PRISM-World: Teaching Video World Models the Physics of Retail Environments

Amirreza Rouhi Rajat Aggarwal Parikshit Sakurikar Anoop M. Namboodiri Sashi P. Reddi
DreamVu
{amirreza.rouhi, rajat.aggarwal, parikshit.sakurikar, anoop.namboodiri, sashi.reddi}@dreamvu.ai

Figure 1. **Qualitative comparison on two held-out clips.** Frames sampled at $t \in \{0.3, 1.9, 3.7, 5.4\}$ s; rows show Base Cosmos vs. PRISM-World. Both models are conditioned on the same first frame, so any divergence between the rows is attributable to the diffusion rollout alone. The prompt for each panel is printed in the panel header. **(a)** Base Cosmos hallucinates a second subject next to the prompted woman at the produce shelf and never opens the fridge; PRISM-World preserves a single subject, opens the glass fridge door, and reaches in to browse. **(b)** Base Cosmos breaks fixed-camera continuity (the viewpoint drifts into a close-up by t=2.9 s) and the reaching arm contacts the metal shelf in a geometrically implausible pose; PRISM-World preserves the CCTV viewpoint, bends naturally, and reaches into the crate with a coherent hand pose.

Abstract

Text-conditioned video diffusion models such as Cosmos-Predict2 [8] act as world models capable of rolling out plausible future frames from a single observation, but they are trained on broad internet video and lack the domain priors required for embodied retail settings—structured shelves, dense shopper traffic, and the rigid-object physics that govern carts, baskets, and produce crates. We address this gap on two fronts. ***(i) PRISM-World****, a curated exo-centric retail-video dataset of* 15,115 *atomic-action-captioned* 5.8 *s clips derived from* 1,093 *multi-camera 4K recordings (*≈ 437 *h) of five retail stores, with trajectory and atomic-action labels auto-generated by a large vision-language model [3].* ***(ii) A LoRA recipe*** *that adapts Cosmos-Predict2-2B-Video2World to PRISM-World by training only* ≈ 7 *M parameters (*0.35% *of the* 2 *B-parameter DiT) for* ∼ 3.6 *h on* 8×*H100. On the full* 757*-clip held-out test split the adapted model reduces Fréchet Video Distance from* 11.91 *to* 8.98 *(*−24.6%*) and LPIPS by* −12.2 *to* −19.1% *at horizons from* 0.5 *to*

5.5 *s. Qualitatively (Fig. 1), the adapted rollouts execute the prompted action, preserve actor identity, and respect fixed-camera continuity and basic object physics that the unadapted baseline violates.*

1 Introduction

Recent video diffusion foundation models [1, 2, 6, 8] instantiate world models [4, 5] by rolling out future frames from an initial observation and a textual intent. Retail aisles are a high-leverage embodied domain, yet broad internet-trained models lack priors for structured shelves, shopper traffic, staff uniforms, and recurrent stocking and checkout dynamics. Prior retail-video benchmarks consist of short labelled action snippets [10] or coarse surveillance excerpts [11], and are not curated for video world-model training. Recent domain-adapted video diffusion world models, such as GAIA-1 [6] for autonomous driving, target first-person automotive footage rather than the structured third-person retail environments we consider. Multi-second prediction in this setting supports simulation-based policy learning, action ranking via forward rollouts, and pre-action failure anticipation for embodied agents deployed in working stores.

We introduce **PRISM-World**, a curated exo-centric retail-video dataset with trajectory descriptions and atomic-action labels generated by a large vision-language model [3]. We show that LoRA [7] adaptation of Cosmos-Predict2-2B-Video2World [8] on PRISM-World improves FVD [12] by 24.6% and LPIPS [13] by 12 to 19% on the held-out test split at horizons from 0.5 to 5.5 s, without architectural changes. Figure 1 illustrates the qualitative gain: faithful execution of the prompted action, preservation of actor identity, and respect for fixed-camera continuity and basic object physics that the unadapted baseline violates.

2 Method

PRISM-World. The source corpus comprises 1,093 multi-camera 4K fixed-mount recordings from five retail stores (≈ 437 h total); each recording bundles four synchronised camera views mounted at the corners of the store floor plan, providing complementary third-person coverage of every aisle. For each retained window, a large vision-language model [3] produces a free-form English *trajectory description* and a structured *atomic-action* list of the form `label(target), ...` (e.g. `lift_crate(blue crate), push_cart(shopping cart)`). A curation pipeline tags windows by store section (*produce, dry-goods, dairy, beverage, checkout, customer-service, restock/transit*), removes unusable or action-free segments, slices footage into 5.8 s / 16 fps clips matching the Cosmos-Predict2 input format, verifies subject-and-action presence using person and open-vocabulary detectors, and balances clips across sections. The final set contains 15,115 clips, partitioned 13,603/755/757 for train/eval/test, section-stratified.

Held-out test set, n=757					
	FVD ↓	LPIPS ↓ at horizon			
		0.5 s	1.5 s	3.0 s	5.5 s
base Cosmos	11.91	0.244	0.339	0.418	0.460
PRISM-World (ours)	**8.98**	**0.214**	**0.283**	**0.346**	**0.372**
Δ vs. base	−24.6%	−12.2%	−16.6%	−17.3%	−19.1%

Table 1. Results on the 757-clip held-out test split. PRISM-World improves both distribution match (FVD) and per-frame predictive fidelity (LPIPS) at every measured horizon over the unadapted base.

Backbone and adaptation. We adapt Cosmos-Predict2-2B-Video2World [8], an open-weight 2 B-parameter text-conditioned video diffusion model with a DiT backbone [9], which rolls out 5.8 s of future video from a single first frame and a text prompt. We apply rank-16 LoRA [7] to all attention and MLP projections, training ≈ 7 M parameters (0.35% of the backbone) for 2,000 iterations on 8×H100 in ∼ 3.6 h. Base and adapted models use identical prompts, seeds, and inference settings, so quality differences between rollouts are attributable to the adapted weights.

3 Experiments

Setup. We evaluate on the 757-clip held-out test split. For each clip we condition both base Cosmos and PRISM-World on the same first frame and prompt, using identical seeds and inference settings. We report **FVD** [12] on I3D-R50 Kinetics features—the standard distribution-match metric for video generation [2, 8]—and **LPIPS** [13] between the generated frame and the corresponding ground-truth frame at horizons $t \in \{0.5, 1.5, 3.0, 5.5\}$ s, following the protocol of recent embodied world models [1, 2, 5].

Results. Table 1 shows that PRISM-World reduces FVD by 24.6% and improves LPIPS at every horizon, with the gap widening from −12.2% at 0.5 s to −19.1% at 5.5 s. Figure 1 shows the corresponding qualitative gains in prompt following, actor identity, camera continuity, and object handling. The gain is obtained by training only 0.35% of the DiT weights for ∼ 3.6 h on 8×H100, suggesting that domain-specific data is a major bottleneck for retail world modelling, even without architectural changes or large training budgets.

4 Discussion and Limitations

The gains reported here are obtained on multi-second rollouts of up to 5.8 s; extending the recipe to long-horizon prediction over tens of seconds to minutes, as in recent driving world models [6], remains open. PRISM-World covers five stores from a single retail chain, so cross-chain and cross-region generalisation is unverified and likely benefits from a small top-up adaptation on a target corpus. The model is text-conditioned rather than action-conditioned; an action-conditioned variant that ingests a motor or skill vocabulary is the natural next step toward closed-loop embodied use, where forward rollouts can rank candidate actions and feed a downstream value head.

References

[1] Mahmoud Assran et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2503.01928*, 2025. 2

[2] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024. 2

[3] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 2

[4] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 2

[5] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 2

[6] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 2

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[8] NVIDIA. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 1, 2

[9] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[10] Bharat Singh, Tim K. Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[11] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[12] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *arXiv preprint arXiv:1812.01717*, 2018. 2

[13] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2