1X World Model: Evaluating Bits, not Atoms

1X World Model Team *

Abstract

We share this supplementary technical progress report on the 1X World Model released at `https://1x.tech/discover/redwood-ai-world-model`. The 1X World Model (1XWM) is a generative video world model that predicts future robot observations and task-level state values. We train 1XWM to accurately forecast contacts and full-body manipulation given action commands, the first world model for a full-body humanoid. The 1X World Model achieves precise action controllability that enables us to compare different policies under identical observational conditions. We show that the accuracy of 1XWM as an evaluator scales with data from autonomous policy rollouts, and demonstrate high correlation between the world model and real world evaluation. We further show transfer in 1XWM prediction accuracy across different humanoid tasks. With 1XWM powering our evaluation engine, we can quickly iterate on checkpoint selection, architecture comparisons, and long-tail dataset curation.

1 Motivation

In this supplemental progress report, we show 1X World Model (1XWM) exhibits:

1. Precise action controllability that enables us to compare policies that take different decisions given the same observations.
2. Scaling results when leveraging a specific source of data: autonomous policy rollouts.
3. High correlation between the world model and real world evaluations.
4. Multi-task transfer across tasks.

The 1X World Model allows us to:

1. Quickly iterate on architectural decisions.
2. Select the best checkpoint from training runs.
3. Curate datasets of long-tail scenarios in production and re-evaluate models on them.
4. Optimize robot policies at scale through efficient training-evaluation cycles.

We believe that solving the evaluation problem is a major milestone towards deploying real-world robots in general-purpose environments and unstructured homes. For robotics, correlating training-time metrics such as mean-squared error between predicted and expert trajectories is an unreliable proxy to real-world performance.

It takes days to weeks to fully evaluate the performance of a humanoid policy in diverse home environments. For one policy checkpoint, it takes hundreds to thousands of samples to gather statistically significant metrics. Given that one round of training can generate dozens of candidate checkpoints, fully evaluating the models we train at 1X becomes prohibitively expensive. Consequently, we have to carefully select the models that we send for full evaluation in the real world. Reliable offline evaluation would multiply the throughput of model experimentation.

*Contributed by Daniel Ho, Jack Monas, Juntao Ren, Christina Yu.

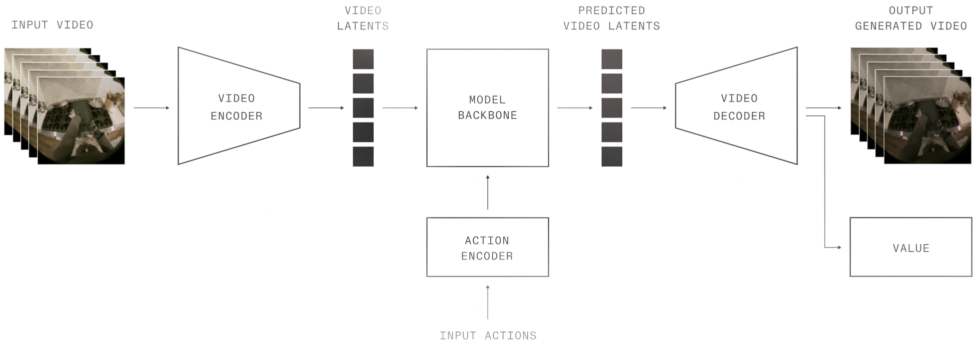


Figure 1: The 1X World Model consists of visual and action encoders, backbone, and decoders for video and state-values. State-value predictions, supervised on success labels, allow us to score the quality of input actions.

Another aspect of evaluation is being able to replay failures and edge-cases we discover in production settings. Consider if a robot fails to brew a cup of coffee using a new brand of coffee maker it has never seen before, that it encounters in a user's house. Can we train a new model and see if it solves this exact situation? With 1XWM, we can generate counterfactual futures and success likelihoods given the actions such a model would take. Furthermore, we can build a dataset of all of our failures in production and evaluate all new policies we train against this production dataset. This gives us more statistical confidence to roll out new policy deployments.

The 1XWM enables us to rollout different policies from *identical initial conditions* for direct comparison, making evaluation more interpretable and giving us more fine-grained "win/loss" comparisons between models. By visually inspecting predicted futures for any particular set of initial conditions, we can also determine specific strengths and failure modes for each policy (e.g. policy A is worse than policy B when the object is oriented in a particular way).

2 Related Work

2.1 Action-Controllable Video Generators

Most current video generators are text-to-video models: Veo, Sora, and Kling (closed source), and Align-Your-Latents, Open-Sora, and HunyuanVideo (open source) create high-resolution clips from text or images ([Google DeepMind, 2024, OpenAI, 2024, Kuaishou Technology, 2025, Blattmann et al., 2023, Kong et al., 2024]). Camera steering is introduced by CameraCtrl and MotionCtrl, which inject camera or object trajectories ([He et al., 2025, Wang et al., 2024]). Game engines like Genie, DIAMOND, and Oasis generate frames after user inputs supporting interactive 2-D or 3-D play ([Bruce et al., 2024, Alonso et al., 2024, Decart et al., 2024]). In driving, Vista, GAIA-2, and Learning to Drive from a World Model generate visual rollouts conditioned on ego-vehicle motion commands ([Gao et al., 2024, Russell et al., 2025, Goff et al., 2025]). Recent work in robotics from Unified World Models and COSMOS accept lower level commands for desktop arm control ([Zhu et al., 2025, NVIDIA, 2025]).

2.2 Policy Evaluation

Progress in robot learning hinges on evaluators that are simultaneously *fast enough* to allow for rapid policy iteration and *faithful enough* to predict deployment performance. Simulators such as CALVIN, ManiSkill, and Isaac Gym enable high-throughput interactions, yet building assets and crafting domain–randomisation schedules are costly, and the sim-to-real gap persists [Mees et al., 2022, Mu et al., 2021, Mittal et al., 2023]. Real-robot evaluations eliminate the sim-to-real gap but remain a bottleneck: even with automated resets and scoring, AUTOEVAL only processes up to $\sim$850 episodes per day, requires periodic motor cool-downs, and is limited to fixed table-top manipulation

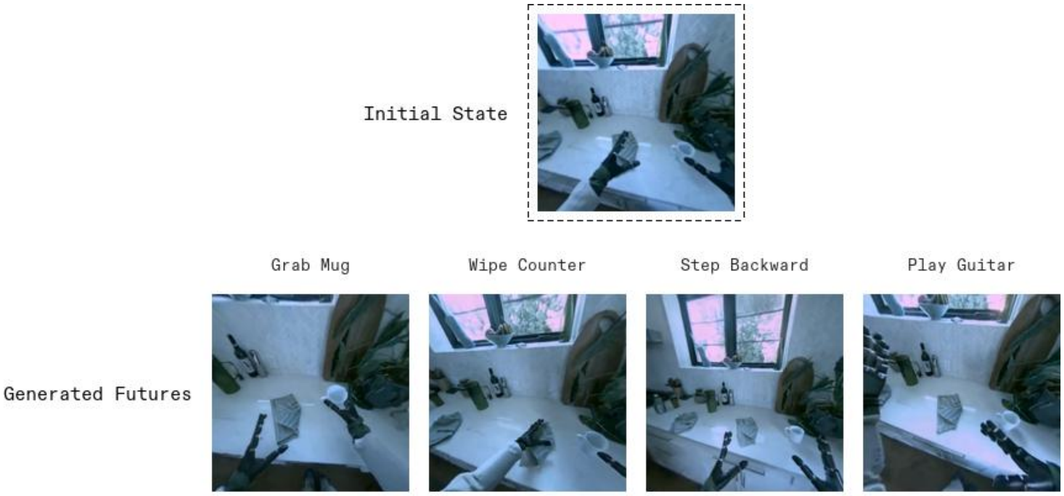


Figure 2: Starting from the same initial state (boxed), 1XWM predicts four distinct futures for Neo when conditioned on **four different low-level action sequences to** grab a mug, wipe the countertop, step backward, and play an imaginary guitar.

scenes [Zhou et al., 2025]. Model-based off-policy methods cut robot time by stitching short k-step roll-outs from a learned dynamics model onto real or logged states, yet their predictions degrade quickly as the imagined horizon grows [Yu et al., 2020, Janner et al., 2019, Hafner et al., 2019]. What remains absent is a middle ground: an *offline* evaluator that runs orders of magnitude faster than physical robots yet still captures contact-rich, full-body dynamics well enough to stand in for real trials.

Our contribution. Unlike prior video-world models that focus on games, desktop arms or driving, the 1X World Model is the first to forecast high-fidelity, contact-rich futures for a full-body humanoid, as well as a learned state value measuring success and progress towards task completion. We show that this evaluator correlates with expensive real-world trials. By providing a fast iteration loop to evaluate candidate models, we can reduce the frequency of needing to evaluate in real.

3 Architecture

1XWM is a generative video model which predicts future states conditioned on image states and input actions. To produce a task-level evaluation score, the model predicts the state value of the generated future states.

We separate clips of videos into a sequence of image frames of four seconds in length. We encode the initial frames into latent space using a temporal image encoder. Separately, we encode input actions and feed both these encodings into our model backbone which predicts latent states representing the full four second clip. Given these latent states, we decode images of the future frames and the state value prediction. By operating in latent space, we are able to co-train on action spaces from multiple embodiments, including both EVE and NEO.

Our pipeline is flexible to video resolution. In practice, we train on both 256×256 and 512×512 resolution data. We pre-train on web-scale video data and post-train on 1X robot data. Our post-training dataset contains teleoperated and autonomous task episodes that we label with binary success values that we interpolate to state value labels Gokmen et al. [2023].

4 Specializing Video Generation Models for Action Controllability

Most video generation models are text-to-video (T2V), using a language prompt to generate the video and, optionally, one or more reference frames for guidance. However, world models for simulating robot policies need to be action-controllable, steered by exact robot trajectories rather than loose

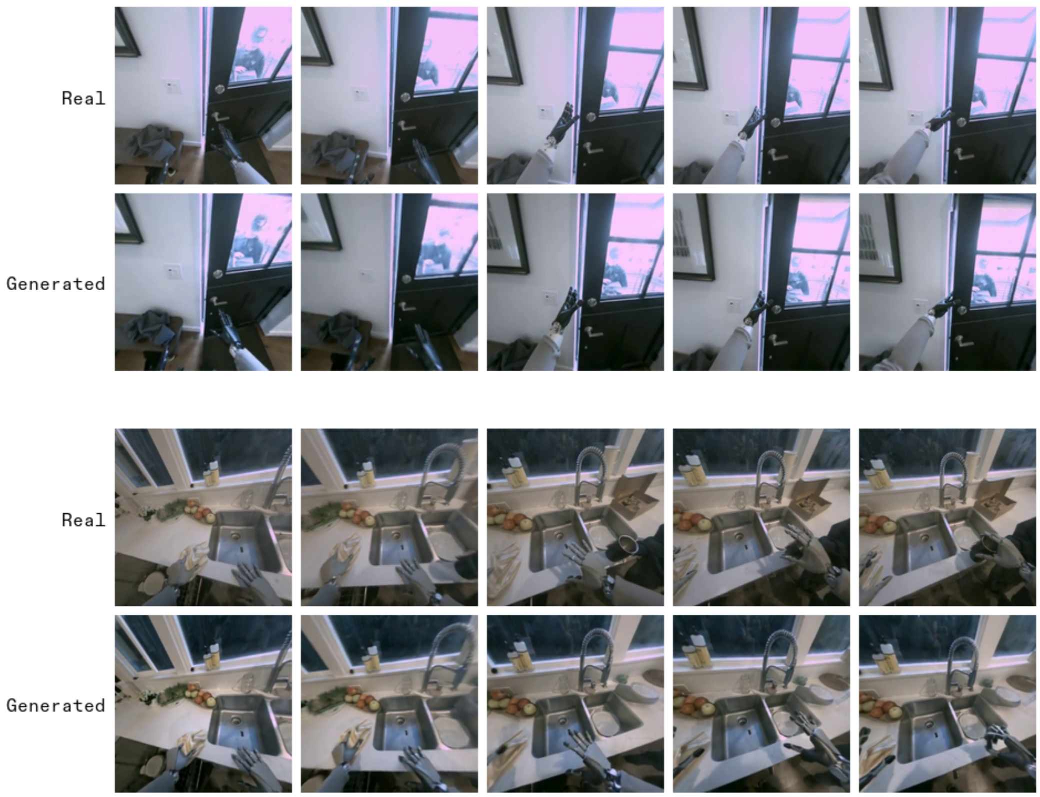


Figure 3: Real versus 1XWM-predicted outcomes when Neo executes an identical sequence of low-level commands. Each stacked strip starts after a brief context clip (omitted for clarity).
Strip 1: The model closely matches the dynamics of Neo opening a door for a guest.
Strip 2: Notice the real scene contains a mug to grasp, but the model, unaware of that object, still moves the gripper along the commanded trajectory and closes on empty space.

directives like "Grab the mug" or "Wipe the countertop." Furthermore, the world model should react to the agent's actions according to the laws of physics.

In Figure 2, 1XWM generates four divergent futures for NEO, all starting from the same initial frame, picking up a mug, wiping the countertop, stepping backward, and playing an imaginary air guitar, each branching into its own realistic outcome.

To showcase the action-controllability and physical alignment of 1XWM, Figure 3 compares generated futures against ground truths. We provide the World Model with a few initial frames of real footage along with the subsequent action trajectories. From this anchor point, the 1XWM simulates the consequences of taking those exact actions, including the physics of objects like doors and cloths being wiped across a countertop.

5 Scaling Results

Scaling laws have recently become a central topic in machine learning, especially in the domain of large language models (LLMs), where clear trends have demonstrated significant performance gains through increased training data capacity [Hoffmann et al., 2022, Muennighoff et al., 2023, Delétang et al., 2024, Isik et al., 2024, Brandfonbrener et al., 2024].

In particular, works such as Isik et al. [2024] measure the importance of specific *sources* of data — in other words, data that most positively correlates with model improvement. Inspired by these findings in LLMs, we explore whether similar scaling relationships hold for robotic world models responsible for accurately predicting and evaluating real-world task outcomes. Does the accuracy of

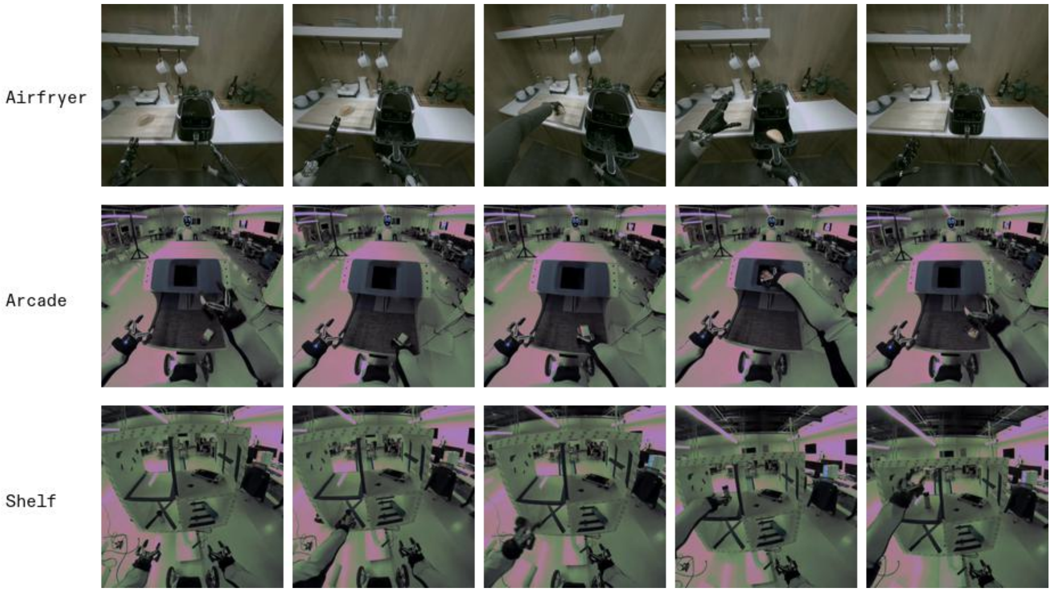
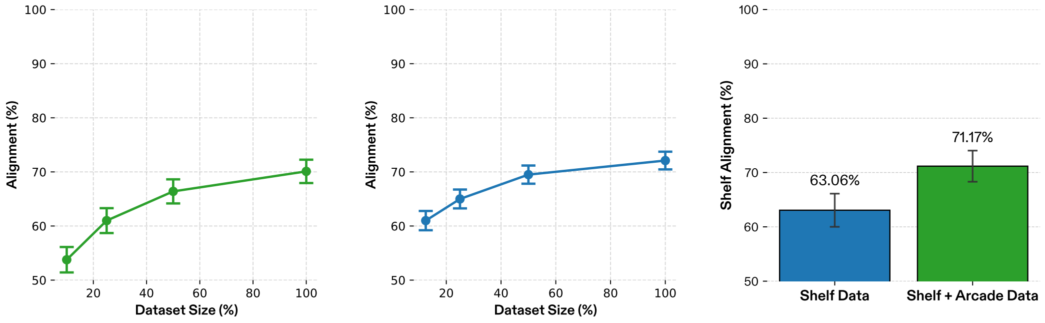


Figure 4: Ground truth sequences for Airfryer, Arcade, and Shelf tasks.



(a) Airfryer alignment

(b) Arcade alignment

(c) Task-transfer improvement

Figure 5: Panels a, b show alignment vs. dataset size. Panel c) shows alignment improvement with cross-task learning.

policy evaluation with 1XWM improve as data is scaled up, and what kinds of data lead to the most improvement?

To study this question, we train 1XWM to not only predict future states and images, but also whether the task attempt succeeded or failed at the end of each generation. We define *alignment* to be how aligned our predictions are with the real world. For a given task, we measure a model's alignment as the accuracy of its success/failure predictions. We see improvements in alignment across the board as we scale up the number of tasks and the diversity of robot behaviors. We explore this in the following tasks: Airfryer, Arcade, and Shelf.

5.1 Experimental Setup

We collect data via teleoperation and autonomous policy rollouts on tasks of interest shown in Figure 4.

1. **Airfryer**: We position NEO in front of an air fryer. The right hand attempts to grasp the air fryer tray handle and pull it out of the base. The left hand then attempts to grasp the object on the cutting board with its left hand, pick it up, and drop it in the airfryer tray. Finally, the

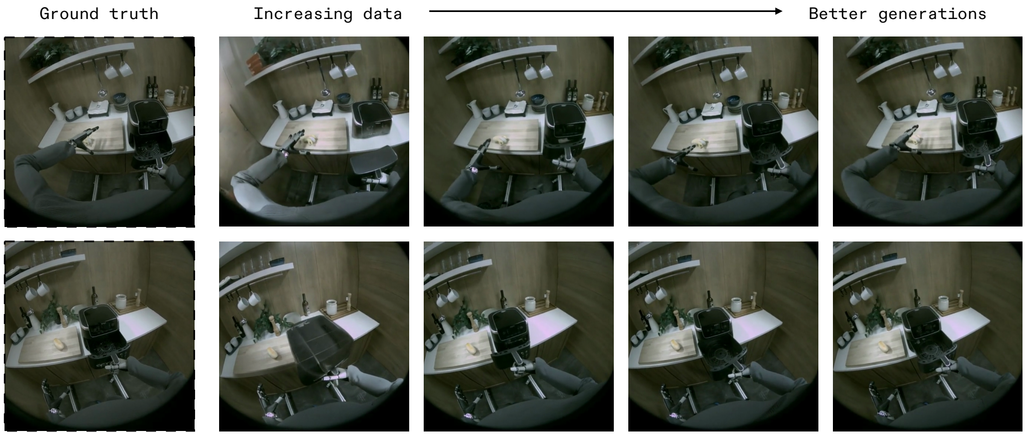


Figure 6: Improvement in Airfryer generations with increasing data.

right hand pushes the tray back into the base. The episode fails if the robot fails at any point in the sequence (e.g. missed handle grasp, missed object grasp, missed drop).

2. **Arcade**: We position EVE in front of an object drop station. The robot attempts to grasp the target object with either left or right gripper. An episode is successful if and only if EVE grasps the object and drops the object into the opening at the top of the station.
3. **Shelf**: We position EVE in front of a station with four shelf quadrants. EVE is language-commanded to pick and place a target object located in one quadrant to another. The object is standing upright in the initial frame and needs to be upright placed in the correct quadrant to be considered successful. This task often involves handing over the object from one gripper.

For each attempt, we assign a value of $+1$ to the final frame if it was a success, and -1 if it was a failure. We use temporal discounting backwards from the final frame to assign value to remaining frames.

As a metric for relative improvement across world models trained on varying amounts and sources of data, we measure the accuracy of the value prediction on a held-out validation set. An accuracy of 50% is no different than random guessing, while an accuracy of 100% suggests a perfect understanding of the world being modeled.

5.2 Quantitative and Qualitative Analysis

Our analysis compares World Model predictions with ground truth observed results, captured across various task checkpoints. We establish a clear correlation between predicted and real-world performance.

We start with the 1XWM base model pretrained on web-scale video data and investigate the effects of adding additional data on Airfryer, a NEO task, and Arcade, an EVE task. From this collection, we randomly subsample datasets ranging from 10% to 100% of the original dataset, where each increase in data is a superset of the previous dataset, and each dataset is balanced in terms of success and failure. We then fine-tune the 1XWM on these datasets, training the model to predict both future states and the corresponding state value.

In Figure 5 (a, b) we show a clear improvement in the accuracy of 1XWM predictions as we scale up the amount of Airfryer and Arcade data.

When confronted with an unfamiliar task and environment, the WM often struggles to model the object interactions exactly, without having knowledge of their specific properties. Training on task-specific data allows the WM to update based on the dynamics of the task at hand.

For example, when trained on smaller amounts of data, 1XWM hallucinates the air fryer tray and body as a single unit, pulling the entire unit off the counter (Figure 6). After adding interaction data with the air fryer, 1XWM gains a better understanding of how the tray separates from the air fryer, and even learns to model subtle interactions such as the confinement of the tray movement within the base of the air fryer.

5.3 Multi-Task Transfer

We find 1XWM improves with the accumulation of more tasks. We investigate whether a modest quantity of data from a new task, Shelf, can be leveraged effectively when paired with a preexisting large corpus collected on the Arcade task. A dataset of $\sim$216 M Shelf video tokens attains an alignment score of 63.06 % on a held-out test. When augmented with an Arcade corpus of $\sim$1.46 B video tokens, the 1XWM trained on the combined datasets achieves an alignment score of 71.17 %. The transfer gain, shown in Figure 5 (c), indicates reuse of latent structure (e.g., grasp kinematics, collision priors, task understanding) learned from Arcade. These results reinforce our belief in the scaling capacity of the 1XWM, as we expect generalization to come from a rich repository of task experience.

5.4 Importance of Data Source

We examine various sources of training data, including robot teleoperation, robot self-supervised exploration or play data, and observational data derived from human videos. Among these diverse sources, we find that *on-policy rollouts from robot policies* are most critical to improving model alignment. In particular, we observed that *failure data* from these autonomous robot rollouts is critical.

Without failure examples, 1XWM generations show optimistic bias towards success, such as having objects reposition for easier manipulation, incorrect estimations of grasp radius, or insufficient modeling of obstructions. Not only does autonomous data collection increase our data throughput compared to teleoperation, allowing us to collect more interesting and diverse failure modes at scale, but autonomous rollouts are also more in-distribution for policy evaluation. Empirically, we found that intentionally teleoperated failures tended to have obvious biases which did not help the model to identify challenging failures in policy rollouts.

6 The Evaluation Problem

An aligned world model can solve the evaluation problem by forecasting the actions of candidate robot policies. Given 1X World Model generations from each policy on datasets of initial states, we can compare their respective performances. Importantly, we can collect datasets of production-setting states and generate counterfactual results from states in which an autonomous policy has previously failed.

For every set of model checkpoint weights, we predict future states and success likelihood, which we show are distributionally aligned to actual real-world futures. This gives us insight into model performance at scale and allows us to make model architecture and checkpoint selection decisions with an instant feedback loop.

Deploying learned robot models in customer settings requires high confidence that policies will be safe and high-performing, specifically in customer settings. Using the 1X World Model, we "snapshot" and then evaluate directly on critical 'test-time' task conditions.

6.1 Evaluation Setup

On the Arcade task, we run the model continuously for ten minutes and reset the cube if the robot fails three grasps in a row or ends up in an unrecoverable state. We count the successes, total number of grasp attempts, and resets needed. We compute a score of: $(\texttt{successes} - 3 * \texttt{resets} + 0.3 * \texttt{attempts})/\texttt{attempts}$ to penalize interventions while rewarding more frequent attempts.

We run all real-world policy evaluations of a particular experiment as a double-blind A/B experiment conducted on one robot run in one setting to minimize confounding variables. Operators execute autonomy with a random model being selected to run, and results are aggregated afterwards.

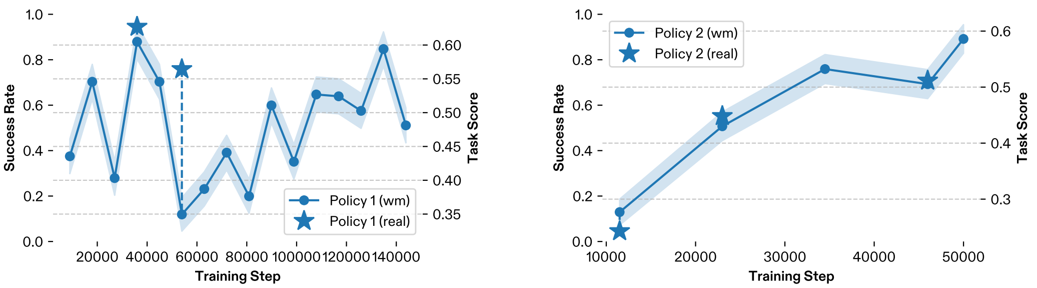
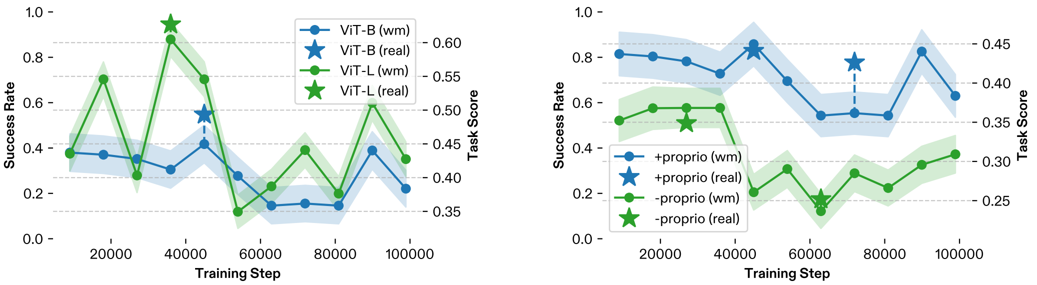


Figure 7: We train two policy architectures (Policy 1 and 2) and use 1XWM to sweep checkpoints on Arcade for checkpoint selection. In practice, we find that a checkpoint that scores significantly higher on the world model evaluation is also likely to score higher in real eval, although the margins are not necessarily similiar.



(a) Comparison of ViT-B vs. ViT-L as an image encoder for a particular policy architecture.

(b) Comparison of including or withholding robot proprioception information as policy input.

Figure 8: We use 1XWM eval scores on Arcade to compare between different policy architectures. Compared to checkpoint selection, we get more signal in seeing trends across different checkpoints.

6.2 Evaluation Experiments

6.2.1 Checkpoint Selection

Checkpoint selection is a direct way to think about evaluation. Given a training run, we can run 1XWM prediction to generate a score for each checkpoint on a task dataset. We can then select the best score as a promising checkpoint to deploy. In Figure 7, we show examples of this in practice for two policy training runs.

Given a true real-world success rate gap of 15% between two checkpoints, a World Model with 70% alignment can accurately predict the better policy with 90% success.

6.2.2 Architecture Selection

Unlike in checkpoint selection, architecture comparisons allow us to leverage more data points to base a decision on: we can compare across all checkpoints in each training one. See Figure 8. In the plot below, we ablate the decision to include proprioception (robot joint states) as input to our robot policy, and plot the 1XWM predicted success rate for each checkpoint. We then run real-world evaluation on the most checkpoints. We find that there is indeed a correlation between the predicted success rates and the true task scores. This allows us to make a likely forecast that proprioception improves policy performance.

As another experiment, we compare using two different image encoders, ViT-B and ViT-L Dosovitskiy et al. [2020], for a policy. We compare the most promising predicted checkpoints from both policies and see that the predicted better ViT-L model indeed performs better in the real world.

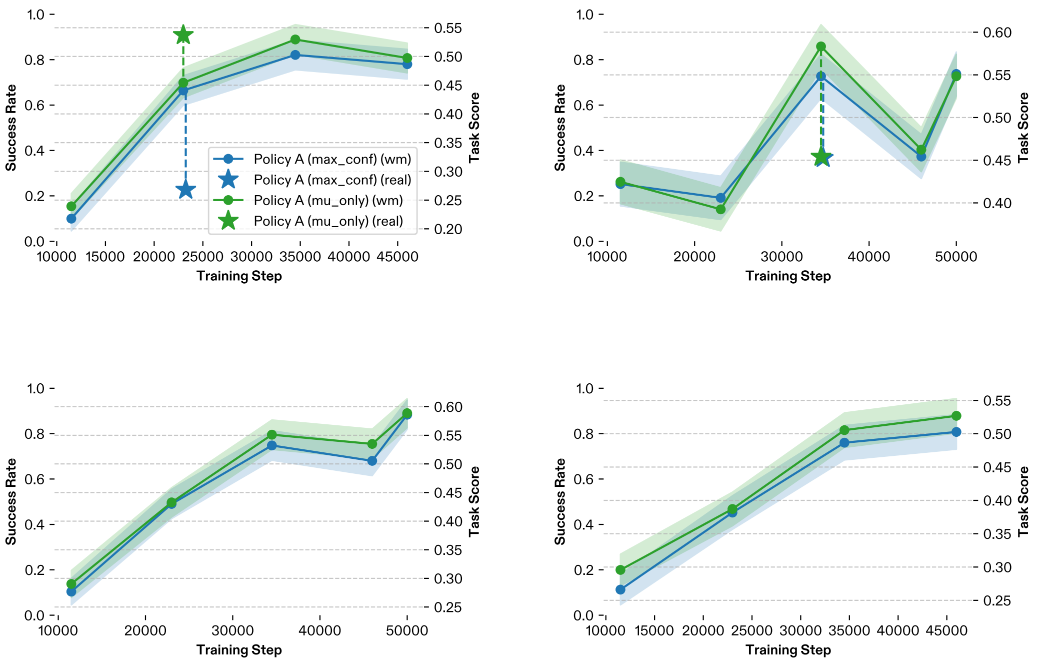


Figure 9: On four different base architectures A-D, we sweep 1XWM evaluation across checkpoints on the Arcade task. We compare two inference-time sampling strategies (green, mu-only vs blue max-conf). We notice a small difference across all experiments suggesting that mu-only sampling is better, and we find this generally holds in real.

We could also compare architecture changes on several different variations of a model, by varying other independent components or simply running the comparison multiple times.

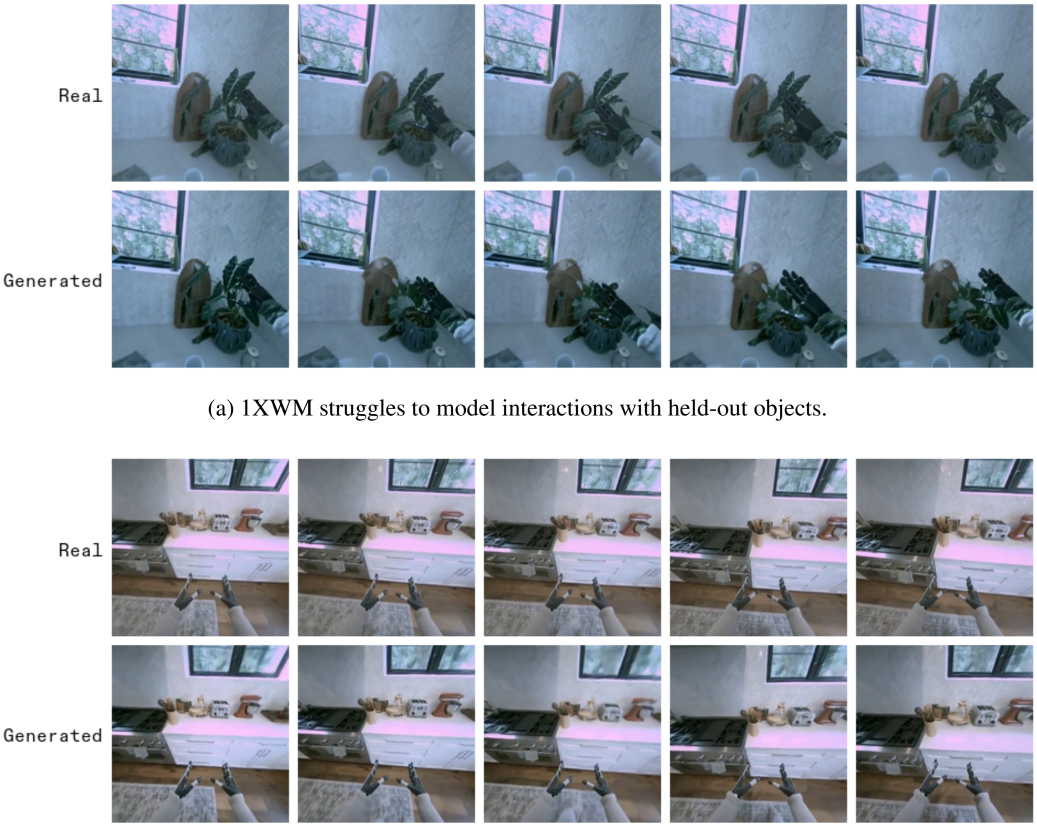
Consider an experiment where we compare the inference-time sampling strategy of our model architecture, shown in Figure 9. While there is a narrow performance gap on any individual checkpoint and run, the consistent gap across all experiments gives us more confidence that mu-only sampling performs better. We do see this hold in practice.

7 Limitations

As shown previously in Figure 6, the 1XWM currently struggles to model interactions with held-out objects not seen in training data. An example is shown in Figure 10 (a).

1XWM exhibits strong frame-level action controllability, such as capturing subtle shaking movements in the neck when walking. However, there is still a small amount of error in the exact predicted pose and location of the robot. This has the biggest impact on predictions involving locomotion, where errors in lower body position accumulate with each step, such as in Figure 10 (b). This limits the capability of our current WM for long-horizon navigation.

As we deploy robots in home, we will need to move away from task-specific evaluation and towards production-level evaluation, capable of handling a wider, more ambiguous array of full-body manipulation tasks and objects. Improving the generalization capability and accuracy of 1XWM will be the first step towards this goal.



(a) 1XWM struggles to model interactions with held-out objects.

(b) Errors in lower-body position accumulate for locomotion.

Figure 10: Real vs. 1XWM-predicted outcomes showing limitations.

References

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In *Thirty-eighth Conference on Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.12399.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

David Brandfonbrener, Nikhil Anand, Nikhil Vyas, Eran Malach, and Sham Kakade. Loss-to-loss prediction: Scaling laws for all datasets. *arXiv preprint arXiv:2411.12925*, 2024. URL https://arxiv.org/abs/2411.12925.

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. *arXiv preprint arXiv:2402.15391*, 2024. URL https://arxiv.org/abs/2402.15391.

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

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. Language modeling is compression. In *ICLR*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024.

Mitchell Goff, Greg Hogan, George Hotz, Armand du Parc Locmaria, Kacper Raczy, Harald Schäfer, Adeeb Shihadeh, Weixing Zhang, and Yassine Yousfi. Learning to drive from a world model. *arXiv preprint arXiv:2504.19077*, 2025.

Cem Gokmen, Daniel Ho, and Mohi Khansari. Asking for help: Failure prediction in behavioral cloning through value approximation, 2023. URL https://arxiv.org/abs/2302.04334.

Google DeepMind. Veo: Next-generation text-to-video model. https://deepmind.google/models/veo/, 2024.

Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *CoRR*, abs/1912.01603, 2019. URL http://arxiv.org/abs/1912.01603.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation, 2025. URL https://arxiv.org/abs/2404.02101.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. Scaling laws for downstream task performance of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Kuaishou Technology. Kling ai 2.0: High-fidelity text-to-video generation. `https://www.scmp.com/tech/big-tech/article/3306631/`, 2025.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks, 2022. URL `https://arxiv.org/abs/2112.03227`.

Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 8(6):3740–3747, June 2023. ISSN 2377-3774. doi: 10.1109/lra.2023.3270034. URL `http://dx.doi.org/10.1109/LRA.2023.3270034`.

Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations, 2021. URL `https://arxiv.org/abs/2107.14483`.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models, 2023. URL `https://arxiv.org/abs/2305.16264`.

NVIDIA. Cosmos world foundation model platform for physical ai, 2025. URL `https://arxiv.org/abs/2501.03575`.

OpenAI. Sora: Creating video from text. `https://openai.com/index/sora/`, 2024.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving, 2025. URL `https://arxiv.org/abs/2503.20523`.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation, 2024. URL `https://arxiv.org/abs/2312.03641`.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization, 2020. URL `https://arxiv.org/abs/2005.13239`.

Zhiyuan Zhou, Pranav Atreya, You Liang Tan, Karl Pertsch, and Sergey Levine. Autoeval: Autonomous evaluation of generalist robot manipulation policies in the real world, 2025. URL `https://arxiv.org/abs/2503.24278`.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets, 2025. URL `https://arxiv.org/abs/2504.02792`.