CAUSAL TRANSFER OF SCENE INFORMATION IN ROBOT POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

An internal edit can move a robot toward a donor's target without identifying what scene information it transferred. We study whether a controlled scene factor can be transported through a policy's internal state into predictable physical behavior. Using target location, we generate counterfactual observations, replace key–value caches at selected recipient layers with donor caches, and execute all branches from a common simulated physical state. Our primary frozen confirmation compares the true cache difference with four fixed channel-sign transformations that preserve each selected layer's K/V perturbation norm within 2% while changing its direction. Across two manipulation tasks, FastWAM and $\pi_{0.5}$, 128 new model rollouts yield 117 reachable and 114 visually eligible states; all four task–model cells meet the joint criterion for bidirectional relative endpoint effects exceeding the equal-norm controls, with control-adjusted means of 51–72 mm. A separate non-target spatial comparison supports three cells, with the fourth limited by a prespecified control-visibility requirement. In bounded two-axis development pilots (four model-visited states per policy family), own-axis effects exceeded cross-axis responses for every tested state–axis pair; these pilots are descriptive rather than independent confirmation. Independent offline reconstruction reproduces the primary effects over 2,340 branches. Together, these results show that, under the tested cache intervention and short simulated horizon, the direction of target-conditioned internal differences—not only their magnitude—can be transferred to produce predictable relative object-motion biases.

1 INTRODUCTION

An internal edit can move a robot toward a donor's target without identifying what was transferred. For vision–language–action (VLA) policies, the scientific question is not simply whether an edit changes behavior, but whether a named, task-relevant scene factor leaves a directional signature that survives a matched alternative edit. The practical problem is therefore evidential: how can we distinguish factor-linked behavioral influence from generic intervention effectiveness?

This ambiguity exposes a gap in current interpretability tools. A probe can show that a scene variable is decodable from an internal state, but decodability does not establish that downstream control uses it. Conversely, an activation edit can change behavior, but a behavioral change under a large edit may reflect generic disruption rather than transfer of a particular piece of world information. We therefore need a test that predicts a behavioral direction and separates content from perturbation magnitude.

We ask: can task-relevant scene information be transplanted into a robot policy while the physical world is held fixed? Our test constructs donor and recipient computations from different rendered target positions, while every branch executes from the same reconstructed center physical state. The rendered target is moved to compute internal values, not to change execution physics. If the internal state carries behaviorally actionable target information, replacing selected recipient cache layers with donor values should produce a predictable, donor-consistent bias in the manipulated object's motion. Reversing the donor direction predicts the opposite relative bias.

We instantiate this test with target location. A target-right donor and target-left recipient define a cache difference $D - R$. We compare this true difference with prespecified transformations that preserve its selected-layer K/V norm while changing its channelwise direction. The experiment

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.

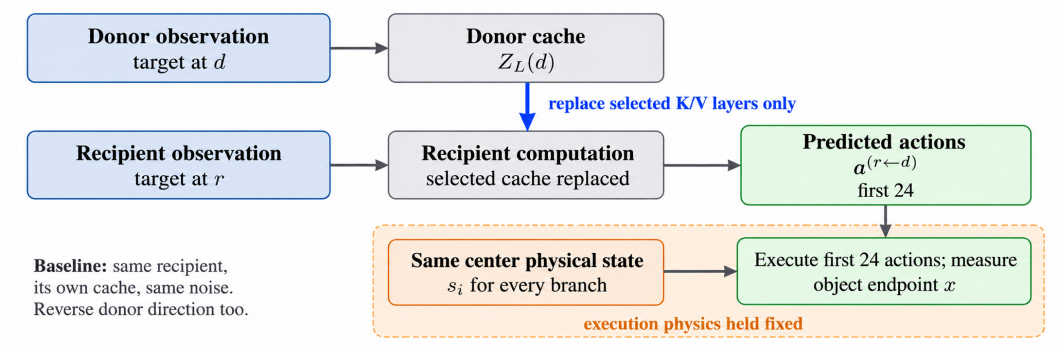

Figure 1: **Separate donor conditioning from execution physics.** A target-position change generates donor internal values. Only selected cache layers enter the recipient computation; baseline and intervention actions execute from the same center physical state. Reversing donor direction predicts the opposite *relative* endpoint bias. Equal-norm controls alter the donor–recipient cache-difference direction. Schematic, not a simulated trajectory.

therefore asks whether the structure of a target-conditioned difference matters, rather than merely whether an equally large edit can move the robot. The donor supplies internal K/V, not an action sequence, and the physical target is never moved to accommodate the transplanted values. A bounded orthogonal-axis diagnostic additionally asks whether the endpoint response distinguishes which target coordinate changed; it is development evidence, not part of the primary support gate.

Internal steering and behavior transfer already have direct precedents (Häon et al., 2025; Grant et al., 2026; Turner et al., 2023; Li et al., 2023; Zou et al., 2023), and controlled input contrasts have been used to construct WAM edits (Hong et al., 2026). We make three narrower contributions. First, we define a factor-controlled transplant test with bidirectional behavioral predictions and common-scene physical execution. Second, we compare the transplant with layerwise equal-norm cache controls, so that a signed endpoint effect is not attributed to perturbation size alone. Third, we freeze the protocol and evaluate new initial conditions and policy-visited states across two tasks and two policy families; all four task–model cells meet the primary direction-over-magnitude criterion. We use *content transfer* operationally for this factor-linked intervention effect, scoped to the tested cache intervention, tasks and short simulated horizon.

2 Related Work

Internal interventions in robot policies. Häon et al. (2025) steer vision–language–action (VLA) policies through internal semantic directions, including direction and speed. Grant et al. (2026) inject source-trajectory activations across scene, task and seed conditions, report source-position effects, and investigate goal-discriminant subspace injection. These are direct antecedents of behavioral transfer. Our factor-controlled cache test is a protocol distinction, not a claim of superior steering or broader applicability.

The policy family studied here sits within a broader progression from large-scale robotic transformers and web-to-robot transfer (Brohan et al., 2022; 2023) to multi-robot pretraining and open generalist policies (Open X-Embodiment Collaboration, 2023; Ghosh et al., 2024; Kim et al., 2024). The recent π_0 line also makes the relationship between a pretrained vision–language backbone and action generation explicit (Black et al., 2024). These works establish capable policy families; they do not by themselves identify which scene-conditioned internal differences are causally used by a downstream physical computation.

More recent methods connect edits to behavioral events (Jin et al., 2026), adapt intervention strength using feedback (Babu et al., 2026), and learn affine activation edits for multirobot control (Das et al., 2026). Hong et al. (2026) derive contrastive WAM directions from camera, gripper and image-noise perturbations and regulate them with WA-LQR. Thus neither stage-aware intervention nor constructing edits from input contrasts is new in isolation. We instead evaluate a single target-conditioned

transplant and a signed object outcome without learning a steering controller or optimizing task success.

Our short action window is an intervention readout rather than a new action-generation method. Action-sequence policies such as ACT and diffusion policies motivate reporting the executed action prefix and physical endpoint rather than treating one predicted command as the behavioral outcome (Zhao et al., 2023; Chi et al., 2023).

Decoding, patching and causal interpretation. Probing internal representations for scene-transition information (Molinari et al., 2025) differs from testing the physical effect of a transplanted value. Causal mediation and causal tracing introduced influential ways to connect internal components to model behavior (Vig et al., 2020; Meng et al., 2022); causal scrubbing and automated circuit discovery further emphasize testing an explicit mechanistic hypothesis rather than treating one patch effect as a complete explanation (Chan et al., 2022; Conmy et al., 2023). Interchange interventions can test correspondence with high-level causal variables (Geiger et al., 2021; 2023), but we do not establish such a complete abstraction here. Activation-patching findings depend on intervention and metric choices (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024), and localization need not predict the success of a downstream edit (Hase & Bansal, 2023); scalable patch-effect estimators also require their own reliability checks (Kramár et al., 2024). These precedents motivate fixed definitions and explicit limitations. Our policies and environment come from FastWAM (Yuan et al., 2026), $\pi_{0.5}$ (Physical Intelligence et al., 2025) and RoboTwin 2.0 (Chen et al., 2025); we introduce no new pretrained architecture.

Our contribution concerns what an intervention identifies. The frozen study tests whether a named scene factor predicts signed behavior beyond specified magnitude-matched edits under fixed execution physics. A separate two-axis development diagnostic asks whether that response distinguishes which target coordinate changed, rather than merely following donor sign. Together these tests distinguish effective steering, directional bias and coordinate-selective bias; they do not equate any of them with identifying a complete internal scene representation.

3 A CONTROLLED TEST OF TARGET-CONDITIONED INFORMATION

3.1 COMMON-STATE COUNTERFACTUALS

At a candidate state s_i reached by a policy's own normal rollout, we render center, target-left (L) and target-right (R) inputs. Left/right moves the task target by $-0.05/+0.05$ m along world x; opposite donor and recipient targets therefore differ by 0.10 m. Robot state, task instruction and every non-target entity remain fixed. A static, non-colliding orange decor object is present throughout each normal trajectory and all three inputs; it is not moved for the primary control. Figure 2 shows saved simulator pixels.

Before every simulated execution we independently reconstruct the center scene and replay the saved normal action prefix. Images, contacts, robot/object states and policy history must match the recorded boundary. Counterfactual target positions are never retained in the execution scene.

3.2 INTERNAL INTERVENTION AND PREDICTION

Let $Z_{\mathcal{L}}(r)$ be the selected layer-band cache computed on recipient input r at state s_i, and let U denote the paired inference randomness. We compute

$$a_i^{r\leftarrow d} = \pi\big(r, U; \mathrm{do}(Z_{\mathcal{L}} = Z_{\mathcal{L}}(d))\big), \qquad x_i(r \leftarrow d) = \mathrm{Exec}_x(s_i, a_i^{r\leftarrow d}). \tag{1}$$

Here do denotes an implemented replacement in the computation, not an identified intervention on a semantic variable. All non-selected cache layers retain recipient values. We compare with $x_i(r)$, the endpoint under the unmodified recipient computation with the same U. Neither $x_i(r)$ nor $x_i(r \leftarrow d)$ substitutes the center-input prediction for the recipient prediction.

All layer numbers are zero-based and inclusive. FastWAM uses layers 10–19 of its visual-conditioning K/V cache, with its verified high-precision action-attention path. $\pi_{0.5}$ uses layers 12–17 of its *entire multimodal prefix cache*, including image and language/state positions. Unchanged language tokens do not imply unchanged contextualized values when images differ. Replacement

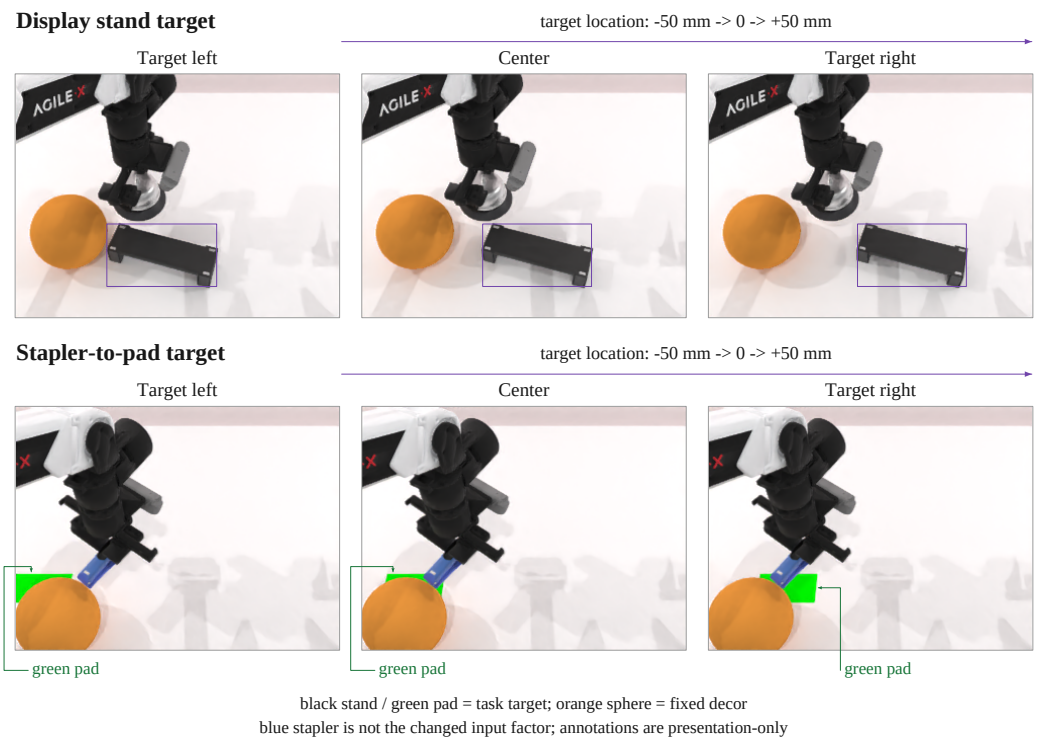

Figure 2: **Saved inputs, not different execution environments.** Head-camera views for display stand (top) and stapler-to-pad (bottom), using the first eligible FastWAM case in each task (episodes 1 and 66), without effect-based selection. The target moves only for input construction; all branches execute in the center physical scene. The orange sphere is fixed decor, not the task target; in the lower row, the colored simulator pad geometry is the task target. The policy receives head and both wrist cameras.

preserves token positions, attention masks, shape and dtype. The interfaces are not assumed to be anatomically equivalent; cross-family replication concerns the intervention question.

Both models use their recorded RoboTwin checkpoints, preprocessing and action normalization. The bands were chosen from earlier evidence and frozen before these states were evaluated; they are not presented as search-free discoveries. Appendix A specifies assets, numerical paths and hook timing. The transplant is made after cache construction and used throughout one action prediction, not later replanning.

3.3 Equal-norm difference-direction controls

For selected-layer recipient and donor tensors R and D, the true replacement produces perturbation $D - R$. Four fixed controls use

$$\widetilde{D}_k = \mathrm{cast}(R + S_k \odot (D - R)), \qquad k = 1, \ldots, 4, \tag{2}$$

where each S_k contains balanced ± 1 channel signs, broadcast over the remaining tensor axes and shared by K and V within a layer. The float32 difference is cast back to the original dtype. Consequently the mathematical norm is unchanged; we additionally require every stored K and V norm ratio to lie in $[0.98, 1.02]$. Patterns are fixed across states before confirmation. They disrupt difference direction while retaining its layerwise scale, but are neither on-manifold donor states nor an exhaustive distribution of all equal-norm edits.

Control hierarchy. The equal-norm controls are the primary control for the new confirmation: they use the same recipient input, cache band, execution state and endpoint metric as the true transplant, while changing the cache-difference direction at approximately fixed layerwise norm.

We keep a different, earlier control separate. In that study, a non-target orange object was moved in the rendered input while the task target remained fixed; its cache transplant was executed from the same center physical scene. This tests one specified non-target spatial intervention, not all low-level visual factors: its projection, edges, occlusion and pixel magnitude were not matched to the target manipulation. We therefore do not use it to claim a pure semantic position code or exhaustive target specificity.

3.4 Action budget and recipe controls

Normal FastWAM predicts 32 actions and executes 24 before replanning; normal $\pi_{0.5}$ predicts and executes 50. Each branch executes at most the first 24 actions of one prediction, retaining native success termination. This measurement does not redefine $\pi_{0.5}$'s normal replanning cadence. We save actual commands, object trajectories, contacts, final states, physics steps and simulated duration.

Every reachable state receives 20 branches: two unmodified executions for each of three inputs (six), two same-input selected-band replacements, two opposite-target true replacements, two opposite-target full-cache replacements, and four equal-norm controls per direction (eight). Model history and subsequent inference randomness are paired. Center continuation, repeats, same-input replacement and full-cache replacement must reproduce their prescribed predictions and physical executions. These are implementation checks, not independent scientific samples.

3.5 Directional effects and frozen decision rule

The endpoint is the world-x coordinate of the manipulated object's actor reference point, not the target, end effector or presumed center of mass. The two signed target effects are

$$E_{i,\rightarrow} = x_i(L \leftarrow R) - x_i(L), \qquad E_{i,\leftarrow} = x_i(R) - x_i(R \leftarrow L). \tag{3}$$

Both positive means opposite, donor-aligned biases relative to their respective recipient baselines. This does not require the object's absolute displacement from s_i to reverse sign.

For direction d and control k, let $x^{\mathrm{scr}}_{i,d,k}$ be the endpoint under Equation 2 and $x_i(r_d)$ its matching unmodified recipient endpoint. We define

$$C_{i,d} = \tfrac{1}{4}\sum_{k=1}^{4} |x^{\mathrm{scr}}_{i,d,k} - x_i(r_d)|, \qquad S_{i,d} = E_{i,d} - C_{i,d}. \tag{4}$$

Absolute control effects are taken before averaging, so opposing control directions cannot cancel. The primary vector is $(E_{\rightarrow}, E_{\leftarrow}, S_{\rightarrow}, S_{\leftarrow})$.

4 Frozen confirmation

4.1 States and eligibility

For each task, we evaluate 32 previously held-out environment instances using a fixed seed list shared across the two models. This gives 64 task–seed initial scenes and 128 normal model rollouts; when the stage is reachable, each model reaches its own candidate state. These seeds are held out from experimental development and selection, not claimed disjoint from all pretraining data.

The candidate is the first native replanning boundary satisfying: object reference-point elevation at least 30 mm above its initial height; contact by both fingers of the same hand; no table, ground or target contact; no success yet; and at least 24 actions remaining before the 400-action cap. Final success and donor effects are not selection criteria. An unreachable stage is reported, without a replacement seed or a later outcome-selected boundary.

All reachable states are executed. Primary eligibility requires the target to occupy at least 25 head-camera pixels in center, left and right inputs, and the manipulated object to be visible in at least one camera. These input-based checks are evaluated without behavioral outcomes. Ineligible reachable cases remain in descriptive plots and are never replaced.

Task	Model	Initial	Reachable	Eligible	Full criterion
Display stand	FastWAM	32	29	29	Yes
Display stand	$\pi_{0.5}$	32	27	27	Yes
Stapler	FastWAM	32	31	29	Yes
Stapler	$\pi_{0.5}$	32	30	29	Yes

Table 1: **Coverage and frozen decisions.** Counts are model rollouts/states, not independent cross-model environments. Every reachable state is executed, including three that fail input-only eligibility. There are 64 task–seed initial scenes, shared across the two models.

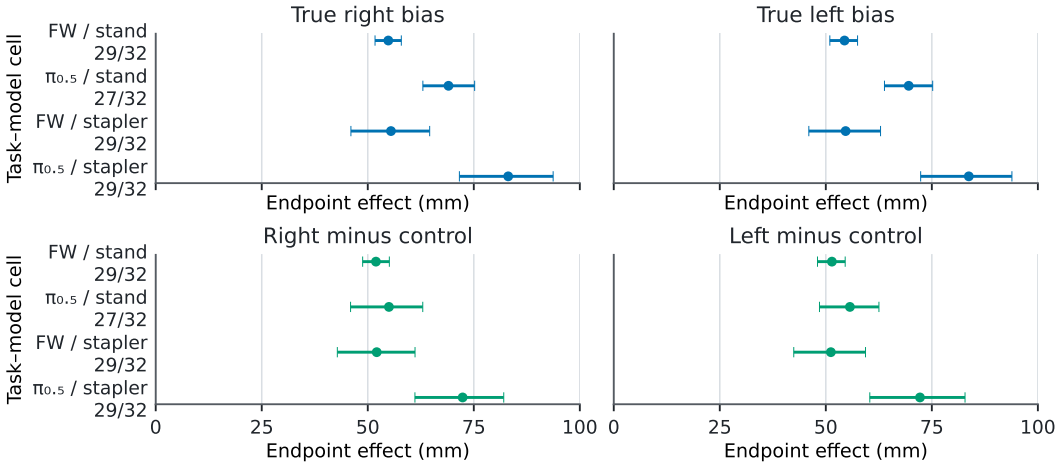

Figure 3: **Four combinations meet the frozen joint criterion.** Means and Bonferroni-adjusted 99.6875% percentile intervals. Blue denotes the signed true-replacement effects in the first two panels; green denotes the control-adjusted effects in the last two panels. The panel titles specify right versus left donor direction. All lower bounds exceed zero.

4.2 Statistics and validation

The resampling unit within each task–model cell is a state with its full four-effect vector; physical branches are not treated as independent samples. We use 20,000 state-paired bootstrap resamples with fixed group-specific seeds, where the resample count controls numerical approximation rather than independent information. Each of the 16 primary intervals is a two-sided 99.6875% percentile interval (Bonferroni correction). Full support requires $n \geq 16$ and all four lower bounds above zero; this is a prespecified support gate, not a claim that smaller cells have zero effect. We neither pool states across cells nor substitute another metric after observing results.

Validation connects frozen predictions to executed prefixes, endpoints to trajectories and final states, and implementation controls to their anchors. An independent script reads the saved records and recomputes every effect and interval without importing the primary analysis. This is code-independent offline reconstruction, not a separate policy/simulator rerun or independent human audit.

5 Results

5.1 All four cells pass the direction-over-magnitude test

Of 128 normal rollouts, 117 reach the fixed stage; all reachable states complete 20 recipes, totaling 2,340 physical branches and 56,103 executed high-level actions. Frozen visibility admits 114 states to the primary analyses. Table 1 reports every denominator. No state, task or seed is selected by intervention effect.

The true signed means range from 54.4 to 83.7 mm. After subtracting the per-state mean absolute equal-norm control effect, means remain 51.2–72.4 mm. The smallest of the sixteen corrected lower

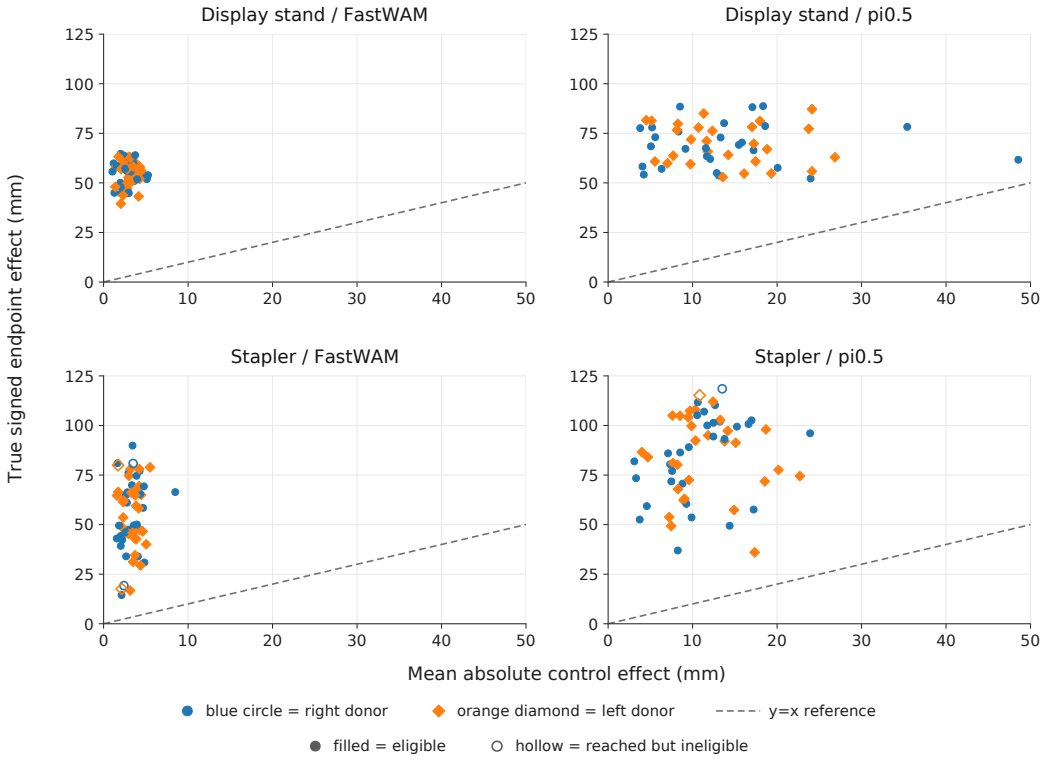

Figure 4: **True direction versus equal-norm controls for every reached state.** Each point is one donor direction; blue/orange denote right/left donors. The ordinate is the signed true-replacement endpoint bias and the abscissa is the mean absolute effect of four sign-transformed differences. Filled points are eligible and hollow points are the retained ineligible reachable states (cell counts: 29/27/31/30 reached, of which 29/27/29/29 are eligible in the displayed row order). The dashed line is the $y = x$ reference; every eligible point lies above it.

bounds is 42.44 mm. Full intervals are given in Appendix D. Thus support does not come from a single direction, task or policy family.

5.2 The result is statewise, not an average rescued by outliers

For every one of the 114 eligible states, both true signed effects are positive and both exceed their corresponding mean absolute controls. Figure 4 shows all reached states; hollow points are the three input-ineligible cases and do not enter primary decisions. The plot also exposes substantial control effects in some $\pi_{0.5}$ states rather than hiding them. This is a statewise diagnostic within the frozen sample, not a claim that every future state will have the same sign. The conclusion is conditional on the four frozen transformations, but is not produced by cancellation or a few favorable examples.

Because two positive relative effects could in principle arise if both crossed endpoints collapsed toward one common point, we also computed the retrospective crossed separation $W_i = x_i(L \leftarrow R) - x_i(R \leftarrow L)$. W_i is positive for all eligible states in every cell; its cell means/minima (mm) are 44.05/30.43 for FastWAM/stand, 66.44/51.38 for $\pi_{0.5}$/stand, 41.06/3.26 for FastWAM/stapler, and 81.17/35.11 for $\pi_{0.5}$/stapler. Thus the saved crossed endpoints retain their donor-direction order rather than exhibiting exact common-endpoint collapse. A stricter one-dimensional check that both crossed endpoints are closer to their respective donor baseline than to the center output passes for 26/29, 27/27, 11/29 and 29/29 states in the same order. These are post-hoc endpoint diagnostics, not additional gates: they strengthen the non-collapse interpretation but do not establish exact donor-trajectory reproduction, absolute direction reversal or natural mediation.

Model	Output axis	Mean O	Mean C	Mean M	Min M
FastWAM	x	30.40	4.62	21.45	10.84
FastWAM	y	27.62	5.13	17.59	16.06
$\pi_{0.5}$	x	39.43	9.72	22.40	5.44
$\pi_{0.5}$	y	36.90	5.61	25.74	7.59

Table 2: **Donor-axis selectivity in development states.** All values are mm. Each row has four states, eight signed O values (all positive), and four C, M values (all $M > 0$). The two axis rows reuse the same states. Models share four task–initial-condition pairs but reach their own states; these are not additional primary-confirmation samples.

5.3 Execution diagnostics and independent reconstruction

All eligible branches in the two FastWAM cells and in $\pi_{0.5}$/stapler execute 24 actions with sampled dual-finger contact throughout. The $\pi_{0.5}$/stand cell contains a few native-success short branches; restricting that cell to the 25 states whose twenty branches all execute 24 actions still passes all four corrected intervals. True-target branches in all four cells have no sampled action end with complete finger-contact loss. Contact is sampled at action boundaries rather than every physics substep, so this check does not prove continuous stable grasping. Height changes occur in baselines and both intervention types, and do not define the signed endpoint metric. The fixed 24-action budget is a short-horizon measurement window, not a common physical-time intervention.

The code-independent offline implementation audits 128 cases, 2,340 physical branches, 56,103 executed commands and 26,376 selected-layer tensor records. It finds all stored norm ratios valid and reproduces every per-state effect and all group statistics exactly. A standalone NumPy supplement reconstructs the four decisions from a 128-row endpoint table. These checks validate the saved evidence chain, not external provenance, simulator determinism on another stack, or a fresh policy/rendering rerun.

Separate spatial-control evidence. The earlier non-target spatial comparison answers a narrower, different question. Target-location effects exceeded the absolute effect of the specified non-target spatial intervention in three supported cells (FastWAM/stand $n = 36$, FastWAM/stapler $n = 38$, and $\pi_{0.5}$/stapler $n = 36$). The $\pi_{0.5}$/stand comparison has only 11 visibility-eligible states and is retained as positive but unconfirmed under the original support rule. Because this control is not appearance-matched, it bounds only the tested spatial nuisance; it is not combined with the equal-norm confirmation or presented as evidence against every low-level visual explanation.

5.4 Does the response distinguish the donor axis? Development evidence

Correctly signed motion alone need not distinguish target coordinates. In separate pilots, each model uses four retained development states (two per task), a center recipient, and target donors shifted ± 50 mm along world x or y. Execution again uses the center physical scene. These are not the opposite-target recipient contrasts or new states of the primary confirmation.

Let $\mathbf{r}^{\sigma}_{i,a}$ be the object's endpoint shift from the center baseline when the donor moves along axis $a \in \{x, y\}$ with sign $\sigma \in \{-1, +1\}$. With $b \neq a$ the other axis, define

$$\begin{aligned} O^{\sigma}_{i,a} &= \sigma\, \mathbf{e}_a^{\top} \mathbf{r}^{\sigma}_{i,a}, \qquad & C_{i,a} &= \max_{\sigma} \left| \mathbf{e}_a^{\top} \mathbf{r}^{\sigma}_{i,b} \right|, \\ M_{i,a} &= \min_{\sigma} O^{\sigma}_{i,a} - C_{i,a}. \end{aligned} \tag{5}$$

The *own-axis effect* O measures signed motion in the changed coordinate. The *cross maximum* C holds that *output coordinate fixed* and uses the largest absolute response to the *other-axis donors*. Thus a positive *selectivity margin* M means even the weaker own-axis direction exceeds either cross-axis donor response. It is not an average-own minus average-cross statistic.

Both models have positive margins for every state–axis pair (Table 2). An axis-blind response that produces the same sign-specific displacement for x and y donors could have positive own effects, but cannot have positive margins for both axes. The pilots therefore provide descriptive evidence of an axis-selective endpoint response, not just a generic sign-linked shift. They do not distinguish

scene-position information from axis-specific action preparation. Both axes concern the task target; this is neither a non-target nor an equal-norm control. Appendix E.2 gives a counterexample, exact definitions and common-22-action checks; Appendix E.3 gives the separate FastWAM protocol. Four states per model provide development evidence, not an independent confirmation of selectivity.

6 Interpretation and limitations

The experiment supports a narrow causal statement: at the tested interfaces, the direction of target-conditioned cache differences carries behavioral consequences beyond their layerwise K/V norm. Common-scene execution removes a direct physical-target explanation. Bidirectional predictions make an unsigned damage-only account insufficient for the observed directional pattern, while equal-norm controls bound a magnitude-only explanation for the four specified transformations. The separate non-target spatial comparison provides a limited nuisance check, not a matched low-level or semantic control.

Stronger claims do not follow. The equal-norm controls do not span every equal-norm direction, and their fixed sign patterns can be off the unmodified activation manifold. The earlier non-target spatial control does not match all image-space or feature-space changes, and its $\pi_{0.5}$/stand coverage is insufficient for a full four-cell claim. Moving a rendered target changes edges, geometry and occlusion, so we do not identify a pure symbolic coordinate or an abstract target-position representation. $\pi_{0.5}$ replaces selected layers of an entire multimodal prefix, not image tokens alone. Nor do we establish that the selected bands are minimal, naturally mediating, or microscopically equivalent across architectures.

All outcomes are simulated and conditional on two tasks, two recorded checkpoints, a model-visited manipulation stage and a maximum 24-action budget. Effects are relative recipient-baseline endpoint shifts, not guaranteed absolute direction reversals, full target switches, long-horizon success improvements or real-robot safety. The millimeter differences between models should not be ranked as representation quality because their states, cache interfaces and action trajectories differ. Earlier non-target spatial controls and a separate damage-recovery study remain in Appendix E; they are supporting but not merged into the equal-norm mechanism claim.

Conclusion. Target-conditioned cache differences can be causally transferred to produce predicted, donor-consistent relative object motion. The new confirmation holds in all four tested task–model cells and exceeds the four prespecified layerwise equal-norm direction controls; a separate, weaker spatial comparison supports three cells and remains incomplete for $\pi_{0.5}$/display stand. Small two-axis development pilots (four model-visited states per policy family) further show positive selectivity margins: the weaker own-axis response exceeds the largest cross-axis response in every tested state–axis pair, while remaining descriptive evidence rather than an independent confirmation. Together, these results show that the tested intervention can distinguish a factor-linked directional effect from the specified magnitude-only and non-target spatial alternatives; they do not establish a general steering algorithm, an abstract world representation, exhaustive visual specificity, or a complete account of natural policy computation.

AI use statement

Generative AI tools assisted with research planning, experiment and analysis code, literature search, statistical interpretation, manuscript drafting and revision, and figure-generation code.

Ethics statement

This study uses simulated manipulation and released robot-policy checkpoints; it involves no new human-subject or real-robot intervention study. Internal edits can induce unsafe behavior, and a signed endpoint effect is not a safety guarantee. We do not recommend deploying these interventions on physical robots without separate safety validation. We retain coverage failures and avoid interpreting source-consistent bias as task success or a semantic ground truth.

Reproducibility statement

Section 3 and Appendix A specify cache interfaces, inputs, execution and frozen inference settings. Section 4 and Appendix D provide the primary selection rules, all denominators, statistics, sensitivity results and execution-recovery provenance. The numerical supplement contains raw endpoint coordinates and a standalone NumPy implementation of the frozen primary analysis. Appendices B and C separately document the earlier spatial-control study. We distinguish offline numerical reproducibility from rerunning a renderer and policy on a new hardware/software stack.

References

Abhijith Babu, Ramneet Kaur, Nathaniel D. Bastian, Olivera Kotevska, Susmit Jha, Yanzhao Wu, Sumit Kumar Jha, and Anirban Roy. Closed-loop neural activation control in vision-language-action models, 2026. URL https://arxiv.org/abs/2606.00269v1. arXiv:2606.00269v1.

Kevin Black, Noah Brown, Danny Driess, et al. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164. arXiv:2410.24164.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, et al. Rt-1: Robotics transformer for real-world control at scale, 2022. URL https://arxiv.org/abs/2212.06817. arXiv:2212.06817.

Anthony Brohan, Noah Brown, Yevgen Chebotar, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. URL https://arxiv.org/abs/2307.15818. arXiv:2307.15818.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses. AI Alignment Forum, 2022. URL https://chanlawrence.me/publication/chan-2022-causal/.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation, 2025. URL https://arxiv.org/abs/2506.18088. arXiv:2506.18088.

Cheng Chi, Zhenjia Xu, Siyuan Feng, et al. Diffusion policy: Visuomotor policy learning via action diffusion, 2023. URL https://arxiv.org/abs/2303.04137. arXiv:2303.04137.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.14997.

Satyajeet Das, Darren Chiu, Shashank Hegde, and Gaurav S. Sukhatme. Steering multirobot behavior via closed-loop affine activation editing, 2026. URL https://arxiv.org/abs/2606.11489v1. arXiv:2606.11489v1.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021. URL https://arxiv.org/abs/2106.02997.

Atticus Geiger, Duligur Ibeling, Amir Zur, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability, 2023. URL https://arxiv.org/abs/2301.04709. arXiv:2301.04709.

Dibya Ghosh, Homer Walke, Karl Pertsch, et al. Octo: An open-source generalist robot policy, 2024. URL https://arxiv.org/abs/2405.12213. arXiv:2405.12213.

Bryce Grant, Xijia Zhao, and Peng Wang. Not all features are created equal: A mechanistic study of vision-language-action models, 2026. URL https://arxiv.org/abs/2603.19233v1. arXiv:2603.19233v1.

Bear Häon, Kaylene Caswell Stocking, Ian Chuang, and Claire Tomlin. Mechanistic interpretability for steering vision-language-action models. In *Conference on Robot Learning*, 2025. URL https://openreview.net/forum?id=YvsUD8C9QS.

Peter Hase and Mohit Bansal. Does localization inform editing? surprising differences in causal tracing and model editing, 2023. URL https://arxiv.org/abs/2301.04213. arXiv:2301.04213.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024. URL https://arxiv.org/abs/2404.15255. arXiv:2404.15255.

Jihoon Hong, Julian Skifstad, Qiyue Dai, Alice Chan, and Glen Chou. Steering robustness into world action models via mechanistic interpretability and optimal control, 2026. URL https://arxiv.org/abs/2607.14943v1. arXiv:2607.14943v1.

Xinchen Jin, Aditya Chatterjee, Pranav Kumar, and Rohan Paleja. Event-grounded sparse autoencoders for vision-language-action policies, 2026. URL https://arxiv.org/abs/2605.17204v1. arXiv:2605.17204v1.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, et al. Openvla: An open-source vision-language-action model, 2024. URL https://arxiv.org/abs/2406.09246. arXiv:2406.09246.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. Atp*: An efficient and scalable method for localizing llm behaviour to components, 2024. URL https://arxiv.org/abs/2403.00745. arXiv:2403.00745.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.03341.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2022. URL https://arxiv.org/abs/2202.05262. arXiv:2202.05262.

Marco Molinari, Leonardo Nevali, Saharsha Navani, and Omar G. Younis. Emergent world representations in OpenVLA, 2025. URL https://arxiv.org/abs/2509.24559. arXiv:2509.24559.

Open X-Embodiment Collaboration. Open x-embodiment: Robotic learning datasets and rt-x models, 2023. URL https://arxiv.org/abs/2310.08864. arXiv:2310.08864.

Physical Intelligence et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054. arXiv:2504.16054.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, et al. Causal mediation analysis for interpreting neural nlp: The case of gender bias, 2020. URL https://arxiv.org/abs/2004.12265. arXiv:2004.12265.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do world action models need test-time future imagination?, 2026. URL https://arxiv.org/abs/2603.16666. arXiv:2603.16666.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.16042.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems*, 2023. URL https://arxiv.org/abs/2304.13705.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

A SHARED IMPLEMENTATION AND EARLIER SPATIAL-CONTROL DETAILS

A.1 EVALUATED ASSETS AND NUMERICAL PATHS

We evaluate released models without new training. FastWAM uses the RoboTwin checkpoint `robotwin_uncond_3cam_384.pt` from the Fast-WAM release (Yuan et al., 2026). Its SHA-256 is

`776475b22566a791854ecf31cf3b50f25e7d8d94c343132ec16eb94994aa9e63`

The recorded action-statistics JSON has SHA-256

`7a02c46cfc8c5e746c0afbe41fca73f723eda34cbc083f8ca54f76d8f7468095`

These SHA-256 values are content fingerprints for provenance and byte-level identity; they are not encryption or an external authenticity claim. The normal and patched computations use the same statistics; a transplant does not change normalization.

The $\pi_{0.5}$ asset is the RoboTwin-finetuned `SidneyXie/pi05_robotwin`, Hugging Face revision `e49e2ab6c11f07511573b67261bd129e88d0a416`. Its 813 tensors are loaded strictly. The model weight SHA-256 is

`9a5381c3260fc58fdb4b90d17b93f9090bcbfbcd8e0287f9bf2acea2477a2bc3`

We use its native pre/post-processors and action statistics, not FastWAM normalization. This evaluated fine-tune is not equated with every checkpoint or the aggregate performance in the original $\pi_{0.5}$ paper (Physical Intelligence et al., 2025).

Setting	FastWAM	$\pi_{0.5}$
Native action prediction / execution	32 / 24	50 / 50
Content execution	At most first 24 actions	At most first 24 actions
Action dimension	14	14
Selected layers (zero-based)	10–19 of 30	12–17 of 18
Selected positions	Visual-conditioning cache	Complete multimodal prefix
Denoising evaluations	10	10
Initial action seed	20260917	20260927
Attention path	bfloat16 model, float64 action attention	Native mixed precision
Core runtime	PyTorch 2.8.0+cu128	PyTorch 2.11.0+cu128
Additional runtime	SAPIEN 3.0.0b1 simulator	LeRobot 0.6.1, Transformers 5.5.4

Table 3: Frozen evaluated implementations. The simulator executes both models' physical commands. A model's 14-dimensional command comprises joint/gripper controls, not an object-position target.

FastWAM constructs one 30-layer visual cache. At the selected layers, both complete K and V arrays are replaced; unselected layers and recipient metadata are retained. Shape, device, dtype and finiteness must match. The cached keys are not assigned new positions by the transplant. For every denoising evaluation, action attention casts Q, K and V to float64, computes scaled dot products, masked softmax and value aggregation in float64, then casts the result back to the input dtype. All 30 layers are checked on each of ten evaluations (300 calls). This is a measured high-precision implementation variant, not a claim of identical behavior to every unmodified low-precision kernel.

$\pi_{0.5}$ captures the prefix after its single native prefill. A hybrid cache takes donor K and V at six layers and recipient K and V elsewhere. It preserves the recipient attention mask, position IDs and cache metadata. During the ten action-denoising evaluations, the hybrid prefix is read by the action computation. The prefix contains image tokens and the language/state sequence. The native language processor incorporates discretized robot state; the actual task tokens, token masks and normalized state are checked equal across inputs. This does not make the corresponding contextualized K/V entries image-independent. The experiment does not isolate individual heads, token classes or semantic dimensions.

The same three RGB camera streams—head, left wrist and right wrist—and the same 14-dimensional robot observation are passed through each model's own preprocessing. We transport saved uint8 camera pixels without image recompression between the simulator and $\pi_{0.5}$ process. All recipes use the random state following verified normal prediction history. Resetting to the same initial seed without reconstructing that history would not be an equivalent pairing procedure.

A.2 Earlier spatial-control tasks, seeds and phase selection

We use RoboTwin's `demo_clean` configuration, with the fixed non-colliding sphere added at initialization. Its configuration SHA-256 is

`3acfcd508078a51e3a48fd7e2f6e61c1b9a6ced0cc44dedd149b55114248e2ab`

The display-stand task manipulates a task object onto a stand; stapler-to-pad manipulates a stapler toward the pad. Assets within a task are not counted as independent tasks. We use environment seeds 9151000–9151039 for each task and instruction seed 20260916 with the environment's unseen-instruction setting. Neither "unseen instruction" nor new evaluation seeds claims that all object assets were held out from model training.

Phase selection is evaluated at the model's native replanning boundaries, not at a fixed action index shared across models. Its rise criterion is the change in the object's actor reference point from its initial normal-rollout height. Both fingers must contact the same object with the same hand; table, ground and target contacts disqualify the boundary. A boundary must precede native success and leave at least 24 actions before the 400-action cap. We select the first qualifying boundary. This is a reproducible stage rule, not proof of stable grasp for the entire future window. Any downstream loss of contact remains part of the measured physical outcome.

A.3 Earlier spatial-control recipe inventory and algorithm

IDs	Input / replacement	Executions
00–01	Center, unmodified	2
02–03	Target-left, unmodified	2
04–05	Target-right, unmodified	2
06–07	Sphere-left, unmodified	2
08–09	Sphere-right, unmodified	2
10–13	Same-input patch, four non-center inputs	4
14	Target-left recipient ← target-right donor	1
15	Target-right recipient ← target-left donor	1
16	Sphere-left recipient ← sphere-right donor	1
17	Sphere-right recipient ← sphere-left donor	1

Table 4: Every reachable state receives all 18 recipes, including states excluded by the geometric rule. Baseline repeat 0 is the paired reference; repeat 1 is a deterministic control.

For each task, model and frozen initial seed:

1. Run the native policy, retaining normal actions, observations at native boundaries and physical states. Find the first qualifying stage. Record unreachable cases without replacement.
2. Construct center, target-left/right and sphere-left/right inputs at that state. Check unchanged non-image input fields and geometry. Keep all reachable cases for execution.
3. Reconstruct the native prediction history; check it against the normal predictions. Save the subsequent random state U. Obtain each input's unmodified cache and action chunk using U.
4. For each patch recipe, restore U, compute the recipient prefix, replace only its designated cache layers, and finish the action prediction. Check repeated and same-input predictions against their anchors.
5. Independently rebuild the center physical scene and replay the saved normal action prefix for each recipe. Require the branch state to match the saved normal state. Execute its

first 24 actions or stop on native success; save executed commands, contacts, trajectory, duration and final state.

6. Check physical repeat/same-input controls, normal center continuation, prediction–execution prefix equality and final-state endpoint identity. Compute the paired effects; apply the input-based eligibility rule to the primary analysis, retaining all other directional outcomes descriptively.

B EARLIER SPATIAL-CONTROL NUMERICAL RESULTS AND SENSITIVITY

All following primary intervals use the frozen 98.75% percentile procedure. The four combinations use separate within-combination state resampling; states from different models are not treated as matched intermediate configurations. No task-averaged or cross-model difference is a primary hypothesis.

Task	Model	Right effect (mm)	Left effect (mm)
Stand	FastWAM	50.43 [47.91, 53.09]	50.82 [48.32, 53.51]
Stand	$\pi_{0.5}$	64.83 [43.36, 76.62]	76.62 [69.91, 83.83]
Stapler	FastWAM	61.50 [56.50, 66.00]	61.77 [56.96, 66.06]
Stapler	$\pi_{0.5}$	84.36 [77.48, 91.28]	84.32 [77.96, 90.59]

Table 5: Bidirectional paired endpoint effects on the geometrically eligible sample. Left is sign-normalized relative to its own right-input baseline.

Task	Model	Target minus sphere (mm)	Sphere absolute mean (mm)
Stand	FastWAM	50.13 [47.66, 52.75]	0.49
Stand	$\pi_{0.5}$	60.08 [18.75, 78.97]	10.64
Stapler	FastWAM	61.14 [56.22, 65.49]	0.50
Stapler	$\pi_{0.5}$	83.66 [77.08, 90.19]	0.68

Table 6: The specified spatial comparison, not a test excluding all low-level visual factors. The $\pi_{0.5}$/stand row remains below the sample-size gate despite positive intervals.

The primary eligible counts are 36, 11, 38 and 36 in table order. Geometry failures among reachable states are: none for FastWAM/stand; 25 sphere-head visibility failures for $\pi_{0.5}$/stand; one sphere-head visibility failure for FastWAM/stapler; and three sphere-head plus one target-head visibility failures for $\pi_{0.5}$/stapler. These are failures of the predefined input/control geometry, not predictions about the sign of the ensuing effect.

The raw RGB L1 magnitude ratio between target and sphere perturbations is not matched: the eligible-cell median ratios are 7.20, 6.69, 1.97 and 3.38 in the same table order, with within-cell maxima from 11.11 to 20.73. These values use concatenated saved RGB streams and are neither perceptual nor cache-space distances. We report them to make the limitation explicit; no ratio-based filtering or reweighting is used.

Task	Model	Branches	Actions	Short branches	Full-24 eligible states
Stand	FastWAM	648	15,552	0	36
Stand	$\pi_{0.5}$	648	15,163	133	9
Stapler	FastWAM	702	16,848	0	38
Stapler	$\pi_{0.5}$	720	17,280	0	36

Table 7: Execution denominators. All 133 short branches end at native success. "Full-24" requires every recipe for a state to execute 24 actions; it is a sensitivity description, not the primary eligibility rule.

For the nine full-24 eligible $\pi_{0.5}$/stand states, the right effect is 63.38 [37.17, 77.67] mm, the left effect 76.21 [68.32, 84.97] mm, and S is 56.82 [6.35, 79.88] mm. This subset still cannot meet

$n \geq 16$. All three fully supported combinations execute all 24 actions, so their full-24 and primary estimates coincide. The sensitivity subset can itself be selected by a post-intervention outcome; it is not promoted to a new primary estimand.

Task	Model	Simulated seconds (min / median / max)	Physics steps (min–max)
Stand	FastWAM	4.71 / 6.46 / 7.57	1178–1893
Stand	$\pi_{0.5}$	2.86 / 5.42 / 8.99	715–2247
Stapler	FastWAM	6.47 / 9.78 / 16.24	1618–4060
Stapler	$\pi_{0.5}$	6.46 / 9.46 / 17.14	1615–4284

Table 8: Actual execution time varies even under a fixed maximum high-level action budget. These are simulator durations, not wall-clock inference or compute costs.

The independent implementation also reports directional intervals using all reachable states, solely to make exclusions inspectable. For $\pi_{0.5}$/stand ($n = 36$), right/left are 61.10 [53.12, 67.83] / 67.35 [61.36, 72.92] mm. For FastWAM/stapler ($n = 39$), they are 61.47 [56.79, 65.83] / 61.76 [57.15, 65.83] mm; for $\pi_{0.5}$/stapler ($n = 40$), 82.96 [76.11, 89.61] / 82.70 [76.34, 88.84] mm. FastWAM/stand has no geometry exclusions. These estimates do not use invisible-sphere cases to establish specificity or replace the frozen primary set.

C Earlier spatial-control provenance and validation scope

The scientific manifest was frozen before the confirmation runs with identifier

`6497495aabd7b3c1dcb14049f11d516141bf63a471e20dda2b28d98fd4984e8e`

It binds sample definitions, original implementation, recipes and statistics. It is an internally recorded prospective design, not an externally registered trial. Subsequent resource continuations and implementation amendments have separate records; they do not retroactively change this identifier or erase failed attempts.

Two $\pi_{0.5}$/stapler cases exposed a reconstruction issue. Replaying only their actions reproduced physical fields but not exactly the saved image. A controlled comparison of sparse and native observation histories gave 12/12 image mismatches for sparse replay and 12/12 exact full-state matches for native-history replay. Subsequent execution matched the normal state but failed after hiding and showing the sphere for an occlusion diagnostic. For these two cases only (153 and 154), an explicit amendment reconstructs each input independently and checks an independent repeat, never reusing hidden-sphere rendering for a later policy input. Both cases complete all 18 branches with the original actions, seeds and exact checks; case 154 remains geometrically ineligible. The other 158 cases retain the original implementation. This isolates an observed rendering side effect; it does not claim to identify or repair the renderer's internal root cause.

All historical execution failures and interrupted inputs were retained in recoverable, hashed archives. Final source/execution validation covers 160 cases, 17 recorded attempts and 11 archive receipts. The final independent saved-data audit checks 2,718 branches, 64,843 executed commands, 30,702 saved images and 80 shared-initial-state pairs. It checks that predictions supply actual command prefixes, object endpoints equal the last saved trajectory and final physical state, and paired controls reproduce. Numerical estimates and intervals agree with the primary implementation within absolute tolerance 10^{-12} with zero relative tolerance.

This validates stored-data consistency, not external authenticity or a new inference/rendering run. A migration diagnostic on another accelerator stack did not reproduce images exactly and was excluded from the scientific data. Exact rendering portability is not established. The anonymous numerical supplement exports all 160 case identities, reachability and geometry flags, and eight raw endpoint coordinates for each reachable state. Its standalone NumPy script recomputes the primary effects, intervals and decisions without importing the original analysis module. It does not validate saved trajectories, geometry masks or model computations.

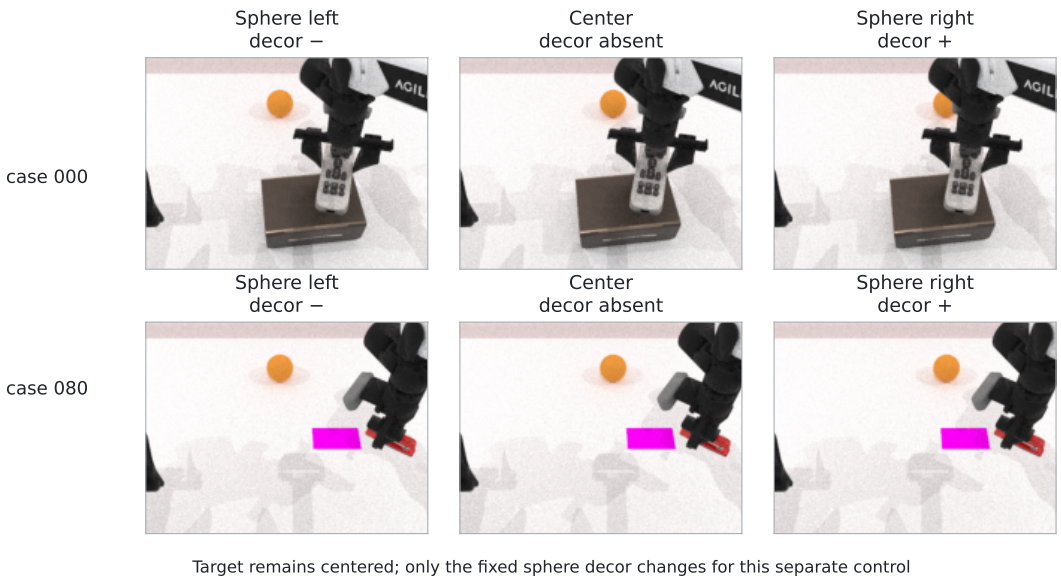

Figure 5: Original saved head-camera pixels from the earlier spatial-control confirmation, showing the sphere contrast. The target is centered in all three columns. Visibility is necessary for a meaningful control but does not imply equivalence of target and sphere visual perturbation magnitudes.

C.1 Retrospective endpoint diagnostics for the spatial control

The earlier spatial control can be challenged by the possibility that both signed effects merely pull the two recipient endpoints toward a common center. We therefore retain a fixed, post hoc diagnostic without changing the primary metric or eligibility rule. Let L and R be the normal target-left and target-right endpoints, and let P and Q be the two crossed target-cache endpoints. With world-x coordinates, $D = R_x - L_x$ is the normal separation and $W = P_x - Q_x$ is the crossed separation; the signed effects satisfy $W = E_{\rightarrow} + E_{\leftarrow} - D$. A pure collapse to one common endpoint would give $W = 0$. The observed W is positive for every eligible state in all four spatial-control cells, with mean/minimum values (mm) of 39.218/31.752 for FastWAM/stand $(36/36)$, 68.097/62.422 for $\pi_{0.5}$/stand $(11/11)$, 45.666/22.389 for FastWAM/stapler $(38/38)$, and 81.455/48.850 for $\pi_{0.5}$/stapler $(36/36)$. This rejects the simple common-collapse explanation, but it does not establish absolute direction reversal or an exact donor trajectory.

For transparency, a stricter endpoint-distance diagnostic asks whether each crossed endpoint is closer to its donor than to the center output. Both directions satisfy this one-dimensional test for 24/36, 10/11, 21/38 and 36/36 states in the same order. The non-uniform counts are retained rather than converted into a new success gate; they are why the paper uses the narrower phrase "donor-consistent relative endpoint bias." These diagnostics use saved terminal endpoints, not equal-time trajectories, and are descriptive checks performed after the frozen analysis.

D Equal-norm cache-difference confirmation

The earlier spatial comparison does not match the magnitude of the induced cache change. The primary experiment tests a narrower competing explanation: whether the physical effect depends on the direction of a target-conditioned cache difference beyond its layerwise norm. For recipient and donor tensors R and D, the true replacement is compared with

$$Z_{\text{scrambled}} = \text{cast}\big(R + S \odot (D - R)\big), \tag{6}$$

where S contains balanced channel signs and the difference is calculated in float32 before casting to the original dtype. Half the final-axis channels receive -1, with the same layer-specific pattern broadcast over the remaining tensor axes and shared by K and V. The four fixed seeds are 2026091901–2026091904; a layer uses its seed plus 1009 times its zero-based index. The patterns

are fixed across states, so the analysis is conditional on these four transformations, not a population of randomly sampled controls. The actual stored K and V perturbation norms must each be within 2% of the corresponding true difference. Recipient input, selected layers, inference noise and execution physics are paired. This control tests the specified sign transformations, not every possible equal-norm perturbation, an on-manifold counterfactual, or an abstract position code.

Development evidence. Eight previously used model states, drawn from four environmental initial states across the two tasks and models, complete all 160 recipes and 3,840 actions. In all eight states, both true signed endpoint effects exceed the mean absolute effect of the four corresponding scrambled controls; the per-state directional margins range from 31.271 to 82.830 mm. An offline implementation independently reconstructs 1,792 selected-layer tensor checks and 56 nontrivial repeat, same-input and full-cache physical pairs. These cases use a cold reset of inference randomness at the branch and are development evidence, not independent confirmation. Scrambled branches can change behavior substantially, including a recorded 102.55 mm height decrease in one $\pi_{0.5}$/stand branch; none is excluded for this reason.

Frozen independent extension. For each task, the extension evaluates 32 previously held-out environment seeds, `9191000-9191031`. The same seed list is shared across the two models, yielding 64 task–seed initial scenes and 128 normal model trajectories. It preserves the native 32/24 and 50/50 prediction/execution cadences, first eligible phase rule, layer bands and maximum 24-action branch budget. Unlike the development test, it reconstructs the actual native inference history and its subsequent random state, requiring the center branch to reproduce normal continuation. The fixed, non-colliding orange decor follows the intervening development configuration: radius 0.065 m, position $[-0.22, -0.12, 0.806]$ m for display stand and $[-0.22, -0.22, 0.806]$ m for stapler, with unchanged RGB $[0.9, 0.35, 0.05]$. These differ from the earlier sphere study's scene configuration and are present throughout each new normal trajectory. Only target visibility and manipulated-object visibility enter this experiment's input eligibility. This is a distinct control question; removing sphere visibility from its eligibility does not repair the earlier sphere-control experiment.

Each reachable state receives 20 recipes: two unmodified executions of each of the three inputs (six), two same-input replacements, two opposite-target selected-band replacements, two full-cache replacements, and four scrambled controls per direction (eight). A full-cache replacement must reproduce its donor's unmodified prediction and physical execution; it is an implementation control, not independent mechanistic evidence. Each comparison uses the matching recipient baseline. For direction d and scrambled endpoint $x^{\mathrm{scr}}_{i,d,k}$, define

$$S^{\mathrm{cache}}_{i,d} = E_{i,d} - \frac{1}{4}\sum_{k=1}^{4} \left| x^{\mathrm{scr}}_{i,d,k} - x_i(r_d) \right|. \tag{7}$$

The primary vector is $(E_{\rightarrow}, E_{\leftarrow}, S^{\mathrm{cache}}_{\rightarrow}, S^{\mathrm{cache}}_{\leftarrow})$. Unlike the earlier sphere comparison, the control and true transplant share the same recipient input, and control magnitude is measured inside the manipulated cache. Absolute effects are taken before averaging, so opposing control directions cannot cancel. The fixed analysis uses 20,000 state-paired bootstrap resamples, two-sided 99.6875% percentile intervals adjusted across 16 primary intervals, and at least 16 eligible states per task–model cell; all four lower bounds must be positive for full support. Every predefined seed is retained in the accounting, including unreachable states, and native-success short branches remain in the endpoint analysis. All four fixed waves finished before primary group statistics and support decisions were computed. Wave-level technical audits reconstructed individual paired effects; three case effects were incidentally displayed during a recovery check. This did not change the frozen samples, interventions, eligibility or decision rules.

D.1 Complete confirmation results

The 128 normal model trajectories contain 117 reachable stages. Primary geometry admits 114 states: 29/29 reached for FastWAM/stand, 27/27 for $\pi_{0.5}$/stand, 29/31 for FastWAM/stapler and 29/30 for $\pi_{0.5}$/stapler. All reachable states receive all twenty recipes, yielding 2,340 physical branches and 56,103 executed commands. No unreachable or ineligible state is replaced.

Across all 117 reachable states, the 15,072 sign-control K/V tensor records have actual stored perturbation norm ratios in [0.999888,1.000143]. Their cosine similarities to the corresponding true

donor difference range from -0.04091 to 0.05096 for FastWAM (median -0.000077), and from -0.24725 to 0.22892 for $\pi_{0.5}$ (median 0.002169). These are intervention diagnostics, not additional independent samples or decision criteria: the applied controls closely match magnitude while changing direction.

Task	Model	$E_{\rightarrow}$ (mm)	$E_{\leftarrow}$ (mm)	$S_{\rightarrow}^{\text{cache}}$ (mm)	$S_{\leftarrow}^{\text{cache}}$ (mm)
Stand	FastWAM	54.84 [51.71, 57.91]	54.41 [50.95, 57.49]	51.93 [48.79, 55.09]	51.44 [48.04, 54.57]
Stand	$\pi_{0.5}$	69.03 [62.99, 75.20]	69.55 [63.83, 75.20]	54.99 [45.94, 62.97]	55.69 [48.51, 62.53]
Stapler	FastWAM	55.47 [46.03, 64.60]	54.67 [46.00, 62.91]	52.11 [42.82, 61.14]	51.19 [42.44, 59.36]
Stapler	$\pi_{0.5}$	83.11 [71.60, 93.71]	83.71 [72.36, 93.88]	72.37 [61.12, 82.06]	72.20 [60.37, 82.83]

Table 9: Primary means and Bonferroni-adjusted 99.6875% percentile intervals. Each row passes the frozen $n \geq 16$ and four-positive-lower-bounds criterion.

Every eligible state has all four paired effects positive: 29/29, 27/27, 29/29 and 29/29 by table row. The corresponding Wilson 95% lower bounds on the statewise joint-positive fractions are 0.883, 0.875, 0.883 and 0.883. This diagnostic was not an additional gate, but shows that the positive means are not produced by a minority of large states. All-reachable descriptive analyses also remain positive; the three input-ineligible states are displayed rather than suppressed.

Full-window and contact sensitivity. Both FastWAM cells and $\pi_{0.5}$/stapler have 24 actions in every eligible branch. For $\pi_{0.5}$/stand, requiring all twenty branches in a state to execute 24 actions leaves 25 states. Its full-window vector is 70.03/70.16/55.54/55.78 mm, with corrected intervals [63.86,76.31], [64.06,76.24], [46.19,63.93] and [47.95,63.32]; it still passes. No true-target branch in any cell has a sampled action end with complete finger-contact loss. Contact is sampled at action boundaries, not at every physics substep, so this is evidence against pervasive terminal contact loss rather than proof of continuous stable grasping. The equal-norm controls are not benign by construction: eleven eligible $\pi_{0.5}$/stand control branches contain sampled contact loss and their mean paired duration difference is 2.09 s, yet the true effects still exceed their absolute endpoint effects. We retain these outcomes rather than filtering controls for smooth behavior.

D.2 Independent audit and execution provenance

The scientific manifest SHA-256 is

`e250485a5c33f5a3c3d4efb0c2bc7f10e96f4a9b297e038386e5ccf7ca61bc69`

It fixes cases, recipes, layer bands, sign seeds, eligibility, metrics and resampling before the four-wave analysis. The run retains platform and worker failures. One wave completed all 600 branches but its final worker was terminated by an inner 3300-second watchdog before the root marker; an offline gate reconstructed only that marker after checking all saved cases. The final wave reached the global 3600-second limit after 554/560 branches. A separately versioned recovery executed only the six missing physical branches from frozen predictions and exact prefix reconstruction, with zero new model inference; the original failed status remains recorded. These recoveries change neither samples nor scientific definitions.

The final independent implementation audits all 128 case identities, 2,340 branch records, 56,103 executed commands and 26,376 selected-layer tensor records. It checks recipient/donor pairing, endpoint identity, actual execution prefixes, implementation controls and every stored norm ratio. Per-state effects and all bootstrap statistics match the primary implementation exactly. This is an offline consistency audit rather than new inference, rendering, external authenticity certification or independent human review.

The anonymous supplement contains one row per normal model trajectory, with reachability, eligibility, recipe endpoints and executed-step counts. A standalone NumPy script reconstructs the four primary decisions and verifies the reference values to absolute tolerance 10^{-12}. It does not rerun geometry, policy inference or simulation.

E EARLIER EVIDENCE, KEPT SEPARATE

E.1 COMMON-STATE DISPLAY-STAND STUDY

Before the frozen expansion, both families were evaluated on 20 common branch states reached by FastWAM at action 72. These states span five manipulated-object assets within one display-stand task, not five independent tasks or 40 independent states. All 1,280 execution branches across the two models execute 24 actions. The FastWAM 10–19 directional means are 18.626/18.702 mm; the target-minus-absolute-lighting effect is 16.400 mm, with 95% interval [12.720, 19.918]. The $\pi_{0.5}$ 12–17 means are 47.245/47.307 mm; its target-minus-lighting effect is 44.852 mm with 98.3333% interval [30.275, 59.451], adjusted for three bands in that study. These estimates are not pooled with the new sample or relabeled with the expansion's intervals.

The earlier study controls intermediate scene differences across models; the expansion tests model-visited states. Its larger effects are not a method improvement: states and controls differ. Earlier band exploration motivates the frozen bands; no further band search uses the final sample.

E.2 $\pi_{0.5}$ ORTHOGONAL TARGET-AXIS DEVELOPMENT PILOT

After the primary bands and states had been fixed, we ran a bounded descriptive pilot on four previously retained $\pi_{0.5}$ development states (two display-stand and two stapler-to-pad cases). At each state, the task target was rendered at ± 0.05 m along world x and y, while execution restored the same center physical scene. The selected $\pi_{0.5}$ cache layers were 12–17, the native prediction chunk was 50 actions, and at most the first 24 actions were executed. Each state received five baselines, four selected-band donor replacements, and five full-cache donor controls.

Shared estimand for both models. For center input 0 and donor a_σ, let $\mathbf{p}_i(0 \leftarrow a_\sigma)$ be the selected-band branch endpoint and $\mathbf{p}_i(0)$ the unmodified center-input endpoint, both in the common center physical scene. Then $\mathbf{r}^{\sigma}_{i,a} = \mathbf{p}_i(0 \leftarrow a_\sigma) - \mathbf{p}_i(0)$ in Equation 5. Thus donor–recipient displacement is 50 mm here, rather than the primary confirmation's 100 mm opposite-target contrast. Neither donor-baseline endpoints nor full-cache controls enter O, C, M; the full-cache branches check implementation. Each model's four states give 16 signed own effects and eight state–axis cross maxima/margins, not 16 margins. Table 2 averages O over eight state–sign pairs per axis and C, M over four states per axis. Margins are computed before averaging.

For output x, $C_{i,x}$ is the maximum absolute x shift under the two y donors; for output y, $C_{i,y}$ is the maximum absolute y shift under the two x donors. It is not the off-axis component of the same donor branch. Absolute values precede the maximum, so opposite cross signs cannot cancel. For a hypothetical axis-blind response $\mathbf{r}^{\pm}_{i,x} = \mathbf{r}^{\pm}_{i,y} = \pm(20, 30)$ mm, all four own effects are positive, yet $C_{i,x} = 20$, $C_{i,y} = 30$ and both margins are zero. This illustrates a distinction that directional signs alone cannot make; it is not an additional robot experiment.

The two pilots share four task–initial-condition pairs (two seeds per task) but use eight model-visited states. Axis rows and cross-model results therefore do not create new independent environmental samples. Both axes move the task target; their raw-pixel differences and cache norms are not matched. A positive margin is a descriptive coordinate-response comparison, not a significance gate or an identification of a semantic coordinate code.

All 56 physical branches passed the saved-state, prediction-prefix and endpoint checks. The 20 full-cache controls exactly reproduced their corresponding donor baseline actions and object trajectories. The 16 signed own-axis effects (two directions for each axis and state) were all positive, with mean 38.17 mm and range 17.38–68.10 mm. There are eight case-axis cross maxima (four states times two axes), with mean 7.66 mm, and eight corresponding descriptive selectivity margins; all were positive (mean 24.07 mm; minimum 5.44 mm). Three case-2 branches ended at 22 or 23 actions because the native simulator reported success; they were retained using the actual executed prefix rather than treated as failed truncations. An earlier running note misreported the eight margins as 16; this denominator correction does not change any endpoint or effect value. As a retrospective horizon check, truncating every saved trajectory to the common 22-action prefix without interpolation kept all 16 effects and all eight margins positive (PI0.5 own/margin means 37.58/23.00 mm); this does not replace the 24-action estimand. The pilot's manifest is

`9b99a8434ba7b0560c148ffb3b8ba722a2e79aee007c133e09e59308d543fab0` and its report is archived with the supplementary research records.

This pilot is development evidence, not an independent confirmation sample or a significance gate. Both x and y are task-target factors, so the orthogonal axis is not a non-target specificity control; image-space magnitudes were not matched. We include it to test whether the donor-consistent endpoint sign survives a second target axis and the native $\pi_{0.5}$ interface, not to claim an abstract position code, natural-manifold intervention or exhaustive visual specificity.

E.3 FastWAM orthogonal target-axis development pilot

The same bounded development protocol was run separately on four previously retained FastWAM states (cases 0, 1, 4 and 5), after the primary bands and states had been fixed. The task target was rendered at ± 0.05 m along world x and y, execution restored the common center scene, the selected cache band was FastWAM layers 10–19, and one native 24-action prediction window was measured. The 56 policy/physical branches passed their saved-state and endpoint checks; 20 full-cache controls reproduced their donor baselines exactly. The 16 own-axis effects had mean 29.01 mm, the eight case-axis cross maxima had mean 4.87 mm, and the eight margins had mean 19.52 mm; all own effects and margins were positive. A retrospective common 22-action prefix check kept the own/margin means at 28.52/19.84 mm (minimum margin 10.68 mm). This is descriptive evidence from previously retained development states, not an independent confirmation or a non-target control. Its manifest is

`476746c865244ef07f8af2bc2f95a3d48fd06cd66c274886e7d118ee11393628`

and the anonymous branch exports and independent recomputation are included in the supplement.

E.4 Functional recovery under a different intervention

A separate study contains 1,351 full closed-loop records across 193 initial states and ten tasks. Under a fixed K/V-mismatch damage background, restoring layers 10–14 exceeds the average of four prespecified equal-budget recovery controls by 7.41 percentage points in a task-equal analysis, with 95% interval [2.60, 12.35]. Relative to the damaged branch, 33 paired initial states are rescued and eight newly fail. This is an average recovery advantage, not superiority to every control or on every task.

Recovery of five layers under sustained damage and transplantation of ten layers during one undamaged-policy prediction do not establish the same circuit. The earlier content states belong to the recovery study's initial-state set: their union is 193, not 233. Functional importance and content-transfer effects remain separate claims.