.1 VLMs fail to represent different world attributes robustly and independently

World models should learn disentangled representations of key perceptual dimensions, like color or spatial position, for flexible compositional

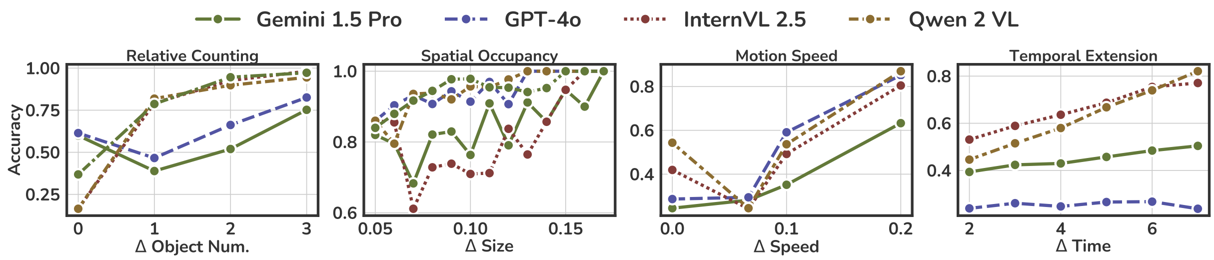


Figure 3: Model performance with respect to increasing stimulus differences, as discussed in Section 5.2. Here, $\Delta_{\text{Object Num.}}$, Δ_{Size}, Δ_{Speed}, and Δ_{Time} represent the differences in the number of objects, object sizes, object speeds, and object movement durations, respectively.

	GPT-4o	Gemini-1.5	Qwen2-VL	InternVL 2.5
$\Delta_{\texttt{collide}}$	-1.09	-0.52	-0.62	0.36
$\Delta_{\texttt{slide}}$	1.67	4.02	1.36	1.15
$\Delta_{\texttt{drop}}$	3.55	1.50	1.94	2.95

Table 3: Accuracy differences ($\Delta(\cdot)\%$) between filtered (correct state perception) and unfiltered inputs across physical transition tasks.

reasoning and categorical distinctions.[1] To assess entanglement, we perturb one dimension while keeping others constant and define the standardized Relative Entanglement (s-RE) as the normalized performance deviation between different perturbations: s-RE $= \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\bar{p}-p_i}{\bar{p}}\right|$, where p_i is model performance under the i_{th} perturbation and $\bar{p}$ is the average performance over all perturbations. Essentially, s-RE calculates the average relative deviation from the mean performance, reflecting model sensitivity to changes in that particular dimension. We control 5 perceptual dimensions (*color*, *shape*, *size*, *position*, *material*) and present the entanglement heatmap in Figure 4. Our results show that VLMs' representations of orthogonal world attributes have significant entanglement. Color and Shape are major sources of entanglement in multiple tasks. For example, color entangles with the Discrete Quantity (DQ) task, where s-RE ranges from 5% (Gemini-1.5 Pro) to 17% (Qwen-2.5 VL). Other dimensions like size and absolute position also exhibit relatively modest entanglement effects.

5.2 VLMs are sensitive to stimulus differences, but not fine details

As shown in Figure 3, we find that VLM performance is positively correlated with the physical differentiability of stimuli, such as differences in size, speed, or moving duration. This aligns with the previously reported "myopia" in model perception (Rahmanzadehgervi et al., 2024). Our observation suggests that VLMs can, to some extent, effectively ground language in corresponding physical attributes. This is evidenced by their strong performance in scenarios with large stimulus differences, which would otherwise be inexplicable. On the other hand, VLMs' perception capabilities are strongly influenced by the magnitude of the stimulus, irrespective of specific physical attributes. This suggests that while they perform well when distinctions are pronounced, they struggle with fine-grained, high-resolution perception, highlighting a significant gap in modeling subtle physical variations.

5.3 VLMs struggle with intuitive physics even under accurate state perception

Accurate next-state prediction relies on two key factors: correctly representing the current state and possessing sufficient mechanistic knowledge for transition simulation. Our analysis reveals that models often make mistakes in perception, which may contribute to their underperformance in prediction tasks, stemming from limited world modeling capabilities or accumulated errors resulting from perception failures. To disentangle perception from prediction, we present a further analysis of the mechanistic simulation from intuitive physics (*collide*, *slide*, *drop*). We retrieve instances where all models correctly answer all relevant perception questions, ensuring an accurate state representation (details provided in Appendix D.2). As Table 3 shows, performance on predicting the next state of *slide* and *drop* increases only marginally or even decreases for *collide*, indicating that limited perception capability is not the only cause. Rather, models lack the foundational physical knowledge to simulate

[1] Human perception also relies on interdependent encoding of visual properties, e.g., object attributes are represented as statistical summaries rather than in isolation (Whitney and Yamanashi Leib, 2018). However, unlike VLMs, human cognition can differentiate these summaries into discrete symbolic states, as evident in our evaluations, enabling precise world representation despite potential interdependencies.

Model	Space	Time		Vision		Motion	Mechanistic Simulation	
	SP	TP	TE	V-C	V-S	MT	Lift	Drop
Open-source Models								
QWen2-VL-72b	32.1	76.7	63.3	100.0	100.0	30.0	53.3	80.0
QWen2.5-VL-72b	35.3	76.7	76.7	100.0	100.0	23.3	76.7	73.3
InternVL2.5-78b	30.0	90.0	63.3	100.0	100.0	40.0	30.0	50.0
Closed-source Models								
Claude 3.5 Sonnet	26.7	70.0	60.0	100.0	100.0	26.7	40.0	66.7
Gemini-1.5-flash	43.3	83.3	60.0	100.0	100.0	30.0	36.7	50.0
Gemini-1.5-pro	53.3	80.0	63.3	100.0	97.7	40.0	53.3	76.7
GPT-4o	50.0	90.0	60.0	100.0	97.7	26.7	60.0	80.0
GPT-4o-mini	33.3	83.3	50.0	100.0	100.0	30.0	60.0	53.3
Random	25.0	33.0	33.0	25.0	25.0	25.0	25.0	25.0

Table 4: Results across 8 dimensions adapted from real-world datasets into `WM-ABench` design. We recast from real datasets and curate approximately 50 data points for each selected dimension.

object interactions accurately.

5.4 `WM-ABench` from Real-World Data

Transferring WM-ABench's controlled settings to real-world scenarios is challenging due to the need for precise attribute manipulations (e.g., altering object colors) and counterfactual state generation (e.g., applying forces to the target object). To draw some insights on whether our simulation findings might generalize to real-world data, we repurpose existing datasets (Walke et al., 2023; Liu et al., 2024a; Lai et al., 2014; Ammirato et al., 2018; Yu et al., 2023; Shangguan et al., 2025; Hwang et al., 2025) and recast them into our evaluation format. For each dimension, we curated approximately 50 data points and evaluated several top-performing models.

Results in Table 4 demonstrate consistent alignment between real-world and simulated performance patterns. For instance, models exhibit poor performance on Spatial Positioning (SP) and Motion Trajectory (MT), mirroring simulation trends. Conversely, they achieve near-perfect accuracy on Color and Shape recognition tasks. Additionally, Temporal Positioning performance (70–90%) consistently exceeds Temporal Extension (50–76%), confirming our earlier findings. These trends suggest that bias from our use of simulation data is likely to be minimal, and our findings tend to be generalizable.

6 Related Work

World Models. World models (WMs) predict how the current state transitions to the next, based on prior states and actions (Tolman, 1948; Battaglia et al., 2013). Traditionally, people train frame-level video-generative models specializing in some narrow domains. For instance, in robotics, WMs enable model-based reinforcement learning and trajectory prediction (Yang et al., 2023; Zhou et al., 2024); in autonomous driving (Wang et al., 2023; Hu et al., 2023), they facilitate path planning; and in gaming (Hafner et al., 2019; Bruce et al., 2024; Ha and Schmidhuber, 2018a,c), they power interactive simulations. Meanwhile, recent work (Brooks et al., 2024; Kang et al., 2024) has explored whether video-generation models can serve as world simulators that go beyond mere pixel-level synthesis. We instead investigate whether Vision-Language Models can capture world dynamics from large-scale training data, enabling them to function as generalist world models.

Benchmarks for Vision Language Models. Previous VLM benchmarks typically take a reductionist approach, measuring a wide range of perceptual capabilities while giving limited attention to how these perceptual dimensions interact and influence one another. For instance, there are works focusing on *visual semantics perception*, e.g., object categories, attributes, actions, agent-object interactions, emotions (Liu et al., 2024a; Li et al., 2024c; Liu et al., 2023), whereas others emphasize *low-level visual perception*, e.g., basic attributes, line segments, optical flow (Johnson et al., 2016; Fu et al., 2024b; Shiono et al., 2025), *spatiotemporal and motion*, e.g., geometry, event ordering, trajectories (Goyal et al., 2020; Mirzaee et al., 2021; Shangguan et al., 2024), or *next-state prediction* (often limited to intuitive physics) (Bear et al., 2022; Yi et al., 2020b). Compared to these efforts, our framework decomposes world modeling into atomic dimensions, offering a precise diagnosis of models' perception and prediction capabilities while establishing a clear checklist for VLMs as world models. Due to the page limit, we leave the comprehensive comparison between `WM-ABench` and existing benchmarks in Table 6 and Appendix B.2.

7 Conclusion

Our study provides the first atomic evaluation of VLMs' internal world modeling abilities with a cognitively inspired framework. While VLMs excel in scenarios with pronounced differences, they struggle with 3D and dynamic perception, fail to differentiate subtle physical distinctions, and exhibit failures in understanding world transitions of transitive and compositional scenarios.

Limitations

While simulators enable compute-scalable and cheap generation of dataset questions and answers, most simulators are difficult to tune or are incapable of photo-realistic image generation. As a result, we may be evaluating VLMs on somewhat out-of-distribution data as the images do not look realistic and the majority of the image data used to train VLMs likely come from real-world videos/images. While ray tracing is used in some of the ManiSkill-generated problems, better photo-realism can be achieved if higher quality assets are used and lighting is tuned better.

Ethics Statement

Human Study. We only involve human subjects to obtain the human performance on our benchmark. The institution's Institutional Review Board (IRB) considered this project exempt from ongoing review. All human evaluations in our study are conducted in Amazon Mechanical Turk (MTurk), with participants fairly compensated according to ethical guidelines. Further details on the evaluation process are provided in the Appendix E.1).

Societal Impact. Our benchmark aims to provide an atomic evaluation of VLMs' fundamental WM abilities. Since our experiments are conducted in simulated environments rather than using real-world data, our work does not raise ethical concerns related to privacy, bias, or societal impact.

Licenses. We strictly adhere to the protocols governing the academic use of all VLMs. Our benchmark is designed to evaluate the performance of VLMs across various dimensions, focusing primarily on inference rather than the training of large-scale models. As a result, it is unlikely to have a significant environmental impact.

AI Assistants In Research. Furthermore, regarding the use of AI tools, we only used ChatGPT for grammar checking and made minor refinements to our visualization code with its assistance, which is entirely unrelated to the scientific content or findings of our work.

Acknowledgments

We would like to thank Angela Shen, Kai Kim, Reventh Sharma, and Prathish Murugan for their contribution to early data collection, and Yi Gu, Yuheng Zha for assistance with setting up simulation environments.

References

Bo Ai, Stephen Tian, Haochen Shi, Yixuan Wang, Cheston Tan, Yunzhu Li, and Jiajun Wu. 2024. Robopack: Learning tactile-informed dynamics models for dense packing. In *Robotics: Science and Systems (RSS)*.

Phil Ammirato, Alexander C. Berg, and Jana Košecká. 2018. Active vision dataset benchmark. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2127–21273.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48.

Anthropic. 2025. Claude 3 & 7 sonnet system card. `https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf`. Accessed May 31, 2025.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Renée Baillargeon, Elizabeth S Spelke, and Stanley Wasserman. 1985. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. 2019. Phyre: A new benchmark for physical reasoning. *Preprint*, arXiv:1908.05656.

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. 2020. Cophy: Counterfactual learning of physical dynamics. *Preprint*, arXiv:1909.12000.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59(1):617–645.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. 2013. Simulation as an engine of physical scene understanding. *PNAS*.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. 2022. Physion: Evaluating physical prediction from vision in humans and machines. *Preprint*, arXiv:2106.08261.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. 2024. Video generation models as world simulators. *OpenAI*, https://openai.com/research/video-generation-models-as-world-simulators.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. 2024. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*.

Peter E. Bryant and Tom Trabasso. 1971. Transitive inferences and memory in young children. *Nature*, 232:456–458.

Haoran Chen, Jianmin Li, Simone Frintrop, and Xiaolin Hu. 2022. The msr-video to text dataset with clean annotations. *Computer Vision and Image Understanding*, 225:103581.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Stanley Coren, Lawrence M Ward, and James T Enns. 2004. *Sensation and perception*. John Wiley & Sons Hoboken, NJ.

Carl Craver, James Tabery, and Phyllis Illari. 2024. Mechanisms in Science. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2024 edition. Metaphysics Research Lab, Stanford University.

Arnaud Delorme, Guillaume A Rousselet, Marc J-M Macé, and Michele Fabre-Thorpe. 2004. Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cognitive Brain Research*, 19(2):103–113.

Stephanie Denison, Pallavi Trikutam, and Fei Xu. 2014. Probability versus representativeness in infancy: Can infants use naïve physics to adjust population base rates in probabilistic inference? *Developmental psychology*, 50(8):2009.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017a. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017b. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. *Preprint*, arXiv:2305.18654.

Johanna Eckert, Hannes Rakoczy, Shona Duguid, Esther Herrmann, and Josep Call. 2021. The ape lottery: Chimpanzees fail to consider spatial information when drawing statistical inferences. *Animal Behavior and Cognition*.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. 2024. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806.

Sagi Eppel. 2024. Do large language vision models understand 3d shapes? *Preprint*, arXiv:2412.10908.

Chris Frith and Raymond J Dolan. 1997. Brain mechanisms associated with top-down processes in perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1358):1221–1230.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024a. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024b. Blink: Multimodal large language models can see but not perceive. *Preprint*, arXiv:2404.12390.

Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Michael Lingelbach, Aidan Curtis, Kevin T. Feigelis, Daniel Bear, Dan Gutfreund, David D. Cox, Antonio Torralba, James J. DiCarlo, Josh Tenenbaum, Josh H. McDermott, and Dan Yamins. 2021. Threedworld: A platform for interactive multi-modal physical simulation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Qiaozi Gao, Shaohua Yang, Joyce Chai, and Lucy Vanderwende. 2018. What action causes this? towards naive physical action-effect prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 934–945.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE.

Rohit Girdhar and Deva Ramanan. 2020. Cater: A diagnostic dataset for compositional actions and temporal reasoning. *Preprint*, arXiv:1910.04744.

E Bruce Goldstein. 1989. *Sensation and perception*. Wadsworth/Thomson Learning.

Google. 2023. Gemini: A family of highly capable multimodal models. Technical report, Google.

Google, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Ankit Goyal, Kaiyu Yang, Dawei Yang, and Jia Deng. 2020. Rel3d: A minimally contrastive benchmark for grounding spatial relations in 3d. In *Advances in Neural Information Processing Systems*, volume 33, pages 10514–10525. Curran Associates, Inc.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2018. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127(4):398–414.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. 2023. Maniskill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations*.

Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Chenming Zhang, Shuai Liu, and Long Chen. 2024. Drivemllm: A benchmark for spatial understanding with multimodal large language models in autonomous driving. *Preprint*, arXiv:2411.13112.

Hyowon Gweon, Joshua B. Tenenbaum, and Laura E. Schulz. 2010. Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, 107(20):9066–9071.

David Ha and Jürgen Schmidhuber. 2018a. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31.

David Ha and Jürgen Schmidhuber. 2018b. World models. *arXiv preprint arXiv:1803.10122*.

David Ha and Jürgen Schmidhuber. 2018c. World models. *arXiv preprint arXiv:1803.10122*.

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

Mary Hegarty. 2004. Mechanical reasoning by mental simulation. *Trends in cognitive sciences*, 8(6):280–285.

Almut Hoffmann, Vanessa Rüttler, and Andreas Nieder. 2011. Ontogeny of object permanence and object tracking in the carrion crow, corvus corone. *Animal behaviour*, 82(2):359–367.

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

Zhiting Hu and Tianmin Shu. 2023. Language models, agent models, and world models: The law for machine reasoning and planning. *Preprint*, arXiv:2312.05230.

Michelle A Hurst and Steven T Piantadosi. 2024. Continuous and discrete proportion elicit different cognitive strategies. *Cognition*, 252:105918.

Minyoung Hwang, Joey Hejna, Dorsa Sadigh, and Yonatan Bisk. 2025. Motif: Motion instruction fine-tuning. *IEEE Robotics and Automation Letters*, 10(3):2287–2294.

Gunnar Johansson. 1975. Visual motion perception. *Scientific American*, 232(6):76–89.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *Preprint*, arXiv:1612.06890.

P.N. Johnson-Laird. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cognitive science series. Harvard University Press.

C. Kadosh and A. Dowker. 2015. *The Oxford Handbook of Numerical Cognition*. Oxford handbooks. Oxford University Press.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. 2024. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*.

Immanuel Kant, John Miller Dow Meiklejohn, Thomas Kingsmill Abbott, and James Creed Meredith. 1934. *Critique of pure reason*. JM Dent London.

Heather Barry Kappes and Carey K Morewedge. 2016. Mental simulation as substitute for experience. *Social and Personality Psychology Compass*, 10(7):405–420.

Edna L Kaufman, Miles W Lord, Thomas Whelan Reese, and John Volkmann. 1949. The discrimination of visual number. *The American journal of psychology*, 62(4):498–525.

Koray Kavukcuoglu. 2025. Gemini 2.5: Our most intelligent ai model. Google Blog. Accessed May 31, 2025.

Bruce E Kendall, Cheryl J Briggs, William W Murdoch, Peter Turchin, Stephen P Ellner, Edward McCauley, Roger M Nisbet, and Simon N Wood. 1999. Why do populations cycle? a synthesis of statistical and mechanistic modeling approaches. *Ecology*, 80(6):1789–1805.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. 2023. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. *Preprint*, arXiv:2306.11290.

David C Knill and Alexandre Pouget. 2004. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719.

Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B. Tenenbaum. 2024. Learning to act from actionless videos through dense correspondences. In *The Twelfth International Conference on Learning Representations*.

Alexander Kuhnle and Ann Copestake. 2017. Shapeworld - a new test methodology for multimodal language understanding. *Preprint*, arXiv:1704.04517.

Kevin Lai, Liefeng Bo, and Dieter Fox. 2014. Unsupervised feature learning for 3d scene labeling. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3050–3057.

Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *Preprint*, arXiv:2307.16125.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark. *Preprint*, arXiv:2311.17005.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024d. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024a. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, page 216–233, Berlin, Heidelberg. Springer-Verlag.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. 2024c. Nvila: Efficient frontier visual language models. *Preprint*, arXiv:2412.04468.

Francesco Mannella and Giovanni Pezzulo. 2024. Transitive inference as probabilistic preference learning. *Psychonomic Bulletin amp; Review*.

David Marr. 2010. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Brendan O McGonigle and Margaret Chalmers. 1977. Are monkeys logical? *Nature*, 267(5613):694–696.

Andrea Mechelli, Cathy J Price, Karl J Friston, and Alumit Ishai. 2004. Where bottom-up meets top-down: neuronal interactions during perception and imagery. *Cerebral cortex*, 14(11):1256–1265.

Maurice Merleau-Ponty. 2004. *The world of perception*. Routledge.

Meta AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. [Online; accessed 15-February-2025].

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online. Association for Computational Linguistics.

Stephen R Mitroff and Brian J Scholl. 2004. Seeing the disappearance of unseen objects. *Perception*, 33(10):1267–1273.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. 2024. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems (RSS)*.

Mary C Olmstead and Valerie A Kuhlmeier. 2015. *Comparative cognition*. Cambridge University Press.

OpenAI. 2025a. Gpt-4-5 system card. http://cdn.openai.com/gpt-4-5-system-card-2272025.pdf. Accessed May 31, 2025.

OpenAI. 2025b. Openai o3 and o4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf. Accessed May 31, 2025.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert

Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. 2022. Cripp-vqa: Counterfactual reasoning about implicit physical properties via video question answering. *Preprint*, arXiv:2211.03779.

Xinyu Pi, Mingyuan Wu, Jize Jiang, Haozhen Zheng, Beitong Tian, Chengxiang Zhai, Klara Nahrstedt, and Zhiting Hu. 2024. Uouo: Uncontextualized uncommon objects for measuring knowledge horizons of vision language models. *arXiv preprint arXiv:2407.18391*.

Ben Prystawski, Michael Y. Li, and Noah D. Goodman. 2023. Why think step by step? reasoning emerges from the locality of experience. *Preprint*, arXiv:2304.03843.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Preprint*, arXiv:2305.13786.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2024. Vision language models are blind. *Preprint*, arXiv:2407.06581.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Rajesh PN Rao and Dana H Ballard. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87.

H. Reichenbach. 2012. *The Philosophy of Space and Time*. Dover Books on Physics. Dover Publications.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. 2020. Intphys: A framework and benchmark for visual intuitive physics reasoning. *Preprint*, arXiv:1803.07616.

Stuart J Russell and Peter Norvig. 2016. *Artificial intelligence: a modern approach*. Pearson.

Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. 2024. Benchmarks as microscopes: A call for model metrology. In *Proceedings of the First Conference on Language Modeling*.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. 2024. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. *Preprint*, arXiv:2410.23266.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. 2025. TOMATO: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *The Thirteenth International Conference on Learning Representations*.

Daiki Shiono, Ana Brassard, Yukiko Ishizuki, and Jun Suzuki. 2025. Evaluating model alignment with human perception: A study on shitsukan in LLMs and LVLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11428–11444, Abu Dhabi, UAE. Association for Computational Linguistics.

K A Smith, J B Hamrick, Adam N Sanborn, P W Battaglia, T Gerstenberg, T D Ullman, and J B Tenenbaum. 2024. *Bayesian models of cognition : reverse engineering the mind*. MIT Press, Boston, MA.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

Elizabeth S Spelke. 2000. Core knowledge. *American psychologist*, 55(11):1233.

David Steinmann, Felix Divo, Maurice Kraus, Antonia Wüst, Lukas Struppek, Felix Friedrich, and Kristian Kersting. 2024. Navigating shortcuts, spurious correlations, and confounders: From origins via detection to mitigation. *arXiv preprint arXiv:2412.05152*.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. 2021a. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. 2021b. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. 2021c. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. 2024. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Ernő Téglás, Edward Vul, Vittorio Girotto, Michel Gonzalez, Joshua B Tenenbaum, and Luca L Bonatti. 2011. Pure reasoning in 12-month-old infants as probabilistic inference. *science*, 332(6033):1054–1059.

Susan P Thompson and Elissa L Newport. 2007. Statistical learning of syntax: The role of transitional probability. *Language learning and development*, 3(1):1–42.

Tongxuan Tian, Haoyang Li, Bo Ai, Xiaodi Yuan, Zhiao Huang, and Hao Su. 2025. Diffusion dynamics models with generative state estimation for cloth manipulation. *arXiv preprint arXiv:2503.11999*.

Edward C Tolman. 1948. Cognitive maps in rats and men. *Psychological review*, 55(4):189.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Joshua B. Tenenbaum, Daniel LK Yamins, Judith E Fan, and Kevin A. Smith. 2023. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *Preprint*, arXiv:2306.15668.

Martijn Van Otterlo and Marco Wiering. 2012. Reinforcement learning and markov decision processes. In *Reinforcement learning: State-of-the-art*, pages 3–42. Springer.

Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. 2023. Bridgedata v2: A dataset for robot learning at scale. In *7th Annual Conference on Robot Learning*.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. 2023. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. *arXiv preprint arXiv:2311.17918*.

Ziyue Wang, Chi Chen, Fuwen Luo, Yurui Dong, Yuanchi Zhang, Yuzhuang Xu, Xiaolong Wang, Peng Li, and Yang Liu. 2024. Actiview: Evaluating active perception ability for multimodal large language models. *Preprint*, arXiv:2410.04659.

Michael Weisberg. 2013. *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.

David Whitney and Allison Yamanashi Leib. 2018. Ensemble perception. *Annual review of psychology*, 69(1):105–129.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Susan Wood, Kathleen M Moriarty, Beatrice T Gardner, and R Allen Gardner. 1980. Object permanence in child and chimpanzee. *Anim. Learn. Behav.*, 8(1):3–9.

Michael Wooldridge and Nicholas R Jennings. 1995. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152.

Barlow C. Wright and Jennifer Smailes. 2015. Factors and processes in children's transitive deductions. *Journal of Cognitive Psychology*, 27(8):967–978. PMID: 26635950.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2024. Star: A benchmark for situated reasoning in real-world videos. *Preprint*, arXiv:2405.09711.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *Preprint*, arXiv:2312.14135.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa:next phase of question-answering to explaining temporal actions. *Preprint*, arXiv:2105.08276.

Fei Xu and Stephanie Denison. 2009. Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, 112(1):97–104.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024. Thinking in space: How multimodal large language models see, remember, and recall spaces. *Preprint*, arXiv:2412.14171.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. 2019. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. *Preprint*, arXiv:1908.02660.

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. 2023. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*.

Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James Matthew Rehg, and Aidong Zhang. 2024. MM-spubench: Towards better understanding of spurious biases in multimodal LLMs. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020a. Clevrer: Collision events for video representation and reasoning. *Preprint*, arXiv:1910.01442.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020b. Clevrer: Collision events for video representation and reasoning. *Preprint*, arXiv:1910.01442.

Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. 2023. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. 2025. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2023. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *Preprint*, arXiv:2207.00221.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. 2024. Robodreamer: Learning compositional world models for robot imagination. *Preprint*, arXiv:2404.12377.

Appendix

A Conceptual Framework Explained

A.1 The Dual-Stage Model

In general, world models (WMs) predict the future states of the world (based on observations of the past and current states) and the next action to be taken.[2] Formally, a WM θ recursively models the transition distribution: $P_\theta(S_t \mid S_1, a_1, \ldots, S_{t-1}, a_{t-1})$.

With theoretical consistency with previous research (Knill and Pouget, 2004; Ha and Schmidhuber, 2018b; Smith et al., 2024), we decouple the world modeling process into two stages. In the first stage, an agent encodes environmental stimuli from sensory signals into internal representations of the external world. In the second stage, the agent performs an extrapolation into possible future states. As time progresses and future states come into embodiment, the agent acquires ground truth data and updates its WM based on the divergence between prediction and reality. Unlike model-free RL, where the agent optimizes its action policy to maximize utility, world modeling is primarily a supervised learning problem, optimized through error reduction.

Given this dual-stage framework, world modeling can fail for two different reasons:

- Perceptual limitations that lead to problematic representations of external states. For example, an agent may encode different colors identically or conflate object size with speed.
- Prediction limitations that arise from insufficient mechanistic knowledge and lead to inaccurate simulations. For example, an agent may fail to follow momentum conservation or struggle with complex multi-object collisions.

Our framework addresses these challenges by considering both perception competency and prediction competency. We discuss how we further decompose the two stages to make this benchmark empirically possible.

A.2 Perception Process

Accurate prediction of future states relies on constructing precise representations of the current environment, a process known as perception (Goldstein, 1989). Perceptions have multitudinous dimensions: force, motion, temperature, sound, magnetics, space, time, quantity, and other agents, to name a few (Merleau-Ponty, 2004; Coren et al., 2004). Agents detect physical signals from external stimuli through their physiological sensors, converting multimodal raw inputs into physio-electronic signals. These signals then undergo complex bottom-up processing and functional transformations until they become recognizable to higher cognitive faculties involved in semantics and reasoning. Subsequently, top-down processing refines and hypothesizes representations of the world state. This bidirectional interaction continues iteratively until convergence. Such top-down processing enables more reliable and robust representations of the world. (Rao and Ballard, 1999; Frith and Dolan, 1997; Delorme et al., 2004; Mechelli et al., 2004) For example, animals such as crows, dolphins, and chimpanzees exhibit an understanding of object permanence, allowing them to maintain stable representations of objects even after they become obscured or disappear (Baillargeon et al., 1985; Hoffmann et al., 2011; Wood et al., 1980; Mitroff and Scholl, 2004). Due to the sensory capabilities of current VLMs, we keep spatial, temporal, quantitative, visual, and motion perceptions in our framework only to accommodate their status quo.

A.3 Prediction Process

To formalize how next-state prediction is processed, we draw on the compositional generalization framework from Dziri et al. (2023) and consider next-state predictions as a topologically sorted computational graph. Given a world environment, each state can be hierarchically decomposed into a structured sequence of nodes, where a node intuitively represents an object as a vector of attributes (e.g., speed, mass, direction, color). The underlying rationale is as follows:

1. Any valid program can be expressed as an equivalent topologically sorted computational graph based on dependency structures;
2. All world simulators function as programs;
3. Therefore, all world simulators have equivalent graphs. The topological order of the graph is determined by time and world dynamics.

Within this framework, future state prediction involves computing values for future nodes based on historical nodes, using the current time step as a cutoff. We then identify three necessary and collectively sufficient conditions.

[2] We regard the corner case of "inaction" as an element of the action space

Atomic Mechanistic Simulation. Mechanism simulation refers to the process of predicting a system's future states via representing and modeling the dynamic interactions among its component parts (Weisberg, 2013; Hegarty, 2004; Barsalou, 2008). This is to be contrasted with statistical predictions, which typically rely on correlational shortcuts and bypass explicit simulations of component behaviors over time (Kendall et al., 1999). Mechanistic knowledge about natural laws and object-specific properties, functions, behavioral patterns, and interactive dynamics are the fuel of mechanistic simulation (Craver et al., 2024), and are mostly learned in a posterior manner (Kant et al., 1934). Here we especially emphasize the atomicity of mechanistic simulation, i.e., single object motion or minimally viable object interactions. More complicated world dynamics are left to the compositional inference part. In terms of computational graphs, extrapolation based on atomic mechanistic knowledge corresponds to predicting the immediate next state based on the current observations and intended action. Formally, the inference process can be expressed as predicting

$$S_t \sim P(S_t \mid S_{t-1}, a_{t-1}, \ldots, S_{t-h}, a_{t-h}),$$

where h denotes the window size of historical states. All of $S_{t-1} \ldots S_{t-h}$ are observed.

Transitive Prediction. A WM that only predicts the immediate next state is hardly useful for complex planning in long-horizon tasks. Given a long hypothetical action sequence generated from any policy, a competent WM should accurately predict the corresponding future state. The statistically less biased way is to perform a step-by-step extrapolation into distant future states (Prystawski et al., 2023). Known as transitive inference, this ability is exhibited by many intelligent animals, including rats, monkeys, and human infants (Wright and Smailes, 2015; Thompson and Newport, 2007; Mannella and Pezzulo, 2024; Bryant and Trabasso, 1971; McGonigle and Chalmers, 1977). Formally, the inference process can be expressed as:

$$\begin{aligned}(\widehat{S_{t+q}}, \widehat{S_{t+q-1}}, \ldots, \widehat{S}_t) \\ \sim P_\theta(\widehat{S_{t+q}}, \widehat{S_{t+q-1}}, \ldots, \widehat{S}_t \mid \\ a_t, \ldots, a_{t+q-1}, a_{t+q}, \\ S_{t-1}, a_{t-1}, \ldots, S_{t-h}, a_{t-h}),\end{aligned}$$

and the most natural chain of thoughts (CoT) paradigm can be expressed recursively with:

$$\begin{aligned}\widehat{S_{t+i}} \sim P_\theta(\widehat{S_{t+i}} \mid a_{t+i}, \widehat{S_{t+i-1}}, a_{t+i-1}, \ldots, \\ \widehat{S}_t, a_t, S_{t-1}, a_{t-1}, \ldots, S_{t-h}, a_{t-h}).\end{aligned}$$

All $\widehat{S}_i$ are inferred, not observed. In terms of computational graphs, the transitive extrapolation through time takes the form of a forward chain.

Compositional Prediction. Previous research in cognitive psychology provides strong evidence that humans and intelligent animals can adjust their statistical expectations of outcome distributions when domain-specific mechanisms (e.g., agentic preferences, intuitive physics, or sampling procedures) conflict with the base rates of the population (Xu and Denison, 2009; Eckert et al., 2021; Denison et al., 2014; Gweon et al., 2010; Téglás et al., 2011). Extending from sampling processes to general next-state predictions, we identify another advanced inference mechanism: **integrating two or more known conflicting or synergistic mechanisms into a unified effect**.

We provide a motivating example to demonstrate what we mean by compositional inference: Consider a 2D plane with three balls, A, B, and C, each of equal weight.

- **Observation 1**: Ball A strikes Ball C from the lower left at a specific speed and angle, causing C to move upper right.
- **Observation 2**: Ball B strikes Ball C from the vertically symmetric lower right at the same speed, causing C to move upper left.

Now, suppose the agent has never observed a scenario where a single ball is simultaneously struck by two others. However, with basic physical intuition, the agent should infer that the leftward and rightward motion components cancel each other out, leading to a prediction that Ball C will move vertically upward.

Formally expressing compositional inference under the standard Markov Decision Process will be tricky (Van Otterlo and Wiering, 2012), because the monolithic state denotation S in the MDP formalism fails to capture a crucial fact that complex states can be decomposed into multiple atomic states. To bridge the gap, we define a conceptual-level composition function (Marr, 2010) $S_t = \text{Compose}\big[S_t^{(1)}, \ldots S_t^{(n)}\big]$ to denote the whole relationship between n component states and complex states. Then compositional inference can be expressed as:

$$S_t \sim \text{Compose}\big[S_t^{(1)}, \dots S_t^{(n)}\big)$$

where $S_t^{(i)} \sim P_\theta(S_t^{(i)} \mid S_{t-1}^{(i)}, a_{t-1}, \dots S_{t-h}^{(i)}, a_{t-h}\big]$. The compositional extrapolation through time takes the form of a collider with two or more parent node.

A.4 Final Remarks

In this work and the conceptual framework section, we adopt a specific interpretation of a world model, which we refer to as a *mechanistic world model*. For example, a statistical model (e.g., multi-linear regression, XGBoost) used to predict a client's risk of loan default based on features like yearly income, number of children, ethnicity, and medical history is undeniably a form of future-state prediction. However, such a model does not faithfully simulate world dynamics, as it lacks any representation of temporal progression. While it may provide useful predictive insights, it does not qualify as a mechanistic model because it blackboxes causal mechanisms and interactive kinetic dynamics. Thus, it is important to recognize that mechanistic simulation is not the only approach to extrapolating future states.

B Benchmark Details and Comparisons

B.1 Benchmark Statistics

Table 5 provides a structured overview of our benchmark tasks across different perceptual and predictive dimensions.

B.1.1 Perception

Spatial Perception

1. Spatial Relation (SR): Given the front and top view of two objects on a table, let the model infer the relative position of one object with respect to the other. This task evaluates whether the model can accurately discern spatial relationships based on visual cues.
2. Spatial Vacancy (SE-V): Given an object and a structure with a hollow space in the middle, let the model infer whether the object can fit into the hollow space. Thus testing the model's understanding of spatial constraints.
3. Spatial Occupancy (SE-O): Given two objects with different sizes, let the model infer which object is larger. Thus testing the model's understanding of object size.
4. Spatial Positioning (SP): Given the front and side view of a set of object arrangements, let the model infer the top view of the objects.

These multi-view tasks emphasize the model's ability to synthesize distinct viewpoints into a coherent three-dimensional representation of object arrangements.

Temporal Perception

1. Temporal positioning (TP): Compare two episodes of object motions from different views, and let the model differentiate which motion started first.
2. Temporal extension (TE): Compare two episodes of object motions from different perspectives, and let the model differentiate which motion lasts longer.

Collectively, these tasks investigate the aptitude of a model to maintain consistent temporal representations, estimate durations, and infer the correct order of events.

Visual Perception

1. Color (V-C): Differentiating whether two objects have the same color, or identifying which color the object is.
2. Shape (V-S): Determine the object's shape; another task involves differentiating whether two objects have the same shape.
3. Material (V-M): Differentiating whether the two objects have the same material or the model determine the material of the given object.

By isolating these fundamental visual features, the tasks provide targeted evaluations of how effectively a model can parse and differentiate basic object attributes, separate from any contextual or motion-based confounding factors.

Motion Perception

1. Motion Identification (MI): Given an episodes of motions of an object and a set of static objects, let the model decide which object is moving. This setup tests the model's capacity to identify the moving action of objects.
2. Motion Speed (MS): Given episodes of motions of two objects, let the model decide which object moves faster. This setup tests the model's capacity to track position changes over time and estimate relative velocity.
3. Motion Direction (MD): Given one moving object and a set of static objects, let the model determine which static object the moving object is heading towards. This setup tests the model's

capacity to track position changes over time and estimate relative moving direction.

4. Motion Trajectory (MT): Given two episodes of object motions, let the model decide whether their motion trajectories are the same, assessing its aptitude for higher-level spatiotemporal pattern recognition and object-specific path tracking across multiple frames.

Quantitative Perception

1. Discrete Quantity (DQ): Given the top view of objects on the table, let the model count the objects. This setup evaluates the model capacity for discrete numerical estimation.
2. Continuous Quantity (CQ): Given the top view of two objects with the same color theme, let the model determine which object has a darker shade.
3. Relative Quantity (RQ): Given the top view of objects with different colors, determine which color group has more objects. This setup probes counting skills, numerical reasoning, and perceptual comparisons in a visual context.

B.1.2 Prediction

Mechanistic Knowledge

1. Intuitive Physics (M-IP): Given a sequence of images showing consecutive states of the environment in which two objects move towards each other, let the model choose the most probable prediction of the next state. This setup evaluates the model capacity in physical reasoning.
2. Agent Navigation (M-Nav): Given an image of the start state, let the model choose what is most likely to be the final state after the robot/vehicle moves in a certain direction. This setup evaluates the model capacity in predictive reasoning.
3. Agent Manipulation (M-Man): Given an image of the start state, let the model choose what is most likely to be the final state after the robot arm does certain movements toward the object. This setup evaluates the model capacity in predictive reasoning for robot manipulation.

Transitivity

1. Agent Navigation (T-Nav): Given an image of the start state, let the model choose what is most likely to be the final state after the robot/vehicle moves through multiple directions in sequence. This setup evaluates the model capacity in predictive reasoning for autonomous navigation.
2. Agent Manipulation (T-Man): Given an image of the start state, let the model choose what is most likely to be the final state after the robot arm performs two actions in sequence. This setup evaluates the model capacity in multi-step predictive reasoning for robotic manipulation.

Compositionality

1. Multi-Object Intuitive Physics (C-IP): Given several images of two balls colliding with a third object at the same time, let the model predict the state after the collision occurred. This task evaluates the model's ability to perform compositional inferences about physical causality and object behavior.
2. Multi-Agent Manipulation (C-Man): Given an image of the start state, let the model choose what is most likely to be the final state after two robot arms do certain actions on one object simultaneously. This setup evaluates the model capacity in concurrent action predictive reasoning for robotic manipulation.

B.2 Relevant Benchmarks

We summarize and compare `WM-ABench` to existing benchmarks in Table 6, and provide more descriptive details below.

Visual semantics perception benchmark. This line of work primarily tests multimodal models' static (image) and dynamic (video) recognition competency of object class, object existence, object components, object properties, postures, actions, agent-object interactions, activities, emotions, social relations, functions, image quality, image style, scene, OCR, layouts. Such competency requires models to have diverse schematic knowledge and common sense about the anthropocentric world, and strong pattern recognition capability (i.e., recognize unobserved instances of known categories). However, the foundational perceptual and cognitive functions as enumerated in our work are out of coverage. Representative works in this line include MMBench(Liu et al., 2024a), MvBench (Li et al., 2024c), VSR(Liu et al., 2023), VQA v2 (Goyal et al., 2018), UOUO (Pi et al., 2024), MME (Fu et al., 2024a), STAR (Wu et al., 2024), Perception Test (Pătrăucean et al., 2023), MSRVTT-QA (Chen et al., 2022), NExT-QA (Xiao et al., 2021).

Visual perception Benchmark. This line of work primarily focuses on testing models'

Dimension	Subdimension	Total Cases	# Images per Case
Spatial Perception	Spatial Relation (SR)	1600 (ManiSkill) 8640 (ThreeDWorld)	2 (ManiSkill) 2 (ThreeDWorld)
	Spatial Vacancy (SE-V)	2880 (ManiSkill) 324 (ThreeDWorld)	3 (ManiSkill) 1 (ThreeDWorld)
	Spatial Occupancy (SE-O)	1152 (ManiSkill) 941 (ThreeDWorld)	1 (ManiSkill) 1 (ThreeDWorld)
	Spatial Positioning (SP)	6400 (ManiSkill) 2000 (ThreeDWorld)	2 (ManiSkill) 2 (ThreeDWorld)
Temporal Perception	Temporal Positioning (TP)	2000 (ManiSkill) 5646 (ThreeDWorld)	5 (ManiSkill) 6 (ThreeDWorld)
	Temporal Extension (TE)	2000 (ManiSkill) 7680 (ThreeDWorld)	5 (ManiSkill) 6 (ThreeDWorld)
Visual Perception	Color (V-C)	648 (ManiSkill) 1440 (ThreeDWorld)	3 (ManiSkill) 3 (ThreeDWorld)
	Shape (V-S)	1000 (ThreeDWorld)	3 (ThreeDWorld)
	Material (V-M)	1728 (ThreeDWorld)	1 (ThreeDWorld)
Motion Perception	Motion Identification (MI)	2688 (ManiSkill) 1913 (ThreeDWorld)	6 (ManiSkill) 4 (ThreeDWorld)
	Motion Speed (MS)	2304 (ManiSkill) 2304 (ThreeDWorld)	4 (ManiSkill) 8 (ThreeDWorld)
	Motion Direction (MD)	2688 (ManiSkill) 1913 (ThreeDWorld)	6 (ManiSkill) 6 (ThreeDWorld)
	Motion Trajectory (MT)	6912 (ManiSkill) 671 (ThreeDWorld)	7 (ManiSkill) 8 (ThreeDWorld)
Quantitative Perception	Discrete Quantity (DQ)	2520 (ManiSkill) 1134 (ThreeDWorld)	1 (ManiSkill) 1 (ThreeDWorld)
	Continuous Quantity (CQ)	756 (ThreeDWorld)	1 (ThreeDWorld)
	Relative Quantity (RQ)	2880 (ManiSkill) 1701 (ThreeDWorld)	1 (ManiSkill) 1 (ThreeDWorld)
Mechanistic Knowledge (Prediction)	Intuitive Physics (M-IP)	2160 (Physion: Slide) 3456 (Physion: Drop) 1728(Physion: Collide)	3 (Physion: Slide) 3 (Physion: Drop) 3 (Physion: Collide)
	Agent Navigation (M-Nav)	1575 (Carla) 1728 (Habitat-Lab)	1 (Carla) 1 (Habitat-Lab)
	Agent Manipulation (M-Man)	2304 (ManiSkill: Drop) 2304 (ManiSkill: Lift) 2304 (ManiSkill: Push)	1 (ManiSkill: Drop) 1 (ManiSkill: Lift) 1 (ManiSkill: Push)
Transitivity (Prediction)	Agent Navigation (T-Nav)	700 (Carla) 1956 (Habitat-lab)	1 (Carla) 1 (Habitat-lab)
	Agent Manipulation (T-Man)	1728 (ManiSkill: Push&Pick) 576 (Maniskill: Pick&Rotate)	1 (ManiSkill: Push&Pick Pick&Rotate)
Compositionality (Prediction)	Multi-Object Intuitive Physics (C-IP)	1296 (ThreeDWorld)	3 (ThreeDWorld)
	Multi-Agent Manipulation (C-Man)	2304 (ManiSkill: Push) 2304 (ManiSkill: Lift)	1 (ManiSkill: Push) 1 (ManiSkill: Lift)

Table 5: Benchmark Task Breakdown

competency of low-level, semantic-scarce visual perceptions, such as elementary visual attributes recognition (e.g. color, material, shape, size, texture), line segments, lighting, optical flow, insection, segmentation, overlapping area, corresponding points across different perspectives. CLEVR (Johnson et al., 2016), BLINK (Fu et al., 2024b), BlindTest (Rahmanzadehgervi et al., 2024), V* (Wu and Xie, 2023), KITTI (Geiger et al., 2012), ActiView (Wang et al., 2024), Shitsukan-eval (Shiono et al., 2025).

Spatiotemporal and motion perception Benchmark. This spatial perception tasks focus on evaluating models' ability to understand spatial relationships, configurations, and arrangements within static and dynamic contexts. Spatial perception tasks require recognizing geometric and topological relationships between objects, including proximity, alignment, containment, intersection, adjacency, relative positions, orientation, and distances, as well as multi-perspective alignment and integration. Rel3D

Benchmark	Spatial				Temporal		Quantity			Visual			Motion				Mechanistic			Transitivity		Compositionality	
	SP	MV	SO	SE	TP	TE	DQ	CQ	RQ	AR-S	AR-M	AR-C	MD	DI	SC	MT	IP	Nav	Mani	Nav	Mani	IP	Mani
BLINK (2024b)	✔	✔					✔	✔						✔									
SpatialRGPT (2024)	✔		✔																				
VSI-Bench (2024)	✔		✔				✔																
VL-CheckList (2023)	✔		✔								✔	✔											
CLEVR (2016)	✔		✔				✔			✔	✔	✔											
CLEVRER (2020a)	✔		✔				✔			✔	✔	✔					✔					✔	
NLVR (2017)	✔		✔							✔		✔											
NLVR2 (2019)	✔						✔																
ShapeWorld (2017)	✔						✔			✔		✔											
VALSE (2022)	✔						✔												✔				
MVBench (2024d)	✔						✔			✔		✔		✔									
MMBench (2024b)	✔						✔			✔	✔	✔											
MME (2024a)	✔						✔					✔											
VQA(v2) (2018)	✔						✔					✔											
NExT-QA (2021)	✔						✔																
V* (2023)	✔						✔				✔	✔											
ActiView (2024)	✔						✔																
SEED-Bench (2023)	✔						✔			✔	✔	✔											
Perception Test (2023)	✔									✔			✔				✔					✔	
BlindTest (2024)							✔					✔											
SHAPES (2016)										✔		✔											
TOMATO (2024)										✔				✔	✔	✔							
STAR (2024)																			✔		✔		
Ours	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

Table 6: Comparison between different benchmarks.

(Goyal et al., 2020), SPARTQA (Mirzaee et al., 2021), SpatialSense (Yang et al., 2019), 3D-Shape-Test (Eppel, 2024), DriveMLLM (Guo et al., 2024), VSI-Bench (Yang et al., 2024).

Temporal and motion perception go hand-in-hand. Since perception of time typically relies on changes and motions, temporal and motion perception tasks lack clear boundaries. Temporal perceptions typically involve: action count, attribute change, action sequence and procedure understanding, event order, scene transition, character order, and action antonym. Motion perception typically involves: direction (e.g. left, clockwise, up, outward), speed, and trajectory. Representative works include: TOMATO (Shangguan et al., 2024), CATER (Girdhar and Ramanan, 2020).

Next state prediction benchmark. Next-state prediction is different from the perception, visual-semantic inference (typically about static, state-intrinsic specifications, such as categories, properties, functions, relations, emotions, and intentions), and verbal-logical reasoning (e.g. comparison, logic operations) tasks introduced above. Emphasizing extrapolating objective world states, next-state prediction requires grounded knowledge (in contrast with verbal knowledge) of world mechanics and dynamics. That is, how the environment changes and transits. The agent-centric next-state prediction would additionally emphasize action-transition and interactive dynamics. Representative works includes: Physion (Bear et al., 2022), Physion++ (Tung et al., 2023), Phyre (Bakhtin et al., 2019), CLEVRER (Yi et al., 2020b), IntPhys (Riochet et al., 2020), CoPhy (Baradel et al., 2020), CRIPP-VQA(Patel et al., 2022), SEED-Bench (Li et al., 2023).

C Simulator Setup

We describe how we set up each simulator for generating test cases for the proposed world model benchmark.

C.1 ThreeDWorld

We use the ThreeDWorld simulator (Gan et al., 2021) to generate images for the perception tasks, with intuitive physics (e.g., collisions) modeled using the Physion framework (Bear et al., 2022). We first select a curated set of pre-packaged scenes, objects, and materials from ThreeDWorld. For most questions, we spawn selected objects like cubes and spheres onto the floor or a table within one of the selected scenes, then randomly varying their color, size, position, and material to ensure diversity across samples. We then render images of the scene from multiple viewing angles, including top-down, front, and side views, to ensure diverse perspectives for question generation. For tasks requiring a sequence of images to demonstrate object movement, we first render an initial image capturing the objects in their original spawned positions. We then teleport the objects to new locations, rendering an image after each movement until the desired movement is complete. After generating all images for a given scene, we clear the scene by removing all objects. We then repeat the process in a newly selected scene, beginning with the random selection of object color, size, position, and material to ensure diversity across scenes.

C.2 ManiSkill

We use both ManiSkill framework version 2 (Gu et al., 2023) and 3 (Tao et al., 2024)) and we do the following to generate the images used in the dataset. For most questions, we spawn objects such as cubes and spheres onto a table in the ReplicaCAD apartment scene (Szot et al., 2021b) and render an image of the scene. For tasks requiring before-and-after images, we first render the scene, then teleport objects to new locations, and render again. For static tasks, following a similar approach to ThreeDWorld, we place several cubes and spheres on a table within the provided background. To ensure sample diversity, these objects vary randomly in color, size, and position. We then render images from multiple viewpoints (e.g., top-down, front, and side angles) to provide diverse perspectives for subsequent question generation For tasks requiring a sequence of images to capture object movements (similar to ThreeDWorld), we first render an initial image depicting the objects in their original positions. We then relocate the objects, rendering an image after each repositioning, until the intended motion is complete. For agent manipulation questions, we use the RoboCasa dataset (Nasiriany et al., 2024). One scene features two Franka Panda arm robots, while another includes only one. Using ManiSkill, we randomize several dimensions in our test cases, including object geometry, object color, apartment layout, and visual style. Teleoperation tools collect demonstrations of both successful and failed trajectories for each object geometry, with physics simulation enabled. Other randomized dimensions are synthetically generated. For the question-answer pairs, we render the first and last frames of each demonstration

C.3 Physion

We use various setups (e.g. collide, slide, drop) provided by Physion (Bear et al., 2022) and instantiate various objects and run the physics simulation to simulate various physical effects such as dropping or rolling, testing mechanistic state transition knowledge. An image is captured before the simulation starts and frames are iteratively being captured after a simulation has run.

C.4 Carla

Using Carla (Dosovitskiy et al., 2017b) simulator we do the following to generate the images used in the dataset. We first select a curated set of pre-packaged towns, weather, and car agents from Carla. For the majority questions we spawn a car agent at a random position onto one of the selected scenes, instantiated with a random weather from the selected weathers. We refined the control mechanisms of the car agent to enhance realism, ensuring that actions such as moving forward and turning exhibit natural and physically plausible behavior that aligns with their corresponding natural language descriptions. We then provide commands instructing the car to move in a specified direction or make a turn at an intersection. The simulator subsequently renders and captures a sequence of images depicting the car's actions, which are used to construct our dataset.

C.5 Habitat

We use Habitat 2 (Szot et al., 2021c) to render the HSSD (Khanna et al., 2023) dataset, which includes a large number of simulated indoor scenes, to create navigational transition action-state pairs. We use discrete actions that enable the agent in the simulation to move around and change viewing directions. Pre-condition and post-condition images are generated for the dataset.

D Addendum to Results

D.1 Entanglement in Perception Tasks

Figure 4 provides a heatmap showing Relative Entanglement (s-RE) scores. These scores represent the average performance deviation (s-RE) across a subset of the highest-performing models (GPT-4o, Gemini-1.5 Pro, Qwen2-VL, Qwen2.5-VL, and InternVL-2.5).

D.2 Evaluation via Perceptual Queries

We devised three targeted perceptual queries to rigorously assess the models' understanding of dynamic scene attributes:

- Does the scene contain a moving object?
- What is the observed color of the moving object?
- What is the observed shape of the moving object?

These queries are intended to verify that the models accurately detect and interpret the moving object—a critical prerequisite for successful task performance. Notably, this stringent evaluation protocol resulted in the exclusion of approximately

Figure 4: Heatmap showing s-RE scores that quantify the relative impact of different physical dimensions on perception task performance. Higher values (darker red) indicate stronger entanglement between a physical dimension and task performance, while lower values (lighter colors) suggest weaker relationships.

30% of the test instances, thereby underscoring the robustness and effectiveness of our filtering approach.

E Evaluation and Reproducibility

E.1 Human Evaluation

We recruited Mechanical Turk Masters on Amazon Mechanical Turk. Annotators were required to have a 98% HIT approval rate, at least 100 approved HITs, and reside in the United States. Each problem was evaluated by three annotators, with the final label determined by majority vote (ties were resolved by randomly selecting an answer). Workers were paid $1 per HIT (10 examples per HIT, where each example took about 20 to 30 seconds). For each task, we provide brief task instructions. Here is an example:

> You will be provided with six images, each representing evenly spaced frames from a video. Two moving objects are visible in the frames. Your task is to determine which object started moving first, `$object_name1` or `$object_name2`?

E.2 VLM Evaluation

We develop a general prompt framework to support both open-source and closed-source models under a unified design. The system is built around a general evaluator class capable of loading and executing multiple model types. Our data management strategy involves categorizing datasets and assigning each task a unique identifier. This allows seamless retrieval of the corresponding datasets for varied evaluation scenarios. To streamline prompt creation and ensure consistency, we maintain prompt template files containing different formats for a wide range of question types. Below is the system prompt that we use to regularize the output format from the models:

> You are a helpful assistant. You will be given a question to answer. If it is a multichoice question, return the index of your choice 1,2,3,4 or A,B,C,D depending on the question, and then followed by any explanation necessary. If it is a yes/no question, clearly answer “yes” or “no” at first, and then follow with your explanation if needed. If you are asked to choose the images, please note that the last four images of all the given images are your choices. And your answer should be 1 or 2 or 3 or 4 pointing to these last four images.

E.3 Computational Resource

We use H100 GPUs to run the experiments, with an estimated total runtime of 200 GPU hours for model inference.

E.4 License and Research Artifacts

Simulators. Regarding the licenses or terms for the use and distribution of artifacts, our paper utilizes the following simulation environments: ThreeDWorld (Gan et al., 2021), ManiSkill (Tao et al., 2024), Physion (Bear et al., 2022), Carla (Dosovitskiy et al., 2017b), and Habitat

2 (Szot et al., 2021c). Detailed documentation on these artifacts is provided in Appendix C. Their corresponding licenses are listed in table 7.

Simulators	URL	License
ThreeDWorld (TDW)	Link	BSD-2-Clause
ManiSkill	Link	Apache v2.0
Physion	Link	MIT license
Carla	Link	MIT license
Habitat 2	Link	MIT license

Table 7: License information for the simulators used.

Having reviewed the rights and terms of these licenses, we confirm that we have fully complied with their requirements. Our work will be released under the MIT License, ensuring no legal issues arise. As discussed in the ethics statement of section 7, our benchmark is designed to provide a fundamental evaluation of the core world modeling abilities in VLMs, which generally do not involve social aspects or social reasoning. We stipulate that our work should be used strictly for academic purposes. Since our data is collected from simulators, it is fully anonymized, does not contain personally identifiable information, and does not require additional measures to verify the absence of sensitive information relevant to individuals.

Software. We have utilized existing software packages provided by the creators of the simulation environments and Transformers (Wolf et al., 2020). The packages and versions are listed in Table 8.

Package	Version
absl-py	2.1.0
accelerate	0.30.1
ai2thor	5.0.0
aiohttp	3.9.5
aiosignal	1.3.1
altgraph	0.17.4
annotated-types	0.7.0
antlr4-python3-runtime	4.9.3
anyio	4.4.0
asttokens	2.4.1
async-timeout	4.0.3
attrs	23.2.0
aws-requests-auth	0.4.3
backcall	0.2.0
bitsandbytes	0.43.1
blinker	1.8.2
boto3	1.34.118
botocore	1.34.118
cachetools	5.3.3
certifi	2024.2.2
cffi	1.16.0
charset-normalizer	3.3.2
chex	0.1.82
click	8.1.7
cloudpickle	3.0.0
comm	0.2.2
contextlib2	21.6.0
contourpy	1.2.1
cryptography	42.0.8
cycler	0.12.1
datasets	2.17.1
debugpy	1.6.7
decorator	5.1.1
decord	0.6.0
Deprecated	1.2.14
diffusers	0.30.0
dill	0.3.8
distro	1.9.0
docker-pycreds	0.4.0
einops	0.8.0
entrypoints	0.4
etils	1.5.2
exceptiongroup	1.2.1
executing	2.0.1
ffmpeg	1.4
filelock	3.14.0
Flask	3.0.3
flax	0.7.0
fonttools	4.51.0
frozenlist	1.4.1
fsspec	2023.10.0
google-ai-generativelanguage	0.6.6
google-auth	2.29.0
google-cloud-storage	2.16.0
grpcio	1.64.1
huggingface-hub	0.24.5
imageio	2.34.1
ipykernel	6.29.3
ipython	8.12.0
jupyter-client	7.3.4
jupyter-core	5.7.2
matplotlib	3.9.0
ml-collections	0.1.1
numpy	1.26.4
nvidia-ml-py	12.535.161
openai	1.30.3
opencv-python	4.10.0.84
pandas	2.2.2
protobuf	4.25.3
psutil	5.9.0
pyarrow	16.1.0
Pygments	2.18.0
requests	2.32.2
rich	13.7.1
scipy	1.10.1
torch	-
torchaudio	-
torchvision	-
tqdm	4.66.4
transformers	4.29.2

Table 8: Python Packages