GENERALIZABLE WORLD SIMULATOR: REAL-TIME & LONG-HORIZON INTERACTION WITH DIVERSE OBJECTS

Anonymous authors
Paper under double-blind review

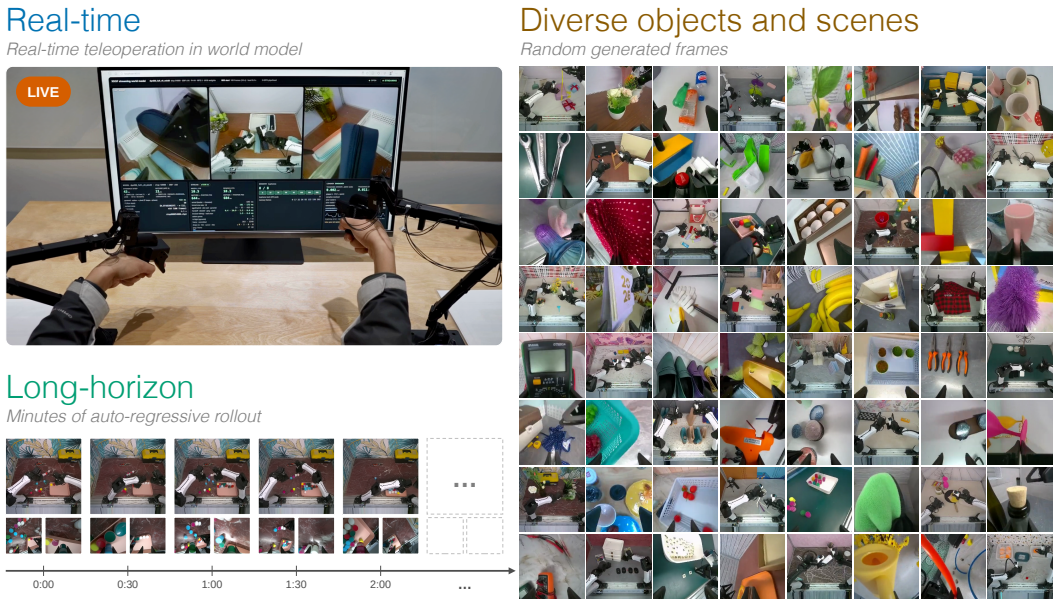

Figure 1: **GWS simulates minutes of bimanual manipulation at interactive rates.** *Left*: live teleoperation of the simulator with two GELLO leader arms at 10 Hz. *Bottom left*: a two-minute rollout after a 15-s seed of real frames, every later frame predicted from the model's own past. *Right*: generated frames from rollouts on unseen scenes.

ABSTRACT

Interactive world models have shown great potential in serving as a simulator for robot manipulation. However, existing models are either overfit to a single environment or lack ability to simulate long-horizon contact-rich events. We present Generalizable World Simulator (GWS), a world model whose dynamics model is trained from scratch on a large robot-native teleoperation dataset, with no internet-video pretraining. By focusing on developing the optimal model architecture and learning recipe to fit robot-native data, we managed to learn a world model that is orders of magnitude smaller than SOTA video generation models, yet shows significantly better understanding of contact and causality, and runs 10 FPS on an commercial GPU. Experimentally, our model significantly outperforms previous SOTA methods across various tasks including future prediction, policy evaluation, and mixed-data co-training. We further demonstrated various potential applications where our model can be used in a robotics setting. Project page: `gws-anonymous.github.io`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

A world model, in its simplest form, can be formulated as a model to predict the future conditioned on an action input. It is also by far, the most powerful abstraction for control, allowing robots to plan many steps into the future, simulate the outcomes of their behavior and like humans (Flanagan & Wing, 1997), allow them to search for the optimal actions in order to achieve their goals. *Given these strengths, why are world models not the de facto standard tool for solving contact-rich manipulation tasks?*

One of the major shortcomings of today's models is their ability to do *long-horizon prediction* (Gao et al., 2026). Rolling out future predictions for long time horizons, even for tens of seconds, often leads catastrophic error accumulation. And even when models do achieve long-range prediction (Wang et al., 2026), they often struggle with *generalization*, as they are unable to properly model the diverse range of objects and scenes that robots commonly interact with.

To address these challenges, we propose Generalizable World Simulator (GWS), a world model designed to enable real-time, long-horizon, and contact-rich interaction with many objects and scenes, enabled by the following contributions:

- **Long–short-term memory context.** We introduce a context design that combines a short-term window of recent frames with a set of long-term memory keyframes. The short-term context captures ongoing motion, while the long-term memory retains scene information from earlier interactions, allowing the model to condition on both recent dynamics and distant history without continually expanding its attention context.
- **Contact-based memory selection.** Similar to haptic memory in humans (Hutmacher & Kuhbandner, 2018), we introduce a memory mechanism that selects *contact-making or breaking moments* as long-term memory keyframes to be added as context to the world model. Contact event prediction of generated future frames is also added as an auxilliary output of the model. Such contact-aware keyframe selection directs model's attention to moments when robot–object interactions change the scene structure and the auxiliary contact prediction allows the model to update its memory during rollout without external contact annotations.
- **Streaming dynamics with a persistent attention sink.** We develop a causal transformer that combines hybrid spatial–temporal attention with pipelined denoising over a rolling KV cache. A learned sink frame, never noised or evicted, provides a persistent attention target even when memory slots are empty or dropped. Together, these designs support continuous action-conditioned generation with constant computation per step throughout extended rollouts.

Across diverse bimanual manipulation tasks, we demonstrate that GWS generalizes to unseen tasks, sustains up to **ten minutes of open-loop rollouts** under recorded actions, and **supports live teleoperation at 10 Hz**. On 1,000 unseen scenes held-out from training, our model achieves significantly higher average rollout PSNR and slower visual degradation over one minute than state-of-the-art approaches when all models are trained on the same robot dataset (Fig. 3). We also evaluate GWS by the correlation between policy rollouts inside our model and the real world: across two policy checkpoints and three tasks, simulated and real-robot mean task progress have a Pearson correlation of $r = 0.71$ (Fig. 6). Finally, we demonstrate that data collected inside the simulator can improve real-robot performance. On a chips-in-basket task, co-training a policy on mixed real-world and simulator demonstrations increases the task progress by two-fold compared to a policy trained on real-world demonstrations alone.

2 RELATED WORK

Video-generator world models. Most recent robot world models pretrain a large video generator on internet-scale videos, followed by post-training on robot specific data (Agarwal et al., 2025; Ali et al., 2025; Guo et al., 2025; Ye et al., 2026; Gao et al., 2026). Their size follows internet video (2B to 14B parameters (Agarwal et al., 2025; Ali et al., 2025)), and that prior's physical grounding is in question: scale alone does not teach physical law (Kang et al., 2024; Motamed et al., 2025), and chained clips drift past the training horizon (Song et al., 2025; Zhang et al., 2025b). GWS takes the opposite route: its dynamics are trained causally from scratch on robot-native data, need no conversion and stream frame by frame; We ablate the effects of internet video pretraining in Sec. 4.3.

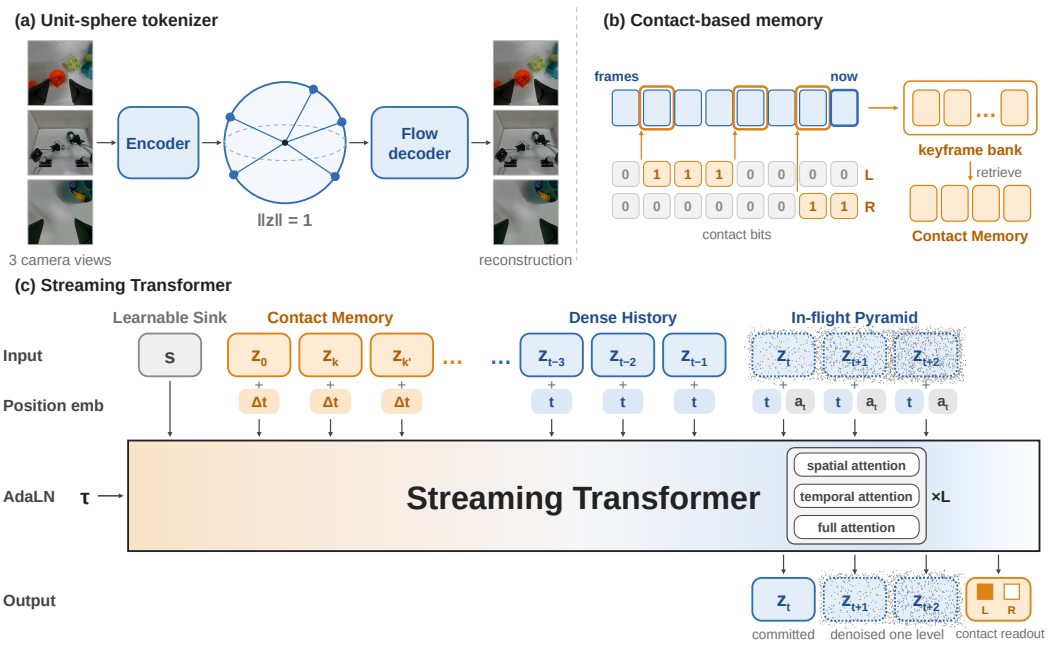

Figure 2: **Overview of *Generalized World Simulator*:** (a) Each view is encoded as per-position latents projected onto the unit sphere and decoded by a flow decoder. (b) When the robot comes into (or out of) contact with an object, we insert the corresponding keyframe into *contact memory*, which are index by end effector pose. (c) We use a transformer that has a number of design decisions that make it well suited to streaming and long-horizon generation.

Causal world models on robot or agent data. A parallel line of works train on robot-specific data from the start. Pioneered by Ha & Schmidhuber (2018), World Models jointly learns action-conditioned future observation prediction by factorizing the problem into representation learning (mapping high dimensional observations into latent state vectors), dynamics prediction (future state prediction conditioned on current state and actions) and future image generation (mapping a low dimensional future state into a high dimensional observation). Follow up works like DREAMER (Hafner et al., 2019) and many others use this framework for planning, RL and control (Hafner et al., 2025; Alonso et al., 2024; Bruce et al., 2024). Model-based RL (Silver et al., 2017; Deisenroth & Rasmussen, 2011; Levine & Koltun, 2013) works predate this recent interest in world modeling. On robots, action-conditioned video prediction from robot-only data dates to Finn et al. (2016) and continues with autoregressive, diffusion and feature-space predictors (Wu et al., 2024a; Zhu et al., 2024; Zhou et al., 2024; Li et al., 2026). Closest to ours are WEAVER (Jain et al., 2026), a causal world model trained on robot-native manipulation data, and the Interactive World Simulator (IWS) (Wang et al., 2026), which adds real-time, long-horizon interactive generation. GWS keeps their recipe and adds a unit-sphere latent, a learned sink frame and a contact-based memory.

Long-horizon and memory mechanisms. Two fundamental mechanisms allow modern video generators to model worlds; first, *generation* which controls how errors compound, e.g., per-frame noise levels, KV-cached or rolling-window denoising, training on the model's own rollouts (Chen et al., 2024; Yin et al., 2025; Huang et al., 2026; Liu et al., 2026a), and a frame-level attention sink on the initial generated frame(s) (Yang et al., 2026; Liu et al., 2026a). Second, *conditioning* which decides what to remember: compressed history, frame banks retrieved by camera pose or field-of-view overlap, or explicit 3-D maps (Zhang et al., 2025b; Xiao et al., 2025; Yu et al., 2025a; Wu et al., 2025); on robots, Ctrl-World and WEAVER keep pose-conditioned time-strided history frames (Guo et al., 2025; Jain et al., 2026). GWS adapts both to manipulation: the per-frame noise levels become a streaming pyramid of in-flight frames served from a rolling KV cache, the sink is one learned, never-noised, never-evicted frame, and memory keeps pose-similarity retrieval but changes what is eligible: keyframes are the contact transitions the model itself predicts (Sec. 3.2).

Policy evaluation and data generation. Physics simulators with real-to-sim assets rank policies in agreement with real trials (Li et al., 2024; Zhang et al., 2025a; Badithela et al., 2025; Jain et al.,

2025), but each scene is built by hand. Learned world models skip the assets: policies are rolled out and scored inside the model (Quevedo et al., 2025; Li et al., 2025; Tseng et al., 2026) or trained on generated data (Jang et al., 2025), and IWS reports real-to-simulated correlation of policy outcomes (Wang et al., 2026). GWS targets the same two uses as a streaming simulator in which a policy, or a human through leader arms, acts closed-loop at 10 Hz; Sec. 4.4 defines the protocols, whose results are pending.

3 METHOD

Let x_t and a_t denote the visual observation and robot action at time t, and $h_t = (x_{0:t}, a_{0:t-1})$ their history. We learn an action-conditioned world model that generates K future observations autoregressively:

$$p_\theta(x_{t+1:t+K} \mid h_t, a_{t:t+K-1}) = \prod_{k=t}^{t+K-1} p_\theta(x_{k+1} \mid h_k, a_k). \tag{1}$$

During rollout, h_k incorporates the model's own predictions, enabling long-horizon simulation driven by recorded actions, a policy, or a human operator.

To achieve this goal, we propose GWS, a latent flow-matching model that combines a visual tokenizer to encode observations into a compact latent space with a dynamics transformer to predict action-conditioned state transitions. We describe how unit-sphere normalization constrains latent scale (Sec. 3.1), how predicted contact transitions and pose-based retrieval select long-term memories to complement recent context (Sec. 3.2), and how hybrid attention, pipelined denoising over a rolling KV cache, and a persistent attention sink enable streaming prediction (Sec. 3.3).

3.1 UNIT-SPHERE TOKENIZER

Our tokenizer uses an SD3-based image encoder (Esser et al., 2024) to map each observation x_t to a compact latent z_t for dynamics modeling. We apply unit-sphere normalization (Davidson et al., 2018; Loshchilov et al., 2024) to the encoded tokens and re-normalize predicted tokens before decoding and reuse as context to limit scale drift.

A flow-matching decoder v_ϕ reconstructs each camera view $x_t^{(v)}$ conditioned on its latent $z_t^{(v)}$. For Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ and flow time $\tau \in [0, 1]$, its objective is (Lipman et al., 2022; Liu et al., 2022)

$$x_{t,\tau}^{(v)} = (1-\tau)x_t^{(v)} + \tau\epsilon, \mathcal{L}_{\text{FM}}^{\text{dec}} = \mathbb{E}_{x_t,v,\tau,\epsilon}\left[\left\| v_\phi(x_{t,\tau}^{(v)}, \tau, z_t^{(v)}) - (\epsilon - x_t^{(v)}) \right\|_2^2\right]. \tag{2}$$

Here t indexes observation time and τ indexes flow time; decoding integrates from noise at $\tau = 1$ to an image at $\tau = 0$. We jointly fine-tune the encoder from pretrained encoder (**?**), and train the decoder from scratch, then freeze them for dynamics training. Architecture and auxiliary training losses are detailed in App. A.

3.2 CONTACT-BASED MEMORY

A short context window can forget earlier interactions, such as where an object was placed after it leaves view. We therefore represent the visual history in h_t using both a window of T recent latents $z_{t-T+1:t}$ and a long-term memory set $\mathcal{M}_t$. The recent window captures ongoing motion, while memory retains selected earlier observations without expanding the transformer's context with rollout length.

We select memory keyframes based on predicted contact transitions, which mark the onset or end of robot–object interactions. An auxiliary head on the dynamics transformer predicts two contact bits $\hat{c}_t \in \{0, 1\}^2$, one per arm; a change in either predicted contact adds the corresponding frame to the keyframe bank. The head is trained with binary cross-entropy against labels from a fine-tuned vision-language model (Clark et al., 2026), with gradients detached from the dynamics backbone. During rollout, the model updates the bank using its own contact predictions, requiring no external contact labels.

To construct $\mathcal{M}_t$, we retain an initial anchor frame and retrieve keyframes preceding the recent window, ranked by end-effector-pose similarity to that window and recency. Poses are computed

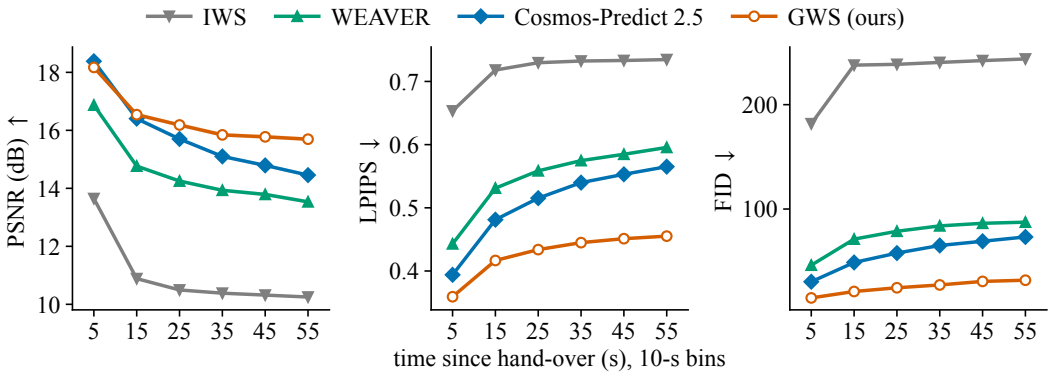

Figure 3: **Drift on ABC over one minute.** PSNR, LPIPS and FID in 10-s bins over 1,000 held-out episodes (15-s real seed, 60-s rollout under the recorded actions, three views). GWS degrades the slowest on all three metrics; Cosmos-Predict 2.5 has the highest PSNR in the first bin and falls below GWS afterwards.

from predicted joint states via forward kinematics. Retrieved frames carry relative time offsets, and unused slots are masked, giving a fixed memory budget of eight frames. Labeling, retrieval, and training details are provided in App. A.

3.3 Streaming Dynamics Transformer

Our causal dynamics transformer predicts z_{t+1} conditioned on action a_t, recent latents $z_{t-T+1:t}$, and memory $\mathcal{M}_t$, while jointly predicting the robot's joint state, as well as contact. Hybrid attention combines spatial attention within frames and causal temporal attention across frames, with selected layers using joint spatio-temporal attention to capture motion across spatial positions. A learned sink frame Xiao et al. (2024), never noised or evicted, provides a persistent attention target even when memory slots are empty. We train with latent flow matching and independently sampled noise levels per frame, following diffusion forcing (Chen et al., 2024).

For streaming inference, we maintain P frames at staggered flow times and denoise them in parallel using a rolling KV cache (Liu et al., 2026b; Jain et al., 2026). Each model call advances all in-flight frames, commits the oldest to the recent context, evicts the oldest context frame, and admits a new noise frame conditioned on the next action. Caching the committed context, memory, and sink keeps the dynamics computation per step constant in rollout length, while pipelining introduces a delay of P steps between an action and its rendered frame.

4 Experiments

4.1 Experimental Setup

Dataset. GWS is trained on an internal bimanual teleoperation dataset collected on YAM robots: two 6-DoF arms with parallel grippers, a 14-D state and action space, and three camera views. We process all episodes to 10 Hz and crop each of the three views to a square that is resized to 256×256. Actions are the 14-D joint targets at the same rate. For the baseline comparison, we instead train every model on the public ABC dataset () for reproducibility: about 131k episodes over 201 tasks (roughly 3,500 hours). We hold out a fixed set of 1,000 ABC episodes for evaluation.

Implementation details. Our tokenizer follows the architecture of SSDD-M (), with a 46 M-parameter convolutional encoder and a 50 M-parameter decoder (); each 256×256 view becomes a 16×16 grid of 16-dimensional unit-norm latents. Both are initialized from pretrained SSDD checkpoint and fine-tuned jointly on our dataset. The dynamics model of GWS is a 1.9 B-parameter transformer with 28 blocks of hidden size 2048 and 16 attention heads; every block applies spatial attention within a frame followed by causal temporal attention across frames, and five blocks use full spatio-temporal attention. Each frame is represented by 193 tokens. The model sees a 5-s window of 50 frames at 10 Hz together with a fixed budget of eight memory slots and one learned sink frame.

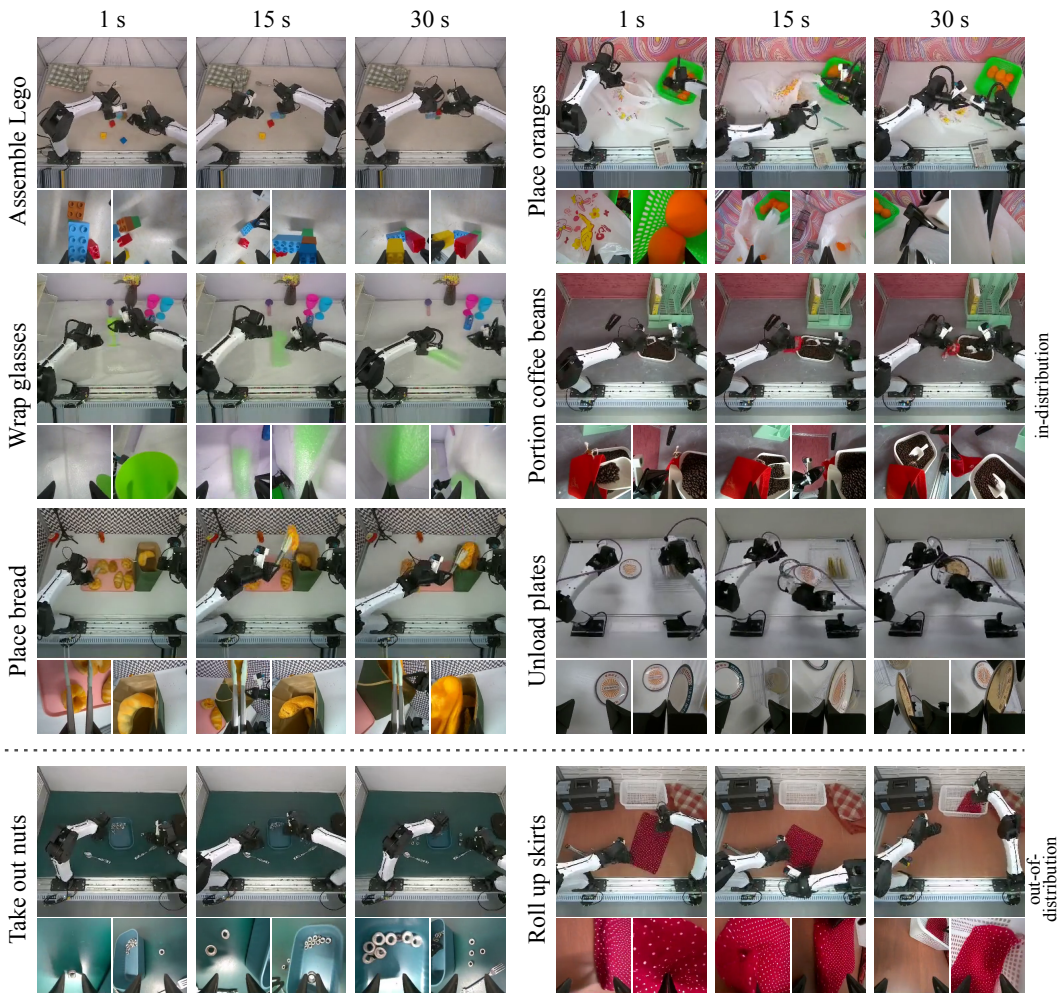

Figure 4: **Rollouts on diverse interactions.** GWS keeps objects and scene stable across compositional assembly (Lego), tool and container use (bags, bread, plates) and deformable objects (skirt).

Training is conducted on 256 H200 GPUs for six days with AdamW, a learning rate of 10^{-3} with 1,000 warm-up steps and no decay, a batch size of 2,048, and independently sampled logit-normal flow times per frame (Esser et al., 2024; Chen et al., 2024). Inference uses $P = 16$ frames in flight, a context-noise level of 0.24 on committed frames.

4.2 Qualitative Results

As is shown in Fig. 4, GWS produces stable rollouts across interaction types: compositional assembly of Lego bricks, tool use, and deformable objects. Objects keep their identity and position over the 30 s, the arms follow the recorded actions, and contact events such as grasps and placements are rendered consistently across the top and wrist views. The rollouts remain coherent in the OOD episodes, indicating that the learned dynamics transfer to unseen objects and task compositions.

4.3 Baseline Comparison

We compare GWS with three baselines: IWS (Wang et al., 2026), WEAVER (Jain et al., 2026), and Cosmos-Predict 2.5 (Ali et al., 2025). For a fair comparison, we implement WEAVER and GWS at a matched scale of about 0.8 B parameters. We size IWS to 355 M to match the streaming step time. All models are trained on the same ABC split and evaluated on the 1,000 held-out episodes: each model is seeded with 15 s of real frames and then rolled out autoregressively for 60 s. We report PSNR, LPIPS, and FID in 10-s bins, averaged over episodes.

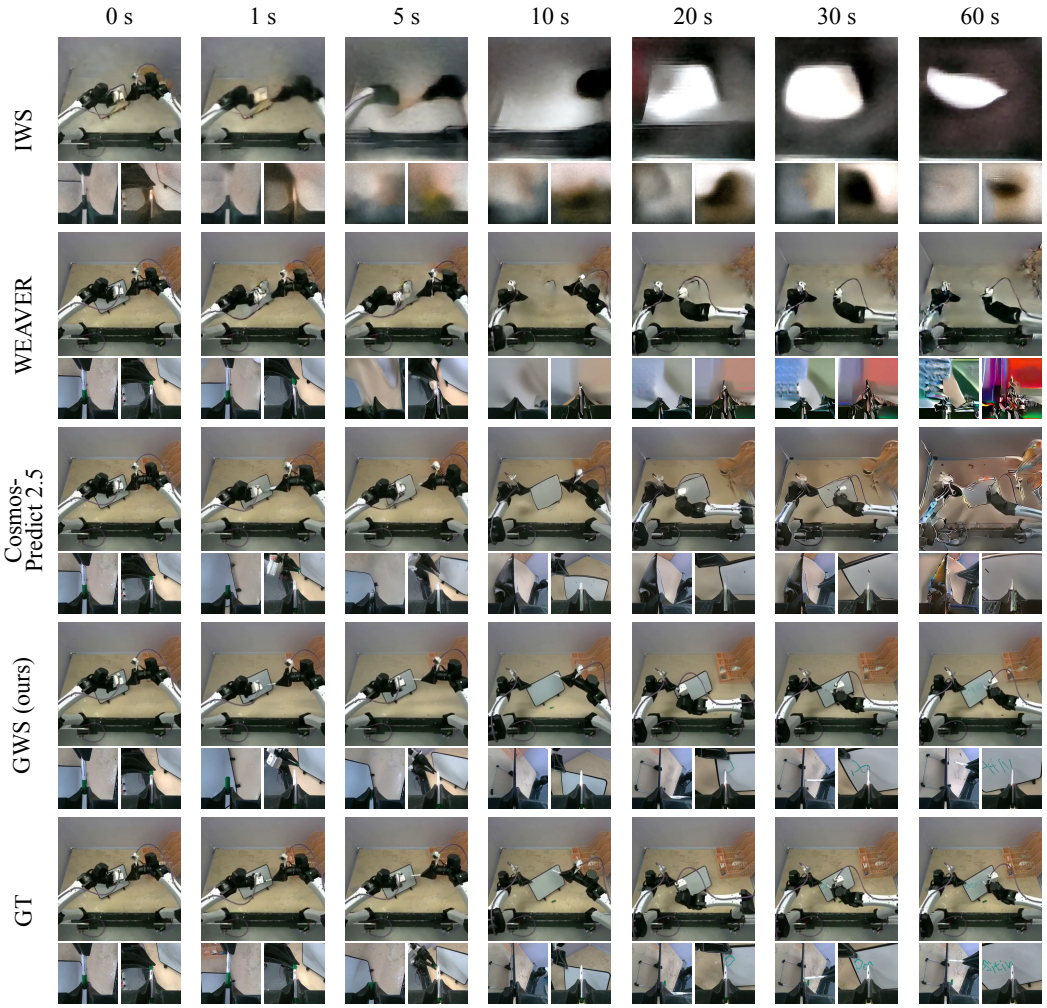

Figure 5: **Visual comparison against the baselines on ABC.** GWS keeps the scene layout, the arm pose and the manipulated object over the whole minute; Cosmos-Predict 2.5 drifts in arm geometry and texture, WEAVER hallucinates texture in the wrist views, and IWS collapses within seconds.

Fig. 5 shows one held-out episode over one minute. GWS produces plausible frames throughout: the table layout, the arm pose and the manipulated object remain consistent with the real frames after 60 s of rollout. Cosmos-Predict 2.5 progressively changes the arm geometry and the scene texture; WEAVER keeps the top view roughly coherent while the wrist views fill with hallucinated texture; and IWS degrades into a blurry mean image within the first 10 s.

Fig. 3 confirms that GWS drifts much more slowly than every baseline on all three metrics. Over the 60-s rollout, GWS averages 16.4 dB PSNR against 15.8 for Cosmos-Predict 2.5, 14.5 for WEAVER and 11.0 for IWS. The gap widens further on the perceptual and distributional metrics: GWS has the lowest LPIPS in every bin and its FID rises from 15 to 32 over the minute, whereas Cosmos-Predict 2.5 rises from 30 to 73, WEAVER from 46 to 87 and IWS stays above 180 throughout.

4.4 Policy Evaluation

We evaluate whether GWS can serve as a faithful proxy for real-robot policy evaluation. We define three tasks: placing plates into a bin, throwing plastic bottles into a bin, and folding a t-shirt. The policy is π_0 (Black et al., 2024) fine-tuned on teleoperation demonstrations of our tasks, and we compare its final checkpoint (15k steps) against an early checkpoint (2.5k steps). For every policy–task pair we run 32 episodes on both real robot and inside GWS, with the same prompt; each episode

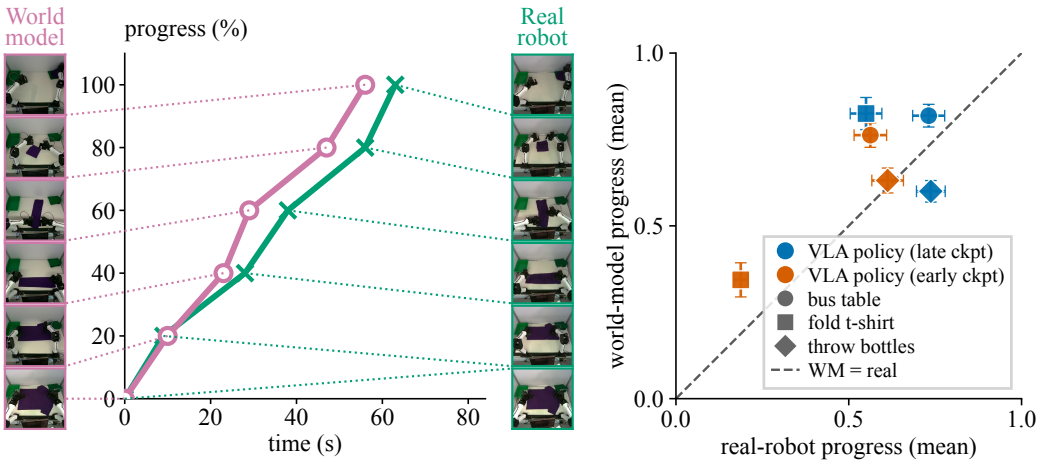

Figure 6: **Policy evaluation inside GWS against the real robot.** (a) Task progress over time for one t-shirt-folding episode of the fine-tuned policy, on the real robot (×) and inside GWS (○). (b) Mean progress per policy and task. The simulator is generally optimistic in absolute progress and preserves the checkpoint ranking on two of three tasks.

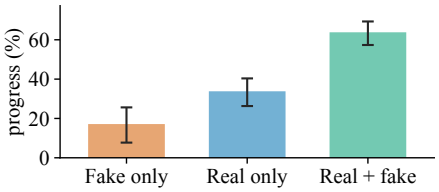

Figure 7: **Co-training with demonstrations collected inside GWS.** Real-robot progress (fraction of three chips bags placed in the basket within 60 s) of π_0 policies. Combining real and simulated episodes nearly doubles the performance.

is hand-labelled with a task progress in $[0, 1]$, the fraction of the task completed within the episode. Fig. 6 summarizes the results. Per-task mean progress in the simulator correlates with the real robot at Pearson $r = 0.71$ over the six policy–task points. The simulator is optimistic in absolute terms: for plates and t-shirt folding it reports higher progress than the real robot, and in the episode of Fig. 6(a) the simulated rollout reaches every stage of the fold 5–10 s before the real one.

4.5 Co-Training with Generated Demonstrations

Since GWS runs in real time from teleoperation input (Fig. 1), an operator can collect demonstrations inside the simulator without a robot: the GELLO leader arms (Wu et al., 2024b) drive the simulated robot at 10 Hz, and each episode is recorded in the training-data format, with the rendered views, the predicted joint state and the commanded actions. We test whether such simulated demonstrations are useful for policy learning on a single task, placing three bags of chips into a basket. We fine-tune the π_0 policy (Black et al., 2024) on three datasets: 75 real demonstrations, 75 simulated demonstrations, or their union (150 episodes). Each policy is evaluated in 10 real-robot trials of 60 s. Fig. 7 reports the results. Combining the two nearly doubles the real-only result to 63%. This study shows that demonstrations collected inside GWS carry useful signal for the real robot, rather than isolating the value of a simulated episode against a real one.

4.6 Ablation on Contact-Based Memory

We ablate the memory design of Sec. 3.2 on a task where objects leave and re-enter view: organising condiment bottles into a bin. Three models differ only in memory: no memory, sparse memory with keyframes taken at a fixed 1-s stride, and our contact-based memory. To isolate the effect, all three use a short 1-s recent context and two memory slots. Fig. 8 shows the three rollouts and the real observation when the left wrist camera leaves and looks into the bin again. Without memory, the bin is rendered empty: the bottles placed earlier are no longer in the context and the model produces an unoccupied bin. With sparse memory, the bin contains bottles, but the wrong ones. With contact-based memory, the bin contents match the real observation. This follows from the mechanism: a

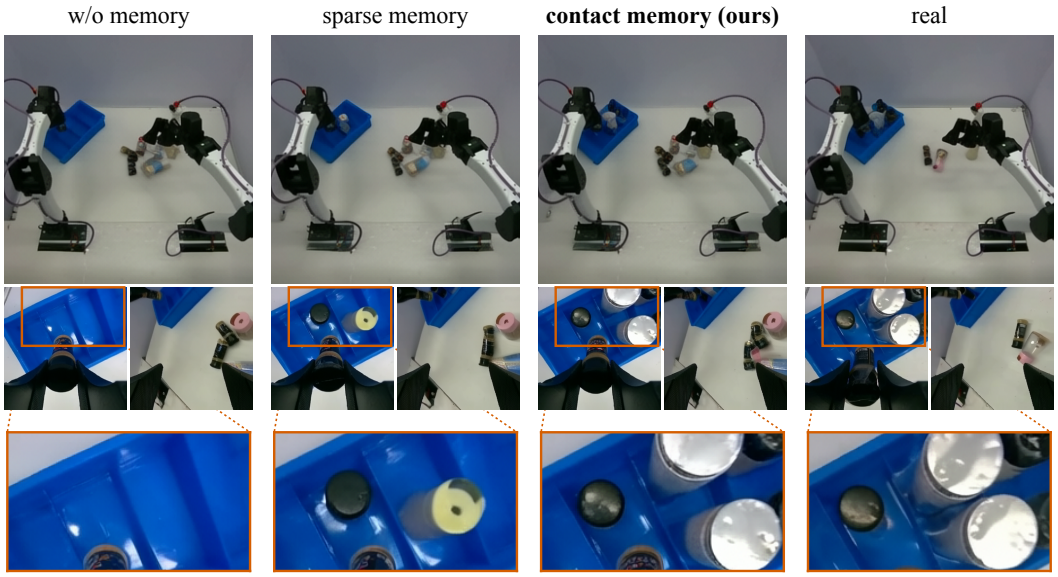

Figure 8: **Contact-based memory keeps placed objects persistent.** Three models with the same architecture but different memory selection. Without memory the bin is empty, sparse time-strided memory hallucinates its contents, and contact-based memory matches the real observation.

predicted contact transition, here the release of a bottle, writes exactly the frame that shows the new scene state into the keyframe bank, and when the end-effector returns to a similar pose the frame is retrieved as clean context. The memory therefore makes placed objects persistent across the rollout at a fixed cost in context length, enabling long rollouts of Secs. 4.3 and 4.2.

5 Discussion and Limitations

We presented GWS, a world model for real-time, long-horizon, contact-rich manipulation across diverse objects and tasks. By combining long–short-term memory, keyframe selection based on predicted contact transitions, and streaming dynamics with a persistent attention sink, our framework supports extended interactive rollouts and applications in policy evaluation and data generation.

Scene diversity. Our model is limited to interactions within a YAM station. Future work can extend this framework to open-world environments with more diverse backgrounds, layouts, and scenes.

Robot embodiment. Our model captures a single robot embodiment. Learning a world model across multiple embodiments is an important next step toward simulating interactions across robots with different kinematics and action spaces.

Manipulation dexterity. The YAM station's grippers offer less dexterity than multi-fingered hands, limiting the range of manipulation behaviors studied here. Future work can extend our model to more dexterous robots and fine-grained manipulation tasks, such as in-hand object reorientation.

6 Conclusion

We presented GWS, a world simulator whose dynamics are trained from scratch on robot-native teleoperation data rather than adapted from an internet-video generator. A unit-sphere latent, a streaming hybrid-attention transformer with a learned sink and a KV-cache pyramid, and a contact-based memory let it run in real time for interactive teleoperation, sustain two-minute rollouts and transfer to held-out tasks; in a preliminary equal-data comparison on the public ABC dataset, an ABC-trained instance matches or exceeds a post-trained video foundation model and a prior causal design. Our experience suggests that minute-scale autoregressive latent rollouts benefit from a norm-constrained latent, a sink and content-keyed memory, and that, in a preliminary equal-data comparison, an internet-video prior did not help. We release code and model weights; closing the loop to real-world policy outcomes is our next step.

Ethics Statement

GWS is trained on recordings of robot teleoperation: camera views of a bimanual robot workspace together with the robot's joint states and commanded actions. The recordings contain no personal data, no human subjects were studied, and no identity, biometric or demographic information is collected or used; the teleoperators are not identified and appear in no released artefact. The public ABC dataset is used under its Apache-2.0 licence and cited by its paper. The simulator is trained on predominantly successful demonstrations and may over-estimate the success of a policy; it must not be the sole gate for deploying a policy, and policies evaluated inside it must still be validated on hardware, since simulator outcomes are not yet shown to track real outcomes. Training the model required a substantial multi-GPU compute budget; releasing the weights avoids repeating it.

Reproducibility Statement

The architecture, training recipe and inference procedure are described in Sec. 3; the evaluation protocols, baseline adaptations and inference settings are described in Sec. 4. We release the code for training, inference and the interactive demo, and the model weights of the full GWS and of the ABC-trained GWS. With the release, the equal-data comparison and the speed measurements can be reproduced: we release our ABC split (training tasks, held-out episodes and the 10 held-out tasks), the benchmark episode list, the square-crop preprocessing and the evaluation scripts.

AI Use Statement

In this work, we used generative AI tools to implement methods (writing and refactoring parts of the training, inference and evaluation code) and to support the interpretation of results. We have not used generative AI tools to generate synthetic data sets, develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, propose or refine hypotheses, or design the research methodology or experiments; translation, dataset cleaning and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to edit parts of the paper for readability. We have reviewed all AI-assisted work: LLM-generated code was verified and tested for correctness by the authors, and all claims, numbers, figures and citations were checked by the authors against our experiment records and the cited sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

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

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

Apurva Badithela, David Snyder, Lihan Zha, Joseph Mikhail, Matthew O'Kelly, Anushri Dixit, and Anirudha Majumdar. Reliable and scalable robot policy evaluation with imperfect simulators. *arXiv preprint arXiv:2510.04354*, 2025.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22669–22679. IEEE, 2023.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

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

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pp. 7480–7512. PMLR, 2023.

Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML-11)*, pp. 465–472, 2011.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29, 2016.

J Randall Flanagan and Alan M Wing. The role of internal models in motion planning and control: evidence from grip force adjustments during movements of hand-held loads. *The Journal of Neuroscience*, 17(4):1519–1528, 1997.

Zelin Fu, Bin Tan, Changjiang Sun, Shaohui Liu, Kecheng Zheng, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Vision pretraining for dense spatial perception. *arXiv preprint arXiv:2607.05247*, 2026.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-World: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

Fabian Hutmacher and Christof Kuhbandner. Long-term memory for haptically explored objects: fidelity, durability, incidental encoding, and cross-modal transfer. *Psychological science*, 29(12): 2031–2038, 2018.

Arhan Jain, Mingtong Zhang, Kanav Arora, William Chen, Marcel Torne, Muhammad Zubair Irshad, Sergey Zakharov, Yue Wang, Sergey Levine, Chelsea Finn, et al. Polaris: Scalable real-to-sim evaluations for generalist robot policies. *arXiv preprint arXiv:2512.16881*, 2025.

Arnav Kumar Jain, Yilin Wu, Jesse Farebrother, Gokul Swamy, and Andrea Bajcsy. Weaver, better, faster, longer: An effective world model for robotic manipulation. *arXiv e-prints*, pp. arXiv–2606, 2026.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024. ICML 2025.

Sergey Levine and Vladlen Koltun. Guided policy search. In *International conference on machine learning*, pp. 1–9. PMLR, 2013.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024.

Yaxuan Li, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. Worldeval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations (ICLR)*, 2026a.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. nGPT: Normalized transformer with representation learning on the hypersphere. *arXiv preprint arXiv:2410.01131*, 2024.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation. *arXiv preprint arXiv:2506.00613*, 2025.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.

Wei-Cheng Tseng, Gashon Hussein, Yuzhu Dong, Allen Z Ren, Lucy X Shi, XuDong Wang, Sergey Levine, Zhaoshuo Li, Jinwei Gu, Florian Shkurti, et al. Sc3-eval: Evaluating robot foundation models via self-consistent video generation. *arXiv preprint arXiv:2606.18610*, 2026.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Yixuan Wang, Rhythm Syed, Fangyu Wu, Mengchao Zhang, Aykut Onol, Jose Barreiros, Hooshang Nayyeri, Tony Dear, Huan Zhang, and Yunzhu Li. Interactive world simulator for robot policy training and evaluation. *arXiv preprint arXiv:2603.08546*, 2026.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119, 2024a.

Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12156–12163. IEEE, 2024b.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time interactive long video generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies, 2026. URL https://arxiv.org/abs/2602.15922.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models, 2025. URL https://arxiv.org/abs/2412.07772.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *SIGGRAPH Asia 2025 Conference Papers*, 2025a. doi: 10.1145/3757377.3763833.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025b.

Kaifeng Zhang, Shuo Sha, Hanxiao Jiang, Matthew Loper, Hyunjong Song, Guangyan Cai, Zhuo Xu, Xiaochen Hu, Changxi Zheng, and Yunzhu Li. Real-to-sim robot policy evaluation with Gaussian splatting simulation of soft-body interactions. *arXiv preprint arXiv:2511.04665*, 2025a.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A fine-grained world model for robot manipulation. *arXiv preprint arXiv:2406.14540*, 2024.

A Architecture and Training Details

Tokenizer recipe. Each observation x_t contains three 256×256 RGB views. A shared convolutional encoder with stride 16 (46.1 M parameters) maps each view to a 16×16 grid of 16-dimensional vectors; z_t collects the latent tokens from all three views. The decoder is a U-ViT flow-matching decoder (50.3 M parameters) (Bao et al., 2023). The encoder output is L2-normalised per spatial position; there is no KL term, and one learned scalar inside the decoder rescales the latent, so the served latent is exactly unit-norm (measured mean norm 1.000, standard deviation 10^{-4}). The decoder reconstructs each view from noise conditioned on its latent tokens. Its loss is the flow-matching error (Lipman et al., 2022; Liu et al., 2022) plus LPIPS (weight 0.5) (Zhang et al., 2018), representation alignment (0.1) against a frozen vision-transformer teacher with a warm-up (Yu et al., 2025b; Fu et al., 2026), and a uniformity term (0.1), a hinged squared-MMD reformulation of the loss of Wang & Isola (2020). Decoding takes two Euler steps, about 14 ms per view-frame; for video, one noise tensor is drawn per episode and reused for every frame. Tab. 1 reports reconstruction on held-out episodes; on held-out tasks PSNR is 26.51 / 30.06 / 30.65 dB for the top / left-wrist / right-wrist views. The wrist views reconstruct better than the top view in single frames, whereas in rollouts they trail it (Sec. 5); the rollout wrist gap is therefore a dynamics limitation, not a tokenizer one.

Table 1: **Tokenizer reconstruction on held-out episodes.** Single-frame reconstruction at 256×256 per view; PSNR in dB.

View	PSNR ↑	SSIM ↑	LPIPS ↓
Top	27.95	0.722	0.085
Left wrist	31.26	0.664	0.100
Right wrist	31.55	0.670	0.100

Sequence layout and positions. The three $16 \times 16 \times 16$ view latents are stacked into a $48 \times 16 \times 16$ tensor per frame and patchified with patch size 2 (192 tokens) plus a state token carrying the 14-D joint positions, 193 tokens per frame. The backbone has 28 blocks with hidden size 2048 and 16 heads of width 128. Spatial attention uses 2-D rotary positions and temporal attention rotary positions on the frame index; the sink frame (193 learned vectors) receives no rotation, memory frames are rotated by a log-compressed time offset Δt to the current frame, never by a slot index, and unused slots are masked. The five joint-attention blocks use a frame-causal block-sparse mask. Every attention applies qk-RMSNorm (Dehghani et al., 2023); the flow time enters through adaptive layer normalisation with zero initialisation (Peebles & Xie, 2023), and each frame's 20-D action is added to that frame's modulation through a small MLP.

Dynamics training. The flow loss is applied on the context positions with independent logit-normal flow times per frame (Esser et al., 2024; Chen et al., 2024); the state token carries a flow loss with weight 0.1 and the contact readout a BCE loss with weight 0.1. Memory slots are dropped with probability 0.2. We use AdamW with learning rate 10^{-3}, 1k warm-up steps and then a constant schedule, $\beta = (0.9, 0.95)$, weight decay 0.01, gradient clipping at 0.5 and an EMA of the weights with decay 0.999; the EMA weights are used at inference. The effective batch size is 2048 windows.

The model was initialised, weights only, from an earlier training run of the same architecture taken at 14.5k of its steps.

Contact labeler. Each call of the fine-tuned vision-language model (Clark et al., 2026) receives, for one arm, an 8-frame clip at 5 Hz of composites of that arm's wrist view and the base view, and returns per-frame booleans with a soft probability from the token logits; two passes offset by one frame are interleaved to 10 Hz and smoothed with a 3-frame box filter, a hysteresis (enter above 0.6, leave below 0.4) and a minimum dwell. The model was fine-tuned on human-corrected labels with task-disjoint splits. In a visual audit, contact onsets lag gripper closing by 0.03–0.38 s, releases coincide with re-opening, and sustained holds stay at high probability with no gripper motion, so the detector is not a gripper proxy. A re-run reproduces the reference predictions at 99.2% per-frame agreement, a determinism check rather than a validity measure. Rejected design choices: a gripper open/close hysteresis as labeler (blind to non-prehensile contact; misses about 60% of grasps on the achieved gripper state), field-of-view-overlap retrieval keys (degenerate under a static top camera), adaptive contact thresholds (they leak future statistics), and co-denoising contact as a flow modality (the detached readout was better calibrated).