FROM PHYSICAL CHANGE TO ACTION: CAUSAL ATTRIBUTION IN WORLD ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

World Action Models (WAMs) couple future prediction with robot action generation, yet how prediction causally contributes to action and whether this contribution survives reduced predictive computation remain unclear. We use matched physical counterfactuals to define action responses and representation interventions to trace their transmission across four frozen WAM configurations. Model-specific interfaces strongly transmit these responses, although Fast-WAM-IDM shows that strong transmission need not reconstruct the complete counterfactual action. In Fast-WAM-Joint, fixing the action decoder's future inputs sharply reduces the effect of replacing representations associated with the current observation, revealing substantial propagation through downstream future representations. Under blocking, restoring future inputs in a predefined late window of decoder layers recovers substantially more of the response than an equally sized early window. Targeted RoboTwin experiments likewise find larger propagation-allowed effects than future-fixed current effects, with the additional propagated effect opposing the latter in signed mean. Reducing Fast-WAM-Joint's world-branch computation lowers cumulative policy-computation time by a median of 23.7% while all evaluated LIBERO continuations still complete. However, matched action responses change, with little held-out improvement from a shared scalar rescaling. Together, these findings show that prediction's causal contribution to action can change with the inference procedure even when task completion is preserved.

1 INTRODUCTION

World Action Models (WAMs) couple predictions of future scene evolution with robot action generation (Shen et al., 2026). This coupling differs across architectures. Some models decode actions from generated futures, while others use intermediate predictive representations or jointly model future states and actions. Explicit multi-frame generation adds inference cost, motivating designs that reduce or bypass future rollout while retaining strong task performance (Yuan et al., 2026; Zhang et al., 2026c). Similar success, however, need not imply the same use of prediction. End-task performance alone cannot distinguish a policy that consumes generated futures from one that benefits from predictive training without reading an explicit future at inference (Zhang et al., 2026a). The causal role of future prediction in action generation, and how it varies across architectures and inference procedures, therefore remains unclear.

Clarifying this role requires addressing two related questions. The first is which internal representations transmit the effect of a physical change to action across WAM configurations and how predictive information contributes to this process. The second is whether this contribution is preserved when predictive computation is reduced in models that generate future information at inference time. These questions are linked because reducing computation may alter the representations through which prediction influences action. Together, they motivate assessing lower-cost inference in terms of both task performance and the preservation of prediction's causal contribution to action.

The main challenge is that world-modeling capability, causal use of prediction, and task performance need not align. WAM surveys identify joint evaluation of prediction and action as an open problem (Shen et al., 2026; Wang et al., 2026b). Across architectures, representation form, information flow, and generation procedure often vary together, making performance differences difficult to attribute. Internal interventions offer more direct evidence, but their interpretation depends on the

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.

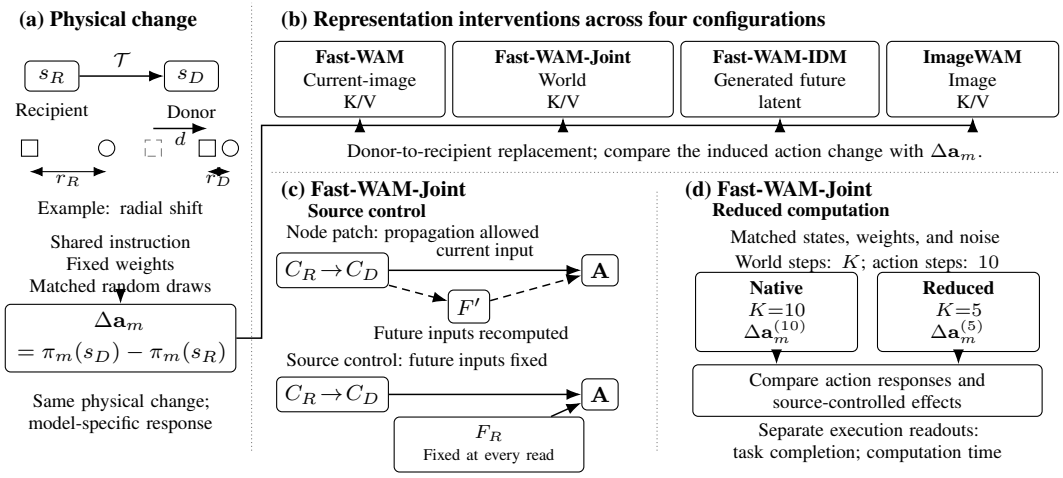

Figure 1: **Physical responses, source control, and reduced computation.** (a) Physical counterfactuals define action responses $\Delta\mathbf{a}_m$ (square: robot; circle: object). (b) Representation replacements test response transmission through four model-specific interfaces; K/V denotes attention keys and values. (c) In Fast-WAM-Joint, recipient current inputs C_R are replaced with donor inputs C_D, with future inputs either recomputed (F') or fixed to the recipient reference (F_R). $\mathbf{A}$ denotes the action chunk. (d) Reducing world steps from ten to five, with ten action steps retained, tests response preservation alongside task completion and computation time.

intervention boundary and the downstream computation left active (Geiger et al., 2021; Zhang & Nanda, 2024). In particular, changing a current-state representation may affect action directly or first alter a predicted future that the action decoder consumes. A strong upstream effect therefore does not establish equally strong direct dependence. A common external reference is needed to compare these effects while distinguishing their origin from their downstream use.

We address these questions using controlled physical counterfactuals as a common reference for action responses in four frozen WAM configurations. Under matched inference conditions, representation interventions trace how these responses are transmitted through model-specific interfaces. In Fast-WAM-Joint, we independently control current- and future-origin inputs at the action decoder, allowing current-side effects with future inputs fixed to be compared with effects that include downstream propagation. Selective restoration further probes the contribution of future inputs under blocking. We then repeat the matched response and source-controlled comparisons under reduced world-branch computation, assessing response preservation alongside task completion and computation time. Experiments on LIBERO and targeted RoboTwin analyses distinguish the origin of an action effect from its downstream transmission. They further show that, in Fast-WAM-Joint, the contribution of predictive inputs to action can change under reduced computation despite preserved completion of the evaluated LIBERO continuations.

Our contributions are threefold.

- **A common physical reference.** Matched simulator counterfactuals enable causal comparison across four WAM configurations without aligning latent spaces.
- **Source-conditioned attribution.** In Fast-WAM-Joint, consumer-side controls distinguish current-node influence from downstream predictive propagation, with location-dependent recovery and targeted RoboTwin validation.
- **Response-level evaluation of efficient inference.** Matched physical-response tests distinguish preserved recipient completion from preserved action responses when world-branch computation is reduced.

2 Related Work

Prediction–action coupling and efficient inference. WAMs couple prediction and action through generated observations, joint video–action modeling, or compact predictive representations (Du et al., 2023; Zhu et al., 2025; Shen et al., 2026). These designs differ in whether and how actions use explicitly generated futures. Fast-WAM retains video co-training without test-time future generation, while ImageWAM replaces video rollout with image-editing representations (Yuan et al., 2026; Zhang et al., 2026c). Efficient-WAM reduces video-side denoising, and GlanceWAM uses sparse, asynchronous lookahead (Li et al., 2026a; Wang et al., 2026a). RIFT intervenes on future-cache values, spatiotemporal organization, and evolution to examine their roles in execution, then learns rollout-free cache generation while retaining the future-read interface (Zhang et al., 2026a). These studies show that changing how predictive information is produced and supplied to action can preserve task performance. Whether reduced predictive computation also preserves prediction's causal contribution to the policy's response to physical changes remains unclear.

Causal use of representations in embodied policies. Behavioral and representational diagnostics characterize policy sensitivity and relate internal features to execution (Zhang et al., 2026d; Mai et al., 2026). Intervention-based studies examine whether modality-specific and historical information causally affects action (Grant et al., 2026; Shi et al., 2026; Liao & Cao, 2026). In WAMs, related analyses probe spatial grounding, object binding, and the use of activation structure for feedback steering (Qiu et al., 2026; Liu et al., 2026; Hong et al., 2026). Counterfactual memory audits further test whether histories that warrant different decisions guide action at a fixed present (Zhang et al., 2026b). BadWAM demonstrates action drift while penalizing changes to predicted futures (Li et al., 2026b), motivating closer examination of how prediction influences action. For WAMs, a more specific question concerns how the response to a controlled physical change is transmitted through current representations and downstream future inputs. We use matched physical counterfactuals to define this response and source control in Fast-WAM-Joint to compare current-representation effects with future inputs recomputed or held fixed.

Causal interventions and pathway attribution. Interchange interventions and causal tracing test whether representations affect model outputs, complementing probes of what they encode (Geiger et al., 2021; Meng et al., 2022; Belinkov, 2022). Mediation analysis formalizes direct and indirect effects, while path patching and circuit discovery identify computational routes (Pearl, 2001; Vig et al., 2020; Wang et al., 2023; Goldowsky-Dill et al., 2023; Conmy et al., 2023). Studies of activation patching and downstream self-repair show that attribution depends on intervention design and the computation left active (Zhang & Nanda, 2024; McGrath et al., 2023). This distinction matters for WAMs because an intervention on a current representation may also alter the future inputs consumed by the action decoder. Strong upstream influence therefore need not imply equally strong direct dependence at the decoder.

3 Method

We use a controlled physical change to define the action effect traced by each intervention.

3.1 Physical Counterfactuals and Action Responses

Let m index a frozen WAM configuration and i a recipient–donor comparison with recipient simulator state $s_{R,i}$. Subscripts R and D denote recipient and donor conditions. Given simulator state s, instruction ℓ_i, and inference randomness $\xi_{m,i}$, policy $\pi_m(s; \ell_i, \xi_{m,i})$ returns a de-normalized action chunk in $\mathbb{R}^{H_m \times d_m}$, where H_m is its prediction horizon and d_m its number of action channels. The state supplies both visual observations and robot proprioception. For physical factor f, manipulation phase p, and signed perturbation magnitude d, transformation $\mathcal{T}_{f,p,d}$ defines

$$s_{D,i} = \mathcal{T}_{f,p,d}(s_{R,i}), \qquad \Delta\mathbf{a}_{m,i} = \pi_m(s_{D,i}; \ell_i, \xi_{m,i}) - \pi_m(s_{R,i}; \ell_i, \xi_{m,i}). \tag{1}$$

The instruction, sampling schedule, and random draws are matched within each model. Observations and proprioception are regenerated after the physical transformation. Thus, each model is evaluated on the same physical comparison using its own action response $\Delta\mathbf{a}_{m,i}$.

Physical factors and phases. We consider three task phases, PREGRASP (approaching the object before grasping), TRANSPORT (carrying the grasped object), and PREPLACE (preparing for placement). We use five physical factor families, denoted F1–F5. F1 changes the robot's radial position relative to the object during PREGRASP. F2 moves the object alone or rigidly translates the object and goal together. F3 changes object–goal radial progress during TRANSPORT and PREPLACE. F4 and F5 apply tangential displacement and object rotation, respectively. The goal-only operator F3-G moves the target basket while keeping the robot and grasped object fixed. The signed magnitude d denotes rotation for F5 and translation otherwise. Appendix A gives the evaluated settings.

Action readouts. Full-action scores use all entries of the de-normalized command tensor. On LIBERO, a fixed linear map $\mathcal{P}_i$ selects the action channels and projects them onto the coordinate associated with the physical factor. The F1 profiles use gripper-closing commands; radial analyses project translation commands and retain all predicted chunk positions. On RoboTwin, each condition's joint targets are first mapped through forward kinematics to end-effector target positions, then projected onto a radial direction fixed from the recipient state. Differences are computed after this mapping. These readouts measure predicted commands and targets.

3.2 REPRESENTATION INTERCHANGE AND RESPONSE TRANSMISSION

Following interchange interventions and activation patching (Geiger et al., 2021; Meng et al., 2022; Zhang & Nanda, 2024), we replace recipient activations with donor activations at a fixed set of internal locations $\mathcal{L}$, then rerun downstream computation with recipient external inputs and matched inference conditions. Figure 1 and Table 1 identify the primary interfaces. Fast-WAM-Joint world K/V contain the attention keys and values associated with current and generated future inputs. Supplementary interface measurements and Fast-WAM-IDM reconstruction references are given in Appendix B.

Let $\mathbf{a}^{R}_{m,i}$ and $\mathbf{a}^{D}_{m,i}$ denote the native recipient and donor action chunks, and $\mathbf{a}^{R\leftarrow D}_{m,i,\mathcal{L}}$ the recipient output after replacement at $\mathcal{L}$. Donor-aligned transfer and relative reconstruction error are

$$
\begin{aligned}
T_{m,i}(\mathcal{L}) &= \frac{\langle \mathbf{a}^{R\leftarrow D}_{m,i,\mathcal{L}} - \mathbf{a}^{R}_{m,i}, \Delta\mathbf{a}_{m,i}\rangle_F}{\|\Delta\mathbf{a}_{m,i}\|_F^2 + \varepsilon}, \\
E_{m,i}(\mathcal{L}) &= \frac{\|\mathbf{a}^{R\leftarrow D}_{m,i,\mathcal{L}} - \mathbf{a}^{D}_{m,i}\|_F}{\|\Delta\mathbf{a}_{m,i}\|_F + \varepsilon}.
\end{aligned}
\tag{2}
$$

Here $\varepsilon > 0$ stabilizes the denominators. The Frobenius inner product $\langle\cdot,\cdot\rangle_F$ and norm $\|\cdot\|_F$ operate over the full action tensor. T measures the component of the intervened action change along $\Delta\mathbf{a}_{m,i}$; E measures the residual to the complete donor action relative to the size of that physical response. An orthogonal change affects E but not T, so directional transmission and full reconstruction are distinct outcomes. T is a projection coefficient with no restriction to $[0, 1]$. Identity replacements test replay, structure-destroying replacements test content sensitivity, and changed-value replacements at alternative locations test location sensitivity. Control constructions are reported with the corresponding model results.

3.3 CONDITIONAL SOURCE EFFECTS IN FAST-WAM-JOINT

In Fast-WAM-Joint, replacing a current-state node can affect action through the current inputs and through the future inputs subsequently computed from them. We compare this propagation-allowed node replacement with a 2×2 source intervention at the action decoder (Pearl, 2001; Vig et al., 2020; Goldowsky-Dill et al., 2023). Let $e = (h, k)$ identify a consumer event, where h is the layer index and k the action-denoising step. The event set $\mathcal{I}$ covers all 30 layers and ten steps, both indexed from zero. Let C_e and F_e denote the current- and future-origin K/V at event e, and let g map the complete input sequences to an action chunk. Suppressing m, i, define

$$
\mathbf{A}_{ab} = g\Big(\{C_e^{(a)}, F_e^{(b)}\}_{e\in\mathcal{I}}; \mathbf{z}_R, \xi\Big), \qquad a, b \in \{0, 1\}, \tag{3}
$$

Here a selects the current source and b the future source, with 0 denoting recipient values and 1 donor values. Thus, $C_R = \{C_e^{(0)}\}_{e\in\mathcal{I}}$ and $C_D = \{C_e^{(1)}\}_{e\in\mathcal{I}}$ are the complete current-input sequences; F_R, F_D are defined analogously. The context $\mathbf{z}_R$ contains the remaining recipient inputs, including

instruction and proprioception, and ξ is the matched inference randomness. Assigning both sources at every event keeps the selected inputs fixed throughout action decoding. With $\boldsymbol{\delta}_{ab} = \mathbf{A}_{ab} - \mathbf{A}_{00}$,

$$\begin{aligned} \mathbf{U}_{F|C_R} &= \boldsymbol{\delta}_{01}, & \mathbf{U}_{F|C_D} &= \boldsymbol{\delta}_{11} - \boldsymbol{\delta}_{10}, \\ \mathbf{U}_{C|F_R} &= \boldsymbol{\delta}_{10}, & \mathbf{U}_{C|F_D} &= \boldsymbol{\delta}_{11} - \boldsymbol{\delta}_{01}, \\ \mathbf{J} &= \boldsymbol{\delta}_{11} - \boldsymbol{\delta}_{10} - \boldsymbol{\delta}_{01}. \end{aligned} \tag{4}$$

Each $\mathbf{U}$ measures the effect of replacing one source with the other fixed; $\mathbf{J}$ measures its dependence on that background. Thus, $\mathbf{U}_{C|F_R}$ is a controlled direct effect with future inputs fixed to F_R (Pearl, 2001). Cell $\mathbf{A}_{10}$ combines donor current and recipient future inputs, which need not co-occur during native inference. Matched-cell checks localize the difference to this cell but do not exclude decoder sensitivity to the resulting input combination. Geometric contrasts apply the readouts in Section 3.1 before differencing.

Let F' denote the future-input sequence recomputed after current-node replacement. In contrast, F_D is captured during native donor inference. Under the same current-input assignment and recipient context, the propagation-allowed node effect $\mathbf{N}$ satisfies

$$\begin{aligned} \mathbf{N} &= g(C_D, F'; \mathbf{z}_R, \xi) - g(C_R, F_R; \mathbf{z}_R, \xi), \\ \mathbf{N} - \mathbf{U}_{C|F_R} &= g(C_D, F'; \mathbf{z}_R, \xi) - g(C_D, F_R; \mathbf{z}_R, \xi). \end{aligned} \tag{5}$$

The difference $\mathbf{N} - \mathbf{U}_{C|F_R}$ isolates the additional action effect obtained by admitting F' instead of F_R while retaining donor current and recipient context. This is the matched contrast evaluated on RoboTwin in Section 4.6.

Localized restoration. Let P denote the future K/V captured from propagation-allowed current-node replacement, corresponding to F'. The blocked condition B retains donor current inputs and assigns recipient future inputs at every event. Condition P_W restores P only at layers in window W, across all ten action steps, while P_{all} restores the same P at every layer. The primary comparison uses a late window $W = \{25, \ldots, 29\}$ and an equally sized early window $V = \{0, \ldots, 4\}$. For a blocking or restoration condition X, let $\mathbf{v}_{i,X} \in \mathbb{R}^{32}$ be its predicted radial command vector. The root-mean-square (RMS) recovery error is $e_{i,X} = \|\mathbf{v}_{i,X} - \mathbf{v}_{i,P_{\text{all}}}\|_2/\sqrt{32}$, where $\|\cdot\|_2$ is the Euclidean norm. The common reference is full restoration P_{all}. Appendix C.2 includes the subsequent six-window scan.

3.4 Responses under Reduced World-Branch Computation

Let K denote the number of executed world-branch denoising steps. The primary comparison reduces Fast-WAM-Joint from its native $K = 10$ to $K = 5$, retaining all ten action-denoising steps (Li et al., 2026a). A matched extension adds $K = 8$ on the same cohort. Remaining action steps reuse that call's last complete cache; omitted world steps are not executed, and subsequent calls rebuild their caches. Replacing π_m by $\pi_m^{(K)}$ in Equation 1 defines $\Delta\mathbf{a}_{m,i}^{(K)}$. We match weights, physical inputs, instructions, and corresponding random draws across budgets. Each budget supplies its own generated inputs to the source-controlled comparisons. Response differences, recipient completion, and cumulative policy-computation time are measured separately.

Shared-gain comparison. We test whether scalar rescaling explains response changes, fitting each benchmark, budget, and response set separately. Let $\mathbf{r}_{t,j}^{(K)} \in \mathbb{R}^{32}$ denote the radial response for source t and dose–contrast combination $j = 1, \ldots, n_t$. Its coordinates are predicted chunk positions, not independent samples. LIBERO's primary fits pool signed doses and $\boldsymbol{\delta}_{10}$, $\boldsymbol{\delta}_{01}$, and $\boldsymbol{\delta}_{11}$ for each $K \in \{5, 8\}$; natural responses are fitted separately. On RoboTwin, natural responses and $\mathbf{U}_{F|C_D}$ are fitted separately at $K = 5$, with one $+1$ cm donor per source ($n_t = 1$).

We use five-fold source-level cross-fitting. Let $\mathcal{S}$ be the source set and $\mathcal{S}_{-q}$ exclude fold q. The zero-intercept gain is

$$\widehat{\lambda}_{-q}^{(K)} = \arg\min_{\lambda \in \mathbb{R}} \sum_{t \in \mathcal{S}_{-q}} \frac{w_{t,-q}}{n_t} \sum_{j=1}^{n_t} \left\| \mathbf{r}_{t,j}^{(K)} - \lambda \mathbf{r}_{t,j}^{(10)} \right\|_2^2. \tag{6}$$

LIBERO weights sources equally: $w_{t,-q} = 1/|\mathcal{S}_{-q}|$ and $w_t = 1/|\mathcal{S}|$. RoboTwin weights tasks equally and sources equally within task: $w_{t,-q} = 1/(5N_{\tau(t),-q})$ and $w_t = 1/(5N_{\tau(t)})$, where $\tau(t)$ identifies the task and $N_{\tau,-q}$ and N_τ count its training-subset and full-set sources.

With $q(t)$ denoting source t's held-out fold, the squared-discrepancy reduction is

$$Q_K = 1 - \frac{\sum_t \frac{w_t}{n_t} \sum_{j=1}^{n_t} \left\| \mathbf{r}_{t,j}^{(K)} - \widehat{\lambda}_{-q(t)}^{(K)} \mathbf{r}_{t,j}^{(10)} \right\|_2^2}{\sum_t \frac{w_t}{n_t} \sum_{j=1}^{n_t} \left\| \mathbf{r}_{t,j}^{(K)} - \mathbf{r}_{t,j}^{(10)} \right\|_2^2}. \tag{7}$$

Positive Q_K indicates improvement over the unscaled native response. RMSE divides the corresponding weighted squared discrepancy by 32 before taking the square root. Bootstrap replicates retain fold assignments and refit gains. Appendices D and E provide implementation details.

4 Experiments

4.1 Experimental Setup

We evaluate Fast-WAM, Fast-WAM-Joint, Fast-WAM-IDM (Yuan et al., 2026), and Image-WAM (Zhang et al., 2026c) on LIBERO-Object (Liu et al., 2023), with all model weights frozen. The first three belong to the Fast-WAM family and are labeled Direct, Joint, and IDM in figures, respectively. The four-configuration LIBERO analysis uses 20 shared held-out base states, with intervention choices fixed during discovery. Restoration and reduced-computation evaluation use 20 and 50 source trajectories, respectively. On RoboTwin (Mu et al., 2025; Chen et al., 2025), Fast-WAM and Fast-WAM-Joint are evaluated for interface–context effects, with additional source-control and reduced-computation analyses in Fast-WAM-Joint. The RoboTwin shared-gain analyses use 34 registered sources across five tasks; a separate closed-loop comparison uses 29 preflight-eligible task–seed pairs.

Confidence intervals use 10,000 bootstrap replicates, with repeated observations grouped by source. Shared-gain analyses use five-fold source-level cross-fitting. Appendices A and E detail cohort construction, task coverage, sampling, and weighting. Code is available online.[1]

4.2 Physical Response Profiles across Models

The discovery profiles indicate systematic responses to controlled physical changes in all four WAMs (Figure 2). During PreGrasp, robot-side radial displacement produces a signed, dose-dependent gripper-closing response, with asymmetry at larger displacements. For F3, PrePlace radial-response slopes range from 0.39 to 0.94 per cm, compared with 0.02–0.17 during Transport. On the 20 held-out states, PrePlace slopes remain positive in all models, whereas Transport responses are weak, with every interval including zero (Appendix A). These phase-specific effects provide the response targets for the following interventions.

4.3 Model-Specific Interfaces Transmit Physical Responses

Across 544 held-out cases per model, mean donor-aligned transfer ranges from 0.946 to 0.999, compared with 0.006 or below for the model-specific controls (Table 1). The mean relative reconstruction error is 0.306 for Fast-WAM-IDM, compared with 0.026 for Fast-WAM and 0.024 for Fast-WAM-Joint; ImageWAM's reported error rounds to zero. Strong transfer along the physical-response direction therefore need not recover the complete donor action. A prospective evaluation with fixed interface assignments similarly yields mean transfer of 0.955–1.000 across 374 geometrically valid counterfactuals (Appendix B.1). In Fast-WAM-Joint, node replacement permits downstream future inputs to be recomputed, so its near-unit transfer does not isolate the current-input effect with future inputs held fixed.

[1] https://github.com/HY2027/wam-causal-audit

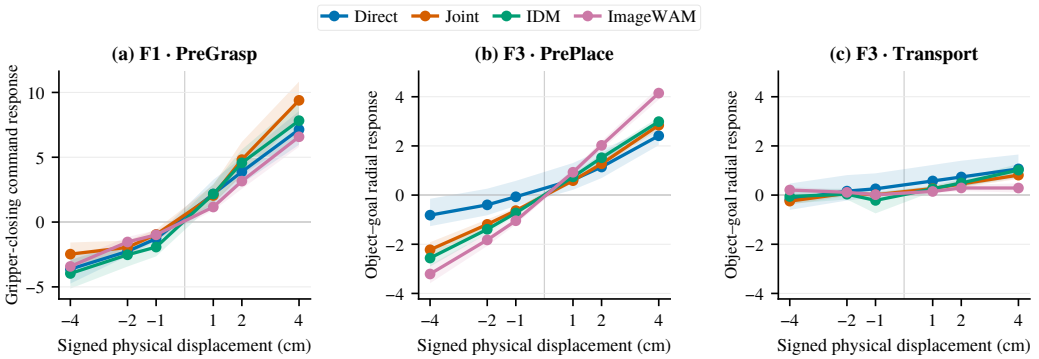

Figure 2: **Physical responses vary with displacement and phase.** (a) Robot-side perturbations during PREGRASP. (b,c) Object–goal perturbations during PREPLACE and TRANSPORT. Shading denotes 95% confidence intervals.

Table 1: **Held-out interface transmission, 544 cases per model.** Values are means [95% CI]. Fast-WAM-Joint's primary measurement permits downstream propagation. Controls are model-specific (Appendix B.2).

Model	Registered interface	Primary T	Control T	Residual E
Fast-WAM	Current-image K/V	.998 [.997, .999]	.006 [.003, .012]	.026 [.022, .031]
Fast-WAM-Joint	World K/V	.999 [.998, 1.000]	.000 [.000, .000]	.024 [.018, .030]
Fast-WAM-IDM	Generated future latent	.946 [.914, .971]	.000 [.000, .000]	.306 [.266, .352]
ImageWAM	Image K/V	.998 [.996, .999]	.000 [.000, .001]	.000 [.000, .000]

4.4 SOURCE ATTRIBUTION AND LOCAL RESTORATION IN FAST-WAM-JOINT

Fixing future inputs sharply reduces the current-side effect in Fast-WAM-Joint (Figure 3a,b). Across the tested F1 doses, the future-fixed current effect is 0.045–0.074 of the joint radial effect, compared with 0.927–0.954 for the current-fixed future effect. Across 42 paired cases, $\mathbf{A}_{00}$, $\mathbf{A}_{01}$, and $\mathbf{A}_{11}$ are elementwise identical between node replacement and source control; only the current-only output, $\mathbf{A}_{10}$, changes. The strong current-node effect therefore depends on downstream future inputs. Target-basket displacement (F3-G) yields the same ordering, with a donor-conditioned future-minus-current magnitude gap of 0.0587 in the radial-command readout on the validation split. Current–future interactions are reported in Appendix C.1.

Late restoration recovers substantially more of the blocked response than the equally sized early window (Figure 3c). Across 103 F1 donors from 20 source trajectories, mean radial RMS error is 0.248 under blocking, 0.247 after early restoration, and 0.038 after late restoration. The paired early-minus-late error difference is 0.210 (95% CI $[0.184, 0.238]$). Both windows were fixed before cohort evaluation and restore the same captured future inputs under identical blocking. A subsequent six-window scan on the same cohort yields errors of 0.235–0.247 for the other five windows covering layers 0–24, placing the large recovery in layers 25–29 at five-layer resolution (Appendix C.2).

4.5 EFFECTS OF REDUCED WORLD-BRANCH COMPUTATION

Reducing world-branch steps from $K = 10$ to $K = 5$ yields a median paired reduction of 23.7% in cumulative policy-computation time, while all 50 evaluated LIBERO recipient continuations complete (Figure 4a). Although completion does not distinguish the budgets in this cohort, matched physical responses change across 200 F3-G counterfactuals, evaluated using each budget's own generated caches. The mean norm of the conditional future effect $\mathbf{U}_{F|C_D}$ changes by -10.87% (95% CI $[-18.39\%, -0.66\%]$). In the illustrated $+1$ cm response, later action-chunk positions approach or cross zero at $K = 5$, while native responses remain positive (Figure 4b).

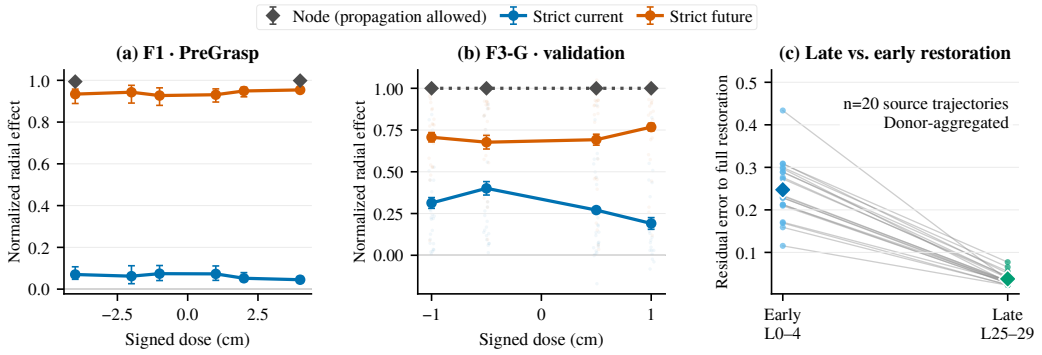

Figure 3: **Source conditioning and localized recovery in Fast-WAM-Joint.** (a,b) Fixing future inputs reduces the current-side effect for F1 and F3-G. (c) Late restoration outperforms the size-matched early control. Error bars in (a,b) are 95% CIs; lines in (c) pair donor-aggregated sources. Node effects in (a) are shown at ± 4 cm. The six-window scan appears in Appendix C.2.

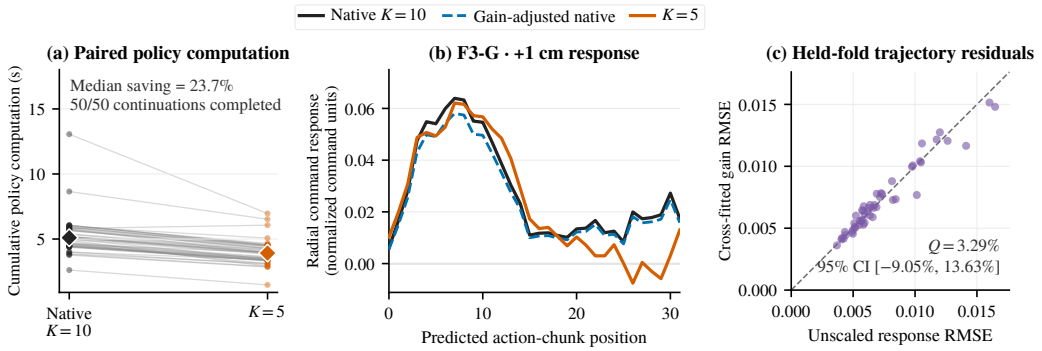

Figure 4: **Lower world-branch cost coexists with changed control responses.** (a) All 50 recipient continuations complete. (b) Matched F3-G responses differ at $K = 5$. (c) Cross-fitted gain adjustment explains little of the discrepancy.

A shared scalar gain explains little of the change in the factorial response vectors: RMSE decreases from 0.0079 to 0.0078, removing only 3.29% of held-out squared discrepancy (95% CI $[-9.05\%, 13.63\%]$; Figure 4c). Thus, the response differences are not well explained by uniform rescaling. A subsequent $K = 8$ extension on the same cohort yields a natural-response RMSE of 0.00473 relative to native, compared with 0.00950 at $K = 5$, again with little held-out improvement from a shared gain. Extension statistics and static-input timing are reported in Appendix D.

4.6 Targeted Cross-Benchmark Analysis on RoboTwin

On 34 Fast-WAM-Joint sources, the propagation-allowed current-node effect $\mathbf{N}$ exceeds the future-fixed current effect $\mathbf{U}_{C|F_R}$ in mean radial RMS magnitude (7.71 versus 2.67 mm; Figure 5a). Their matched contrast in Equation 5 has signed mean $+3.30$ mm (95% CI $[0.29, 6.93]$ mm), opposing the future-fixed current effect's mean of -1.73 mm. Decoder context also counteracts the interface effect: replacing the interface and then the command-state input yields signed increments of $+7.6$ and -8.7 mm in Fast-WAM, and $+1.6$ and -6.4 mm in Fast-WAM-Joint (Figure 5b). Replacing both inputs reconstructs the independently evaluated donor action.

Reducing world-branch steps from $K = 10$ to $K = 5$ changes natural-response vectors, with RMSE 2.98 mm (95% CI $[2.33, 3.60]$ mm). A cross-fitted scalar gain removes only 7.9% of held-out squared discrepancy (95% CI $[-18.5\%, 23.7\%]$; Figure 5c), consistent with the limited improvement observed on LIBERO. The conditional future effect $\mathbf{U}_{F|C_D}$ has a mean-norm change of only -0.2% (95% CI

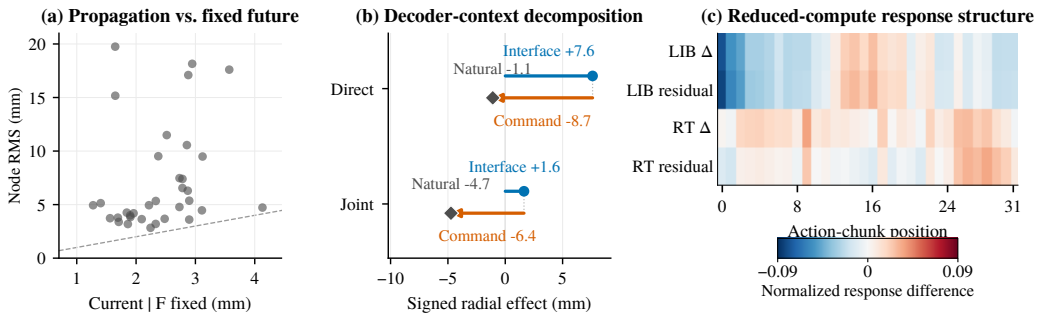

Figure 5: **Propagation, context, and computation shape prediction–action coupling.** (a) RoboTwin node effects exceed future-fixed current effects for most sources; the dashed line denotes equality. (b) Interface and command-state effects oppose in signed mean. (c) Natural-response differences (Δ, $K = 5$ minus $K = 10$) and gain-adjusted residuals on LIBERO (LIB) and RoboTwin (RT), shown on normalized display scales.

$[-8.3\%, 6.1\%]$), yet its vector RMSE is 3.22 mm. Thus, a small change in the mean response norm does not imply preservation of the response vector.

A separate closed-loop evaluation on 29 task–seed pairs passing a shared preflight before either budget runs yields a median paired reduction of 47.6% in cumulative policy-computation time at $K = 5$ (95% CI $[42.1\%, 50.3\%]$). Native and $K = 5$ complete 18 and 19 episodes, respectively, with a task-equal paired completion difference ($K = 5$ minus native) of -0.027 (95% CI $[-0.267, 0.160]$). Appendix E.2 details eligibility, weighting, and timing.

5 Discussion

The results connect the origin of a physical change to the representations through which it influences action. In Fast-WAM-Joint, a current-state node can strongly affect action while much of its influence reaches the decoder through future inputs. On RoboTwin, the future-fixed current effect and the additional propagated effect oppose in signed mean, as do the interface and command-state effects. Natural responses can therefore conceal stronger conditional responses. The final-window recovery further identifies where the captured future inputs can restore the blocked response.

Task completion and matched response preservation assess different properties of efficient inference. A continuation can succeed while the policy responds differently to the same physical change, and a nearly unchanged response norm can conceal a changed vector. Matched physical responses make these changes measurable. They complement task outcomes rather than require an efficient policy to reproduce identical actions; a response difference alone is not evidence of worse control.

Our scope is limited to three Fast-WAM variants and ImageWAM, with detailed source conditioning in Fast-WAM-Joint. The donor-current/recipient-future cells specify controlled intervention effects rather than natural mediation. Extending source conditioning to other architectures and relating response changes to control quality are natural next steps.

6 Conclusion

Matched physical counterfactuals provide a common reference for tracing action effects across WAMs. Source controls in Fast-WAM-Joint distinguish propagated current-node influence from conditional source use, while localized restoration identifies a strong late-over-early recovery difference. Reduced world-branch computation changes matched response vectors despite successful LIBERO recipient continuations. Together with targeted RoboTwin validation, these results show why evaluating prediction's role requires tracing its effect on action, not task success alone.

AI USE STATEMENT

Generative AI tools assisted with translation, language polishing, and reference formatting. The authors take full responsibility for the final manuscript.

REFERENCES

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

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. *arXiv preprint arXiv:2506.18088*, 2025. URL https://arxiv.org/abs/2506.18088.

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*, volume 36, pp. 16318–16352, 2023. doi: 10.52202/075280-0719.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 9156–9172, 2023. URL https://arxiv.org/abs/2302.00111.

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.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023. URL https://arxiv.org/abs/2304.05969.

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

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

Jiajun Li, Tiecheng Guo, Yifan Ye, Rongyu Zhang, Xiaowei Chi, Qianpu Sun, Ying Li, Yunfan Lou, Yan Huang, Zhihe Lu, Meng Guo, and Shanghang Zhang. Efficient-WAM: A 1B-parameter world-action model with low-cost future imagination. *arXiv preprint arXiv:2606.10040*, 2026a. URL https://arxiv.org/abs/2606.10040.

Qi Li, Xingyi Yang, and Xinchao Wang. BadWAM: When world-action models dream right but act wrong. *arXiv preprint arXiv:2607.15207*, 2026b. URL https://arxiv.org/abs/2607.15207.

Chih-Ting Liao and Xin Cao. Present but not remembered: Auditing how frozen VLAs encode, deploy, and steer visual history. *arXiv preprint arXiv:2607.03372*, 2026. URL https://arxiv.org/abs/2607.03372.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44776–44791, 2023. URL https://arxiv.org/abs/2306.03310.

Yushan Liu, Peibo Sun, Shoujie Li, Yifan Xie, Lingfeng Zhang, Xintao Chao, Shiyuan Dong, Fang Chen, Xiao-Ping Zhang, and Wenbo Ding. OA-WAM: Object-addressable world action model for robust robot manipulation. *arXiv preprint arXiv:2605.06481*, 2026. URL https://arxiv.org/abs/2605.06481.

Hung Mai, Bin Zhu, and Tuan Do. Beyond task success: Behavioral and representational diagnostics for WAM and VLA. *arXiv preprint arXiv:2606.01095*, 2026. URL https://arxiv.org/abs/2606.01095.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023. URL https://arxiv.org/abs/2307.15771.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. URL https://arxiv.org/abs/2202.05262.

Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, Lunkai Lin, Zhiqiang Xie, Mingyu Ding, and Ping Luo. RoboTwin: Dual-arm robot benchmark with generative digital twins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27649–27660, 2025. URL https://arxiv.org/abs/2504.13059.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, 2001.

Lu Qiu, Yizhuo Li, Yi Chen, Yuying Ge, Yixiao Ge, and Xihui Liu. Making foresight actionable: Repurposing representation alignment in world action models. *arXiv preprint arXiv:2606.12217*, 2026. URL https://arxiv.org/abs/2606.12217.

Qiuhong Shen, Shihua Zhang, Yue Liao, Qi Li, Zhenxiong Tan, Shizun Wang, Shuicheng Yan, and Xinchao Wang. World action models: A survey. *arXiv preprint arXiv:2606.20781*, 2026. URL https://arxiv.org/abs/2606.20781.

Haoyuan Shi, Xiancong Ren, Yingji Zhang, Qinfan Zhang, Jiayu Hu, Haozhe Shan, Han Dong, Jinpeng Lu, Yinda Chen, Yi Zhang, Yong Dai, and Xiaozhu Ju. VLA-Trace: Diagnosing vision-language-action models through representation and behavior tracing. *arXiv preprint arXiv:2605.30117*, 2026. URL https://arxiv.org/abs/2605.30117.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Linhan Wang, Zijian An, Mingyuan Zhang, Chen Dai, Yi Xu, Can Cui, Zichong Yang, Yinlin Chen, Lifeng Zhou, and Chang-Tien Lu. GlanceWAM: Sparse test-time imagination for world-action models. *arXiv preprint arXiv:2608.23927*, 2026a. URL https://arxiv.org/abs/2608.23927.

Siyin Wang, Junhao Shi, Zhaoyang Fu, Xinzhe He, Feihong Liu, Chenchen Yang, Yikang Zhou, Zhaoye Fei, Jingjing Gong, Jinlan Fu, Mike Zheng Shou, Xuanjing Huang, Xipeng Qiu, and Yu-Gang Jiang. World action models: The next frontier in embodied AI. *arXiv preprint arXiv:2605.12090*, 2026b. URL https://arxiv.org/abs/2605.12090.

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

Chushan Zhang, Jinguang Tong, Xuesong Li, Yikai Wang, and Hongdong Li. Keep the future, drop the rollout: RIFT for world action models. *arXiv preprint arXiv:2608.11521*, 2026a. URL https://arxiv.org/abs/2608.11521.

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://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Jiajie Zhang, Yankai Xiang, and Changhao Chen. Memory that changes action is not memory that guides it: Counterfactual auditing of history-conditioned robot policies. *arXiv preprint arXiv:2609.27247*, 2026b. URL https://arxiv.org/abs/2609.27247.

Yuyang Zhang, Wenyao Zhang, Zekun Qi, He Zhang, Haitao Lin, Jingbo Zhang, Yao Mu, Xiaokang Yang, Wenjun Zeng, and Xin Jin. ImageWAM: Do world action models really need video generation, or just image editing? *arXiv preprint arXiv:2606.19531*, 2026c. URL https://arxiv.org/abs/2606.19531.

Zhanguang Zhang, Zhiyuan Li, Behnam Rahmati, Rui Heng Yang, Yintao Ma, Amir Rasouli, Sajjad Pakdamansavoji, Yangzheng Wu, Lingfeng Zhang, Tongtong Cao, Feng Wen, Xinyu Wang, Xingyue Quan, and Yingxue Zhang. Do world action models generalize better than VLAs? a robustness study. *arXiv preprint arXiv:2603.22078*, 2026d. URL https://arxiv.org/abs/2603.22078.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025. URL https://arxiv.org/abs/2504.02792.

A Additional Evaluation Details and Held-Out Responses

The held-out LIBERO study uses 20 base states from tasks 5 and 9, shared by all four models and disjoint from discovery. The prospective study constructs a separate set of counterfactuals. Restoration uses 20 source trajectories from tasks 0 and 1; the budget study uses 50 trajectories from the five tasks indexed 5–9. Reported budget results use a confirmation cohort distinct from the trajectories used for pipeline development.

Physical comparisons. The factor labels and phases are defined in Section 3.1. F1 uses signed displacements $d \in \{-4, -2, -1, 1, 2, 4\}$ cm. The F3-G source-control study uses $d \in \{-1, -0.5, 0.5, 1\}$ cm, moving only the target basket and retaining the robot configuration and grasped object. F2 includes object-only and object–goal rigid-shift conditions; Table 2 reports the object-only condition. Counterfactuals are screened for geometric validity before policy evaluation, including the requirement that no new collision is introduced. Readouts follow Section 3.1.

Table 2: Held-out physical-response slopes on 20 shared base states from tasks 5 and 9. Entries show estimates and reported 95% intervals in each registered command readout per cm; F2 is object-only. These estimates are separate from the discovery curves in Figure 2.

Model	F1 PreGrasp	F2 PreGrasp	F3 PrePlace	F3 Transport
Fast-WAM	2.409	1.638	.302	.019
	[1.373, 3.798]	[.814, 3.067]	[.239, .369]	[−.064, .080]
Fast-WAM-Joint	3.183	2.003	.545	.005
	[1.696, 4.976]	[1.006, 3.356]	[.390, .677]	[−.061, .091]
Fast-WAM-IDM	2.397	1.872	.546	−.007
	[1.863, 3.082]	[1.290, 2.648]	[.446, .651]	[−.141, .216]
ImageWAM	1.905	1.150	.788	−.041
	[1.400, 2.603]	[.911, 1.522]	[.520, 1.047]	[−.140, .057]

Uncertainty. All intervals are 95% bootstrap intervals from 10,000 draws. The broad audit resamples tasks and paired base states; gain analysis resamples task–trajectory blocks and refits gains. Restoration uses donor-within-source means and resamples sources within each of two fixed tasks. RoboTwin closed-loop estimates weight the five tasks equally (Appendix E.2).

B Interface Transmission and Reconstruction

B.1 Prospective Transmission

The prospective study freezes interface assignments, directional rules, recipient states, doses, and geometric tolerances before counterfactual construction. Of 560 registered cases, 374 pass collision checks; all 186 exclusions precede policy evaluation. The same valid cases are evaluated in all four models. Table 3 separates transmission of the realized action effect from prediction of its sign.

Table 3: Prospective results on 374 valid counterfactuals per model. These are reported point estimates. Values printed as .000 or 1.000 are rounded summaries.

Model	Transfer T	Relative residual E	Sign accuracy
Fast-WAM	.999	.011	.810
Fast-WAM-Joint	1.000	.007	.765
Fast-WAM-IDM	.955	.136	.767
ImageWAM	1.000	.000	.816

Transfer remains high, while sign accuracy is .765–.816. The interface test measures transmission of an observed response, independently of whether the frozen directional rule predicts it correctly.

B.2 Replacement Sets and the Fast-WAM-IDM Reference

Supplementary held-out transfer is .926 [.912, .940] for Fast-WAM-Joint's future-node replacement and .003 [.001, .004] for ImageWAM's non-image prefix. The latter separates image and non-image inputs. Controls use model-specific same-value checks or changed-value replacements; structural absence is distinct from a measured zero effect.

Fast-WAM-IDM's discovery future-only interventions give $T = .961$ [.930, .991] and $E = .418$ [.352, .487] across 1,727 cases. The future latent is only part of Fast-WAM-IDM's action-conditioning inputs and does not provide the separate current/future consumer assignments used in Fast-WAM-Joint.

C Fast-WAM-Joint Source Control and Local Restoration

C.1 Source Conditioning and Interaction

The F3-G comparison uses 132 registered states and 1,054 valid geometric records out of 1,056. Source assignments follow Equation 3. Table 4 reports the future-minus-current donor-conditioned magnitude gap and its association with target distance. The future effect is larger in both splits. The distance slope summarizes the association between the magnitude gap and target distance.

Table 4: Additional F3-G estimates in the registered radial-command readout. The slope is per meter of target distance.

Split	Magnitude gap	Distance slope	95% interval
Development	.0793	$-.4046$	$[-.4727, -.3420]$
Validation	.0587	$-.3264$	$[-.4462, -.2021]$

Strict current–future interaction. We compute the current–future interaction $\mathbf{J}$ in Equation 4 from complete four-cell outputs. For F1, a state contributes a displacement magnitude only when both signs pass geometric checks, retaining 174 of 300 registered signed-dose grids from 35 states. For F3-G, 526 of 528 grids pass geometric screening, spanning 132 states in 47 source-trajectory groups. Each analysis spans five tasks. All 700 tensors satisfy $\boldsymbol{\delta}_{11} = \boldsymbol{\delta}_{10} + \boldsymbol{\delta}_{01} + \mathbf{J}$ exactly.

Table 5: Strict current–future interaction in LIBERO Fast-WAM-Joint. Values are in 10^{-3} de-normalized translation-command units; intervals are marginal 95% bootstrap intervals.

Factor	Signed radial mean [95% CI]	Mean case radial RMS
F1	$0.0585\ [-0.7738, 0.7663]$	9.9216
F3-G	$-0.0829\ [-0.2007, 0.0311]$	3.2264

F1 uses the end-effector–object direction in three dimensions; F3-G uses the horizontal object–goal direction, both fixed from the recipient. The signed radial mean averages the 32 projected entries of $\mathbf{J}$; radial RMS measures their root-mean-square magnitude. Dose signs are retained separately from the projection direction. Donor-dose summaries are averaged within state. The F1 bootstrap resamples tasks and registered states; the F3-G bootstrap resamples tasks and trajectory groups while retaining state weights. Both use 10,000 draws. The F3-G interaction uses global min–max de-normalization. Main-effect RMS comparisons below use the same readout as the corresponding interaction.

Every dose-specific signed interval includes zero. Table 5 reports RMS as a descriptive magnitude, not a separate test of exact additivity. For scale, the mean radial RMS values of $\mathbf{J}$, $\boldsymbol{\delta}_{10}$, $\boldsymbol{\delta}_{01}$, and $\boldsymbol{\delta}_{11}$ are 0.00992, 0.01762, 0.18381, and 0.19217 for F1, and 0.00323, 0.00541, 0.01868, and 0.02100 for F3-G, respectively. The pooled F3-G interaction includes both development and validation cases. In both analyses, the interaction is smaller than the future and joint effects but comparable in scale to the future-fixed current effect. Since $\mathbf{U}_{C|F_D} - \mathbf{U}_{C|F_R} = \mathbf{J}$, an unresolved signed mean of $\mathbf{J}$ does not imply identical current-effect vectors across future backgrounds.

C.2 Restoration Assignments and Paired Errors

We use the conditions defined in Section 3.3. P is the captured propagated future K/V; B is blocking, P_V and P_W restore the early and late windows, and P_{all} restores every layer. The windows were fixed before evaluating 20 source trajectories in tasks 0 and 1; geometric screening retains 103 of 120 F1 donors. For each donor i, the error $e_{i,X}$ compares the radial command vector in condition X with full restoration. Donor errors are first averaged within source, then sources are weighted equally within each equally weighted task. In Table 6, e_B, e_V, e_W denote the resulting aggregate errors for B, P_V, P_W. Full restoration and the natural donor action have a separately measured RMS difference of .00264 [.00216, .00333].

Table 6: Restoration errors and paired reductions on the same 20-source cohort. Units are de-normalized translation commands; intervals are marginal 95% paired-bootstrap intervals. Early and late windows were prespecified; † marks the additional windows in the six-window scan on the same cohort.

Condition or contrast	Estimate	95% interval
Blocked error e_B	.24767	[.21836, .27834]
L0–4, early e_V	.24730	[.21802, .27793]
L5–9†	.24693	[.21774, .27753]
L10–14†	.24385	[.21486, .27423]
L15–19†	.23529	[.20717, .26494]
L20–24†	.23904	[.21008, .26994]
L25–29, late e_W	.03772	[.03220, .04408]
Blocked-to-late $e_B - e_W$	.20995	[.18419, .23813]
Early-to-late $e_V - e_W$	.20958	[.18387, .23775]

Six-window depth scan. The scan evaluates six fixed five-layer windows on the same 103 donors. Recomputing the early- and late-window errors for the scan gives the values in Table 6 exactly. Statistics use donor-within-source aggregation and 10,000 paired source draws within task. The five windows spanning layers 0–24 leave errors of .2353–.2473, whereas L25–29 leaves .0377. The adjacent L20–24 minus L25–29 contrast is .2013 [.1758, .2296]. Recovery is concentrated in the final window at this resolution. Intervals are marginal.

D Response Changes under Reduced World-Branch Computation

All LIBERO budget comparisons use the 50-trajectory, 200-donor confirmation cohort described in Section 4.1. Responses use stepwise de-normalized translation commands.

Table 7 compares shared-gain fits for LIBERO's three basis contrasts and RoboTwin's natural response. All gains are evaluated out of fold. The complementary strict-future response is summarized below.

Table 7: Complementary response and shared-gain estimates relative to $K = 10$. LIBERO RMSE uses translational-command units; RoboTwin RMSE uses mm after forward kinematics. Q is reported in percent. Brackets are 95% intervals; RMSE entries marked † are point estimates.

Response family	Unscaled RMSE	Gain-adjusted RMSE	Q (%)
LIBERO, three contrasts, $K = 8$	.004210†	.004218†	−0.37 [−2.07, .44]
LIBERO, three contrasts, $K = 5$	.00794 [.00623, .00985]	.00781 [.00617, .00964]	3.29 [−9.05, 13.63]
RoboTwin, natural	2.979 [2.327, 3.603]	2.860 [2.162, 3.647]	7.9 [−18.5, 23.7]

Matched $K = 8$ extension. The three-budget comparison uses 50 trajectories from tasks 5–9, with four signed doses per trajectory. Separate budget fits use the same five trajectory folds and 10,000

task–trajectory draws with gain refitting. Table 8 reports natural donor–recipient response vectors, which equal $\mathbf{A}_{11} - \mathbf{A}_{00}$ in the radial readout. Table 7 uses one gain shared across the three basis contrasts defined in Section 3.4.

Table 8: Matched three-budget extension. Future-norm change is the relative change in the mean $\|\mathcal{P}_i \mathbf{U}_{F|C_D}\|_2$. RMSE uses de-normalized translation-command units. Static savings are median paired totals over five identical endpoint inputs per source, not closed-loop savings.

K	Natural-response RMSE	Future-norm change [95% CI]	Static saving
10	0	—	—
8	.00473	-1.20% [-3.37%, 1.10%]	7.95%
5	.00950	-10.87% [-18.39%, -0.66%]	22.24%

The $K = 8$ norm-change interval includes zero; the $K = 5$ interval is negative. Timing uses 250 matched static inputs in counterbalanced budget order, excludes preparation, and retains ten action steps. Static totals are separate from the 23.7% continuation saving; $K = 8$ is evaluated only on static inputs.

On RoboTwin, the natural-response mean norm decreases by 6.30% [1.94%, 10.40%]. For the strict future effect given donor current, $\mathbf{U}_{F|C_D}$, the norm changes by -0.19% [$-8.28\%, 6.07\%$], while the vector RMSE is 3.222 [2.302, 4.182] mm. Its shared-gain fit gives $Q_5 = -9.9\%$ [$-134.3\%, 11.3\%$]. Thus, a nearly unchanged mean norm coexists with a changed response vector.

Figure 5c displays normalized natural-response differences and gain-adjusted residuals. The two rows for each benchmark show the natural-response difference and its residual after gain adjustment. The plotted LIBERO natural-response analysis is separate from the three-contrast fit in Table 7.

E RoboTwin Attribution, Decoder Context, and Paired Execution

Matched node-minus-current contrast. Equation 5 is evaluated on 34 Fast-WAM-Joint sources at $+1$ cm. The two conditions share donor current inputs, recipient context, and the recipient baseline, while the future inputs change from F_R to F'.

Table 9: RoboTwin Fast-WAM-Joint effects on 34 sources. Values are task-equal summaries in mm after condition-wise forward kinematics, with 95% hierarchical paired-bootstrap intervals.

Effect	Signed radial mean [95% CI]	Mean radial RMS [95% CI]
$\mathbf{N}$	1.577 [-1.426, 4.906]	7.707 [5.418, 10.859]
$\mathbf{U}_{C\|F_R}$	-1.727 [-2.289, -1.277]	2.666 [2.161, 3.312]
$\mathbf{N} - \mathbf{U}_{C\|F_R}$	3.304 [0.285, 6.933]	7.439 [4.642, 11.229]

The analysis uses 10,000 task–source bootstrap draws. The difference-vector RMS, 7.439 mm, differs from the 5.040 mm difference of RMS summaries.

E.1 Conditional Context Effects

The interface–command-state analysis evaluates Fast-WAM and Fast-WAM-Joint on the same 34 sources. The decoder's command-state input is a 14-dimensional vector of joint-drive targets and gripper commands, encoded as an action-context token. It also conditions interface generation.

For this separate two-factor intervention, a selects the interface and b the command state, with recipient (0) or donor (1) values. Let $\mathbf{y}_{ab} \in \mathbb{R}^{32}$ denote the resulting radial target vector after forward kinematics. The ordered decomposition is

$$\mathbf{y}_{11} - \mathbf{y}_{00} = (\mathbf{y}_{10} - \mathbf{y}_{00}) + (\mathbf{y}_{11} - \mathbf{y}_{10}). \tag{8}$$

The first term changes the interface under recipient command state; the second changes command state under the donor interface. Table 10 shows the corresponding signed means. Each effect is conditional on the input assignment of the preceding step.

Table 10: Ordered conditional effects on 34 RoboTwin sources per model (mm). Positive interface and negative command-state estimates counteract in the mean natural response.

Model	Interface	Command state	Natural
Fast-WAM	7.614	−8.703	−1.088
Fast-WAM-Joint	1.633	−6.367	−4.734

The interface–command-state interaction has signed mean .203 [.068, .339] mm for Fast-WAM and .310 [.065, .551] mm for Fast-WAM-Joint, indicating background dependence. Replacing both inputs matches the independently evaluated natural donor action elementwise in all 68 model–source pairs, linking the combined replacement to the natural donor reference.

The interface–command-state interaction is $\mathbf{y}_{11} - \mathbf{y}_{10} - \mathbf{y}_{01} + \mathbf{y}_{00}$. Its two factors differ from the current/future sources defining $\mathbf{J}$ in Appendix C.1.

E.2 Closed-Loop Eligibility and Completion

The closed-loop study registers 50 task–seed pairs. A shared benchmark preflight evaluates environment reset and expert eligibility before either budget runs. Twenty-nine pairs pass and have technically complete episodes under both budgets; 21 fail the preflight. Table 11 gives the task-level pairing.

Table 11: RoboTwin paired completion conditional on the shared preflight. "Only" denotes success under that budget and failure under the other.

Task	Frozen	Valid	K=10	K=5	10 only	5 only
adjust_bottle	10	3	2	1	1	0
beat_block_hammer	10	8	5	5	2	2
hanging_mug	10	3	1	1	0	0
move_can_pot	10	10	5	7	0	2
move_pillbottle_pad	10	5	5	5	0	0
Total	50	29	18	19	3	4

The task-equal paired difference ($K = 5$ minus native) is -0.027 [-0.267, .160], using 10,000 hierarchical paired-bootstrap replicates. Its sign differs from the pooled difference because valid counts vary by task. The 21 excluded pairs retain their preflight status; completion is evaluated among eligible pairs.