LOCAL COUNTERFACTUAL GEOMETRY IN WORLD MODELS: PROPOSE GLOBALLY, CORRECT LOCALLY

Anonymous authors
Paper under double-blind review

ABSTRACT

A world model can predict how changing an action will change its outcome. We test whether frozen action-conditioned latent predictors get these changes right. We study local counterfactual geometry, the mapping from perturbations around an action plan to predicted future changes conditioned on observation and action history. Comparing predicted and realized effects of the same perturbations tests whether this geometry reflects the environment and where it remains valid. Our experiments show that strong action responses can misrepresent real consequences, and that predicted effects may not transfer across interactions. In control, this source mismatch becomes more consequential as optimization is allowed to move farther from the proposal. The same geometry also supports action refinement: an existing controller supplies a proposal, while the frozen world model computes bounded, goal-directed corrections using the current interaction. We compare released models on the same base action plans and perturbations to measure effect accuracy across perturbation sizes, horizons, and interactions. This offers a complementary view of world-model capability: beyond proposal performance, we ask whether the model can reliably guide local improvement. Local counterfactual geometry thus connects observation and control: propose globally, correct locally.[1]

1 INTRODUCTION

Planning with a learned model allows an agent to reconsider an action before committing to it. A planner can predict the outcome of a candidate sequence, modify the actions, and compare the resulting prediction with its goal (Hafner et al., 2023; Hansen et al., 2024; Bharadhwaj et al., 2020). For action-conditioned latent predictors used in visual planning, these comparisons take place in representations of future observations, as in V-JEPA 2-AC (Assran et al., 2025), LeWorldModel (Maes et al., 2026), and DINO-WM (Zhou et al., 2025). When a planner uses such a comparison to revise its actions, it relies on the predicted improvement carrying over to execution. The relevant question is therefore not only where a plan will lead, but whether changing that plan will change the outcome in the way the model predicts.

This requirement is easy to overlook when a model predicts observed trajectories well or supports successful planning. Neither observation, by itself, tells us how faithfully the model describes the consequences of changing an action. Planning success reflects the combined representation, predictor, and planner (Terver et al., 2026). Accurate predictions along observed trajectories leave a different ambiguity. If the data contain little action variation at a given state, they may not determine what would have happened under an alternative action (Zhang et al., 2026). A model can therefore fit the futures it has observed while describing their alternatives incorrectly. For a planner choosing between nearby action sequences, those errors matter because the predicted difference is precisely what motivates the change.

We study this difference through *local counterfactual geometry*, the mapping from perturbations around an action plan to predicted future changes conditioned on observation and action history. Keeping model parameters fixed, we test this relation by executing original and perturbed plans from the same simulator state and comparing predicted and actual changes (Fig. 1). We also compare predictions from two situations with execution of the same action alternatives in each.

[1] Anonymous code is available in this repository.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

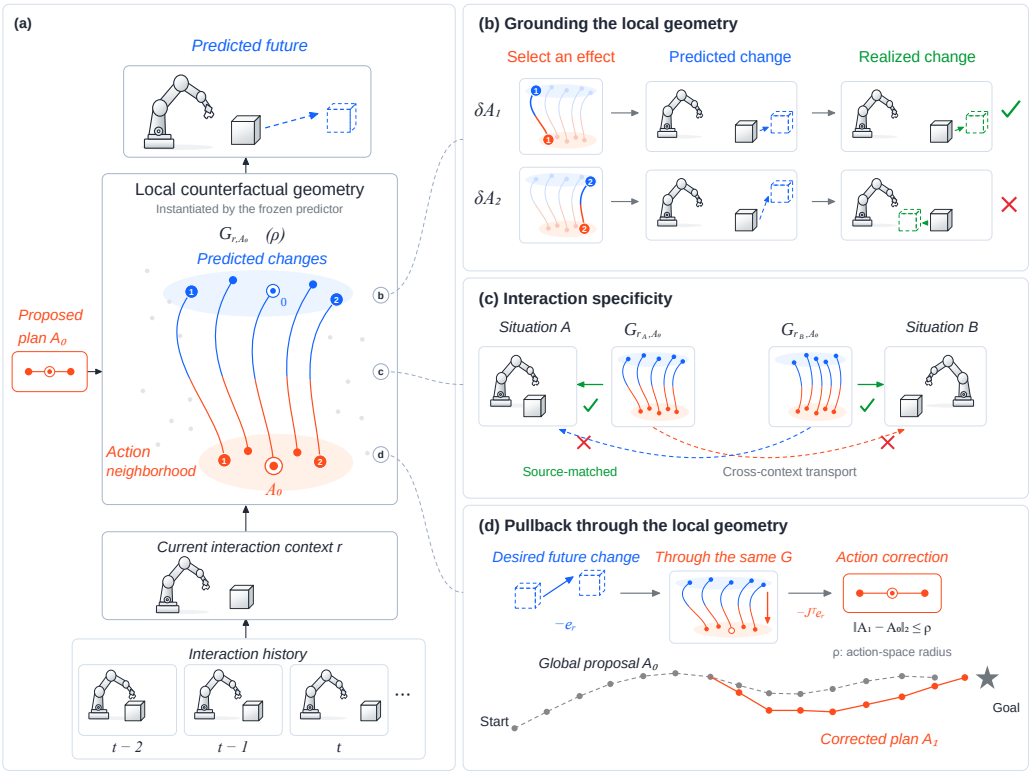


Figure 1: Local counterfactual geometry as an observational lens and a control interface. (a) A frozen predictor associates action changes around proposal A_0 with predicted future changes conditional on the current interaction r. (b) Same-state replay compares predicted and realized effects. (c) Predicted effects from two situations are checked against actual effects in each, testing accuracy before and after reuse. (d) The Jacobian transpose maps the predicted difference from the goal to the correction direction $-J_A^\top e_r$, constrained by action-space radius ρ. Curves, paths, and crosses are schematic.

This distinguishes effects already inaccurate at their source from those that lose accuracy when reused elsewhere. Strong model responses can misrepresent real consequences, and current-history predictions are more accurate on average than reused effects.

For control, an existing controller supplies a plan (Nguyen et al., 2026; Sun et al., 2026); the frozen world model computes a bounded, goal-directed correction from the current history. Holding initial plans fixed across comparisons isolates refinement gains. Allowing larger plan changes widens the performance gap between current-history guidance and guidance from another situation.

Finally, this relation provides an observational lens on world models. Using the same plans and perturbations, we compare effect accuracy across models, perturbation sizes, horizons, and situations, complementing task success. We also compare re-prediction from a new observation with reuse of an earlier effect. Thus *propose globally and correct locally* means refining supplied plans using current history for control, and distinguishing queries to the same model in new situations from reuse of earlier effects for observation.

Our contributions are:

1. We test local counterfactual geometry against executed action effects, separating response strength from effect accuracy and showing how accuracy changes when predicted effects are reused across situations.
2. We improve existing plans through bounded, goal-directed correction and show how larger allowed changes amplify the cost of using guidance from a different situation.
3. We compare how world models predict the effects of the same perturbations to recorded action plans, across perturbation sizes, prediction horizons, and situations.

2 Related Work

Functional representations and interventions. Amnesic Probing removes decodable information to test its role in model behavior (Elazar et al., 2021). Causal abstraction and distributed alignment methods intervene on neural representations to test their correspondence with high-level variables (Geiger et al., 2021; 2024). J-space uses a Jacobian lens to identify representations associated with verbalization in language models (Gurnee et al., 2026). These studies motivate testing what representations do, rather than only what they encode. Our interventions change action plans, so environment execution provides an external reference for the predicted change.

Predictive fit and planning diagnosis. Zhang et al. (2026) connect identifiability of representations and controlled dynamics to spectral separation and conditional action variation under their theoretical assumptions. Their analysis explains why fitting observed trajectories can leave alternative-action effects undetermined. Terver et al. (2026) instead compare architectures, training objectives, and planning algorithms to identify design choices that improve JEPA-based planning. Together, these studies distinguish what training data can determine from what makes a planning system effective. Our replay tests examine a trained predictor at the intervening step, asking whether the action changes it evaluates produce the predicted consequences.

Learning and evaluating action effects. Action-adaptive policies encode observed action impacts to handle changing execution effects (Zeng et al., 2023). Causal-JEPA predicts masked object latents from surrounding context to encourage interaction-dependent representations (Nam et al., 2026). ActSWM keeps alternative-action futures distinguishable and their actions recoverable (Gan et al., 2026), while SelfWAM jointly predicts actions, future images, and robot self-masks to strengthen action sensitivity (Pan et al., 2026). These approaches make action consequences more explicit in learned representations. Dueling World Models isolates them by subtracting the mean prediction over actions, removing action-shared variation even as a post-processing step for frozen models (Li et al., 2026). CoCo uses inverse-action, zero-action, and mirror consistency to reduce statistical shortcuts, and evaluates same-state action alternatives and planning (Shi et al., 2026). Building on this concern with action effects, we cross prediction and execution contexts with the action alternatives fixed. This tests whether an effect agrees with execution at its source and whether that agreement survives reuse elsewhere.

Gradient-based planning. CEM and MPPI optimize plans through sampling (de Boer et al., 2005; Williams et al., 2017); Bharadhwaj et al. (2020) combine CEM with gradient refinement. Ma et al. (2024) design world-model architectures with more direct gradient paths for policy learning. GRASP optimizes virtual states under soft dynamics constraints to parallelize planning (Psenka et al., 2026), while Parthasarathy et al. (2025) use training-time data synthesis to improve models for gradient-based planning. These methods improve how gradients are obtained or used. We keep the predictor fixed and test how the source of its guidance and the permitted departure from a proposal affect execution.

Proposals and local refinement. POMP initializes plans with a policy and makes conservative updates through its D3P planner (Zhu et al., 2023), providing a precedent for refining supplied plans. A goal-conditioned inverse dynamics model (GC-IDM) maps the current latent state, goal, and remaining horizon directly to an action, amortizing planning (Nguyen et al., 2026). INTACT learns an intent-to-action mapping and optionally refines its direct plans with local CEM (Sun et al., 2026). We use GC-IDM and INTACT as proposal sources rather than introducing another proposal mechanism. Holding proposals fixed lets us isolate the contribution of bounded, current-context correction and compare it with local search, while replay separately checks the predicted action effects.

3 Local Counterfactual Geometry in a Frozen World Model

The same action–future relation supports observation by predicting the effects of action changes and control by mapping a desired future change back to actions. The predictor and goal-distance objective follow LeWM and INTACT (Maes et al., 2026; Sun et al., 2026); Taylor expansion, the chain rule, and projected optimization are standard. Our tests compare predicted and executed effects, separate source accuracy from transferability, and examine bounded refinement of existing plans.

3.1 FROM NATIVE ACTION INTERVENTION TO LOCAL GEOMETRY

The *interaction* r contains the observation and action histories and the cached encoded state used by the predictor. The *proposal* A_0 is the initial action plan, supplied by a controller in the control experiments. A plan $A \in \mathbb{R}^{H\times d}$ has H model steps, each containing b primitive actions of dimension d_0, so $d = bd_0$. Norms and Jacobians flatten the plan to $\mathbb{R}^{Hd}$. The terminal latent and goal z_g lie in $\mathbb{R}^n$. Appendix A summarizes the notation and experimental dimensions; Appendix B gives the local derivations. Frozen means that model parameters remain fixed; predictions and action derivatives still vary with r and A. The predictor returns

$$\hat{z}_H = F_H(r, A). \tag{1}$$

Holding r fixed and changing only the action defines a same-context counterfactual effect:

$$\Delta_F(r, A, \delta A) = F_H(r, A + \delta A) - F_H(r, A). \tag{2}$$

We call Δ_F the model-predicted action effect. Its *environmental grounding* is agreement with the effect of executing the same action change.

For a radius ρ around the current plan A, these paired displacements define

$$\mathcal{G}_{r,A}(\rho) = \left\{ \left(\delta A, \Delta_F(r, A, \delta A)\right) : \|\delta A\|_2 \le \rho \right\}. \tag{3}$$

We call $\mathcal{G}_{r,A}$ the local counterfactual geometry. It records which future changes the predictor associates with nearby action changes at the current interaction and plan.

Under a locally Lipschitz action Jacobian, the standard first-order expansion is

$$\Delta_F(r, A, \delta A) = J_A(r, A)\, \delta A + O(\|\delta A\|^2), \tag{4}$$

where the action Jacobian is

$$J_A(r, A) = \frac{\partial F_H(r, A)}{\partial A}. \tag{5}$$

The Jacobian is the tangent map at (r, A). We test its approximation against finite effects, then compare those effects with execution. Its singular values (the *spectrum*) measure response strength along orthogonal action directions. In supplementary diagnostics, *damping* downweights weak-response directions to test their role in correction.

3.2 PULLING A DESIRED FUTURE CHANGE THROUGH THE GEOMETRY

For control, we use geometry to reduce the released latent planning objective

$$\mathcal{L}_r(A) = \|F_H(r, A) - z_g\|_2^2. \tag{6}$$

The residual $e_r = F_H(r, A) - z_g$ is the predicted difference from the goal. The *pullback* maps this goal discrepancy to an action gradient by the chain rule:

$$g_r(A, z_g) = \nabla_A \mathcal{L}_r(A) = 2J_A(r, A)^\top e_r. \tag{7}$$

The associated local least-squares problem is

$$\delta A^* = \arg\min_{\delta A} \|J_A(r, A)\delta A + e_r\|_2^2. \tag{8}$$

Its steepest direction at $\delta A = 0$ is proportional to $-J_A^\top e_r$: the same map used to predict action effects now converts the goal discrepancy into a correction direction. We use this gradient in projected Adam rather than solving the least-squares problem directly; Appendix B gives the derivations.

3.3 TESTING INTERACTION LOCALITY BY SOURCE SWAPS

Both the interaction r and plan A index the geometry. An effect's *source* is the interaction used to predict it. *Transport* means reusing an effect or correction at another interaction; its accuracy there is an empirical question. The execution interaction r is the *recipient*; another interaction r' used to compute guidance is the *donor*. A source swap applies $g_{r'} = \nabla_A \mathcal{L}_{r'}(A)$ to an action executed under

r, holding the proposal, goal, update rule, action bounds, and simulator realization fixed. Donor guidance is rescaled to the matched norm:

$$\widetilde{g}_{r'\to r} = \|g_r\|_2 \frac{g_{r'}}{\|g_{r'}\|_2}. \tag{9}$$

We take the normalized direction to be zero when its source vector is zero. Matched uses g_r; Donor uses the norm-matched gradient from r'. Comparing them tests whether guidance transfers across the tested interactions. Random uses an equal-norm isotropic direction as a magnitude-matched null, while reciprocal environment replay provides the direct source-validity test (section 4).

3.4 Controlling action locality by a trust radius

A controller first supplies $A_0 = P(r, z_g)$. With the encoder, predictor, and controller fixed, we update only the action tensor:

$$A_{k+1} = \left(\Pi_{\mathcal{A}} \circ \Pi_{B(A_0,\rho)}\right)\left[\mathrm{Adam}\big(A_k, g_r(A_k, z_g)\big)\right], \tag{10}$$

where $\mathcal{A}$ is the official action box and $B(A_0, \rho)$ is an L2 trust region around the proposal. Each step is projected radially, then clamped to the action box. The radius ρ bounds the permitted departure from the proposal, which we call *optimization authority*. Increasing it lets locally recomputed model guidance move the plan farther from the proposal, allowing us to test how far that guidance remains useful.

4 Experimental Design

4.1 Systems, proposals, and paired evaluation

We evaluate all four released LeWM environments: Push-T, Reacher, OGBench-Cube, and Two-Room (Maes et al., 2026). We repeat the transport intervention under INTACT using its native proposal and predictor. Same-state grounding uses Push-T and Reacher, where the simulator state can be restored. Transport by optimization authority spans both systems and all four tasks.

Controller-based evaluations begin with stored LeWM proposals, INTACT's native proposal, or an independently trained GC-IDM proposal. The encoder, predictor, and proposal source remain frozen. Behavioral comparisons share the proposal, goal, model checkpoint, and environment realization. Complete manifests, sample sizes, and action-interface adaptations are in Appendix C.

4.2 Calibrating and grounding the local operator

For tangent calibration, draw an isotropic unit action direction u and set $\delta A_\epsilon = \Pi_{\mathcal{A}}(A + \epsilon u) - A$. We compare the exact frozen-model displacement

$$\Delta_F(\epsilon) = F_H(r, A + \delta A_\epsilon) - F_H(r, A) \tag{11}$$

with its Jacobian–vector product (JVP) approximation

$$\Delta_J(\epsilon) = J_A(r, A)\delta A_\epsilon. \tag{12}$$

We measure directional agreement and relative error between these two displacements.

The grounding test uses a normalized matched-descent perturbation and replays its post-projection displacement δA outside the model. Starting twice from an identical simulator state, we execute A and $A + \delta A$. We compare the predicted effect with both the released encoder's terminal difference and a predefined task-coordinate change. Replay uses the evaluator's task-terminal endpoint: each branch stops on termination or at the control-step budget T obtained by decoding the H model action steps. Write this endpoint as $s^A_{\tau_T(A)}$, where $\tau_T(A) \le T$ is its executed length. Let $D \in \mathbb{R}^{n_q \times n}$ be the linear part of the fixed task readout and let q extract the same standardized coordinates from the simulator state. Both replay tasks use $n_q = 4$. Readout offsets cancel when taking effects. We compare

$$\begin{aligned} \Delta_q^{\mathrm{model}} &= D(F_H(r, A + \delta A)) - D(F_H(r, A)), \\ \Delta_q^{\mathrm{env}} &= q\Big(s^{A+\delta A}_{\tau_T(A+\delta A)}\Big) - q\Big(s^{A}_{\tau_T(A)}\Big). \end{aligned} \tag{13}$$

We fit the readout on factual trajectories and freeze it before replay; the environment effect uses the simulator block pose or joint configuration directly. We report directional cosine and relative magnitude error (Table 8). We compare Matched, near-donor, and far-donor predictions with an equal-norm isotropic Random effect vector in the measured latent or task space. Near and Far donors maximize and minimize mean cosine similarity across state, goal-intent, and proposal features, excluding the recipient's episode.

Reciprocal replay instead uses a gradient-independent isotropic action direction. We fix the base plan A and post-projection perturbation δA while crossing prediction and execution sources over recipient r and donor r'. Writing $c_{\cdot,\cdot}$ for effect cosine, with prediction source first and execution source second, we use

$$\Omega = \tfrac{1}{2}\big[(c_{r,r} + c_{r',r'}) - (c_{r,r'} + c_{r',r})\big]. \tag{14}$$

Positive Ω indicates greater own-source than crossed alignment. We report post-projection response strength $\|\Delta_F\|_2/\|\delta A\|_2$ alongside environmental alignment. A matched-only analysis examines high-response effects. A 2×2 factorial then separates the recipient and donor Jacobians from their goal residuals, using one normalized descent step for each combination (Appendix D.2.2).

4.3 Testing transport and optimization authority

Matched and Donor recompute guidance along their optimization trajectories, using the recipient and a fixed donor history, respectively. Each solve takes 30 projected Adam steps and accepts the final iterate, without a cost-based acceptance check. Proposal, goal, trust radius, action bounds, and execution recipient are shared. We norm-match donor gradients to recipient gradients before each Adam step. Random redraws an equal-norm isotropic direction at each update.

All conditions share a trust-radius grid from a tight proposal neighborhood to unbounded updates. The primary statistic is the increase in the Matched–Donor success gap between these endpoints. Here, extrapolation means allowing repeatedly recomputed guidance to move the action plan farther from A_0. Donor ranks test how transport varies with interaction similarity. A LeWM control crosses Donor and Random guidance with fixed and refreshed schedules (Appendix D.3.2).

4.4 Testing bounded residual correction

We test bounded Matched correction on independent GC-IDM proposals in Push-T and native INTACT proposals in Reacher. Each uses one projected Adam step and reverts to the proposal if recipient latent cost increases. A strict-local cross-entropy method (CEM) frontier compares derivative-free search in the proposal neighborhood at six frozen-model evaluation budgets under the same acceptance rule. We report paired success changes, recovery and harm counts, and realized action displacement. Full settings are in Appendix C.

4.5 Comparing action effects across released systems

Released LeWM, DINO-WM, and JEPA-WM share recorded-action base plans on Push-T. Shared Near and Far states are the nearest and most distant other recipients in standardized pusher-position and block-pose coordinates. Direct reuse compares the original situation's predicted effect with execution at the Far state; re-prediction recomputes it from the Far history, with the model, readout, and action intervention fixed. Here ρ sets perturbation size, not optimization authority. For model m, source c, radius ρ, primitive-step horizon h, recipient i, and direction d, $\widehat{\Delta}_{m,c,\rho,h,i,d}$ and $\Delta^E_{c,\rho,h,i,d}$ are predicted and realized symmetric plus-minus effects in standardized physical task coordinates. The pooled cell skill is

$$S_{m,c,\rho,h} = 1 - \frac{\sum_{i,d} \|\widehat{\Delta}_{m,c,\rho,h,i,d} - \Delta^E_{c,\rho,h,i,d}\|_2^2}{\sum_{i,d} \|\Delta^E_{c,\rho,h,i,d}\|_2^2}. \tag{15}$$

Here $S = 1$ is perfect; $S = 0$ matches a zero-effect predictor and $S < 0$ is worse. We average equally over the prespecified 60-cell interaction–radius–horizon grid $\mathcal{G}$:

$$\overline{S}^{\mathcal{G}}_m = \frac{1}{|\mathcal{G}|} \sum_{(c,\rho,h)\in\mathcal{G}} S_{m,c,\rho,h}. \tag{16}$$

Full protocol details are in Appendix C.4.

5 RESULTS

5.1 ACTION-RESPONSE STRENGTH AND ENVIRONMENTAL ALIGNMENT SEPARATE

The Jacobian closely approximates nearby model-side effects (Table 5), but response strength alone does not establish effect accuracy. On Reacher (Table 1), Matched, Near, and Far have almost identical response magnitudes, but their planning-space alignment falls from .661 to .328 to $-.058$. The same ordering persists in joint coordinates.

Table 1: Response magnitude and environmental alignment at $\rho = 0.25$. Planning and task columns compare the predicted effect with the realized effect in each coordinate system. Bold marks the largest alignment within each task.

Task	Response source	Response strength	Planning align.	Task align.
Push-T	Matched	12.44	**.570**	**.368**
Push-T	Near donor	9.15	.057	.108
Push-T	Far donor	2.00	.016	$-.035$
Reacher	Matched	18.99	**.661**	**.718**
Reacher	Near donor	18.89	.328	.602
Reacher	Far donor	18.90	$-.058$	.343

Across all three radii, Matched exceeds Near, Far, and Random in both coordinate systems (Tables 6 and 7). Matched predictions also show high-response effects with the wrong task-space direction (Appendix D.1.3).

5.2 ACTION EFFECTS DEPEND ON THEIR SOURCE INTERACTION

Reciprocal replay crosses both tasks with Near and Far donors (Table 2). Source-matched pairs have higher aggregate alignment on both tasks in both coordinate systems. In task coordinates, the Near Reacher pair retains more cross-context alignment than the other pairs (Table 10).

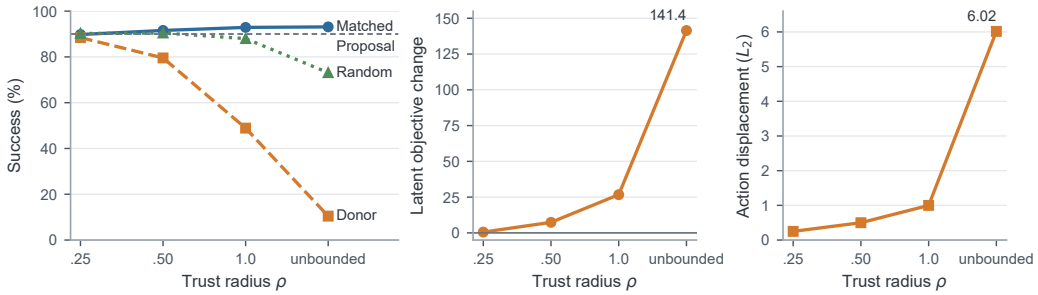
Table 2: Reciprocal source pairing has greater aggregate validity than crossed pairing. The first and fourth pairing columns are own-source comparisons; the middle two are crossed comparisons. Ω_z and Ω_q measure the paired contrast in the planning representation and task coordinates.

Task	Donor	$r \to E_r$	$r \to E_{r'}$	$r' \to E_r$	$r' \to E_{r'}$	Ω_z	Ω_q
Push-T	Near	**.451**	.078	.086	**.421**	**.354**	**.286**
Push-T	Far	**.451**	.007	.004	**.167**	**.304**	**.254**
Reacher	Near	**.701**	.448	.395	**.852**	**.355**	**.104**
Reacher	Far	**.701**	$-.101$	$-.102$	**.807**	**.855**	**.414**

In the 2×2 decomposition, swapping the Jacobian changes the correction direction more than swapping the predicted difference from the goal. At comparable action displacements, Jacobian-only swaps reduce cosine alignment with the matched gradient to .046 on Push-T and .006 on Reacher. After execution, mean squared error between the endpoint and goal encodings increases by .052 and .187, respectively. Residual-only swaps retain more alignment (.522 and .908), with no detectable increase in endpoint error (Table 11). Additional random-probe controls examine residual similarity and operator agreement beyond the goal direction (Appendix D.2.3).

5.3 OPTIMIZATION AUTHORITY AMPLIFIES MISMATCH COSTS

Increasing optimization authority enlarges the Matched–Donor success gap in all eight task–system settings (Table 3), with a significant increase in each setting (Table 13). Donor similarity further modulates transport on Reacher; Push-T shows less gradation near the proposal (Appendix D.2.4).



(a) Task success (b) Recipient objective (c) Action displacement

Figure 2: Source mismatch under increasing optimization authority in LeWM Push-T. (a) Task success; (b) recipient latent-objective change and (c) action displacement under Donor guidance. Results aggregate three actor replications over 150 recipient identities.

Table 3: Source mismatch across eight task–system settings. Matched, Donor, and Random give success at the unbounded endpoint. Gap amplification is the increase in the Matched–Donor gap from $\rho = 0.25$ to that endpoint. Paired uncertainty estimates are in Table 13.

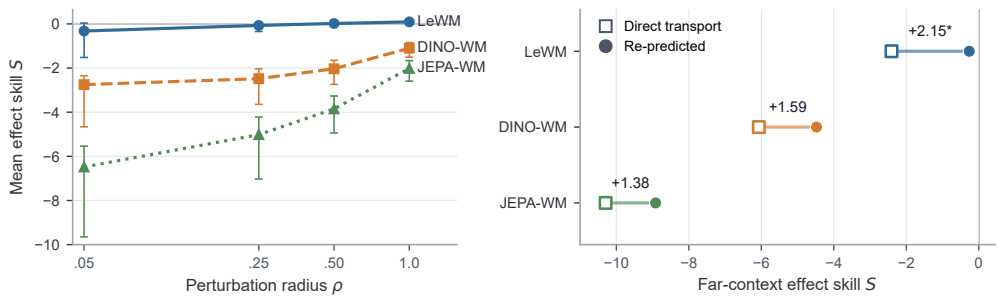
System	Task	Success (%)				Gap amp. (pp)
		Base	Matched	Donor	Random	
LeWM	Push-T	90.00	93.11	10.44	73.11	**+81.33**
LeWM	Reacher	60.22	75.78	25.56	46.22	**+31.78**
LeWM	Cube	96.44	93.11	56.44	92.22	**+36.89**
LeWM	Two-Room	98.67	99.78	89.78	96.67	**+9.11**
INTACT	Push-T	79.67	80.33	10.00	67.33	**+67.33**
INTACT	Reacher	93.67	99.67	33.67	68.67	**+55.67**
INTACT	Cube	100.00	98.33	59.00	99.33	**+39.33**
INTACT	Two-Room	86.67	80.00	68.67	83.33	**+11.33**

Push-T makes the trajectory visible (Fig. 2): Matched and Donor are similar in the tight neighborhood, then separate as the radius grows; Matched remains effective, Random degrades, and Donor collapses as recipient cost rises. Gap amplification also occurs in all three INTACT Reacher checkpoints (Appendix D.3.1). In a separate LeWM control, Matched exceeds Fixed and Fresh Donor at $\rho = \infty$ on both tasks (Table 14). Random redraws its direction at each update, whereas Donor retains its source; the paired refresh control shows that this schedule affects their comparison. Spectral and damping diagnostics test whether disagreement lies in weak-response action directions and whether downweighting those directions improves correction (Appendices D.6 and D.7).

5.4 Using geometry for correction and model observation

One bounded, source-matched correction uses two forward candidate evaluations and one backward call per decision. At $\rho = .25$, it raises GC-IDM Push-T success from 84.33% to 87.67%. Across three INTACT Reacher checkpoints, mean success rises from 95.67% to 97.00%. The paired gain is statistically resolved on Push-T and remains uncertain on Reacher. On Reacher, its success point estimate exceeds all six tested CEM configurations, which attain 95.56%–96.56% using 5–385 forward candidate evaluations. On Push-T, CEM reaches 88.83% at nine evaluations. Table 15 reports the full paired intervals, CEM frontier, recovery and harm counts, and realized action displacements.

As an observational lens, the geometry separates action scale from interaction locality. Relative effect skill increases from the smallest to the largest tested radius in all three released Push-T systems (Fig. 3a). Their 60-cell grid-mean point estimates are $-.107$ for LeWM, -2.279 for DINO-WM, and -4.864 for JEPA-WM, below the zero-effect baseline; all three pairwise differences are significant under the prespecified test. At the Far interaction, with the checkpoint and readout fixed, re-predicting from the new observation raises LeWM's mean effect skill by 2.145 over direct effect reuse, to $-.262$ (Fig. 3b). Only LeWM shows a statistically resolved re-prediction gain. This within-system



(a) Action-effect prediction

(b) Cross-context effect prediction

Figure 3: A geometry lens on action-effect prediction in released systems. (a) Effect skill at four displayed perturbation radii, averaged over interactions and horizons; error bars show pointwise 95% recipient-bootstrap intervals. (b) Direct effect transport and re-prediction at the Far interaction, averaged over radii and horizons. Labels indicate the gain from re-prediction; the asterisk denotes an adjusted 95% interval above zero. Scores use each system's frozen factual readout to map model outputs to standardized $(x, y, \sin\theta, \cos\theta)$ coordinates. Complete surfaces are in Fig. 4.

contrast tests the value of recomputing the action–future relation after the interaction changes. Complete surfaces, prespecified comparisons, and the predictor–readout error decomposition are in Appendix D.5.

6 Discussion

Action locality concerns changes around a plan; interaction locality concerns the history supplying its predicted effects. Measurement scale and optimization authority answer different questions: larger perturbations can have higher relative effect skill while wider optimization ranges amplify the cost of source mismatch.

Propose globally, correct locally connects control and observation. For control, it separates finding an initial plan from refining it using the current history. For observation, it separates reusing a world model from assuming that one situation's predicted action effects remain valid elsewhere. Re-prediction tests this distinction without changing model parameters. In ten LeWM Push-T recipients, spectral diagnostics locate source disagreement in directions with strong predicted responses; damping shows no resolved donor benefit. These results concern the learned relation, whereas identifiability theory concerns what training data can determine (Zhang et al., 2026).

Limitations. Grounding covers terminal effects in Push-T and Reacher in released planning representations and selected task coordinates. The cross-system evaluation spans different training data, budgets, observation histories, inputs, and representation dimensions. Physical effect scores describe each predictor together with its fixed representation and readout. Spectral and null-control diagnostics cover ten LeWM Push-T recipients at one actor seed. Correction can saturate near performance ceilings or fail when latent objective and task utility diverge. Using measured effect fidelity to select a correction or radius at deployment remains open.

7 Conclusion

Local counterfactual geometry relates action changes to predicted future changes and connects them to environment replay. Its locality concerns both the neighborhood of a proposal and the interaction supplying the prediction. Bounded correction puts this relation to use with few candidate evaluations; the released-system assay reveals its fidelity across scales, horizons, and interactions.

AI USE STATEMENT

We used generative AI tools to assist with manuscript writing and polishing, literature search and discovery, and research ideation and execution, including assistance with code, analysis, and figure drafting. Verification included code tests and record-based numerical consistency checks. The authors take responsibility for the final manuscript, methods, and reported results.

REPRODUCIBILITY STATEMENT

Our experiments build on publicly available datasets and pretrained world-model checkpoints. The accompanying code provides the proposed action-correction methods, model adapters, task configurations, and experiment scripts. Appendix C, Appendix C.4, and Appendix D.3.2 detail the data and model sources, donor construction, optimization settings, evaluation protocols, and statistical procedures.

REFERENCES

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction, and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Homanga Bharadhwaj, Kevin Xie, and Florian Shkurti. Model-predictive control via cross-entropy and gradient-based optimization. In *Learning for Dynamics and Control*, 2020.

Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134:19–67, 2005.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Zhenfeng Gan, ZiTong Zeng, Jiajun Cheng, Yeke Song, Yongyi Tang, and Xueqian Wang. ActSWM: Action-sensitive world models for long-horizon planning in open-world games. arXiv:2607.26712v2, 2026. URL https://arxiv.org/abs/2607.26712v2.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Conference on Causal Learning and Reasoning*, 2024.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *arXiv preprint arXiv:2607.15495*, 2026.

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

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Jiazhuo Li, Yiming Fei, Zhiruo Zhou, and Heikichi Hayashi. Dueling world models: Advantage-style action channels for common-mode distractor rejection. arXiv:2608.06706v1, 2026. URL https://arxiv.org/abs/2608.06706v1.

Michel Ma, Tianwei Ni, Clement Gehring, Pierluca D'Oro, and Pierre-Luc Bacon. Do transformer world models give better policy gradients? In *Proceedings of the 41st International Conference on Machine Learning*, pp. 33855–33879, 2024.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorld-Model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-JEPA: Learning world models through object-level latent masking. *arXiv preprint arXiv:2602.11389v2*, 2026. URL https://arxiv.org/abs/2602.11389v2.

Hoang Nguyen, Xiaohao Xu, and Xiaonan Huang. Latent geometry beyond search: Amortizing planning in world models. *arXiv preprint arXiv:2605.08732*, 2026.

Bikang Pan, Fan Liu, Haotao Lu, Jingya Wang, and Ye Shi. SelfWAM: A self-grounded unified world action model for fast robot control. arXiv:2608.00725, 2026. URL https://arxiv.org/abs/2608.00725.

Arjun Parthasarathy, Nimit Kalra, Rohun Agrawal, Yann LeCun, Oumayma Bounou, Pavel Izmailov, and Micah Goldblum. Closing the train–test gap in world models for gradient-based planning. *arXiv preprint arXiv:2512.09929*, 2025. URL https://arxiv.org/abs/2512.09929.

Michael Psenka, Michael Rabbat, Aditi S. Krishnapriyan, Yann LeCun, and Amir Bar. Parallel stochastic gradient-based planning for world models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.00475.

Yuhong Shi, Zhenhao Chu, Jie Wei, Jun Hao, Jianyi Liu, and Jingwen Fu. Overcoming statistical bias in action-controllable world models. arXiv:2608.04653v1, 2026. URL https://arxiv.org/abs/2608.04653v1.

Junhan Sun, Hao Zhao, and Guofeng Zhang. INTACT: Isomorphic intent-to-action learning for search-free world models. *arXiv preprint arXiv:2607.26056*, 2026.

Basile Terver, Tsung-Yen Yang, Jean Ponce, Adrien Bardes, and Yann LeCun. What drives success in physical planning with joint-embedding predictive world models? *Transactions on Machine Learning Research*, 2026.

Grady Williams, Nolan Wagener, Brian Goldfain, Paul Drews, James M. Rehg, Byron Boots, and Evangelos A. Theodorou. Information theoretic MPC for model-based reinforcement learning. In *IEEE International Conference on Robotics and Automation*, 2017.

Kuo-Hao Zeng, Luca Weihs, Roozbeh Mottaghi, and Ali Farhadi. Moving forward by moving backward: Embedding action impact over action semantics. In *the International Conference on Learning Representations*, 2023.

Xiangteng Zhang, Yang Guan, Bo Zhang, Hongyang Li, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models. arXiv:2607.22430v2, 2026. URL https://arxiv.org/abs/2607.22430v2.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 79115–79135, 2025.

Jinhua Zhu, Yue Wang, Lijun Wu, Tao Qin, Wengang Zhou, Tie-Yan Liu, and Houqiang Li. Making better decision by directly planning in continuous control. In *International Conference on Learning Representations*, 2023.

A Notation

Table 4 collects the local-geometry and observation-assay notation. Action norms and inner products use the flattened plan. Concrete dimensions refer to LeWM Push-T; the released-system assay uses each model's native representation and a common four-dimensional physical readout.

Table 4: Notation and dimensions. Model steps H and primitive environment horizons h are distinct; ρ is an optimization bound or a measurement perturbation radius according to the experiment.

Symbol	Domain / size	Meaning
r, r'	interaction contexts	Recipient and donor histories and encoded state
A, A_0	$\mathbb{R}^{H\times d}$	Action plan and initial proposal
H, b, d_0	$5, 5, 2$	Model steps, primitives per step, primitive dimension
d, Hd, Hb	$10, 50, 25$	Coordinates per model step, plan size, environment steps
$F_H, \hat{z}_H, z_g$	output in $\mathbb{R}^n$	Frozen predictor, predicted terminal latent, encoded goal
n, n_q	$192, 4$	LeWM latent dimension and physical task dimension
Δ_F	$\mathbb{R}^n$	Finite model-predicted action effect
J_A	$\mathbb{R}^{n\times Hd}$	Action Jacobian at the current interaction and plan
$\mathcal{G}_{r,A}(\rho)$	set of displacement pairs	Local action–future relation
e_r	$\mathbb{R}^n$	Predicted terminal-goal residual
g_r	$\mathbb{R}^{Hd}$	Goal-cost gradient $2J_A^\top e_r$
ξ	$\mathbb{R}^{Hd}$	Generic transferred update direction in the derivations
u, w	unit action / latent vectors	Random tangent and cotangent probes
ρ	positive radius or ∞	Proposal-centered bound; finite radius in the effect assay
K	30 or 1	Updates in the transport sweep or bounded correction
η	positive scalar	Step size in the local descent analysis
σ_k, v_k	scalar, $\mathbb{R}^{Hd}$	Singular value and right singular vector of J_A
λ	nonnegative scalar	Spectral damping, relative to σ_1^2
D, q	maps to $\mathbb{R}^{n_q}$	Fixed linear readout and simulator task coordinates
$s^A_{\tau_T(A)}$	terminal simulator state	Endpoint at termination or control-step budget T
$c_{r,r'}, \Omega$	effect cosine, source contrast	Source pairing statistics in a specified coordinate space
h	$5, 10, 15, 25$	Primitive-step horizons in the released-system assay
$S_{m,c,\rho,h}$	$(-\infty, 1]$	Physical-effect skill for model m and source c
$\overline{S}^{\mathcal{G}}_m$	$(-\infty, 1]$	Equal-cell mean over the specified 60-cell grid

B TECHNICAL DERIVATIONS

We formalize three local facts used to interpret the experiments: a JVP remainder bound, a first-order criterion for recipient transfer, and the role of the proposal-centered trust radius. Multi-step transport is measured empirically.

B.1 LOCAL REMAINDER OF THE ACTION-EFFECT LINEARIZATION

Fix a recipient r and suppress it in the notation. Let $F(A) = F_H(r, A)$. The first-order action effect around A is $J_A(r, A)\delta A$. The following standard remainder bound makes it precise why the finite-difference assay calibrates a local object.

Proposition B.1 (Quadratic remainder under a local Lipschitz Jacobian)**.** *Assume F is differentiable on the line segment $\{A + t\delta A : t \in [0, 1]\}$ and its action Jacobian is β-Lipschitz there in operator norm:*

$$\|J_A(r, A_1) - J_A(r, A_2)\|_{\mathrm{op}} \le \beta\|A_1 - A_2\|_2. \tag{17}$$

Then

$$\|F_H(r, A + \delta A) - F_H(r, A) - J_A(r, A)\delta A\|_2 \le \frac{\beta}{2}\|\delta A\|_2^2. \tag{18}$$

Proof. By the fundamental theorem of calculus along the action segment,

$$F(A + \delta A) - F(A) = \int_0^1 J_A(r, A + t\delta A)\,\delta A\,dt. \tag{19}$$

Subtracting $J_A(r, A)\delta A$ and applying the triangle inequality gives

$$\|F(A + \delta A) - F(A) - J_A(r, A)\delta A\|_2 \tag{20}$$

$$\leq \int_0^1 \|J_A(r, A + t\delta A) - J_A(r, A)\|_{\text{op}} \, \|\delta A\|_2 \, dt \tag{21}$$

$$\leq \int_0^1 \beta t \|\delta A\|_2^2 \, dt = \frac{\beta}{2}\|\delta A\|_2^2. \tag{22}$$

□

Interpretation. The proposition concerns the frozen predictor itself. Under local smoothness, the JVP approximates its exact alternative-action difference with a second-order residual. Appendix D.1.1 measures this approximation. Same-state replay evaluates environmental validity separately.

B.2 Recipient transfer as a local descent criterion

The goal-directed pullback is $g_r = \nabla_A \mathcal{L}_r(A)$. Consider a small update $A^+ = A - \eta\xi$ evaluated by the execution recipient r.

Proposition B.2 (First-order effect of a norm-matched transferred direction)**.** *Assume that $\nabla_A \mathcal{L}_r$ is L-Lipschitz on the segment between A and $A - \eta\xi$. Then, for any direction ξ,*

$$\mathcal{L}_r(A - \eta\xi) = \mathcal{L}_r(A) - \eta g_r^\top \xi + R_r(\eta, \xi), \qquad |R_r(\eta, \xi)| \leq \frac{L}{2}\eta^2\|\xi\|_2^2. \tag{23}$$

For the matched direction $\xi_M = g_r$,

$$\mathcal{L}_r(A - \eta\xi_M) = \mathcal{L}_r(A) - \eta\|g_r\|_2^2 + R_r(\eta, g_r), \tag{24}$$

with $|R_r(\eta, g_r)| \leq \frac{L}{2}\eta^2\|g_r\|_2^2$. For nonzero g_r and $g_{r'}$, define the norm-matched donor direction

$$\xi_D = \|g_r\|_2 \frac{g_{r'}}{\|g_{r'}\|_2}. \tag{25}$$

Its first-order contribution is

$$-\eta\|g_r\|_2^2 \cos(g_r, g_{r'}) \, . \tag{26}$$

Proof. The standard descent lemma for an L-smooth function gives the remainder bound in (23). Substituting $\xi_M = g_r$ yields (24). For ξ_D,

$$g_r^\top \xi_D = \|g_r\|_2 \frac{g_r^\top g_{r'}}{\|g_{r'}\|_2} \tag{27}$$

$$= \|g_r\|_2^2 \cos(g_r, g_{r'}) \, , \tag{28}$$

which gives (26). □

Corollary B.3 (Local transfer criterion)**.** *For nonzero g_r and $g_{r'}$, the matched direction maximizes the first-order decrease of $\mathcal{L}_r$ among directions with norm $\|g_r\|_2$. A norm-matched donor gives first-order descent when $\cos(g_r, g_{r'}) > 0$, no first-order change when the cosine is zero, and first-order ascent when it is negative.*

Proof. The first claim follows from Cauchy–Schwarz: $g_r^\top \xi \leq \|g_r\|_2\|\xi\|_2 = \|g_r\|_2^2$, with equality for $\xi = g_r$. The sign statements follow directly from (26). □

Equal-norm Random as a local null control. Conditional on g_r, draw u uniformly on the unit sphere and set $\xi_R = \|g_r\|_2 u$. Symmetry gives

$$\mathbb{E}[g_r^\top \xi_R \mid g_r] = 0, \tag{29}$$

so the expected first-order change of $\mathcal{L}_r$ under the Random direction is zero. This defines the local null comparison. Multi-step behavior also depends on the refresh schedule, measured in Appendix D.3.2.

Relation to the radius experiment. In the source-swap stress test, Matched and Donor recompute their own local pullbacks at every optimization step. Their trajectories therefore visit different parts of the frozen model's action landscape. The cosine criterion measures alignment at a single action point. The mismatch–by–radius experiment also measures how optimization traverses that landscape.

B.3 Trust radius as an extrapolation constraint

For a feasible proposal $A_0 \in \mathcal{A}$, radial projection followed by box clamping places every iterate in

$$\mathcal{A} \cap B(A_0, \rho), \tag{30}$$

so, by construction,

$$\|A_k - A_0\|_2 \leq \rho \qquad \text{for every iterate } A_k. \tag{31}$$

Thus ρ bounds how far locally recomputed responses may move the plan from the proposal. In the main experiment, "extrapolation" refers to enlarging this visited action region while recomputing the response along the trajectory. The trust-radius sweep measures this range. Proposition B.1 instead concerns first-order fidelity at one fixed action point.

C Experimental protocols

C.1 Systems, proposals, and action interfaces

The four LeWM tasks are Push-T, Reacher, OGBench-Cube, and Two-Room; the transport intervention is repeated under INTACT on the corresponding tasks. LeWM uses a learned actor with optional demonstration retrieval, INTACT uses its native Direct proposal, and GC-IDM supplies an independently trained inverse-controller proposal on Push-T. The encoder, forward model, and proposal source remain frozen throughout evaluation. Radii and action displacements are measured in each model's standardized action coordinates, including the GC-IDM correction.

GC-IDM emits a primitive two-dimensional action whereas LeWM consumes a five-step macro block. The frozen-model query therefore uses a constant-action local surrogate, and the controller replans after executing the first primitive. This adaptation is fixed across the GC-IDM comparison.

C.2 Paired evaluation and simulator control

All core behavioral comparisons use fixed recipient lists and deterministic simulator seeds. Matched, Donor, Random, and unrefined conditions share the evaluation episode, start step, proposal, goal, model checkpoint, and simulator realization. The source swap moves observation history, action history, and cached representation together. The external execution recipient remains fixed, as does one donor within each solve. Official action bounds are applied before execution.

For the LeWM transport study, each task crosses three actor seeds with 150 unique recipients per seed. INTACT task-wise results preserve native paired recipients and simulator realizations. Training variation is tested on three independently trained INTACT Reacher checkpoints using the same 300 recipient identities per checkpoint.

C.3 Same-state environment replay

For each replay recipient, both branches restore the same simulator state and seed. The model query first produces a post-projection perturbation δA. The environment executes plans A and $A + \delta A$ with a 25-control-step budget, stopping each branch at task termination. Each completed branch retains its terminal observation. The realized effect is the difference between their terminal encoder latents. Push-T and Reacher each cross three actor seeds with the same 100 recipient identities.

For task-coordinate grounding, a fixed ridge readout maps the frozen 192-D latent to standardized task state. Its fitting data consists of factual current and goal frames from the task's training episodes; evaluation uses disjoint development episodes. Push-T coordinates are normalized block x, y and periodic orientation $(\sin\theta, \cos\theta)$; Reacher coordinates are periodic joint configurations. Each coordinate is standardized on the training split before computing vector cosine and relative magnitude error. The frozen readout is applied to predicted endpoints to measure task coordinates. Per-coordinate

development R^2 is $(.982, .975, .911, .932)$ for Push-T and $(.9997, .9998, .9993, .9996)$ for Reacher; standardized root mean squared error (RMSE) is $(.136, .161, .297, .262)$ and $(.017, .014, .027, .019)$, respectively.

Nearest and farthest donors must come from different episodes. Donors are ranked by the mean cosine similarity of state, goal-intent, and proposal features. The nearest donor therefore resembles the recipient in the quantities available to the assay while remaining a distinct interaction.

For reciprocal replay, each recipient receives one isotropic Gaussian direction in plan space, normalized to unit ℓ_2 norm and scaled to the requested radius before action projection, independently of the model gradient. The reciprocal assay uses the same base plan and actual post-projection δA in all four source–environment cells. The simulator is restored separately to E_r and $E_{r'}$. Each state executes both A and $A + \delta A$ under its paired deterministic seed. Reacher action bounds can shorten the requested perturbation, so the assay records the perturbation after projection and fixes it across all cells. The diagonal and off-diagonal contrasts then change only the relation between prediction source and realized effect. Cosines use a 10^{-12} floor on each effect norm; pairs with a zero predicted or observed effect receive cosine zero and remain in the reported averages.

C.4 Released-model operating-envelope assay

We compare the released LeWM, DINO-WM, and JEPA-WM Push-T checkpoints (Maes et al., 2026; Zhou et al., 2025; Terver et al., 2026). The common bank contains 240 distinct episodes sampled from the official evaluation split after excluding a prior 150-episode fixed-duration validity bank and a separate 400-episode capability bank. One start is sampled per episode subject to 25 future recorded actions lying inside the official $[-1, 1]^2$ action box.

The next 25 recorded physical actions define a model-neutral base plan. We standardize them with factual-training action statistics and group the first five primitive actions into a 10-D block. Four isotropic unit directions per recipient are reused at radii $\{.05, .1, .25, .5, 1\}$. Each plus-minus chord modifies only that first block; actions 6–25 remain identical. Symmetric feasibility uses the smaller two-sided distance to the action bound. All evaluated perturbations satisfy the action bounds. Simulator dynamics continue through task success and are measured at horizons $\{5, 10, 15, 25\}$.

Matched, Near, and Far initial states are fixed across models. Near and Far are the closest and farthest other recipients under standardized pusher x, y, block x, y, and block $(\sin\theta, \cos\theta)$. Each model re-encodes the selected interaction and predicts both branches. Separately fitted linear ridge readouts map the released representations to standardized block $(x, y, \sin\theta, \cos\theta)$; all are frozen before intervention and have development-coordinate R^2 between .911 and .996. The environment effect comes directly from simulator block pose.

The primary local score uses all 240 recipients. On a fixed subset of 60 recipients, observed branch images are also re-encoded to estimate native-space effect grounding and decoder-only physical-effect error. A post-hoc decomposition aligns these decoded effects with the model predictions (Appendix D.5). We resample recipient episodes, retain all four directions in each cluster, and pair bootstrap draws across models and cells. The prespecified family contains three global model comparisons and, for each model, far-minus-matched, far-local-minus-direct, $h = 25$ minus $h = 5$, and $\rho = 1$ minus $\rho = .05$ contrasts. We use 5,000 draws and 15-test Bonferroni-adjusted percentile intervals.

C.5 Transport, proposal correction, and planner controls

The transport study recomputes guidance for $K = 30$ updates at $\rho \in \{0.25, 0.5, 1.0, \infty\}$. Matched uses the execution interaction, Donor uses a fixed cyclic source interaction, and Random draws a fresh equal-norm isotropic direction. The primary transport statistic is the growth in the Matched–Donor success gap from the tight to the unbounded condition.

Action refinement uses Adam with learning rate .05, $(\beta_1, \beta_2) = (.9, .999)$, $\epsilon = 10^{-8}$, zero weight decay, and AMSGrad disabled. Optimizer moments start at zero for each proposal to solve. LeWM transport clips each sample's gradient L2 norm to 1; INTACT transport uses no gradient clipping. At every Donor or Random iterate, the recipient gradient is recomputed at that condition's current action. After any clipping, the source direction is rescaled per sample to this recipient norm, assigned as the

Adam gradient, and followed by radial projection around A_0 and clamping to the official action box. Norm matching therefore applies before Adam and projection; realized update lengths can differ. Transport accepts the final iterate without a latent-cost acceptance test.

The GC-IDM and INTACT comparisons use one Adam step with learning rate .05 and trust radius $\rho = .25$, without gradient clipping. This is the tight endpoint of the transport sweep, fixed across both proposal settings and all CEM budgets. Both the pullback and CEM retain the original proposal if the final candidate increases its latent cost. The CEM frontier uses populations $4, 8, 16, 32, 64, 128$ with respectively $2, 2, 2, 4, 8, 16$ elites and initial sampling scale .25. The largest setting uses three iterations; the others use one. Every population contains the original proposal, every candidate is projected individually, and the final elite mean is rescored and compared with the best sample retained across all rounds. The lower-cost candidate is selected.

C.6 Bootstrap intervals

The transport statistic is the change in the Matched–Donor success gap from the tight-radius to the unbounded condition. LeWM intervals use a crossed paired bootstrap over actor seed and unique recipient. The INTACT aggregate jointly resamples the checkpoint and shared recipient index. GC-IDM intervals resample proposal-training split and recipient hierarchically.

For same-state replay, Push-T and Reacher each cross three actor seeds with the same 100 recipient identities. Confidence intervals resample actor seed and recipient jointly. Reciprocal source specificity uses the same crossed unit. Near and Far donor identities remain inside the actor-seed factor because they may vary with the actor proposal. Contrasts are paired whenever the design supplies a paired realization.

D Supporting results

D.1 Action-effect measurement and environmental grounding

D.1.1 Finite-difference calibration

Finite differences calibrate the Jacobian against model-side effects (proposition B.1).

Table 5: Finite-difference validation of local action effects. Medians over 100 recipients per task.

Task	Radius	Cosine ↑	Normalized residual ↓
Push-T	0.01	0.999999	0.0020
Push-T	0.25	0.999462	0.0390
Push-T	1.00	0.991326	0.1539
Reacher	0.01	0.999998	0.0019
Reacher	0.25	0.998962	0.0462
Reacher	1.00	0.984406	0.1770

D.1.2 Same-state replay and response strength

Table 6: Cosine alignment between predicted and environment-realized terminal latent effects.

Task / ρ	Matched	Near	Far	Random	Matched − Near [95% CI]
Push-T / 0.05	0.512	0.069	0.023	-0.002	$0.443\ [0.358, 0.523]$
Push-T / 0.25	0.570	0.057	0.016	0.001	$0.512\ [0.435, 0.588]$
Push-T / 0.50	0.690	0.095	0.016	0.004	$0.595\ [0.524, 0.666]$
Reacher / 0.05	0.730	0.376	-0.083	0.000	$0.354\ [0.273, 0.437]$
Reacher / 0.25	0.661	0.328	-0.058	-0.002	$0.333\ [0.244, 0.421]$
Reacher / 0.50	0.761	0.395	-0.066	-0.006	$0.366\ [0.285, 0.449]$

All Matched-minus-Near, Far, and Random paired intervals are above zero.

Table 7: The same replay expressed in standardized task coordinates through a frozen factual-trajectory readout. Push-T readout mean development $R^2 = .950$ (standardized RMSE .214); Reacher $R^2 = .9996$ (RMSE .019).

Task / ρ	Matched	Near	Far	Random	Matched − Near [95% CI]
Push-T / .05	.440	.171	.002	.003	.270 [.166, .375]
Push-T / .25	.368	.108	-.035	-.002	.260 [.146, .374]
Push-T / .50	.512	.168	.032	-.062	.344 [.236, .454]
Reacher / .05	.822	.688	.415	-.013	.134 [.075, .203]
Reacher / .25	.718	.602	.343	-.054	.116 [.062, .180]
Reacher / .50	.830	.701	.393	.006	.129 [.072, .195]

Matched also exceeds Far and Random in every task-coordinate row, with all paired intervals above zero. Source matching therefore persists in direct simulator task coordinates. We also measure

$$\epsilon_{\text{mag}} = \frac{\left| \|\Delta_q^{\text{model}}\|_2 - \|\Delta_q^{\text{env}}\|_2 \right|}{\max(\|\Delta_q^{\text{env}}\|_2, 10^{-12})}. \tag{32}$$

Table 8: Median relative magnitude error at $\rho = .25$. Directional alignment is the primary grounding statistic; this table reports the complementary effect magnitude check.

Task	Matched	Near donor	Far donor
Push-T	.607	.779	.882
Reacher	.113	.115	.108

We report medians because 10.7% of Push-T replays have a near-zero realized effect, for which the mean relative error is unstable.

At $\rho = 0.25$, we pair model-side response magnitude with cosine alignment to the environment-realized effect. On Reacher, the response-strength intervals overlap tightly for Matched, Near, and Far: $[18.66, 19.32]$, $[18.54, 19.23]$, and $[18.46, 19.32]$. Their environmental alignment separates; Matched is positive $[.555, .754]$, while Far is below zero $[-.098, -.016]$.

D.1.3 Natural matched counterexamples

At $\rho = .25$, we select the top matched-sensitivity quartile within each actor seed and count wrong-direction effects. Wrong direction means an environment cosine at or below zero, and robust wrong cases also exceed the seed's bottom quartile of environment-effect norm.

Table 9: Matched-only sensitivity analysis in the planning representation. Wrong/high-response counts are out of 75 top-quartile effects per task. Robust wrong additionally excludes weak environment changes.

Task	Spearman ρ [95% CI]	Top / bottom alignment	Wrong / high response	Robust wrong
Push-T	0.252 $[0.062, 0.419]$	0.693 / 0.389	4 / 75	4
Reacher	0.119 $[-0.112, 0.336]$	0.713 / 0.643	8 / 75	1

The robust cases in Table 9 are distinct recipient identities. In task coordinates, the same fixed definition yields 13 robust wrong-direction observations from 7 Push-T recipients and 2 from 2 Reacher recipients. The clearest task-space cosines are $-.884$ and $-.853$, respectively. These cases place the sensitivity–meaning dissociation under the native matched source as well as donor transport.

D.2 Source dependence and operator–residual decomposition

D.2.1 Reciprocal source grounding

For the crossing in Table 2, planning-space Ω_z has paired 95% intervals $[.290, .413]$ and $[.259, .347]$ for Push-T Near and Far, and $[.287, .425]$ and $[.781, .925]$ for Reacher. It is positive for every actor

seed. All eight individual source advantages, $c_{r,r} - c_{r,r'}$ and $c_{r',r'} - c_{r',r}$, also have intervals above zero. Task-coordinate intervals appear in Table 10.

Table 10: Reciprocal ownership in standardized task coordinates.

Task	Donor	$r \to E_r$	$r \to E_{r'}$	$r' \to E_r$	$r' \to E_{r'}$	Ω_q [95% CI]
Push-T	Near	.394	.108	.102	.389	.286 [.197, .368]
Push-T	Far	.394	-.053	.031	.093	.254 [.177, .334]
Reacher	Near	.780	.827	.686	.940	.104 [.059, .157]
Reacher	Far	.780	.471	.389	.908	.414 [.325, .508]

The task-coordinate ownership contrast is positive for every actor seed and all four paired intervals exclude zero. The Near Reacher recipient effect has slightly higher alignment in the donor environment (.827 versus .780), but the donor effect has a larger own-source advantage (.940 versus .686). The complete crossing therefore favors the two own-source cells in aggregate, while also revealing partial transport between similar interactions.

D.2.2 Jacobian–residual decomposition

For the goal-directed pullback, write $e_r = F_H(r, A) - z_g$ and $v_r = J_r^\top e_r$. We recombine recipient and donor operators and residuals to separate two sources of interaction dependence. For each combination v, we execute $A^+ = (\Pi_{\mathcal{A}} \circ \Pi_{B(A,.25)})[A - .25v/\|v\|_2]$ from the same recipient state. This is one normalized descent step, with no Adam history. Across the four cells, mean realized displacement is .250 on Push-T and .2483–.2486 on Reacher; action-box clipping accounts for the small Reacher differences. Directional cosine and the environment terminal latent mean squared error (MSE) relative to the encoded goal are the primary diagnostics. Binary success supplies a behavioral check.

Table 11: Operator–residual 2×2 decomposition. Cosine is measured against the matched pullback $J_r^\top e_r$.

Task	Pullback	Cosine to matched	Success
Push-T	$J_r^\top e_r$	1.000	85.33%
Push-T	$J_{r'}^\top e_r$	.046	82.00%
Push-T	$J_r^\top e_{r'}$	.522	84.67%
Push-T	$J_{r'}^\top e_{r'}$	.103	81.33%
Reacher	$J_r^\top e_r$	1.000	31.00%
Reacher	$J_{r'}^\top e_r$	.006	26.00%
Reacher	$J_r^\top e_{r'}$	.908	30.33%
Reacher	$J_{r'}^\top e_{r'}$	-.043	21.00%

Swapping only the Jacobian worsens environment endpoint error by $+0.052$ $[0.006, 0.101]$ on Push-T and $+0.187$ $[0.139, 0.237]$ on Reacher relative to Matched. The corresponding residual-only intervals include zero. Binary-success differences follow the same ordering; their paired intervals include zero.

D.2.3 Random-probe controls for the source decomposition

We extend the operator–residual decomposition with random-probe controls on ten LeWM Push-T recipients at actor seed 0 and the same proposal. The diagnostic uses 32 isotropic action probes and 64 random latent cotangents per recipient. Table 12 reports the four pullback cells and the additional source-agreement measurements.

Residuals have mean cosine .356, so the residual intervention changes the latent direction appreciably even though its pullback retains cosine .695. Operator effects have low mean agreement under isotropic action probes (.050) and along the matched pullback (.019). The Jacobian-swap cosine (.078) is close to the random-cotangent mean (.077); their paired difference is .001 $[-.327, .325]$. These probes extend the observed disagreement beyond the particular goal residual, with substantial uncertainty in the comparison between probe families.

Table 12: Source-swap and random-probe controls on ten LeWM Push-T recipients. The first four rows give cosine to the matched pullback. Subsequent rows explicitly specify the compared vectors. Intervals resample paired recipients.

Quantity	Value
Matched $J_r^\top e_r$	1.000
Jacobian swap $J_{r'}^\top e_r$	.078
Residual swap $J_r^\top e_{r'}$	.695
Full donor $J_{r'}^\top e_{r'}$	-.037
Residual control $\cos(e_r, e_{r'})$	.356
Operator control $\cos(J_r u, J_{r'} u)$, isotropic u	.050
Operator control along $v_r/\|v_r\|$	.019
Cotangent null $\cos(J_r^\top w, J_{r'}^\top w)$, isotropic w	.077
Jacobian swap minus cotangent null	.001 [-.327,.325]
Gradient control $\cos(g_r, g_{r'})$	-.037
Random-direction null for $\cos(g_r, \cdot)$	-.001
Gradient control minus random null	-.036 [-.358,.287]
Fraction with negative $\cos(g_r, g_{r'})$	.500

At the proposal, the full donor gradient has mean cosine $-.037$ to the recipient gradient, and half the recipients have negative cosine. Its difference from the random-direction null is $-.036$ $[-.358, .287]$. This initial-direction measurement leaves both adverse individual steps and multi-step accumulation possible. The separate refresh experiment (Appendix D.3.2) measures schedule effects on complete solves.

D.2.4 Donor similarity and partial transport

Across donor-similarity quantiles, Reacher has slope 0.313 $[0.227, 0.395]$ and nearest–farthest difference 0.386 $[0.283, 0.486]$. Push-T's slope is near zero at $\rho = .25$ and positive at $\rho = .5$, showing less consistent gradation under the feature-based similarity criterion.

D.3 Optimization authority and guidance controls

D.3.1 Paired uncertainty across trust radii

Table 13 gives task-wise uncertainty. A separate INTACT Reacher replication uses the same 300 recipients at each of three checkpoint seeds. Gap amplification is 55.67, 57.00, and 53.00 percentage points at seeds 0, 42, and 3072, respectively. Jointly resampling the checkpoint and recipient gives 55.22 $[49.78, 60.44]$. The task-wise Reacher row uses checkpoint 0; it is distinct from this three-checkpoint aggregate.

Table 13: Paired uncertainty for the gap amplification in Table 3. Intervals use the resampling units specified in Appendix C.

System	Task	Gap amplification (pp)	Paired 95% interval
LeWM	Push-T	81.33	[74.67, 87.56]
LeWM	Reacher	31.78	[20.67, 43.11]
LeWM	Cube	36.89	[29.33, 44.67]
LeWM	Two-Room	9.11	[4.44, 14.22]
INTACT	Push-T	67.33	[61.33, 73.00]
INTACT	Reacher	55.67	[49.33, 62.00]
INTACT	Cube	39.33	[34.00, 45.00]
INTACT	Two-Room	11.33	[5.67, 17.00]

D.3.2 GUIDANCE SOURCE AND REFRESH SCHEDULE

A separate LeWM evaluation on Push-T and Reacher crosses guidance source (Donor or Random) with refresh schedule (Fixed or Fresh), alongside Matched and unrefined-proposal references, at $\rho \in \{.25, \infty\}$. Each task uses the same 150 recipient identities crossed with three actor replications, giving 450 paired outcomes per condition and radius. All six conditions share the same proposals and actor checkpoints. The refinement conditions use 30 Adam updates, learning rate .05, norm matching, clipping, and action bounds specified in Appendix C.

Fixed Donor retains one source identity within each solve while recomputing its gradient at the current action. Fresh Donor cycles through other nonterminated evaluation episodes without self-pairing; with two such episodes, it has one donor. At $\rho = \infty$, every donor direction comes from another live episode. Fixed Random retains its sampled orientation, while Fresh Random redraws at every update; both share the initial draw and rematch norms at each step. Matched computes guidance at the execution interaction.

Let $p_{D,\mathrm{fix}}$ and $p_{D,\mathrm{fresh}}$ denote Donor success under the two schedules, with analogous notation for Random. The source–schedule interaction is

$$I_{\mathrm{refresh}} = (p_{D,\mathrm{fix}} - p_{R,\mathrm{fix}}) - (p_{D,\mathrm{fresh}} - p_{R,\mathrm{fresh}}).$$

Table 14 reports success contrasts at $\rho = \infty$. Pointwise 95% intervals use 20,000 crossed bootstrap draws over actor replication and recipient identity, preserving condition pairing.

Table 14: Source and refresh-schedule control at $\rho = \infty$. All entries are success differences in percentage points with crossed 95% confidence intervals, using 150 recipients crossed with three actor seeds.

Contrast	Push-T	Reacher
Matched − Fixed Donor	$+83.8$ $[77.1, 89.8]$	$+49.1$ $[39.6, 58.2]$
Matched − Fresh Donor	$+78.9$ $[72.4, 85.1]$	$+52.7$ $[44.7, 60.7]$
Fixed Donor − Fixed Random	-7.8 $[-14.4, -1.1]$	$+4.9$ $[-3.6, 13.3]$
Fresh Donor − Fresh Random	-58.0 $[-66.7, -48.9]$	-23.3 $[-31.6, -15.1]$
I_{refresh}	$+50.2$ $[38.7, 60.9]$	$+28.2$ $[17.3, 38.9]$

Matched exceeds both donor schedules on both tasks. Fresh Donor also falls below Fresh Random, whereas the Fixed Donor–Fixed Random contrast is smaller on Push-T and includes zero on Reacher. Both source–schedule interaction intervals are positive: Donor and Random's relative performance depends on the refresh schedule. At $\rho = .25$, the interaction is $-.9$ $[-4.7, 2.9]$ points on Push-T and -11.6 $[-21.8, -1.3]$ on Reacher.

At $\rho = \infty$, Fixed/Fresh Random action displacement is $10.61/2.22$ on Push-T and $8.96/2.10$ on Reacher, with consecutive-direction cosine 1 versus approximately 0. These refinement-batch averages are consistent with accumulation under fixed directions and cancellation under fresh draws. The refresh contrasts use paired outcomes from this source–schedule evaluation; Table 3 uses the radius sweep.

D.4 BOUNDED CORRECTION AND SEARCH COMPARISONS

Table 15 reports six CEM budgets with paired proposals, recipients, seeds, action bounds, and trust radius $\rho = .25$. Forward counts include the final elite-mean evaluation. Recovery and harm counts report repaired proposal failures and lost proposal successes, respectively. The common radius permits different realized displacements, reported in the final column.

Table 15: Success and model-query budgets for bounded correction. GC-IDM uses 600 paired outcomes (two training splits, 300 recipients each); INTACT uses 900 (three checkpoints, 300 recipients each). F counts forward candidate evaluations and B counts backward calls; candidates may be batched. In CEM, P is population size and I is the number of iterations. Rec./harm reports repaired proposal failures and lost proposal successes, respectively.

Setting	Method	Compute	Success (%)	Δ proposal (pp) [95% CI]	Rec./harm	$\|\Delta A\|_2$
GC-IDM Push-T	Proposal	0F	84.33	+0.00 [0.00, 0.00]	0/0	0.000
GC-IDM Push-T	One-step correction	2F+1B	87.67	+3.33 [1.33, 5.33]	28/8	0.070
GC-IDM Push-T	CEM $P = 4, I = 1$	5F	88.00	+3.67 [−0.33, 7.67]	41/19	0.168
GC-IDM Push-T	CEM $P = 8, I = 1$	9F	88.83	+4.50 [0.17, 8.83]	50/23	0.208
GC-IDM Push-T	CEM $P = 16, I = 1$	17F	86.67	+2.33 [−2.00, 6.67]	48/34	0.224
GC-IDM Push-T	CEM $P = 32, I = 1$	33F	86.50	+2.17 [−2.83, 6.84]	50/37	0.230
GC-IDM Push-T	CEM $P = 64, I = 1$	65F	87.00	+2.67 [−2.00, 7.33]	52/36	0.231
GC-IDM Push-T	CEM $P = 128, I = 3$	385F	87.00	+2.67 [−2.00, 7.33]	50/34	0.230
INTACT Reacher	Proposal	0F	95.67	+0.00 [0.00, 0.00]	0/0	0.000
INTACT Reacher	One-step correction	2F+1B	97.00	+1.33 [−0.22, 3.00]	19/7	0.166
INTACT Reacher	CEM $P = 4, I = 1$	5F	95.56	-0.11 [−1.67, 1.33]	9/10	0.202
INTACT Reacher	CEM $P = 8, I = 1$	9F	95.78	+0.11 [−1.67, 1.89]	13/12	0.233
INTACT Reacher	CEM $P = 16, I = 1$	17F	95.78	+0.11 [−1.56, 2.00]	11/10	0.238
INTACT Reacher	CEM $P = 32, I = 1$	33F	95.89	+0.22 [−1.33, 2.11]	13/11	0.238
INTACT Reacher	CEM $P = 64, I = 1$	65F	96.22	+0.56 [−0.78, 2.11]	12/7	0.242
INTACT Reacher	CEM $P = 128, I = 3$	385F	96.56	+0.89 [−0.89, 2.67]	19/11	0.241

D.5 Released-system action-effect evaluation

Tables 16 and 17 report grid means and prespecified contrasts; Fig. 4 resolves the full surface. Table 18 reports native effect skill and base prediction/readout errors before the paired error decomposition below.

Table 16: Grid-mean effect-skill profiles. Global intervals are pointwise 95% recipient-bootstrap intervals, and interaction columns are 20-cell point estimates. The final column counts cells with $S > 0$.

System	Global S [95% CI]	Matched	Near	Far	$S > 0$ cells
LeWM	-0.107 [-0.553,0.071]	-0.176	0.116	-0.262	37/60
DINO-WM	-2.279 [-3.383,-1.915]	-1.408	-0.953	-4.477	0/60
JEPA-WM	-4.864 [-6.641,-4.235]	-3.125	-2.549	-8.917	1/60

Table 17: All 15 prespecified effect-skill contrasts. Intervals are paired recipient-bootstrap percentile intervals with Bonferroni correction over the complete family. Horizon and radius contrasts subtract the smaller-endpoint score from the larger-endpoint score.

Scope	Contrast	Estimate	Adjusted 95% interval	Resolved
All	LeWM - DINO-WM	2.172	[1.291,3.846]	yes
All	LeWM - JEPA-WM	4.756	[3.616,7.304]	yes
All	DINO-WM - JEPA-WM	2.584	[1.759,4.242]	yes
LeWM	Far - Matched	-0.086	[-0.805,2.427]	no
LeWM	Far local - direct	2.145	[1.069,12.088]	yes
LeWM	$h = 25$ - $h = 5$	-0.356	[-3.587,0.320]	no
LeWM	$\rho = 1$ - $\rho = .05$	0.415	[0.038,2.749]	yes
DINO-WM	Far - Matched	-3.069	[-7.614,-1.090]	yes
DINO-WM	Far local - direct	1.591	[-0.424,4.709]	no
DINO-WM	$h = 25$ - $h = 5$	2.614	[-1.281,4.422]	no
DINO-WM	$\rho = 1$ - $\rho = .05$	1.651	[1.256,3.999]	yes
JEPA-WM	Far - Matched	-5.793	[-13.359,-2.395]	yes
JEPA-WM	Far local - direct	1.379	[-2.396,6.991]	no
JEPA-WM	$h = 25$ - $h = 5$	7.103	[0.767,11.493]	yes
JEPA-WM	$\rho = 1$ - $\rho = .05$	4.465	[3.218,8.983]	yes

Table 18: Native effect skill and physical-state errors. Native skill and decoder MSE use the fixed 60-recipient subset; model MSE uses all 240 recipients. Native skill averages the complete grid in within-model coordinates. Base MSE measures error in standardized physical coordinates under the unperturbed plan, using predicted latents (Model) or encoded observed outcomes (Decoder).

System	Native effect S	Model base MSE	Decoder base MSE
LeWM	0.310	0.091	0.064
DINO-WM	0.117	0.121	0.036
JEPA-WM	0.121	0.133	0.078

Predictor–readout error decomposition. Let $\hat{d}$ denote a predicted physical effect, d_R^{obs} the same frozen readout applied to the two real outcome images and differenced, and d the simulator effect. Aligning these vectors gives

$$e_P = \hat{d} - d_R^{\text{obs}}, \qquad e_R = d_R^{\text{obs}} - d, \qquad \|\hat{d} - d\|^2 = \|e_P\|^2 + \|e_R\|^2 + 2e_P^\top e_R. \tag{33}$$

Here e_P is the readout-projected predictor discrepancy and e_R is the observed-state encoding/readout error. The post-hoc decomposition uses all conditions on the same 60-recipient subset.

Table 19: Physical-effect error decomposition on the same 60 recipients. Equal-cell values average 60 cellwise error/effect-energy ratios; pooled values divide summed errors by summed effect energy. Total = Predictor + Readout + Cross = $1 - S$ under each aggregation.

	Equal-cell normalized energy				Energy-pooled normalized energy			
System	Total	Predictor	Readout	Cross	Total	Predictor	Readout	Cross
LeWM	1.224	6.668	6.209	-11.653	0.949	1.230	0.835	-1.116
DINO-WM	3.537	85.813	83.783	-166.059	1.539	5.273	4.356	-8.089
JEPA-WM	6.764	104.764	101.896	-199.896	2.192	6.273	5.084	-9.165

Both component errors are substantial and oppose each other (Table 19). Energy-pooled cross terms are -1.116 $[-1.458, -.884]$, -8.089 $[-11.506, -5.740]$, and -9.165 $[-12.778, -6.721]$ for LeWM, DINO-WM, and JEPA-WM, respectively (pointwise 95% paired recipient-bootstrap intervals, 5,000 draws). This cancellation accounts for the smaller end-to-end error relative to the two component energies. Native scores use model-specific coordinates: on the same 60 recipients, the DINO-WM–JEPA-WM contrast is 3.227 in physical space (pointwise paired 95% interval $[2.018, 5.559]$) and $-.004$ in native space ($[-.038, .031]$).

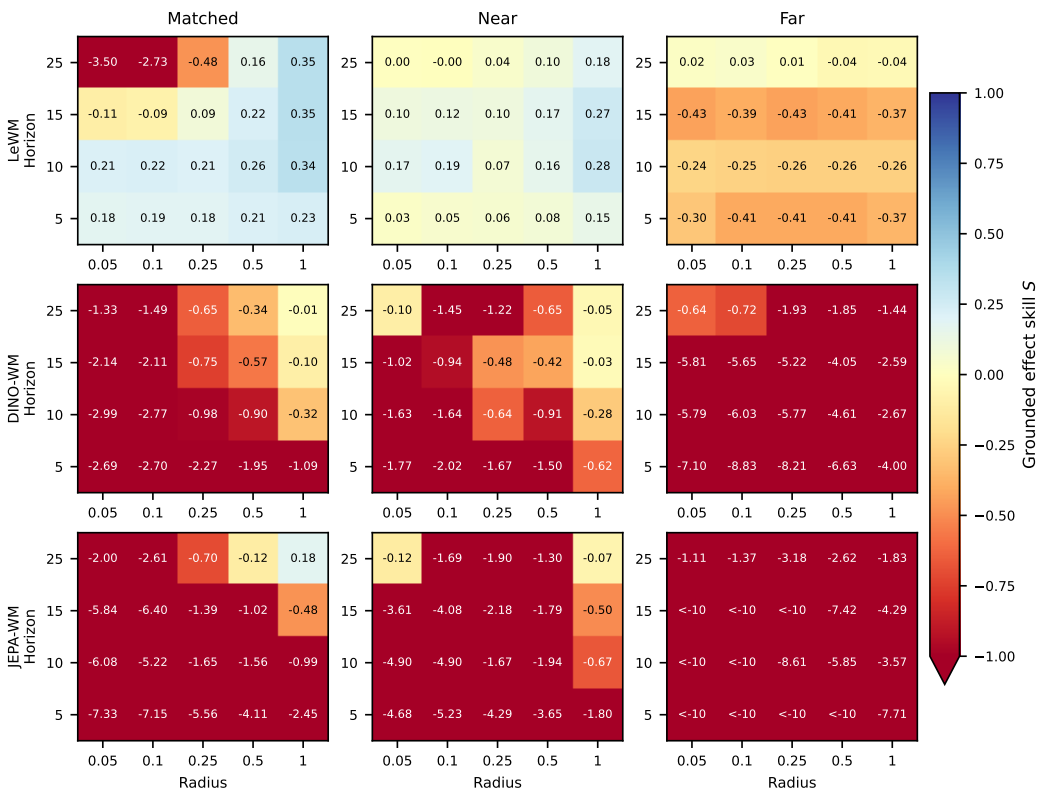


Figure 4: Complete local-effect skill surface for the released Push-T systems. Rows are models, columns are interaction sources, and each panel crosses horizon with perturbation radius. All panels use the same color scale; values below -1 saturate visually but retain their printed value. Printed values are point estimates.

D.6 Proposal-spectrum and optimization diagnostics

The following diagnostics use ten LeWM Push-T recipients at actor seed 0. They examine whether source disagreement is concentrated in weak-response directions of the frozen model. The proposal Jacobian maps 50 action coordinates to 192 latent coordinates. Its effective rank is $\exp(-\sum_k p_k \log p_k)$, with $p_k = \sigma_k^2 / \sum_j \sigma_j^2$. Table 20 shows effective ranks 2.04 and 2.56 for recipient and donor operators. The leading singular directions therefore dominate response energy, while later plan blocks have smaller Jacobian norms.

For each recipient right singular vector v_k, we measure cross-source agreement $\cos(J_r v_k, J_{r'} v_k)$ and squared operator discrepancy $\|(J_{r'} - J_r)v_k\|^2$. Band shares normalize these unit-probe discrepancy energies over all 50 directions; they describe operator differences rather than the error of a particular executed correction. Table 21 reports low source agreement throughout the spectrum, including the leading band. That band accounts for .647 of the operator-discrepancy energy, compared with .004 in ranks 26–50. Source differences are thus also present in directions of large model response.

The target effect for a unit singular direction is $.01\sigma_1$, giving a required step $.01\sigma_1/\sigma_k$. Probes at ranks $1, 2, 3, 13, 26, 50$ compare the Jacobian with central differences at the smaller of this step and 1. The linearized target requires a step within the unit-norm budget for all recipients at ranks 1–3, 30% at rank 13, and none at ranks 26 or 50. The first three probes have low linearization error. This norm-budget test does not measure feasibility inside the coordinate-wise action box.

We next run the same 30-update optimizer as the authority study and project the achieved displacement onto the recipient proposal's singular basis. Core share uses its first two directions; excited share uses $\sigma_k \geq .05\sigma_1$. The nonlinear source-discrepancy ratio is $\|\Delta z_{r'} - \Delta z_r\| / \|\Delta z_r\|$, where both effects are computed by the frozen predictor on the achieved displacement. Linearization error instead compares Δz_r with $J_r \delta A$. Tables 22 and 23 report the measurements and paired source contrasts.

Table 20: Proposal Jacobian spectrum on LeWM Push-T. Action-block norms are Frobenius norms restricted to the coordinates of one model plan step.

Quantity	Value
Effective rank, recipient	2.04
Effective rank, donor	2.56
Condition number, recipient	1.87e+05
Condition number, donor	2.57e+05
Singular value σ_1	13.518
Singular value σ_2	5.932
Singular value σ_3	2.839
Singular value σ_4	1.538
Singular value σ_5	.682
Action-block Jacobian norm, block 0	9.74
Action-block Jacobian norm, block 1	7.46
Action-block Jacobian norm, block 2	5.82
Action-block Jacobian norm, block 3	4.25
Action-block Jacobian norm, block 4	3.45
Finite-difference relative error	.0011

Table 21: Rank-band operator agreement and discrepancy, selected finite-difference probes, and a damped least-squares reference. Probe-budget fractions count recipients whose required step is at most 1; finite-difference errors use the required step capped at 1.

Rank band	$\bar{\sigma}_k$	$\cos(J_r v_k, J_{r'} v_k)$	Operator share	Step for fixed effect
1	13.52	.027	.647	.01
$2-3$	4.39	.117	.270	.04
$4-5$	1.11	.078	.032	.15
$6-10$	.21	.026	.024	.68
$11-25$	.05	.023	.024	3.69
$26-50$	4.1×10^{-3}	.008	.004	215.54
Probed ranks: target norm and linearity at the capped probe norm				
Rank	Step	Lin. error	Within budget	
1	.01	.000	1.00	
2	.02	.000	1.00	
3	.05	.001	1.00	
13	1.52	.170	.30	
26	9.33	.335	.00	
50	1865.37	.232	.00	
Damped least-squares reference, λ relative to σ_1^2				
λ	$\|\delta A\|$	Core share	Excited share	Linear source ratio
10^{-6}	153.64	.000	.001	4.27
10^{-4}	22.69	.002	.055	2.45
10^{-2}	2.51	.123	.847	1.86
10^{-1}	.84	.433	.963	1.72

The donor-minus-matched cost contrast increases from 13.85 $[8.17, 20.61]$ at $\rho = .25$ to 193.62 $[124.13, 265.85]$ without a radius bound. Core-share contrasts remain uncertain. The donor's mean nonlinear source-discrepancy ratio is 1.72 at the smallest radius and 1.37 without a bound, whereas its linearization error rises from .19 to 2.18. These observations do not support an explanation based solely on increasing relative source discrepancy in weak-response directions; the larger updates also leave the proposal's accurate linear regime.

The least-squares reference in Table 21 solves $(J_r^\top J_r + \lambda I)\delta A = -J_r^\top e_r$ before action projection. Its reported source-discrepancy ratio uses the two linearized operators. Increasing λ reduces the required action norm and concentrates the solution in higher-gain directions. This algebraic reference motivates the iterated damping diagnostic below.

Table 22: Achieved corrections by radius and guidance source. Shares use the recipient proposal's right singular basis. Source discrepancy and linearization error are model-side quantities.

Cell	$\|\delta A\|$	Core share	Excited share	Source ratio	Lin. error	Δcost
$\rho = 0.25$, donor	.25	.470	.665	1.72	.190	-3.56
$\rho = 0.25$, matched	.25	.553	.781	1.52	.153	-17.41
$\rho = 0.5$, donor	.50	.522	.725	1.75	.409	-3.06
$\rho = 0.5$, matched	.50	.473	.801	1.48	.278	-31.67
$\rho = 1.0$, donor	1.00	.461	.699	1.57	.662	12.90
$\rho = 1.0$, matched	.99	.394	.777	1.37	.495	-60.50
$\rho = \infty$, donor	6.00	.324	.580	1.37	2.182	84.52
$\rho = \infty$, matched	2.65	.203	.624	1.33	.897	-109.10

Table 23: Donor-minus-matched contrasts at each radius, with paired 95% recipient-bootstrap intervals. Each row compares guidance sources at one radius, rather than differences between radii.

Radius	Quantity	Donor	Matched	Donor − matched
$\rho = 0.25$	Core energy share	.47	.55	-.08 [-.26,.10]
$\rho = 0.25$	Cost change Δc	-3.56	-17.41	13.85 [8.17,20.61]
$\rho = 0.25$	Step norm $\|\delta A\|$	.25	.25	.00 [.00,.00]
$\rho = 0.25$	Nonlinear model-source discrepancy ratio	1.72	1.52	.20 [-.03,.47]
$\rho = 0.25$	Linearization error	.19	.15	.04 [-.02,.11]
$\rho = 0.5$	Core energy share	.52	.47	.05 [-.15,.24]
$\rho = 0.5$	Cost change Δc	-3.06	-31.67	28.61 [17.39,42.20]
$\rho = 0.5$	Step norm $\|\delta A\|$	.50	.50	.00 [.00,.00]
$\rho = 0.5$	Nonlinear model-source discrepancy ratio	1.75	1.48	.27 [.05,.60]
$\rho = 0.5$	Linearization error	.41	.28	.13 [-.02,.36]
$\rho = 1.0$	Core energy share	.46	.39	.07 [-.13,.27]
$\rho = 1.0$	Cost change Δc	12.90	-60.50	73.40 [42.96,105.21]
$\rho = 1.0$	Step norm $\|\delta A\|$	1.00	.99	.01 [.00,.02]
$\rho = 1.0$	Nonlinear model-source discrepancy ratio	1.57	1.37	.20 [.00,.39]
$\rho = 1.0$	Linearization error	.66	.50	.17 [-.07,.49]
$\rho = \infty$	Core energy share	.32	.20	.12 [-.01,.26]
$\rho = \infty$	Cost change Δc	84.52	-109.10	193.62 [124.13,265.85]
$\rho = \infty$	Step norm $\|\delta A\|$	6.00	2.65	3.35 [2.12,4.59]
$\rho = \infty$	Nonlinear model-source discrepancy ratio	1.37	1.33	.04 [-.44,.34]
$\rho = \infty$	Linearization error	2.18	.90	1.29 [.54,2.08]

D.7 Spectral damping inside the refinement solver

We filter the selected gradient in the recipient proposal's right singular basis using weights $\sigma_k^2/(\sigma_k^2 + \lambda)$, with λ relative to σ_1^2. We then renormalize the filtered vector to the original gradient norm before Adam. The same ten recipients, proposals, and 30 updates are used at $\rho = .25$ and $\rho = \infty$. Adam and action projection can change the resulting displacement norm, which we report alongside the outcome in Table 24.

At the largest tested damping, $\lambda = .1\sigma_1^2$, core share rises in all four guidance–radius cells, with paired intervals above zero. For unbounded Matched guidance it rises by .121 $[.012, .256]$, while mean displacement falls from 2.65 to 1.87 and cost reduction shrinks by 7.19 $[3.11, 12.57]$. Donor cost-change contrasts remain uncertain at both radii: -1.87 $[-4.42, .28]$ at $\rho = .25$ and 20.75 $[-6.19, 48.48]$ without a bound. Spectral concentration is therefore achieved but does not yield a resolved donor benefit in this sample. This comparison evaluates the filtered optimizer as a whole; equal input gradient norms do not isolate spectral direction from achieved step length.

Table 24: Spectral filtering with matched pre-Adam gradient norms. Core and excited shares describe the achieved correction in the recipient proposal's singular basis. The lower block gives damped-minus-gradient contrasts with paired 95% recipient-bootstrap intervals.

Radius	Guidance	λ	Δc	$\|\delta A\|$	Core	Excited
$\rho = 0.25$	Matched	—	-17.41	.25	.553	.781
$\rho = 0.25$	Matched	$1\times10^{-3}\sigma_1^2$	-17.27	.25	.579	.804
$\rho = 0.25$	Matched	$1\times10^{-2}\sigma_1^2$	-17.25	.25	.613	.810
$\rho = 0.25$	Matched	$1\times10^{-1}\sigma_1^2$	-16.57	.25	.675	.863
$\rho = 0.25$	Donor	—	-3.56	.25	.470	.665
$\rho = 0.25$	Donor	$1\times10^{-3}\sigma_1^2$	-4.61	.25	.571	.784
$\rho = 0.25$	Donor	$1\times10^{-2}\sigma_1^2$	-5.47	.25	.657	.808
$\rho = 0.25$	Donor	$1\times10^{-1}\sigma_1^2$	-5.43	.25	.693	.811
$\rho = \infty$	Matched	—	-109.10	2.65	.203	.624
$\rho = \infty$	Matched	$1\times10^{-3}\sigma_1^2$	-106.98	2.37	.216	.728
$\rho = \infty$	Matched	$1\times10^{-2}\sigma_1^2$	-104.46	2.13	.281	.789
$\rho = \infty$	Matched	$1\times10^{-1}\sigma_1^2$	-101.92	1.87	.324	.801
$\rho = \infty$	Donor	—	84.52	6.00	.324	.580
$\rho = \infty$	Donor	$1\times10^{-3}\sigma_1^2$	89.10	5.67	.378	.759
$\rho = \infty$	Donor	$1\times10^{-2}\sigma_1^2$	99.34	5.56	.461	.881
$\rho = \infty$	Donor	$1\times10^{-1}\sigma_1^2$	105.26	5.42	.540	.916

Damped minus gradient, paired 95% intervals

Radius	Guidance	λ	Δc	Core	Excited
$\rho = 0.25$	Matched	$1\times10^{-3}\sigma_1^2$	.139 [-.015,.329]	.025 [.002,.062]	.023 [-.002,.057]
$\rho = 0.25$	Matched	$1\times10^{-2}\sigma_1^2$	.153 [-.154,.476]	.059 [.018,.110]	.029 [-.003,.070]
$\rho = 0.25$	Matched	$1\times10^{-1}\sigma_1^2$	.837 [.387,1.343]	.121 [.068,.178]	.082 [.038,.131]
$\rho = 0.25$	Donor	$1\times10^{-3}\sigma_1^2$	-1.059 [-3.342,.401]	.101 [.046,.161]	.119 [.065,.177]
$\rho = 0.25$	Donor	$1\times10^{-2}\sigma_1^2$	-1.919 [-4.608,.212]	.187 [.084,.299]	.143 [.075,.211]
$\rho = 0.25$	Donor	$1\times10^{-1}\sigma_1^2$	-1.874 [-4.422,.276]	.223 [.110,.340]	.145 [.080,.216]
$\rho = \infty$	Matched	$1\times10^{-3}\sigma_1^2$	2.125 [.733,3.803]	.013 [-.011,.039]	.105 [.067,.149]
$\rho = \infty$	Matched	$1\times10^{-2}\sigma_1^2$	4.643 [1.638,8.461]	.079 [.008,.165]	.165 [.112,.225]
$\rho = \infty$	Matched	$1\times10^{-1}\sigma_1^2$	7.185 [3.109,12.567]	.121 [.012,.256]	.177 [.108,.253]
$\rho = \infty$	Donor	$1\times10^{-3}\sigma_1^2$	4.586 [-11.308,22.871]	.054 [-.003,.108]	.178 [.101,.251]
$\rho = \infty$	Donor	$1\times10^{-2}\sigma_1^2$	14.826 [-12.356,44.541]	.137 [.078,.206]	.301 [.186,.428]
$\rho = \infty$	Donor	$1\times10^{-1}\sigma_1^2$	20.745 [-6.192,48.483]	.215 [.092,.368]	.335 [.225,.465]