Navigation World Models as Group Actions

Zijie Wang[1,2]* **Wei Zhang**[3] **Weiming Zhang**[3] **Fanqi Zhang**[1]
Xiao Tan[3] **Yipeng Qin**[4] **Guanbin Li**[1,2,5]†
[1]Sun Yat-sen University [2]Shenzhen Loop Area Institute [3]Baidu Inc. [4]Cardiff University
[5]Guangdong Key Laboratory of Big Data Analysis and Processing
wangzj75@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn

Abstract

Navigation video world models have achieved strong visual realism, but this does not ensure that their dynamics are truly governed by actions. In this work, we argue that action faithfulness should be understood through the compositional structure of actions, which in many embodied settings follows a group structure (e.g., SE(2) for navigation). Based on this insight, we formalize action-conditioned navigation world modeling as realizing a *group action* on the state space, providing a principled criterion for evaluating dynamics beyond visual quality. To operationalize this framework, we propose a unified approach that enforces identity, inverse, and composition consistency via latent-space regularization with synthesized supervision, avoiding additional data collection. We further introduce two metrics: Group-Action Consistency (GAC) and Group-Action Robustness (GAR), to evaluate structural correctness and rollout stability. Extensive experimental results show that our method consistently improves both GAC and GAR across representative video world model backbones in the navigation setting without degrading perceptual quality.

1 Introduction

Video world models are expected to serve as action-conditioned simulators for planning [1, 2, 3], decision making [4, 5, 6], and embodied intelligence [7, 8, 9]. Recent diffusion-based video models have greatly improved visual realism and long-horizon coherence [10, 11, 12, 13]. However, visual realism alone does not ensure that a model behaves as a faithful simulator. A simulator must generate dynamics that *are governed by actions*, not merely correlated with them.

In this work, we argue that the above-mentioned requirement can be understood through the compositional structure of actions. In many embodied settings, actions are not arbitrary inputs but correspond to transformations with well-defined algebraic properties. For instance, navigation actions can be interpreted as planar rigid motions, which form the Lie group $SE(2)$. Such structure imposes fundamental constraints on valid dynamics: a null action should leave the state unchanged, an action followed by its inverse should cancel, and different decompositions of the same transformation should produce consistent outcomes. These properties are intrinsic to how such structured actions govern state evolution, rather than artifacts of any specific model. Fig. 1 illustrates these consequences as observable conditions for action-faithful world modeling.

Building on this perspective, we propose a novel theoretical framework that formalizes **action-conditioned world modeling as realizing a *group action* of the action space on the underlying state space**. It provides a general principle for action-conditioned world models whose controlled transformations admit a suitable group structure, and we further instantiate it in the navigation setting

*Work done during an internship at Baidu.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

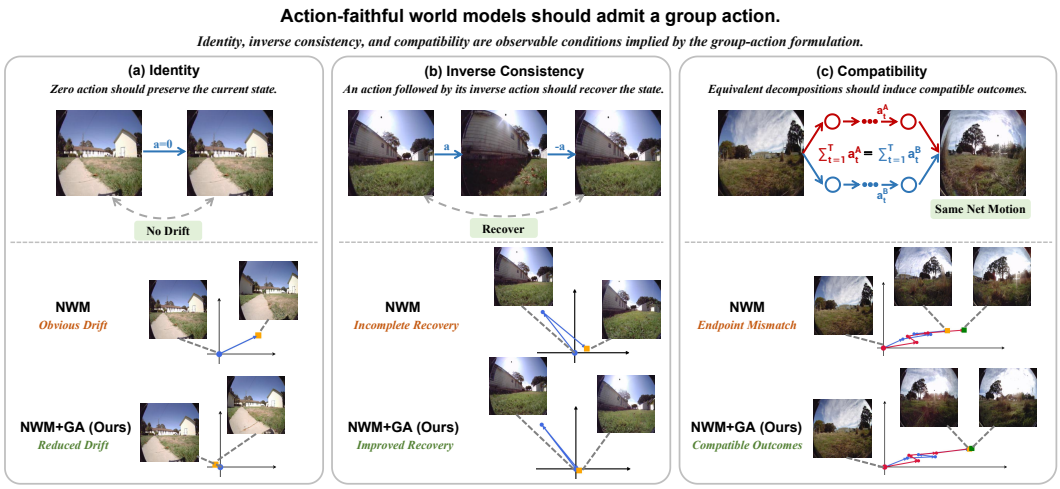


Figure 1: **Consequences of group-action-faithful dynamics.** Identity, inverse consistency, and compatibility are observable conditions implied by the group-action formulation. Representative examples illustrate that GA regularization reduces zero-action drift, improves forward–inverse recovery, and yields more compatible outcomes under equivalent action decompositions.

of video world models, where the action space is grounded in $SE(2)$ and the resulting structure can be directly analyzed through generated rollouts. This formulation introduces a new dimension for understanding and evaluating world models (beyond visual realism) by assessing whether their dynamics respect the underlying group action structure. Leveraging this perspective, we conduct a preliminary study on a recent state-of-the-art video world model, NWM [13]. Despite producing visually convincing rollouts, NWM exhibits clear violations of the group-action structure.

To address this gap, we introduce a novel framework for enforcing and evaluating action-faithful dynamics under the proposed group-action formulation. Specifically, it addresses the challenges of computational cost and lack of group-action supervision by moving the regularization from explicit state-space to efficient latent-space, where identity, inverse, and composition consistency are enforced directly on latent representations of model rollouts. To enable this, the method synthesizes supervision signals from existing trajectories by constructing zero-action, forward-inverse, and composition-equivalent action segments, avoiding the need for additional data collection. Finally, it defines two complementary evaluation metrics: Group-Action Consistency (GAC), which measures violations of structural properties in recovered state space, and Group-Action Robustness (GAR), which quantifies rollout stability under stochastic generation. Together, these components form a unified training and evaluation pipeline that operationalizes group-action structure in modern video world models.

Extensive experiments demonstrate that our framework consistently reduces both GAC and GAR errors across representative video world model backbones, while preserving perceptual generation quality. In summary, our contributions include:

- Conceptually, we formalize action-conditioned world modeling as a *group action*, providing a principled perspective that characterizes action faithfulness through identity, inverse, and composition consistency.
- Practically, we propose a unified framework that enforces these group-action constraints via latent-space regularization with synthesized supervision, along with corresponding evaluation metrics.
- Extensive experiments show that our approach consistently improves local consistency and long-horizon rollout stability across representative video world model backbones in the navigation setting, while maintaining visual generation quality.

2 Related Work

Learning-based visual navigation aims to map visual observations, goals, or task specifications to actions that accomplish navigation tasks. Representative approaches include goal-conditioned

navigation policies such as NoMaD [14], general-purpose navigation models such as GNM [15] and ViNT [16], and more recent systems that leverage large-scale data and vision-language models for long-horizon navigation and reasoning [17, 18, 19, 20]. These methods are primarily policy-centric, focusing on predicting effective actions from observations or goals. In contrast, our work studies a complementary question: whether an action-conditioned world model produces rollouts whose induced dynamics remain structurally consistent with the actions that condition them.

World models instead learn predictive dynamics for planning, control, or future-state generation. Latent world models such as Dreamer and its successors [21, 22] demonstrate that imagined rollouts in compact latent spaces can support effective decision making, while recent diffusion-based video models provide increasingly expressive backbones for controllable visual prediction. Navigation World Models [13] instantiate this paradigm for action-conditioned visual navigation, and recent systems such as Vid2World [23] and DreamZero [24] further adapt large-scale video generators toward interactive or embodied world modeling. Related efforts explore geometry-aware prediction, language-conditioned future modeling, and general-purpose embodied simulation [25, 26, 27]. These advances substantially improve the realism, controllability, and utility of generated rollouts, while our focus is specifically on whether the resulting dynamics preserve the compositional structure of navigation actions.

Evaluation of world models has accordingly expanded beyond perceptual quality and short-horizon prediction error toward action control, spatial consistency, memory, and rollout reliability [28, 29, 30, 31, 32, 33]. Recent work also studies state alignment and action-following consistency over multi-step rollouts [34, 35, 36]. These evaluations reveal important failures that may remain hidden under visual metrics, but they generally treat action inconsistency as a benchmark dimension or a rollout-level outcome. Our formulation instead starts from the compositional structure of navigation motion and derives identity preservation, inverse cancellation, and local composition consistency as observable consequences of the underlying group action. GAC probes these local structural relations, while GAR measures their rollout-level manifestation under stochastic generation. This connects trajectory-level instability to specific violations of action composition rather than treating it only as an aggregate rollout symptom. A more comprehensive review is provided in Appendix G.

3 Group-Action Formulation of World Models

A central challenge in learning world models is ensuring that predictions are structurally consistent with the underlying transformations induced by actions. Standard approaches typically optimize one-step prediction accuracy, treating actions as independent inputs. As a result, the learned dynamics may fail to respect the compositional structure inherent in many physical systems. Addressing this issue, we take a structured perspective and posit that, in many domains, actions possess an underlying algebraic structure, forming a group G equipped with a composition law, identity element, and inverses. Rather than learning an unconstrained transition function, we require the world model $f_\theta : \mathcal{S} \times G \to \mathcal{S}$ to realize a *group action* of G on the state space:

Theorem 1. *Let G be a group with identity e, composition law $\circ$, and inverses a^{-1}. Suppose the world model $f_\theta : \mathcal{S} \times G \to \mathcal{S}$ satisfies, for all $s \in \mathcal{S}$ and $a_1, a_2, a \in G$,*

$$\text{(Compatibility)} \qquad f_\theta(f_\theta(s, a_1), a_2) = f_\theta(s, a_1 \circ a_2), \tag{1}$$

$$\text{(Identity)} \qquad f_\theta(s, e) = s, \tag{2}$$

$$\text{(Inverse consistency)} \qquad f_\theta(f_\theta(s, a), a^{-1}) = s. \tag{3}$$

*Then f_θ defines a **group action** of G on $\mathcal{S}$. In particular, for any finite sequence $a_1, \dots, a_n \in G$,*

$$f_\theta(\cdots f_\theta(f_\theta(s, a_1), a_2) \cdots, a_n) = f_\theta(s, a_1 \circ \cdots \circ a_n). \tag{4}$$

Please see Appendix A for the proof of Theorem 1. Notably, in the video world models for embodied intelligence, actions often correspond to physical motions in the plane and can be naturally modeled as elements of the Lie group $SE(2)$:

Corollary 1.1. *If $G = SE(2)$, then a group-action-faithful model preserves the algebraic structure of planar rigid motions, so that the recovered motion induced by a rollout is consistent with the composed ego-motion, up to stochastic generation and state-estimation error.*

In this case, the proposed formulation characterizes action faithfulness through consistency with planar rigid-motion composition.

3.1 Instantiating Group Actions in Video World Models

Video world models predict future observations conditioned on past observations and actions. Given an initial observation $\mathbf{o}_0$ and an action sequence $\mathbf{a}_{1:T} = (\mathbf{a}_1, \ldots, \mathbf{a}_T)$, the model generates a rollout

$$\mathbf{o}_{1:T} \sim p_\theta(\mathbf{o}_{1:T} \mid \mathbf{o}_0, \mathbf{a}_{1:T}), \tag{5}$$

where each observation $\mathbf{o}_t$ is generated by an agent with state $\mathbf{s}_t$ navigating in a given scene at time t. According to Theorem 1 and Corollary 1.1, this induces state dynamics of the agent in the form:

$$\mathbf{s}_{t+1} = f_\theta(\mathbf{s}_t, \mathbf{a}_t), \tag{6}$$

where $\mathbf{s}_t = (R_t, \mathbf{p}_t)$, R_t is the rotation and $\mathbf{p}_t$ is the translation; $\mathbf{a}_t = (\Delta x_t, \Delta y_t, \Delta\theta_t)$ parameterizes a local planar ego-motion increment. Note that for clarity, we suppress stochasticity in the generation process and focus on the deterministic transition dynamics, as our goal is to characterize the structural properties of the underlying dynamics.

Estimating Agent States from Observations. We estimate the agent state $\mathbf{s}_t$ (Eq. 6) from observations $\mathbf{o}_t$ using a pose estimator $\Phi(\cdot)$:

$$\mathbf{s}_{1:T} = \Phi(\mathbf{o}_{1:T}), \tag{7}$$

Based on the recovered states, we define a distance metric in the state space as:

$$d(\mathbf{s}_i, \mathbf{s}_j) = \|\mathbf{p}_i - \mathbf{p}_j\|_2 + \alpha\, d_R(R_i, R_j), \tag{8}$$

where $\mathbf{p}_i$ and R_i denote the translation and rotation components of $\mathbf{s}_i$, respectively, $d_R(\cdot,\cdot)$ measures rotational discrepancy, and α balances the contributions of translation and rotation.

Instantiation of Group Action Conditions. Under the above ego-motion action parameterization, the composition law $\circ$ is operationally implemented by accumulating action increments over a segment, while the identity and inverse actions are represented by $\mathbf{e} = \mathbf{0}$ and $\mathbf{a}^{-1} = -\mathbf{a}$, respectively. The relation between this operational form and exact $SE(2)$ composition is discussed in Appendix A.4. Based on this structure, we instantiate the three group action conditions in Theorem 1 using the state distance defined in Eq. 8 as follows:

- *Compatibility*: $d(\mathbf{s}_T^A, \mathbf{s}_T^B) = 0$, where initial state $\mathbf{s}_0^A = \mathbf{s}_0^B$, action sequence $\mathbf{a}_{1:T}^A \neq \mathbf{a}_{1:T}^B$, and their accumulated increments satisfy $\sum_{t=1}^T \mathbf{a}_t^A = \sum_{t=1}^T \mathbf{a}_t^B$.
- *Identity*: $d(\mathbf{s}_e, \mathbf{s}) = 0$, where $\mathbf{s}_e = f_\theta(\mathbf{s}, \mathbf{0})$.
- *Inverse consistency*: $d(\mathbf{s}_{inv}, \mathbf{s}) = 0$, where $\mathbf{s}_{inv} = f_\theta(f_\theta(\mathbf{s}, \mathbf{a}), -\mathbf{a})$.

3.2 Do Native Video World Models Respect Group Action Structure?

Despite the group action structure outlined above in Sec. 3.1, it remains unclear whether existing action-conditioned video world models, which primarily emphasize visual realism, respect this structure. To investigate this, we conduct preliminary experiments. As shown in Fig. 2, our results indicate that the state-of-the-art NWM [13] violates all three group action conditions (identity, inverse, and compatibility), even though the generated videos appear visually plausible. These findings reveal a clear gap between visual realism and dynamical consistency, motivating the development of methods to enforce such structure and systematic evaluation protocols to quantify adherence to group action properties in these video world models.

4 Methodology

The group-action formulation in Sec. 3 provides a principled state-space criterion for action-faithful world modeling, specifying the conditions required for rollouts to remain consistent with motion composition. In principle, these state-space conditions can be directly used as training objectives and evaluation metrics. However, doing so introduces two key challenges:

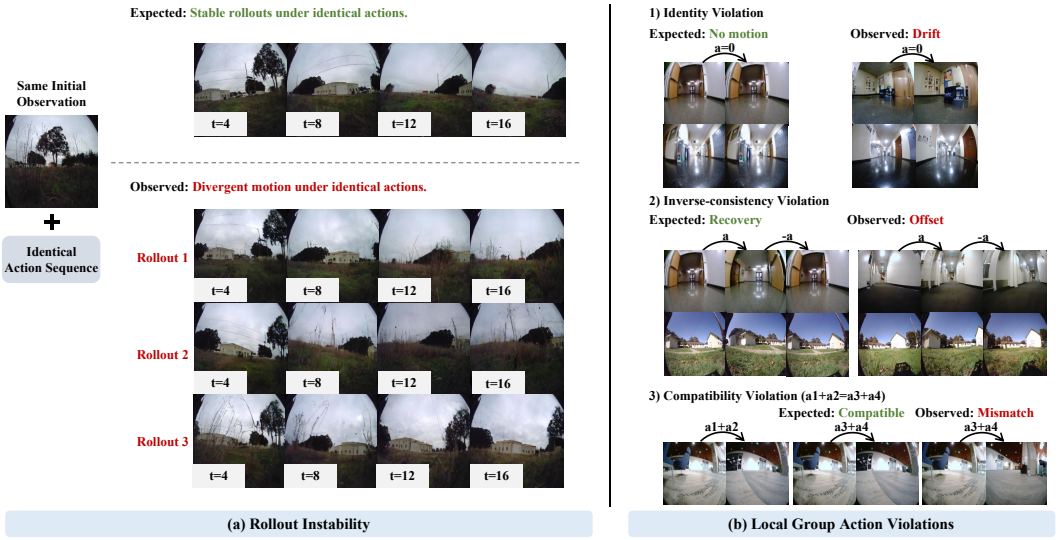


Figure 2: **Preliminary evidence of action inconsistency.** (a) Repeated stochastic rollouts from the same initial observation and identical action sequence exhibit divergent motion. (b) Controlled probes reveal violations of identity, inverse consistency, and compatibility, manifested as zero-action drift, incomplete inverse recovery, and endpoint mismatch under equivalent action decompositions.

- **Computational cost:** enforcing these constraints naively requires repeated rollout, decoding, and pose estimation, resulting in substantial computational overhead;
- **Lack of supervision:** there is no explicit training data corresponding to the identity, inverse, and compatibility conditions.

To address these challenges, we develop a unified framework comprising three novel components: (1) we enforce group-action constraints as latent regularization, directly shaping the model's internal transition dynamics without requiring explicit supervision (Sec. 4.1); (2) we synthesize group-action training signals from existing video world model datasets to approximate supervision for identity, inverse, and compatibility conditions (Sec. 4.2); (3) in addition, we introduce an evaluation protocol in the recovered state space to systematically quantify the extent to which world models adhere to group action structure (Sec. 4.3).

4.1 Latent Group-Action Supervision

The group-action conditions in Sec. 3.1 are defined over agent states. A straightforward training approach would therefore be to decode generated latents into image frames, estimate states with the pose estimator Φ, and compute group-action errors on the recovered trajectories. While conceptually faithful, this objective is impractical for optimization. It would repeatedly invoke the decoder and an external pose estimation module inside the multi-step rollout loop, leading to heavy computation and brittle gradient propagation. More critically, keeping this decode-and-estimate pipeline in the differentiable graph exceeds GPU memory for large video world models, making direct state-space optimization infeasible in our setting.

To address this issue, we propose a training objective that operates directly in the *latent space* of video world models. Let ϕ_{enc} denote the visual encoder. For an observation $\mathbf{o}_t$, its latent representation is

$$\mathbf{z}_t = \phi_{\text{enc}}(\mathbf{o}_t). \tag{9}$$

Accordingly, we define the latent group-action training objective as follows.

Latent Group-Action Training Objective. Let $D_z(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2^2$ denote the latent discrepancy. We design the three components of the training objective according to the group-action conditions (identity, inverse, and compatibility) as follows:

$$\mathcal{L}_{\text{id}} = D_z\big(\widehat{\mathbf{z}}^{\text{id}}, \mathbf{z}_t\big), \qquad \mathcal{L}_{\text{inv}} = D_z\big(\widehat{\mathbf{z}}^{\text{inv}}, \mathbf{z}_t\big), \qquad \mathcal{L}_{\text{comp}} = D_z\big(\widehat{\mathbf{z}}^{A}, \widehat{\mathbf{z}}^{B}\big), \tag{10}$$

where the latent endpoints $\widehat{\mathbf{z}}^{\mathrm{id}}$, $\widehat{\mathbf{z}}^{\mathrm{inv}}$, $\widehat{\mathbf{z}}^{A}$ and $\widehat{\mathbf{z}}^{B}$ are defined as:

$$\begin{aligned} \widehat{\mathbf{z}}^{\mathrm{id}} &= \mathrm{End}_\theta\big(\mathbf{z}_t, \mathbf{u}^{\mathrm{id}}\big), \\ \widehat{\mathbf{z}}^{\mathrm{inv}} &= \mathrm{End}_\theta\big(\mathbf{z}_t, \mathbf{u}^{\mathrm{inv}}\big), \\ \widehat{\mathbf{z}}^{A}, \widehat{\mathbf{z}}^{B} &= \mathrm{End}_\theta\big(\mathbf{z}_t, \mathbf{u}^{A}\big), \mathrm{End}_\theta\big(\mathbf{z}_t, \mathbf{u}^{B}\big). \end{aligned} \tag{11}$$

where $\mathrm{End}_\theta(\mathbf{z}_t, \cdot)$ denotes the latent endpoint obtained by rolling out the model from $\mathbf{z}_t$ under a given action segment $\mathbf{u}$ using its native transition dynamics. The overall group-action objective is therefore defined as:

$$\mathcal{L}_{\mathrm{GA}} = \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}} + \lambda_{\mathrm{inv}}\mathcal{L}_{\mathrm{inv}} + \lambda_{\mathrm{comp}}\mathcal{L}_{\mathrm{comp}}, \tag{12}$$

The revised diffusion training objective becomes:

$$\mathcal{L} = \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{GA}}\mathcal{L}_{\mathrm{GA}}. \tag{13}$$

Eq. 13 describes the full objective. To reduce the computation and memory cost of multiple free-running rollouts, we sample one constraint type from $\{\mathrm{id}, \mathrm{inv}, \mathrm{comp}\}$ in each batch and optimize the corresponding stochastic estimate of Eq. 13. Here, $\mathcal{L}_{\mathrm{diff}}$ preserves visual realism, while $\mathcal{L}_{\mathrm{GA}}$ regularizes latent dynamics. Existing video world model datasets do not contain paired trajectories that explicitly realize zero-motion preservation, forward–inverse cancellation, or alternative decompositions of the same local motion. Therefore, the corresponding group-action relations are synthesized from existing trajectories.

Remark. The latent variables represent the internal rollout states from which future observations are generated. Constraining their evolution therefore regularizes the model transition before visual decoding. Although the group-action conditions are evaluated in recovered state space, their latent-space counterparts are constructed from the same algebraic relations: identity preservation, inverse cancellation, and local composition consistency. The latent objective should therefore be viewed as an efficient surrogate for state-space regularization, rather than an exact replacement. As an endpoint-based objective, it does not provide a formal guarantee on every intermediate state along the rollout. It encourages the internal rollout dynamics to follow the same structural relations that are later measured on recovered trajectories, while avoiding the memory-prohibitive decode-and-estimate loop during training.

4.2 Group-Action Supervision Synthesis

To address the lack of supervision described above, one approach is to collect additional trajectories that explicitly realize pauses, forward–inverse cycles, or multiple decompositions of the same local displacement. While such data would provide direct supervision, it is costly to acquire and difficult to scale. We therefore propose a novel strategy to synthesize group-action supervision from existing training trajectories by leveraging the algebraic operations of the action space. Specifically, given a trajectory from an existing dataset consisting of observations and actions:

$$(\mathbf{o}_{1:T}, \mathbf{a}_{1:T-1}), \tag{14}$$

then, for any $1 \le t < T-1$, we construct action segments $\mathbf{u}^{\mathrm{id}}_{1:l}$, $\mathbf{u}^{\mathrm{inv}}_{1:2l}$ and ($\mathbf{u}^{A}_{1:l}$, $\mathbf{u}^{B}_{1:l}$), where the segment length satisfies $1 \le l \le T-t$.

Identity Action Segment. The action segment for the identity condition should induce no state change and is implemented as a zero-action segment:

$$\mathbf{u}^{\mathrm{id}}_{1:l} = (\mathbf{0}, \ldots, \mathbf{0}). \tag{15}$$

As shown in Eq. 10, this condition enforces that executing the zero-action segment $\mathbf{u}^{\mathrm{id}}_{1:l}$ preserves the latent state, keeping the endpoint close to $\mathbf{z}_t$.

Inverse Action Segment. The inverse condition requires a motion followed by its inverse to recover the starting state. Therefore, let $\mathbf{u}_{1:l} = (\mathbf{a}_t, \ldots, \mathbf{a}_{t+l-1})$ be an action segment from the original trajectory, we construct the inverse action segment as a forward–inverse cycle:

$$\mathbf{u}^{\mathrm{inv}}_{1:2l} = (\mathbf{u}_{1:l}, \hat{\mathbf{u}}_{1:l}) = (\mathbf{a}_t, \ldots, \mathbf{a}_{t+l-1}, -\mathbf{a}_{t+l-1}, \ldots, -\mathbf{a}_t), \tag{16}$$

As shown in Eq. 10, the endpoint should also be close to $\mathbf{z}_t$, as the inverse action segment cancels the forward action.

Compatibility Action Segment. The compatibility condition requires that different decompositions of the same intended local motion yield compatible endpoints. Let $\mathbf{u}^A_{1:l} = (\mathbf{u}^A_1, \dots, \mathbf{u}^A_l) = (\mathbf{a}_t, \dots, \mathbf{a}_{t+l-1})$ denote an action segment from the original trajectory. We construct an alternative decomposition $\mathbf{u}^B_{1:l} = (\mathbf{u}^B_1, \dots, \mathbf{u}^B_l)$ that:

$$\mathbf{u}^B_i = w_i \sum_{j=1}^{l} \mathbf{u}^A_j, \qquad i = 1, \dots, l. \tag{17}$$

where $\mathbf{w} = (w_1, \dots, w_l) \sim \mathrm{Dir}(\alpha)$ are non-negative weights sampled from a Dirichlet distribution. By construction, the accumulated increments are identical:

$$\sum_{i=1}^{l} \mathbf{u}^A_i = \sum_{i=1}^{l} \mathbf{u}^B_i. \tag{18}$$

Accordingly, the loss in Eq. 10 enforces that rollouts from the same starting latent under $\mathbf{u}^A_{1:l}$ and $\mathbf{u}^B_{1:l}$ produce matching endpoints.

4.3 Group-Action Metrics

While group-action conditions are enforced in latent space during training, the resulting dynamics are ultimately assessed through motion in generated videos. We therefore evaluate in the recovered state space. Following the group-action conditions defined in Sec. 3.1, we introduce two complementary metrics: *Group-Action Consistency* (GAC), which measures violations of these conditions, and *Group-Action Robustness* (GAR), which assesses rollout stability under stochastic generation.

Group-Action Consistency. Group-Action Consistency (GAC) measures the extent to which the recovered motion induced by generated rollouts satisfies the group-action conditions. Specifically, we construct a state-space analogue of Eq. 10 and evaluate each condition under a fixed family of controlled probe configurations. For identity and inverse consistency, a probe configuration is indexed by (k, l), where k denotes the number of inserted or evaluated segments and l denotes the length of each segment. For local composition consistency, the probe configuration is indexed by the local window length l. Within each configuration, errors are averaged over fixed valid starting positions and evaluation sequences, yielding configuration-level errors $\Delta^{(k,l)}_{\mathrm{id}}$, $\Delta^{(k,l)}_{\mathrm{inv}}$, and $\Delta^{(l)}_{\mathrm{comp}}$. Detailed definitions of these configuration-level errors are provided in Appendix C.5. The reported component errors are obtained by averaging over the evaluated probe configurations:

$$\bar{\Delta}_{\mathrm{id}} = \frac{1}{|\mathcal{K}_{\mathrm{id}}|} \sum_{(k,l)\in\mathcal{K}_{\mathrm{id}}} \Delta^{(k,l)}_{\mathrm{id}}, \; \bar{\Delta}_{\mathrm{inv}} = \frac{1}{|\mathcal{K}_{\mathrm{inv}}|} \sum_{(k,l)\in\mathcal{K}_{\mathrm{inv}}} \Delta^{(k,l)}_{\mathrm{inv}}, \; \bar{\Delta}_{\mathrm{comp}} = \frac{1}{|\mathcal{L}_{\mathrm{comp}}|} \sum_{l\in\mathcal{L}_{\mathrm{comp}}} \Delta^{(l)}_{\mathrm{comp}}. \tag{19}$$

These terms measure identity drift, inverse recovery error, and composition mismatch, respectively. The aggregated GAC error is defined as:

$$\mathcal{E}_{\mathrm{GAC}} = \frac{1}{3}\left(\bar{\Delta}_{\mathrm{id}} + \bar{\Delta}_{\mathrm{inv}} + \bar{\Delta}_{\mathrm{comp}}\right), \tag{20}$$

where a lower $\mathcal{E}_{\mathrm{GAC}}$ indicates stronger adherence to the group-action conditions.

Group-Action Robustness. Group-Action Robustness (GAR) measures the stability of motion under stochastic rollout generation. Given the same initial observation and identical action sequence, we sample N stochastic rollouts and recover their state trajectories $\{\mathbf{s}^{(i)}_{1:T}\}_{i=1}^{N}$. The GAR error is then defined as:

$$\mathcal{E}_{\mathrm{GAR}} = \frac{2}{N(N-1)} \sum_{1\le i<j\le N} \frac{1}{T} \sum_{t=1}^{T} d\big(\mathbf{s}^{(i)}_t, \mathbf{s}^{(j)}_t\big). \tag{21}$$

where lower $\mathcal{E}_{\mathrm{GAR}}$ indicates that repeated stochastic rollouts remain more consistent under the same action sequence. GAR measures rollout reproducibility rather than fidelity to the ground-truth trajectory.

Additional details on free-running latent rollout, constraint synthesis and sampling, action-space approximations, and the practical per-batch training objective are provided in Appendix B.

5 Experiments

5.1 Experimental Setup

Our experiments are conducted on the RECON [37], SCAND [38], and HuRoN [39] datasets, following the preprocessing protocol of NWM [13]. We report the main results on RECON, with additional results on SCAND and HuRoN provided in Appendix D. We use NWM [13] as the primary testbed and include DIAMOND [40] as an additional baseline. For GAC evaluation, we use a fixed set of controlled probes shared by all methods. For identity and inverse probes, configurations are indexed by the number of local segments k and the segment length l. Increasing k tests whether violations accumulate when the same local relation is applied repeatedly, while increasing l tests whether violations grow with the temporal extent of each segment. We vary one factor while fixing the other, using $k, l \in \{1, 3, 5\}$ for the main RECON analysis. For local composition probes, we vary the window length with $l \in \{2, 4, 6\}$, since the probe compares two decompositions of the same local motion. Within each configuration, errors are averaged over fixed valid starting positions and evaluation sequences. For GA training, the local rollout horizon is sampled from $l \sim \mathcal{U}\{1, \ldots, L\}$ with $L = 4$ by default, and one group-action constraint type is sampled per batch for efficiency. Evaluation is conducted using the metrics defined in Sec. 4.3. Additional details of dataset statistics, model adaptation, training hyper-parameters, pose recovery, GAC probe configurations, GAR rollout settings, and metric aggregation procedures are provided in Appendix C.

Table 1: Group-Action Consistency (GAC) errors on the RECON dataset. Columns 2–4 report the individual error terms, and the last column reports the aggregated GAC error E_{GAC}. "+GA" denotes applying the proposed GA regularization. E_{GAC} is computed from the mean component errors.

Method	$\bar{\Delta}_{\mathrm{id}} \downarrow$	$\bar{\Delta}_{\mathrm{inv}} \downarrow$	$\bar{\Delta}_{\mathrm{comp}} \downarrow$	$\mathcal{E}_{\mathrm{GAC}} \downarrow$
DIAMOND [40]	2.47 ± 1.16	2.38 ± 1.08	1.12 ± 0.70	1.99
DIAMOND + GA	2.25 ± 1.03	2.16 ± 0.97	0.86 ± 0.55	1.76
NWM [13]	2.10 ± 0.99	2.29 ± 0.99	0.79 ± 0.52	1.73
NWM + GA	$\mathbf{1.95 \pm 0.86}$	$\mathbf{1.95 \pm 0.84}$	$\mathbf{0.60 \pm 0.40}$	**1.50**

5.2 Group-Action Consistency Results

Table 1 reports the Group-Action Consistency (GAC) errors on RECON. Compared with their GA-regularized counterparts, both baseline models show consistently larger errors across identity, inverse, and local composition consistency. This indicates that strong visual rollout quality does not by itself guarantee adherence to the underlying group-action structure. Applying GA regularization reduces the aggregate GAC error for both DIAMOND and NWM, with gains observed across all three components. These results suggest that the proposed objective improves group-action consistency beyond a single backbone architecture. Fig. 3 provides a probe-wise view of the same effect. Across most probe settings, GA regularization shifts the error curves downward and reduces the standard deviation indicated by the marker size. The improvement is particularly clear for inverse and local composition consistency, where errors tend to increase when the model must compose multiple action-conditioned transitions. For identity consistency, the curve is less monotonic because repeated or longer zero-action segments can reinforce the stationary condition, but larger settings impose a stricter preservation requirement across more pause locations or longer pause durations. Additional GAC results and probe-wise analysis are provided in Appendices D.2 and D.3.

5.3 Group-Action Robustness Results

Table 2 reports GAR errors on RECON, with additional results on SCAND and HuRoN provided in Appendix D.1. The errors are computed from repeated rollouts ($N = 5$) generated from the same initial observation and action sequence drawn from the test set. It can be observed that the baseline NWM exhibits clear horizon-dependent degradation. In particular, its non-aligned error increases from 0.65 at 16 frames to 1.52 at 64 frames, indicating that stochastic variations accumulate into substantial global trajectory drift over longer rollouts. In contrast, GA training consistently reduces GAR error across both horizons and alignment settings, with more pronounced improvements at 64 frames where accumulated transition errors are larger. The reduction in non-aligned error indicates

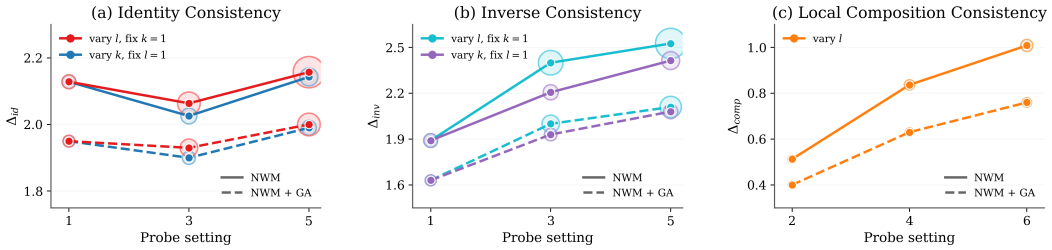


Figure 3: Group-Action Consistency (GAC) errors on RECON under different probe configurations. For identity and inverse probes, k and l denote the number and length of probed segments; for composition probes, l denotes the window length. Markers indicate mean errors, with size proportional to the standard deviation. Lower values indicate stronger group-action consistency.

Table 2: Group-Action Robustness (GAR) errors on the RECON dataset. "Aligned" removes global pose drift before computing trajectory error, while "Non-aligned" measures raw trajectory deviation.

Method	16 Frames		64 Frames	
	Aligned ↓	Non-aligned ↓	Aligned ↓	Non-aligned ↓
DIAMOND [40]	0.44 ± 0.36	0.89 ± 0.58	0.93 ± 0.67	2.48 ± 1.91
DIAMOND + GA	0.37 ± 0.29	0.74 ± 0.49	0.76 ± 0.55	2.03 ± 1.46
NWM [13]	0.32 ± 0.26	0.65 ± 0.41	0.62 ± 0.44	1.52 ± 1.27
NWM + GA	$\mathbf{0.26 \pm 0.20}$	$\mathbf{0.49 \pm 0.35}$	$\mathbf{0.46 \pm 0.33}$	$\mathbf{1.28 \pm 0.82}$

that stochastic rollouts become more consistent in their global motion under the same action sequence, rather than only after trajectory alignment. Overall, these results demonstrate that the proposed GA regularization enhances robustness to stochastic rollout variability, especially over longer horizons.

Table 3: Ablation of group-action constraints on RECON using NWM as the base model. Single-loss rows apply only the indicated group-action constraint. Full GA uses all three constraints. GAC reports local group-action consistency, while GAR reports rollout-level robustness at 64 frames. Lower is better.

Method	GAC				GAR 64 Frames	
	$\bar{\Delta}_{\text{id}}$ ↓	$\bar{\Delta}_{\text{inv}}$ ↓	$\bar{\Delta}_{\text{comp}}$ ↓	$\mathcal{E}_{\text{GAC}}$ ↓	Aligned ↓	Non-aligned ↓
Baseline	2.10 ± 0.99	2.29 ± 0.99	0.79 ± 0.52	1.73	0.62 ± 0.44	1.52 ± 1.27
$\mathcal{L}_{\text{id}}$	1.93 ± 0.87	2.18 ± 0.96	0.79 ± 0.50	1.62	0.57 ± 0.41	1.44 ± 1.16
$\mathcal{L}_{\text{inv}}$	2.03 ± 0.94	1.99 ± 0.85	0.73 ± 0.48	1.58	0.53 ± 0.38	1.38 ± 1.06
$\mathcal{L}_{\text{comp}}$	2.06 ± 0.95	2.08 ± 0.90	0.62 ± 0.41	1.59	0.50 ± 0.35	1.33 ± 0.94
Full GA	$\mathbf{1.95 \pm 0.86}$	$\mathbf{1.95 \pm 0.84}$	$\mathbf{0.60 \pm 0.40}$	**1.50**	$\mathbf{0.46 \pm 0.33}$	$\mathbf{1.28 \pm 0.82}$

5.4 Ablation Analysis

Table 3 presents an ablation of the three group-action conditions defined in Eq. 10. Each individual condition improves over the NWM baseline in both GAC and GAR, indicating that identity preservation, inverse cancellation, and compatibility enforcement each address distinct failure modes in world model dynamics. The full objective achieves the best overall performance, suggesting that the three conditions are complementary. Notably, the GAC results show that improvements are not strictly confined to the directly targeted component: optimizing inverse or compatibility also reduces other errors, reflecting coupling among the conditions through shared rollout dynamics. The GAR ablation results further show that group-action consistency improvements translate into more stable repeated rollouts. Among the individual terms, the compatibility term yields the largest gains at longer horizons, consistent with its role in multi-step dynamics. Identity and inverse constraints also contribute by reducing stationary drift and residual motion after cancellation.

Additional ablations on rollout span, loss weight, and free-running versus teacher-forced training are provided in Appendix E.

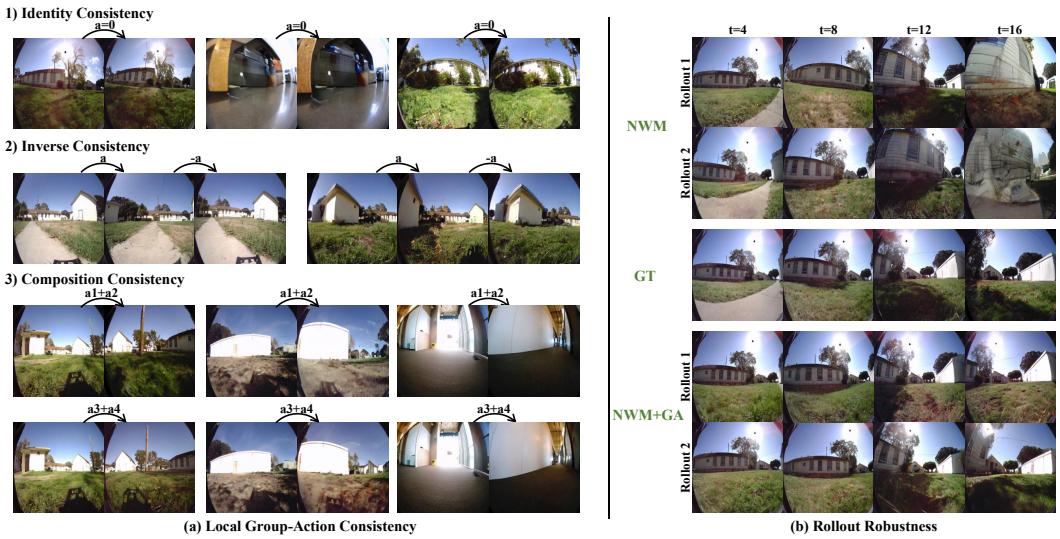


Figure 4: Qualitative results for Group-Action Consistency (GAC) and Group-Action Robustness (GAR). (a) Representative examples of identity, inverse, and compatibility probes, where GA regularization improves local action consistency. (b) Repeated stochastic rollouts from the same initial observation and identical action sequence. Although visual appearance varies across samples, the GA-regularized model exhibits more consistent motion evolution.

5.5 Qualitative Results

Fig. 4 presents qualitative results for both evaluation metrics. For GAC, the GA-regularized model better preserves stationary observations under zero action, more accurately recovers after forward–inverse segments, and produces more consistent outcomes under locally equivalent decompositions. For GAR, although samples generated from the same initial observation and action sequence may still vary in appearance, their motion evolution remains more consistent. Overall, these results corroborate the quantitative findings, showing that the proposed constraints improve action faithfulness while maintaining visual realism. Additional qualitative examples are provided in Appendix F. A complementary image-quality evaluation in Appendix D.4 further shows that these consistency gains do not come at the cost of perceptual generation quality.

6 Conclusion

In this work, we argue that consistency with the compositional structure of actions provides a principled criterion for assessing action faithfulness in navigation video world models beyond visual realism. By formalizing navigation world modeling as a group action, we provide a principled framework to diagnose and understand failures of existing video world models. We further introduce a unified method to enforce these structural constraints in latent space, along with metrics that evaluate both group action consistency and rollout stability. Extensive experimental results show that our approach consistently improves action-consistent dynamics while preserving visual quality. The current formulation applies to settings with suitable group-structured transformations and does not directly cover arbitrary or intrinsically irreversible world dynamics.

Acknowledgments and Disclosure of Funding

This paper is supported in part by the National Natural Science Foundation of China (NO. 62322608) and in part by the Shenzhen Loop Area Institute under grant FPF10120260001. The authors would like to thank National Supercomputer Center in Guangzhou for providing high performance computational resources.

References

[1] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2998–3009, 2023.

[2] Xiaodong Li, Guohui Tian, Yongcheng Cui, Xuyang Shao, and Zhiwei Wang. Grhp: Graph-fused hierarchical planning for embodied long-horizon robotic task. *Engineering Applications of Artificial Intelligence*, 165:113413, 2026.

[3] Xiang Li, Ning Yan, and Masood Mortazavi. Embodied task planning via graph-informed action generation with large language model. *arXiv preprint arXiv:2601.21841*, 2026.

[4] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li E Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534, 2024.

[5] Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11142–11152, 2025.

[6] Wenlong Liang, Rui Zhou, Yang Ma, Bing Zhang, Songlin Li, Yijia Liao, and Ping Kuang. Large model empowered embodied ai: A survey on decision-making and embodied learning. *arXiv preprint arXiv:2508.10399*, 2025.

[7] Yang Liu, Xinshuai Song, Kaixuan Jiang, Weixing Chen, Jingzhou Luo, Guanbin Li, and Liang Lin. Meia: Multimodal embodied perception and interaction in unknown environments. *arXiv preprint arXiv:2402.00290*, 2024.

[8] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics*, 2025.

[9] Kaixuan Jiang, Yang Liu, Weixing Chen, Jingzhou Luo, Ziliang Chen, Ling Pan, Guanbin Li, and Liang Lin. Beyond the destination: A novel benchmark for exploration-aware embodied question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9091–9101, 2025.

[10] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

[11] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[12] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[13] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15791–15801, 2025.

[14] Ajay Sridhar, Dhruv Shah, Catherine Glossop, and Sergey Levine. Nomad: Goal masked diffusion policies for navigation and exploration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 63–70. IEEE, 2024.

[15] Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. Gnm: A general navigation model to drive any robot. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7226–7233. IEEE, 2023.

[16] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. Vint: A foundation model for visual navigation. *arXiv preprint arXiv:2306.14846*, 2023.

[17] Xinshuai Song, Weixing Chen, Yang Liu, Vincent Chan, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[18] Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, et al. Embodied navigation foundation model. *arXiv preprint arXiv:2509.12129*, 2025.

[19] Zhangyang Qi, Zhixiong Zhang, Yizhou Yu, Jiaqi Wang, and Hengshuang Zhao. Vln-r1: Vision-language navigation via reinforcement fine-tuning. *arXiv preprint arXiv:2506.17221*, 2025.

[20] Teng Wang, Xinxin Zhao, Wenzhe Cai, and Changyin Sun. Imaginenav++: Prompting vision-language models as embodied navigator through scene imagination. *arXiv preprint arXiv:2512.17435*, 2025.

[21] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

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

[23] Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025.

[24] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

[25] Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8535–8546, 2025.

[26] Yifei Dong, Fengyi Wu, Yilong Dai, Lingdong Kong, Guangyu Chen, Xu Zhu, Qiyu Hu, Tianyu Wang, Johnalbert Garnica, Feng Liu, et al. Language-conditioned world modeling for visual navigation. *arXiv preprint arXiv:2603.26741*, 2026.

[27] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023.

[28] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024.

[29] Chun-Kai Fan, Xiaowei Chi, Xiaozhu Ju, Hao Li, Yong Bao, Yu-Kai Wang, Lizhang Chen, Zhiyuan Jiang, Kuangzhi Ge, Ying Li, et al. Wow, wo, val! a comprehensive embodied world model evaluation turing test. *arXiv preprint arXiv:2601.04137*, 2026.

[30] Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026.

[31] Ao Liang, Lingdong Kong, Tianyi Yan, Hongsi Liu, Wesley Yang, Ziqi Huang, Wei Yin, Jialong Zuo, Yixuan Hu, Dekai Zhu, et al. Worldlens: Full-spectrum evaluations of driving world models in real world. *arXiv preprint arXiv:2512.10958*, 2025.

[32] Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. Mind: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025*, 2026.

[33] Kewei Lian, Shaofei Cai, Yilun Du, and Yitao Liang. Toward memory-aided world models: Benchmarking via spatial consistency. *arXiv preprint arXiv:2505.22976*, 2025.

[34] Qimin Zhong, Hao Liao, Haiming Qin, Mingyang Zhou, Rui Mao, Wei Chen, and Naipeng Chao. Toward consistent world models with multi-token prediction and latent semantic enhancement. *arXiv preprint arXiv:2604.06155*, 2026.

[35] Zhen Li, Zian Meng, Shuwei Shi, Wenshuo Peng, Yuwei Wu, Bo Zheng, Chuanhao Li, and Kaipeng Zhang. Wildworld: A large-scale dataset for dynamic world modeling with actions and explicit state toward generative arpg. *arXiv preprint arXiv:2603.23497*, 2026.

[36] Han Yan, Zishang Xiang, Zeyu Zhang, and Hao Tang. Mwm: Mobile world models for action-conditioned consistent prediction. *arXiv preprint arXiv:2603.07799*, 2026.

[37] 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.

[38] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, 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.

[39] 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.

[40] 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.

[41] 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.

[42] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[43] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

[44] 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*, pages 586–595, 2018.

[45] 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. *Advances in Neural Information Processing Systems*, 36:50742–50768, 2023.

[46] Vaishnav Vaidheeswaran, Dilith Jayakody, Samruddhi Mulay, Anand Lo, Md Mahbub Alam, and Gabriel Spadon. Goal-conditioned reinforcement learning for data-driven maritime navigation. *arXiv preprint arXiv:2509.01838*, 2025.

[47] Wangtian Shen, Ziyang Meng, Jinming Ma, Mingliang Zhou, and Diyun Xiang. An efficient and multi-modal navigation system with one-step world model. *arXiv preprint arXiv:2601.12277*, 2026.

[48] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7606–7623, 2022.

[49] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649, 2024.

[50] Wei Xue, Mingcheng Li, Xuecheng Wu, Jingqun Tang, Dingkang Yang, and Lihua Zhang. Profocus: Proactive perception and focused reasoning in vision-and-language navigation. *arXiv preprint arXiv:2603.05530*, 2026.

[51] Wangtian Shen, Pengfei Gu, Haijian Qin, and Ziyang Meng. Effonav: An effective foundation-model-based visual navigation approach in challenging environment. *IEEE Robotics and Automation Letters*, 2025.

[52] Michele Moriconi, Stefan Laible, and Carmine Recchiuto. Foundation-model-based action selection for behavior trees in navigation. In *2025 European Conference on Mobile Robots (ECMR)*, pages 1–7. IEEE, 2025.

[53] Muhammad Tayyab Khan and Ammar Waheed. Foundation model driven robotics: A comprehensive review. *arXiv preprint arXiv:2507.10087*, 2025.

[54] Lazar Milikic, Manthan Patel, and Jonas Frey. Vld: Visual language goal distance for reinforcement learning navigation. *arXiv preprint arXiv:2512.07976*, 2025.

[55] Maeva Guerrier, Karthik Soma, Jana Pavlasek, and Giovanni Beltrame. Can vision foundation models navigate? zero-shot real-world evaluation and lessons learned. *arXiv preprint arXiv:2603.25937*, 2026.

[56] Fei Liu, Shichao Xie, Minghua Luo, Zedong Chu, Junjun Hu, Xiaolong Wu, and Mu Xu. Navforesee: A unified vision-language world model for hierarchical planning and dual-horizon navigation prediction. *arXiv preprint arXiv:2512.01550*, 2025.

[57] Zijie Wang, Jingyi Xue, Aichun Zhu, Yifeng Li, Mingyi Zhang, and Chongliang Zhong. Amen: Adversarial multi-space embedding network for text-based person re-identification. In *Chinese conference on pattern recognition and computer vision (PRCV)*, pages 462–473. Springer, 2021.

[58] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM international conference on multimedia*, pages 209–217, 2021.

[59] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *Proceedings of the 30th ACM international conference on multimedia*, pages 1984–1992, 2022.

[60] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5314–5322, 2022.

[61] Zijie Wang, Aichun Zhu, Jingyi Xue, Daihong Jiang, Chao Liu, Yifeng Li, and Fangqiang Hu. Sum: Serialized updating and matching for text-based person retrieval. *Knowledge-Based Systems*, 248:108891, 2022.

[62] Aichun Zhu, Zijie Wang, Jingyi Xue, Xili Wan, Jing Jin, Tian Wang, and Hichem Snoussi. Improving text-based person retrieval by excavating all-round information beyond color. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):5097–5111, 2024.

[63] Zijie Wang, Jingyi Xue, Xili Wan, Aichun Zhu, Yifeng Li, Xiaomei Zhu, and Fangqiang Hu. Aspd-net: Self-aligned part mask for improving text-based person re-identification with adversarial representation learning. *Engineering Applications of Artificial Intelligence*, 116:105419, 2022.

[64] Wenhao Li, Zhibin Wu, Chong Xiao, and Qiangchang Wang. Semmsa: Latent semantic-aided robust multimodal sentiment analysis with incomplete data. *arXiv preprint arXiv:2609.30238*, 2026.

[65] Guan-Nan Dong, Yumeng Yang, Zijie Wang, Aichun Zhu, Hao Wang, Yuanfei Dai, Tian Wang, and Hichem Snoussi. Taking astray domain back home for single-source domain generalizable text-to-image person retrieval. *IEEE Transactions on Image Processing*, 2026.

[66] Mingcheng Ni, Zijie Wang, Aichun Zhu, Jingyi Xue, Guan-Nan Dong, and Yong Cheng. Multi-agent consultation and uncertainty-guided voting for text-to-image person retrieval. *Engineering Applications of Artificial Intelligence*, 165:113298, 2026.

[67] Wenhao Li, Qiangchang Wang, Xianjing Meng, Zhibin Wu, and Yilong Yin. Vt-fsl: Bridging vision and text with llms for few-shot learning. In *NeurIPS*, volume 38, pages 113697–113727, 2025.

[68] Wenhao Li, Xianjing Meng, Qiangchang Wang, Zhongyi Han, Zhibin Wu, and Yilong Yin. Dvla-rl: Dual-level vision-language alignment with reinforcement learning gating for few-shot learning. In *ICLR*, 2026.

[69] Masashi Okada and Tadahiro Taniguchi. Dreaming: Model-based reinforcement learning by latent imagination without reconstruction. In *2021 ieee international conference on robotics and automation (icra)*, pages 4209–4215. IEEE, 2021.

[70] Yao Mu, Yuzheng Zhuang, Bin Wang, Guangxiang Zhu, Wulong Liu, Jianyu Chen, Ping Luo, Shengbo Li, Chongjie Zhang, and Jianye Hao. Model-based reinforcement learning via imagination with derived memory. *Advances in Neural Information Processing Systems*, 34:9493–9505, 2021.

[71] Pengxuan Yang, Yupeng Zheng, Deheng Qian, Zebin Xing, Qichao Zhang, Linbo Wang, Yichen Zhang, Shaoyu Guo, Zhongpu Xia, Qiang Chen, et al. Dreamerad: Efficient reinforcement learning via latent world model for autonomous driving. *arXiv preprint arXiv:2603.24587*, 2026.

[72] Zhiyi Li, Peilin Wu, Xiaoshen Han, Ruojin Cai, and Yilun Du. 4d latent world model for robot planning. *OpenReview preprint*, 2026.

[73] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[74] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[75] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

[76] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[77] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023.

[78] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[79] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

[80] Yifei Dong, Fengyi Wu, Guangyu Chen, Zhi-Qi Cheng, Qiyu Hu, Yuxuan Zhou, Jingdong Sun, Jun-Yan He, Qi Dai, and Alexander G Hauptmann. Unified world models: Memory-augmented planning and foresight for visual navigation. *arXiv preprint arXiv:2510.08713*, 2025.

[81] Yangcheng Yu, Xin Jin, Yu Shang, Xin Zhang, Haisheng Su, Wei Wu, and Yong Li. Mowm: Mixture-of-world-models for embodied planning via latent-to-pixel feature modulation. *arXiv preprint arXiv:2509.21797*, 2025.

[82] Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. Gigaworld-policy: An efficient action-centered world–action model. *arXiv preprint arXiv:2603.17240*, 2026.

[83] Taiyi Su, Jian Zhu, Yaxuan Li, Chong Ma, Jianjun Zhang, Zitai Huang, Hanli Wang, and Yi Xu. Towards high-consistency embodied world model with multi-view trajectory videos. *arXiv preprint arXiv:2511.12882*, 2025.

[84] Zijie Wang, Weiming Zhang, Wei Zhang, Xiao Tan, Hongxing Liu, Yaowei Wang, and Guanbin Li. Lanediffusion: Improving centerline graph learning via prior injected bev feature generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 27052–27062. IEEE, 2025.

[85] Zijie Wang, Wei Zhang, Weiming Zhang, Xiao Tan, Jie Feng, Xiaonan Luo, and Guanbin Li. Lanediffusion++: Feature-level generative priors for structured lane graph learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[86] Zedong Dan, Zijie Wang, Wei Zhang, Xiangru Lin, Weiming Zhang, Xiao Tan, Jingdong Wang, Liang Lin, and Guanbin Li. Optimvmap: Offline vectorized map construction via optimal multi-vehicle perspectives. *arXiv preprint arXiv:2604.17135*, 2026.

[87] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306. IEEE, 2025.

[88] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10371–10381. IEEE, 2024.

[89] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024.

[90] Zijie Wang, Wei Zhang, Weiming Zhang, Xiao Tan, Weikai Chen, Xiaoxu Li, and Guanbin Li. Ardepth: Auto-regressive monocular depth estimation with progressive visual conditioning. *arXiv preprint arXiv:2607.12433*, 2026.

Appendix

Appendix organization. This appendix provides supplementary material for the theoretical formulation, implementation details, evaluation protocol, additional results, ablations, qualitative examples, and related work. Appendix A gives the proof of the group-action formulation and clarifies its connection to $SE(2)$ navigation. Appendix B details the latent rollout objective, constraint synthesis, action-space approximation, and the relation between latent training and recovered-state evaluation. Appendix C describes datasets, model variants, training setup, pose recovery, and metric protocols. Appendices D–F provide additional quantitative results, ablations, and qualitative examples. Appendix G reviews related work, and Appendix H discusses broader impact.

A Additional Details of the Group-Action Formulation

This appendix provides additional details for the group-action formulation in Sec. 3. We first prove Theorem 1, then discuss why $SE(2)$ is the natural group for planar navigation, and finally clarify how the exact group-action ideal is operationalized through local observable conditions in video world models.

A.1 Proof of Theorem 1

Proof. We use the action-order convention adopted in Sec. 3, where applying a_1 followed by a_2 corresponds to the composed action $a_1 \circ a_2$. Under this convention, the compatibility condition

$$f_\theta(f_\theta(s, a_1), a_2) = f_\theta(s, a_1 \circ a_2) \tag{22}$$

and the identity condition

$$f_\theta(s, e) = s \tag{23}$$

are exactly the axioms of a right group action of G on $\mathcal{S}$. Therefore, f_θ defines a group action under this convention.

It remains to show the finite-sequence property. For $n = 1$, the statement is trivial. Assume that for some $n - 1$,

$$f_\theta(\cdots f_\theta(f_\theta(s, a_1), a_2) \cdots, a_{n-1}) = f_\theta(s, a_1 \circ \cdots \circ a_{n-1}). \tag{24}$$

Applying a_n and using compatibility gives

$$\begin{aligned} &f_\theta(f_\theta(s, a_1 \circ \cdots \circ a_{n-1}), a_n) \\ &\quad = f_\theta(s, (a_1 \circ \cdots \circ a_{n-1}) \circ a_n) \\ &\quad = f_\theta(s, a_1 \circ \cdots \circ a_n), \end{aligned} \tag{25}$$

where the last equality follows from associativity of the group operation. Thus, by induction,

$$f_\theta(\cdots f_\theta(f_\theta(s, a_1), a_2) \cdots, a_n) = f_\theta(s, a_1 \circ \cdots \circ a_n). \tag{26}$$

The inverse condition follows from the group-action axioms. Indeed, for any $a \in G$, compatibility and identity imply

$$f_\theta(f_\theta(s, a), a^{-1}) = f_\theta(s, a \circ a^{-1}) = f_\theta(s, e) = s. \tag{27}$$

We state inverse consistency explicitly in the main text because it is one of the most direct observable consequences of action cancellation in navigation rollouts. □

A.2 Why $SE(2)$ for Navigation

In embodied navigation, actions typically represent local ego-motion increments in the plane. An idealized planar motion can be written as an element of the special Euclidean group $SE(2)$:

$$g = (R, \mathbf{p}), \qquad R \in SO(2), \quad \mathbf{p} \in \mathbb{R}^2, \tag{28}$$

or equivalently as the homogeneous transformation

$$G = \begin{bmatrix} R & \mathbf{p} \\ \mathbf{0}^\top & 1 \end{bmatrix}. \tag{29}$$

The group operation is rigid-motion composition:

$$(R_1, \mathbf{p}_1) \circ (R_2, \mathbf{p}_2) = (R_1 R_2, \mathbf{p}_1 + R_1 \mathbf{p}_2). \tag{30}$$

The identity element is

$$e = (I, \mathbf{0}), \tag{31}$$

and the inverse of a motion is

$$(R, \mathbf{p})^{-1} = (R^\top, -R^\top \mathbf{p}). \tag{32}$$

This group structure captures the algebraic relations that navigation actions should satisfy. A zero motion should preserve the state, a motion followed by its inverse should cancel, and a sequence of motions should have an effect determined by their composition. The group-action formulation in Sec. 3 uses this structure as the ideal state-space criterion for action-faithful world modeling.

In practical video world models, actions are often represented by normalized local ego-motion increments such as

$$\mathbf{a}_t = (\Delta x_t, \Delta y_t, \Delta \theta_t). \tag{33}$$

The main text uses this ego-motion action parameterization to instantiate observable local consequences of the $SE(2)$ group-action ideal. The exact action-space implementation used for training is described separately in Appendix B.

A.3 From Exact Group Action to Observable Local Conditions

The group-action formulation defines an ideal state-space property. If the state $\mathbf{s}_t$ were directly observable and the transition map were deterministic, one could test whether

$$\begin{aligned} f_\theta(f_\theta(s, a_1), a_2) &= f_\theta(s, a_1 \circ a_2), \\ f_\theta(s, e) &= s, \\ f_\theta(f_\theta(s, a), a^{-1}) &= s, \end{aligned} \tag{34}$$

hold exactly. In stochastic video rollout, however, the underlying state is not directly observed. The model produces visual observations, and the induced motion must be recovered through an estimator Φ. Exact equality in the abstract state space is therefore not directly available as an evaluation target.

The main paper consequently focuses on local observable consequences of the group-action ideal. The identity condition becomes zero-action preservation: a rollout under the neutral action should not introduce spurious motion. The inverse condition becomes recovery after cancellation: executing a motion and then its inverse should return to the starting state. The compatibility condition becomes local composition consistency: different short action decompositions with the same intended composed motion should induce compatible endpoints.

These conditions are not introduced as task-specific heuristics. They are operational tests of whether the model-induced rollout preserves the basic algebraic consequences of planar motion composition. Their violations reveal failure modes that can remain hidden when evaluating only visual plausibility.

A.4 Scope of the Local Composition Approximation

The group-action formulation is grounded in the SE(2) structure of planar rigid motions. Exact SE(2) composition is generally non-commutative: the effect of a translation followed by a rotation is not identical to the effect of the same rotation followed by the same translation. Therefore, the composition condition used in this paper should not be interpreted as assuming that SE(2) composition is commutative, nor as claiming that arbitrary trajectories with the same cumulative displacement are equivalent.

Our use of composition consistency is a local operational approximation for short ego-motion windows. In the datasets and training setup considered here, actions are represented as normalized local ego-motion increments. For short windows with small rotations, redistributing the same intended local displacement provides a practical probe for whether the model is overly sensitive to the temporal realization of a local motion segment. The goal is not to replace the exact SE(2) operation with a globally valid additive rule. Rather, the approximation asks whether two short action decompositions with matched intended motion induce compatible endpoint states.

This distinction is important for interpreting both the metric and the training objective. The composition probe should be read as a local consistency test for action-conditioned rollout, not as a statement that all trajectories with equal cumulative displacement should coincide. Accordingly, all composition probes are restricted to short windows, and the window lengths used in evaluation are chosen to remain within this local regime. More generally, the present formulation is intended for controlled transformations that admit meaningful group structure. It does not assume that arbitrary world dynamics, particularly intrinsically irreversible transitions, can be modeled as a group action.

B Additional Details on Group-Action Training

This appendix provides implementation details for the group-action training objective introduced in Sec. 4. We describe the free-running latent rollout setting, the synthesis of training constraints from existing trajectories, the action-space approximations used during training, the practical per-batch objective, and the relation between latent training and recovered-state evaluation. Evaluation protocols and metric aggregation are described separately in Appendix C.

B.1 Free-Running Latent Rollout

The group-action objective is computed on latent endpoints generated by the model's own rollout process. Starting from a sampled latent state $\mathbf{z}_t$, the model is rolled out under a synthesized action segment without injecting ground-truth observations inside the constrained interval. This setting matches the object being regularized, namely the action-conditioned transition induced by the model itself.

There are two possible rollout paradigms for applying such constraints. Teacher-forced rollout supplies ground-truth visual context at intermediate steps, while free-running rollout lets the model condition on its own generated latent states. Since the identity, inverse, and composition conditions characterize how the learned transition behaves under self-composition, we adopt free-running latent rollout inside the constrained interval. This allows the group-action objective to act directly on the internal dynamics used during generation and exposes violations that teacher forcing may hide.

For memory efficiency and training stability, gradients are propagated only through the sampled local interval. Latent states before the interval are detached from the computation graph. This truncated rollout is used as an optimization device and does not change the definition of the group-action constraints. We examine the effect of this free-running design in the ablation study.

B.2 Training Constraint Synthesis and Per-Batch Sampling

For each training trajectory, a starting index t is sampled from valid temporal positions. The local rollout horizon is sampled as

$$l \sim \mathcal{U}\{1, \dots, L\}, \tag{35}$$

where L controls the maximum span over which a group-action constraint is imposed. Short horizons emphasize immediate transition behavior, while longer horizons expose errors that appear only after repeated transition composition. Unless otherwise specified, $L = 4$ is used in our experiments.

Given the native local action segment $\mathbf{u}_{1:l}$ defined in Sec. 4.2, the identity, inverse, and composition constraints can in principle all be evaluated for the same sampled segment. In practice, computing all three constraints in every batch increases memory and computation because each constraint requires a separate latent rollout. We therefore sample one constraint type per training batch. The active group-action loss is one of

$$\mathcal{L}_{\mathrm{GA}}^{(b)} \in \{\mathcal{L}_{\mathrm{id}}, \mathcal{L}_{\mathrm{inv}}, \mathcal{L}_{\mathrm{comp}}\}. \tag{36}$$

This corresponds to a stochastic realization of the full objective in Eq. 13: in each batch, only the sampled group-action term is activated, while the other two terms have zero effective weight. Equivalently, the per-batch objective can be written as

$$\mathcal{L} = \mathcal{L}_{\mathrm{diff}} + \lambda \mathcal{L}_{\mathrm{GA}}^{(b)}. \tag{37}$$

Across training, this stochastic sampling exposes the model to all three group-action constraints while keeping the rollout-based objective tractable.

B.3 Action-Space Approximation for Training

The theoretical formulation in Sec. 3 is defined in terms of the group operation of planar rigid motions. In the video world models considered in this work, however, actions are provided as normalized local ego-motion increments. The training constraints are therefore instantiated under this action parameterization, following the operational form used in Sec. 3.1.

For the inverse constraint, let

$$\mathbf{u}_{1:l} = (\mathbf{a}_t, \dots, \mathbf{a}_{t+l-1}) \tag{38}$$

be a local action segment. The reverse part is constructed by reversing the temporal order and negating each normalized increment:

$$\hat{\mathbf{u}}_{1:l} = (-\mathbf{a}_{t+l-1}, \dots, -\mathbf{a}_t). \tag{39}$$

The full inverse training segment is then the forward–inverse cycle

$$\mathbf{u}^{\mathrm{inv}}_{1:2l} = (\mathbf{u}_{1:l}, \hat{\mathbf{u}}_{1:l}) = (\mathbf{a}_t, \dots, \mathbf{a}_{t+l-1}, -\mathbf{a}_{t+l-1}, \dots, -\mathbf{a}_t). \tag{40}$$

This implements the cancellation relation used by the inverse constraint under the normalized ego-motion parameterization.

For the composition constraint, exact equality under $SE(2)$ composition would require composing rigid motions in the full group representation. Instead, consistent with the main text, we construct local composition probes by accumulating normalized action increments over short ego-motion windows. Given a segment $\mathbf{u}^A_{1:l}$, the alternative segment $\mathbf{u}^B_{1:l}$ is constructed as in Eq. 17, and satisfies

$$\sum_{i=1}^{l} \mathbf{u}^A_i = \sum_{i=1}^{l} \mathbf{u}^B_i. \tag{41}$$

Thus, the two segments preserve the same accumulated increments while changing the temporal allocation of the intended local motion.

This approximation is used only for local constraint synthesis. It should not be interpreted as replacing exact $SE(2)$ composition with a globally valid additive rule. Rather, it provides an efficient way to construct alternative short-horizon decompositions from existing trajectories, enabling the model to be regularized against sensitivity to different temporal realizations of the same intended local motion.

The Dirichlet redistribution in Eq. 17 controls the diversity of these alternative decompositions. The concentration parameter α determines how evenly the accumulated increments are distributed across the segment. Smaller values produce more uneven allocations, while larger values produce smoother redistributions.

B.4 Latent Training versus State-Space Evaluation

Latent training and recovered-state evaluation serve different roles. The group-action objective is imposed in latent space because this is where the video world model composes its internal rollout dynamics before decoding future observations. Applying the same objective directly in recovered state space would require decoding generated frames and running an external pose estimator inside the training loop, which is computationally expensive and unstable for gradient-based optimization.

Recovered-state evaluation is used because action faithfulness is ultimately a property of induced motion. Recovered trajectories provide an interpretable representation for measuring identity drift, inverse recovery error, composition mismatch, and repeated-rollout dispersion.

The method does not assume that latent distance is identical to recovered-state distance. The latent losses and the recovered-state metrics are both derived from the same group-action conditions. The former provide a practical optimization signal, while the latter measure whether generated rollouts exhibit the expected motion-level consequences.

C Additional Experimental Details

This appendix provides additional details for the experiments in Sec. 5. We describe the datasets, model variants, training setup, pose recovery, GAC evaluation protocol, GAC probe design and aggregation, and GAR rollout protocol.

C.1 Datasets

Experiments are conducted on RECON [37], SCAND [38], and HuRoN [39], following the preprocessing and data split protocol of NWM [13]. Relative actions are derived from ground-truth pose differences and normalized so that action scales remain comparable across datasets.

RECON. RECON is an outdoor robotics dataset collected using a Clearpath Jackal UGV platform. It contains approximately 40 hours of data across 9 open-world environments. Following NWM [13], 9,468 video segments are used for training and 2,367 video segments for testing. Its emphasis on long-range traversal in visually diverse open-world environments makes it particularly suitable for evaluating whether rollout instability accumulates over extended action sequences.

SCAND. SCAND is a robotics dataset of socially compliant navigation demonstrations collected using a wheeled Clearpath Jackal and a legged Boston Dynamics Spot in both indoor and outdoor settings at UT Austin. The full dataset contains 8.7 hours of data, 138 trajectories, and 25 miles of demonstrations. Following NWM [13], 484 video segments are used for training and 121 video segments for testing. Compared with RECON, SCAND contains richer local interaction structure and denser socially conditioned motion variation, making it useful for evaluating whether local group-action consistency is preserved in socially complex settings.

HuRoN. HuRoN is a robotics dataset of social navigation interactions collected using a Robot Roomba in indoor environments at UC Berkeley. The dataset contains more than 75 hours of data across 5 environments and includes over 4,000 human interactions. Following NWM [13], 2,451 video segments are used for training and 613 video segments for testing. Its emphasis on socially unobtrusive navigation makes it relevant for evaluating whether action-conditioned rollout remains stable in interaction-heavy and behaviorally sensitive environments.

C.2 Model Variants and Adaptation

NWM [13] is used as the main testbed throughout the experiments. It is a representative diffusion-based navigation world model with stochastic action-conditioned rollout, making it suitable for studying whether explicit group-action constraints improve rollout stability. Following its default setting, a CDiT-XL backbone with approximately 1B parameters is used together with a 4-frame context window. Visual observations are tokenized using the Stable Diffusion VAE [41], following the design adopted in NWM.

DIAMOND [40] is included as an additional action-conditioned world-model baseline. Since DIAMOND was originally introduced in a different domain, we adapt its public implementation to the embodied navigation setting using a training setup similar in spirit to NWM. This additional baseline is included to test whether the proposed GA framework generalizes beyond a single world-model architecture.

C.3 Training Setup

Optimization is performed with AdamW [42]. For GA training, full fine-tuning is used, while the pretrained VAE is kept frozen. Keeping the VAE fixed allows us to evaluate the effect of GA under a common latent representation. The initial learning rate is set to 1×10^{-6}. Training is conducted on 8 A800 GPUs with a per-GPU batch size of 1 for approximately 100k optimization steps. Unless otherwise specified, the remaining optimization settings follow the corresponding base model as closely as possible. The group-action objective is applied only during fine-tuning. The maximum local rollout horizon is set to $L = 4$ by default. The group-action loss weight λ is selected from $\{0.1, 0.5, 1.0\}$, and $\lambda = 0.5$ is used by default. As described in Appendix B.2, one group-action constraint type is sampled in each training batch for efficiency.

C.4 Pose Recovery and Trajectory Alignment

Generated videos are mapped to recovered state trajectories using the pose estimator Φ defined in Sec. 3.1, implemented with the off-the-shelf DROID-SLAM [43] toolkit. All motion-level errors are computed using the state distance $d(\cdot, \cdot)$ in Eq. 8.

For GAR evaluation, both aligned and non-aligned trajectory errors are reported. Aligned errors remove global pose drift before computing trajectory discrepancy, while non-aligned errors measure raw deviation in the recovered trajectory. The aligned setting emphasizes local trajectory shape, whereas the non-aligned setting additionally reflects global drift in position and heading.

C.5 GAC Evaluation Protocol

This subsection specifies the evaluation configurations used to compute the Group-Action Consistency (GAC) error in Sec. 4.3. Generated videos are first mapped to recovered state trajectories using Φ. The main text defines Δ_{id}, Δ_{inv}, and Δ_{comp} as the averaged errors for identity, inverse, and local composition consistency. Here we spell out the corresponding probe families.

Identity evaluation. Identity consistency is evaluated by inserting zero-action segments into the action stream. Let $\mathcal{T}_{\mathrm{id}}^{(k,l)}$ denote the set of evaluated zero-action segments, where k is the number of inserted segments and l is the length of each segment. For a segment starting at timestep t, the identity drift is measured by comparing the recovered states before and after the zero-action interval:

$$\Delta_{\mathrm{id}}^{(k,l)} = \frac{1}{|\mathcal{T}_{\mathrm{id}}^{(k,l)}|} \sum_{t \in \mathcal{T}_{\mathrm{id}}^{(k,l)}} d\big(\mathbf{s}_{t+l}^{\mathrm{id}}, \mathbf{s}_t\big). \tag{42}$$

A larger value indicates stronger drift under the neutral action.

Inverse evaluation. Inverse consistency is evaluated by executing a local action segment followed by its inverse sequence. Let $\mathcal{T}_{\mathrm{inv}}^{(k,l)}$ denote the set of evaluated forward–inverse segments, where k is the number of evaluated segments and l controls the temporal scale of each segment. The inverse recovery error is

$$\Delta_{\mathrm{inv}}^{(k,l)} = \frac{1}{|\mathcal{T}_{\mathrm{inv}}^{(k,l)}|} \sum_{t \in \mathcal{T}_{\mathrm{inv}}^{(k,l)}} d\big(\mathbf{s}_{t,\mathrm{rec}}^{\mathrm{inv}}, \mathbf{s}_t\big), \tag{43}$$

where $\mathbf{s}_{t,\mathrm{rec}}^{\mathrm{inv}}$ is the recovered state after executing the action segment and its inverse. A larger value indicates weaker inverse consistency.

Composition evaluation. Local composition consistency is evaluated through alternative decompositions of the same local motion. Given a local action segment $\mathbf{a}_{1:l}$, an alternative segment $\mathbf{b}_{1:l}$ is constructed by redistributing the same cumulative displacement:

$$\mathbf{b}_i = w_i \sum_{j=1}^{l} \mathbf{a}_j, \qquad \mathbf{w} \sim \mathrm{Dir}(\alpha), \tag{44}$$

which satisfies

$$\sum_{i=1}^{l} \mathbf{b}_i = \sum_{i=1}^{l} \mathbf{a}_i. \tag{45}$$

This construction changes the temporal allocation of the local motion while preserving its cumulative displacement. The model is rolled out under the original segment and the recomposed segment. Let $\mathcal{T}_{\mathrm{comp}}^{(l)}$ denote the set of evaluated composition segments. The composition mismatch is

$$\Delta_{\mathrm{comp}}^{(l)} = \frac{1}{|\mathcal{T}_{\mathrm{comp}}^{(l)}|} \sum_{t \in \mathcal{T}_{\mathrm{comp}}^{(l)}} d\big(\mathbf{s}_{t+l}^{A}, \mathbf{s}_{t+l}^{B}\big), \tag{46}$$

where $\mathbf{s}_{t+l}^{A}$ and $\mathbf{s}_{t+l}^{B}$ are the recovered endpoint states under the original and recomposed segments. A larger value indicates weaker local composition consistency.

C.6 GAC Probe Design and Metric Aggregation

The GAC metric is implemented as a fixed controlled probe suite. All probe configurations are specified before evaluation and shared across methods, so the reported differences are not affected

by method-specific probe selection. The probe suite evaluates local observable consequences of the group-action ideal at multiple difficulty levels.

For identity and inverse consistency, each configuration is indexed by (k, l). The parameter k controls the number of inserted or evaluated local segments. It tests whether small violations accumulate when the same local group-action relation is applied repeatedly. The parameter l controls the temporal length of each segment. It tests whether the violation grows as the local rollout span becomes longer. In the main analysis, one factor is varied while the other is fixed, using $k, l \in \{1, 3, 5\}$. This produces small, medium, and large local probes while keeping the evaluation interpretable. We do not use a dense grid over all (k, l) pairs because it would mix the two sources of difficulty and introduce redundant configurations.

For local composition consistency, the probe compares two decompositions of the same local motion window. The natural scale parameter is therefore the window length l. We use $l \in \{2, 4, 6\}$ to cover increasing local composition difficulty while staying within the short-horizon regime where the local composition approximation is intended to apply. The goal is not to claim that long-horizon trajectories with the same cumulative displacement are equivalent. Rather, the probe tests whether the model is overly sensitive to different temporal realizations of a short local motion.

For each group-action condition, configuration-level errors are first averaged over fixed valid starting positions and evaluation sequences. The component errors are then averaged over the corresponding evaluation family:

$$\bar{\Delta}_{\mathrm{id}} = \frac{1}{|\mathcal{K}_{\mathrm{id}}|} \sum_{(k,l)\in\mathcal{K}_{\mathrm{id}}} \Delta_{\mathrm{id}}^{(k,l)}, \qquad \bar{\Delta}_{\mathrm{inv}} = \frac{1}{|\mathcal{K}_{\mathrm{inv}}|} \sum_{(k,l)\in\mathcal{K}_{\mathrm{inv}}} \Delta_{\mathrm{inv}}^{(k,l)}. \tag{47}$$

For local composition consistency, we average over the evaluated window lengths:

$$\bar{\Delta}_{\mathrm{comp}} = \frac{1}{|\mathcal{L}_{\mathrm{comp}}|} \sum_{l\in\mathcal{L}_{\mathrm{comp}}} \Delta_{\mathrm{comp}}^{(l)}. \tag{48}$$

The aggregate GAC error reported in the experiments is

$$\mathcal{E}_{\mathrm{GAC}} = \frac{1}{3}\left(\bar{\Delta}_{\mathrm{id}} + \bar{\Delta}_{\mathrm{inv}} + \bar{\Delta}_{\mathrm{comp}}\right). \tag{49}$$

This aggregation is well-defined because all components are measured using the same recovered-state distance $d(\cdot, \cdot)$. Equal weighting avoids introducing a task-specific preference among identity preservation, inverse cancellation, and local composition consistency. We report both the aggregate score and the component errors, so changes in $\mathcal{E}_{\mathrm{GAC}}$ can be traced to the corresponding local group-action conditions.

C.7 GAR Evaluation Protocol

Group-Action Robustness (GAR) is evaluated under repeated stochastic rollouts conditioned on the same initial observation and identical action sequence. Unless otherwise specified, each model is sampled $N = 5$ times. The corresponding generated videos are mapped to recovered state trajectories $\{\mathbf{s}_{1:T}^{(i)}\}_{i=1}^{N}$. The GAR error is computed using Eq. 21. Mean and standard deviation are reported across evaluation sequences. Both 16-frame and 64-frame horizons are evaluated to measure short-range and longer-range accumulation of rollout errors.

C.8 Assets and Licenses

This work uses existing research datasets and model implementations, including RECON [37], SCAND [38], HuRoN [39], NWM [13], DIAMOND [40], and DROID-SLAM [43]. We credit the original creators through citation and use these assets for research purposes following their released terms and licenses. No new dataset is collected or redistributed in this work.

D Additional Quantitative Results

This appendix provides supplementary quantitative results beyond the main RECON analysis in Sec. 5.

D.1 Additional GAR Results

Table 4 reports additional GAR results on SCAND and HuRoN under repeated stochastic rollouts with identical initial observations and action sequences. The same evaluation protocol as in Sec. 5.3 is utilized on both datasets. On both SCAND and HuRoN, GA training reduces rollout dispersion for both DIAMOND and NWM, showing that the improvement in rollout-level robustness is not limited to the main RECON setting. The gains are consistently observed under both aligned and non-aligned evaluation, indicating improved local trajectory agreement as well as reduced global drift.

Table 4: Additional Group-Action Robustness (GAR) results on SCAND and HuRoN under repeated stochastic rollouts with identical action sequences. Lower is better. "Aligned" removes global pose drift before computing trajectory error, while "Non-aligned" measures raw trajectory deviation.

Dataset	Method	16 Frames		64 Frames	
		Aligned ↓	Non-aligned ↓	Aligned ↓	Non-aligned ↓
SCAND	DIAMOND [40]	0.63 ± 0.42	1.24 ± 0.86	2.06 ± 1.31	6.42 ± 4.16
	DIAMOND + GA	0.53 ± 0.35	1.03 ± 0.72	1.68 ± 1.05	5.28 ± 3.36
	NWM [13]	0.46 ± 0.28	0.91 ± 0.64	1.45 ± 0.85	4.08 ± 2.69
	NWM + GA	$\mathbf{0.37 \pm 0.23}$	$\mathbf{0.72 \pm 0.50}$	$\mathbf{1.12 \pm 0.68}$	$\mathbf{3.39 \pm 2.04}$
HuRoN	DIAMOND [40]	0.65 ± 0.23	1.37 ± 0.56	2.34 ± 0.92	7.36 ± 3.20
	DIAMOND + GA	0.55 ± 0.19	1.13 ± 0.46	1.91 ± 0.75	6.04 ± 2.58
	NWM [13]	0.47 ± 0.16	0.98 ± 0.35	1.70 ± 0.58	5.13 ± 2.22
	NWM + GA	$\mathbf{0.39 \pm 0.13}$	$\mathbf{0.77 \pm 0.29}$	$\mathbf{1.33 \pm 0.47}$	$\mathbf{4.15 \pm 1.70}$

D.2 Additional GAC Results

Table 5 reports aggregated GAC errors on SCAND and HuRoN. Component errors are averaged over the evaluated probe configurations, and $\mathcal{E}_{\mathrm{GAC}}$ is computed as the mean of identity, inverse and composition errors. On both datasets, GA training reduces the aggregate GAC error for both DIAMOND and NWM. This supports the conclusion that the latent group-action objective improves local action faithfulness beyond the main RECON dataset and beyond a single backbone architecture.

Table 5: Aggregated Group-Action Consistency (GAC) error on SCAND and HuRoN. Component errors are averaged over evaluated probe configurations. The last column reports $\mathcal{E}_{\mathrm{GAC}}$ as the mean of the three component errors. Standard deviations are shown for component metrics. Lower is better.

Dataset	Method	$\bar{\Delta}_{\mathrm{id}} \downarrow$	$\bar{\Delta}_{\mathrm{inv}} \downarrow$	$\bar{\Delta}_{\mathrm{comp}} \downarrow$	$\mathcal{E}_{\mathrm{GAC}} \downarrow$
SCAND	DIAMOND [40]	6.12 ± 2.95	5.92 ± 2.81	3.56 ± 2.12	5.20
	DIAMOND + GA	5.56 ± 2.71	5.43 ± 2.53	2.96 ± 1.82	4.65
	NWM [13]	5.38 ± 2.63	5.51 ± 2.55	2.68 ± 1.61	4.52
	NWM + GA	$\mathbf{4.84 \pm 2.32}$	$\mathbf{4.67 \pm 2.16}$	$\mathbf{2.06 \pm 1.31}$	**3.86**
HuRoN	DIAMOND [40]	7.08 ± 2.86	6.77 ± 2.91	3.31 ± 1.84	5.72
	DIAMOND + GA	6.36 ± 2.66	6.08 ± 2.55	2.79 ± 1.50	5.08
	NWM [13]	6.07 ± 2.48	6.39 ± 2.64	2.53 ± 1.37	5.00
	NWM + GA	$\mathbf{5.36 \pm 2.25}$	$\mathbf{5.49 \pm 2.32}$	$\mathbf{1.99 \pm 1.15}$	**4.28**

D.3 Probe-wise Analysis of GAC Errors

Fig. 3 provides a probe-wise view of GAC errors on RECON. The probe parameters are not intended to define a strictly monotonic difficulty scale. Instead, they expose complementary failure modes of group-action consistency, including repeated local relations, longer temporal extent, and alternative decompositions of the same local motion.

For inverse consistency, errors tend to increase as the evaluated segment becomes longer or is repeated more times. This is expected because the model must compose a forward motion with its inverse and

recover the starting state. Residual transition errors accumulated during the forward segment can therefore remain after cancellation. GA regularization consistently lowers the inverse error, indicating improved action cancellation under composed rollouts.

For local composition consistency, longer windows lead to larger endpoint mismatch. Longer windows involve more transition steps and give two equivalent local decompositions more opportunities to diverge. The reduction after GA regularization shows that the model becomes less sensitive to the particular temporal decomposition of a local motion, which is the desired behavior under the group-action formulation.

Identity consistency follows a less monotonic pattern. Repeated or longer zero-action segments can make the stationary condition more explicit, which may reduce mean drift in intermediate settings. At larger settings, however, the probe also imposes a stricter preservation requirement, since the model must remain static across more pause locations or longer pause durations. The increased error or variance in these cases suggests residual motion under extended zero-action rollouts, rather than exact realization of the identity transformation. GA regularization reduces identity errors in most settings, showing improved but still imperfect preservation under zero action.

D.4 Image-Quality Evaluation

Table 6 reports image-quality metrics for generated rollouts on RECON. We evaluate LPIPS [44], DreamSim [45], and PSNR between generated frames and the corresponding ground-truth frames following NWM [13], using the same preprocessing and sampling pipeline across methods. This evaluation checks whether the gains in GAC and GAR come at the cost of visual prediction quality. All methods are evaluated under the same preprocessing, sampling, and metric-computation pipeline. The results show that GA regularization keeps image-quality metrics comparable to the base model. This suggests that the proposed objective improves action-conditioned dynamics without noticeably degrading perceptual or pixel-level generation quality.

Table 6: Image-quality evaluation on RECON under the same evaluation pipeline. LPIPS and DreamSim measure perceptual dissimilarity, while PSNR measures pixel-level reconstruction quality. Lower is better for LPIPS and DreamSim, while higher is better for PSNR.

Method	LPIPS ↓	DreamSim ↓	PSNR ↑
NWM	0.386 ± 0.004	0.141 ± 0.002	13.86 ± 0.08
NWM + GA	$\mathbf{0.364 \pm 0.003}$	$\mathbf{0.130 \pm 0.002}$	$\mathbf{14.29 \pm 0.07}$

D.5 Direct Trajectory Fidelity

To complement GAC and GAR, we directly compare the recovered generated trajectory with the corresponding ground-truth trajectory at every timestep using the state distance in Eq. 8. We report the mean error over the full rollout, the maximum intermediate error excluding the final step, and the final-step error. GA reduces all three trajectory errors at both horizons. In particular, the reduction in maximum intermediate error indicates that the improvement is not restricted to the final endpoint, providing complementary evidence to the rollout-stability results measured by GAR.

E Additional Ablation Studies

This appendix provides supplementary ablation studies on RECON. We analyze the effect of the maximum rollout span, the group-action loss weight, and the training rollout paradigm.

E.1 Effect of Maximum Rollout Span

Table 8 studies the effect of the maximum rollout span L used in group-action training. Increasing the rollout span generally improves GAR, indicating that enforcing group-action constraints over multi-step transitions is important for reducing accumulated rollout error. The performance becomes similar once L reaches a moderate range, with only marginal differences between $L = 4$ and $L = 6$.

Table 7: Direct trajectory-to-ground-truth errors on RECON using NWM. Lower is better.

Frames	Method	Mean ↓	Max intermediate ↓	Final ↓
16	NWM	1.37	2.43	2.21
16	NWM + GA	**1.11**	**2.02**	**1.82**
64	NWM	3.21	5.68	5.27
64	NWM + GA	**2.69**	**4.81**	**4.48**

We therefore use $L = 4$ as the default setting, which offers a good balance between performance and training cost.

Table 8: Effect of the maximum rollout span L used in group-action training on RECON. Lower is better. "Aligned" removes global pose drift before computing trajectory error, while "Non-aligned" measures raw trajectory deviation.

L	16 Frames		64 Frames	
	Aligned ↓	Non-aligned ↓	Aligned ↓	Non-aligned ↓
2	0.27 ± 0.22	0.53 ± 0.37	0.49 ± 0.36	1.35 ± 0.89
4	$\mathbf{0.26 \pm 0.20}$	$\mathbf{0.49 \pm 0.35}$	0.46 ± 0.33	$\mathbf{1.28 \pm 0.82}$
6	0.26 ± 0.21	0.50 ± 0.35	$\mathbf{0.45 \pm 0.33}$	1.29 ± 0.83

E.2 Effect of Group-Action Loss Weight

Table 9 studies the effect of the group-action loss weight λ. Beyond hyperparameter sensitivity, this experiment tests whether the observed gain could be attributed simply to continued fine-tuning. The $\lambda = 0$ case corresponds to training under the same pipeline without group-action constraints. Its performance is almost unchanged relative to the baseline, showing only marginal variation. Introducing a nonzero group-action weight leads to clear improvements, with the best overall performance obtained at an intermediate setting. When λ becomes larger, the gain saturates slightly, suggesting that excessively strong group-action training may start to compete with the base diffusion objective.

Table 9: Effect of the group-action loss weight λ on RECON. Lower is better. "Aligned" removes global pose drift before computing trajectory error, while "Non-aligned" measures raw trajectory deviation.

λ	16 Frames		64 Frames	
	Aligned ↓	Non-aligned ↓	Aligned ↓	Non-aligned ↓
Baseline	0.32 ± 0.26	0.65 ± 0.41	0.62 ± 0.44	1.52 ± 1.27
0	0.33 ± 0.23	0.61 ± 0.42	0.60 ± 0.44	1.47 ± 1.21
0.1	0.27 ± 0.21	0.52 ± 0.37	0.48 ± 0.35	1.34 ± 0.88
0.5	$\mathbf{0.26 \pm 0.20}$	$\mathbf{0.49 \pm 0.35}$	$\mathbf{0.46 \pm 0.33}$	$\mathbf{1.28 \pm 0.82}$
1.0	0.27 ± 0.20	0.50 ± 0.36	0.47 ± 0.34	1.31 ± 0.86

E.3 Teacher-Forcing versus Free-Running Training

Table 10 compares two ways of applying group-action constraints during training. In the teacher-forcing variant, the same constraints are constructed, but each transition is conditioned on ground-truth history rather than the model's own previously generated states. In the free-running variant, the model rolls forward under its own generated latent states, and the group-action losses are imposed on the resulting rollout. The teacher-forcing variant improves over the baseline, indicating that the group-action relations are meaningful even when applied under locally corrected context. However, its gain remains consistently smaller than that of free-running training, especially at longer horizons. This supports the free-running latent rollout design discussed in Appendix B.1.

Table 10: Teacher-forcing versus free-running group-action training on RECON. Lower is better. "Aligned" removes global pose drift before computing trajectory error, while "Non-aligned" measures raw trajectory deviation.

Training scheme	16 Frames		64 Frames	
	Aligned ↓	Non-aligned ↓	Aligned ↓	Non-aligned ↓
Baseline	0.32 ± 0.26	0.65 ± 0.41	0.62 ± 0.44	1.52 ± 1.27
GA w/ Teacher-Forcing	0.30 ± 0.24	0.60 ± 0.39	0.56 ± 0.40	1.43 ± 1.05
GA w/ Free-Running	$\mathbf{0.26 \pm 0.20}$	$\mathbf{0.49 \pm 0.35}$	$\mathbf{0.46 \pm 0.33}$	$\mathbf{1.28 \pm 0.82}$

F Additional Qualitative Results

This appendix provides additional qualitative evidence for the two levels of evaluation used in the main paper. We first show failure cases of the baseline model, including weak rollout-level robustness under identical actions and local violations of the group-action conditions. We then provide additional examples after GA training, showing that the improvements measured by GAC and GAR are also visible in generated rollouts.

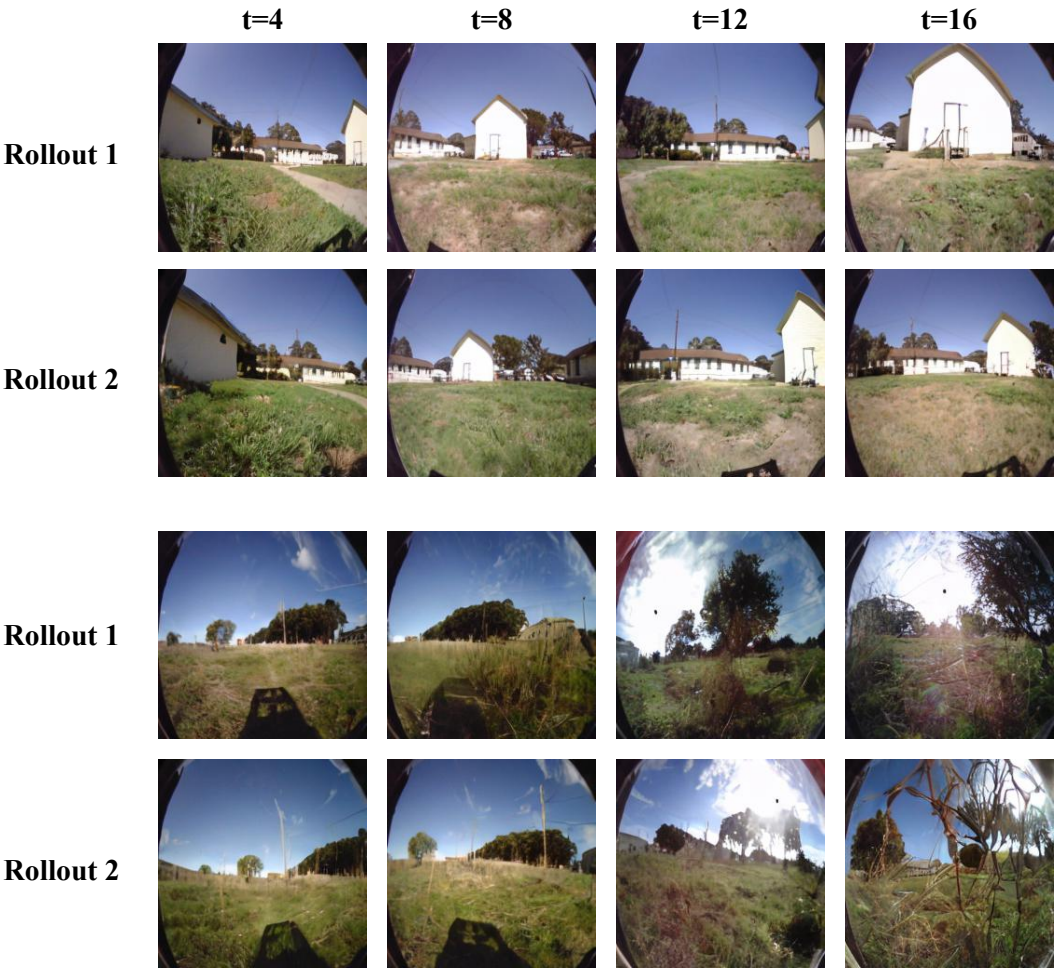


Figure 5: Additional qualitative examples of weak rollout-level robustness in the baseline NWM. Each example shows two stochastic rollouts generated from the same initial observation and identical action sequence. Although the frames can remain visually plausible, the induced motion may diverge over time.

F.1 Baseline Rollout Robustness Failures

Figure 5 shows additional qualitative examples of weak rollout-level robustness in the baseline NWM. Each example compares two stochastic rollouts generated from the same initial observation and identical action sequence. Although individual frames remain visually plausible, the induced motion can diverge noticeably over time, leading to inconsistent trajectory behavior across repeated rollouts.

These examples complement the GAR results in Sec. 5.3. They show that the dispersion measured by $\mathcal{E}_{\mathrm{GAR}}$ is directly visible in generated sequences rather than being only a property of recovered trajectory metrics. The same behavior appears across multiple cases, consistent with the quantitative GAR results.



Figure 6: Additional qualitative examples of local GAC failures in the baseline NWM. The figure includes representative cases for identity consistency, inverse consistency, and local composition consistency. Observed behaviors include drift under zero-action probes, incomplete recovery after forward–inverse action sequences, and mismatched outcomes under locally equivalent action decompositions.

F.2 Baseline Local GAC Failures

Figure 6 shows additional qualitative examples of local GAC failures in the baseline NWM. The examples cover identity consistency, inverse consistency, and local composition consistency. They make the three local failure modes visible in generated rollouts, complementing the quantitative GAC errors in Sec. 5.2.

For identity consistency, inserted zero-action segments do not reliably preserve the current state. Even when the action input indicates no motion, the generated rollout may still exhibit visible drift or viewpoint change. For inverse consistency, a local action segment followed by its inverse often fails to recover the starting state, leaving a residual offset after the forward–inverse sequence. For local composition consistency, locally equivalent action decompositions can lead to noticeably different rollout outcomes, indicating sensitivity to temporal realization beyond cumulative displacement alone.

Together, these examples show that the local errors measured by Δ_{id}, Δ_{inv}, and Δ_{comp} correspond to concrete failures in generated action-conditioned behavior. They also support the interpretation that poor rollout robustness is connected to violations of local group-action conditions.

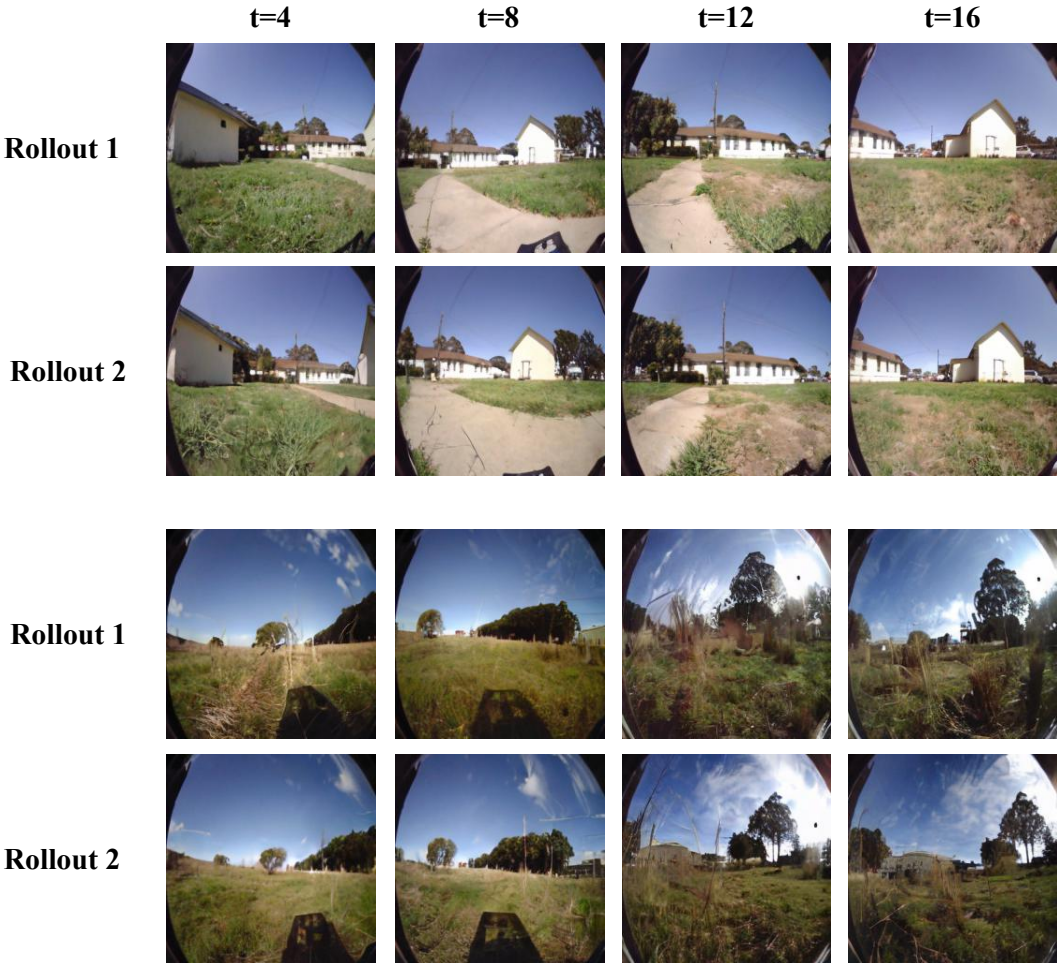


Figure 7: Additional qualitative examples of repeated-rollout robustness after GA training. Each example shows stochastic rollouts generated from the same initial observation and identical action sequence. Although visual details may vary across samples, the induced motion remains more consistent.

F.3 Additional Qualitative Results after GA Training

Figures 7 and 8 provide additional qualitative examples after GA training. Figure 7 shows repeated stochastic rollouts under identical initial observations and action sequences. Although individual samples may still differ in visual details, their motion evolution remains more consistent across rollouts. This complements the GAR improvements reported in Sec. 5.3 and Appendix D.1.

Figure 8 shows representative local group-action behavior under identity, inverse, and composition transformations. The GA-trained model exhibits limited drift under zero action, approximate recovery

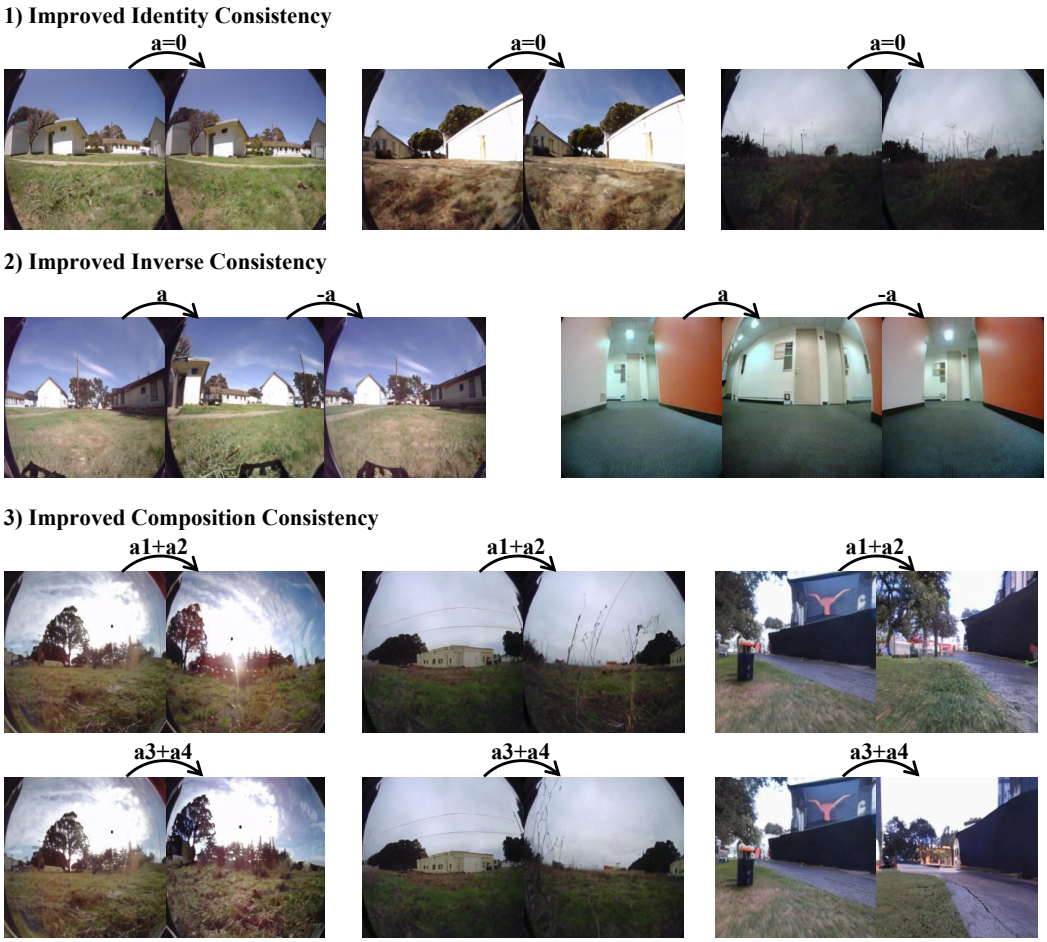


Figure 8: Additional qualitative examples of local group-action behavior after GA training. The examples show reduced drift under zero action, approximate recovery under forward–inverse action sequences, and more compatible outcomes under locally equivalent action decompositions.

after forward–inverse action sequences, and compatible outcomes under locally equivalent action decompositions. These examples complement the GAC improvements in Sec. 5.2 and Appendix D.2.

Taken together, the qualitative results show both sides of the empirical story. The baseline model can generate visually plausible videos while violating local group-action conditions and diverging under repeated rollout. After GA training, the generated rollouts remain more consistent with the action-composition structure, and the improvement is visible not only in aggregate metrics but also in the generated sequences themselves.

G More Comprehensive Related Work

G.1 Visual Navigation and Embodied Action Prediction

Learning-based visual navigation studies how agents select actions from visual observations, goals, or task specifications. This problem has been widely explored in goal-conditioned navigation [14, 46, 47], vision-and-language navigation [48, 49, 50], and foundation-model-based navigation [51, 52, 53]. Early learning-based approaches typically combined visual perception with policy learning for goal reaching or waypoint prediction. More recent methods increasingly adopt large-scale data, end-to-end training, and expressive action distributions to improve generalization across scenes and embodiments. Representative examples include GNM [15] and ViNT [16], which learn generalizable

visual navigation policies from diverse data, and NoMaD [14], which uses diffusion modeling to represent multi-modal navigation actions.

Recent work further pushes navigation toward longer horizons, stronger semantic grounding, and broader task generalization. LH-VLN [17] studies long-horizon vision-language navigation with new benchmarks and evaluation protocols, while NavFoM [18] scales navigation modeling across tasks and embodiments. Large-model-based navigation methods have also become increasingly common. VLN-R1 [19] explores reinforcement fine-tuning for vision-language navigation, and ImagineNav++ [20] uses imagined future views and large vision-language models to improve planning and reasoning. Other recent methods investigate foundation-model-based action selection, goal-distance estimation, and proactive prediction for navigation [54, 55, 56].

These methods are primarily policy-centric. Their main objective is to map observations, instructions, or goals to actions that complete a navigation task. The present work studies a different but complementary question. Rather than asking whether a model predicts a good action, it asks whether an action-conditioned world model produces rollouts whose induced dynamics remain faithful to the composition structure of the actions themselves. This distinction matters because a policy may be useful for goal reaching, and a generated rollout may appear visually plausible, while the underlying motion evolution still fails to preserve identity, inverse, and composition relations under repeated rollout. Our focus is therefore not visual navigation policy learning, but the structural action faithfulness of action-conditioned dynamics.

G.2 Latent and Diffusion-Based World Models

More broadly, multi-modal representation learning has received increasing attention, with extensive studies on cross-modal alignment, fine-grained matching, and robust text-image representation learning [57, 58, 59, 60, 61, 62, 63]. Recent work further extends multi-modal learning toward incomplete-data robustness, domain generalization, uncertainty-aware reasoning, and few-shot vision-language transfer [64, 65, 66, 67, 68].

The study of action-conditioned dynamics is closely related to model-based reinforcement learning, where learned transition models are used for planning, policy learning, and long-horizon decision making. Latent world models such as Dreamer and its successors [21, 22] demonstrate that imagined rollouts in a compact latent space can support efficient control across diverse domains. Related work studies reconstruction-free latent imagination [69], memory-augmented latent dynamics [70], and recent applications of latent world models to embodied planning and autonomous driving [71, 72]. These methods establish that learned latent dynamics can be useful for downstream control and planning, and they motivate the use of latent rollout as an efficient computational substrate.

However, the objectives in model-based reinforcement learning are usually organized around control performance, sample efficiency, or planning utility. They do not explicitly ask whether the learned transition respects the algebraic relations of the action space. In contrast, the present work focuses on a structural property of the rollout dynamics themselves. The question is not only whether latent imagination is useful, but whether the action-conditioned transition remains compatible with the composition, identity, and inverse relations induced by physical motion.

Modern visual world models are also shaped by the rapid development of diffusion-based image and video generation. Denoising diffusion models [73] and latent diffusion models [74] provide scalable foundations for high-fidelity visual synthesis. Video diffusion models extend this paradigm to temporal prediction and controllable motion, including early video diffusion work [75, 76], latent video generation [77], motion adaptation [78], Stable Video Diffusion [41], and large-scale video generators such as CogVideoX [11] and Open-Sora [79]. As these models improve in fidelity, duration, and temporal coherence, they become increasingly viable backbones for action-conditioned visual rollout.

Several recent systems adapt pretrained or large-scale video diffusion models to interactive or embodied settings. Vid2World [23] transfers pretrained video diffusion models into interactive world models, while DreamZero [24] builds a world-action model on top of a pretrained video diffusion backbone. These systems reflect a broader trend: video generators are increasingly expected to function not only as visual synthesizers, but also as simulators conditioned on actions. The present work builds on this transition but targets a different property from perceptual quality. Visual fidelity, temporal smoothness, and controllability do not by themselves guarantee that generated motion is

governed by the input actions. Our focus is therefore not generative realism alone, but whether learned rollout dynamics preserve the local observable consequences of action composition.

G.3 Embodied World Models for Navigation and Simulation

World models approach embodied decision making by predicting future observations conditioned on candidate actions. This paradigm has a long history in model-based reinforcement learning, where latent dynamics models such as Dreamer [21, 22] support planning and policy learning through imagined rollouts. More recently, similar ideas have been extended to visual navigation and embodied simulation, where future-state generation is used not only for representation learning but also for planning, action selection, and counterfactual evaluation.

Navigation World Models [13] show that controllable video generation can support action-conditioned rollout and planning in navigation. This line of work is especially relevant to our setting because navigation actions correspond to physical motion in the plane, making the structure of action composition directly meaningful. Recent work further broadens embodied world modeling through geometry-aware prediction, language-conditioned future modeling, general-purpose simulation, and action-centered world-action modeling. Aether [25] combines geometry-aware reconstruction, action-conditioned prediction, and visual planning in a unified framework. Language-conditioned world modeling for visual navigation [26] studies future prediction, language grounding, and action generation jointly. UniSim [27] explores large-scale video models as general-purpose embodied simulators. Other systems such as UniWM [80], NavForesee [56], MoWM [81], GigaWorld-Policy [82], DreamZero [24], and MTV-World [83] investigate unified world modeling, planning-oriented future imagination, or action-centered world-action modeling for embodied control. Complementary progress in autonomous-driving and geometric perception has also emphasized structured scene representations, including lane-graph learning and vectorized map construction [84, 85, 86], as well as general geometric reconstruction and depth estimation [87, 88, 89, 90]. Although these works do not directly study action-conditioned world dynamics, they highlight the importance of structured geometric representations for scene understanding and motion reasoning in embodied systems.

These studies demonstrate the promise of world models for embodied decision making. At the same time, visual plausibility and planning utility do not by themselves establish that learned dynamics faithfully follow actions. A model may generate realistic-looking futures and still allow the same action sequence to induce inconsistent motion across stochastic rollouts. It may also drift under zero action, fail to cancel a motion with its inverse, or produce incompatible endpoints under locally equivalent decompositions. The issue examined in the present work is therefore more specific than embodied world modeling in general. We ask whether action-conditioned rollout remains consistent with action composition, and how failures of this consistency appear as trajectory instability. This places the proposed framework between generative world modeling and dynamical consistency analysis.

G.4 Action Consistency and World-Model Evaluation

As world models become more capable, their evaluation has expanded beyond perceptual metrics, reconstruction quality, and short-horizon prediction error. Recent work has begun to ask whether generated worlds are useful for downstream decision making and whether they remain reliable under interaction. WorldSimBench [28] evaluates predictive video models from perceptual and embodied perspectives, including action-level consistency in downstream tasks. WoW-World-Eval [29] studies embodied world models through perception, planning, prediction, generalization, and execution. WorldArena [30] highlights the gap between perceptual quality and embodied functional utility, while WorldLens [31] evaluates driving world models along multiple axes including action-following and downstream reliability. These benchmarks reflect an important shift from asking whether generated videos look realistic to asking whether they function as useful and reliable simulations.

A closely related line of work focuses more directly on controllability, memory, and rollout reliability. MIND [32] evaluates action control and memory consistency in closed-loop rollouts. Toward Memory-Aided World Models [33] emphasizes spatial consistency under long-horizon navigation rollouts. Recent consistency-oriented world modeling work [34, 35] studies state alignment and action following beyond one-step prediction quality, while Mobile World Models [36] highlights the gap between visually plausible prediction and action-conditioned consistency and proposes strategies

to improve multi-step rollout behavior. Together, these studies show growing awareness that visual realism is insufficient for world modeling, especially when generated rollouts are used for planning or decision making.

The present work is aligned with this emerging direction but differs in how the problem is formulated. Existing evaluations often treat action consistency as a benchmark dimension, an action-following score, or a global rollout symptom. Such evaluations are valuable, but they do not explicitly formalize the local action relations whose violations may contribute to long-horizon rollout instability. In contrast, this work asks which local relations should be preserved for a navigation world model to be action-faithful. The group-action view decomposes action consistency into identity preservation, inverse cancellation, and local composition consistency. These properties are not arbitrary probes; they are local observable consequences of the group action induced by planar ego-motion. The proposed GAC metric evaluates these local conditions, while GAR measures whether repeated stochastic rollouts remain robust under the same action sequence. This provides a structured explanation of why repeated rollouts can diverge even when generated frames remain visually plausible.

G.5 Positioning of the Proposed Work

The proposed work is closest to embodied world modeling, action-conditioned video generation, and recent efforts on evaluating world-model consistency. Existing studies have begun to move beyond perceptual realism by measuring spatial coherence, memory consistency, action controllability, or trajectory-level stability. These directions reveal important failures of visually plausible simulators. Our work shares this motivation, but places a different criterion on action-conditioned dynamics: rather than treating trajectory inconsistency only as a rollout-level outcome, we ask whether the model preserves the local action-composition structure from which stable rollouts should arise.

This criterion distinguishes the proposed framework from broad embodied world-model benchmarks and trajectory-consistency evaluations. Trajectory dispersion measures whether repeated stochastic rollouts under the same action sequence remain close in recovered state space, but it does not specify which relations of the action space are violated. We instead start from the group action of planar ego-motion and instantiate its local observable consequences through identity consistency, inverse consistency, and local composition consistency. GAC probes these local relations under controlled transformations, while GAR separately measures rollout-level robustness.

The contribution is therefore a group-action formulation of action faithfulness and its operationalization for navigation visual world models. The same local conditions that support GAC are also converted into latent rollout constraints, providing a practical training signal that improves action-conditioned rollout stability without changing the backbone architecture. In this sense, the proposed work complements existing world-model evaluation efforts by connecting rollout-level instability to local action-composition structure and by showing that this structure can guide model improvement.

H Discussion on Broader Impact

The broader impact of this work is primarily conceptual. It argues that action-conditioned world models should be viewed not only as video generators conditioned on action tokens, but as simulators whose generated dynamics should remain organized by the actions or transformations they receive. This adds a structural dimension to world-model evaluation beyond visual realism and task performance. The proposed metrics do not certify real-world deployment safety or replace closed-loop validation. Rather, they expose action-faithfulness failures that may remain hidden under perceptual metrics or short-horizon prediction losses, providing an additional lens for analyzing world models before they are used in planning, embodied evaluation, or simulation-based testing.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses scope and limitations in Conclusion and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The theoretical result is stated with its assumptions in Sec. 3. A complete proof of Theorem 1 is provided in Appendix A.1.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: The paper provides the information needed to reproduce the main experiments, including datasets, model variants, training setup, pose recovery, GAC/GAR evaluation protocols, probe configurations, and metric aggregation details in Sec. 5.1 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code is not yet publicly released, but the paper provides detailed experimental settings, datasets, training protocols, evaluation metrics, and aggregation procedures in Sec. 5.1 and Appendix C. We plan to release the code and reproduction scripts with the final publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies the experimental setup in Sec. 5.1, including datasets, baselines, GAC probe configurations, GAR evaluation, and training protocol. Additional details on data splits, model adaptation, hyper-parameters, optimizer, pose recovery, and metric aggregation are provided in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All quantitative experiments report standard deviations in the main paper and appendix, including GAC, GAR and ablation results. The deviations are computed across evaluation sequences or repeated stochastic rollouts, as described in Sec. 5.1 and Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the compute setup in Appendix C.3, including GPU type, number of GPUs, batch size and training configuration.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses broader impacts in Appendix.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new high-risk models, scraped datasets, or general-purpose generative systems requiring special safeguards. The work uses existing research datasets and focuses on evaluation and training protocols for action-conditioned world models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper credits the original creators of the datasets, models, and tools used in the experiments, including RECON, SCAND, HuRoN, NWM, DIAMOND, and DROID-SLAM. The assets and their use are summarized in Appendix C.8, and no new dataset is collected or redistributed.

    Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new assets at submission time. The proposed method and evaluation protocol are documented in the paper and appendices, and code release is planned upon acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or new research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or new research with human subjects, so IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as important, original, or non-standard components. Any use of LLMs for writing, editing, or formatting, if applicable, does not affect the methodology, experiments, or scientific claims.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.