Newton - A Small Benchmark for Interactive Foundation World Models

Anonymous authors
Paper under double-blind review

Abstract

Foundation world models (FWMs) are an emerging class of generative model that aim to generate realistic, interactive worlds from pre-training on video data. FWMs in particular promise to provide an online, stable environment for training generalist embodied agents. However, contemporary models suffer from several drawbacks, including poor object permanence, and struggle to apply physical principles consistently. Unlike large language models (LLMs) and video models, no benchmarks currently exist to specifically evaluate foundation world models' performance in the context of interactivity. We present Newton, a series of datasets and benchmarks for training and evaluating small interactive FWMs, particularly on long-context memory and physics tasks. Newton-OP includes 5,000 examples of occlusion and camera rotation, aiming to evaluate models' ability to recall objects in 3D space over long time periods. Newton-Physics additionally includes 5,000 examples of interactive rigid body physics, evaluating both action following and physical accuracy. We additionally release code to evaluate models, and demonstrate the performance of common baselines[1].

1 Introduction

1.1 Foundation World Models

Foundation world models, (also occasionally called "interactive world models" (Wu et al., 2024), and "open-ended models" (Hughes et al., 2024)), are generative models that aim to simulate diverse 3D environments. While there is not yet a consensus definition, paradigmatic examples include Genie 1 (Bruce et al., 2024) and 2 (DeepMind, 2024a), Cosmos (NVIDIA et al., 2025), Oasis (Decart, 2024) and "The Matrix" (Feng et al., 2024). In general, these models are trained on large amounts of video and game data, and predict frames of video, optionally conditional on actions or text, for arbitrarily long durations. This distinguishes FWMs from the more general category of video models, such as Sora (OpenAI, 2024) and CogVideoX (Yang et al., 2024) which often predict 5-30 seconds of video at a time, and are not typically interactive. Both approaches have been explored and evaluated as physical world models, but FWMs uniquely promise interactive feedback to agents, which is useful for training robotics and embodied AI, in addition to applications for generative games and media.

1.2 Physical Benchmarks

The question of whether modern video models learn physical principles is a matter of hot scientific debate: while models are often evaluated on perceptual benchmarks, such as VBench (Huang et al., 2023) and FVD (Unterthiner et al., 2019), it is unclear whether perceptual improvement correlates with strong physical understanding. Sora and its replications showed poor performance upon release. However, DeepMind (2024b) recently demonstrated both perceptual quality and strong physical understanding, suggesting that the difference may be a matter of scale. Nonetheless, there has been a flurry of recent work on physics

[1]Code is available at https://anonymous.4open.science/r/newton-348F

benchmarks, including Physion (Bear et al., 2022), IntPhys (Riochet et al., 2020), VideoPhy (Bansal et al., 2024), Physics-RW (Zhao et al., 2024), PhyGenBench (Bansal et al., 2024), PhysBench (Chow et al., 2025), and others. These build on a long history of physics benchmarks in robotics, arguably themselves inheriting from "block world" demos of the 20th century. Of these, some focus on vision-language models' ability to understand textual prompts, or score model outputs using language models themselves. To our knowledge, no benchmarks currently involve precise engagement with the physics of the model in an interactive context. With this in mind, we propose two novel design goals for Newton:

1. Interactivity: FWMs must be able to understand actions much more granularly than general video models, and maintain a persistent world-state. This is particularly important for RL use-cases, where inaccurate physics can harm real-world transfer, and inapplicable to video models that may be generating several scenes at once in response to a high-level prompt.
2. Simplicity: Many current benchmarks focus on diversity and physical realism. While this is necessary, due to the scale of video data required, realistic FWMs can be prohibitively expensive to train for many researchers. In the vein of projects such as TinyStories (Eldan & Li, 2023), Newton aims to provide a simple, low-cost train and test set that can be used to quickly develop new FWM research, and to provide a baseline for future work.

2 Object Permanence

As discussed above, object permanence is a fundamental aspect of physical understanding, often lacking in current models. Empirically, Oasis loses track of objects after just a few frames out-of-view, and Genie 2 reports to lose coherence after about a minute of generation. More so than LLMs, where long-context retrieval is desirable but not necessary, persistent state is essential for world models. Models that cannot do this for arbitarily long durations have arguably failed to internalise the 3D nature of their training data. We propose a simple benchmark to evaluate this, where a model is asked to remember the location of objects after a user moves them out of view. For simplicity, we choose a single camera with only two actions - rotate 90 degrees left or right. The camera observes a scene containing randomly generated terrain, and 1-3 colored cubes [2]. Cubes are kept at constant size, so the model can, in theory, estimate their depth monoscopically. Newton-OP contains a training set of 5000 15-second videos, a finetuning set of 1000 30-second videos, and a test set of 500 30-second videos. Examples are shown below. A camera motion takes 15 frames, and occurs with a likelihood of 0.005 per frame.

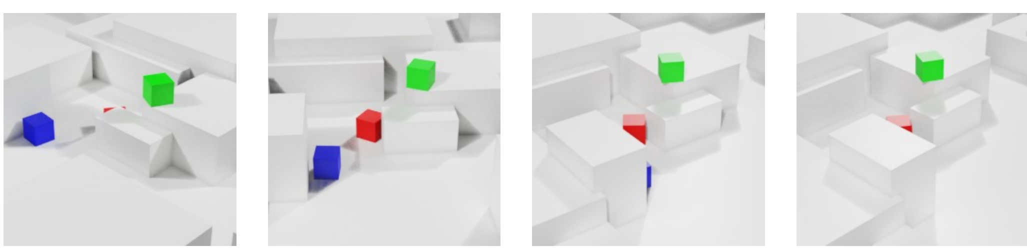

Figure 1: Camera movement and occlusion

The test set exclusively contains videos with partially occluded cubes, visible at the end of the example. The model is prompted with the first 5 frames and the actions, and must produce a final frame. This final frame is then evaluated using three weighted metrics (20%, 20%, 60%):

1. Pixel-level: The images are compared using a standard MSE reconstruction metric.
2. Feature-level: The images are compared using a learned feature metric, LPIPS.

[2]Data is generated in Blender 4.3.

3. Object-level: Since we are using known geometry, we can accurately estimate the 6D pose of each cube in either image using non-neural techniques[3]. We can then compute the mean absolute error of the cube locations.

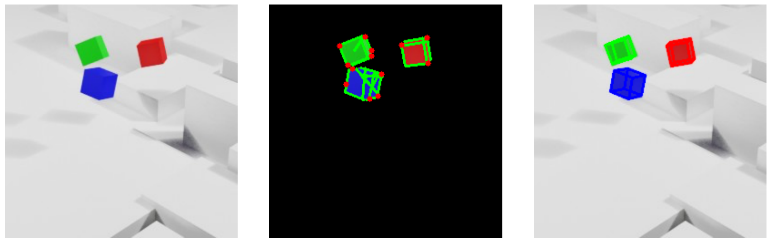

Figure 2: Reconstructing object-level poses

3 Interactive Physics

We extend this benchmark to evaluate models on interactive rigid body physics. While the aforementioned benchmarks cover a vastly more complex set of physical phenomena, we instead choose to focus on "action-following", the FWM equivalent of instruction adherence in LLMs. Re-using the above pipeline, we add an additional "click" action, which causes an impulse force from the point of the click to the center of the cube's mass, with a slight upward bias. We choose this task in particular because of its simplicity - the interaction will naturally "feel right" to a user if performed correctly, and its precision - the model must understand exactly where on the cube the user clicked in order to simulate the correct response. We ensure that all actions are deterministic. We generate a training set of 5000 15-second videos, and a test set of 500 15-second videos. Examples are shown below. A click occurs approximately one in every 10 frames, and of these, approximately 1 in 5 hit a cube and thus have a physical effect.

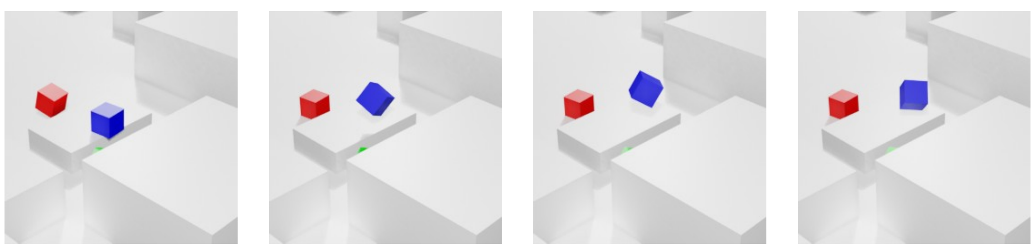

Figure 3: User clicks on the left side of the blue cube

4 Baselines and Results

We develop a simple autoregressive model to use as a baseline. Unfortunately, no pretrained FWM currently available can be readily finetuned on our dataset - most use a form of temporal compression, which renders them unable to provide realtime interactivity out-of-the-box, although we plan to accomodate these in the future. In general, we expect FWMs to become more adaptable and eventually able to zero-shot this task.

For our autoregressive baseline, we modify Llama-3.1 (Grattafiori et al., 2024) to predict sequential video tokens. We use the Cosmos-16x16 image tokenizer, and thus generate 256 tokens per frame. We use adaLN-Zero conditioning for actions, following Peebles & Xie (2023). Currently, after training for 10000 steps with 300M parameters, we score 23.90 on Newton-OP, largely due to MSE. We expect to improve this score significantly with further training and experimentation with spatiotemporal and axial attention schemes.

[3]Specifically, we estimate our known geometry by using corner and edge detection, and then refine the pose using differential evolution.

Table 1: Newton-OP Preliminary Results

Model	Parameters	MSE ↓	LPIPS ↓	Pose ↓	Total ↑
AR	300M	0.03	16.07	60.00	23.9

5 Future Work

We intend to complete evaluation of our autoregressive baseline, and additionally introduce an autoregressive-diffusion baseline in line with Cosmos. In addition, we will compare recent techniques such as Diffusion Forcing (Chen et al., 2024), which promise to greatly improve performance on object permanence. As mentioned above, we plan to evaluate recent and forthcoming FWMs in the same manner.

In addition, we aim to complete the Newton suite by producing a variety of resolutions, aspect ratios and framerates for the current tasks, in addition to more complex tasks focusing on interactivity, such as first-person navigation, free cameras and non-player characters, which are beginning to be explored in the literature.

Finally, we hope to develop Newton as a test-bed for new architectural ideas in FWMs, particularly as it encourages 3D priors and persistent worlds, and to encourage the community to develop new techniques for evaluating and improving these models.

References

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

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. Physion: Evaluating physical prediction from vision in humans and machines, 2022. URL https://arxiv.org/abs/2106.08261.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments, 2024. URL https://arxiv.org/abs/2402.15391.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion, 2024. URL https://arxiv.org/abs/2407.01392.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding, 2025. URL https://arxiv.org/abs/2501.16411.

Quinn McIntyre Spruce Campbell Xinlei Chen Robert Wachen Decart, Julian Quevedo. Oasis: A universe in a transformer. 2024. URL https://oasis-model.github.io/.

DeepMind. Genie 2: A large-scale foundation world model, 2024a. URL https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/.

DeepMind. Veo 2, 2024b. URL https://deepmind.google/technologies/veo/veo-2/.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023. URL https://arxiv.org/abs/2305.07759.

Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control, 2024. URL https://arxiv.org/abs/2412.03568.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana

Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models, 2023. URL https://arxiv.org/abs/2311.17982.

Edward Hughes, Michael Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktaschel. Open-endedness is essential for artificial superhuman intelligence, 2024. URL https://arxiv.org/abs/2406.04268.

NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. URL https://arxiv.org/abs/2501.03575.

OpenAI. Video generation models as world simulators, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning, 2020. URL https://arxiv.org/abs/1803.07616.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation, 2019. URL https://openreview.net/forum?id=rylgEULtdN.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models, 2024. URL https://arxiv.org/abs/2405.15223.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2024. URL https://arxiv.org/abs/2408.06072.

Pengyu Zhao, Ning Cheng, Huiqi Hu, Xue Zhang, Xiuwen Xu, Zijian Jin, Fandong Meng, Jie Zhou, Jinan Xu, and Wenjuan Han. Bridging the reality gap: A benchmark for physical reasoning in general world models with various physical phenomena beyond mechanics, 2024. URL https://openreview.net/forum?id=vsYt8UHGzI.