



GrndCtrl: Grounding World Models via Self-Supervised Reward Alignment

<https://rlwg-grndctrl.github.io/>

Haoyang He^{1,2}

Jay Patrikar²

Dong-Ki Kim²

Max Smith²

Daniel McGann²

Ali-akbar Agha-mohammadi²

Shayegan Omidshafiei²

Sebastian Scherer^{1,2}

¹Carnegie Mellon University

²FieldAI

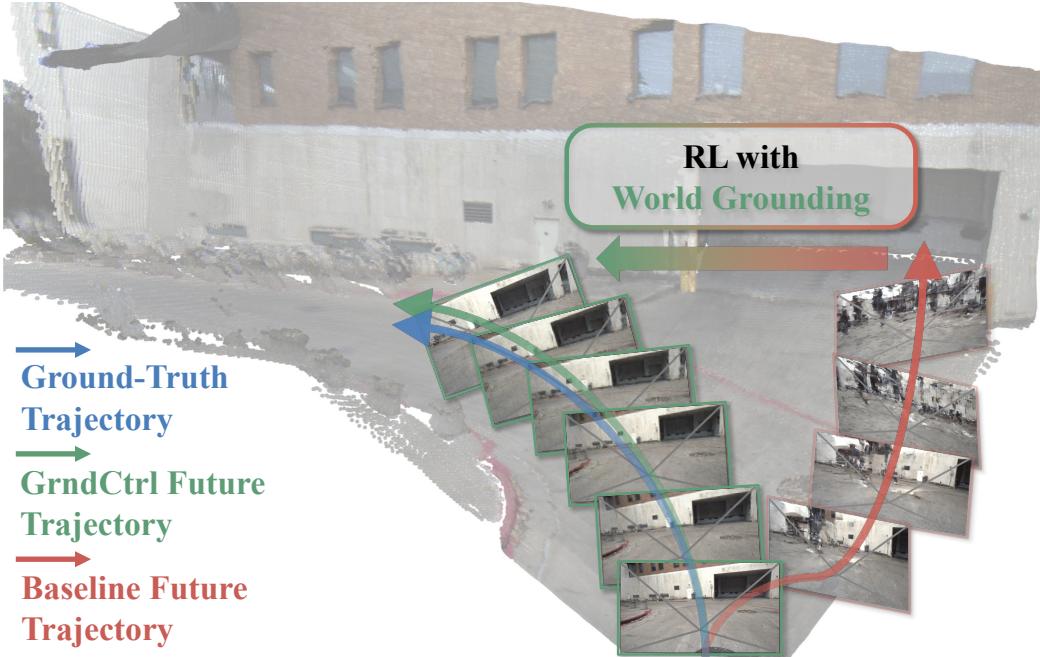


Figure 1. **Reinforcement Learning with World Grounding (RLWG)** addresses geometric inconsistencies in pretrained video world models through self-supervised post-training with verifiable rewards. Instead of reconstruction losses, RLWG grounds models using geometric and perceptual rewards from frozen evaluators. *GrndCtrl* instantiates RLWG using Group Relative Policy Optimization (GRPO), enabling physically consistent rollouts essential for reliable world generation.

Abstract

Recent advances in video world modeling have enabled large-scale generative models to simulate embodied environments with high visual fidelity, providing strong priors for prediction, planning, and control. Yet, despite their realism, these models often lack geometric grounding, limiting their use in navigation tasks that require spatial coherence and long-horizon stability. We introduce **Reinforcement Learning with World Grounding (RLWG)**, a self-supervised post-training framework that aligns pretrained world models with a physically verifiable structure through geometric and perceptual rewards. Analogous to reinforcement learning from verifiable feedback (RLVR) in lan-

guage models, RLWG can use multiple rewards that measure pose cycle-consistency, depth reprojection, and temporal coherence. We instantiate this framework with **GrndCtrl**, a reward-aligned adaptation method based on Group Relative Policy Optimization (GRPO), yielding world models that maintain stable trajectories, consistent geometry, and reliable rollouts for embodied navigation. Like post-training alignment in large language models, **GrndCtrl** leverages verifiable rewards to bridge generative pretraining and grounded behavior, achieving superior spatial coherence and navigation stability over supervised fine-tuning in outdoor environments.

1. Introduction

Large-scale video world models have emerged as powerful priors for modeling perception and control for embodied agents [2, 4, 9, 17, 19]. By learning to predict future observations from past frames and actions, these models approximate the transition dynamics of the physical world, enabling simulation, planning, and policy evaluation. Operating in the pixel domain aligns them with real-world sensors and exploits the vast implicit supervision available in video, allowing unified modeling across domains such as manipulation, driving, and navigation. Yet despite their impressive generative fidelity, these models are often incentivized to capture the appearance of motion more than its structure. Their rollouts remain visually plausible but geometrically and temporally inconsistent: poses drift, depths wobble, and trajectories lose alignment over time. Even subtle deviations in inferred geometry accumulate into compounding spatial errors corrupting metric structure. These instabilities limit the use of current models for closed-loop tasks such as localization, mapping, and planning, where physically consistent representation is essential.

We define **world model grounding** as aligning learned dynamics with physically verifiable spatial and temporal invariants, so that rollouts honor geometry and time in addition to reproducing surface appearance. Grounding shifts the objective of world modeling from visual plausibility to structural consistency, ensuring that the model’s internal dynamics respect the constraints of real motion and scene structure. To this end, we introduce **Reinforcement Learning with World Grounding (RLWG)**, a self-supervised post-training framework that refines pretrained world models using verifiable geometric and perceptual rewards derived from model rollouts. RLWG extends the principle of Reinforcement Learning with Verifiable Rewards (RLVR) from language models [14] to the embodied domain, replacing text-based logical verification with geometric and temporal verification. In RLWG, a pretrained world model is treated as a policy that generates multiple candidate rollouts from the same context; each rollout is automatically scored using verifiable grounding rewards that quantify spatial and temporal coherence, such as pose cycle-consistency, depth reprojection agreement and action adherence. Unlike reconstruction losses that only penalize pixel error, these rewards measure physical correctness of the rollouts.

To optimize these verifiable rewards efficiently, we adopt Group Relative Policy Optimization (GRPO) [21] as our training mechanism, yielding our algorithm, *GrndCtrl*. For each context (and actions when available), the model generates a group of rollouts that are ranked by their grounding rewards; relative advantages are computed within the group, and the latent transition operator is updated using a clipped policy gradient objective regularized toward the pretrained model. This formulation preserves visual qual-

ity while progressively aligning the model’s dynamics with measurable structure in the real world. The process requires no human annotations or external simulators, operating entirely through self-supervised reinforcement grounded in the model’s own predictions. Conceptually, *GrndCtrl* extends the success of GRPO-based alignment in generative modeling to the geometric domain, grounding visual world models in verifiable 3D and temporal coherence.

This paradigm reframes the role of post-training in world modeling. Rather than optimizing for perceptual fidelity or next-frame likelihood, RLWG drives the model toward internal representations that are self-consistent and physically grounded. It establishes a structural analogue to the self-alignment processes that have improved reasoning in large language models: where RLVR grounds language in logic, RLWG grounds world models in geometry. The resulting models are self-grounded, spatially coherent, and dynamically stable—capable not only of rendering the world vividly, but of representing it in actionable, physically consistent form. Through this lens, we move beyond visually coherent generation toward structurally consistent simulation, bridging the gap between generative video modeling and physical world understanding, and opening a path toward world models that can both imagine and inhabit the real world.

The main contributions of this work are:

1. We introduce **Reinforcement Learning with World Grounding (RLWG)**, a self-supervised grounding framework using verifiable geometric and temporal rewards from frozen evaluators without labels or simulators.
2. We construct *GrndCtrl*, a method that extends GRPO to the RLWG regime by multi-reward alignment over stochastic rollouts optimizing Translation, Rotation, Depth Temporal Reprojection Inlier ratio, and perceptual quality with pretrained frozen evaluators.
3. We provide a comprehensive evaluation of *GrndCtrl* across multiple datasets showing reduced pose error means and variances, with strong gains under counterfactual rollouts and generalization to unseen inputs.

2. Related Work

2.1. Controlling World Models

Recent progress in large-scale video foundation models has transformed video prediction into controllable world simulation. Models such as Cosmos-Predict [17], and V-JEPA [2] unify multi-modal conditioning for long-horizon prediction and control. These models achieve impressive simulation fidelity but still exhibit spatial drift, geometric misalignment, and temporal incoherence over extended rollouts, revealing limitations in geometric grounding. Architectural innovations like flow matching, conditional dif-

fusion transformers, and masked latent prediction have improved realism but not physical consistency.

Controlling these models can be categorized into *action-conditioned* and *camera-conditioned* paradigms. The first paradigm trains models to predict futures from discrete actions, and can be further defined by the nature of the action and observation frame. Some methods [29, 30] assume a static camera observing an embodiment performing actions, while others [3, 4, 6, 23] assume a fixed camera on the embodiment and train models to predict egocentric views. Jointly, they aim to enable model-predictive control, but face challenges with physical realism. The second paradigm explicitly decouples viewpoint from embodiment, allowing the model to generate future observations from arbitrary camera poses. Early works [7, 25] injected pose embeddings into diffusion models to achieve this control, but lacked explicit geometric alignment. Subsequent methods [9, 19] improved consistency via self-supervised warping, 3D-informed point cloud conditioning, achieving more precise viewpoint control. Recent work [5] bridges the two paradigms by jointly training on multiple viewpoints, and improved spatial-temporal consistency with pose-conditioned memory retrieval. Despite these advances, most approaches rely on supervised learning or one-step consistency objectives and remain open-loop, with no mechanism to evaluate or optimize physical correctness.

2.2. Reward Learning for Post-Training

Reinforcement-based post-training has become central to aligning large generative models. In language systems, Reinforcement Learning from Human Feedback (RLHF) [18] and Reinforcement Learning with Verifiable Rewards (RLVR) [14] replace imitation with objective-driven alignment, while Group Relative Policy Optimization (GRPO) [21] stabilizes learning via stochastic rollouts comparative updates. Extensions to vision, such as Dance-GRPO [27], demonstrate that rewards on visual quality can fine-tune video diffusion models effectively.

Building on this foundation, RLWG adapts RLVR to world modeling, optimizing pretrained world models using physically verifiable rewards including cycle-consistency, depth reprojection, and trajectory stability. *GrndCtrl* instantiates RLWG as a multi-objective GRPO over grounded rewards. This process enforces geometric coherence without human supervision, significantly reducing long-horizon drift. Parallel advances in 3D perception emphasize similar constraints: VGGT [24] and MapAnything [13] predict depth and camera pose for consistent scene reconstruction, while SpaTracker [26] integrates rigidity priors for robust 3D tracking. Together, these efforts point toward reward-informed geometry as a unifying principle, where physical correctness acts as an alignment signal bridging generative modeling, simulation, and control.

3. GrndCtrl

3.1. Problem Definition

We consider a pretrained *video world model* W_θ , a policy parameterized by θ , that predicts future observations conditioned on a visual history and optionally actions. Let x_0 denote the observed frame and $a_{0:T-1} = (a_0, \dots, a_{T-1})$ the associated control inputs. The model samples a rollout $\hat{x}_{1:T} \sim W_\theta(\cdot | x_0, a_{0:T-1})$, where $a_t = (R_t, \mathbf{t}_t)$ controls translation $\mathbf{t}_t \in \mathbb{R}^3$ and rotation $R_t \in \text{SO}(3)$. Our goal is to *post-train* W_θ using self-supervised reinforcement learning to improve the *spatial coherence* and *embodied reliability* of its rollouts.

To obtain *verifiable* feedback without supervision, we use a frozen *feed-forward 3D evaluator* \mathcal{E} that provides relative pose estimates $(\Delta R_{1:T}, \Delta \mathbf{t}_{1:T})$ and per-frame depth maps D_t , where $\Delta R_t \in \text{SO}(3)$ and $\Delta \mathbf{t}_t \in \mathbb{R}^3$. Additionally, we obtain feedback on the visual and motion quality of the overall video using a frozen *feed-forward video evaluator* \mathcal{V} that provides overall visual quality scoring.

Our objective is to optimize θ such that sampled rollouts maximize a set of verifiable rewards $U(\hat{x}_{1:T})$ comprising translation (r_{trans}), rotation (r_{rot}), depth temporal reprojection (r_{dtr}), and video quality (r_v). These rewards are constructed from the 3D evaluator \mathcal{E} and video evaluator \mathcal{V} , and are defined in detail in Section 3.2.

3.2. Verifiable Self-Supervised Rewards

Let \mathcal{T} denote the set of evaluated timesteps. Each reward term measures a distinct aspect of spatial and temporal consistency.

(1) Translation Reward. We compute the Euclidean deviation in translation:

$$\mathbf{e}_t^{\text{trans}} = \Delta \mathbf{t}_t - \mathbf{t}_t. \quad (1)$$

and define the sum of mean squared trajectory error and squared final error as translation reward:

$$r_{\text{trans}} = - \left(\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \|\mathbf{e}_t^{\text{trans}}\|_2^2 + \|\mathbf{e}_{|\mathcal{T}|}^{\text{trans}}\|_2^2 \right). \quad (2)$$

When metric scale is ambiguous, a normalization factor is applied from the evaluator’s scale estimate.

(2) Rotation Reward. We compute the minimum angular deviation between predicted and evaluator rotations:

$$\mathbf{e}_t^{\text{rot}} = \arccos \left(\frac{\text{tr} (\Delta R_t \cdot R_t^T) - 1}{2} \right), \quad (3)$$

and define the axis-angle cumulative error as rotation reward:

$$r_{\text{rot}} = - \sum_{t \in \mathcal{T}} \mathbf{e}_t^{\text{rot}}. \quad (4)$$

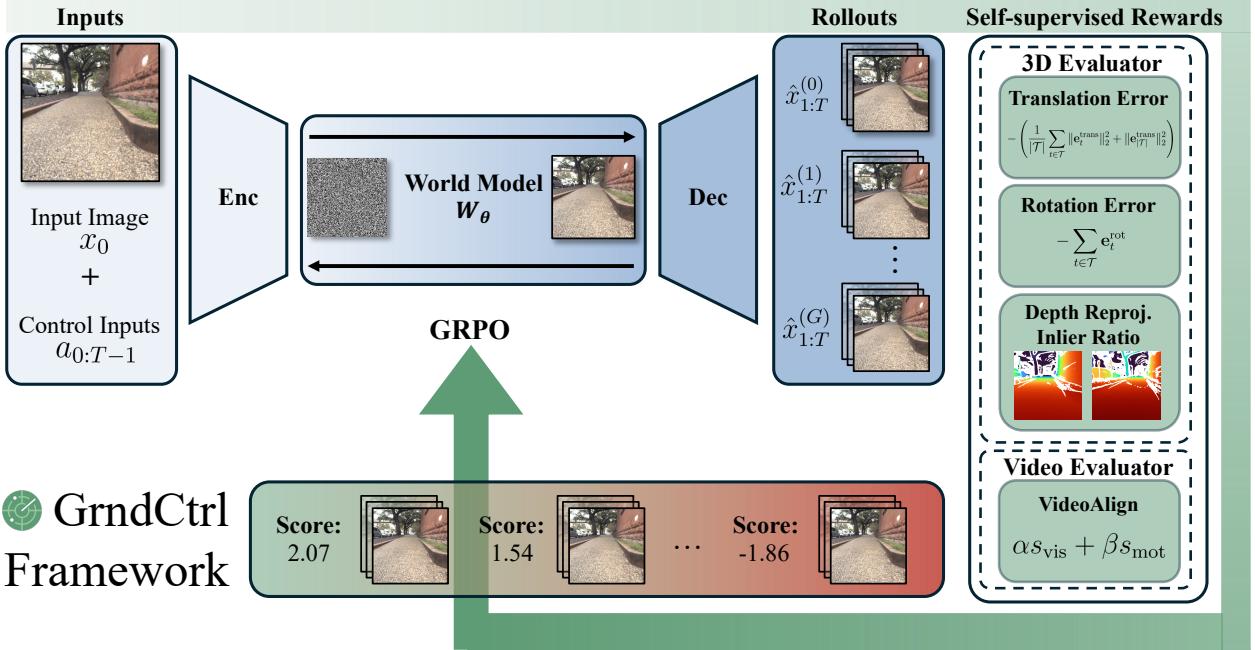


Figure 2. *GrndCtrl* framework architecture. Given conditioning context $c = (x_0, a_{0:T-1})$, the world model generates multiple stochastic rollouts $\{\hat{x}_{1:T}^{(i)}\}_{i=1}^G$. Frozen evaluators compute verifiable rewards for each rollout. Relative advantages are calculated within each group, and GRPO updates model parameters using a clipped policy gradient objective regularized toward the pretrained model, favoring physically consistent rollouts.

(3) Depth Temporal Reprojection Reward. We adopt the **depth inlier ratio** from MapAnything [13] evaluations as a verifiable geometric reward for depth temporal reprojection. For each pixel $p \in \Omega$, define the reprojected correspondence and expected depth via the evaluator geometry

$$(\hat{p}, d_{t \rightarrow t+1}^{\text{exp}}(p)) = \Phi(p; D_t, \Delta R_t, \Delta \mathbf{t}_t, K_t, K_{t+1}), \quad (5)$$

where Φ back-projects p using D_t , applies $(\Delta R_t, \Delta \mathbf{t}_t)$, and projects into frame $t+1$. The per-pair depth inlier ratio (threshold $\gamma=0.0103$) is

$$\text{DTRI}_t^{(\gamma)} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \mathbf{1} \left[\frac{|D_{t+1}(\hat{p}) - d_{t \rightarrow t+1}^{\text{exp}}(p)|}{d_{t \rightarrow t+1}^{\text{exp}}(p)} < \gamma \right]. \quad (6)$$

We use the average inlier ratio as the reward:

$$r_{\text{dtr}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \text{DTRI}_t^{(\gamma)}. \quad (7)$$

(4) Video Quality Reward. We use the frozen VideoAlign [15] as evaluator \mathcal{V} , which returns three sequence-level scores in $[0, 1]$ —visual quality s_{vis} , motion quality s_{mot} , and text alignment s_{txt} —for a rollout $\hat{x}_{1:T}$ (optionally conditioned on a prompt y):

$$(s_{\text{vis}}, s_{\text{mot}}, s_{\text{txt}}) = \mathcal{V}(\hat{x}_{1:T}; y). \quad (8)$$

Our visual reward uses only visual and motion quality as a convex combination,

$$r_v = \alpha s_{\text{vis}} + \beta s_{\text{mot}}, \quad \alpha, \beta \geq 0, \quad \alpha + \beta = 1, \quad (9)$$

with $\alpha = \beta = \frac{1}{2}$ by default.

3.3. GRPO for RLWG Post-Training

For each *conditioning context* $c = (x_0, a_{0:T-1})$, a group of G candidate rollouts $\{\hat{x}_{1:T}^{(i)}\}_{i=1}^G$ is sampled from W_θ . Each rollout is evaluated by every reward from the verifiable reward set $U(\hat{x}_{1:T})$, obtaining $\{r_{\text{trans}}^{(i)}, r_{\text{rot}}^{(i)}, r_{\text{dtr}}^{(i)}, r_v^{(i)}\} = U(\hat{x}_{1:T}^{(i)})$. We compute the normalized reward for every verifiable reward $\tilde{r}^{(i)}$, and obtain a multi-objective normalized group advantage A_i :

$$\tilde{r}^{(i)} = \frac{r^{(i)} - \text{mean}(r^{(1)}, \dots, r^{(G)})}{\text{std}(r^{(1)}, \dots, r^{(G)})}, \quad A_i = \frac{\tilde{r}^{(i)} - \text{mean}(\tilde{r}^{(1)}, \dots, \tilde{r}^{(G)})}{\text{std}(\tilde{r}^{(1)}, \dots, \tilde{r}^{(G)})}, \quad (10)$$

The GRPO objective optimizes θ using the clipped surrogate:

$$J(\theta) = \mathbb{E} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{T} \sum_{t=1}^T \min (\rho_{t,i} A_i, \text{clip}(\rho_{t,i}, 1-\epsilon, 1+\epsilon) A_i) \right], \quad (11)$$

where ϵ is the clip ratio, $\rho_{t,i}$ is the per-step likelihood ratio under the current and reference policies. This formulation

stabilizes policy optimization for continuous video generation tasks. An illustration of the detailed *GrndCtrl* framework is shown in Figure 2. To form each group for GRPO in practice, we generate multiple stochastic rollouts.

Groups are instantiated by sampling multiple candidate rollouts from the diffusion generator under controlled stochasticity. Diffusion sampling can be formulated as a reverse-time *stochastic differential equation* (SDE) or, equivalently in time-marginals, as a *probability-flow ordinary differential equation* (ODE) [1, 22]. Let $x_t \in \mathbb{R}^d$ denote the latent state at time $t \in [0, 1]$ (noisiest at $t=1$, data manifold at $t \rightarrow 0$). A standard variance-preserving forward SDE is

$$dx_t = -\frac{1}{2}\beta(t)x_t dt + \sqrt{\beta(t)}dw_t, \quad (12)$$

with w_t as a standard Brownian motion and $\beta(t) > 0$ as the noise rate. The corresponding *reverse* SDE is

$$dx_t = \left(-\frac{1}{2}\beta(t)x_t - \beta(t)\nabla_x \log p_t(x_t) \right) dt + \sqrt{\beta(t)}d\bar{w}_t, \quad (13)$$

which is stochastic due to the $d\bar{w}_t$ term and uses the learned score $\nabla_x \log p_t$, where p_t is the forward SDE solution distribution of x_t . The *probability-flow ODE*, which shares the same p_t as (13) but is deterministic, is

$$\frac{dx_t}{dt} = -\frac{1}{2}\beta(t)x_t - \frac{1}{2}\beta(t)\nabla_x \log p_t(x_t). \quad (14)$$

In practice, sampler stochasticity is controlled by $\eta \in [0, 1]$ (ODE limit at $\eta=0$; SDE-style sampling with per-step Gaussian perturbations at $\eta=1$). For each context c , we draw G rollouts $\{\hat{x}_{1:T}^{(i)}\}_{i=1}^G$ with identical conditioning and independent noise governed by η , providing the within-group diversity required by GRPO.

4. Experimental Setup

4.1. Datasets

We train and evaluate our methods on three datasets spanning diverse embodiments and scenarios. **CODa** [28] is a campus navigation dataset collected on wheeled robots with pseudo-ground truth poses. **SCAND** [10] is a social navigation dataset also collected on campus, featuring embodiments of both a wheeled robot and a quadruped. **City-Walk** [16] is an egocentric urban navigation dataset of a person walking in crowded city streets, collected from in-the-wild YouTube city walking videos and reprocessed with MapAnything [13] to obtain pose estimates. We subsample each dataset to 2k non-overlapping 13-frame sequences of pose-action pairs for training, ensuring temporal diversity and avoiding data leakage.

4.2. Baseline Model

We obtain the baseline pretrained world model W_θ via supervised fine-tuning (SFT) on Cosmos-Predict2-2B-Video2World [17], a diffusion model with a latent VAE

backbone and temporal attention. We use a modified version of its action-conditioned video predictor post-training pipeline, adapting the action space from 7 to 6 degrees of freedom to match our navigation setting: $(x, y, z, \text{roll}, \text{pitch}, \text{yaw})$ as 6×12 action embeddings. We initialize the action embeddings randomly while fine-tuning the entire DiT backbone for 20k steps with an effective batch size of 64. Training follows the EDM framework [11, 12], using the weighted expectation of denoising score matching loss over noise levels. We perform full SFT over all DiT backbone parameters keeping the visual encoder and decoder frozen, as our experiments showed that only full SFT or full-size LoRA yielded meaningful changes in motion dynamics. Additional implementation details are provided in Appendix.

4.3. GRPO Post-Training

Post-training applies GRPO with self-supervised verifiable rewards as described in Sec. 3.3. We use MapAnything [13] as \mathcal{E} to obtain rewards r_{trans} , r_{rot} , and r_{dtr} . We use VideoAlign [15] as \mathcal{V} to obtain the reward r_v . We perform an ablation study of combinatorial rewards as defined in Sec. 3.2 to evaluate multi-objective GRPO. For each reward configuration, we train for 100 steps with an effective batch size of 8, generating $G = 8$ stochastic rollouts $\{\hat{x}_{1:T}^{(i)}\}_{i=1}^G$ per conditioning context $c = (x_0, a_{0:T-1})$. To obtain diverse rollouts, we use the same initial noise for reverse-SDE diffusion with the same starting frame x_0 and action trajectory $a_{0:T-1}$, but inject stochastic Brownian noise at each diffusion step as described in Sec. 3.3. We compute per-step likelihood ratios only over the first 60% of diffusion timesteps to focus training on the most relevant denoising steps to improve training stability. Additional training details are provided in Appendix.

4.4. Evaluation Regimes

We evaluate on three regimes that progressively test generalization capabilities:

- **Seen:** Start frames and scene domains seen during SFT with matching action distributions. This regime tests in-distribution performance.
- **Counterfactual:** Scenes seen during SFT but with counterfactual actions (e.g., mirrored or directionally inverted action sequences). This regime tests the model’s ability to extrapolate geometric structure to novel motion patterns.
- **Unseen:** Both scenes and actions novel at test time, with scenes from different domains and action sequences with different motion characteristics. This regime tests full generalization to unseen scenarios.

4.5. Metrics

We report four metrics averaged across an evaluation set of 200 non-overlapping sequences for each regime: Trans-

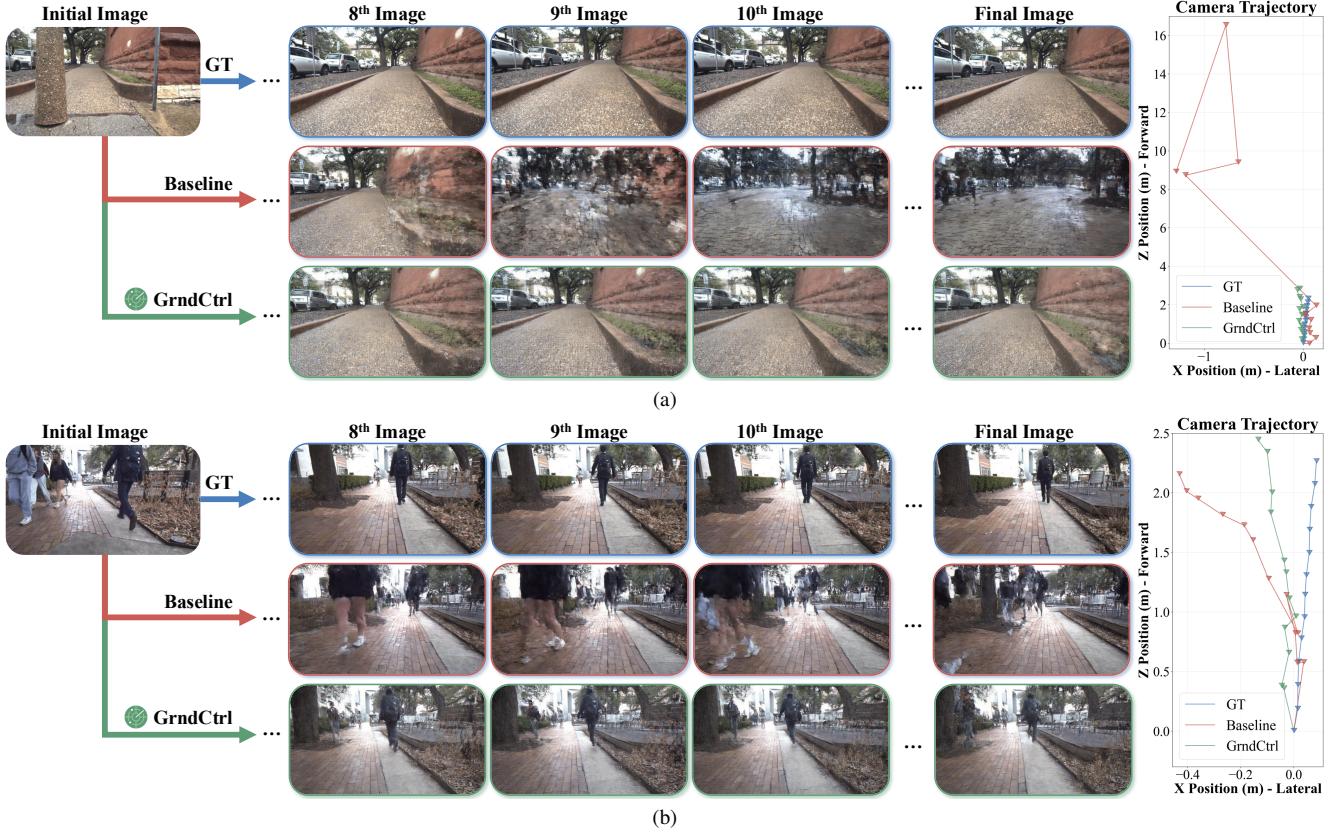


Figure 3. Qualitative results: (a) *GrndCtrl* mitigates scene drift on counterfactual rollouts, maintaining spatial coherence where baseline diverges. (b) *GrndCtrl* successfully follows directionally inverted actions, generating geometrically consistent rollouts where baseline fails.

lation error ($T = -r_{\text{trans}}$) in meters, Rotation error ($R = -r_{\text{rot}}$) in radians, Video Quality ($V = r_v$) combining visual and motion quality scores, and Depth Temporal Reprojection Inlier Ratio ($\text{DTRI} = r_{\text{dtr}}$) as a percentage. Lower values indicate better performance for T and R , while higher values are better for V and DTRI .

5. Analysis and Discussion

5.1. Impact of *GrndCtrl* on World Model Failures

O1: Baselines show poor counterfactual performance but good generalization within familiar motion manifolds. Table 1 reports quantitative results on CODa, SCAND, and CityWalk datasets across three evaluation regimes. Across all datasets, the baseline performs well in Seen but degrades significantly in Counterfactual, indicating limited transfer to out-of-distribution action sequences. On CODa, baseline translation error increases by 24% from Seen to Counterfactual, while on SCAND the degradation is more severe, increasing by 70%. On CityWalk, the increase is more modest at 12%. Unseen performance, however, remains comparable to Seen across datasets, suggesting that the pretrained model does generalize within familiar motion

manifolds. This contrast in performance between Counterfactual and Unseen indicates a fundamental limitation in pretrained video world model’s ability to extrapolate geometric structure to counterfactual rollouts, a key property expected of grounded world simulators.

O2: Translation and rotation rewards improve spatial alignment with largest gains under counterfactual motion. Introducing translation and rotation rewards r_{trans} and r_{rot} (**T+R**) improves spatial alignment across all datasets. On CODa, T+R decreases translation error by 20% in Seen and by 29% in Counterfactual relative to baseline, while rotation error improves by 19% in Seen. On SCAND, the improvements are substantial: translation error decreases by 15% in Seen and by 20% in Counterfactual, with rotation error improving by 4% in Seen. CityWalk shows the strongest counterfactual gains, with Counterfactual translation error dropping by 63% relative to baseline, while Seen improves by 24%. These results demonstrate that explicit pose-based feedback enhances motion consistency and stabilizes rollouts $\hat{x}_{1:T}$ across diverse embodiments. Notably, the improvement in Counterfactual highlights that verifiable motion alignment encourages general-

Method	Seen				Counterfactual				Unseen			
	$T \downarrow$	$R \downarrow$	$V \uparrow$	DTRI \uparrow	$T \downarrow$	$R \downarrow$	$V \uparrow$	DTRI \uparrow	$T \downarrow$	$R \downarrow$	$V \uparrow$	DTRI \uparrow
CODa [28]												
Baseline [17]	57.8	1.77	7.40	38.9	71.5	1.55	7.41	39.1	56.9	1.71	7.40	38.3
+T+R	46.4	1.44	7.32	38.4	50.5	1.53	7.34	38.7	54.3	1.75	7.36	39.3
+T+R+DTRI	65.7	1.74	7.43	37.0	57.7	1.86	7.42	36.8	42.6	1.74	7.40	37.1
+T+R+DTRI+V	39.9	1.27	7.35	37.5	40.7	1.42	7.34	37.4	31.0	1.53	7.37	38.0
SCAND [10]												
Baseline [17]	186.3	3.76	7.16	23.6	315.9	4.24	7.13	21.4	117.0	4.02	6.99	18.4
+T+R	158.2	3.61	7.19	23.7	251.2	4.34	7.18	21.7	131.1	3.95	7.04	19.1
+T+R+DTRI	157.9	3.65	7.10	22.1	288.6	4.45	7.17	20.1	118.6	4.07	7.03	17.9
+T+R+DTRI+V	133.4	3.30	7.11	24.5	220.1	4.23	7.08	22.8	123.4	3.62	6.98	19.4
CityWalk [16]												
Baseline [17]	11.7	3.13	7.96	46.9	13.1	3.27	7.94	47.4	20.8	4.47	7.90	44.5
+T+R	8.9	3.31	7.90	44.9	4.8	4.42	7.91	45.6	10.2	3.47	7.87	42.8
+T+R+DTRI	8.4	3.36	7.84	43.5	4.7	4.40	7.83	44.1	10.9	3.68	7.79	41.4
+T+R+DTRI+V	8.8	3.37	7.84	42.6	4.7	4.37	7.85	43.3	9.9	3.74	7.80	40.8

Table 1. Quantitative evaluation across three datasets (**CODa**, **SCAND**, **CityWalk**) and three regimes: **Seen**, **Counterfactual**, and **Unseen**. We compare baseline against progressive reward combinations (T+R, T+R+DTRI, T+R+DTRI+V). *GrndCtrl* achieves substantial improvements. Metrics: T (Translation Error, m), R (Rotation Error, rad), V (Video Quality), DTRI (Depth Temporal Reprojection Inliers).

ization to directionally inverted action sequences $a_{0:T-1}$, an essential feature for embodied reasoning.

Table 2 further demonstrates that GRPO training systematically improves model reliability: the baseline model shows high variance across both translation and rotation errors, indicating unstable rollouts sensitive to diffusion noise. With *GrndCtrl* training, both mean errors and variance decrease substantially: at 200 iterations, translation error means reduce by 77% relative to baseline across experiments, with standard deviations reduced by 75%. Rotation error also improves, with means reducing by 39% and standard deviations reduced by 32%, achieving reliable and consistent rollouts across all evaluation regimes.

O3: Depth reward enforces local coherence but trades off global alignment. Adding the depth reprojection reward r_{dtr} (**T+R+DTRI**) enforces local geometric coherence but introduces a trade-off in global rollout alignment. On CODa, compared to T+R, translation error increases by 42% in Seen and by 14% in Counterfactual, while Unseen benefits substantially with a 22% improvement. On SCAND, Seen shows minimal change, Counterfactual degrades by 15%, while Unseen improves by 10%. CityWalk shows a different pattern where DTRI maintains or slightly improves performance across most regimes. This suggests

the depth reward enforces short-horizon consistency and local geometric smoothness, with benefits most apparent in unseen scenarios.

O4: Full reward set produces most balanced and robust performance. Incorporating perceptual feedback through the full reward set (**T+R+DTRI+V**) produces the most balanced and robust performance across all datasets. On CODa, the full objective achieves translation error reductions of 31% (Seen), 43% (Counterfactual), and 45% (Unseen) relative to baseline, with rotation error improving by 28% (Seen), 8% (Counterfactual), and 11% (Unseen). On SCAND, translation error reduces by 28% (Seen), 30% (Counterfactual), and 5% (Unseen), while CityWalk shows consistent improvements with Counterfactual achieving a 64% reduction. The full objective recovers and extends the rollout accuracy of T+R while preserving local stability from DTRI. Perceptual alignment via video-based evaluators acts as a long-horizon stabilizer, promoting rollouts that are both physically consistent and visually coherent. Figure 3 showcases qualitative comparisons of *GrndCtrl* rollouts $\hat{x}_{1:T}$ against the baseline. *GrndCtrl* significantly mitigates the scene drift failures frequently observed when queried with mirrored input action sequences $a_{0:T-1}$, and improves rollout consistency.

Method	Seen		Counterfactual		Unseen	
	$T \downarrow$	$R \downarrow$	$T \downarrow$	$R \downarrow$	$T \downarrow$	$R \downarrow$
Baseline [17]	73.2 ± 243.7	2.38 ± 3.88	75.8 ± 253.9	2.38 ± 3.90	71.2 ± 251.2	2.88 ± 4.28
<i>GrndCtrl</i> T+R 100	72.0 ± 283.6	1.95 ± 3.38	75.9 ± 311.7	1.85 ± 3.22	58.4 ± 231.1	2.57 ± 4.08
<i>GrndCtrl</i> T+R 150	26.7 ± 101.5	1.54 ± 2.84	24.8 ± 99.1	1.53 ± 2.80	26.5 ± 104.3	2.08 ± 3.33
<i>GrndCtrl</i> T+R 200	18.4 ± 68.1	1.40 ± 2.49	16.8 ± 63.0	1.36 ± 2.57	16.3 ± 56.5	1.97 ± 3.11

Table 2. Reliability analysis showing error statistics (mean \pm standard deviation) across multiple stochastic rollouts for different GRPO iterations. Baseline exhibits high variance, while GRPO training progressively reduces both mean errors and variance, achieving consistent rollouts. Metrics: T (Translation Error, m), R (Rotation Error, rad).

5.2. Training Insights and Stability

I1: Pretraining bias requires full fine-tuning for meaningful motion adaptation. Cosmos-Predict2’s pretraining on videos with mostly short clips and slight forward movements biases the model toward scene stability over dynamics. This pretraining bias explains why we observed negligible changes in motion behavior when using parameter-efficient methods like LoRA (except full-size variants), suggesting that pretrained world models may require substantial capacity updates to adapt to navigation-specific action distributions.

I2: Early diffusion timesteps are most critical for GRPO training stability. *GrndCtrl* training is significantly more memory intensive than SFT, as it requires maintaining computation graphs over all diffusion steps to track per-timestep log-probabilities for likelihood ratio computation. Early timesteps are primarily responsible for content denoising while later timesteps refine details with exponentially smaller likelihood values, making the 60% timestep limitation both memory-efficient and beneficial for training stability by focusing on the most relevant denoising steps.

I3: *GrndCtrl* requires sufficient reward variance and careful checkpoint selection. Most critically, *GrndCtrl* training stability depends on having sufficient reward variance across stochastic rollouts. *GrndCtrl* focuses optimization on the reward component with highest variance within each group, which is beneficial for addressing scene drift when some rollouts fail visual odometry while others succeed. However, when checkpoints are overfitted or undertrained, producing uniformly poor or uniformly good rollouts, the normalization step in advantage calculation can amplify subtle differences, leading to degraded training where the model learns arbitrary patterns that add noise to predictions. We mitigate this through KL regularization toward the pretrained model and careful checkpoint selection, ensuring the baseline checkpoint generates distinguishably good and bad rollouts before applying GRPO.

6. Limitations

RLWG introduces a new paradigm for grounding world models using verifiable rewards, and our study focuses on establishing core principles rather than exhaustively scaling the approach. Our experiments use moderate budgets and limited datasets, but already demonstrate that meaningful grounding can be achieved efficiently. A practical constraint of *GrndCtrl* is its reliance on sufficient rollout variance for stable relative-policy optimization; understanding the variance dynamics of reward-aligned world model training remains an open direction. Additionally, while RLWG optimizes a multi-objective reward, we do not explore alternative weighting strategies or adaptive weighting schemes, which could further shape the trade-off between global alignment, local consistency, and visual fidelity. Investigating principled multi-reward weighting and larger-scale training holds promise for pushing RLWG toward increasingly stable, generalizable, and physically grounded world models.

7. Conclusions

Our work demonstrates that **RLWG** enables pretrained video world models to be effectively grounded in physical structure through self-supervised post-training, without requiring human labels or external simulators. Instantiated as *GrndCtrl*, the key insight is that verifiable geometric and perceptual rewards, when optimized via relative policy methods, systematically improve spatial coherence while preserving the visual quality that makes these models powerful priors. The most significant improvements emerge under counterfactual scenarios. *GrndCtrl* achieves substantial improvements under counterfactual rollouts, with up to 64% reduction in translation error, demonstrating that reward-based alignment addresses a fundamental limitation of current world models: their tendency to prioritize visual plausibility over structural consistency. By explicitly optimizing for both, **RLWG** enables models that are not only visually coherent but also geometrically consistent, opening new possibilities for reliable long-horizon planning and control in real-world environments.

Acknowledgements

We thank NVIDIA and the authors of Cosmos-Predict2 [17] for publicly releasing their code and checkpoints, upon which we build our post-training pipeline. We thank Nikhil Keetha and collaborators for MapAnything [13], and the authors of VideoAlign [15], for making their code and checkpoints publicly available, enabling our verifiable self-supervised rewards and evaluation metrics. We also thank Amir Bar and collaborators for releasing Navigation World Models [4] code and checkpoints and for kindly responding to our inquiry email. Finally, we thank Cherie Ho for early-stage discussions and Jay Karhade for discussions on geometric consistency.

References

- [1] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2025. 5
- [2] Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 2
- [3] Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction, 2025. 3
- [4] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models, 2025. 2, 3, 9, 11
- [5] Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation, 2025. 3
- [6] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024. 3
- [7] 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. 3
- [8] Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 9(1):49–56, 2023. 11
- [9] Bu Jin, Weize Li, Baihan Yang, Zhenxin Zhu, Junpeng Jiang, Huan ang Gao, Haiyang Sun, Kun Zhan, Hengtong Hu, Xueyang Zhang, Peng Jia, and Hao Zhao. Posepilot: Steering camera pose for generative world models with self-supervised depth, 2025. 2, 3
- [10] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warrell, Soeren Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation, 2022. 5, 7, 11
- [11] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. 5
- [12] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models, 2024. 5
- [13] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kotschieder. Mapanything: Universal feed-forward metric 3d reconstruction, 2025. 3, 4, 5, 9, 11, 15
- [14] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. 2, 3
- [15] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025. 4, 5, 9, 15
- [16] Xinhao Liu, Jintong Li, Yicheng Jiang, Niranjan Sujay, Zhicheng Yang, Juexiao Zhang, John Abanes, Jing Zhang, and Chen Feng. Citywalker: Learning embodied urban navigation from web-scale videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6875–6885, 2025. 5, 7, 11
- [17] NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefański, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. 2, 5, 7, 8, 9
- [18] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob

- Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. 3
- [19] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3
- [20] Dhruv Shah, Benjamin Eysenbach, Gregory Kahn, Nicholas Rhinehart, and Sergey Levine. Rapid exploration for open-world navigation with latent goal models. *arXiv preprint arXiv:2104.05859*, 2021. 11
- [21] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. 2, 3
- [22] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. 5
- [23] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines, 2025. 3
- [24] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3
- [25] 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. 3
- [26] Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. Spatialtrackerv2: 3d point tracking made easy, 2025. 3
- [27] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing grp on visual generation, 2025. 3
- [28] Arthur Zhang, Chaitanya Eranki, Christina Zhang, Ji-Hwan Park, Raymond Hong, Pranav Kalyani, Lochana Kalyanaraman, Arsh Gamare, Arnav Bagad, Maria Esteva, et al. Towards robust robot 3d perception in urban environments: The ut campus object dataset. *arXiv preprint arXiv:2309.13549*, 2023. 5, 7, 11
- [29] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning, 2025. 3
- [30] Chunling 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. 3



GrndCtrl: Grounding World Models via Self-Supervised Reward Alignment

<https://rlwg-grndctrl.github.io/>

Appendix

A. Qualitative Results

We include additional counterfactual generation results of *GrndCtrl* against baseline and ground truths in SCAND [10] (Figure 4) and CityWalk [16] (Figure 5), in addition to CODa [28] (Figure 3). The counterfactual rollouts of *GrndCtrl* and baseline are expected to mirror the trajectory shown in ground truths in each scene. We observe *GrndCtrl* improves trajectory-following and mitigates scene drift in counterfactual rollouts in all three datasets.

B. Comparisons with NWM [4]

While RLWG is a post-training framework agnostic to the baseline world model, and in *GrndCtrl* we obtain the baseline with supervised fine-tuning, we acknowledge other potential candidates that perform similar tasks as an experimental baseline. This is exemplified by NWM [4], which operates a similar world models scenario in navigation tasks. However, significant differences in experimental settings between *GrndCtrl* and NWM results in infeasible quantitative comparisons:

- NWM is jointly-trained on multiple datasets, including RECON [20] and HuRoN [8], which are collected using fish-eye cameras. This results in frequent twisted artifacts in rollouts conditioned on rectilinear images such as SCAND [10]. This biases NWM to perform poorly with visual odometry metrics based on a *feed-forward 3D evaluator* such as MapAnything [13].
- *GrndCtrl* is conditioned on actions defined in 6-DOF space ($x, y, z, \text{roll}, \text{pitch}, \text{yaw}$), whereas NWM’s actions are simplified to 3-DOF space (x, y, yaw).

Since *GrndCtrl* adopts MapAnything as an evaluator to obtain rewards, we are unable to directly use NWM as our baseline pretrained world model W_θ . Nevertheless, we include qualitative samples of NWM counterfactual rollouts in Figure 6. We used the same image-action sequences from Figures 4 in SCAND, the dataset shared by both our experiments and NWM. We used 3-DOF actions as conditions, then performed the same inversion for the rest of the trajectory in 3-DOF to generate counterfactual rollouts. We observe similar failures in scene drift and trajectory-following.

C. Failure Modes

While *GrndCtrl* demonstrates meaningful grounding in world models, we identify three main failure modes.

First, the lack of pixel-level supervision in our post-training results in gradual increase in pixel-level noises in our rollouts. While we mitigate this with KL-regularization towards the pretrained model, we still observe a gradual increase in visual noises as our post-training increases overall rewards with by improving trajectory-following or mitigating scene drift in rollouts. When unconstrained in post-training iterations, the added visual noise may eventually exceed the noise tolerance of the *feed-forward 3D evaluator*, resulting in invalid rewards used for training, and the model rollouts gradually collapse to pure noises.

Second, when a set of rollouts produce similarly good or similarly bad results, our multi-objective normalized group advantage treats the slightly worse good results as negative samples, or the slightly better bad results as positive samples. This results in counterproductive learning in post-training, which causes training instability and gradual collapse to noises. We leave the investigation of rollout variance in world models as future work.

Third, we occasionally observe reward-hacking behaviors, where rewards increase despite generating noises. Due to the black-box nature of the evaluators and our pipeline stochasticity, we mitigate this by retraining.

D. Implementation Details

We first describe supervised fine-tuning of the baseline world model, followed by GRPO post-training and the evaluator setup. Table 3 summarizes the key hyperparameters used in both stages.

D.1. Supervised Fine-Tuning

We fine-tune the Cosmos-Predict2-2B Video2World backbone starting from the released 2B-720p-16fps checkpoint. The visual encoder and decoder of the latent VAE are kept frozen, and we update all parameters in the DiT backbone.

Before settling on this configuration, we experimented with several alternatives: (i) full supervised fine-tuning (all DiT parameters), (ii) LoRA with ranks 16 and 64, (iii) a full-size LoRA configuration with 2048×2048 hidden states, and (iv) training only the action embeddings. Due to Cosmos-Predict2’s large-scale pretraining on short clips with mostly mild forward motion, the model is strongly biased toward scene stability rather than rich trajectory dynamics. In practice, only full SFT and the full-size LoRA variant produced meaningful changes in motion behavior; lower-rank LoRA and action-only updates had negligible



Figure 4. Counterfactual generation example in SCAND. Baseline and *GrndCtrl* are conditioned on left-right mirrored actions, and the rollouts are expected to invert the movements of the ground truth. Baseline generation still follows similar left-forward movement as GT (trajectory-following failure), while *GrndCtrl* generation successfully follow inverted right-forward movement.

effect. For simplicity and robustness, we therefore adopt full SFT for all reported experiments.

Supervision is applied in the latent space using an MSE loss with EDM regularization. During SFT, we perform

single-step diffusion and backpropagate only through the DiT backbone, which keeps training computationally efficient. We train for 20k steps with Fully Sharded Data Parallel (FSDP) across eight A100 GPUs. All remaining optimi-



Figure 5. Counterfactual generation example in CityWalk. Baseline and *GrndCtrl* are conditioned on left-right mirrored actions, and the rollouts are expected to invert the movements of the ground truth. Baseline generation still follows similar left-turn movement as GT (trajectory-following failure), while *GrndCtrl* generation successfully follow inverted right-turn movement.

mization and diffusion hyperparameters follow the default action-conditioned Cosmos-Predict2 post-training configuration and are listed in Table 3.

D.2. GRPO Post-Training

GRPO post-training is substantially more GPU-memory intensive than SFT because it requires multiple full video roll-



Figure 6. Counterfactual generation example in SCAND. NWM and *GrndCtrl* are conditioned on left-right mirrored actions, and the rollouts are expected to invert the movements of the ground truth. NWM generation shows clear sign of scene drift in frames 1-3, ending in a more consistent but different scene than the conditional image.

Component	Hyperparameter	Value
Optimization		
SFT/GRPO	Optimizer	fused AdamW
SFT/GRPO	Learning rate	1.0×10^{-4}
SFT/GRPO	Weight decay	0.1
SFT/GRPO	Betas	(0.9, 0.99)
SFT/GRPO	Epsilon	1.0×10^{-8}
SFT/GRPO	LR schedule	constant
Diffusion and sampling		
SFT/GRPO	Precision	bfloat16
SFT	Classifier-free guidance	7.0
SFT/GRPO	Diffusion timesteps	35
SFT	EDM loss scale	10.0
SFT	EDM σ_{cond}	1.0×10^{-4}
SFT	EDM σ_{data}	1.0
SFT	EDM high- σ ratio	0.0
GRPO	Classifier-free guidance	0.0
GRPO	Timesteps backpropagated	21
Batching and hardware		
SFT/GRPO	GPUs	8 × A100
SFT	Batch size / GPU	8
SFT	Effective batch size	64
GRPO	Batch size / GPU	1
GRPO	Effective batch size	8
GRPO	Rollouts per context G	8

Table 3. Key hyperparameters for supervised fine-tuning (SFT) and GRPO post-training.

outs $\hat{x}_{1:T}$ per conditioning context while maintaining computation graphs over all diffusion steps to track per-timestep log-probabilities. To fit within memory constraints, we train with a batch size of 1 per GPU on eight A100 GPUs and compensate by sampling $G = 8$ stochastic rollouts for each context $c = (x_0, a_{0:T-1})$, which provides sufficient diversity for group-relative advantage estimation.

A single GRPO update accumulates gradients through the diffusion trajectory to compute per-step likelihood ratios. However, diffusion timesteps do not contribute equally: early steps primarily denoise the main scene content, whereas later steps refine fine-grained details and contribute exponentially smaller likelihood values. Computing likelihood ratios over all steps causes the cumulative log-likelihood to vanish and destabilizes training. In practice, we compute the per-step likelihood ratio only over the first 60% of diffusion timesteps while still running the full sampler, which improves both numerical stability and memory efficiency.

D.3. Counterfactual Actions

We define the action conditioning of our model as $(x, y, z, \text{roll}, \text{pitch}, \text{yaw})$. In practice, we use the absolute

poses with respect to the conditional image’s camera frame for each frame we generate. We follow the standard camera coordinate convention with x right, y down, and z forward.

We perform left-right mirroring of the ground truth actions to obtain the counterfactual actions. Mirroring a trajectory with respect to the image center corresponds to reflecting motion across the plane $x = 0$ and flipping the viewing direction around the optical axis. For the translational part, this reflection is $(x, y, z)^\top \mapsto (-x, y, z)^\top$, which inverts only the lateral component. For small angular magnitudes, the viewing direction can be linearized as $\mathbf{d}(\text{pitch}, \text{yaw}) \approx (\text{yaw}, \text{pitch}, 1)^\top$, so mirroring around the optical axis sends $\mathbf{d} \mapsto (-\text{yaw}, -\text{pitch}, 1)^\top$, corresponding to $(\text{pitch}, \text{yaw}) \mapsto (-\text{pitch}, -\text{yaw})$. Combining these, a 6-DOF action $(x, y, z, \text{roll}, \text{pitch}, \text{yaw})$ has the counterfactual action $(-x, y, z, \text{roll}, -\text{pitch}, -\text{yaw})$, which is the mirror of the original motion with respect to the conditioning image’s center axis.

This method of obtaining counterfactual actions ensure realistic movement of the embodiment as those in the datasets. When conditioning using arbitrary actions, the model rollouts from the pretrained model are more likely to have poorer quality.

D.4. Evaluators and Rewards

MapAnything [13] serves as our frozen 3D evaluator E , providing relative pose estimates $(\Delta R_t, \Delta \mathbf{t}_t)$ and per-frame depth maps D_t for each rollout $\hat{x}_{1:T}$. VideoAlign [15] serves as our frozen video evaluator V , providing sequence-level visual quality, motion quality, and text alignment scores. We use only visual and motion quality scores (with equal weighting $\alpha = \beta = 0.5$) for the video quality reward r_v , ignoring text alignment since our rollouts are not text-conditioned. When GPU memory is constrained, model parameters of the evaluators are temporarily offloaded to CPU, and re-loaded only when used.