A Planner-Centered Study of Robust Latent World Models

Anonymous Authors

Abstract—Latent world models are reusable predictive components for robot planning, but visual nuisance shifts can change the latent costs used by model-predictive planners. We study whether generic latent-invariance objectives make LeWorldModel-style joint-embedding planning robust under observation shift. In a controlled evaluation, all rows share the same architecture, data source, optimizer family, and CEM planner; auxiliary rows add a paired clean/shifted training branch and differ by the alignment objective, while rows marked "dec." are explicit schedule ablations. The result is conditional. On a low-signal navigation task, equivariant InfoNCE and Max-sliced-Wasserstein alignment improve robust competence by approximately +0.037 and +0.019, respectively, under uncorrected exact paired tests, but these gains also improve clean success and are best interpreted as representation regularization rather than pure visual-shift repair. On Reacher and Reacher-Distractor, the same objectives do not yield statistically reliable robust-planning gains and can regress. A planner-margin diagnostic explains the failure mode: alignment reduces clean–shift latent separation, but the half-margin ratio $R_{\text{margin}} = 2\delta_J/\delta_a$ remains between 42 and 249 for the clean-non-collapsed methods. Robust latent world-model planning therefore requires planner-level action stability, not only representation-space invariance.

I. Introduction

Foundation models now appear in many robot-planning stacks, including language, vision-language, diffusion, and world-model components. A latent world model plays a specific role in this stack: after self-supervised training on image-action sequences, it gives planners a reusable predictive state space for scoring candidate actions. This is useful because planning queries can be long horizon, interactive, and counterfactual. It also makes robustness hard to ignore. If a benign visual shift changes the learned latent dynamics, a planner may rank the wrong action sequence even though the physical task has not changed.

We study this issue for joint-embedding world models, using LeWorldModel (LeWM) as the base architecture [15]. LeWM learns from pixels with next-embedding prediction and a projection-based Gaussian regularizer, avoiding reconstruction decoders, exponential-moving-average targets, and stop-gradient heuristics. That simplicity makes it a useful testbed for a narrow question: when added to the same LeWM training and planning stack, do auxiliary invariance losses make latent world-model planning robust under observation shift?

We call this controlled evaluation framework Robust LeWM (RoLeWM). The LeWM baseline is the clean-observation LeWM objective from the same implementation. Auxiliary rows use the same encoder, predictor, data source, optimizer family, and CEM planner, but add the paired shifted-observation branch needed to evaluate an invariance loss. They should therefore be read as same-architecture auxiliary-training variants, not as a claim that every training loss term except $\mathcal{L}_{\text{aux}}$ is identical to vanilla LeWM. Rows explicitly marked "dec." are schedule ablations. The core transport variant aligns clean and shifted action-conditioned latent rollout increments with a sliced-Wasserstein loss; the evaluation compares it with direction-concentrated variants and same-stack non-transport objectives such as VICReg-seq, dynamics-bisimulation-style alignment, and equivariant InfoNCE. The main metric is robust competence,

$$S_{\text{robust}} = \frac{1}{n}\sum_i \min(S_{\text{clean},i}, S_{\text{shift},i}), \tag{1}$$

where success is the fraction of evaluation episodes in which the planned trajectory reaches the specified goal within the task tolerance, $S_{\text{clean},i}$ is success on unmodified observations, and $S_{\text{shift},i}$ is success averaged over brightness, contrast, blur, and noise shifts for seed i. This metric avoids the failure mode of pure retention scores, which can reward methods that collapse both clean and shifted success. For example, a planner with S_{clean}=S_{shift}=0.02 has perfect retention but poor robust competence.

The results show a boundary, but not a simple "visual shift breaks planning and invariance fixes it" story. In displayed means, LeWM is not consistently worse under the averaged visual shifts: on tworoom, S_{clean}=0.017 and S_{shift}=0.040; on reacher, S_{clean}=0.483 and S_{shift}=0.498. The low-signal tworoom gains therefore reflect a broader representation-learning benefit: equivariant InfoNCE improves both clean and shifted success. In the competent reacher regime, none of the n=16 headline auxiliary rows significantly improves S_{robust}, and the n=8 supporting controls also show no reliable gain on matched seeds. On reacher_distractor, no method yields a statistically reliable improvement over LeWM across n=20 paired training seeds; the best mean, Eq. InfoNCE + dec., is only +0.006 in S_{robust} with p=0.83. Top-16 sliced-Wasserstein shows an uncorrected regression signal, and VICReg-seq robustly regresses. A planner-margin diagnostic explains why latent alignment does not guarantee retained planning: the residual shift-induced cost drift remains far larger than the gap between the best and second-best CEM candidates. Throughout, "low-signal" and "competent-baseline" refer to the empirical clean-planning competence

of the same LeWM stack with latent dimension d=128 on a task, not to different model sizes or separately selected checkpoints.

Contributions. First, we provide a same-stack protocol for testing whether auxiliary latent-invariance losses improve world-model planning under visual shifts. Second, we use robust competence, a seed-wise minimum of clean and shifted success, to avoid retention scores that reward uniformly poor planners. Third, we show empirically that invariance helps only in a weak-baseline regime and does not reliably transfer to competent or distractor-heavy regimes. Fourth, we provide a planner-margin diagnostic explaining why reduced latent separation can still fail to preserve CEM action selection.

Relation to prior work. Unlike robust visual-control work based on domain randomization, adversarial augmentation, or pretrained visual encoders, this study keeps the LeWM-style architecture, data source, and planner fixed while comparing auxiliary shift-paired training variants. Unlike representation-only alignment studies, our primary outcome is closed-loop planning success and candidate-ranking stability. These diagnostics are relevant to language, vision-language, diffusion-policy, and generalist robot systems [1, 5–7, 29] because high-level semantic planners and policies still rely on low-level perception or predictive components whose action rankings can be visually unstable.

II. Method

World-model stack. Given image observations o_t, actions a_t, an encoder f_θ produces $z_t = f_\theta(o_t) \in \mathbb{R}^d$, and a predictor g_ϕ predicts the next latent $\hat{z}_{t+1} = g_\phi(z_t, a_t)$. LeWM trains with a next-embedding prediction loss and SIGReg, a projection-based isotropic Gaussian regularizer that keeps latent batches close to an isotropic distribution [3, 15]. For non-decoupled auxiliary variants, training minimizes

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{rob}}\mathcal{L}_{\text{rob-pred}} + \lambda_{\text{sig}}\mathcal{L}_{\text{sig}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}, \quad (2)$$

where

$$\mathcal{L}_{\text{pred}} = \frac{1}{B}\sum_{b=1}^{B}\left\| g_\phi(f_\theta(o_t^{(b)}), a_t^{(b)}) - f_\theta(o_{t+1}^{(b)})\right\|_2^2. \quad (3)$$

$\mathcal{L}_{\text{rob-pred}}$ is the same prediction loss on shifted observations, $\mathcal{L}_{\text{sig}}$ is the LeWM Gaussian regularizer on clean and shifted latent batches, and $\mathcal{L}_{\text{aux}}$ is one of the invariance losses below. The LeWM baseline omits the shifted branch and uses only $\mathcal{L}_{\text{pred}} + \lambda_{\text{sig}}\mathcal{L}_{\text{sig}}$ on clean observations. Given a goal observation o^g with latent $z^g = f_\theta(o^g)$ and an action sequence $a_{0:H-1}$, CEM uses

$$J_H(z, z^g, a_{0:H-1}) = \min_{1\le h\le H} \|\hat{z}_{t+h}(a_{0:h-1}) - z^g\|_2^2, \quad (4)$$

where $\hat{z}_{t+h}$ is obtained by recursively applying g_ϕ from $z = f_\theta(o_t)$. At shifted evaluation, the current observation is transformed by the selected shift family before encoding, while the goal observation remains clean. This tests the harder current-view shift setting rather than jointly shifting current and goal images.

Shifted paired rollouts. For each training snippet, we sample a visual shift $\mathcal{T}_\xi$ and form paired observations $\tilde{o}_t = \mathcal{T}_\xi(o_t)$. The clean and shifted branches share the same action sequence. This pairing compares the model's predicted rollouts for an identical candidate sequence; after CEM re-scores candidates under shifted observations, the selected action sequence can still change. Starting from the encoded latent z_t, the clean rollout-increment stack for horizon H is

$$r_t^{(H)} = [\hat{z}_{t+1} - z_t; \hat{z}_{t+2} - \hat{z}_{t+1}; \ldots; \hat{z}_{t+H} - \hat{z}_{t+H-1}], \quad (5)$$

and $\tilde{r}_t^{(H)}$ is defined analogously for shifted observations. Aligning increments rather than absolute latent states is intended to reduce sensitivity to static nuisance offsets caused by photometric shifts.

Sliced-Wasserstein dynamics alignment. Each rollout-increment stack lies in $\mathbb{R}^{Hd}$. Let $\mathbb{P}^{(H)} = \frac{1}{B}\sum_{b=1}^{B}\delta_{r_{t,b}^{(H)}}$ and $\widetilde{\mathbb{P}}^{(H)} = \frac{1}{B}\sum_{b=1}^{B}\delta_{\tilde{r}_{t,b}^{(H)}}$ be the minibatch empirical distributions of clean and shifted stacks. The transport loss is

$$\mathcal{L}_{\text{sw}} = \frac{1}{M}\sum_{m=1}^{M} W_2^2\Big(\langle\omega_m,\cdot\rangle_\#\mathbb{P}^{(H)}, \langle\omega_m,\cdot\rangle_\#\widetilde{\mathbb{P}}^{(H)}\Big), \quad (6)$$

where $\omega_m \in \mathbb{S}^{Hd-1}$ are unit projection directions, $\langle\omega_m,\cdot\rangle_\#$ denotes the pushforward distribution under projection onto ω_m, and W_2^2 is the squared 2-Wasserstein distance, computed exactly in one dimension by sorting. For top-k SW, we sample M_{cand} directions, compute projected costs c_m, and use

$$\mathcal{L}_{\text{top-k}} = \frac{1}{k}\sum_{m\in\text{TopK}(c_1,\ldots,c_{M_{\text{cand}}})} c_m.$$

For MaxSW, the projection direction is optimized for $T_{\max}$ gradient steps from R random restarts, and the largest projected cost is used. All non-decoupled auxiliary variants keep the same encoder, predictor, optimizer, data, shift sampler, and CEM planner.

Other auxiliary losses. VICReg-seq applies the VICReg invariance, variance, and covariance penalties to paired clean/shifted rollout stacks, with weights $25, 25, 1$. Equivariant InfoNCE treats $(f_\theta(o_t), f_\theta(\tilde{o}_t))$ from the same physical state as positives and other minibatch states as negatives, uses cosine-normalized encoder outputs, a one-way in-batch cross-entropy, and temperature τ=0.1. DBC-style alignment penalizes disagreement between pairwise encoder-latent distances and pairwise one-step predicted-dynamics distances under the shared action. Because the offline goal-reaching data do not provide a reward label for this auxiliary objective, this row is a dynamics-bisimulation-style baseline rather than a full reward-aware DBC implementation.

Planner-margin condition. Let $\mathcal{A}_K$ be a finite candidate set and write the clean and shifted costs as $J_H(z, z^g, a)$ and $\widetilde{J}_H(\tilde{z}, \tilde{z}^g, a)$, where $z = f_\theta(o)$, $\tilde{z} = f_\theta(\mathcal{T}_\xi(o))$, and $\tilde{z}^g = z^g$ in our current-view shift evaluation.

Proposition 1 (Fixed-candidate action preservation). Let the clean cost have a unique minimizer a^* with runner-up margin

$$\delta_a = \min_{a \in \mathcal{A}_K \setminus \{a^*\}} J_H(z, z^g, a) - J_H(z, z^g, a^*) > 0.$$

If

$$\eta = \max_{a \in \mathcal{A}_K} \left| \widetilde{J}_H(\tilde{z}, \tilde{z}^g, a) - J_H(z, z^g, a) \right|$$

satisfies $2\eta < \delta_a$, then a^* is also the unique minimizer of $\widetilde{J}_H$ over $\mathcal{A}_K$.

Sketch: For any $a \neq a^*$,

$$\widetilde{J}_H(a) - \widetilde{J}_H(a^*) \geq J_H(a) - J_H(a^*) - 2\eta \geq \delta_a - 2\eta > 0.$$

Thus no candidate overtakes a^* under the shifted cost. ■

This condition is only sufficient and applies to the fixed evaluated candidate set, not to the full continuous action space or to independently resampled CEM runs. Nevertheless, it identifies the planner-facing quantity a robust world model must control: shift-induced cost drift relative to the action margin, rather than average latent alignment alone.

III. Experiments and Results

Protocol. We evaluate image-based goal-reaching tasks at 64×64 resolution: reacher, tworoom, and reacher_distractor. tworoom is a continuous 2D point-navigation task with two rooms separated by a wall and doorway, action space $[-1, 1]^2$, episode horizon 40, and success radius 0.08. reacher is Gymnasium MuJoCo Reacher-v5 rendered from pixels, with a 2-DoF arm, action space $[-1, 1]^2$, episode horizon 50, and success defined by fingertip-target distance $\|x_{\text{tip}} - x_{\text{goal}}\|_2 < 0.05$. reacher_distractor uses the same Reacher dynamics and success criterion but overlays a per-episode random RGB background. Training data are reward-free exploratory trajectories: tworoom actions follow a scripted goal-directed controller with Gaussian exploration noise of standard deviation 0.05, while reacher and reacher_distractor use low-pass-filtered random torques. No rewards or success labels are used during representation training. The train/validation episode counts are 800/120 for tworoom, 1000/150 for reacher, and 500/80 for reacher_distractor.

All non-decoupled methods use the same four-stage convolutional encoder, action-conditioned MLP predictor, AdamW optimizer, batch size, shift sampler, training schedule, and CEM planner. Rows marked "dec." are schedule ablations and are reported separately: stage 1 trains the LeWM objective for 33 epochs, then stage 2 freezes the encoder and trains the predictor for 17 additional epochs with the losses listed in Appendix B.

TABLE I

Main planning results under clean and shifted observations. Cells report mean ± s.e.; p is the uncorrected exact paired permutation test against LeWM on S_{robust}. Bold in S_{robust} columns marks paired-significant gains over LeWM on shared seeds. The matched n=8 baseline is shown because supporting controls are paired against the same seed subset. Deltas and p-values are computed from unrounded per-seed values.

Method	reacher				tworoom			
	S_{clean}	S_{shift}	S_{robust}	p	S_{clean}	S_{shift}	S_{robust}	p
LeWM (n=16)	$0.483_{\pm.045}$	$0.498_{\pm.024}$	$0.405_{\pm.029}$	–	$0.017_{\pm.010}$	$0.040_{\pm.009}$	$0.009_{\pm.005}$	–
LeWM matched (n=8)	$0.450_{\pm.055}$	$0.498_{\pm.036}$	$0.394_{\pm.041}$	–	$0.025_{\pm.018}$	$0.042_{\pm.012}$	$0.015_{\pm.010}$	–
Uniform SW (n=8)	$0.467_{\pm.033}$	$0.519_{\pm.019}$	$0.448_{\pm.027}$	.219	$0.025_{\pm.012}$	$0.058_{\pm.006}$	$0.017_{\pm.008}$	1.000
Top-16 SW (n=8)	$0.508_{\pm.047}$	$0.471_{\pm.025}$	$0.440_{\pm.024}$	.547	$0.042_{\pm.025}$	$0.067_{\pm.012}$	$0.025_{\pm.012}$	.625
MaxSW (n=16)	$0.467_{\pm.018}$	$0.454_{\pm.025}$	$0.424_{\pm.018}$	.620	$0.033_{\pm.009}$	$0.064_{\pm.010}$	$\mathbf{0.028}_{\pm.008}$	.016
VICReg-seq (n=8)	$0.467_{\pm.065}$	$0.444_{\pm.060}$	$0.390_{\pm.058}$	.969	$0.008_{\pm.008}$	$0.012_{\pm.004}$	$0.002_{\pm.002}$	.500
DBC-style (n=8)	$0.458_{\pm.041}$	$0.458_{\pm.017}$	$0.402_{\pm.024}$	.875	$0.008_{\pm.008}$	$0.046_{\pm.010}$	$0.008_{\pm.008}$	1.000
Equiv. InfoNCE (n=16)	$0.408_{\pm.039}$	$0.404_{\pm.018}$	$0.325_{\pm.020}$	.062	$0.087_{\pm.023}$	$0.084_{\pm.008}$	$\mathbf{0.046}_{\pm.011}$	.006
Eq. InfoNCE + dec. (n=16)	$0.487_{\pm.036}$	$0.497_{\pm.022}$	$0.445_{\pm.032}$	.365	$0.042_{\pm.012}$	$0.037_{\pm.008}$	$0.019_{\pm.007}$	.375

The CEM planner uses 256 candidates, 5 iterations, horizon 12, and the latent goal-distance cost above. Each evaluation uses clean observations and current-view shifted observations under brightness, contrast, blur, and noise; S_{shift} averages over these four families. Training and evaluation use the same shift families and severity ranges but independent episodes and independently sampled shift parameters, so the results test held-out robustness within the specified nuisance distribution, not extrapolation to unseen shift types. Each reported value is mean ± standard error over paired training seeds; the intervals do not include additional binomial uncertainty from the finite episode budget. Statistical comparisons against LeWM use exact paired two-sided permutation tests on per-seed S_{robust} differences with shared training and evaluation seeds. For n=8 supporting controls, tests use the LeWM values restricted to the same eight seeds. Because multiple auxiliary losses are tested against the same baseline, uncorrected p-values are shown in the main tables and interpreted as exploratory; Holm-corrected values are summarized in Appendix B.

Low-signal regime. On tworoom, equivariant InfoNCE improves S_{robust} by approximately +0.037 over LeWM (p=0.006) and MaxSW improves S_{robust} by +0.019 (p=0.016) under uncorrected exact paired tests. Equivariant InfoNCE also improves clean success from 0.017 to 0.087 and shifted success from 0.040 to 0.084. Because both terms improve together, the gain indicates better task competence under the same planner, not only improved retention under visual shift. Since the LeWM shifted score is already higher than its clean score, this result is not evidence that visual shift broke the baseline planner and the invariance loss repaired it. Under Holm correction over the seven within-task Table I auxiliary comparisons, equivariant InfoNCE remains significant and MaxSW should be read as an uncorrected exploratory signal. The more accurate interpretation is that, in a weak-baseline regime, the auxiliary invariance objective regularizes the representation and improves overall planning competence.

Competent-baseline regimes. On reacher, LeWM al-

TABLE II

reacher_distractor boundary at n=20. Cells report mean ± s.e.. Random RGB backgrounds make visual invariance useful in principle, but LeWM remains hard to beat under this protocol. No method significantly improves S_{robust}; top-16 SW has an uncorrected regression signal and VICReg-seq remains regressive after Holm correction over the six distractor comparisons. Bold marks the best mean in each metric, not a significant gain.

Method	S_{clean}	S_{shift}	$S^{\min}_{\text{shift}}$	S_{robust}	p vs. LeWM
LeWM	0.560±.026	0.512±.024	0.370±.034	0.488±.021	–
MaxSW	**0.570**±.032	0.492±.017	0.327±.025	0.471±.019	.571
Eq. InfoNCE + dec.	0.563±.030	**0.534**±.018	**0.417**±.026	**0.494**±.019	.828
DBC-style	0.500±.030	0.484±.018	0.353±.024	0.432±.021	.097
Equiv. InfoNCE	0.470±.028	0.486±.016	0.343±.023	0.422±.020	.062
Top-16 SW	0.477±.024	0.457±.019	0.317±.025	0.408±.017	.018 (reg.)
VICReg-seq	0.317±.028	0.331±.013	0.220±.015	0.267±.017	< .001 (reg.)

ready reaches S_{clean}=0.483 and S_{robust}=0.405, and its shifted success is slightly above clean success. Among the n=16 headline Reacher rows, none significantly improves S_{robust} over LeWM. The n=8 supporting controls also show no reliable gain on matched seeds. Equivariant InfoNCE alone reduces S_{robust} to 0.325, consistent with over-regularizing the encoder away from the clean prediction optimum. On reacher_distractor (Table II), LeWM remains competitive at n=20; MaxSW and equivariant InfoNCE plus decoupled training do not significantly improve it, while top-16 sliced-Wasserstein shows an uncorrected regression signal and VICReg-seq robustly regresses.

Mechanism. Figure 2 shows the core diagnostic. Latent alignment succeeds at its local goal: on reacher, top-16 SW reduces the normalized clean-vs-shift centroid distance from 0.089 to 0.0076, and VICReg-seq to 0.0018. We compute this distance as

$$D_{\text{cent}} = \frac{\|\mu_{\text{clean}} - \mu_{\text{shift}}\|_2}{\frac{1}{2}(\sigma_{\text{clean}} + \sigma_{\text{shift}}) + \epsilon},$$

where μ is the rollout-stack centroid and σ is the root-mean-square radius of the corresponding cloud. However, the planner needs action stability. The runner-up diagnostic compares the clean-best candidate drift

$$\delta_J = \left| \widetilde{J}_H(\tilde{z}, \tilde{z}^g, a^*) - J_H(z, z^g, a^*) \right|, \tag{7}$$

where a^* is the clean-best CEM candidate sequence. We report the normalized half-margin ratio

$$R_{\text{margin}} = \frac{\delta_J}{\delta_a/2} = \frac{2\delta_J}{\delta_a}.$$

This is a lower-bound diagnostic for the worst-case drift η: if $R_{\text{margin}} > 1$, then the sufficient certificate $2\eta < \delta_a$ is already impossible because $\eta \geq \delta_J$. Across clean-non-collapsed methods, R_{margin} remains between 42 and 249, so even the clean-best drift alone is far above the sufficient half-margin threshold. Thus the latent clouds can overlap much better while the CEM argmin still changes under shift. This is the main robustness lesson: representation alignment is not enough unless shift-induced cost drift is small relative to the planner's action margin.

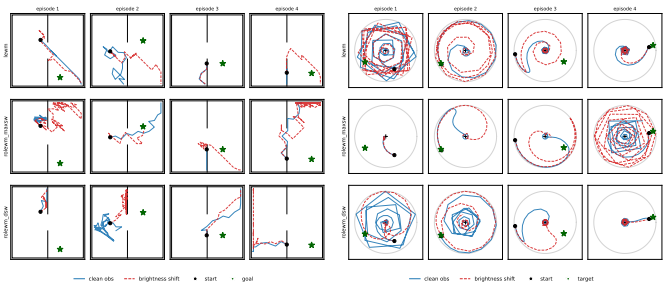

Fig. 1. Qualitative planning behavior. Left: tworoom rollouts show the low-signal regime where invariance improves planning competence together with clean success. Right: reacher rollouts show the higher-signal regime where the baseline planner already produces useful actions and extra invariance can trade off against clean predictive quality. These examples motivate reporting clean success, shifted success, and planner-margin diagnostics together rather than ranking methods by one robustness number.

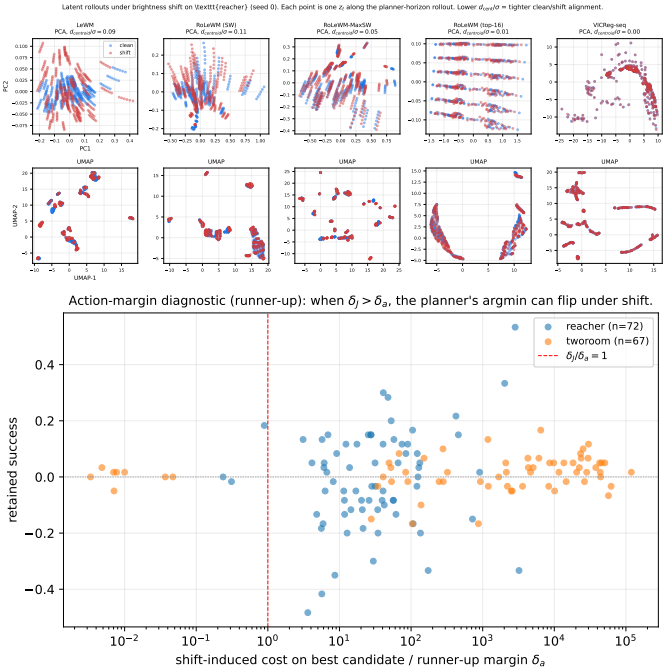

Fig. 2. Why latent tightening is not enough. Top: clean and shifted latent rollouts on reacher; lower normalized centroid distance means better clean-shift overlap. Alignment variants can shrink this geometric separation by one to two orders of magnitude. Bottom: CEM runner-up-margin diagnostic. The plotted ratio is $R_{\text{margin}} = 2\delta_J/\delta_a$; values larger than one already rule out the sufficient certificate $2\eta < \delta_a$. Larger values, farther right, mean the shift is large relative to the planner's decision margin and action rankings are easier to flip.

Qualitative examples in Fig. 1 illustrate the difference between the low-signal navigation regime and the competent Reacher regime.

IV. Discussion and Future Work

The results give a practical lesson for robot planning with learned world models: robustness should be evaluated through the planner, rather than through representation geometry alone. Sliced transport and contrastive align-

TABLE III

Planner-margin diagnostic on reacher. D_{cent} measures clean–shift rollout-cloud separation; $R_{\text{margin}} = 2\delta_J/\delta_a$ measures clean-best cost drift relative to the half-margin. VICReg-seq is shown for completeness but is not evidence of robust planning because task competence collapses.

Method	D_{cent}	R_{margin}	Action preservation
LeWM	0.089	171.2	0.534
Top-16 SW	0.0076	83.6	0.474
Equiv. InfoNCE	–	42.3	0.725
Eq. InfoNCE + dec.	–	249.1	0.467
VICReg-seq	0.0018	0.68	0.969

TABLE IV

Condensed design targets. The same results suggest different fixes depending on the observed failure mode.

Observation	Interpretation	Next target
Clean and shifted success rise together	Invariance regularizes a weak planner	Improve representation and data coverage
Latents align but CEM changes actions	Cost drift exceeds action margin	Train with margin-aware losses
Clean success drops under alignment	Auxiliary loss competes with prediction	Stage or reweight objectives

ment can make clean and shifted rollouts much closer, but CEM still needs enough action margin to preserve the selected sequence. This separates appearance robustness from planning robustness. The former asks whether shifted images map to similar latents; the latter asks whether the planner keeps choosing the same or equally successful actions.

The experiments also separate failure modes that should not be collapsed into one robustness score. When clean success is near zero, retained performance is not meaningful. When clean and shifted success improve together, as in tworoom, the gain is better interpreted as representation regularization than as pure shift repair. When the baseline is already competent, as in reacher, auxiliary invariance can reduce latent shift while leaving CEM rankings unstable or even lowering clean competence.

Future work should therefore make robustness objectives planner-aware. A margin-aware loss could penalize shift-induced cost drift most strongly near states with small CEM runner-up margins. Larger latent capacity, stronger data collection, curriculum or expert trajectories, and uncertainty-aware CEM should be evaluated by whether they increase action stability, not only by whether they reduce latent distance. In foundation-model robot-planning stacks, semantic planners can propose goals and subgoals, but the low-level world model still needs to expose when visual shifts make its candidate rankings unreliable.

Design implications. The same diagnostic suggests three concrete interventions, summarized in Table IV. First, training losses should be local to the planner: global latent alignment is useful only when it reduces cost drift for action sequences that CEM might actually select. Second, robustness objectives should preserve clean competence, because over-regularizing the encoder can reduce prediction quality even when shifted latents look better aligned. Third, data coverage must be checked before robustness claims are made; if offline data rarely visits successful goal-reaching behavior, every method can fail for reasons unrelated to visual shift.

Benchmark implications. A useful benchmark should report clean success, shifted success, robust competence, and action-preservation diagnostics together. These quantities answer different questions: whether the model can solve the unshifted task, whether the visual shift actually changes task performance, whether a method improves both clean and shifted competence, and whether CEM candidate rankings are stable. This also makes comparisons to stronger visual backbones or semantic foundation-model planners more meaningful: the relevant test is whether they increase planner margins and retained action quality, not only whether they improve visual features.

Deployment implications. In a receding-horizon robot controller, the planner can evaluate multiple augmented views of the current observation and measure candidate-ranking disagreement before acting. High-margin actions can be executed normally; low-margin actions can trigger additional sensing, uncertainty-aware replanning, or a higher-level semantic query. This turns the offline margin diagnostic into a runtime confidence signal for world models used inside larger robot-planning systems.

This framing is intentionally conservative. It does not treat any single invariance objective as the robust-planning solution. Instead, it asks whether an objective improves competence, whether it preserves CEM rankings, and whether it does so without reducing clean prediction quality. Those checks are the minimum needed before scaling the approach to larger encoders, longer horizons, or manipulation tasks with harder data-coverage requirements.

Open questions. The main open question is how to train invariance without erasing information that is useful for prediction. One possibility is to separate appearance-sensitive latents from dynamics latents, allowing the planner to query a stable subspace while preserving visual details for contact, object identity, or replanning. Another is to apply invariance only near candidate trajectories whose CEM costs are close, because those are the states where small visual perturbations are most likely to change the selected action.

A second question is how much of the observed failure is caused by model scale versus data coverage. Larger encoders may reduce prediction error and increase action margins, but they may also make the planner more sensitive to nuisance correlations if the data do not cover the relevant goals. The evaluation should therefore couple scaling studies with stronger goal-reaching data, including scripted, expert, or curriculum trajectories.

A third question is how to expose these diagnostics to

higher-level planners. Language or vision-language models can choose subgoals, but they usually do not know when the low-level latent dynamics are visually unstable. Passing action-margin or candidate-ranking uncertainty upward would let a semantic planner ask for a different view, a safer intermediate goal, or a new low-level plan before failure occurs.

Scope of claims. The evidence supports a conditional conclusion rather than a universal method ranking. The positive tworoom result shows that invariance can help a weak same-stack planner, while the reacher and reacher_distractor results show that the same pressure does not automatically transfer to competent or distractor-heavy regimes. The mechanistic diagnostic explains this pattern in planner terms: reducing average latent separation is useful, but robust planning depends on whether the remaining cost drift is small compared with the action margin used for selection.

This scope is also useful for future foundation-model evaluations. A larger visual encoder, a pretrained representation, or a language-conditioned goal proposal may improve clean competence, but it should still be tested under the same planner-centered criteria. The relevant question is not only whether representations become more invariant, but whether the planner's selected action remains stable and successful under the observation changes expected at deployment.

V. Conclusion

Same-stack invariance improves planning competence on a low-signal navigation task, but the improvement is not purely a robustness effect: clean and shifted success improve together. In competent-baseline regimes, the same class of objectives does not reliably improve planning under shift and can reduce clean competence. The mechanistic reason is concrete: latent alignment reduces representation shift, yet the remaining cost drift can exceed the CEM action margin by orders of magnitude. Robust world-model planning therefore requires planner-level action stability, not just latent invariance.

Limitations

The study uses compact simulated tasks, 64×64 observations, and LeWM-style ConvNet encoders. It does not certify worst-case robustness, does not include real-robot deployment, and does not establish that the conclusions hold for large pretrained VLA/world-model systems. The visual shifts are controlled nuisance shifts rather than a full deployment distribution. Reported standard errors are over training seeds; with 15 episodes per seed and shift-family cell, finite-evaluation noise remains a residual source of uncertainty.

AI Usage Statement

Generative AI tools were used for language editing and code-review assistance. All experiment design, implementation choices, result verification, and scientific claims were checked and approved by the authors.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances, 2022. URL https://arxiv.org/abs/2204.01691.

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499.

[3] Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics, 2025. URL https://arxiv.org/abs/2511.08544. arXiv: 2511.08544.

[4] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. Transactions on Machine Learning Research, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=QaCCuDfBk2. Featured Certification.

[5] Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X. Lee, Maria Bauza, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, Antoine Laurens, Claudio Fantacci, Valentin Dalibard, Martina Zambelli, Murilo Martins, Rugile Pevceviciute, Michiel Blokzijl, Misha Denil, Nathan Batchelor, Thomas Lampe, Emilio Parisotto, Konrad Zolna, Scott Reed, Sergio Gomez Colmenarejo, Jon Scholz, Abbas Abdolmaleki, Oliver Groth, Jean-Baptiste Regli, Oleg Sushkov, Tom Rothorl, Jose Enrique Chen, Yusuf Aytar, Dave Barker, Joy Ortiz, Martin Riedmiller, Jost Tobias Springenberg, Raia Hadsell, Francesco Nori, and Nicolas Heess. RoboCat: A self-improving foundation agent for robotic manipulation. Transactions on Machine Learning Research, 2023. URL https://openreview.net/forum?id=vsCpILiWHu.

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics transformer for real-world control at scale. In Robotics: Science and Systems, 2023. URL https://roboticsproceedings.org/rss19/p025.html.

[7] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. Int. J. Rob. Res., 44(10–11):1684–1704, September 2025. ISSN 0278-3649. doi: 10.1177/02783649241273668. URL https://doi.org/10.1177/02783649241273668.

[8] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. IEEE Trans. Pattern Anal. Mach. Intell., 39(9):1853–1865, September 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2615921. URL https://doi.org/10.1109/TPAMI.2016.2615921.

[9] David Ha and Jürgen Schmidhuber. World models. arXiv preprint arXiv:1803.10122, 2018. doi: 10.5281/ZENODO.1207631. URL https://zenodo.org/record/1207631.

[10] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024. URL https://arxiv.org/abs/2301.04104. arXiv: 2301.04104.

[11] Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=Oxh5CstDJU.

[12] Soheil Kolouri, Kimia Nadjahi, Umut Şimşekli, Roland Badeau, and Gustavo K. Rohde. Generalized sliced wasserstein distances. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2019. Curran Associates Inc.

[13] Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde. Sliced wasserstein auto-encoders. In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=H1xaJn05FQ.

[14] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrep-

ancy for unsupervised domain adaptation. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10277–10287, 2019. doi: 10.1109/CVPR.2019.01053.

[15] Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels, 2026. URL https://arxiv.org/abs/2603.19312. arXiv: 2603.19312.

[16] Sloan Nietert, Ritwik Sadhu, Ziv Goldfeld, and Kengo Kato. Statistical, robustness, and computational guarantees for sliced wasserstein distances. In Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[17] Joseph Ortiz, Antoine Dedieu, Wolfgang Lehrach, J Swaroop Guntupalli, Carter Wendelken, Ahmad Humayun, Sivaramakrishnan Swaminathan, Guangyao Zhou, Miguel Lazaro-Gredilla, and Kevin Patrick Murphy. DMC-VB: A benchmark for representation learning for control with visual distractors. In The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2024. URL https://openreview.net/forum?id=0NQzQVu9tY.

[18] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In 2018 IEEE International Conference on Robotics and Automation (ICRA), page 1–8. IEEE Press, 2018. doi: 10.1109/ICRA.2018.8460528. URL https://doi.org/10.1109/ICRA.2018.8460528.

[19] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, page 2817–2826. JMLR.org, 2017.

[20] Sai Prasanna, Karim Farid, Raghu Rajan, and André Biedenkapp. Dreaming of many worlds: Learning contextual world models aids zero-shot generalization. In Seventeenth European Workshop on Reinforcement Learning, 2024. URL https://openreview.net/forum?id=zHt4K5zX4P.

[21] Max Rudolph, Caleb Chuck, Kevin Black, Misha Lvovsky, Scott Niekum, and Amy Zhang. Learning action-based representations using invariance, 2024. URL https://arxiv.org/abs/2403.16369.

[22] Vlad Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim G. J. Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models. In The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2026. URL https://openreview.net/forum?id=hTZ0SJCGQX.

[23] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), page 23–30. IEEE Press, 2017. doi: 10.1109/IROS.2017.8202133. URL https://doi.org/10.1109/IROS.2017.8202133.

[24] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In International Conference on Learning Representations, 2018. URL https://openreview.net/forum?id=HkL7n1-0b.

[25] Leonardo F. Toso, Davit Shadunts, Yunyang Lu, Nihal Sharma, Donglin Zhan, Nam H. Nguyen, and James Anderson. Learning invariant visual representations for planning with joint-embedding predictive world models, 2026. URL https://arxiv.org/abs/2602.18639.

[26] Jialong Wu, Haoyu Ma, Chaoyi Deng, and Mingsheng Long. Pre-training contextualized world models with in-the-wild videos for reinforcement learning. In Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[27] Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction, 2021. URL https://arxiv.org/abs/2006.10742.

[28] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum?id=D5RNACOZEI.

[29] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Quan Vuong, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Jialin Wu, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In Proceedings of The 7th Conference on Robot Learning, volume 229 of Proceedings of Machine Learning Research, pages 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

Appendix A
Related Work

World models and latent planning. World models learn compact predictive state representations for planning and control [9–11]. Recent latent-planning systems avoid pixel reconstruction and instead predict in feature space, including DINO-WM and reward-free latent planning [22, 28]. Joint-embedding predictive architectures such as I-JEPA, V-JEPA, LeJEPA, and LeWorldModel further remove contrastive negatives or pixel decoders from the core predictive objective [2–4, 15].

Robust robot learning. Robustness to visual shift has been studied through domain randomization, adversarial training, contextual world models, and visual-control benchmarks with distractors [17–20, 23, 26]. The present study differs by holding the architecture and planner fixed and asking whether shift-paired auxiliary training improves retained planning.

Bisimulation, equivariance, and transport. Bisimulation metrics align states with similar rewards and transition behavior [21, 25, 27]. Equivariant or contrastive state-pairing losses align clean and shifted views of the same underlying state. Optimal transport and sliced-Wasserstein distances have been used for domain alignment and representation learning [8, 12–14, 16, 24]. RoLeWM applies sliced transport to action-conditioned rollout increments rather than static feature clouds.

Connection to foundation-model robot planning. Learned world models complement language- and vision-language-based planners. Semantic models can propose symbolic subgoals, while a latent world model scores whether a candidate low-level action sequence is dynamically plausible from the current observation. This division makes robustness important. A symbolic planner can replan after a failed high-level step, but a model-predictive controller that repeatedly queries a brittle latent model may introduce visual-shift errors at every control cycle.

Why same-stack comparison matters. Robustness studies often combine several changes at once: more data, larger encoders, external pretraining, data augmentation, modified planning costs, or new policy heads. Those changes are useful for systems, but they obscure the mechanism. The same-stack protocol keeps the world-model capacity and planner fixed, while auxiliary rows add the paired shifted branch needed for their alignment losses. This explains the paired positive and negative results: an invariance loss can regularize a weak baseline on tworoom and hurt a competent baseline on reacher without contradicting its local geometric effect.

Appendix B
Additional Protocol Details

The main experiments use 64×64 RGB observations, a four-stage convolutional encoder with latent dimension d=128, and a three-layer action-conditioned MLP predictor. Training uses AdamW with learning rate 3×10^{-4}, batch size 128, and 15 epochs unless a decoupled schedule is explicitly named. Metric definitions are summarized in Table VIII.

The CEM planner samples 256 action sequences, runs 5 refinement iterations, and evaluates horizon 12. For each method and seed, success is measured on clean observations and current-view shifted observations. The main evaluation shift families are brightness, contrast, blur, and noise. At default severity, brightness and contrast factors are sampled from $[0.5, 1.5]$, Gaussian blur uses $\sigma \in [0.1, 1.5]$ with a 5×5 kernel, and Gaussian noise uses standard deviation 0.05 on $[0, 1]$ image intensities. Only the current observation is transformed at shifted evaluation; the goal image remains clean. We do not evaluate the easier setting where the same transform is applied to both current and goal images. S_{shift} averages over these families; $S^{\min}_{\text{shift}}$ is the worst family. Tables V, VI, and VII give the decoupled schedule, hyperparameters, and multiple-comparison check.

TABLE V
Decoupled training details. Exact stage-2 losses and trainable modules for rows marked "dec.". The encoder is frozen after stage 1, so the planner changes only through the predictor trained in stage 2.

Variant	Stage 1	Stage 2 losses	Stage 2 trainable modules	Planner uses
Eq. InfoNCE + dec.	LeWM, 33 epochs	$\mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{rob-pred}} + \mathcal{L}_{\text{nce}}$	predictor only; encoder frozen	frozen encoder z and updated g_ϕ
Top-16 SW + dec.	LeWM, 33 epochs	$\mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{rob-pred}} + \mathcal{L}_{\text{top-k}}$	predictor only; encoder frozen	frozen encoder z and updated g_ϕ

In the Eq. InfoNCE + dec. row, $\mathcal{L}_{\text{nce}}$ is evaluated on the frozen encoder outputs and therefore has no trainable encoder path during stage 2. This row is a schedule ablation: its planner changes through the predictor update under clean and shifted prediction losses, not through a learned projection head used by CEM.

TABLE VI
Training and planning hyperparameters.

Parameter	Value
Observation resolution	64×64 RGB
Latent dimension	128
Encoder	four-stage ConvNet, channels 32–256
Predictor	3-layer action-conditioned MLP, hidden 256
Training epochs	15; decoupled rows use 50
Optimizer / learning rate	AdamW / 3×10^{-4}
Batch size	128
SIGReg projections / weight	32 / 1.0
Robust prediction weight λ_{rob}	1.0
SW-family λ_{aux}	0.25
InfoNCE / DBC-style λ_{aux}	0.25 / 0.25
VICReg weights and global multiplier	$(25, 25, 1)$; multiplier 0.25
Auxiliary rollout horizon H_{aux}	5
SW projections M	32 uniform; top-16 uses 64 sampled directions
MaxSW inner steps / restarts / LR	5 / 4 / 0.1
InfoNCE temperature	0.1
CEM candidates / elites / iterations	256 / 32 / 5
Planning horizon	12
Evaluation episodes	15 per seed and shift-family cell

Baselines. Uniform SW uses fresh random projection directions in each minibatch. MaxSW performs a short inner optimization to find the projection direction with largest projected Wasserstein discrepancy. Top-16 SW samples a bank of directions and averages the sixteen largest projected costs. VICReg-seq uses a variance-invariance-covariance loss on paired rollout stacks. DBC-style align-

TABLE VII

Multiple-comparison check. Holm correction is applied separately within each task over the seven Table I auxiliary comparisons, using exact paired permutation p-values on S_{robust}. For reacher_distractor, correction is applied over the six Table II comparisons.

Comparison	Uncorrected p	Holm-corrected p
tworoom Equiv. InfoNCE vs LeWM	.006	.042
tworoom MaxSW vs LeWM	.016	.094
All reacher improvements	$\geq .062$	$\geq .438$
reacher_distractor VICReg-seq regression	$< .001$	$< .001$
reacher_distractor Top-16 SW regression	.018	.089

TABLE VIII

Metric definitions. The primary metric is competence under clean and shifted observations, not retained delta, because retained delta can be high when both clean and shifted performance are poor.

Quantity	Definition and interpretation
S_{clean}	Mean success on clean observations.
S_{shift}	Mean success under the four main shift families.
$S_{\text{shift}}^{\min}$	Worst shift-family success; diagnostic for brittle families.
S_{robust}	$\frac{1}{n}\sum_i \min(S_{\text{clean},i}, S_{\text{shift},i})$; competence under clean and shifted observations.
Δ_{ret}	$S_{\text{shift}} - S_{\text{clean}}$; retention diagnostic, not a primary claim metric.
δ_a	Clean CEM cost gap between the best and runner-up candidates.
δ_J	Shift-induced drift in the clean-best candidate cost.
Action preservation	Fraction of evaluated states for which clean and shifted planning select the same first action over a fixed candidate set.

ment matches latent distances to one-step predicted-dynamics distances because the reward-free offline data do not provide a task reward for this auxiliary loss. Equivariant InfoNCE applies a state-paired contrastive loss to clean and shifted encoder outputs with in-batch negatives.

A. Auxiliary Loss Definitions

For InfoNCE, let $q_i = \text{normalize}(f_\theta(o_i))$ and $\tilde{q}_i = \text{normalize}(f_\theta(\tilde{o}_i))$. The one-way in-batch objective is

$$\mathcal{L}_{\text{nce}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(q_i^\top \tilde{q}_i/\tau)}{\sum_{j=1}^{B}\exp(q_i^\top \tilde{q}_j/\tau)},$$

with $\tau = 0.1$. Negatives are other states in the same minibatch.

VICReg-seq acts on clean and shifted rollout-increment stacks $r_i^{(H)}$ and $\tilde{r}_i^{(H)}$. It uses

$$\mathcal{L}_{\text{vic}} = 25\,\text{MSE}(r,\tilde{r}) + 25\,v(r,\tilde{r}) + c(r,\tilde{r}),$$

where v is the VICReg variance hinge penalty and c is the off-diagonal covariance penalty, both applied to the clean and shifted stack batches. The global multiplier in the total loss is $\lambda_{\text{aux}} = 0.25$.

The DBC-style row uses a reward-free dynamics distance target:

$$\mathcal{L}_{\text{dbc}} = \frac{1}{B^2}\sum_{i,j}\left(\|z_i - z_j\|_2 - \text{sg}[\|\hat{z}_{i,t+1} - \hat{z}_{j,t+1}\|_2]\right)^2,$$

where sg denotes stop-gradient and $\hat{z}_{i,t+1} = g_\phi(z_i, a_i)$. Because no reward labels are used, this is not a full reward-aware DBC objective.

Scope. The central results are those most relevant to robot planning: whether a learned latent dynamics model preserves closed-loop planning success under visual shifts. Additional per-family and severity diagnostics are omitted from the main paper for space; the main claims rely only on Tables I–II and the planner-margin diagnostics.

Data-coverage negatives. Several candidate robot-planning benchmarks were excluded from the main table because the common random-policy data protocol produced near-zero or unstable clean planning success for every method. These included PushT-style planar manipulation, OGBench cube manipulation, DMC cartpole balance, DMC cheetah, DMC-VB walker, and MetaWorld reach probes under the compact model and data budget. In those cases, robustness metrics are not meaningful: a method cannot retain task performance that the baseline planner never obtains. These cases separate robustness failures from data-coverage failures.

Appendix C
Additional Diagnostics

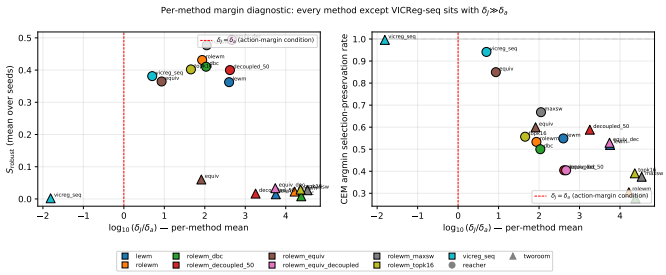

Fig. 3. Planner-margin diagnostic across methods. Left: competence under clean and shifted observations versus the mean $\log_{10} R_{\text{margin}}$, where $R_{\text{margin}} = 2\delta_J/\delta_a$ is the clean-best candidate drift relative to the half-margin. Right: CEM action-preservation rate versus the same ratio. Larger values mean the shift-induced cost change is large relative to the clean action margin, so being farther right is worse for action stability. The ratio predicts action preservation, but high task competence also requires clean predictive quality; methods with good preservation can still fail if clean planning is weak.

Reading the margin diagnostic. The action-margin condition does not claim that every violation causes an action flip. It states that if the worst candidate-wise shift drift is less than half the clean runner-up margin, then clean and shifted planning must agree over the fixed sampled candidate set. The plotted ratio $R_{\text{margin}} = 2\delta_J/\delta_a$ uses clean-best drift, so it is a lower-bound diagnostic for $2\eta/\delta_a$: large values rule out the sufficient certificate, while small values would not prove it. Figure 3 shows that this ratio is strongly related to action preservation, while robust task success also depends on whether the clean planner is competent.

Why latent overlap can still fail. The latent-overlap figure compares clean and shifted rollout clouds after

projection. This is an average geometric diagnostic. CEM, however, acts on a small set of candidate action sequences at a particular state. A method can reduce average cloud separation while leaving enough candidate-specific cost drift to reorder the top candidates. This explains the empirical pattern in which top-16 SW and VICReg-seq look strong under geometric diagnostics but do not reliably improve planning success.

A small capacity probe in Table IX is reported as a diagnostic and is not used to support the main claims.

TABLE IX
Capacity probe on reacher with d=256. Values are mean ± s.e. over n=8 paired seeds; p is an exact paired permutation test on S_{robust} against the matched d=256 LeWM row. This diagnostic is not used to support the main claims.

Method	S_{clean}	S_{shift}	S_{robust}	p
LeWM, d=256	**0.517**$\pm$.035	0.460$\pm$.037	0.431$\pm$.026	–
Equiv. InfoNCE, d=256	0.425$\pm$.025	0.398$\pm$.022	0.373$\pm$.022	.141
Top-16 SW + dec., d=256	0.508$\pm$.050	0.481$\pm$.023	0.433$\pm$.036	1.000
Eq. InfoNCE + dec., d=256	0.508$\pm$.053	**0.506**$\pm$.035	**0.438**$\pm$.032	.891

Interpretation of the capacity probe. The d=256 runs suggest that capacity and invariance may be complementary levers, but the probe is too small for a central claim. Increasing latent dimension improves the clean model relative to a matched d=128 subset, but the gain is marginal at the tested seed budget and does not dominate the best d=128 invariance results. A larger world model may improve the clean planner, but robustness still requires controlling the shift-induced cost drift measured by the action-margin diagnostic.

Appendix D
Limitations

The study is intentionally controlled and therefore limited. The benchmarks use compact image resolution and latent sizes; larger models may change the clean-robustness trade-off. The data are collected by the same protocol across methods, which makes comparisons clean but leaves several manipulation benchmarks near zero clean success because the data rarely visit goal states. The present experiments do not include full-scale external foundation encoders or real-robot validation; the conclusions should therefore be read as controlled evidence about compact LeWM-style planners under visual nuisance shifts. The theory gives a sufficient action-margin condition for fixed CEM candidate sets; it does not certify worst-case robustness.

The visual shifts are useful probes, but they are not a complete model of real robotic deployment. Real robots encounter viewpoint changes, calibration errors, occlusion, object rearrangement, actuation uncertainty, and delayed observations. The experiments isolate visual nuisance shifts so that the learned dynamics and planning margins can be studied cleanly. A deployment-ready system must combine these representation-level tests with closed-loop hardware or high-fidelity simulation evaluations.

For double-blind review, repository links and implementation identifiers are omitted. The submission includes the training protocol, hyperparameters, metrics, and statistical tests needed to assess the central claims; code and run logs will be released with the camera-ready version.

Appendix E
Practical Implications

A world model used for robust robot planning should expose planner-relevant uncertainty and margins. A representation loss can reduce shift sensitivity, but planners select discrete candidates from finite samples, so small cost perturbations near a narrow runner-up margin can flip actions. Future systems should train latent dynamics with action-margin-aware objectives, use uncertainty-aware CEM or model predictive path integral planning, and collect data that makes robust goal-reaching identifiable before evaluating visual robustness.

One concrete path is to add an action-margin regularizer during training. For clean and shifted views of the same state, the model could penalize cases where the predicted costs of the clean-best and shifted-best candidates disagree by more than a margin-scaled tolerance. Another path is to make CEM uncertainty-aware by scoring candidates with both predicted goal distance and model disagreement under clean-shift augmentations. A third path is to separate dynamics latents from appearance latents, so nuisance information can remain available for perception while the planner queries a dynamics-focused subspace.

For planning, a world model should be more than a general visual representation. It should be a queryable simulator with calibrated invariances. Standard invariance losses move part of the way toward that goal, while the remaining gap is planner-specific and measurable through action margins.

Appendix F
Future Evaluation

Further evaluation should separate three confounded axes. The first is model scale: larger encoders and predictors may improve clean planning and increase action margins. The second is data quality: expert, scripted, or curriculum data can ensure that the planner observes meaningful goal-reaching transitions during training. The third is planner robustness: uncertainty-aware CEM or model-predictive path integral planning can avoid over-committing to candidates whose latent costs are unstable under clean-shift perturbations.

A useful benchmark suite should include both low-signal and high-signal regimes. Low-signal tasks reveal whether invariance acts as a helpful inductive bias. High-signal tasks reveal whether the same invariance pressure harms an already competent world model. Distractor tasks are especially valuable because they test whether the model can ignore nuisance visual information without losing the action-conditioned dynamics needed for planning.

Appendix G
Summary of Claims

The supported claims are deliberately narrow. Same-stack invariance can improve planning competence on tworoom, with a Holm-robust gain for equivariant InfoNCE and an uncorrected exploratory gain for MaxSW, but those gains improve clean success as well as shifted success and should be interpreted as representation regularization. These gains do not transfer uniformly to reacher or reacher_distractor. Latent alignment is a real geometric effect, but it is not sufficient for retained planning when the remaining shift-induced cost drift exceeds the planner's action margin. Robust world-model systems should train and evaluate invariance in planner-relevant terms, rather than in representation-space terms alone.