IDENTIFYING THE APPEARANCE–KINEMATICS TRADE-OFF IN CROSS-EMBODIMENT TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-embodiment transfer changes both what a robot can do and what a vision-based policy sees. Separating these effects requires interventions that vary collision geometry and silhouette independently, which are unavailable under the manufacturing constraint linking a robot's outline to its collision body. We introduce a strict 2×2 simulator design that supplies these counterfactual interventions. Under the stated identification assumptions, including invariance of the outcome mechanism between the natural and simulator models on and off the hardware manifold, we prove that the four cells identify the natural model's mediation estimands. Procedural shells, source-calibrated frozen dose families, and field-level checks of bit-identical physics implement the design. We express the local appearance–kinematics exchange rate as the percentage change in link length equivalent to one appearance just-noticeable difference (JND), at the frozen operating point and silhouette distance. On the simulated panel, the preregistered pooled primary estimate is 1.50 (95% Fieller set $[1.09,\ 2.11]$). Stratum estimates are 2.71 $[1.74,\ 5.29]$ for tasks tolerating centimetre-scale placement errors and 0.84 $[0.50,\ 1.24]$ for tasks requiring millimetre-scale precision; these two stratum intervals are 95% Fieller sets. The pooled interaction is -1.3 percentage points, with its interval $[-4.0,\ +1.4]$ lying entirely within the prespecified ±5-percentage-point margin. These results quantify how the balance between appearance and kinematic sensitivity varies with task contact precision.

1 INTRODUCTION

Robot swaps change both capability and visual input. Cross-embodiment transfer asks a policy trained on one robot to control another. Morphology distances and morphology-conditioned policies address the resulting change in capability (Abe et al., 2026; Mishra et al., 2026; Suzuki et al., 2026; Li et al., 2026). For a visual policy, however, the robot also changes inside its own observations. Masking an unfamiliar end-effector can improve transfer (Piseno et al., 2026), while matched-data experiments can show a smaller response to morphology (Yang et al., 2026). Recolouring a real arm can alter success without changing its kinematics (Li et al., 2024). These findings motivate a shared measurement question: how much transfer loss comes from appearance, how much from kinematics, and how does their relative importance depend on the task?

Hardware comparisons couple silhouette and collision geometry. A robot's appearance includes material and silhouette. Material can be changed independently by recolouring, whereas the outer-surface relation considered here ties silhouette to collision geometry. Comparing robots therefore changes both along the hardware manifold. The two missing contrasts—a new silhouette on the source skeleton and a new skeleton with the source silhouette—lie outside that manifold (Fig. 1). This is a support constraint: adding robots cannot supply these counterfactuals. It concerns silhouette and outer-surface geometry; inertia and joint and torque limits remain free.

Four simulator interventions supply the missing contrasts. MuJoCo exposes collision geometry, visual geometry and material as separate fields (Todorov et al., 2012). We use this separation to construct a strict 2×2: the source robot, the ordinary robot swap and the two missing counterfactuals. Four interventions identify the natural model's mediation effects under the stated assumptions, including a shared outcome mechanism between the natural and simulator models both on and off

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.

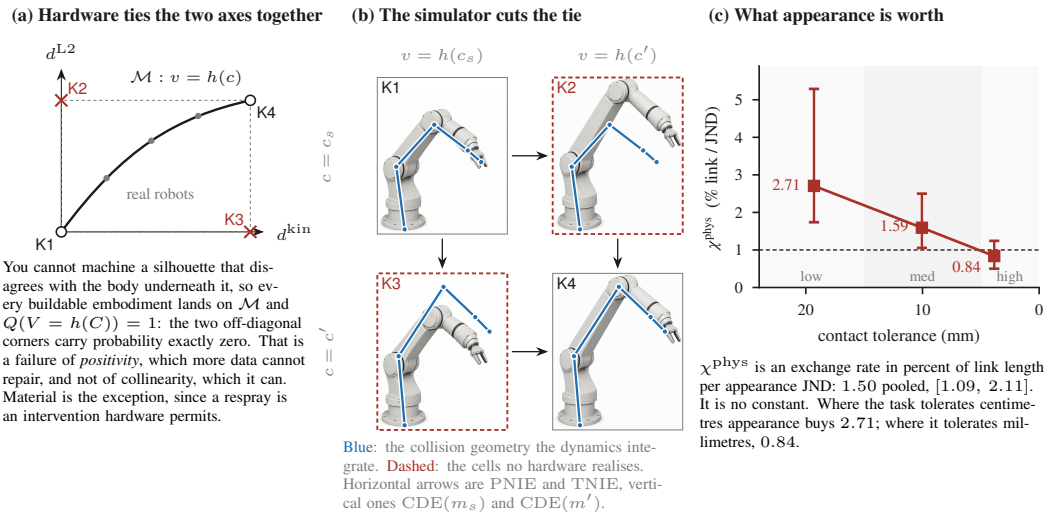


Figure 1: Appearance and kinematics cannot be separated on hardware, can be separated in a simulator, and once separated the trade-off between them has a number. **(a)** Every manufacturable robot lies on $\mathcal{M}$, so the chimeras K2 and K3 fall outside it and no observational panel reaches them (Eq. (2)). **(b)** MJCF stores collision geometry, visual geometry and material as independent fields and the dynamics read only the first, so both chimeras compile. Columns fix visual geometry, while rows fix collision geometry. These geometries disagree in the two dashed cells, which cannot be realised on hardware. The arrows are two of the contrasts the square identifies (Eq. (3)). **(c)** The exchange rate χ^{phys} against contact tolerance, with 95% Fieller sets.

the hardware manifold (A4). We build the cells with procedural shells, calibrate dose families on the source before freezing them, and check physical equivalence field by field for twins sharing a skeleton. We report the interaction that reallocates pathway contributions between the two decomposition orders while preserving their common total effect.

The measured trade-off varies across contact-precision strata. We express the local appearance–kinematics exchange rate, χ^{phys}, in percent of link length per appearance just-noticeable difference (JND). On our simulation panel, the pooled estimate is 1.50 with a 95% Fieller interval of $[1.09, 2.11]$. The corresponding estimates are 2.71 ($[1.74, 5.29]$) for low-contact-precision tasks and 0.84 ($[0.50, 1.24]$) for high-contact-precision tasks. All intervals are 95% Fieller sets. Fitted slopes also predict residual transfer gaps and estimate a boundary for appearance-randomisation sufficiency (Section 5).

Our contributions are threefold.

- **Identifiable mediation contrasts.** We establish the conditions under which four simulator interventions identify the natural model's appearance and kinematic mediation effects, and show how their interaction governs order-dependent pathway allocations (Section 4).
- **An executable intervention protocol.** We construct the four cells using procedural shells, source-calibrated frozen dose families and field-level checks of twin-pair physical equivalence (Section 4.3).
- **A calibrated measure of transfer sensitivity.** We define a local exchange rate in interpretable units, quantify its variation across contact-precision strata and evaluate predictions of residual transfer gaps (Sections 4.4 and 5).

2 Related Work

Morphology-aware transfer measures changes in capability. Morphology descriptors and Fused Gromov–Wasserstein similarity relate physical differences to transfer behaviour in state-based

locomotion (Mishra et al., 2026; Abe et al., 2026). Morphology conditioning and procedural embodiment variation expand the robots a policy can control (Suzuki et al., 2026; Li et al., 2026; Parakh et al., 2025). Our design measures the appearance and kinematic pathways jointly in visual control, with a fixed control interface that makes their contrasts comparable.

Robot-image interventions establish appearance sensitivity. Masking, repainting and augmentation change the robot pixels received by a policy (Lepert et al., 2025; Piseno et al., 2026; Chen et al., 2024; 2025; Ji et al., 2026); material interventions also reach real hardware (Li et al., 2024). Studies varying both appearance and morphology provide complementary comparisons (Yang et al., 2026; Chen et al., 2026). Image editing generally acts on material and silhouette together, while observational hardware panels retain the coupling between collision geometry and silhouette. We assign these fields separately and measure the dose on each axis, enabling the crossed contrasts needed for a pathway decomposition.

Causal simulator design makes the required factors assignable. Interventions on generative factors and structured simulation benchmarks support causal analysis and representation evaluation (Suter et al., 2019; Ahmed et al., 2021; Zhang et al., 2020). We use this capability to remove the deterministic edge from collision geometry to silhouette. Under the stated transport condition, mediation analysis (Robins & Greenland, 1992; Pearl, 2009; VanderWeele, 2015; Richardson & Robins, 2013) then connects the four simulator cells to the natural model's estimands. Appendix L gives the full taxonomy and detailed comparisons.

3 Problem Formulation

Embodiment separates physical and visual variables. A task τ specifies a scene distribution and success criterion. An embodiment $e = (c, v, t)$ consists of collision geometry with inertia, joint and torque limits (c), visual geometry (v), and material and texture (t). A frozen source-trained policy π_{e_s} generates an episode of horizon K through

$$\begin{aligned} x_{k+1} &= F\big(x_k,\ \mathrm{clip}_c(a_k);\ c,\ \varsigma,\ \varepsilon\big), & o_k &= R\big(x_k;\ v,\ t,\ \varsigma\big), \\ a_k &= \pi_{e_s}\big(o_{0:k},\ a_{0:k-1};\ \varepsilon\big), & Y &= \Sigma_\tau\big(x_{0:K}, \varsigma\big) \in \{0,1\}, \end{aligned} \tag{1}$$

and $\mathrm{SR}_\tau(\pi_{e_s}; e) = \mathbb{E}_{\varsigma,\varepsilon}[Y]$. The implementation ensures that the transition F reads c but not v or t, while the success criterion Σ_τ reads physical state but not images. Appearance enters through the renderer R. We sample the exogenous scene variables ς and policy and solver randomness ε for every intervention.

Hardware support excludes the off-diagonal contrasts. The manufacturing map $h : \mathcal{C} \to \mathcal{V}$ assigns each collision body its outer surface. Under this constraint, real embodiments lie on $\mathcal{M} = \{(c, v, t) : v = h(c)\}$. Thus, any hardware distribution Q satisfies

$$Q\big(V = h(C)\big) = 1, \qquad \text{so} \qquad Q\big(V \in B \mid C = c\big) = 0 \ \text{ whenever } h(c) \notin \overline{B}, \tag{2}$$

The restriction concerns silhouette and outer-surface geometry; inertia and joint and torque limits remain free. With $m_s := h(c_s)$, $m' := h(c')$ and material fixed at t_s, define $\theta(c, m) := \mathbb{E}_{\varsigma,\varepsilon}[Y(c, m, t_s)]$. The target contrasts require all four means $\theta(c_s, m_s)$, $\theta(c_s, m')$, $\theta(c', m_s)$ and $\theta(c', m')$, including the two off $\mathcal{M}$.

Five operators separate the intervention pathways. $T_{\mathrm{L1}} = \mathrm{do}(T = t')$ changes material; $T_{\mathrm{L2}} = \mathrm{do}(V = m')$ at $C = c_s$ changes silhouette alone; $T_{\mathrm{kin}}^{\mathrm{sync}} = \mathrm{do}(C = c', V = h(c'))$ is the ordinary robot swap; and $T_{\mathrm{kin}}^{\mathrm{froz}} = \mathrm{do}(C = c', V = m_s)$ changes the skeleton while retaining source visual geometry. The silhouette-only and frozen-geometry operators are unavailable in hardware under the manufacturing constraint. The fifth operator, T_{replay}, forces the source's achieved end-effector path onto the target using absolute poses; replaying delta commands on a different arm would accumulate trajectory drift. Table 2 lists the operators and their hardware availability.

Three distances measure the assigned changes. Material distance d^{app} is the cosine distance between mask-weighted DINOv2 patch tokens inside the robot mask (Oquab et al., 2024), varied

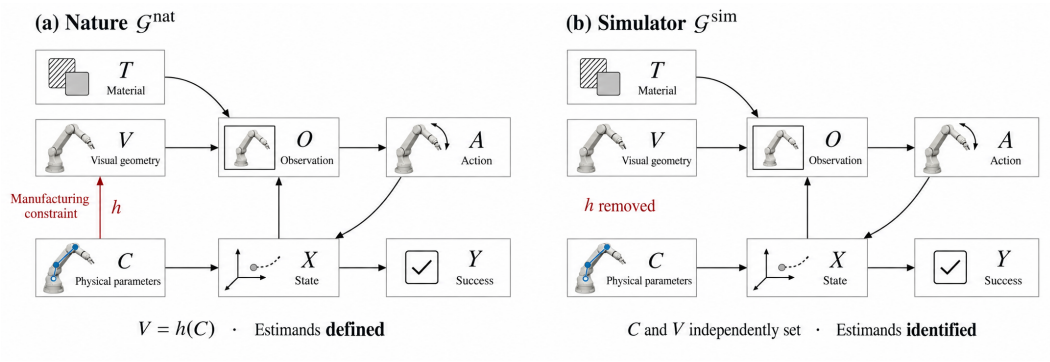


Figure 2: The two models differ only in the edge $C \xrightarrow{h} V$ (red), the manufacturing constraint. That edge makes V a mediator, and it is deterministic, which is what makes the mediation quantities well defined but unidentifiable in nature. MuJoCo's dynamics never read V or T. Seeds ς, ε omitted.

through T_{L1}. Silhouette distance d^{L2}, varied through T_{L2}, averages $1 - \text{IoU}$ and symmetric contour Chamfer distance over matched poses. The kinematic axis d^{kin} is the pre-registered control parameter, percent of link length; five plugins provide descriptive read-outs, with one nominated in advance to invert the dose grid (Appendix D). Interventions separate the axes; the distances measure their changes and need not themselves be unconfounded instruments. Although d^{app} can respond to shape, the material family changes colour only. Build-time checks keep depth and segmentation bit-identical, fixing its mask (Appendix F); silhouette changes enter through the separate regressor d^{L2}. The three success-rate slopes at the pre-registered working point are $\beta^{\text{app}}, \beta^{\text{L2}}, \beta^{\text{kin}}$.

4 Identification by Construction

4.1 Counterfactual design

Two causal models connect the target effects to interventions. In the natural model $\mathcal{G}^{\text{nat}}$, the edge $C \xrightarrow{h} V$ makes visual geometry a mediator and defines the natural direct and indirect effects. Its deterministic constraint prevents estimation of the missing contrasts. The simulator model $\mathcal{G}^{\text{sim}}$ omits that edge: MJCF assigns collision geometry, visual geometry and material independently, and MuJoCo's dynamics read only the first (Todorov et al., 2012). Figure 2 depicts the models. Assignable simulator inputs have precedent (Suter et al., 2019); here they supply the contrasts required by the natural model's mediation estimands.

Four cells vary collision and visual geometry independently. Holding material at t_s, the design pairs the source robot $\mathsf{K1}$ with the ordinary swap $\mathsf{K4}$ and two chimeras: $\mathsf{K2}$ retains source physics with target visual geometry; $\mathsf{K3}$ retains source visual geometry with target physics. Writing $\mathsf{K1} = \theta(c_s, m_s)$, $\mathsf{K2} = \theta(c_s, m')$, $\mathsf{K3} = \theta(c', m_s)$ and $\mathsf{K4} = \theta(c', m')$ gives

$$\begin{gathered} \text{TE} = \mathsf{K4} - \mathsf{K1}, \quad \text{PNIE} = \mathsf{K2} - \mathsf{K1}, \quad \text{TNIE} = \mathsf{K4} - \mathsf{K3}, \\ \text{CDE}(m_s) = \mathsf{K3} - \mathsf{K1}, \quad \text{CDE}(m') = \mathsf{K4} - \mathsf{K2}, \quad \text{INT} = (\mathsf{K4} - \mathsf{K2}) - (\mathsf{K3} - \mathsf{K1}). \end{gathered} \tag{3}$$

Figure 6 diagrams the contrasts. $\mathsf{K2}$ requires $v = h(c')$ exactly: intermediate silhouette doses estimate β^{L2}, not PNIE, and occupy a separate arm. INT is estimated from cell success rates on the probability scale; a logit-link interaction coefficient represents a different quantity.

4.2 Identification and assumptions

Identification requires a shared outcome mechanism. Assumption (A4) equates the outcome mechanisms of $\mathcal{G}^{\text{nat}}$ and $\mathcal{G}^{\text{sim}}$, including off $\mathcal{M}$. The simulation experiments cannot establish this equivalence. Assumption (A0) fixes the dimensions and semantics of the action and proprioceptive interface; ablation A2 examines the additional channel from joint-angle inputs. Assumption (A3) requires independent episodes conditional on the compiled model. Six further conditions concern

construction and bookkeeping (Appendices B and F). Common random numbers support paired estimation and reduce variance without being required for identification.

Proposition 1 (Identification of the mediation quantities)**.** *Let $c_s \to c'$ be a kinematic contrast with the material held at t_s, and let $\theta(c, m) = \mathbb{E}_{\varsigma,\varepsilon}[Y(c, m, t_s)]$. Under* (A0)–(A2)*,* (A4) *and* (A5)–(A6)*:*

(i) ***(Twin identification.)*** *For every cell (c, v, t) of the design, $\mathbb{E}[Y(c, v, t)] = \mathbb{E}[Y \mid C = c, V = v, T = t]$. Consistency, unconfoundedness and positivity all hold by construction, and no cross-world condition is needed, because h is deterministic and we can apply it ourselves.*

(ii) ***(Indirect effect from the silhouette twin.)*** *With $m' = h(c')$, the chimera at (c_s, m') identifies $\theta(c_s, m') - \theta(c_s, m_s) = \mathbb{E}[Y(c_s, V_{c'}) - Y(c_s, V_{c_s})] = \mathrm{PNIE}$. The controlled direct effect $\mathrm{CDE}(m) = \theta(c', m) - \theta(c_s, m)$ is a different quantity and requires the other chimera, at (c', m_s). Because the mediator is deterministic, $\mathrm{PNDE} = \mathrm{CDE}(m_s)$ and $\mathrm{TNDE} = \mathrm{CDE}(m')$.*

(iii) ***(A residual appearance channel survives freezing v and t.)*** *Even with v and t fixed, the path $c \to x_k \to o_k \to a_k \to x_{k+1} \to \cdots \to Y$ is open, so $\mathrm{CDE}(m_s)$ is not a pure capability effect. Forcing the source's achieved end-effector path, T_{replay}, closes every path through o_k, and $\beta^{\mathrm{phys}} = \mathbb{E}[Y(c', \bar{x}^{\mathrm{ee}}) - Y(c_s, \bar{x}^{\mathrm{ee}})]$ carries no appearance channel* conditional on $\bar{x}^{\mathrm{ee}}$*; as a marginal quantity it remains a functional of (m_s, t_s), since the law of $\bar{x}^{\mathrm{ee}}$ depends on them.*

(iv) ***(The split is order-dependent.)*** *With $\mathrm{TE} = \theta(c', m') - \theta(c_s, m_s)$ and $\mathrm{TNIE} = \theta(c', m') - \theta(c', m_s)$, both decompositions hold exactly,*

$$\begin{gathered} \mathrm{TE} = \mathrm{CDE}(m_s) + \mathrm{TNIE} = \mathrm{PNIE} + \mathrm{CDE}(m'), \\ \boxed{\mathrm{TNIE} - \mathrm{PNIE} = \mathrm{CDE}(m') - \mathrm{CDE}(m_s) =: \mathrm{INT}} \end{gathered} \tag{4}$$

so "pure kinematic effect = kinematic effect minus silhouette effect" is well defined only when $\mathrm{INT} = 0$.

(v) ***(Observational non-identifiability.)*** *If Q is supported on $\mathcal{M}$, so that $Q(V = h(C)) = 1$, then each of* CDE*,* PNIE*,* TNIE *and* INT *involves at least one argument off $\mathcal{M}$ and is not a functional of the observational law. The only contrast whose arguments both lie on $\mathcal{M}$ is the total effect along $c \mapsto (c, h(c))$, and it is identified only if the panel's embodiment assignment is in addition unconfounded.*

Frozen observations separate feedback from execution. Proofs are in Appendix A. Part (iii) retains a configuration-induced appearance channel because different trajectories change the pixels even with visual geometry fixed. Arm F substitutes the source's observation stream while leaving execution to the target dynamics. Writing its outcome as $\theta_{\mathrm{F}}(c)$ yields

$$\underbrace{\mathrm{CDE}(m_s) - \big[\theta_{\mathrm{F}}(c') - \theta_{\mathrm{F}}(c_s)\big]}_{\text{configuration-induced appearance}}, \quad \underbrace{\big[\theta_{\mathrm{F}}(c') - \theta_{\mathrm{F}}(c_s)\big] - \beta^{\mathrm{phys}}}_{\text{execution gap}}. \tag{5}$$

4.3 Constructing geometry twins

Procedural shells make the geometry map explicit. Admitted links have separate contact and visible surfaces, with inertia recomputed from contact surfaces alone (Appendix C). This prevents compiler fallbacks from changing mass during silhouette interventions. We generate visual geometry as a convex offset surface with joint-bridging spheres under globally frozen parameters, making h known and deterministic as required by (A6). Borrowed visual meshes would not guarantee these properties or close joint gaps. Each dose family is calibrated once on the source skeleton, then frozen across embodiments; per-embodiment recalibration would change treatment intensity with the embodiment.

Extent control and physics checks isolate the assigned axes. A Helmert decomposition of log link scales separates extent a from shape $z \perp \mathbf{1}$. Constraining $|a| \leq 0.03$ keeps projected size within the camera matcher's adjustment range while permitting the $\pm 30\%$ shape amplitude required by the power analysis (Appendix D). Build-time assertions verify that 25 compiled physics arrays are bit-identical within each twin pair, K1–K2 and K3–K4 (Appendix F).

4.4 MEASURING THE EXCHANGE RATE

A local slope ratio measures the trade-off. Locally around the source, hold silhouette distance at its working-point value and assume $\mathrm{SR}_\tau(\pi_{e_s}; e) = \mathrm{Sr}(d^{\mathrm{app}}, d^{\mathrm{kin}}; d^{\mathrm{L2}})$ for a C^2 map on an open convex $\mathcal{D} \subseteq \mathbb{R}^2$. Then $\chi_\tau = (\partial \mathrm{Sr}/\partial d_a)/(\partial \mathrm{Sr}/\partial d_k) = \beta^{\mathrm{app}}/\beta^{\mathrm{kin}}$ is the marginal rate of substitution. Both derivatives are evaluated at the same point, so the logit-link derivative $p(1-p)$ cancels in the ratio. Geometrically, $-\chi$ is the slope of an iso-success contour and is undefined when $\beta^{\mathrm{kin}} = 0$. We calibrate success to $[0.2, 0.8]$ to retain slope information. A panel-wide ratio additionally requires a single-index response, assessed by ablation A4 and all four drift diagnostics (Appendix A.3); the local estimate does not. Lemma 1 formalises these properties and the reference-population dependence of standardised ratios.

Physical units give the ratio an interpretable scale. The primary summary is

$$\chi^{\mathrm{phys}} := \left. \frac{\partial \mathrm{SR}/\partial(\text{appearance JNDs})}{\partial \mathrm{SR}/\partial(\text{percent of link length})} \right|_{d_0}, \tag{6}$$

evaluated at the pre-registered working point d_0, using the pre-registered control parameter in the denominator and the psychophysical rescaling of β^{app} in the numerator (Appendix D). One appearance JND has the same local success-rate effect as χ^{phys} percent of link length. The five kinematic plugins are descriptive read-outs outside the primary fit; ablation A5 replaces the control parameter with each plugin and reports the change in the joint fit. We also report $\vartheta = \mathrm{atan2}(\beta^{\mathrm{app}}, \beta^{\mathrm{kin}})$, with $\chi = \tan\vartheta$, to interpret weak-denominator cases without division. Remark 1 separately decomposes $1 - \mathrm{SR}_\tau(\pi_{e_s}; e')$ into capability, observability and policy deficits.

5 EXPERIMENTS

We evaluate the four-cell decomposition, its dependence on contact precision, and its ability to predict residual transfer gaps.

5.1 EXPERIMENTAL SETUP

Robot panel and tasks. The evaluation uses simulated manipulation. Auditing 23 MuJoCo Menagerie (Zakka et al., 2022) arms admits 16 under the geometry-separation criterion; 15 mount on the bench, giving 12 training skeletons and 3 held out of every fit (Appendix C). Six robosuite (Zhu et al., 2020) tasks span three measured contact-precision strata. A body-free gripper on a mocap body follows the frozen reference while each segment's object-frame anchor is rigidly perturbed. We define τ^{prec} as the perturbation at which success reaches 50%: low contact above 15 mm, medium between 5 and 15 mm, and high below 5 mm. Appendix M discusses the scope of this evaluation.

Policies and paired evaluation. Demonstrations come from MimicGen (Mandlekar et al., 2023). The primary policy is a CNN-UNet diffusion policy on RGB (Chi et al., 2025); ACT (Zhao et al., 2023) is an equal second architecture, and a point-cloud policy (Ze et al., 2024) provides the depth control. Each condition runs $N_{\mathrm{ep}} = 100$ episodes under five training seeds. All cells share a pre-registered scene-seed list, making cell differences paired. At design time, camera-distance bisection matches robot pixel fraction within 2%, and base placement matches the source's log-manipulability profile. These post-treatment quantities are matched rather than included as regression covariates, which would open a collider path.

Intervention arms and controls. The cells are crossed with the material-dose axis; a separate silhouette-dose ladder estimates β^{L2} without requiring K2, whose mediator must equal $h(c')$. Replay arm R, frozen-observation arm F (Eq. (5)), and privileged planner Π (Remark 1) separate the remaining channels. Calibration uses a target-trained oracle, a sham varying only the rollout random stream, four negative controls, and two positive controls. Each positive control must change success by 30 pp; otherwise the pipeline is declared blind. Appendices E and J specify the arms and comparison tiers.

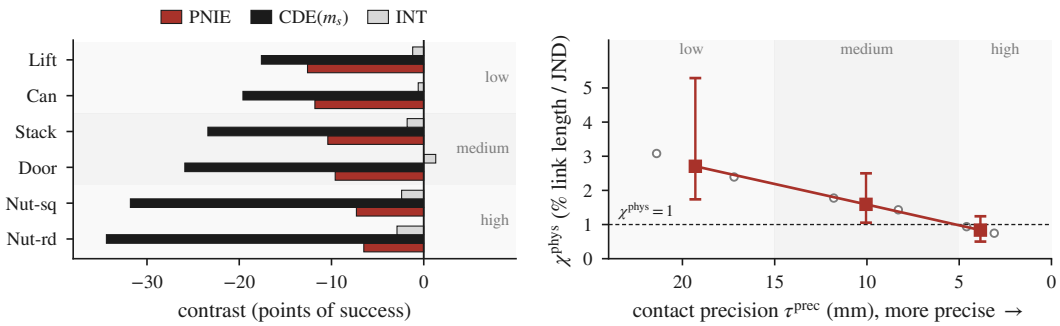


Figure 3: Left: four-cell contrasts (Eq. (3)) at each source skeleton's farthest embodiment pair, ordered by increasing contact precision. Right: exchange rates versus measured contact precision. Squares show the three preregistered strata with Fieller sets; open circles show six tasks for context. The analysis fits one slope pair per stratum. Cell means and the angular form appear in Fig. 7.

Table 1: The exchange rate and the mediation contrasts it is built from. The shaded row is the primary endpoint; the three below it are the pre-registered confirmatory stratification, which carries the single directional hypothesis. Mediation quantities are differences of cell means on the probability scale, in points. χ^{phys} is in percent of link length per appearance JND; its interval is a Fieller set and so is not symmetric about the point estimate, and ϑ is the same quantity read as a direction.

		mediation contrast (pp)			exchange rate (% link / JND)		
contact	τ^{prec} (mm)	PNIE	CDE(m_s)	INT	χ^{phys}	95% Fieller	ϑ (deg)
pooled	11.1	−9.7	−25.4	−1.3	1.50	[1.09, 2.11]	−124
low	19.3	−12.2	−18.6	−0.9	2.71	[1.74, 5.29]	−110
medium	10.1	−10.0	−24.6	−0.2	1.59	[1.05, 2.50]	−122
high	3.8	−6.9	−33.1	−2.6	0.84	[0.50, 1.24]	−140

Estimators and uncertainty. The estimand, models, decision rules and negligibility margin were fixed before evaluation (Appendix K). Three co-primary estimators are retained regardless of outcome: a binomial mixed model with crossed random effects for skeleton, variant, twin pair, source, target, task and training seed; a two-stage cluster-robust regression on cell success rates with CR2 correction (Pustejovsky & Tipton, 2018); and episode-level conditional logit within matched scene seeds (Appendix G). The mixed model yields χ^{phys} and its Fieller set, with cases and interpretations prespecified. Intervals are resampled at the skeleton level, and power counts independent skeletons. No-effect claims use a 5 pp equivalence margin. For a randomiser covering axes $\mathcal{A}$, the fitted slopes predict the residual gap without running the corresponding method on the target embodiment as $\hat{R}^{(\mathcal{A})} = |\sum_{j \notin \mathcal{A}} \hat{\beta}^j d^j|$; the criterion declares randomisation sufficient when this is at most 5 pp.

5.2 Decomposing transfer loss

Both pathways contribute to transfer loss. Transfer to the farthest embodiment has a pooled total effect of -36.4 percentage points. The appearance and kinematic paths contribute -9.7 and -25.4, respectively (Fig. 3 and Table 1). From low to high contact precision, the appearance contribution falls to a little over half its magnitude while the kinematic contribution grows by about three quarters. The pooled interaction, INT $= -1.3$ points with interval $[-4.0, +1.4]$, lies within ± 5 points. This satisfies the criterion for a negligible order-dependent shift in path contributions. As preregistered, the interaction is reported without gating the analysis; ratio stability is assessed separately below. Calibration holds source success at the same dynamic-range threshold across tasks by selecting the reset distribution and horizon. Differences in target-cell success therefore measure transfer from a common starting level. Both chimera cells lose success on every task, showing that changing either visual geometry or collision geometry can disrupt the source policy (Table 5).

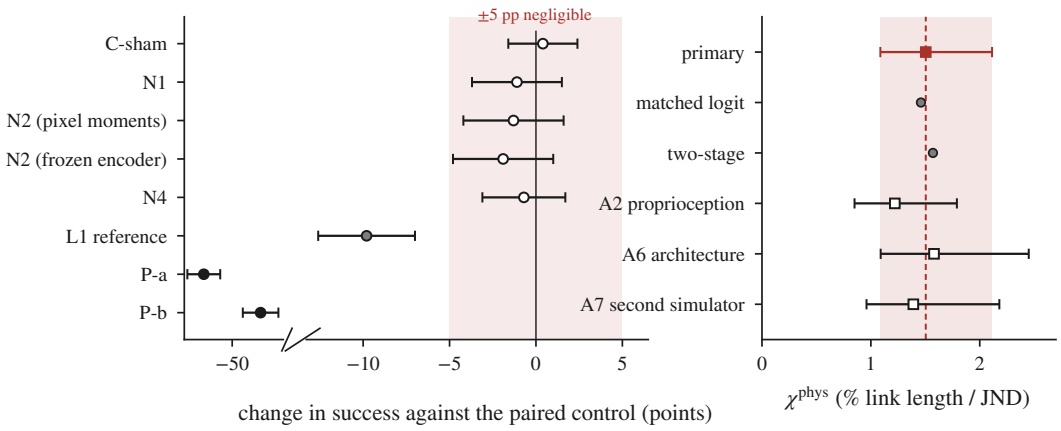


Figure 4: Left: the control arms against the 5 pp negligibility margin, on a broken axis because the positive controls sit an order of magnitude outside it. Open circles are the pre-registered no-effect claims, all inside the margin; the grey circle applies the same amplitude to the robot rather than the scene; filled circles are the positive controls. Right: the rate re-estimated under a different estimator, observation space, policy class and simulator, against the primary Fieller set (shaded); the two alternative estimators are point estimates, the ablations carry their own sets. Numbers in Tables 6 and 7.

Configuration changes carry a measurable appearance effect. The replay arm attributes -20.6 points of the kinematic contrast to physics, with -3.1 points mediated by configurations visited by the reshaped arm and -1.7 points of execution gap (Eq. (5)). The frozen-observation arm isolates the middle term, which otherwise would be misattributed (Proposition 1(iii)).

Target embodiments retain the ability to perform the tasks. Across tasks, the privileged planner has a small capability deficit, while the policy term accounts for most of the total deficit even at its lower bound. The target-trained image policy supplies the lower bound on attainable performance with the available observations. The exchange rate measures the local sensitivity of success to appearance and kinematic changes at the frozen operating point (Table 8).

5.3 Exchange rate across contact precision

Point estimates decrease with contact precision. Pooled over the panel, $\chi^{\text{phys}} = 1.50$ with a 95% Fieller set of $[1.09, 2.11]$. The ordering across all three strata follows the preregistered directional hypothesis (Fig. 3 and Table 1). The same appearance difference corresponds to 2.71 percent of link length at low contact precision and 0.84 at high precision. The low-contact set, $[1.74, 5.29]$, excludes unity; the high-contact set, $[0.50, 1.24]$, does not. The angular form (Fig. 7) supports the same interpretation without division. The wide low-contact Fieller set reflects weak determination of β^{kin}, while the appearance slope remains stable.

5.4 Robustness and controls

Alternative estimates remain within the primary Fieller set. Matched conditional logit gives 1.46, two-stage cluster-robust regression 1.57, and adding joint-angle observations 1.22. Architecture and second-simulator variants (Tao et al., 2025) produce smaller changes than adding proprioception (Table 7). The four second-order drift measures (Lemma 2) reach at most 0.19, below the preregistered threshold 0.25, satisfying the criterion for a panel-wide ratio. Direction symmetry and the single-index null are not rejected (Table 7). Together, these checks support a shared ratio over the tested range at the stated operating point. Figure 4 collects the control and robustness diagnostics.

Matched controls localise the response to robot appearance. Both positive controls change success by more than 30 percentage points. The sham, source-robot masking, material changes in the robot-masked point-cloud arm, and both preregistered scene-perturbation criteria satisfy equivalence

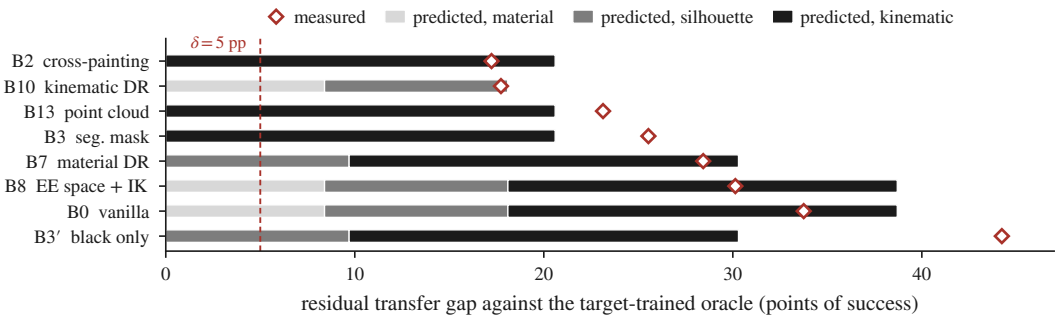


Figure 5: Residual transfer gaps. Bars predict gaps from fitted slopes and unchanged axes without running the corresponding method on the target embodiment; diamonds show measurements. B3′ (black-only masking) removes information instead of randomising it and lies outside the predictor's method class (Table 9).

to zero within ± 5 points. Applying the matched perturbation amplitude to the robot reduces success by 9.8 points, localising the appearance response under the tested perturbations. The two scene controls separately match pixel moments and spatial frequencies, or the response of a frozen encoder unavailable to the policy. Both remain equivalent to zero; the matched robot perturbation has an interval excluding zero (Table 6).

5.5 Predicting residual transfer gaps

Axis coverage predicts the remaining gap. Robot-masked point clouds make the material axis inert by construction and reduce the gap to the target-trained oracle from 33.7 to 23.1 points. Test-time cross-painting closes the most gap among the evaluated appearance methods; kinematic domain randomisation leads the kinematic methods, leaving 17.7 points (Fig. 5). Unadjusted predictions differ from measured gaps by 3.8 percentage points on average. Regressing measured on predicted gaps gives a calibration slope of 0.61 and a residual mean absolute error of 2.4 points: the predictor overestimates the largest gaps.

The fitted slopes define a randomisation boundary. At the fitted kinematic slope, the sufficiency criterion places the boundary at a 5.0% change in link length (Appendix H). Cross-painting approximates K3 with a repaint residual of 5.5 percentage points, within the preregistered intermediate band (Appendix J.3), while retaining a measurable difference from the exact chimera intervention.

Held-out validation supports screening candidate transfers. Sensitivity and specificity are estimated on the held-out skeletons; predictive values are then computed at the declared deployment prevalences of insufficient randomisation because the panel is enriched for large embodiment differences. The negative predictive value stays high across those prevalences. At low prevalence, a flagged pair is less conclusive, so the criterion is most useful for clearing pairs whose predicted residual lies within the margin and selecting the remaining pairs for further evaluation. The full predictive values and skeleton-level intervals are given in Table 4.

6 Conclusion

We make the appearance–kinematics trade-off measurable through four counterfactual interventions. Independently setting collision geometry and silhouette supplies the missing contrasts and, under the stated identification conditions, separates mediation effects in the natural model. The construction protocol turns this design into executable geometry twins, while the local exchange rate expresses transfer sensitivity in percent of link length per appearance JND. The pooled estimate is 1.50, with point estimates of 2.71 and 0.84 at low and high contact precision. The fitted slopes predict residual transfer gaps and an estimated boundary for appearance randomisation. Together, these results provide a way to determine which pathway contributes to transfer loss and how that balance changes with task tolerance.

AI USE STATEMENT

AI assistants supported manuscript organization, reference summarization, inspection of supplied code and experiment records, bilingual drafting and editing, LaTeX preparation, and generation and refinement of the schematic in Fig. 2. The authors are responsible for the final technical content, figures, and references.

REPRODUCIBILITY STATEMENT

The claims in Section 4 are proved in Appendix A from the assumptions tabulated in Appendix B, each of which is marked as holding by construction or as substantive, together with the arm or assertion that checks it. The design is specified at a level intended to be followed rather than paraphrased: the admission predicate and the audit outcome for every candidate model are in Appendix C; the three dose families, the field-level kinematic recipe and the calibration procedure that freezes them are in Appendix D; the invariance assertions that make the central bit-identity claim testable, including what each one covers and the name-based geometry matching, are in Appendix F; the estimators, the Fieller case analysis and the power calculation in independent skeletons are in Appendix G; and Appendix K states what was frozen, when, and the order in which the calibration stages had to run. The embodiment models, the seed lists and the frozen pre-registration files will be released with the camera-ready version.

ETHICS STATEMENT

All robot experiments are conducted in simulation, and the work collects no personally identifying data and deploys nothing on physical hardware. Its one human-subjects component is the psychophysical calibration that gives χ^{phys} its numerator: a two-alternative forced-choice procedure on rendered images, which collects no personal data beyond anonymous responses and is run under the relevant institutional approval. The panel of robot models is drawn from a publicly released, permissively licensed collection, used at a pinned revision and cited as such.

REFERENCES

Haruki Abe, Takayuki Osa, Yusuke Mukuta, and Tatsuya Harada. Cross-embodiment offline reinforcement learning for heterogeneous robot datasets. In *International Conference on Learning Representations*, volume 2026, pp. 21596–21617, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/24a0f56452202e8433d582dbf5095c32-Paper-Conference.pdf.

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Stefan Bauer, and Manuel Wüthrich. CausalWorld: A robotic manipulation benchmark for causal structure and transfer learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=SK7A5pdrgov.

Bo Ai, Liu Dai, Nico Bohlinger, Dichen Li, Tongzhou Mu, Zhanxin Wu, K. Fay, Henrik I. Christensen, Jan Peters, and Hao Su. Towards embodiment scaling laws in robot locomotion. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3483–3515. PMLR, 2025. URL https://proceedings.mlr.press/v305/ai25a.html.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Mächler, and Benjamin M. Bolker. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400, 2017. doi: 10.32614/RJ-2017-066.

Paul-Christian Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008. doi: 10.1162/rest.90.3.414.

Lawrence Yunliang Chen, Karthik Dharmarajan, Kush Hari, Chenfeng Xu, Quan Vuong, and Ken Goldberg. MIRAGE: Cross-embodiment zero-shot policy transfer with cross-painting. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.069.

Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. RoVi-Aug: Robot and viewpoint augmentation for cross-embodiment robot learning. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 209–233. PMLR, 2025. URL https://proceedings.mlr.press/v270/chen25a.html.

Yixiang Chen, Jiabing Yang, Yuan Xu, Qisen Ma, Keji He, Peiyan Li, Kai Wang, Ziheng He, Xiangnan Wu, Jing Liu, Nianfeng Liu, Yan Huang, and Liang Wang. XEWorld: Can action-conditioned world models generalize to unseen robot embodiments? *arXiv preprint arXiv:2608.05799*, 2026. URL https://arxiv.org/abs/2608.05799.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion Policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10–11):1684–1704, 2025. doi: 10.1177/02783649241273668.

Haoran Ding, Liang Ma, Yaxun Yang, Wen Yang, Tianyu Liu, Anqing Duan, Xiaodan Liang, Dezhen Song, Ivan Laptev, and Yoshihiko Nakamura. Choose what to observe: Task-aware semantic-geometric representations for visuomotor policy. *arXiv preprint arXiv:2603.07875v2*, 2026. URL https://arxiv.org/abs/2603.07875v2.

Ioannis Exarchos, Yifeng Jiang, Wenhao Yu, and C. Karen Liu. Policy transfer via kinematic domain randomization and adaptation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 45–51. IEEE, 2021. doi: 10.1109/ICRA48506.2021.9561982.

Arnaud Fickinger, Samuel Cohen, Stuart Russell, and Brandon Amos. Cross-domain imitation learning via optimal transport. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nDITHIIblzh.

Edgar C. Fieller. Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2):175–185, 1954. doi: 10.1111/j.2517-6161.1954.tb00159.x.

Peter Flach and Meelis Kull. Precision-Recall-Gain curves: PR analysis done right. In *Advances in Neural Information Processing Systems*, volume 28, pp. 838–846. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/hash/33e8075e9970de0cfea955afd4644bb2-Abstract.html.

Florian Hartig. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*, 2016. URL https://CRAN.R-project.org/package=DHARMa.

Guanhua Ji, Harsha Polavaram, Lawrence Yunliang Chen, Sandeep Bajamahal, Zehan Ma, Simeon Adebola, Chenfeng Xu, and Ken Goldberg. OXE-AugE: A large-scale robot augmentation of OXE for scaling cross-embodiment policy learning. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2512.13100.

Kuno Kim, Yihong Gu, Jiaming Song, Shengjia Zhao, and Stefano Ermon. Domain adaptive imitation learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5286–5295. PMLR, 2020. URL https://proceedings.mlr.press/v119/kim20c.html.

Marion Lepert, Ria Doshi, and Jeannette Bohg. SHADOW: Leveraging segmentation masks for cross-embodiment policy transfer. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3536–3550. PMLR, 2025. URL https://proceedings.mlr.press/v270/lepert25a.html.

Hao Li, Ganlong Zhao, Yufei Liu, Haotian Hou, Guoquan Ye, Tongyan Fang, Chunxiao Liu, Siyuan Huang, Jianbo Liu, Xiaogang Wang, and Hongsheng Li. ACE-Ego-0: Unifying egocentric human and robotic data for VLA pretraining. *arXiv preprint arXiv:2606.17200*, 2026. URL https://arxiv.org/abs/2606.17200.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941v1*, 2024. URL https://arxiv.org/abs/2405.05941v1.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. MimicGen: A data generation system for scalable robot learning using human demonstrations. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1820–1864. PMLR, 2023. URL https://proceedings.mlr.press/v229/mandlekar23a.html.

Ruth K. Meyer and Christopher J. Nachtsheim. The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, 37(1):60–69, February 1995. doi: 10.1080/00401706.1995.10485889.

Reihaneh Mirjalili, Tobias Jülg, Florian Walter, and Wolfram Burgard. Augmented reality for RObots (ARRO): Pointing visuomotor policies towards visual robustness. *IEEE Robotics and Automation Letters*, 11(4):4785–4792, 2026. doi: 10.1109/LRA.2026.3665444.

Prakhar Mishra, Amir Hossain Raj, Xuesu Xiao, and Dinesh Manocha. MorFiC: Fixing value miscalibration for zero-shot quadruped transfer. *arXiv preprint arXiv:2603.14554*, 2026. URL https://arxiv.org/abs/2603.14554.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024. doi: 10.1109/ICRA57147.2024.10611477.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Meenal Parakh, Alexandre Kirchmeyer, Beining Han, and Jia Deng. AnyBody: A benchmark suite for cross-embodiment manipulation. *arXiv preprint arXiv:2505.14986*, 2025. URL https://arxiv.org/abs/2505.14986.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, second edition, 2009. ISBN 9780521895606. doi: 10.1017/CBO9780511803161.

Michael Piseno, Guy Tevet, and C. Karen Liu. Cloak: Zero-shot cross-embodiment manipulation by masking the end-effector from the VLA. In *10th Annual Conference on Robot Learning*, 2026. URL https://openreview.net/forum?id=7KV78k8rcv.

James E. Pustejovsky and Elizabeth Tipton. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4):672–683, 2018. doi: 10.1080/07350015.2016.1247004.

Thomas S. Richardson and James M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Working Paper 128, Center for Statistics and the Social Sciences, University of Washington, April 2013. URL https://csss.uw.edu/Papers/2013/wp128.pdf.

James M. Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992. doi: 10.1097/00001648-199203000-00013.

I. M. Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967. doi: 10.1016/0041-5553(67)90144-9.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6056–6065. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/suter19a.html.

Kei Suzuki, Jing Liu, Ye Wang, Chiori Hori, Matthew Brand, Diego Romeres, and Toshiaki Koike-Akino. Embedding morphology into transformers for cross-robot policy learning. In *International Conference on Learning Representations (ICLR) Workshop on Efficient Spatial Reasoning*, April 2026. URL https://www.merl.com/publications/TR2026-048.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-Kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Viswesh Nagaswamy Rajesh, Yong Woo Choi, Yen-Ru Chen, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. Demonstrating GPU parallelized robot simulation and rendering for generalizable embodied AI with ManiSkill3. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.021.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Tyler J. VanderWeele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, 2015. ISBN 9780199325870. URL https://global.oup.com/academic/product/explanation-in-causal-inference-9780199325870.

Andrew J. Vickers and Elena B. Elkin. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006. doi: 10.1177/0272989X06295361.

Matthew D. Webb. Reworking wild bootstrap-based inference for clustered errors. *Canadian Journal of Economics/Revue canadienne d'économique*, 56(3):839–858, 2023. doi: 10.1111/caje.12661.

Tong Wu, Shoujie Li, Junhao Gong, Changqing Guo, Xingting Li, Shilong Mu, and Wenbo Ding. CEI: A unified interface for cross-embodiment visuomotor policy learning in 3D space. *IEEE Robotics and Automation Letters*, 11(3):3486–3493, March 2026. doi: 10.1109/LRA.2026.3656802.

Jonathan Yang, Chelsea Finn, and Dorsa Sadigh. Data analogies enable efficient cross-embodiment transfer. *arXiv preprint arXiv:2603.06450*, 2026. URL https://arxiv.org/abs/2603.06450.

Kevin Zakka, Yuval Tassa, and MuJoCo Menagerie Contributors. MuJoCo Menagerie: A collection of high-quality simulation models for MuJoCo, 2022. URL https://github.com/google-deepmind/mujoco_menagerie.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3D Diffusion Policy: Generalizable visuomotor policy learning via simple 3D representations. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.067.

Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. Invariant causal prediction for block MDPs. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11214–11224. PMLR, 2020. URL https://proceedings.mlr.press/v119/zhang20t.html.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.016.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Kevin Lin, Abhiram Maddukuri, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020. URL https://arxiv.org/abs/2009.12293.

A PROOFS

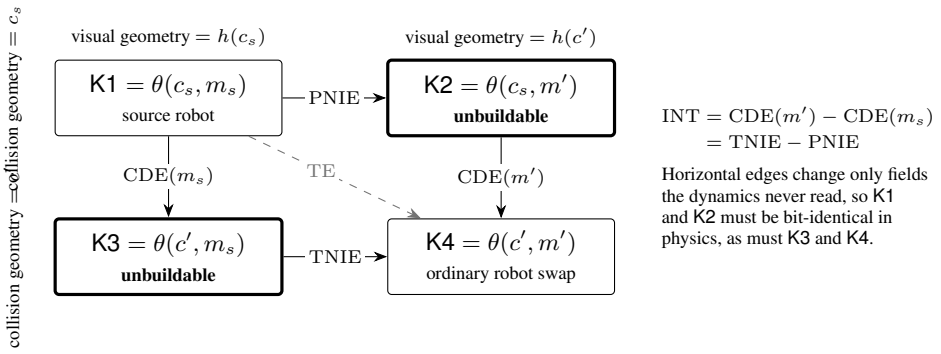


Figure 6: A strict 2×2 design in (c, v) with material fixed at t_s. Only the diagonal cells can be realised on hardware; the off-diagonal cells separate the two axes. Both decompositions sum exactly to TE. Their path contributions depend on intervention order: $\mathrm{TNIE} - \mathrm{PNIE} = \mathrm{CDE}(m') - \mathrm{CDE}(m_s) = \mathrm{INT}$.

Throughout, the simulator is a deterministic map: given a compiled model (c, v, t) and seeds (ς, ε), the recursion Eq. (1) has a unique solution $x_{0:K}(c, v, t; \varsigma, \varepsilon)$, and we *define* the potential outcome as its success value,

$$Y(c, v, t)(\varsigma, \varepsilon) \;:=\; \Sigma_\tau\big(x_{0:K}(c, v, t; \varsigma, \varepsilon),\ \varsigma\big). \tag{7}$$

This is a definition, not an assumption: the potential outcome for a cell we did not run is still a well-defined function, because the compiled model exists and the recursion is deterministic. Expectations are over (ς, ε) only, and $\theta(c, m) := \mathbb{E}[Y(c, m, t_s)]$, the material being held at t_s except where a dose is written explicitly.

Parts (i)–(iv) are statements about $\mathcal{G}^{\mathrm{sim}}$, where Y is the map just defined. The estimands of Section 4.1 are statements about $\mathcal{G}^{\mathrm{nat}}$, and the two are connected by (A4): the structural function generating Y from (c, v, t) is the same in both models, including at arguments off $\mathcal{M}$. (A4) is the price of the whole construction and is not checkable inside the simulator; it is what a reader who rejects simulation would reject, and we prefer to state it as a numbered assumption rather than leave it implicit in the word "twin". Part (v) is a statement about $\mathcal{G}^{\mathrm{nat}}$ alone and does not use it.

A.1 PROOF OF PROPOSITION 1

(i) Twin identification. Let A denote the assignment variable, taking values in the finite set of compiled cells. Three facts hold by construction in $\mathcal{G}^{\mathrm{sim}}$. *Consistency* is exact rather than almost sure: on the event $\{A = (c, v, t)\}$ the executed run solves the same recursion as Eq. (7) with the same seeds, so the observed outcome equals $Y(c, v, t)(\varsigma, \varepsilon)$ pointwise. *Unconfoundedness* holds

because (ς, ε) are drawn from a pre-registered seed list that is fixed before assignment and shared across cells, so $(\varsigma, \varepsilon) \perp A$; there is no latent common cause of A and the seeds, since A is set by the experimenter. *Positivity* holds because the allocation puts positive probability on every cell: $\Pr(A = (c, v, t)) > 0$ for each cell of the design, which is what distinguishes $\mathcal{G}^{\mathrm{sim}}$ from the observational case of part (v), where one sub-block of (C, V) has probability zero. Therefore $\mathbb{E}[Y \mid A = (c, v, t)] = \mathbb{E}[Y(c, v, t) \mid A = (c, v, t)] = \mathbb{E}[Y(c, v, t)]$, the first equality by consistency and the second by unconfoundedness.

The usual fourth condition of mediation analysis, cross-world independence, is not needed here. A cross-world quantity such as $Y(c_s, V_{c'})$ asks for the outcome under one value of the treatment and the mediator value that a *different* treatment value would have produced. Under (A6) the mediator is the deterministic known map h, so $V_{c'} = h(c') = m'$ with probability one and $Y(c_s, V_{c'}) = Y(c_s, m')$, a single-world quantity, which (A5) says is constructible and which we execute.

(ii) The silhouette twin gives the indirect effect. The mediation quantities are defined in $\mathcal{G}^{\mathrm{nat}}$, where the treatment is $C \in \{c_s, c'\}$, the mediator is V and the edge $C \to V$ carries h, so $V_{c_s} = m_s$ and $V_{c'} = m'$ almost surely. Under (A4) each potential outcome below is the one computed in $\mathcal{G}^{\mathrm{sim}}$ by Eq. (7), so

$$\mathrm{PNIE} = \mathbb{E}\big[Y(c_s, V_{c'}) - Y(c_s, V_{c_s})\big] = \theta(c_s, m') - \theta(c_s, m_s), \tag{8}$$

$$\mathrm{PNDE} = \mathbb{E}\big[Y(c', V_{c_s}) - Y(c_s, V_{c_s})\big] = \theta(c', m_s) - \theta(c_s, m_s) = \mathrm{CDE}(m_s), \tag{9}$$

$$\mathrm{TNDE} = \mathbb{E}\big[Y(c', V_{c'}) - Y(c_s, V_{c'})\big] = \theta(c', m') - \theta(c_s, m') = \mathrm{CDE}(m'), \tag{10}$$

$$\mathrm{TNIE} = \mathbb{E}\big[Y(c', V_{c'}) - Y(c', V_{c_s})\big] = \theta(c', m') - \theta(c', m_s). \tag{11}$$

The first line is exactly the operator T_{L2} of Table 2 applied to the source: it moves the mediator to the value the target's skeleton would have induced, while holding the skeleton at the source. It is therefore an indirect effect, and reading it as a direct effect of appearance on capability is a category error. The coincidence of natural and controlled direct effects in the second and third lines is a consequence of the mediator being degenerate: with a point-mass mediator distribution there is nothing to average over, which is also why no cross-world assumption appears.

(iii) A residual appearance channel, and what closes it. Fix $V = m_s$ and $T = t_s$ and consider the directed path

$$C \to x_1 \to o_1 \to a_1 \to x_2 \to \cdots \to x_K \to Y. \tag{12}$$

Every edge is present in Eq. (1): C enters F; R reads x; π_{e_s} reads o; F reads a; and Σ_τ reads $x_{0:K}$. By (A0) the only edge from C into o_k runs through x_k, so Eq. (12) is the whole of the residual channel and not a proper subset of it. The path is unblocked because nothing on it is conditioned on. Changing c changes the configuration trajectory, hence the rendered pixels, hence the actions, even though the silhouette and material fields were held fixed. So $\mathrm{CDE}(m_s)$ is not a pure capability contrast, and the intuition that freezing v and t isolates physics is false.

Now apply $T_{\mathrm{replay}} = \mathrm{do}(C = c', X^{\mathrm{ee}}_{0:K} = \bar{x}^{\mathrm{ee}})$, writing $Y(c, \bar{x}^{\mathrm{ee}})$ for the potential outcome of Eq. (7) at (c, m_s, t_s) with the end-effector path additionally forced to $\bar{x}^{\mathrm{ee}}$: the visual and material fields stay at their source values throughout, so the two displayed arguments are the only ones that move. Under this intervention the action at every step is produced by a tracking controller whose input is $\bar{x}^{\mathrm{ee}}$, an exogenous constant, so the edge $o_k \to a_k$ is removed for all k. Every directed path from V or T to Y passes through some $o_k \to a_k$, because V and T enter the system only through R (assumption (A1)), o_k influences the state only through the policy, and the success criterion does not read the image (assumption (A2)), so no path from V or T reaches Y except through the actions. All such paths are therefore blocked, and

$$\beta^{\mathrm{phys}} = \mathbb{E}_{\varsigma, \varepsilon, \bar{x}^{\mathrm{ee}}}\big[Y(c', \bar{x}^{\mathrm{ee}}) - Y(c_s, \bar{x}^{\mathrm{ee}})\big] \tag{13}$$

contains no appearance channel conditional on $\bar{x}^{\mathrm{ee}}$. The qualification is not removable: $\bar{x}^{\mathrm{ee}}$ is the path achieved by π_{e_s} on the source embodiment, whose law depends on (m_s, t_s), so as a marginal quantity β^{phys} is a functional of the source appearance. We state this instead of suppressing it, and the design carries a reference-trajectory replay arm, whose $\bar{x}^{\mathrm{ee}}$ comes from a fixed pre-planned path instead, as the version of the contrast that does not depend on the source policy.

(iv) Order dependence and the interaction term. Both decompositions are algebraic identities in the four cell means:

$$\begin{aligned}\mathrm{CDE}(m_s) + \mathrm{TNIE} &= \big[\theta(c', m_s) - \theta(c_s, m_s)\big] + \big[\theta(c', m') - \theta(c', m_s)\big] \\ &= \theta(c', m') - \theta(c_s, m_s) = \mathrm{TE}, \\ \mathrm{PNIE} + \mathrm{CDE}(m') &= \big[\theta(c_s, m') - \theta(c_s, m_s)\big] + \big[\theta(c', m') - \theta(c_s, m')\big] \\ &= \theta(c', m') - \theta(c_s, m_s) = \mathrm{TE}.\end{aligned}$$

Subtracting,

$$\mathrm{TNIE} - \mathrm{PNIE} = \theta(c', m') - \theta(c', m_s) - \theta(c_s, m') + \theta(c_s, m_s) = \mathrm{CDE}(m') - \mathrm{CDE}(m_s) = \mathrm{INT}, \tag{14}$$

so both decompositions always sum to TE, and they assign the same value to the direct part, and hence to the indirect part, if and only if $\mathrm{INT} = 0$. The phrase "the kinematic effect, with the appearance effect subtracted out" therefore names a unique quantity only under $\mathrm{INT} = 0$; otherwise it names two different quantities depending on which order the analyst had in mind. Note also that INT is a difference of differences of *means*, so it is defined on the probability scale; the coefficient of a product term in a logit-link model estimates a different functional and can be non-zero when $\mathrm{INT} = 0$ and vice versa.

(v) Non-identification on an observational panel. Let Q be any distribution over embodiments with $Q(\mathcal{M}) = 1$, and let the observed data be (C, V, T, Y) with $(C, V, T) \sim Q$. Then for $m \neq h(c)$ the conditioning event $\{C = c, V = m\}$ has probability zero, so $\mathbb{E}[Y \mid C = c, V = m]$ is not determined by the law of the observed data. Concretely, let θ_1 and θ_2 be any two response surfaces that agree on $\mathcal{M}$ and differ at (c_s, m'). The two induced observational laws are identical, since Q places no mass off $\mathcal{M}$, yet PNIE differs between them by $\theta_1(c_s, m') - \theta_2(c_s, m')$, which may be any value in $[-1, 1]$. Hence PNIE is not a functional of the observational law, and neither is any of $\mathrm{CDE}(m_s)$, $\mathrm{CDE}(m')$, TNIE or INT, each of which involves at least one off-manifold argument. The one contrast whose arguments both lie on $\mathcal{M}$ is $\theta(c', h(c')) - \theta(c_s, h(c_s)) = \mathrm{TE}$, and even that is identified only under the further condition that the panel's embodiment assignment is unconfounded with the outcome, which an observational panel does not supply by construction. Enlarging the panel enlarges the support of Q_C but leaves $Q(\mathcal{M}) = 1$ intact, so no amount of additional real hardware repairs the failure. □

A.2 Properties of the exchange rate

Each part below fixes a condition under which a single rate may be quoted, and Section 4.4 says which. Throughout, Sr is the local response surface of Section 4.4 and $\chi = \beta^{\mathrm{app}}/\beta^{\mathrm{kin}}$ its marginal rate of substitution.

Lemma 1 (Properties of χ)**.** *Let* $\mathrm{Sr} \in C^2(\mathcal{D})$ *and* $d_0 \in \mathcal{D}$.

(a) If $\beta^{\mathrm{kin}}(d_0) \neq 0$ *there is a unique* C^1 *curve* ψ *with* $\mathrm{Sr}(d_a, \psi(d_a)) \equiv \mathrm{Sr}(d_0)$ *near* d_0*, and* $\psi' = -\chi$.

(b) $\chi \equiv \lambda$ *on* $\mathcal{D}$ *if and only if* $\mathrm{Sr}(d_a, d_k) = H(\lambda d_a + d_k)$ *for some* $H \in C^1$ *with* $H' \neq 0$.

(c) Under smooth strictly increasing reparameterisations of the axes, $\tilde{\chi} = \chi \cdot \varphi_k'/\varphi_a'$*; for linear standardisation,* $\tilde{\chi} = \chi \cdot \sigma_{\mathrm{app}}/\sigma_{\mathrm{kin}}$.

(a). $\mathrm{Sr} \in C^2$ and $\partial_{d_k}\mathrm{Sr}(d_0) = \beta^{\mathrm{kin}}(d_0) \neq 0$, so the implicit function theorem applies to $\Psi(d_a, d_k) = \mathrm{Sr}(d_a, d_k) - \mathrm{Sr}(d_0)$ at d_0: there is a neighbourhood of d_a^0 and a unique C^1 function ψ with $\psi(d_a^0) = d_k^0$ and $\Psi(d_a, \psi(d_a)) \equiv 0$. Differentiating, $\partial_{d_a}\mathrm{Sr} + \partial_{d_k}\mathrm{Sr} \cdot \psi' = 0$, hence $\psi' = -\beta^{\mathrm{app}}/\beta^{\mathrm{kin}} = -\chi$. Where $\beta^{\mathrm{kin}} = 0$ the level set has a vertical tangent and χ is undefined; this is the case in which the Fieller construction returns an unbounded set, and it is why we also report ϑ.

(b). If $\mathrm{Sr} = H(\lambda d_a + d_k)$ with $H' \neq 0$ then $\partial_{d_a}\mathrm{Sr} = \lambda H'$ and $\partial_{d_k}\mathrm{Sr} = H'$, so $\chi \equiv \lambda$. Conversely suppose $\chi \equiv \lambda$ on $\mathcal{D}$, which by definition of χ presupposes $\partial_{d_k}\mathrm{Sr} \neq 0$ there. Then $\partial_{d_a}\mathrm{Sr} - \lambda\partial_{d_k}\mathrm{Sr} = 0$, that is, the directional derivative of Sr along the constant field $w = (1, -\lambda)$

vanishes identically. Since $\mathcal{D}$ is open and convex, each line segment $\{p + sw\} \cap \mathcal{D}$ is connected, so Sr is constant on it. The linear functional $\ell(d_a, d_k) = \lambda d_a + d_k$ satisfies $\ell(p + sw) = \ell(p)$ and has $\ell(\mathcal{D})$ an interval, and its level sets in $\mathcal{D}$ are exactly those segments. Hence Sr factors as $\mathrm{Sr} = H \circ \ell$ for a well-defined H on that interval, and $H' = \partial_{d_k} \mathrm{Sr} \neq 0$, so $H \in C^1$.

(c). Let $\tilde{d}_a = \varphi_a(d_a)$ and $\tilde{d}_k = \varphi_k(d_k)$ with φ_a, φ_k smooth and strictly increasing, and let $\widetilde{\mathrm{Sr}}(\tilde{d}_a, \tilde{d}_k) = \mathrm{Sr}(\varphi_a^{-1}(\tilde{d}_a), \varphi_k^{-1}(\tilde{d}_k))$. By the chain rule $\partial_{\tilde{d}_a} \widetilde{\mathrm{Sr}} = \partial_{d_a} \mathrm{Sr}/\varphi_a'$ and $\partial_{\tilde{d}_k} \widetilde{\mathrm{Sr}} = \partial_{d_k} \mathrm{Sr}/\varphi_k'$, so $\tilde{\chi} = \chi \cdot \varphi_k'/\varphi_a'$; for the linear standardisation $\varphi_a(u) = u/\sigma_{\mathrm{app}}$, $\varphi_k(u) = u/\sigma_{\mathrm{kin}}$ this is $\tilde{\chi} = \chi \cdot \sigma_{\mathrm{app}}/\sigma_{\mathrm{kin}}$. If a unit change $d_a \mapsto \alpha d_a$ is accompanied by recomputing σ_{app} on the *same* reference population, then χ is divided and σ_{app} multiplied by α, and $\tilde{\chi}$ is unchanged; so $\tilde{\chi}$ is invariant under any positive diagonal rescaling provided the scales are recomputed on a fixed $\mathcal{Q}$, and is not invariant to a change of $\mathcal{Q}$. Since the scales are otherwise estimated on whichever population the analyst supplies, $\tilde{\chi}$ is a property of the pair (response surface, reference population): widening the kinematic amplitude of the panel by a factor κ multiplies σ_{kin} by approximately κ and divides $\tilde{\chi}$ by approximately κ. This is why $\mathcal{Q}$ is pinned from outside, to the Menagerie arms together with the embodiments in Open X-Embodiment (O'Neill et al., 2024), and why the headline number is χ^{phys}, whose units are fixed by the design rather than by whichever panel happens to have been run. □

A.3 Curvature and the four drift numbers

Lemma 2 (Second-order behaviour of χ). *With* $\mathrm{Sr} \in C^2$ *and* $d_0 = (0, 0)$, $\mathrm{Sr} = \kappa_0 + \beta^{\mathrm{app}} d_a + \beta^{\mathrm{kin}} d_k + \frac{1}{2}\gamma_{aa} d_a^2 + \gamma\, d_a d_k + \frac{1}{2}\gamma_{kk} d_k^2 + O(\|d\|^3)$ *and*

$$\chi(d_a, d_k) = \frac{\beta^{\mathrm{app}} + \gamma_{aa} d_a + \gamma d_k}{\beta^{\mathrm{kin}} + \gamma d_a + \gamma_{kk} d_k} + O(\|d\|^2). \tag{15}$$

Over a panel of half-widths (D_a, D_k),

$$\frac{\chi(0, D_k)}{\chi(0,0)} \approx \frac{1 + \gamma D_k/\beta^{\mathrm{app}}}{1 + \gamma_{kk} D_k/\beta^{\mathrm{kin}}}, \qquad \frac{\chi(D_a, 0)}{\chi(0,0)} \approx \frac{1 + \gamma_{aa} D_a/\beta^{\mathrm{app}}}{1 + \gamma D_a/\beta^{\mathrm{kin}}}, \tag{16}$$

so the drift of χ *away from* d_0 *is governed by the four dimensionless quantities* $|\gamma D_k/\beta^{\mathrm{app}}|$, $|\gamma_{kk} D_k/\beta^{\mathrm{kin}}|$, $|\gamma_{aa} D_a/\beta^{\mathrm{app}}|$ *and* $|\gamma D_a/\beta^{\mathrm{kin}}|$.

Proof. Taylor expansion about d_0 with $\gamma_{aa} = \partial^2_{d_a} \mathrm{Sr}$, $\gamma = \partial_{d_a}\partial_{d_k} \mathrm{Sr}$ and $\gamma_{kk} = \partial^2_{d_k} \mathrm{Sr}$ evaluated at d_0. Differentiating gives $\partial_{d_a} \mathrm{Sr} = \beta^{\mathrm{app}} + \gamma_{aa} d_a + \gamma d_k + O(\|d\|^2)$ and $\partial_{d_k} \mathrm{Sr} = \beta^{\mathrm{kin}} + \gamma d_a + \gamma_{kk} d_k + O(\|d\|^2)$; taking the ratio and evaluating at $(0, D_k)$ and $(D_a, 0)$ gives the two displays. □

All four numbers are reported. A value near 0.1 means χ is trustworthy across the panel to roughly ten percent; a value near 1 means the panel is too wide for a single number to describe, and ablation A3 fits the full quadratic so that the reader can see which of the two situations holds.

A.4 The working point

The requirement that success rates lie in $[0.2, 0.8]$ is often justified by appeal to a $0/0$ that does not occur. Under the single-index model of Lemma 1(b) with $\mathrm{Sr} = H(\lambda d_a + d_k)$, saturation means $H' \to 0$, so $\beta^{\mathrm{app}} = \lambda H'$ and $\beta^{\mathrm{kin}} = H'$ vanish together and their ratio remains exactly λ. The estimand does not degenerate; its estimator does. The argument is on the link scale, which is where Eq. (19) is fitted: in a generalised linear model with mean p and canonical logit link, the Fisher information for the coefficient vector is $\sum_i p_i(1 - p_i) x_i x_i^\top$, so the variance of each link-scale slope estimate grows like $1/\{p(1-p)\}$. By Appendix G.3 the Fieller set for $\chi = \beta^{\mathrm{app}}/\beta^{\mathrm{kin}}$ is bounded if and only if $\hat{\beta}_k^2 > z^{*2} \widehat{\mathrm{Var}}(\hat{\beta}_k)$, that is, only if the denominator is significantly non-zero at the chosen level. As $p \to 0$ or $p \to 1$ this inequality fails at any fixed sample size. Reading the same statement on the probability scale is not the argument: there the slopes themselves shrink with $p(1-p)$, and it is the signal, not the variance, that disappears. The working point is thus an information requirement on the estimator, not a regularity requirement on the estimand, and the calibration stage that enforces it is part of the design rather than a post-hoc filter.

A.5 DEFICIT DECOMPOSITION

Remark 1 (Deficit decomposition). Let Π be a pre-registered privileged-state planner that reads simulator state and never images, $\mathrm{SR}^{\mathrm{plan}}(c) = \mathrm{SR}_\tau(\Pi; e)$, and $\mathrm{SR}^\star_{\mathrm{obs}}(e) = \sup_{\pi:\mathcal{O}^*\to\mathcal{A}} \mathrm{SR}_\tau(\pi; e)$ over history-dependent policies. Then $\mathrm{SR}^{\mathrm{plan}}$ depends on c alone and is invariant to T_{L1} and T_{L2} by construction; for every observation-only transform Φ, $\mathrm{SR}_\tau(\pi_{e_s}\circ\Phi; e') \le \mathrm{SR}^\star_{\mathrm{obs}}(e')$; and

$$1 - \mathrm{SR}_\tau(\pi_{e_s}; e') = \underbrace{\big(1 - \mathrm{SR}^{\mathrm{plan}}(c')\big)}_{\text{capability deficit}} + \underbrace{\big(\mathrm{SR}^{\mathrm{plan}}(c') - \mathrm{SR}^\star_{\mathrm{obs}}(e')\big)}_{\text{observability deficit}} + \underbrace{\big(\mathrm{SR}^\star_{\mathrm{obs}}(e') - \mathrm{SR}_\tau(\pi_{e_s}; e')\big)}_{\text{policy deficit}}. \tag{17}$$

The identity is a telescoping sum of the three bracketed differences and holds for any three real numbers. Two claims in it are substantive. First, $\mathrm{SR}^{\mathrm{plan}}$ depends on c alone: Π reads simulator state and never reads images, so its input distribution is a functional of F, which by Eq. (1) reads c but not v or t; hence $\mathrm{SR}^{\mathrm{plan}}$ is invariant under T_{L1} and T_{L2} by construction, and any change in it under those operators indicates an implementation fault and not an effect. Second, for any observation-only transform Φ, the composition $\pi_{e_s} \circ \Phi$ is itself a map from observation histories to actions, so it belongs to the class over which $\mathrm{SR}^\star_{\mathrm{obs}}$ takes its supremum and $\mathrm{SR}_\tau(\pi_{e_s} \circ \Phi; e') \le \mathrm{SR}^\star_{\mathrm{obs}}(e')$. A randomised policy poses no difficulty: its value is a mixture over the values of the deterministic maps it selects among, hence at most the supremum of those values, so writing $\mathrm{SR}^\star_{\mathrm{obs}}$ as a supremum over deterministic maps loses nothing. The supremum must range over history-dependent policies: success criteria of the form "was lifted at some time and is placed at the final time" are functionals of the trajectory, and the corresponding optimal control problem is not in general solved by any memoryless policy, so a supremum over memoryless policies would not be an upper bound on what a sequence model can do. Finally, $\mathrm{SR}^{\mathrm{plan}}(c')$ is the value achieved by one frozen planner, not a supremum over state-feedback policies, so it may be exceeded by an image-based policy on particular tasks and the middle term of Eq. (17) may be negative. We therefore report the three terms as a decomposition and not as a budget, and avoid the word "ceiling". □

B ASSUMPTIONS AND NOTATION

B.1 THE INTERVENTION OPERATORS

Table 2: The five operators of Section 3. "Buildable" refers to physical hardware; all five are realisable in the simulator, the first four as edits to a model's fields and the last as a controller-level intervention.

Operator	Definition	Changes	Buildable
T_{L1}	$\mathrm{do}(T = t')$	material t	yes
T_{L2}	$\mathrm{do}(V = m')$ with $C = c_s$	silhouette v only	**no**
$T^{\mathrm{sync}}_{\mathrm{kin}}$	$\mathrm{do}(C = c',\ V = h(c'))$	c and v together	yes
$T^{\mathrm{froz}}_{\mathrm{kin}}$	$\mathrm{do}(C = c',\ V = m_s)$	c only, v held at source	**no**
T_{replay}	$\mathrm{do}(C = c',\ X^{\mathrm{ee}}_{0:K} = \bar{x}^{\mathrm{ee}})$	c, source end-effector path forced	yes

B.2 THE ASSUMPTIONS, WITH THEIR STATUS

The status column distinguishes substantive assumptions from conditions that are *verified* by construction. The fourth column lists the checks for each condition.

Of the three substantive assumptions, (A0) is most likely to fail in other deployments. Joint angles give a policy kinematic information through a channel unaffected by appearance interventions. Ablation A2 in Appendix E measures the contribution of that channel on our panel. Section 4.2 says why common random numbers are absent from the table.

Table 3: The assumptions of the design. Proposition 1 uses (A0)–(A2), (A4) and (A5)–(A6); (A0-base), (A3) and (A7) support estimation and reporting rather than identification.

	Statement	Status	Checked by
(A0)	The action space and the proprioceptive channel have fixed dimension and semantics across the family, so c has no edge into o_k except through x_k.	substantive	ablation A2 (proprioception on/off)
(A0-base)	Base placement is not a function of c.	enforced by design	the placement rule of Appendix C, plus a sensitivity arm with a fixed base
(A1)	The renderer reads only (v, t, x, ς).	by construction	Appendix F, assertion 4
(A2)	The success criterion reads only x.	by construction	task definitions; no image enters Σ_τ
(A3)	Episodes are independent given the compiled model.	substantive	fresh simulator instance per episode; no state carry-over
(A4)	The structural function generating Y from (c, v, t) is the same in $\mathcal{G}^{\mathrm{nat}}$ and $\mathcal{G}^{\mathrm{sim}}$, including at arguments off $\mathcal{M}$.	substantive	not testable within the simulator; this is a scope condition of Proposition 1 and motivates the proposed hardware test of the material axis (Appendix M.1)
(A5)	Both chimeras compile and run.	by construction	the build stage reports an error otherwise
(A6)	h is deterministic and known.	by construction	procedural shells (Section 4.3)
(A7)	The version of the material dose family is declared with every estimate.	bookkeeping	recorded with every estimate

B.3 Notation

$e = (c, v, t)$	embodiment: collision, visual, material	d^{app}	material distance
h	outer-surface map, $v = h(c)$	d^{L2}	silhouette distance
$\mathcal{M}$	$\{(c, v, t) : v = h(c)\}$	d^{kin}	kinematic distance
m_s, m'	$h(c_s)$, $h(c')$	$\beta^{\mathrm{app}}, \beta^{\mathrm{L2}}, \beta^{\mathrm{kin}}$	slopes of SR in the three distances
$\theta(c, m)$	$\mathbb{E}[Y(c, m, t_s)]$	χ_τ	$\beta^{\mathrm{app}}/\beta^{\mathrm{kin}}$
K1..K4	the four cells of Fig. 6	χ^{phys}	Eq. (6)
ς, ε	scene seed, policy and solver seed	ϑ	$\mathrm{atan2}(\beta^{\mathrm{app}}, \beta^{\mathrm{kin}})$
τ^{prec}	measured contact precision	$\mathcal{Q}$	external reference population
RPF, OPF	robot, object pixel fraction	G	number of independent skeletons
K, N_{ep}	episode horizon; episodes per condition	Π	privileged-state planner

C The embodiment panel

C.1 The admission predicate

A MuJoCo Menagerie model is admitted if, for every moving link on the serial chain that the kinematic intervention will touch, both of the following hold.

(P1) The visual and collision roles are decidable. The link must ship both contact geometries and non-contact geometries, with no geometry serving both roles. The collision bitmasks in the model file distinguish them: a geometry with either bitmask non-zero is a collision geometry, one with both zero is a visual geometry. If a link's only geometry collides, the silhouette operator would also modify the contact surface and therefore change the physics.

(P2) An equal-density inertia is derivable from collision geometry alone. The kinematic recipe discards each link's declared inertial properties so that the compiler recomputes them under a single density. This makes inertia a consequence of geometry and keeps it consistent with geometry inter-

ventions. If a link has a visual mesh and no collision geometry, the compiler falls back on the visual mesh, and then editing the silhouette would change the link's mass. We reject such models even if they meet the other requirements.

Applied at a pinned Menagerie revision, the predicate audits 23 arms and admits 16: eight need no repair, seven need a documented repair (five a per-link split of a fused mesh, two a burned-in inertia), and one, the TidyBot arm, meets the geometry requirements but cannot be mounted on our bench because of a limitation of the rig. Seven are rejected: the xArm7, whose every link geometry defaults to colliding, violates (P1); the Unitree Z1 has two links with a visual mesh and no collision primitive and violates (P2), as does the i2rt YAM; Stretch and Stretch 3 are mobile bases whose chain walk finds several moving children at the root, so there is no serial arm to audit; the remaining two fail (P1) for the same reason as the xArm7.

The final panel contains 15 mountable skeletons: 12 for training and three held out for evaluation, the Sawyer, ARX L5 and UR10e. Confidence intervals in Appendix G use the skeleton as the independent unit. The original power analysis specified 16 skeletons; the available mountable set reduced the hold-out from four to three while preserving the 12 training skeletons. UR10e shares a design family with a training skeleton, so its hold-out result evaluates an unseen skeleton within that family. The hold-out is therefore not fully independent at the design-family level. External validity on physical hardware requires separate evaluation.

C.2 Two-stage selection

Distances are outputs of an irreversible map, so a panel cannot be sampled directly in distance space. Stage one draws $10n$ candidate kinematic variants per skeleton from a Sobol sequence (Sobol', 1967) in *parameter* space, where n is the number of scaled links, and computes all three distances for every candidate; this is forward kinematics, reach sampling and offline rendering, with no rollouts, so oversampling is cheap. Stage two selects the panel by maximin coordinate exchange (Meyer & Nachtsheim, 1995) on the standardised distance vectors, subject to $|r(d^{\mathrm{app}}, d^{\mathrm{kin}})| < 0.2$ and $|r(d^{\mathrm{L2}}, d^{\mathrm{kin}})| < 0.4$, with at most four variants per skeleton. The selected panel satisfies both constraints with a margin, achieving $r(d^{\mathrm{app}}, d^{\mathrm{kin}}) = 0.11$ and $r(d^{\mathrm{L2}}, d^{\mathrm{kin}}) = 0.31$, with a largest variance inflation factor of 1.42 across the three distance columns. These three values allow readers to assess axis separation without rerunning the design.

The de-correlation constraint applies to the silhouette-dose and panel-regression rows only. Applying it to K1 through K4 would conflict with the definitions of those cells. Chimeras are constructed *within* a skeleton, with c' an axial-scaling variant of c_s, so the link correspondence used by the silhouette operator is the identity map and no cross-robot mesh transplant is required. Cross-skeleton chimeras are declared as an external-validity extension, with the link correspondence written down and frozen in advance.

C.3 Base placement

The base transform is solved so that the manipulability profile of the target over the task workspace matches the source's in log ratio, subject to the reachability of every task keypoint. We solve for placement before making any optical adjustments. Cells that share a skeleton share a base by construction, so no cell contrast can be attributed to placement. A sensitivity arm holds the base at the source's transform for all embodiments, trading matched manipulability for exact positional identity; it is pre-registered and runs alongside the main grid.

D Dose families and the kinematic intervention

D.1 Why every axis is calibrated once and then frozen

The nominal knob of an intervention and the distance it produces are related non-linearly, and in our pilot sweeps the relation differs across models by a factor of two to three: the same LAB displacement can land two robots at very different measured d^{app}, because their base materials differ. A slope fitted against the nominal knob therefore carries a skeleton effect. Two repairs are available and only one is correct. Re-tuning the knob per embodiment so that the measured distance matches

a target makes the treatment intensity a function of the embodiment. Instead we run one closed loop per axis on the *source* skeleton alone, freeze one shared parameter table, and regress on the measured distance:

the amplitude of one pre-registered direction is swept on the source skeleton over a frozen grid, a monotone response from amplitude to measured distance is fitted to that sweep, and the fit is inverted at the 25th, 50th, 75th and 100th percentiles of the range the sweep covered, each inverted amplitude being re-measured to verify the inversion. What survives is one frozen amplitude per pair of direction and dose. The lookup takes no embodiment argument, so treatment intensity cannot be a function of the embodiment, and nothing downstream reads the nominal amplitude. The loop is run once per axis, kinematic as well as material, and the kinematic one is not optional. β^{kin} is the *denominator* of χ; a Fieller interval is bounded only when its denominator is significantly non-zero (Appendix G.3); and a denominator diluted by a skeleton effect is exactly how that condition fails. Calibrating the appearance axis and leaving the kinematic axis on its nominal knob would buy a tidy numerator at the cost of the one quantity the interval depends on.

D.2 Material

Doses interpolate each link's mean colour in CIE-LAB along six pre-registered directions ($\pm L^*$, $\pm a^*$, $\pm b^*$) at $q \in \{0.25, 0.5, 0.75, 1.0\}$, with at least four independent draws per level so that the level effect is not one arbitrary colour. Specular, shininess and reflectance are frozen at the source values, so the intervention is chromatic and not photometric. The primary measured distance is the cosine distance between mask-weighted DINOv2 patch tokens; a LAB chroma histogram χ^2 and a Gabor texture contrast are recorded alongside as interpretable low-level descriptors and are used only descriptively.

The JND scale. The numerator of χ^{phys} is calibrated by a two-alternative forced-choice procedure that is independent of any policy: rendered pairs at a range of amplitudes are presented, the proportion correct is fitted with a psychometric function, and one just-noticeable difference is the amplitude at which discrimination reaches 75%. This anchors the appearance axis in a unit that does not move when we change encoders, and it is the reason χ^{phys} can be compared across papers that use different appearance features.

D.3 Silhouette

The silhouette ladder replaces each link's visual geometry by successively coarser proxies: full mesh, convex hull, capsule fit, box fit. Coarsening changes apparent size, so each rung is size-matched until the mean robot pixel fraction on oracle trajectories is within 2% of the full-mesh rung. Collision geometry is untouched at every rung, which Appendix F verifies. d^{L2} is measured on the rendered silhouettes as the mean of $1 - \mathrm{IoU}$ and a symmetric contour Chamfer distance over a matched pose set. Shape moments are kept out of the material distance, so that a coarsened rung does not register twice; but the separation the design actually relies on is the one the interventions provide, not the one the descriptors promise, since a mask-weighted encoder read inside a silhouette that has just changed will move whatever features it is built from.

D.4 Kinematics

Reparameterising away the dominant confounder. Scaling link lengths changes how much of the image the robot occupies, and the robot pixel fraction is the largest confounder in the design. Writing the per-link log scales as $\log s \in \mathbb{R}^n$, we split them in a Helmert basis B of $\mathbf{1}^{\perp}$,

$$\log s = a\mathbf{1} + Bz, \qquad a = \tfrac{1}{n}\textstyle\sum_i \log s_i, \tag{18}$$

so that a is the projected-extent component and z the shape component. We constrain $|a| \le 0.03$, which puts the predicted camera-distance factor in $[0.970, 1.030]$, comfortably inside the window the pixel-fraction matcher can absorb, and leave z free to reach the $\pm 30\%$ per-link amplitude the power analysis needs. Because the kinematic distance reads link lengths, joint travel and torque ceilings, all of which move under z, the axis the paper needs is the one that is now unconstrained, while the axis that drives the confounder is pinned near unity. The Helmert basis is chosen because it is written in closed form, so two machines cannot disagree about a column's sign; the split is

applied *after* the Sobol draw, not by drawing the two blocks separately, which keeps the sequence dimension equal to the number of degrees of freedom and preserves its balance properties on the axes stage one explores.

Which distance the dose grid is inverted on. Five plugins compute d^{kin}: a reach-set Chamfer distance under per-set-reach normalisation, a directional Chamfer distance on point–normal pairs, an attributed Fused Gromov–Wasserstein distance between kinematic trees (Abe et al., 2026), a cosine distance between morphology parameter vectors, and a normalised link-scale distance. All five are descriptive, since the regressor of Eq. (19) is the design knob; the one place a single metric has to be nominated is the dose grid, which cannot be inverted on five at once. The inversion uses the reach-set Chamfer distance under per-set-reach normalisation. It is nominated rather than averaged with the other plugins because per-set-reach normalisation is the only one that removes overall scale outright, which keeps the calibration from tracking "this arm is bigger", the very quantity the extent constraint pins near unity. The directional Chamfer distance on point–normal pairs is measured at every swept amplitude as a second opinion, and its rank agreement with the inversion metric is checked: a family calibrated on a metric the others disagree with would make the frozen amplitudes plugin-dependent. The morphology-parameter cosine cannot serve here, because its scaler is a root-mean-square over a population and the only population available during calibration is one arm at twenty amplitudes, on which two feature coordinates have exactly zero spread.

What the kinematic operator edits. The operator moves link position offsets; the size and placement of collision geometry; visual mesh scales; joint travel; actuator force and control ranges; and wrist-camera mounting frames. Inertia is not edited but *derived*: each link's declared inertial properties are discarded and the compiler recomputes mass and the inertia tensor from collision geometry at a single density, so the equal-density rule is enforced by the compiler and not asserted by us.

The scaling itself acts on each link's *long axis* only, with the cross-section frozen, which is what makes the torque rule come out the way it does. Under equal density, a link whose length scales by s at fixed cross-section has mass proportional to s and its centre of mass at a distance proportional to s, so the gravitational load about the joint proximal to it grows as s^2. Scaling torque limits as s^2 therefore holds the normalised gravity-load margin fixed. The exponent is specific to this parameterisation and we state it rather than inherit it: had the cross-section scaled with the length, mass would grow as s^3, the load as s^4, and an s^2 limit would leave the larger arms systematically weaker relative to their own weight, which is exactly the kind of confound a kinematic dose family exists to avoid. Operational-space controller gains are renormalised at the nominal configuration to critical damping against the task-space inertia $M_{\mathrm{task}} = (JM^{-1}J^{\top})^{-1}$. Two pre-registered sensitivity arms retain fixed gains and frozen inertia respectively, so that neither normalisation is load-bearing.

E ARMS, CONTROLS AND ABLATIONS

E.1 ARMS

Arm	Operator	What it is for
K1–K4	T_{L2}, $T_{\mathrm{kin}}^{\mathrm{froz}}$, $T_{\mathrm{kin}}^{\mathrm{sync}}$	the strict 2×2 of Section 4.1
K1′–K4′	the same, plus T_{L1}	the material dose axis crossed with the cells
L2dose	T_{L2} at intermediate rungs	estimates β^{L2}; deliberately not K2, whose mediator value must be exactly $h(c')$
R	T_{replay}	β^{phys}: the contrast with every path through the image blocked
$\mathrm{R}_{\mathrm{ref}}$	T_{replay} on a pre-planned path	the same contrast with a reference trajectory that does not depend on the source policy
F	$\mathrm{do}(O_{0:K} = \bar{o})$	separates the configuration-induced appearance channel from the execution gap, Eq. (5)
O	target-trained policy	the achievable upper reference per target embodiment
Π	privileged planner	$\mathrm{SR}^{\mathrm{plan}}$ in Remark 1; reads state, never images
S	rollout seed stream only	the sham: the pipeline's own noise floor
P-a, P-b	action permutation; 90° camera rotation	positive controls; each must move success by ≥ 30 pp with an interval excluding zero

E.2 NEGATIVE CONTROLS

N1, masking on the source embodiment. The source robot's pixels are replaced by its own segmentation mask, and the policy is trained and evaluated on its own robot. Masking must cost nothing here, as it does not in Lepert et al. (2025); if it does, every masked arm downstream is measuring a damaged policy rather than an appearance effect. Note what this control does *not* test. Whether the appearance axis has dynamic range under transfer is settled elsewhere, by the black-only baseline B3′ and by the robot-side reference row of Table 6, and those are the rows to read for it.

N2, equal-amplitude scene perturbation. A perturbation of the *scene* rather than the robot, matched in amplitude to a robot-appearance perturbation under two independently pre-registered criteria: first-and-second pixel moments together with a radially averaged power spectrum, and cosine distance in a frozen encoder the policy never sees. Both criteria are run, because a single one would leave the match a choice of the analyst. If matched scene noise moves success as much as matched robot-appearance change, what we are measuring is sensitivity to image perturbation in general and not to the robot's appearance in particular.

N3, occlusion-amount matching. The object's visible pixel fraction is matched at *design* time. Post-hoc matching would condition on a post-treatment variable, since occlusion is a consequence of the kinematic intervention, and would bias the slope it was intended to protect.

N4, robot-masked point cloud. A depth policy that never sees robot appearance. The material axis must be inert here by construction, so a non-zero material slope in this arm is a pipeline fault rather than a finding.

E.3 ABLATIONS

A1, direction symmetry. Is $\Delta(e_s \to e)$ equal to $\Delta(e \to e_s)$? Asymmetry means the slopes are properties of the source as well as of the distance. **A2, proprioception.** The direct test of (A0): a policy reading joint angles receives kinematic information through a channel no appearance intervention touches, and this ablation bounds that leak. The primary analysis is pre-registered on the proprioception-free configuration. **A3, curvature.** The full quadratic of Lemma 2, from which the four drift numbers are read. **A4, single index.** A direct test of Lemma 1(b): if the single-index fit is rejected, one χ may not be reported for the whole panel. **A5, the Fused Gromov–Wasserstein trade-off parameter.** Substitutes one plugin for the design knob in Eq. (19) and sweeps its structure-versus-attribute weight, so that the spread it induces in χ^{phys} can be set against the primary interval. **A6, architecture.** Diffusion policy against ACT, to test whether χ is a property of the task or of the policy class. **A7, second simulator.** Replication in ManiSkill3, which tests whether the result is a property of MuJoCo's renderer and contact model.

F BIT-LEVEL INVARIANCE ASSERTIONS

The central structural claim of Section 4.1 is that K1 and K2 differ only in fields the dynamics never read, and likewise K3 and K4. In prose it is an appeal to how MuJoCo is documented to behave. As a test it is a fact about the two models the build stage actually compiled. Every pair is put through the following five assertions before a single rollout starts.

Assertion 1: the physics arrays are bit-identical. Twenty-five compiled arrays must agree element by element, with no tolerance. They cover everything the integrator reads about the body of the robot: each link's mass, inertia tensor and centre-of-mass frame; the pose of every body in its parent's frame; each joint's axis, anchor, travel and torque range, together with whether it is limited at all; the damping, armature inertia and friction loss on every degree of freedom; the force and control ranges, gain and bias parameters, and gear ratio of every actuator; the type, operands and parameters of every equality constraint; and the rest length and stiffness of every tendon. Six scalar solver options must also agree: the integrator and its timestep, the iteration cap, and the three medium parameters of the fluid model. The body pose fields matter most and are the easiest to leave out of a list like this, because nothing about them looks like a contact parameter. They are exactly what a kinematic intervention moves, so omitting them would let a kinematic change pass as an appearance twin.

Assertion 2: contact geometry agrees, matched by name. Twelve properties of every colliding geometry must agree: its primitive type, size, position and orientation; its friction coefficients, contact dimensionality and the two solver-reference parameter sets that set contact stiffness and damping; its margin and gap; and its priority and mixing weight. The last two are included because they decide how contact parameters are combined between the two geometries of a contact pair, so two models agreeing on everything else would still make different contacts without them. The comparison is by geometry *name*, not by index. Changing the visual geometry changes the geometry count, and therefore the index of every collision geometry after the edit; an index-wise comparison would compare a link's shoulder against another link's elbow and pass or fail for reasons unrelated to the intervention.

Assertion 3: the visual geometry is physically inert. Field equality is not enough on its own, because a geometry that is excluded from contact by its collision bitmasks can still reach the dynamics. Every route by which it could is asserted closed rather than assumed closed: no visual geometry may appear in an explicit contact pair or be referenced by a rangefinder sensor, the fluid model must be switched off in all three of its medium parameters, and no ellipsoid fluid coefficient may be set on any body. The last two hold by MuJoCo's defaults, which is precisely why they are asserted: the ellipsoid fluid model sizes drag from a geometry's dimensions and never consults the collision bitmasks.

Assertion 4: the rendered difference is confined to the robot. For a fixed pose set and fixed seeds, a material intervention must leave the pixels outside the robot mask identical to its base, and must leave depth and segmentation bit-identical everywhere. This is the testable form of (A1), and it catches the failure modes that field comparison cannot: a light attached to the robot body, a shadow-casting flag that moved with the visual geometry, a skybox reference that changed with the material. A pair that fails is dropped, and because high-reflectance materials are both the likeliest to fail and the ones with the largest d^{app}, dropping would bias β^{app} downwards; freezing reflectance and specular across the material family (Appendix D) is what keeps the check from becoming a selection mechanism.

Assertion 5: the reach sets coincide. The sampled reach set of a twin and its base must agree at Chamfer distance exactly zero. Since the sampler is deterministic given the compiled model and its seed, this holds by construction once assertions 1 and 2 pass. The redundancy is deliberate, because the assertion exercises the reach-set pipeline itself: a non-zero answer with the first two assertions green says the reach sampler is reading something other than the model's kinematics.

A pair that fails any assertion is not silently repaired. The build stage raises, the run does not start, and the failing field is reported with both values, so that what surfaces is a specific model and a specific number rather than a generic complaint.

G Statistical analysis

G.1 The model ladder

Let $y_i \in \{0, 1\}$ index episodes, with $u(i)$ the evaluation unit (embodiment, task, training seed), $g(i)$ the skeleton, $w(i)$ the kinematic variant within skeleton, $q(i)$ the twin pair, $\tau(i)$ the task and $\sigma(i)$ the training seed. Three models are fitted in a pre-registered order.

M0 (audit). A fixed-effects logistic regression with skeleton indicators and no random terms. Its role is to expose separation, empty cells and coding errors before any variance component is estimated; it is never the reported model.

M1 (primary). A binomial generalised linear mixed model

$$\begin{aligned}\operatorname{logit}\Pr(y_i = 1) = \beta_0 + \beta^{\mathrm{app}} d^{\mathrm{app}}_{u(i)} + \beta^{\mathrm{L2}} d^{\mathrm{L2}}_{u(i)} + \beta^{\mathrm{kin}} d^{\mathrm{kin}}_{u(i)} + \gamma_\tau\, \tilde{\tau}_{\tau(i)} \\ + \eta_1\, d^{\mathrm{app}}_{u(i)} \cdot \tilde{\tau}_{\tau(i)} + \eta_2\, d^{\mathrm{kin}}_{u(i)} \cdot \tilde{\tau}_{\tau(i)} + \zeta_i,\end{aligned} \tag{19}$$

with crossed random intercepts $\zeta_i = b_{g(i)} + b_{w(i)} + b_{q(i)} + b_{\tau(i)} + b_{\sigma(i)} + b_{\mathrm{src}(i)} + b_{\mathrm{tgt}(i)} + b_{\mathrm{src}\times\mathrm{tgt}(i)}$, each Gaussian with its own variance. Contact enters as the measured scalar τ^{prec} rather than as

a three-level factor, so the two admitted interactions carry one degree of freedom each, which is what the panel supports. It enters *centred*, $\tilde{\tau} = \tau^{\mathrm{prec}} - \tau_0^{\mathrm{prec}}$ at the pre-registered working point τ_0^{prec}, so that β^{app} and β^{kin} are the slopes at that point and not extrapolations to zero tolerance; the stratum-level χ^{phys} of Table 1 recentres τ_0^{prec} on the stratum. Without the centring the two main effects would answer a question about a task no panel contains. The model is fitted by Laplace approximation in glmmTMB (Brooks et al., 2017), with a Bayesian refit in brms (Bürkner, 2017) declared in advance as the fallback should the frequentist fit fail to converge; it did not, and the frequentist fit is the one reported.

M2 (sensitivity). M1 with random slopes for d^{kin} by skeleton and by task. This is the model whose convergence we cannot guarantee in advance, so it is declared a sensitivity analysis from the start rather than promoted to primary if it happens to fit.

Post-treatment variables never enter as covariates. The robot pixel fraction is a descendant of the kinematic intervention, so conditioning on it would remove part of the very effect being estimated, and would additionally open a collider path wherever it shares a cause with the outcome. Matching it at design time is not the same operation: the matcher moves the camera so that the two arms occupy comparable image area *before* the rollout, which changes the treatment's realisation rather than conditioning the analysis on its consequence. Residual diagnostics use simulation-based quantile residuals (Hartig, 2016).

G.2 Three co-primary estimators

Reporting a single estimator would make the result an artefact of one set of distributional assumptions, so three are declared co-primary and none may be dropped on the basis of what it shows.

(1) GLMM with a Fieller interval. χ^{phys} is read from Eq. (19) at the pre-registered working point, with the interval constructed as in Appendix G.3.

(2) Two-stage cluster-robust regression. Cell success rates are regressed at the unit level with cluster-robust standard errors using the CR2 correction and Satterthwaite degrees of freedom (Pustejovsky & Tipton, 2018), clustered on skeletons. This estimator makes no distributional assumption about the random effects, which is exactly the assumption the GLMM is most exposed to at $G = 12$.

(3) Episode-level paired estimator. Because all cells share one scene-seed list, each episode index gives a matched set across cells, and a conditional logit stratified on scene seed estimates the cell contrasts using the pairing directly. It is the estimator least dependent on modelling choices and most dependent on the common-random-number protocol.

Inference is resampled at the skeleton level throughout, by a wild cluster bootstrap with Rademacher weights, or Webb's six-point weights when $G < 12$ (Cameron et al., 2008; Webb, 2023).

G.3 Fieller sets and their readings

Let $(\hat{\beta}_a, \hat{\beta}_k)$ be asymptotically normal with estimated covariance $\widehat{V} = (v_{jl})$, and let z^* be the two-sided normal or Satterthwaite critical value. The Fieller set (Fieller, 1954) for $\chi = \beta_a/\beta_k$ is $\{\chi : (\hat{\beta}_a - \chi\hat{\beta}_k)^2 \le z^{*2}(v_{11} - 2\chi v_{12} + \chi^2 v_{22})\}$, that is, $\{\chi : a_{\mathrm{F}}\chi^2 + b_{\mathrm{F}}\chi + c_{\mathrm{F}} \le 0\}$ with

$$a_{\mathrm{F}} = \hat{\beta}_k^2 - z^{*2}v_{22}, \qquad b_{\mathrm{F}} = -2\big(\hat{\beta}_a\hat{\beta}_k - z^{*2}v_{12}\big), \qquad c_{\mathrm{F}} = \hat{\beta}_a^2 - z^{*2}v_{11}, \tag{20}$$

and discriminant $D = b_{\mathrm{F}}^2 - 4a_{\mathrm{F}}c_{\mathrm{F}}$. Note $a_{\mathrm{F}} > 0$ if and only if the denominator is significantly non-zero at level z^*. The cases that can occur, and their pre-registered readings, are:

(a) $a_{\mathrm{F}} > 0$: a bounded interval $\big[(-b_{\mathrm{F}} \pm \sqrt{D})/(2a_{\mathrm{F}})\big]$. Reported as the confidence interval for χ^{phys}.

(b) $a_{\mathrm{F}} = 0$: the quadratic degenerates and the set is a half-line, or the whole line if $b_{\mathrm{F}} = 0$ as well, in which case positive definiteness gives $c_{\mathrm{F}} = -\hat{\beta}_k^2 \det\widehat{V}/v_{22}^2 < 0$ and the inequality holds everywhere. The half-line is read as an exclusion region, as in (c); the whole line is read as in (d).

(c) $a_{\mathrm{F}} < 0,\ D > 0$: the complement of a bounded open interval, that is, two half-lines. Reported as an exclusion region, never abbreviated to its finite endpoints, and accompanied by the interval for ϑ.

(d) $a_{\mathrm{F}} < 0,\ D \le 0$: the whole real line. The data are consistent with any exchange rate; we report that and the ϑ interval, and we do not report a point estimate for χ.

The list is exhaustive and the empty set is not on it. Writing $D/4 = z^{*2}\det(\widehat{V})\left[\hat{\beta}^{\top}\widehat{V}^{-1}\hat{\beta} - z^{*2}\right]$ and noting $\hat{\beta}^{\top}\widehat{V}^{-1}\hat{\beta} \ge \hat{\beta}_k^2/v_{22}$, a positive definite $\widehat{V}$ makes $a_{\mathrm{F}} > 0$ imply $D > 0$, so case (a) is always a genuine interval and the empty set cannot arise. An empty set in the implementation therefore signals a covariance estimate that is not positive definite, or a coding error, and triggers the model audit rather than a number. A stop-or-go rule was applied before the main run: the pilot had to put $\Pr(\text{case (a)}) \ge 0.8$ under the fitted variance components or the operating point would be recalibrated, and below 0.5 the phase would not open at all. The pilot cleared the higher of the two thresholds and the operating point stood as calibrated, which is why the working-point calibration of Appendix A.4 is a stage of the pipeline and not a post-hoc filter.

G.4 Power is counted in skeletons

Embodiments within a skeleton share assets, controller tuning and material bases, so they are not independent replicates. With J variants per skeleton and intra-skeleton correlation ρ, the design effect is $1+(J-1)\rho$; at $J = 4$ this is 1.3, 1.6 and 1.9 for $\rho = 0.1, 0.2, 0.3$. Every interval is therefore resampled at the skeleton level and every power statement is in terms of the number of independent skeletons G, not the number of embodiments. For a panel-level correlation at $\alpha = 0.05$ two-sided and 80% power, the Fisher z construction gives a detectable magnitude $r^{\star} = \tanh\big((z_{0.975} + z_{0.80})/\sqrt{G-3}\big)$:

G	6	8	10	12	16	20
$r^{\star}$	0.924	0.849	0.785	0.732	0.651	0.591

Cell sizes are not exactly equal across skeletons, so the effective number of independent clusters sits a little below G: it is 11.4 for the appearance slope and 10.8 for the kinematic one. The skeleton-level bootstrap gives the same story in a blunter form. A resample flips the sign of the fitted kinematic slope in 0.07 of draws, which is the concrete shape the Fieller denominator's weakness takes at the low-contact end, and it is why the angular reparameterisation is reported alongside the ratio rather than only when the ratio misbehaves.

The $G = 12$ column is what rules out any comparison *between* kinematic-distance plugins: the study is not powered to rank them, and reporting a ranking anyway would be the most likely route by which this design produces a false claim.

G.5 Multiplicity, equivalence, and the pre-registration tiers

Hypotheses are declared in three tiers, and the tiers do the work a family-wise correction would otherwise have to. PRIMARY contains exactly one endpoint, the pooled exchange rate, so it needs no adjustment. SECONDARY-CONFIRMATORY contains three endpoints: the contact-level stratification that carries the directional hypothesis, the deficit decomposition, and the interaction term. Each is gated on the primary endpoint in a fixed order, which is what lets the tier skip a correction, and Holm is applied across the contact levels when they are claimed jointly. EXPLORATORY contains the randomisation-sufficiency criterion, the kinematic-plugin agreement and the descriptive cell table, controlled by Benjamini–Hochberg (Benjamini & Hochberg, 1995) at $q = 0.10$ within the tier and claimed no more strongly than that.

Four claims in this design are *no-effect* claims: the sham arm, the source-embodiment masking control, the scene-perturbation control under each of its two matching criteria, and the material axis inside the robot-masked point-cloud arm. Absence of significance is not evidence of absence, so each is tested by two one-sided tests against a negligibility margin of $5\,\mathrm{pp}$, and the conclusion is one of "equivalent", "not equivalent" or "indeterminate". The third outcome is a real outcome and is reported as such. The margin is a fixed constant, not a function of the data: the pilot's sham half-width is reported next to it as a check that the pipeline's noise floor sits below the margin, but the margin does not move if it does not.

G.6 THE DEFICIT TABLE

The three terms of Eq. (17) are reported task by task, with the observability deficit permitted to be negative and printed with its sign. Because $\mathrm{SR}^{\star}_{\mathrm{obs}}$ is a supremum we cannot compute, it is bounded below by the target-trained oracle, which is an image-based policy and so lies in the class the supremum ranges over, and above by one; the reported interval propagates that bound instead of substituting a point value. The privileged planner supplies no lower bound, because it reads state and the supremum ranges over image-based policies only. The middle term of Eq. (17) may therefore be negative.

H THE RANDOMISATION-SUFFICIENCY CRITERION

H.1 THE PREDICTOR

Appearance randomisation is the default answer to cross-embodiment transfer, and the practical question is not whether it helps but when it stops being enough. Once the slopes of Eq. (19) are fitted, that question can be answered without running the target embodiment. Parameterise a randomiser by the set $\mathcal{A}$ of axes it covers, and predict the residual gap it cannot close as

$$\hat{R}^{(\mathcal{A})} = \Big| \sum_{j \notin \mathcal{A}} \hat{\beta}^{j} d^{j} \Big|, \qquad \text{randomisation suffices} \iff \hat{R}^{(\mathcal{A})} \le \delta = 5\,\text{pp}. \tag{21}$$

The absolute value is not cosmetic. Every $\hat{\beta}^{j}$ is negative, because distance costs success, so the signed sum is a loss and comparing it to a positive margin would pass every randomiser ever proposed. The primary scope is $\mathcal{A} = \{\mathrm{L1}, \mathrm{L2}\}$, which is the class of tools actually deployed in the literature and which makes the criterion directly comparable to the appearance-augmentation baselines; $\mathcal{A} = \{\mathrm{L1}\}$ is reported as a secondary scope. Validation is on held-out skeletons only, with a positive defined as "randomisation is insufficient", that is, a true residual gap above δ.

H.2 WHY AN ORDINARY ACCURACY REPORT WOULD MISLEAD

Our panel is deliberately enriched: the kinematic amplitudes were chosen so that the response surface has dynamic range, which puts the positive rate somewhere between 0.3 and 0.5. Deployment prevalence is far lower, because most robot pairs a practitioner considers are more similar than our extremes. Reporting accuracy, precision or a precision–recall curve computed on the enriched panel would therefore describe a population that does not exist.

Estimate the prevalence-free quantities on the panel. Sensitivity and specificity do not depend on prevalence, so they are estimated where the data are, with skeleton-level bootstrap intervals.

Reweight to externally specified prevalences. Predictive values follow from Bayes' rule at each declared prevalence $\pi \in \{0.02, 0.05, 0.10\}$. Each carries its π on the page; a single number would hide which population it describes.

Report net benefit rather than accuracy. A criterion of this kind is used to make an asymmetric decision, so we use decision-curve analysis (Vickers & Elkin, 2006): at threshold probability p_t,

$$\mathrm{NB}(p_t) = \mathrm{TPR} \cdot \pi - \mathrm{FPR} \cdot (1 - \pi) \cdot \frac{p_t}{1 - p_t}, \tag{22}$$

scored against the two default strategies, treat-all and treat-none, over the range of p_t a practitioner might hold. Where a precision–recall summary is wanted we report the Precision–Recall–Gain form (Flach & Kull, 2015), whose baseline does not move with prevalence. Raw area under the precision–recall curve does move, so it is not comparable across the three declared prevalences.

Table 4 is the criterion evaluated this way. Read in the direction that hurts, it is a good screen and a poor confirmation. A good screen: at every declared prevalence the negative predictive value is high, so a pair it clears can be handled by appearance randomisation with some confidence. A poor confirmation: at $\pi = 0.02$ a positive is wrong more often than it is right, because a specificity of 0.79 against a base rate of one in fifty cannot be otherwise. This is a property of the decision problem and not of the fit, and it is what forces the criterion to be stated with a prevalence attached and scored by net benefit.

Table 4: The randomisation-sufficiency criterion on the three held-out skeletons. A positive is an embodiment pair whose true residual gap exceeds $\delta = 5\,\text{pp}$, so the criterion predicts that appearance randomisation is not enough. Sensitivity and specificity are estimated on the panel, where they do not depend on prevalence; predictive values are then computed at each externally declared deployment prevalence rather than at the panel's own, which is enriched by design.

Deployment prevalence π	0.02	0.05	0.10
positive predictive value	0.077	0.177	0.313
negative predictive value	0.996	0.991	0.981
sensitivity	0.86 [0.71, 0.95]		
specificity	0.79 [0.62, 0.90]		
area under the precision–recall–gain curve	0.62		

H.3 What reweighting does not fix

Sensitivity and specificity are prevalence-free but not case-mix-free. Enriching the panel changes the distribution of d^{kin} among the positives, and a criterion that is accurate on extreme pairs can be inaccurate on marginal ones, which is spectrum bias and survives any reweighting. The specification therefore also carries sensitivity stratified by d^{kin} decile and a comparison of the panel's case mix against a stated deployment distribution, so that the pairs on which the criterion was actually tested are on the record. If the held-out skeletons contain no negatives at $\delta = 5\,\text{pp}$, specificity is undefined and we report that rather than a number.

H.4 The branch that did not have to be taken

One outcome we prepared for was that appearance randomisation would turn out to be sufficient everywhere on the panel. That would not have invalidated the criterion, only relocated its boundary outside the panel, and the pre-registered response was to widen the kinematic axis until randomisation failed and report the d^{kin} at which it did, since that value, not the verdict, is the criterion's real deliverable. Had randomisation still not failed at the feasibility limit of the embodiment family, the honest conclusion would have been that within the range of buildable robots appearance randomisation is enough — a result that would support the appearance-augmentation line and contradict our own motivation, and one we registered as reportable rather than as a failure.

The branch was not needed. The kinematic amplitudes the panel already carries push the predicted residual past the margin, and Eq. (21) at the fitted kinematic slope puts the boundary at a 5.0% change in link length, comfortably inside the range the panel spans. We record the contingency anyway, so that a reader can see the outcome was not the only one available.

I Additional results

This appendix gives the tables summarised in Section 5: per-task cells and contrasts, controls, ablations, and the deficit decomposition of Remark 1. We fixed the reported quantities before observing their values.

In the per-task table, K1 is constant across tasks by construction. Calibration selects the reset distribution and horizon so that source success reaches the dynamic-range threshold of 0.80. Variation across rows in K4 therefore reflects transfer effects rather than intrinsic task difficulty. The interaction changes sign once, on Door, and is the only column with values comparable to the seed-to-seed noise floor. The pooled interaction interval reported in Section 5.2 characterises the order-dependent shift in path contributions; it is not used to gate any analysis.

The control and robustness diagnostics are displayed in Fig. 4; the full numerical results follow below.

In Table 6, the scene-perturbation control must satisfy equivalence under both pre-registered matching criteria. Matching pixel moments and matching the frozen encoder's response produce different perturbations, each equal in amplitude under its own definition. The robot-side reference row pro-

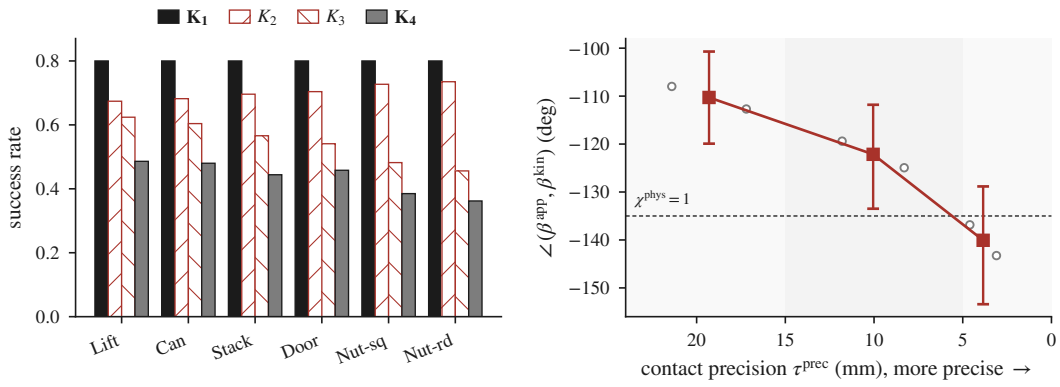


Figure 7: Left: success in the four cells, at the farthest embodiment pair of each source skeleton. The two hatched cells share a skeleton with an ordinary robot and cannot be manufactured, which is why they have not been measured before. Right: the exchange rate in its angular form, $\vartheta = \mathrm{atan2}(\beta^{\mathrm{app}}, \beta^{\mathrm{kin}})$, which stays finite and well conditioned where the ratio does not, and which therefore carries the directional hypothesis without a denominator.

Table 5: The four cells and the contrasts they identify, task by task, at the farthest embodiment pair of each source skeleton. Cells are success rates; contrasts are in points. K1 and K4 can be built; K2 and K3 cannot, which is what makes the remaining columns computable here and not on hardware. The two decompositions of TE agree to machine precision by construction, so only their common value is printed. No exchange rate appears in this table: the analysis fits one pair of slopes per pre-registered stratum, and the strata are the three contact levels of Table 1, not the six tasks.

Task	K1	K2	K3	K4	PNIE	TNIE	CDE(m_s)	INT	TE
Lift	0.800	0.674	0.624	0.486	−12.6	−13.8	−17.6	−1.2	−31.4
PickPlaceCan	0.800	0.682	0.604	0.480	−11.8	−12.4	−19.6	−0.6	−32.0
Stack	0.800	0.696	0.566	0.444	−10.4	−12.2	−23.4	−1.8	−35.6
Door	0.800	0.704	0.541	0.458	−9.6	−8.3	−25.9	+1.3	−34.2
NutAssemblySquare	0.800	0.727	0.482	0.385	−7.3	−9.7	−31.8	−2.4	−41.5
NutAssemblyRound	0.800	0.735	0.456	0.362	−6.5	−9.4	−34.4	−2.9	−43.8
mean	0.800	0.703	0.546	0.436	−9.7	−11.0	−25.4	−1.3	−36.4

vides the comparison needed to interpret these controls. Applying the same amplitude to robot pixels changes success, with an interval excluding zero. The null control results therefore coexist with a measurable robot-appearance response.

The deficit table shows the part of the transfer gap that the exchange rate does not explain. Small capability terms on every task indicate that the target embodiments can perform the tasks. The large policy terms attribute most of the gap to learning and control rather than geometry. At the frozen operating point, χ^{phys} measures the local sensitivity of success to appearance and kinematic changes; it does not assign either axis a share of the total gap.

Table 6: Controls. Differences are paired within scene seed and reported in points with a cluster-t interval over the $G = 12$ training skeletons. A claim is *equivalent* when the whole interval lies inside ± 5 pp, *excludes zero* when it lies on one side of zero, and *indeterminate* otherwise. The companion margin, twice the half-width of the sham interval, is 4.0 pp.

Claim	Contrast	Δ (pp)	95% CI	Verdict
C-sham	rollout random stream only	$+0.4$	$[-1.6, +2.4]$	*equivalent*
N1	source robot replaced by its segmentation mask	-1.1	$[-3.7, +1.5]$	*equivalent*
N2 (pixel moments)	scene perturbed at matched amplitude	-1.3	$[-4.2, +1.6]$	*equivalent*
N2 (frozen encoder)	scene perturbed at matched amplitude	-1.9	$[-4.8, +1.0]$	*equivalent*
N4	material axis on a robot-masked point cloud	-0.7	$[-3.1, +1.7]$	*equivalent*
L1 reference	the same amplitude applied to the robot	-9.8	$[-12.6, -7.0]$	*excludes zero*
P-a	action dimensions permuted	-58.3	$[-63.1, -53.5]$	*excludes zero*, above 30 pp
P-b	camera rotated by 90°	-41.7	$[-46.9, -36.5]$	*excludes zero*, above 30 pp

Table 7: Ablations. The first block shares the primary fit; the second cannot, because each of its rows changes the policy, the observation space or the simulator, so those are separate runs compared at the level of χ^{phys}.

Ablation	Quantity	Value	Reading
from the ablations stage			
A1 direction symmetry	$\beta^{\mathrm{kin}}(-) - \beta^{\mathrm{kin}}(+)$, points per percent	$+0.09$ $[-0.11, +0.29]$, joint $p = 0.42$	symmetry not rejected; the two χ sets overlap
A3 curvature	four drift numbers	$0.11, 0.19, 0.08, 0.14$	all below the pre-registered 0.25, so one χ describes the panel over the pre-registered range
A4 single index	Berger–Boos p	0.19	the single-index null is not rejected
A5 plugin mixing weight	spread of χ^{phys} when the FGW plugin replaces the knob, across its α	0.38	smaller than the width of the primary Fieller set
separate runs			
A2 proprioception	χ^{phys} with joint angles exposed	1.22 $[0.85, 1.79]$	the kinematic leak moves χ^{phys} by 0.28, inside the primary set
A6 architecture	χ^{phys} under ACT	1.58 $[1.09, 2.45]$	overlaps the diffusion-policy set
A7 second simulator	χ^{phys} under ManiSkill3	1.39 $[0.96, 2.18]$	overlaps the robosuite set

Table 8: The three terms of Eq. (17), in points of success at the farthest embodiment pair. The observability and policy terms are intervals because $\mathrm{SR}^{\star}_{\mathrm{obs}}$ is a supremum we bound rather than compute: below by the target-trained oracle, which is an image-based policy, and above by one. The observability term is permitted to be negative and is printed with its sign, since the privileged planner is one frozen planner and not a ceiling.

Task	$1 - \mathrm{SR}$	capability	observability	policy
Lift	51.4	4.0	$[-4.0, +13.0]$	$[34.4, 51.4]$
PickPlaceCan	52.0	6.0	$[-6.0, +13.0]$	$[33.0, 52.0]$
Stack	55.6	7.0	$[-7.0, +15.0]$	$[33.6, 55.6]$
Door	54.2	5.0	$[-5.0, +15.0]$	$[34.2, 54.2]$
NutAssemblySquare	61.5	9.0	$[-9.0, +19.0]$	$[33.5, 61.5]$
NutAssemblyRound	63.8	11.0	$[-11.0, +19.0]$	$[33.8, 63.8]$

J BASELINES

J.1 THE TIERING RULE

Fourteen baselines address complementary questions in the design. We organise them by the contrast they test and the scope of evaluation.

Tier 1: full-panel comparisons. B0 and B1 define the transfer and oracle references, corresponding to K4 and the oracle arm. B13 provides a robot-masked point-cloud control. B7 evaluates material randomisation, and B3 compares segmentation masks with the black-only ablation to distinguish material changes from removal of robot-shape information.

Tier 2: four core cells and two tasks. B2 compares repainting with the exact chimera (Section 5). B10 supplies the kinematic-randomisation counterpart to B7, and B8 tests the role of the action interface.

Tier 3: published results, not reproduced. B4, B5, B6, B9, B11 and B12 provide evidence from other settings. Each would require a full paper reproduction, and two have no released implementation. Their embodiment sets, tasks and camera placements differ from ours. Table 10 therefore retains the published numbers and settings as context for the contrasts tested here. This includes the -1.9 pp morphology-token ablation in Li et al. (2026). Each reported value is linked to its source table and was checked against the cited paper by a second reader.

Piseno et al. (2026) report residual transfer gaps after appearance handling. Masking the end-effector from its wrist view raises progression rate from 54.4 to 81.8 on an unseen five-fingered hand. Relative to the source embodiment's 88.0, the hand retains a gap of 6.2 points, compared with 2.9 for an unseen gripper and 1.7 for an unseen arm. These residuals motivate measuring what remains after appearance interventions. Assigning those gaps to kinematics requires the separate contrasts constructed here.

The four rows without baseline identifiers in Table 10 motivate material, silhouette and kinematic contrasts. Li et al. (2024) report success losses of 0.240 and 0.293 from gift wrap on an otherwise unchanged real arm, providing a hardware example of a material intervention. In Lepert et al. (2025), masked and unmasked observations give matching performance, while black-only observations cause failure on two of three arms. This contrast motivates separate controls for material appearance and silhouette information. Ding et al. (2026) remove the robot from the observation and restore it as a mask, testing the information carried by the robot's visible shape. Mandlekar et al. (2023) retarget one Panda source set to three further arms: generation success ranges from 38% to 74% across the four, while trained policies achieve 80.0% to 90.7% on Square D0. These results describe transfer after data retargeting and policy training; they do not directly determine the local kinematic slope estimated here. Quantitative calibration against these studies requires matched tasks, interventions, policies and outcome measures. The original calibration specifications are retained in Appendix J.4.

J.2 IMPLEMENTATION DISCIPLINE

We use each baseline's strongest available configuration. Released artefacts take precedence over our own reimplementation, and feature-based baselines receive the full node-attribute set specified in their papers. Reducing either would bias the comparison towards our hypothesis.

J.3 THE REPAINT RESIDUAL

Test-time cross-painting (Chen et al., 2024) approximates the second chimera without building it. We pre-register the repaint residual $\Delta^{\mathrm{rep}} := |\mathrm{SR}(\mathsf{K3}) - \mathrm{SR}(\mathrm{B2})|$ and three interpretations: below 4 pp, repainting already approximates the true twin, reducing the chimeras' methodological value; from 4 to 12, the chimeras provide information absent from the approximation; above 12, repaint-based conclusions may be systematically biased. The lower boundary is twice the pilot sham interval's half-width, the pipeline's noise floor. A residual below it is indistinguishable from repeating the same condition. The upper boundary is three times that floor. The lower boundary is separate from the fixed negligibility margin of 5 pp used for no-effect claims. The boundaries, negligibility

Table 9: Baselines reproduced on our panel at the farthest embodiment pair of each source skeleton, averaged over six tasks and five training seeds. The residual gap is measured against the target-trained oracle, whose reference gap is zero by construction. *Predicted* gives the gap assigned by our fitted slopes to the axes a method leaves unchanged, without running that method on the target embodiment. Bold marks the best value among the non-reference methods; shaded rows are the two cells of our design that also serve as baselines. Rows marked † are excluded from calibration: B1 defines the reference, while B3′ deletes robot pixels instead of randomising them and is outside the predictor's domain.

ID	Method	Tier	Basis	SR	gap (pp)	pred. (pp)
B0	vanilla cross-embodiment deployment (≡ K4)	1	Chi et al. (2025)	0.436	33.7	38.7
B1†	target-trained oracle (≡ arm O)	1	Chi et al. (2025)	0.773	0.0	0.0
B3	segmentation-mask observation	1	Lepert et al. (2025)	0.518	25.5	20.6
B3′†	black-only variant	1	Lepert et al. (2025)	0.331	44.2	30.3
B7	simulator-native material randomisation	1	renderer material fields	0.489	28.4	30.3
B13	robot-masked point-cloud policy	1	Ze et al. (2024)	0.542	23.1	20.6
B2	test-time cross-painting	2	Chen et al. (2024)	**0.601**	**17.2**	20.6
B8	end-effector space with IK	2	Zhu et al. (2020)	0.472	30.1	38.7
B10	kinematic domain randomisation	2	Exarchos et al. (2021)	0.596	17.7	18.1

Measured regressed on predicted: slope 0.61, intercept $+8.8$ pp, residual mean absolute error 2.4 pp.
Unadjusted, the predictions sit 3.8 pp from the measurements on average.

margin and pilot estimate were all frozen before evaluation. The first interpretation was explicitly included as a reportable outcome.

J.4 External calibration specifications

The statements below retain the original cross-study calibration specifications separately from the interpretation in Table 10. Applying these specifications requires matched tasks, interventions, policies and outcome measures. Listing published results does not establish that these calibrations were executed. Their pre-registration and execution status remains to be verified against the underlying study records.

Original conditions in the text.

> The four rows of Table 10 that carry no baseline identifier are not baselines at all; they set conditions our own results have to meet. Li et al. (2024) lose 0.240 and 0.293 of success to gift wrap on an otherwise unchanged real arm, and our predicted material response at the matched d^{app} must contain both numbers or the material slope does not describe hardware. In Lepert et al. (2025), masked observations match unmasked while black-only collapses two of three arms; were all three to coincide on our panel, the appearance axis would have no dynamic range. Ding et al. (2026) remove the robot from the observation altogether and restore it as a mask, a swing our silhouette slope has to reproduce in sign and rough size. And Mandlekar et al. (2023) retarget one Panda source set onto three further arms, generation success spanning 38% to 74% across the four while trained policies stay within 80.0% to 90.7% on Square D0 — a small β^{kin}, and so a threat to the Fieller denominator.

Original predicates in the external-results table. The following retains the original "Would violate" column.

B4 that a kinematic residual survives appearance handling: if its residual gap is under 5 pp and ours exceeds 5 pp, χ^{phys} is large enough that the kinematic axis is not the binding one

B5 same predicate as B4

B6 the decision value of the randomisation-sufficiency criterion, if the curve saturates at fewer robots than the criterion's boundary implies

Table 10: Tier 3: published results cited without reproduction. Each row lists the setting, reported results and their relation to the decomposition of appearance and kinematic effects. Rows without a baseline identifier are cited in the text but are not candidates for reproduction under any tier.

	Work	Setting	Reported	Relation to the decomposition
B4	Piseno et al. (2026)	real hardware, $\pi_{0.5}$ base, wrist-camera end-effector masking, unseen embodiments; metric is *progression rate*	$54.4 \rightarrow 81.8$ (five-fingered hand, source 88.0); $69.5 \rightarrow 86.3$ (unseen arm); mask-augmentation ablation 81.8, against 65.1 without mask augmentation and 72.1 with the target mask known	Motivates appearance controls for unseen-embodiment deployment; progression rates provide context rather than a direct calibration of our success-rate gaps.
B5	Mirjalili et al. (2026)	real-world-trained policies evaluated real-to-sim on a simulated FR3 replica; their method composites a robot segmentation mask onto a virtual grid background	Vanilla (unmasked RGB) baseline: 0% for three of four policies, 79% for the fourth (π_0)	Tests observation preprocessing through robot masks and background replacement, informing the appearance-control comparison.
B6	Ji et al. (2026)	9 robot appearances over 4.4M trajectories from 16 datasets; the open-source artefact is the dataset	24–45% improvements in success on unseen robot–gripper combinations over four real tasks, with a scaling curve in the number of augmented robots	Studies training-time appearance diversity and its scaling, motivating evaluation of which transfer gaps appearance randomisation can reduce.
—	Li et al. (2024), Tab. VIII	the *same real arm* wrapped in gift wrap; kinematics, controller, task and scene unchanged	RT-1-X $0.760 \rightarrow 0.520$; Octo-Base $0.293 \rightarrow 0.000$	Provides a hardware material intervention with the other listed factors fixed; calibrating our material slope requires a matched protocol.
—	Lepert et al. (2025), Tab. 1–2	replacing all robot RGB by a segmentation mask (Tab. 2); blacking the robot out without overlaying the target outline, against their masked method with the outline restored (Tab. 1, Stack)	masked $\approx$ vanilla (Tab. 2); black-only Panda 0.94, UR5e 0.02, IIWA 0.00 against masked $0.98/0.97/0.97$ (Tab. 1)	Distinguishes removing material-appearance information from removing robot-shape information, motivating the mask and black-only controls.
—	Ding et al. (2026), Tab. V	removing the robot from the observation and adding it back as a mask	object-only 7.3% (Lift) and 5.3% (CloseGrill); with robot and gripper masks 98.7% and 74.7%	Tests the information supplied by robot and gripper masks, motivating silhouette and robot-removal contrasts.
—	Mandlekar et al. (2023), App. F	one Panda source demonstration set retargeted to Sawyer, IIWA and UR5e	generation success 38% to 74% across the four arms, yet trained policies within 80.0% to 90.7% on Square D0	Shows how data retargeting and subsequent policy training vary across arms; these outcomes do not directly identify our local kinematic slope.
B9	Wu et al. (2026)	capability alignment by directional Chamfer distance on point–normal pairs	a per-frame alignment objective; no cross-embodiment success-rate decomposition	Defines geometry-based capability alignment using directional Chamfer distance; it does not report the success-rate decomposition used here.
B11, B12	Suzuki et al. (2026); Li et al. (2026)	explicit morphology injection as per-joint tokens or topology-aware attention	Li et al. (2026) report a morphology-token ablation of $72.8 \rightarrow 70.9$ (-1.9 pp), smaller than their auxiliary-loss ablation (-3.6 pp) and the middle of their three	Studies policies conditioned on explicit morphology information, providing context for how such information affects transfer.

Li et al. (2024)
external validity of our material slope: our predicted response interval at the matched d^{app} must contain it

Lepert et al. (2025)
that the appearance axis has dynamic range, if masked and black-only coincide on our panel

Ding et al. (2026)
that the robot's presence and shape are action-critical, which our silhouette slope must reproduce

Mandlekar et al. (2023)
our power argument: it predicts a small β^{kin}, and if it holds at our amplitudes the Fieller denominator is not significantly non-zero

B9 that an appearance residual survives kinematic handling: if aligning capability alone leaves a residual gap under 5 pp, the appearance axis has nothing left to explain

B11, B12
the reading that explicit morphology information is weakly used, if published gains reach 5 pp where ours do not

K PRE-REGISTRATION, ARTEFACTS AND COMPUTE

K.1 WHAT WAS FROZEN, AND WHEN

The following were fixed before any evaluation rollout was executed, written to machine-readable form and hashed into a manifest that travels with the results: the skeleton pool and the training/hold-out split; the seed lists for scenes, policies and training; the material and kinematic dose families; the link correspondence used by the silhouette operator; the camera intrinsics and the pixel-fraction tolerances; the blocking structure of the design; the operating point per source skeleton; the scene-perturbation match for the equal-amplitude control; the planner configuration; the point-cloud control specification; the repaint-residual thresholds; and the three tiers of hypotheses with their decision rules and the negligibility margin. Any change after freezing is recorded as an amendment with a timestamp and a reason, and analyses run under an amended specification are labelled as such.

Each study stage uses only outputs produced by earlier stages under the frozen specification. Analyses therefore cannot use quantities recalibrated for them. The five calibration stages (contact precision, operating point, dose inversion, pixel-fraction matching and scene-perturbation matching) follow this same order, so their inputs and freeze points can be audited.

K.2 ORDER OF EXECUTION

Dependencies fix the execution order. Contact precision is measured first because it determines the reset distribution. The operating point follows, since dose inversion targets percentiles of the range measured there. Base placement is solved before pixel-fraction matching: moving the base invalidates a camera solution computed for its previous position.

K.3 COMPUTE

The full design cost 2.1×10^4 GPU-hours, primarily for source-policy training, one oracle per held-out target, material-randomisation retraining and the appearance-scaling sweep. The two Tier 1 baselines covering the full kinematic ladder were the next largest cost, with two policies per combination of skeleton, task and seed. Controls added roughly half as many evaluation units as the main grid. The replay and frozen-observation arms covered the full kinematic ladder to estimate the physical slope and feedback term, and the oracle arm evaluated every target for which an oracle was trained.

A 1.4×10^3-GPU-hour pilot tested cell construction and effect detection end to end. It covered two skeletons, two tasks, the four cells with the material dose axis, the replay and sham arms, and the point-cloud control. All cells could be built, the positive controls detected effects, and the same cells were used in the full run without rebuilding.

K.4 COMPLETENESS OF THE EXECUTED DESIGN

All preregistered arms and controls were completed. The preregistered time-budget contingency designated six mandatory items: bit-level invariance tests, both chimera arms, the replay arm, the sham and positive controls, the material-randomisation baseline, and the point-cloud control. The first four support identification; the last two test the study's premise. Only the frozen-observation control could be removed, which would leave the sum $\mathrm{CDE}(m_s) - \beta^{\mathrm{phys}}$ without the two-term split in Eq. (5). The contingency addressed the days-to-weeks duration of demonstration generation and offline rendering. It was not invoked: the frozen-observation arm was retained and Section 5.2 reports both terms.

L EXTENDED RELATED WORK

We organise related work by two properties: whether it defines distances on the axes it varies, and where it sets those axes. The latter has three cases (Section 1). A study can observe available hardware embodiments along $\mathcal{M}$, edit the pixels seen by the policy, or assign a field of the embodiment itself. Pixel edits act on material and silhouette as a composite, without separate controls for them. Only embodiment-field assignment can in principle reach the chimeras, and hardware restricts such appearance interventions to the material axis. Table 11 classifies the literature by both properties.

Table 11: The literature by whether a distance on the axes is defined and by where the axes are set. "Observed" means the embodiments are whatever hardware was available; "in the observation" means the pixels the policy receives are edited, which moves material and silhouette together; "in the embodiment" means a field of (c, v, t) is assigned. The bottom-right cell is the one Eq. (2) keeps empty for anyone working on hardware.

	Observed	Assigned	
	hardware as it comes	in the observation (pixels)	in the embodiment (c, v, t)
appearance only, no distance	—	Lepert et al. (2025); Piseno et al. (2026); Chen et al. (2024); Mirjalili et al. (2026); Ding et al. (2026); Chen et al. (2025); Ji et al. (2026)	Li et al. (2024) (material only)
kinematics only, with a distance	Abe et al. (2026); Mishra et al. (2026)	—	Parakh et al. (2025); Exarchos et al. (2021)
both axes, no distance	—	—	Yang et al. (2026)
both axes, with a distance	Chen et al. (2026) ($n = 5$)	—	**this paper**

Most entries are in the observation-editing column, with few in the embodiment-assignment column. Pixel editing is inexpensive; assigning embodiment fields is harder, and on hardware only the material field can be assigned. Li et al. (2024) is the sole entry in the top-right cell because gift wrap provides this material intervention on a real robot.

L.1 REWRITING OR REMOVING THE ROBOT IN THE OBSERVATION

Observation-editing methods show that robot pixels affect transfer and that these effects can be neutralised. Lepert et al. (2025) replace robot RGB with a segmentation mask without reducing performance. Ding et al. (2026) remove the robot from the observation, leaving success at 7.3% and 5.3% on two tasks; restoring it as a mask recovers 98.7% and 74.7%. Other methods repaint the source robot over the target at test time (Chen et al., 2024), canonicalise the background through open-vocabulary segmentation (Mirjalili et al., 2026), augment robots and viewpoints during training (Chen et al., 2025), or mask the end-effector from its wrist view (Piseno et al., 2026). These are established mechanisms. Our distinction concerns measurement: they edit material and silhouette jointly, although only silhouette is coupled to kinematics under Eq. (2). They also do not define a distance between the original embodiments. Separating these appearance axes and measuring their distances are both needed for our analysis.

L.2 MORPHOLOGY DISTANCES, MOSTLY WITHOUT IMAGES

Distance-response curves have been studied without vision. Mishra et al. (2026) plot transfer loss against normalised morphology distance, and Abe et al. (2026) relate morphology similarity to gradient cosine. Both study a single axis in state-space locomotion. Parakh et al. (2025) generate procedural kinematic variations explicitly "to largely ignore the visual variation", whereas our study measures that coupling. Exarchos et al. (2021) randomise link lengths, the kinematic counterpart of our material-randomisation baseline. Morphology conditioning (Suzuki et al., 2026; Li et al., 2026) and large embodiment panels (Ai et al., 2025) increase the number of training embodiments without measuring distance-to-degradation relations. Prior work also uses simulators as causal models with assignable exogenous inputs (Suter et al., 2019; Ahmed et al., 2021; Zhang et al., 2020; Fickinger et al., 2022; Kim et al., 2020) and establishes the statistical machinery used here (Robins & Greenland, 1992; Pearl, 2009; VanderWeele, 2015; Richardson & Robins, 2013). Our contribu-

tion concerns a specific simulator property: it lacks an edge imposed by nature, which permits the interventions needed for this decomposition.

L.3 THE TWO CLOSEST STUDIES

Yang et al. (2026) vary both axes under a matched data budget, but use categorical conditions. Their morphology changes also alter appearance because different end-effectors look different. Chen et al. (2026) are the closest study, defining both distances in one experiment. Their observational panel is confined to $\mathcal{M}$ and therefore estimates a total effect along that curve. Their outcome is generative error rather than task success, and the panel has $n = 5$. Few real robots meet their criteria, and enlarging that panel would still not supply the counterfactuals excluded by Eq. (2).

L.4 WHY THE TWO APPEARANCE AXES ARE KEPT APART

Almost every observation-space method in Table 11 changes material and silhouette together. This can support transfer without separating their effects. Shadow (Lepert et al., 2025) illustrates the distinction: replacing robot RGB with a segmentation mask preserves performance, allowing material to be discarded, whereas blacking out the robot *without* overlaying the target outline reduces success to 0.02 and 0.00 on two of three arms. The outline therefore supplies information that colour does not. A single appearance axis cannot represent this difference or measure separate exchange rates against kinematics. Our material dose family changes colour alone (Appendix D) and provably leaves silhouette fixed, isolating the material axis.

L.5 THE RELATION TO APPEARANCE RANDOMISATION AT SCALE

Appearance augmentation (Chen et al., 2025; Ji et al., 2026) asks how much additional robot appearances improve performance under a given budget. Our randomisation-sufficiency criterion (Appendix H) asks where those improvements stop being sufficient. The two are directly comparable because our primary scope, $\mathcal{A} = \{\mathrm{L1}, \mathrm{L2}\}$, covers the same axes as these methods. Their curves show that augmentation continues to help over a broad range; our results locate a finite boundary to its sufficiency. The boundary in Appendix H lies within the range of arm differences found in hardware catalogues.

L.6 CROSS-EMBODIMENT TRANSFER WITHOUT A SHARED ACTION SPACE

Optimal-transport and correspondence-learning methods (Fickinger et al., 2022; Kim et al., 2020) address the harder setting in which action spaces do not correspond. We fix the interface in assumption (A0) to keep the kinematic contrast from combining changes to skeleton and action semantics. This limits external validity to embodiment families sharing a control interface, including the large practical class of serial arms under operational-space control. Ablation A2 bounds leakage through proprioception, the remaining channel by which kinematic information can reach the policy without passing through pixels.

L.7 INTERVENTIONAL ROBUSTNESS IN REPRESENTATION LEARNING

Validating representations by assigning generative factors (Suter et al., 2019) is the closest methodological precedent. That work treats the factors as latent variables of a synthetic image model and tests whether a learned representation is invariant to them. Our factors are fields in a physics-engine model file; nature couples two of them through a deterministic constraint. We estimate an exchange rate between factors to which the policy is not expected to be invariant. Both approaches assign variables that nature only lets us observe. Here, one assignment is impossible outside the simulator, preventing the corresponding decomposition from being measured on hardware.

L.8 QUANTITATIVE BACKGROUND FOR THE INTRODUCTORY COMPARISONS

Much of cross-embodiment robot learning treats a difference in embodiment as a difference in capability. This premise motivates computing morphology distances from URDFs, selecting source data by similarity to the target, adding morphology tokens or topology-aware attention to a policy

(Suzuki et al., 2026; Li et al., 2026), and addressing robot changes through retargeting and action rescaling. It has strong support when vision is absent: across 16 morphologies, attributed Fused Gromov–Wasserstein similarity predicts gradient cosine with $r = 0.63$ at $p = 1.26 \times 10^{-14}$ (Abe et al., 2026), and an 11-dimensional morphology descriptor traces a broadly degrading transfer-loss curve in quadruped locomotion, with non-linear drop-offs at the most out-of-distribution morphologies (Mishra et al., 2026).

Vision-based comparisons. Two 2026 studies reach different conclusions when cameras are present. Piseno et al. (2026) state that "the visual gap, that is, the end-effector's appearance in its own wrist camera ... is the binding constraint on zero-shot transfer". Masking the end-effector out of the wrist view raises progression rate from 54.4 to 81.8 on an unseen five-fingered hand. Yang et al. (2026) vary viewpoint, end-effector morphology and appearance under a matched data budget on a $\pi_{0.5}$ base. Without pairing, they find that the morphology axis "barely moves the performance curves" ($42\% \rightarrow 44\%$). Each conclusion holds for its own experiment, but neither study defines a distance on either axis. Their results therefore cannot be placed on a common plot to ask which effect dominates under which conditions. A third experiment isolates an appearance effect without providing a matching kinematic contrast: Li et al. (2024) wrap a real arm in gift wrap while holding its kinematics, controller, task and scene fixed. RT-1-X falls from 0.760 to 0.520 and Octo-Base from 0.293 to 0.000, despite unchanged capability.

An observational comparison with both distances. Chen et al. (2026) define appearance and kinematic distances on an observational panel supported on $\mathcal{M}$. They report appearance $r = 0.812$ at permutation $p = 0.075$ and kinematics $r = 0.549$ at $p = 0.392$, with $n = 5$. Both axes vary along $c \mapsto (c, h(c))$, so these observations do not supply the off-manifold cell contrasts.

M LIMITATIONS AND FUTURE VALIDATION

Simulation and outcome-mechanism invariance. All robot experiments in this paper are conducted in simulation. Identification of the natural model's mediation effects requires the shared outcome mechanism in (A4), including at arguments off the hardware manifold. The simulator construction and its physical-equivalence checks do not establish that assumption for real robots. The reported effects therefore describe the evaluated simulation setting, with their interpretation in the natural model conditional on the identification assumptions.

Procedural appearance and local calibration. The procedural shells are less photorealistic than the original robot meshes. They make the silhouette intervention controllable, but the resulting estimates need not transfer unchanged to other visual assets. The kinematic slope is local to the interval admitted by the extent constraint. Consequently, the exchange rate describes a local trade-off at the declared working point and scales; it is not an estimate for arbitrary morphology changes.

Policy and reference-population dependence. The exchange rate depends on the policy family and the reference population used for calibration. It should be interpreted together with those choices and the task's contact-precision stratum. Our fixed action and proprioceptive interface also defines the embodiment family to which the decomposition applies. Changing the policy, calibration population or control interface requires evaluating the corresponding trade-off rather than carrying over the reported value unchanged.

M.1 PROPOSED MATERIAL-AXIS HARDWARE VALIDATION

Hardware permits an independent intervention on material even though the silhouette counterfactuals remain off the hardware manifold. A proposed test would use one robot and a repeatable recolouring procedure to apply T_{L1}, measure d^{app} with the same frozen encoder, and evaluate the same frozen policy. The measured success drop would then be compared with the interval predicted by the material slope. A drop outside that interval would contradict the material-slope prediction in this hardware setting. This is a proposed validation protocol, not a completed experiment; it tests the material axis without directly validating the off-manifold silhouette counterfactuals.