

Walk the Painting: Harnessing Internet Priors for Open World Interaction

Supplementary Material

In this supplementary material, we report additional results, details, and visualizations. We begin by discussing qualitative video results of our method reported in the supplementary HTML file (videos.html) in Section A. Next, we report further implementation details of our approach in Section B and provide data, compute, latency, and resolution comparisons with the Navigation World Models (NWM) [2]. Finally, we discuss some failure modes of our method in Section C.

A. Video Results

We provide an extensive set of visualizations in the attached webpage (videos.html).

In Section 4, we compare NWM [2], our SVD-based model [3], our Cosmos-based model [1], and ground-truth trajectories on the test set of the RECON dataset [6]. Each row shows the initial frame, ground-truth rollout, and the predicted rollouts from each method. As time progresses, NWM’s predictions drift substantially farther from the ground truth than either of our models. Between our variants, the SVD model most accurately follows the commanded actions but loses fine-grained visual fidelity (e.g., grass texture), whereas Cosmos preserves texture better but exhibits less precise action following. These observations match our quantitative findings: Cosmos achieves the best LPIPS [8] and DreamSim [4] due to high visual fidelity, while SVD obtains better SCS due to more accurate action following, confirming the utility of our metric.

In Section 3, we show the video visualizations for navigation and manipulation for our 25-DoF humanoid joint-angle control model. In Section 3.1, we show navigation results under the 25-DoF humanoid joint-angle control on the validation split of the 1x dataset. The model navigates precisely even within small spaces where moderate drift would cause failure (e.g., missing a nearby table). For content not visible in the initial frame, the model synthesizes plausible completions despite being different from the ground truth. In Section 3.2, we show manipulation results using 25-DoF humanoid control on the validation split of the 1x dataset. Although most of the robot’s body is occluded in the egocentric view and no link-length information is provided, the model still follows the action sequence accurately. We also include counterfactual sequences with close initial poses but different object configurations. Across these scenarios, the model consistently demonstrates precise action following.

In Section 2, we show **Zero-Shot Open-World** generalization videos on *real-world images captured by us*. Results

with the 3-DoF position control are shown in Section 2.1, where the model follows *unseen* action trajectories while imagining plausible continuations for regions outside the initial field of view; 25-DoF humanoid joint angle control results in Section 2.2 demonstrate that our model can precisely follow complex *unseen* action trajectories across a wide range of real scenes, highlighting its effectiveness as a general egocentric world model.

Finally, we show **Zero-Shot Open-World** generalization to *paintings* in Section 1. Section 1.1 shows 3-DoF navigation inside paintings. Despite the domain shift (paintings never appear during training), the agent follows the *unseen* reference trajectory precisely while preserving the style and texture of the artwork; whereas Section 1.2 shows manipulation of the humanoid’s neck pitch to make the agent “look down” and observe itself within the painting. Although imperfect, the results are striking and demonstrate substantial generalization. To our knowledge, such extreme generalization, **particularly under high-dimensional action spaces**, has not previously been shown.

B. Implementation Details

Training details. We train both our SVD [3] and Cosmos [1] models on 8 A100 GPUs for all variants. For the SVD variant, we take a learning rate of $1e-5$ for all the model parameters except the action projection layers, where we adopt a larger learning rate of $1e-4$. For the Cosmos variant, we use a learning rate of $1e-6$ for the other parameters and $1e-5$ for the action projection layers. Our longest training run takes 8 days for our Cosmos-2B training for 25 DoF manipulation on the 1x dataset, which has 100 hours of training videos at 30 FPS. We subsample the frames at 5 FPS for both training and inference. For 1x training, we select continuous frame sequences with a stride of 5. For training on 3-DoF datasets, we use a stride of 1 for SCAND [5] and Tartan [7], and a stride of 5 for RECON.

Comparison with NWM. Table A lists the action datasets and compute resources used for training each method. NWM employs a higher-resolution version of the Huron dataset, which we did not have access to. While NWM is trained on 64 H100 GPUs, both our SVD and Cosmos variants are trained on only 8 A100s. While NWM predicts frames at 224×224 , we predict at 512×512 for SVD and 480×640 for Cosmos. For fairness, we compute all metrics at 224×224 by downsampling our predictions. Despite using less action data and approximately $8 \times$ less compute, our models outperform NWM.

To generate the latency plot shown in Figure 5 of the

Method	3-DoF Action Data	Compute	Resolution	Avg. 64f Latency (s)
NWM	RECON, SCAND, Tartan-Drive, Huron	64× H100s	224×224	≈ 300
Ours (SVD)	RECON, SCAND, Tartan-Drive	8× A100s	512×512	≈ 200
Ours (Cosmos)			480×640	≈ 50

Table A. Comparison of data, compute, resolution and latency with NWM for 3-DoF position control.

main paper, we perform inference on the same 20 RECON samples across varying prediction horizons, using each model’s native resolution – 224×244 for NWM, 512×512 for Ours (SVD), and 480×640 for Ours (Cosmos). All inferences are executed on a single A100 GPU. Our models achieve up to **6× faster** inference speed while predicting at up to **2× higher resolution**.

C. Failure Modes

Our model struggles to generate small manipulated objects in a physically consistent manner, particularly when object shapes must be preserved across occlusions. Maintaining object permanence during complex manipulation also remains challenging. In Fig A we illustrate one such failure case, where the square shaped block gets distorted during manipulation. We attribute these issues primarily to limitations of the underlying video generation backbones and expect that future advances in video generative modeling will improve these aspects of world-modeling performance.

References

- [1] 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. 1
- [2] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *CVPR*, 2025. 1
- [3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1
- [4] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2023. 1
- [5] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warrnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7(4):11807–11814, 2022. 1
- [6] Dhruv Shah, Benjamin Eysenbach, Gregory Kahn, Nicholas Rhinehart, and Sergey Levine. Rapid exploration for open-world navigation with latent goal models. In *CoRL*, 2021. 1
- [7] Samuel Triest, Matthew Sivaprakasam, Sean J Wang, Wenshan Wang, Aaron M Johnson, and Sebastian Scherer. TartanDrive: A large-scale dataset for learning off-road dynamics models. In *ICRA*, 2022. 1
- [8] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1

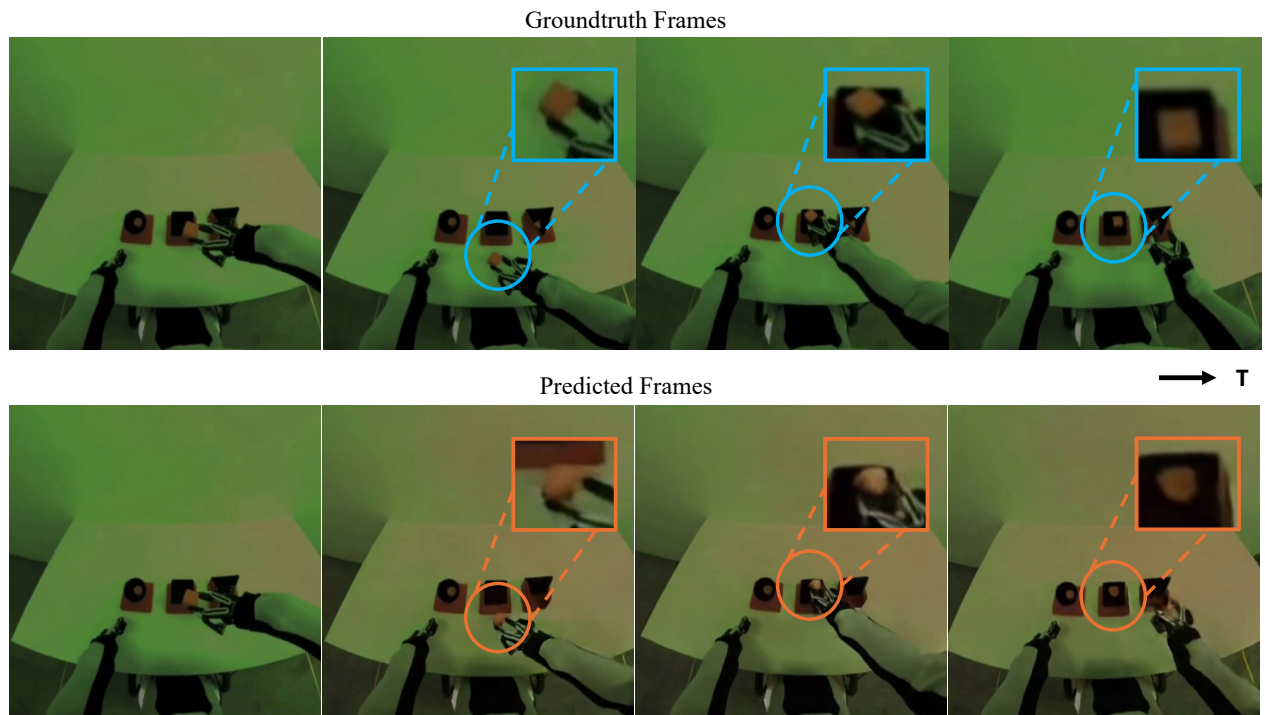


Figure A. Failure modes of our manipulation world model.