DreamZero: World Action Models are Zero-shot Policies

2026-03-01

Paper Info

- **Authors:** Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, et al. (NVIDIA)
- **Link:** arXiv | Project | GitHub
- **Read date:** 2026-03-01

Summary

VLAs (Vision-Language-Action models) generalize well to new objects and semantics but struggle with novel physical motions and environments—they lack spatiotemporal priors from image-text pretraining. DreamZero introduces a **World Action Model (WAM)** built on a pretrained 14B autoregressive video diffusion backbone. By jointly predicting future video frames and actions, DreamZero learns physical dynamics via an inverse-dynamics formulation: align motor commands with predicted visual futures. This shifts action learning from dense state-action imitation to world-model-guided control. Key outcomes: >2× improvement over SOTA VLAs on environment/task generalization, effective learning from diverse heterogeneous robot data (no repetitive demos), real-time control at 7Hz via 38× inference speedup, and two forms of cross-embodiment transfer—video-only demonstrations (+42% with 10–20 min) and few-shot embodiment adaptation (30 min play data to transfer AgiBot G1 → YAM).

Key Ideas

1. **VLAs lack spatiotemporal priors** — VLM backbones use static image-text; VLAs fail at novel skills (e.g. "untie shoelace") not in robot data. Video diffusion encodes how the world evolves.

2. **Joint video-action = inverse dynamics** — DreamZero decomposes as $\pi(\mathbf{o}_{l:l+H}, \mathbf{a}_{l:l+H}) = \pi(\mathbf{o}|\text{context})\,\pi(\mathbf{a}|\mathbf{o}, \text{context})$. Video prediction is implicit visual planning; actions follow predicted futures.
3. **Diverse data > repetitive demos** — Heterogeneous trajectories outperform same-hours multi-task repetitive data. Generalist policies don't need many demos per task.
4. **Autoregressive + KV-cache reset** — After each action chunk, replace predicted frames with ground-truth observations in KV cache to prevent error accumulation; keeps native frame rate for precise video-action alignment.
5. **Policy quality video quality** — Larger pretrained video models → better video predictions → better action execution. Improving robotics reduces to improving video generation.
6. **Cross-embodiment transfer** — Video-only from humans or other robots (no actions) improves target robot by >42% with 10–20 min. Few-shot adaptation: 30 min play on new embodiment preserves zero-shot generalization.

Method

- **Architecture:** 14B autoregressive DiT (flow matching) on pretrained video diffusion (Team Wan 2025). Inputs: visual context (VAE), language (text encoder), proprioception (state encoder). Joint video + action prediction via separate decoders; minimal added parameters.
- **Formulation:** Chunk-wise autoregressive prediction. Each chunk: K latent frames, horizon H. Joint flow-matching objective over video latents $\mathbf{z}_t$ and normalized actions $\mathbf{a}_t$. Teacher forcing: denoise chunk k conditioned on clean chunks $1..k{-}1$.
- **Training objective (Eq. 3):** $\mathcal{L} = \mathbb{E}[\sum_k w(t_k)\|\mathbf{u}_\theta([\mathbf{z}^k_{t_k}, \mathbf{a}^k_{t_k}]; \mathcal{C}_k, \mathbf{c}, \mathbf{q}_k, t_k) - \mathbf{v}^k\|^2]$, with velocity $\mathbf{v}^k = [\mathbf{z}^k_1, \mathbf{a}^k_1] - [\mathbf{z}^k_0, \mathbf{a}^k_0]$.
- **Shared timestep** for video and action (unlike some WAMs) for faster convergence.
- **Inference optimizations:** DreamZero-Flash (decoupled denoising schedules), system-level parallelism/caching, quantization, CUDA tuning → 38× speedup, 7Hz closed-loop.

Results

- **Environment & task generalization:** >2× average task progress vs SOTA pretrained VLAs (RoboArena, PolaRiS). Retains +10% after task-specific post-training.
- **Real-time control:** 7Hz action chunk generation on 14B model via 38× inference speedup.
- **Video-only cross-embodiment:** Human (12 min) or YAM robot (20 min) video demos → +42% relative improvement on AgiBot G1 unseen tasks.
- **Few-shot embodiment adaptation:** Pretrained on AgiBot G1, adapt to YAM with 30 min play data; retains zero-shot generalization.
- **Data:** ~500 hours real-world; shows non-trivial Genie Sim 3.0 (100 tasks) without sim training.

Strengths

- Strong empirical gains on generalization; breaks VLA ceiling for novel skills and environments.
- Data efficiency: diverse heterogeneous data suffices; no need for repeated demonstrations per task.
- Novel cross-embodiment results: video-only transfer and 30-min few-shot adaptation are striking.
- Real-time deployable (7Hz) despite 14B diffusion model—important system contribution.
- Open-source: weights, inference code, benchmarks (RoboArena, PolaRiS, Genie Sim 3.0).
- Clean formulation: policy quality tied to video quality; clear research direction.

Weaknesses / Limitations

- Single embodiment focus (AgiBot G1, YAM); broader cross-embodiment scaling not shown.
- 500 hours of real data still nontrivial; cost of data collection unclear.
- Video diffusion backbone dependency; scaling/generalization may plateau with video model quality.
- Limited comparison to other WAMs (Kim et al., Liang et al., Pai et al.) in main text; ablations could be deeper.
- 7Hz is usable but not high-frequency; some manipulation may need higher rates.

Relevance to My Work

How does this connect to MedEnv, GRPO, or other active projects?

- **World models for RL/planning:** If MedEnv or similar settings involve sequential decision-making, WAMs offer a different backbone than pure policy nets—video as implicit planner, inverse dynamics for actions. Could inform hierarchical or model-based RL.
- **Data efficiency:** DreamZero's diverse-data story (vs repeated demos) parallels sample-efficiency concerns in RL; lessons may transfer to how we structure demonstrations or play data.
- **Cross-domain transfer:** Video-only and few-shot adaptation could inspire transfer setups in non-robotics domains (e.g. sim→real, domain A→domain B) if we have video-like sequential observations.
- **Foundation models for control:** Contrast with language/LLM-based control; video diffusion as alternative pretrained prior for embodied AI.

Key Takeaways

1. WAMs built on video diffusion beat VLAs on motion/skill generalization by inheriting spatiotemporal priors; joint video-action prediction = inverse dynamics.
2. Diverse heterogeneous robot data outperforms repetitive demos; generalist policies don't require many demos per task.
3. Cross-embodiment transfer is surprisingly efficient: video-only (10–20 min) and 30-min few-shot adaptation both work.
4. Policy performance is tied to video generation quality; improving robotics = improving video models.
5. 38× inference speedup (decoupled schedules, caching, quantization) makes 14B diffusion viable for 7Hz real-time control.