Intervene, Calibrate, Discover: Learning Causal Structure from Pixels

Anonymous authors
Paper under double-blind review

Abstract

Learning high-level causal variables from low-level observations is a central problem for artificial intelligence (Schölkopf et al., 2021). With known node semantics but an unknown graph, readout errors can misalign model settings with environment intervention values. With encoder and decoder frozen, verified values from fixed round-robin interventions train dimension-wise invertible calibration; after freezing it, active experiments orient edges, local mechanisms predict consequences, and inverse calibration supports decoding. Across Domino, Pick-and-Place, Balance, Gears, CAUSAL3D, and CausalWorld, we test graph discovery, intervention consequences, and counterfactual images. Controlled domino experiments show that interventions resolve orientation ambiguity left by equal additional observations. In a matched-data domino control, calibration reduces the mean error in predicting how changing an intervention alters other variables within the same factual context, relative to an uncalibrated model.

1 Introduction

Can an agent predict intervention consequences from an image? Beyond reconstructing the domino scene in Figure 1, intervention queries require a target, its new value, and resulting responses (Schölkopf et al., 2021)—a distinction relevant to physical-system counterfactuals (Li et al., 2020) and robotic transfer under environmental changes (Ahmed et al., 2021).

Our goal is to query pixel-derived variables in environment intervention units. Differently directed graphs can share conditional independences, forming a Markov equivalence class (MEC). Interventions can distinguish candidates, but unique recovery requires sufficient coverage and identification conditions (Section 3) (Hauser & Bühlmann, 2012).

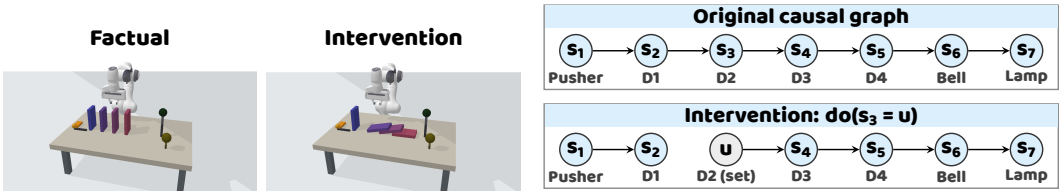


Figure 1: Left: simulator scenes before/after releasing domino 2 from a 20° tilt; the robot arm is fixed. Right: a separate seven-node structural causal model (SCM): pusher, four dominoes, bell, and light. $\mathrm{do}(s_3 = u)$ replaces the target mechanism and removes incoming edges. The physical preview is neither a prediction nor validation of this graph.

Label-aligned readouts can still have scale, offset, or nonlinear errors that corrupt queries despite correct graph directions. A *verified intervention value* is the attained target state; queries assume a valid setting without modeling actuator execution.

Contribution. To address numerical mismatch, we calibrate node-aligned readouts to intervention values with invertible T supervised by verified direct-target values, preserving the frozen visual model. Settings, graph discovery, and mechanism fitting share coordinates; T^{-1} returns predictions

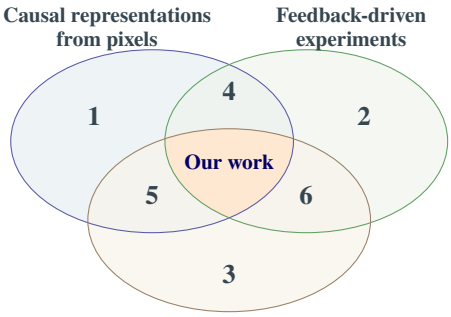
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping 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.

to the decoder. Valid, sufficiently oriented estimated graphs support non-target interventional means and counterfactual images conditioned on factual images.

2 Related Work

Figure 2 summarizes these distinctions.

Visual causal representations. CausalVAE learns concept-supervised representations supporting latent-intervention images (Yang et al., 2021). CITRIS identifies variables from known intervention targets and temporal images, optionally transforming pretrained autoencoder outputs invertibly; iCITRIS handles instantaneous effects, and BISCUIT learns variable–binary-interaction correspondence (Lippe et al., 2022; 2023b;a). Implicit latent causal models (ILCM) learn variables and permit structure extraction from intervention pairs (Brehmer et al., 2022); theory also studies intervention-based identification (Ahuja et al., 2023). These capabilities do not ensure numerical alignment with environment settings. We calibrate known-node readouts against verified values to address scale, offset, or nonlinear errors affecting intervention queries.



1 Ahuja et al. (2023)
2 Causal Curiosity Sontakke et al. (2021)
3 PC-stable Colombo & Maathuis (2014) GIES: Hauser & Bühlmann (2012)
4 ICF Thomas et al. (2018)
5 CausalVAE: Yang et al. (2021) CITRIS: Lippe et al. (2022)
6 AIT: Scherrer et al. (2021) CBED: Tigas et al. (2022)

Figure 2: **Positioning intervention-value queries.** Sets summarize inputs, acquisition, and structure learning; their intersection denotes a joint setting, not a ranking. We align visual estimates with verified values. Numbered work and classification conditions: Appendix H.1.

Active experiments and controllable factors. Causal Curiosity distinguishes factors experimentally; independently controllable factors (ICF) learn separately controllable visual factors (Sontakke et al., 2021; Thomas et al., 2018). Given defined variables, active strategies and AIT choose targets, causal Bayesian experimental design (CBED) also chooses values, and CAASL uses a pretrained policy (Hauser & Bühlmann, 2014; Scherrer et al., 2021; Tigas et al., 2022; Annadani et al., 2024). Visual-readout calibration remains separate: we calibrate round-robin and freeze the map before targeting unoriented edges, giving graph discovery and mechanisms shared intervention-value coordinates.

Response-based coordinate alignment. Rep4Ex uses action–latent invariance for intervention extrapolation (Saengkyongam et al., 2024). Ding et al. study counterfactual retrieval with a CausalVAE world-model plug-in, aligned during training when state supervision is available (Ding et al., 2026). We calibrate frozen, known-node visual models against verified values; a matched-data, fully fitted affine control tests nonlinear calibration's incremental benefit.

3 Problem Formulation

Given target j and environment value u, images inform predictions of settled *post-intervention states*, not frame-by-frame dynamics. Node semantics are known; the direct graph and structural equations are unknown (terminology: Appendix A.4).

Environment and true variables. Let $V = \{1, \dots, K\}$ index states $s = (s_1, \dots, s_K) \in \prod_j \mathcal{S}_j \subseteq \mathbb{R}^K$, with domains $\mathcal{S}_j$ either $\{0, 1\}$ for binary events or real intervals for continuous states. Unknown graph $G^\star = (V, \mathcal{E}^\star)$ has directed edges $\mathcal{E}^\star$; $\mathrm{Pa}_G(j)$ denotes direct parents. Structural equations $f_j^\star$ and disturbances $\epsilon = (\epsilon_1, \dots, \epsilon_K) \sim P_\epsilon$ generate states and images:

$$s_j = f_j^\star(s_{\mathrm{Pa}_{G^\star}(j)}, \epsilon_j), \qquad x = R(s, \nu) \in [0, 1]^{H \times W \times 3}. \tag{1}$$

Renderer R produces size-H, W images with imaging factors ν. Neither R nor $f_j^\star$ generates learner predictions; simulator states, graphs, and paired images serve only permitted supervision/evaluation. Superscript $\star$ marks truth; hats mark estimates.

Known inputs and supervision. Inputs are observational data $\mathcal{D}_{\text{obs}} = \{(x_i, \ell_i)\}_{i=1}^{N_{\text{obs}}}$, eligible targets $\mathcal{J} \subseteq V$, and feasible values $\mathcal{U}_j$. Labels $\ell_i \in \mathbb{R}^K$ supply numerical states and dimension identities, not $\mathcal{E}^\star$, supervising the $z_j = E_\theta(x)_j$–s_j correspondence. Single-frame ambiguity requires protocol-supplied extra frames or action information, equally available to compared methods.

Decoder coordinates z_j and graph/mechanism readouts h_j both contain visual error. The latter standardize $T_j(z_j)$ with observational statistics; T_j neither permutes nor mixes dimensions. Equation (5) calibrates images but maps u directly to target setting $q_j(u)$ without T.

Environment interventions and budget. Operation $a = (j, u)$ specifies target $j \in \mathcal{J}$ and verified value $u \in \mathcal{U}_j$. Hard intervention $\text{do}(s_j = u)$ replaces only the target mechanism:

$$s_j^a = u, \qquad s_k^a = f_k^\star(s^a_{\text{Pa}_{G^\star}(k)}, \epsilon_k) \quad (k \neq j). \tag{2}$$

The environment returns (j, u, x^{int}), not full states. u must match the depicted state; release actions and soft interventions require task-specific mechanism-replacement definitions. Equation (2) is evaluated topologically on acyclic graphs. Population prediction retains the exogenous distribution; individual counterfactuals fix factual (ϵ, ν). Observational selection on $s_j = u$ replaces no mechanism and is not intervention.

Budget $B = B_{\text{cal}} + B_{\text{graph}}$ counts new calibration and graph records, not replay; early stopping reports actual usage. Observational volume is separate. Evidence-dependent targeting precludes treating cross-target records as independent, identically distributed observational samples.

Identification assumptions and scope. Assume $G^\star$ is a directed acyclic graph (DAG), no common causes are omitted, $P_\epsilon = \prod_j P_{\epsilon_j}$, and there is no selection bias. Observational states satisfy causal Markov and faithfulness conditions; single-target interventions preserve other mechanisms and the exogenous distribution (Colombo & Maathuis, 2014; Hauser & Bühlmann, 2012). Relevant states have finite second moments, and standardization variables have nonzero training variance. Encoder training does not ensure these assumptions. Visual error, approximate tests, and inadequate coverage may prevent recovery. Dimension-wise invertible T adds no population conditional-independence information; unique recovery requires sufficient identification conditions.

Calibration requires a dimension-wise monotone approximation of the discrepancy that remains stable across observational/interventional images. Unintervened dimensions receive only observational-preservation constraints.

Learning objectives. We learn encoder E_θ, calibration $T_{\widehat{\phi}}$, graph $\widehat{G}$, mechanisms $\widehat{f}_j$, and decoder D_ψ for non-target responses to node settings. Parameters and normalization statistics use permitted training data only, then freeze (notation and fitted residuals: Appendix A.5).

(1) Structure identification. Compare adjacencies and directions in $\widehat{G}$ with $G^\star$, retaining partial graphs when evidence is insufficient. **(2) Intervention consequence prediction.** For $a = (j, u)$, set node j to verified u and predict each non-target node's mean settled state across contexts:

$$\mu_k(j, u) = \mathbb{E}[s_k \mid \text{do}(s_j = u)] = \mathbb{E}_{\epsilon \sim P_\epsilon}[s_k^a(\epsilon)], \qquad k \neq j.$$

This expectation uses the interventional distribution, not observations selected by $s_j = u$; $s_k^a(\epsilon)$ follows Equation (2). Estimate $\widehat{\mu}_k(j, u)$ averages across contexts without factual-image conditioning. It is an outcome mean, not an average treatment effect between settings. Excluding $k = j$ tests propagation rather than command copying. Tasks define in-distribution (ID) and out-of-distribution (OOD) splits. **(3) Counterfactual image generation.** Compare $\widehat{x}^{\text{cf}}(x, a)$ for factual image x and alternative a with same-context simulator replay. One image need not identify (ϵ, ν), so paired evaluation does not establish individual counterfactual identifiability. DAG-based fitting and queries require valid, sufficiently oriented estimates.

4 METHOD

Three stages connect images to intervention outcomes (Figure 3). **First, learn and calibrate readouts.** Concept supervision aligns latents with nodes. Freeze the encoder/decoder, then update only dimension-wise T using verified values from fixed round-robin acquisition. **Second, recover propagation structure.** Freeze T, estimate an observational candidate skeleton, and orient edges using responses to actively targeted environment interventions. **Third, predict states and images.** Fit local mechanisms on a valid, sufficiently oriented graph for population interventional means and factual-image counterfactuals; inverse calibration returns predictions to decoder coordinates. All modules freeze at test time (query coordinates: Appendix Figure A2).

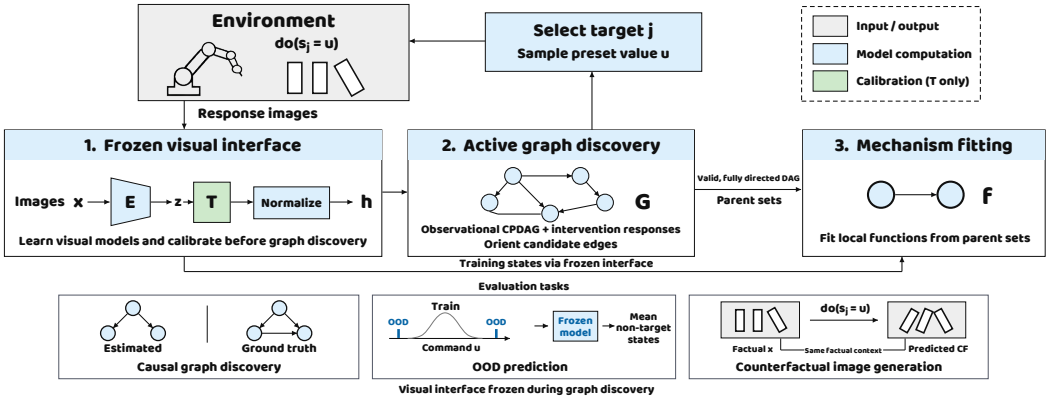


Figure 3: Learning and evaluation. The visual model and calibration freeze before active targeting, orientation, and mechanism fitting. E, T, G, f: encoder, calibration, graph, mechanisms. Gray/blue/green: inputs/outputs, computation, calibration. OOD excludes A/B-only tests lacking defined shift axes. Curves denote continuous commands (query coordinates: Appendix Figure A2).

4.1 NODE-ALIGNED VISUAL LEARNING AND READOUT CALIBRATION

A concept-supervised variational autoencoder learns reconstruction and latents anchored by standardized state labels (Appendix B.3). We use the frozen encoder's posterior mean $z = E_\theta(x)$. Jointly trained decoder D may be refitted solely on ID reconstruction as D'. Only T is calibrated; each control shares D or D'.

Calibrating with verified target states. Verified u and encoding $E_\theta(x^{\text{int}})_j$ describe the same target state. Training-label mean and standard deviation μ_j^s, σ_j^s define anchor units $\alpha_j(u) = (u - \mu_j^s)/\sigma_j^s$. Superscript s marks a fixed training mean, unlike command-dependent $\mu_k(j, u)$. Records (j, u, x^{int}) fit a strictly monotone dimension-wise map

$$T_{\phi,j}(z_j) = g_j z_j + b_j + \sum_{l=1}^{L} w_{jl} \tanh(z_j - \kappa_l), \qquad g_j > \sum_l |w_{jl}|. \tag{3}$$

Parameters ϕ comprise scales g_j, offsets b_j, and weights w_{jl}; L hyperbolic-tangent bases have fixed knots κ_l. The constraint ensures invertibility. For calibration records $\mathcal{D}_{\text{cal}}$, $\lambda \geq 0$ balances target supervision and observational preservation; expectations average finite records:

$$\mathcal{L}_{\text{cal}} = \mathbb{E}_{(j,u,x^{\text{int}})\sim\mathcal{D}_{\text{cal}}} \left[(T_{\phi,j}(E_\theta(x^{\text{int}})_j) - \alpha_j(u))^2 \right] + \lambda \mathbb{E}_{x\sim\mathcal{D}_{\text{obs}}} \frac{\|T_\phi(E_\theta(x)) - E_\theta(x)\|_2^2}{K}. \tag{4}$$

Only targets receive u supervision; the second term limits observational drift. After predetermined round-robin rounds replaying accumulated records, freeze $T = T_{\widehat{\phi}}$ at fitted parameters $\widehat{\phi}$. Training observations supply calibrated means μ^c and standard deviations σ^c:

$$h(x) = \frac{T(E_\theta(x)) - \mu^c}{\sigma^c}, \qquad q_j(u) = \frac{\alpha_j(u) - \mu_j^c}{\sigma_j^c}. \tag{5}$$

Readout $h_j(x)$ and setting $q_j(u)$ share coordinates; division is element-wise. Cross-source stability requires independent testing; neither target supervision nor monotonicity ensures it.

4.2 Observational Structure and Active Orientation

PC-type conditional-independence tests on $h(\mathcal{D}_{\text{obs}})$ estimate the observational equivalence class as a completed partially directed acyclic graph (CPDAG) (Colombo & Maathuis, 2014); finite-sample $\widehat{G}_0$ may be wrong. RGB Domino uses linear partial correlations (MEC backends: Appendix C.1).

Each round selects the target with most incident unoriented edges and remaining allowance. Predetermined tie-breaking and ID value distributions use neither the true graph nor test results (Appendix B.3). New images use frozen h. On RGB Domino candidate adjacencies, single-target records $\mathcal{I}_j$ give

$$e_{j\to k} = \left|\widehat{\text{Corr}}(h_j, h_k; \mathcal{I}_j)\right|. \tag{6}$$

Sample correlation $\widehat{\text{Corr}}$ supplies orientation evidence for initial adjacencies, not direct causal effects; targets remain visual estimates, not commands. Bidirectional responses with threshold $\tau = 0.15$ and gap $\delta = 0.05$ propose orientations. Conflict/acyclicity checks precede Meek propagation (Meek, 1995). Appendix B.1 specifies sample thresholds, one-sided rules, and rejections. Budget exhaustion retains unresolved edges.

Orientation requires reliable adjacencies and detectable responses. Weak correlation does not exclude nonlinear effects, and graph-consistency checks cannot ensure statistical correctness. Appendix B.7 tests edge-order dependence and invalid-evidence safeguards.

4.3 Local Mechanisms and New Commands

On valid, sufficiently oriented estimated graphs, ridge regression with an intercept fits $\widehat{f}_j$ from parent readouts in h coordinates (Appendix B.3), excluding interventions that replace j's mechanism. Root observations and residuals form sample pools. Invertibility does not imply linearity: nonlinear T can change relationships and residual–parent dependence. Readouts, consequences, and counterfactuals require separate tests (Appendix B.6).

Mean-outcome prediction for new commands. For $a = (j, u)$, set $q_j(u)$, resample training roots/residuals, and propagate. Convert samples to label units, then average each non-target node $k \neq j$ for $\widehat{\mu}_k(j, u)$. Inverse-calibrate and decode each sample before averaging images; do not decode a mean latent. Conversions and sampling rule (18) appear in Appendix B.4.

4.4 Decoding Counterfactual Images

Encode, calibrate, and standardize factual x to $h(x)$. For each non-root node, estimate $\eta_j^{\text{fact}} = h_j(x) - \widehat{f}_j(h_{\text{Pa}_{\widehat{G}}(j)}(x))$. Hold these residuals fixed, set $q_j(u)$, and update descendants to obtain h^{cf}. Non-target roots and non-descendants retain factual readouts. Undo standardization and calibration to reach decoder coordinates:

$$\widehat{x}^{\text{cf}} = D_\psi\left[T^{-1}\left(\sigma^c \odot h^{\text{cf}} + \mu^c\right)\right]. \tag{7}$$

Here $\widehat{x}^{\text{cf}}$ is the predicted image, T^{-1} frozen inverse calibration, and $\odot$ element-wise multiplication. Estimated residuals need not equal true exogenous causes. Appendix B.4 also specifies the separately evaluated identity-calibration, conditional-mean configuration without factual residuals.

5 Experiments

Our evidence follows three links: interventions resolve ambiguity left by matched additional observations; same-data calibration lowers mean intervention-effect error in the controlled domino cohort; state and image queries test downstream use. System comparisons assess task performance, not individual components. Decoder-only controls separate rendering gains from calibration (protocol and statistical index: Appendix A.1).

5.1 Experimental Setup

Benchmarks. Tasks comprise Domino, Pick-and-Place (PnP), Balance, Gears, CAUSAL3D Spring and Seesaw (Liu et al., 2025), and CausalWorld (Ahmed et al., 2021). Domino separately includes a seven-node continuous SCM and four-node contact dynamics. The latter uses binary events F_i for whether each domino falls and evaluates fixed/release commands through same-initial-state replay. PnP, Balance, and Gears contain six, five, and six SCM variables. Figure 4 illustrates scenes; Domino configurations and task-table scope appear in Appendices A.9 and A.3.

Baselines. System comparisons include CausalVAE (Yang et al., 2021) and ILCM, which learns from intervention-paired images (Brehmer et al., 2022). Each uses its own representation, graph extraction, latent intervention, and decoder. ICM-VAE is a given-graph generation reference (Komanduri et al., 2024); no-graph controls remove dependencies through their cohort-specific operators (Appendix B.2). Appendix G.2 reports ID quality and model selection. Weak baseline results are not attributed to an absence of intervention data.

Metrics. Graph discovery reports structural Hamming distance (SHD)↓, counting a reversed edge once, and area under the precision–recall curve (AUPRC)↑. Consequence prediction reports training-scale-normalized mean squared error (NMSE)↓ for non-target mean states. Counterfactual images use RGB mean squared error (MSE)↓. For counterfactual images, models receive only a factual image and command; ground truth (GT) replays the same exogenous context or physical initial state. Population prediction instead takes only a command. Errors in the latter two task tables are multiplied by 10^3. Completion, graph recovery, and executable queries have separate denominators (Appendix A.6).

Comparison scope. Our method uses learned T in the three task tables. Appendix A.2 distinguishes unified comparisons, given-graph references, and physical ILCM with different permissions. These descriptive summaries measure conditional error on executed queries, not end-to-end success or superiority on a common query population. Continuous-task columns retain reported ID/amplitude-shift labels; other columns compare test sets A/B (Appendix A.8). Light blue marks our method, bold marks the best point estimate, and "–" means not applicable. Acquisition efficiency uses the total budget $B = B_{\text{cal}} + B_{\text{graph}}$.

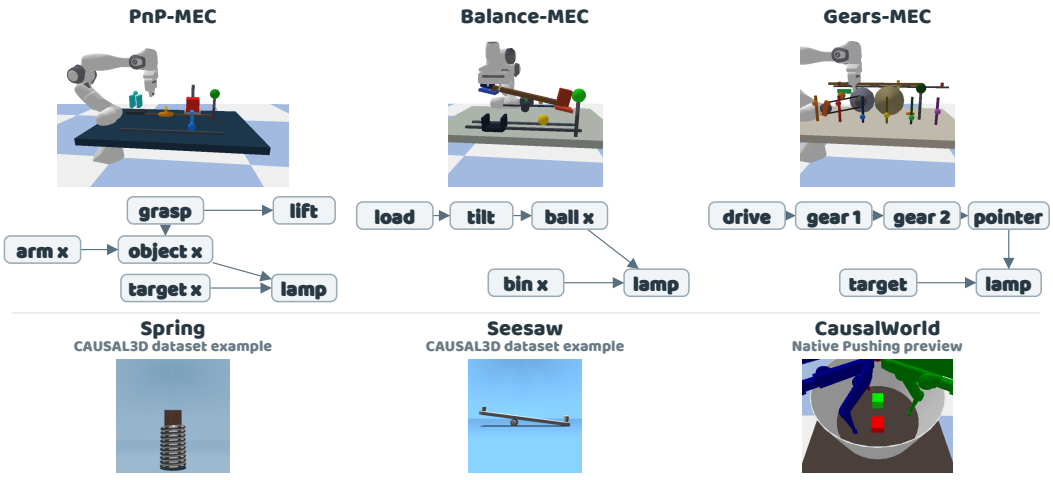


Figure 4: Example environments. Top: controlled MEC scenes and generating graphs for PnP, Balance, and Gears. Bottom: original CAUSAL3D Spring and Seesaw images and a native CausalWorld Pushing illustration. Generating graphs are not estimates; scene illustrations do not represent task-table configurations or model predictions. Sources appear in Appendix A.7.

5.2 Discovering Graphs through Environment Interventions

A three-seed control shares the visual and calibration interface and matches the number of added records for graph discovery: both intervention acquisition strategies recover the true graph with 1800 or 3600 records, while equal additional observations leave 6 unoriented edges. In this controlled task, resolving the remaining orientation ambiguity requires intervention responses; equal additional observations do not suffice. Coverage targeting does not save records over round-robin acquisition. Appendices C.2 and C.1 give the protocols and diagnostics for different orientation criteria.

Table 1: Graph discovery: each cell reports SHD↓/AUPRC↑ on its upper/lower line. Bold marks the best point estimate; light blue marks our method. Bottom: completed/planned runs, not graph-recovery rates. See Appendix G.1 for comparison scope.

	Domino		Other MEC tasks			CAUSAL3D		Causal World
Method	SCM	Physical	PnP	Balance	Gears	Spring	Seesaw	
	4	**0**	**0**	6	7	**9**	**5**	**4**
Ours	**.846**	**.982**	**.862**	.413	**.793**	**.242**	**.587**	**.276**
	6	3	3	8	8	11	7	11
CausalVAE	.313	.284	.516	**.475**	.522	.213	.342	.013
	6	2	5	**4**	**5**	17	8	16
ILCM	.194	.294	.232	.178	.177	.084	.231	.007
Completed/planned runs↑ (not graph-recovery rates)								
Ours	8/8	7/8	8/8	7/8	8/8	8/8	6/8	7/8

In Table 1, our SHD is best in six of eight columns and AUPRC in seven; SHD is 0 for physical Domino and PnP. The complete system yields more accurate structure estimates on most tasks, while the shared-interface control above shows that graph-stage intervention responses supply orientation evidence still absent from additional observations. The advantage does not cover every metric: SHD is higher than ILCM on Balance and Gears, and AUPRC is 0.242 on Spring and 0.276 on Causal-World, leaving substantial full-structure recovery error.

5.3 Intervention-Value Calibration and Effects

Table 2 compares readouts, non-target counterfactual (CF) effects, and RGB errors in the same three-seed cohort for the seven-node domino scene rendered as 64×64 RGB images. Within each seed, configurations share test contexts and commands. The first three rows share an ID-refitted decoder D'. Ordinary joint updating additionally trains a readout head and decoder, so it is included only as a different-permission reference. CF effects compare differences from the factual state and are not pooled with population mean-outcome NMSE.

Table 2: Readouts, intervention effects, and images in the same paired three-seed cohort. Errors are multiplied by 10^3; entries are means $\pm$ sample SD across seeds. The first three rows share each seed's frozen encoder, data, learned graph, and ID-refitted decoder D', fitting mechanisms in their respective coordinates. Ordinary joint updating additionally trains a readout head and decoder and is reference-only. All seeds 5/6/7 are scored; bold marks the lowest mean in the first three rows. Paired differences appear in Appendix D.4.

		Non-target CF effect NMSE↓		RGB MSE↓	
	ID readout				
Configuration	MSE↓	ID	Amp. shift	ID	Amp. shift
Uncalibrated	22.367 ± 3.941	5.463 ± 0.946	6.858 ± 1.290	1.085 ± 0.066	0.874 ± 0.046
Fully fitted affine	14.721 ± 1.177	$\mathbf{4.763} \pm 0.142$	6.656 ± 0.531	$\mathbf{1.043} \pm 0.065$	$\mathbf{0.856} \pm 0.034$
Nonlinear (ours)	$\mathbf{12.864} \pm 0.830$	4.968 ± 0.123	$\mathbf{6.515} \pm 0.418$	1.065 ± 0.063	0.878 ± 0.036
Ordinary joint updating (ref.)	11.361 ± 5.044	9.985 ± 1.611	20.449 ± 3.725	1.058 ± 0.080	0.851 ± 0.053

Calibration places visual readouts and target settings in shared intervention units; its effectiveness must be assessed through intervention consequences, not state-readout MSE alone. Ordinary joint

updating has lower mean readout error than nonlinear calibration (11.361 versus 12.864), but higher ID/amplitude-shift effect errors (9.985/20.449 versus 4.968/6.515). Ordinary updating changes both the readout head and decoder, so this comparison does not by itself establish the contribution of invertibility.

Nonlinear calibration has lower mean error than no calibration for both effects, improving 2/3 seeds in each case. Relative to affine calibration, its amplitude-shift effect error is lower, but ID effect and RGB errors are higher. These results support a reduction in mean intervention-effect error through calibration in this cohort, not nonlinear superiority across all queries. A two-seed fixed-graph Gears control also slightly improves readouts while increasing query error (Appendix D.4). Appendix D.1 defines the effect and within-cohort statistical units.

5.4 Consequence Prediction under Distribution Shifts

With models frozen, we predict the other variables' mean post-intervention states under (j, u). Table 3 separates reported ID/amplitude-shift columns from A/B tests and gives both absolute NMSEs and their ratio r. The second-group error is never smaller than the first-group error. A ratio closer to 1 indicates less relative growth, but increasing the first-group error also raises r; absolute errors must therefore accompany its interpretation.

Table 3: Conditional consequence-prediction errors, excluding blocked queries. Unmarked columns retain the reported ID/amplitude-shift labels; entries are NMSEs ($\times 10^3$) and ratio r. Columns marked † report test sets A, B, and E_A/E_B, without inferring an OOD type. Bold marks the smallest listed error summary; light blue marks our method. Lower errors are better; r alone does not define a ranking. Scope and distribution qualifications: Appendices A.2 and A.8.

	Domino		Other MEC tasks			CAUSAL3D		Causal World
Method	SCM	Physical†	PnP	Balance	Gears	Spring†	Seesaw†	†
	8.573	**6.234**	**1.771**	**2.085**	**.729**	**5.323**	**4.276**	**4.817**
	10.341	**8.291**	**3.461**	**4.101**	**1.091**	**8.064**	**6.786**	**5.378**
Ours	**0.829**	**0.752**	**0.512**	**0.508**	**0.668**	**0.660**	**0.630**	**0.896**
	123.799	228.713	51.105	130.996	75.124	186.127	156.435	293.854
	807.120	791.706	318.545	852.155	503.839	609.486	987.345	974.245
CausalVAE	0.153	0.289	0.160	0.154	0.149	0.305	0.158	0.302
	123.713	162.410	58.106	104.107	101.258	234.144	95.245	124.215
	874.392	515.288	394.825	713.652	699.789	772.011	312.261	332.211
ILCM	0.141	0.315	0.147	0.146	0.145	0.303	0.305	0.374
	34.144	51.205	57.011	102.146	93.292	204.114	98.701	41.124
	245.391	421.363	393.035	700.012	663.595	402.341	244.512	408.946
No graph	0.139	0.122	0.145	0.146	0.141	0.507	0.404	0.101

For executed queries, our point estimates are lowest for both test groups in all eight columns, with r ranging from 0.508–0.896, versus 0.101–0.507 for baselines and the no-graph ablation. Arithmetic differences between the reported second-group conditional errors are approximately 97.8%–99.8% relative to the lower CausalVAE/ILCM entry in each column. These gaps describe the listed implementations and test sets, not end-to-end success or equally large OOD advantages. Interpretation requires the distribution definition and ID fitting quality.

The no-graph rows have higher second-group error throughout. This is a cohort-specific dependency ablation, not necessarily an unchanged-mechanism propagation switch (Appendix B.2). The factual-conditioned effect NMSE in Section 5.3 evaluates a different query from the population mean error here.

5.5 Counterfactual RGB Image Generation

We fix the factual context and generate images through latent interventions and each model's decoder. Among the reported summaries in Table 4, our pixel MSE is numerically lowest in seven of eight columns: 68.6% below CausalVAE on PnP and 26.8% below ILCM on Spring, but 57.9% above ILCM on Seesaw. Whole-image MSE combines reconstruction and consequence errors. Causal interpretation requires non-target state effects and independent image readouts of target execution, descendant response, and non-descendant preservation; sharper images alone are insufficient.

Table 4: Counterfactual RGB generation: pixel MSE↓ ($\times 10^{-3}$) for executed queries only, excluding blocked queries. Bold marks the smallest listed summary; light blue marks our method; "–" means not applicable. Physical ILCM is a single-seed reference with different permissions and budget, excluded from unified ranking. Appendix A.2 specifies the comparison scope.

	Domino		Other MEC tasks			CAUSAL3D		Causal World
Method	SCM	Physical	PnP	Balance	Gears	Spring	Seesaw	
Ours	**1.473**	**.880**	**1.118**	**2.180**	**.938**	**2.384**	8.264	**5.261**
CausalVAE	2.525	1.287	3.563	4.603	1.859	6.345	11.745	7.761
ILCM	9.009	5.707	7.025	15.477	8.187	3.256	**5.234**	9.664
ICM-VAE (given graph)	–	–	–	–	–	7.342	9.753	–
No graph	4.215	2.314	2.161	5.014	1.859	15.234	32.219	19.281

Figure 5A summarizes the proposed counterfactual query path. Panel B compares the uncalibrated conditional-mean control, CausalVAE, and ILCM in one physical context; its measured images are not outputs of panel A. Appendix Figure A4 gives paired images with learned calibration and factual residuals, and Appendix D.3 reports independent image-readout evaluation.

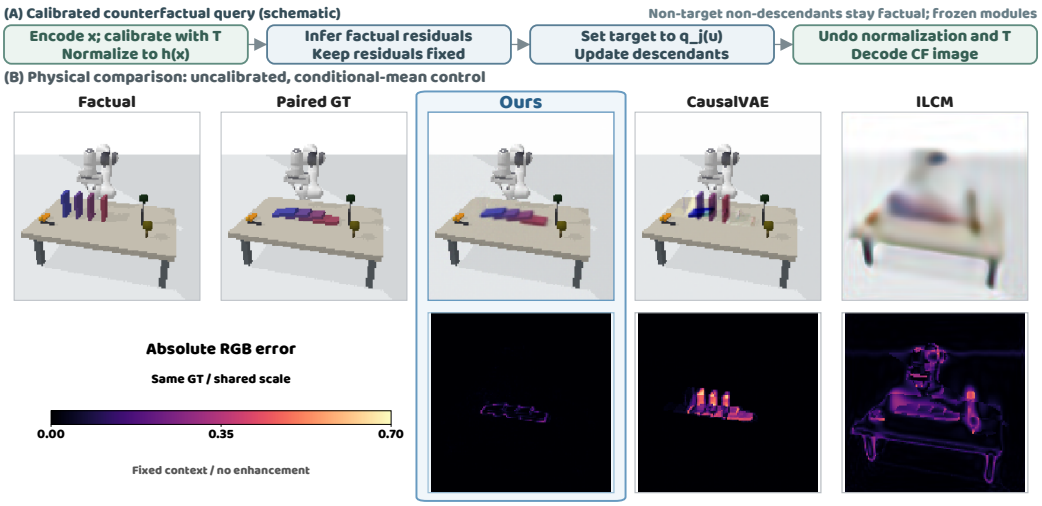


Figure 5: (A) Schematic of learned calibration, factual-residual propagation, and inverse-calibrated decoding. (B) Physical comparison for making the first domino fall ($F_1 = 1$); blue marks the uncalibrated conditional-mean control. Factual images, paired GT, and model outputs are above RGB mean absolute errors with a shared scale. Images are clipped to $[0, 1]$ for display only. ILCM is a single-seed reference with different training permissions (Appendix A.9).

Paired decoder control. Replacing only the original D with the ID-refitted decoder D' lowers amplitude-shift CF pixel MSE by 49.32% (3/3), with unchanged state predictions. Uncalibrated and affine models also benefit (Appendix E.1); Table 2 instead shares D' to compare calibration. Appendix F.3 gives other component and budget controls.

6 Conclusion

With known node semantics, verified intervention values calibrate visual readouts so that node settings, causal mechanisms, and decoding use consistent intervention units. Matched controls support improvements for some intervention queries, but the relative benefits of affine and nonlinear calibration depend on the setting; better readouts do not automatically improve downstream predictions. Environment interventions provide direction evidence absent from additional observations. This finding is distinct from targeting efficiency and gains in image reconstruction.

AI Use Statement

Generative AI tools assisted result analysis, writing and translation, methodological and experimental discussions, code implementation, and figure preparation. AI-assisted analyses were subsequently checked by the authors. The authors led the research motivation, methodology, and scientific decisions, and take responsibility for the experiments, reported results, and final manuscript.

Reproducibility Statement

Experimental settings, evaluation protocols, and disaggregated results are described in the appendix. Appendix A.2 summarizes comparison permissions and statistical units; settings and results are interpreted within their respective experimental groups.

References

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Manuel Wüthrich, and Stefan Bauer. 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.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 372–407. PMLR, 2023. URL https://proceedings.mlr.press/v202/ahuja23a.html.

Yashas Annadani, Panagiotis Tigas, Stefan Bauer, and Adam Foster. Amortized active causal induction with deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1403. URL https://papers.neurips.cc/paper_files/paper/2024/hash/4e2c6423d724370b36c3a7612f25b78c-Abstract-Conference.html.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco S. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/fa567e2b2c870f8f09a87b6e73370869-Abstract-Conference.html.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(116):3921–3962, 2014. URL https://www.jmlr.org/papers/v15/colombo14a.html.

Patrick Cooper and Alvaro Velasquez. Active causal experimentalist (ACE): Learning intervention strategies via direct preference optimization. arXiv:2602.02451, 2026. URL https://arxiv.org/abs/2602.02451v2. Version 2, June 2026; preprint.

Ziyi Ding, Xianxin Lai, Weiyu Chen, Xiao-Ping Zhang, and Jiayu Chen. CausalVAE as a plug-in for world models: Towards reliable counterfactual dynamics. In *Computer Vision – ECCV 2026*, volume 17017 of *Lecture Notes in Computer Science*, pp. 278–293. Springer, 2026. doi: 10.1007/978-3-032-37252-9_16. URL https://link.springer.com/chapter/10.1007/978-3-032-37252-9_16.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012. URL https://www.jmlr.org/papers/v13/hauser12a.html.

Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014.

Aneesh Komanduri, Yongkai Wu, Feng Chen, and Xintao Wu. Learning causally disentangled representations via the principle of independent causal mechanisms. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 4308–4316, 2024. doi: 10.24963/ijcai.2024/476. URL https://www.ijcai.org/proceedings/2024/476.

Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, and Animesh Garg. Causal discovery in physical systems from videos. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/6822951732be44edf818dc5a97d32ca6-Abstract.html.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 13557–13603. PMLR, 2022. URL https://proceedings.mlr.press/v162/lippe22a.html.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. BISCUIT: Causal representation learning from binary interactions. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216, pp. 1263–1273. PMLR, 2023a. URL https://proceedings.mlr.press/v216/lippe23a.html.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *International Conference on Learning Representations*, 2023b. URL https://arxiv.org/abs/2206.06169.

Disheng Liu, Yiran Qiao, Wuche Liu, Yiren Lu, Yunlai Zhou, Tuo Liang, Yu Yin, and Jing Ma. CAUSAL3D: A comprehensive benchmark for causal learning from visual data. *arXiv preprint arXiv:2503.04852*, 2025. URL https://arxiv.org/abs/2503.04852.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 403–410, 1995. URL https://arxiv.org/abs/1302.4972.

Rafael Rodriguez-Sanchez, Cameron Allen, and George Konidaris. From pixels to factors: Learning independently controllable state variables for reinforcement learning. arXiv:2510.02484, 2025. URL https://arxiv.org/abs/2510.02484. Preprint.

Sorawit Saengkyongam, Elan Rosenfeld, Pradeep Ravikumar, Niklas Pfister, and Jonas Peters. Identifying representations for intervention extrapolation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aa9508b2577e0d9c824da9e2e27fe9d6-Abstract-Conference.html.

Nino Scherrer, Olexa Bilaniuk, Yashas Annadani, Anirudh Goyal, Patrick Schwab, Bernhard Schölkopf, Michael C. Mozer, Yoshua Bengio, Stefan Bauer, and Nan Rosemary Ke. Learning neural causal models with active interventions. arXiv:2109.02429, 2021. URL https://arxiv.org/abs/2109.02429. Revised March 2022.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. URL https://arxiv.org/abs/2102.11107.

Sumedh A Sontakke, Arash Mehrjou, Laurent Itti, and Bernhard Schölkopf. Causal curiosity: RL agents discovering self-supervised experiments for causal representation learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 9848–9858. PMLR, 2021. URL https://proceedings.mlr.press/v139/sontakke21a.html.

Valentin Thomas, Emmanuel Bengio, William Fedus, Jules Pondard, Philippe Beaudoin, Hugo Larochelle, Joelle Pineau, Doina Precup, and Yoshua Bengio. Disentangling the independently controllable factors of variation by interacting with the world. arXiv:1802.09484, 2018. URL https://arxiv.org/abs/1802.09484.

Panagiotis Tigas, Yashas Annadani, Andrew Jesson, Bernhard Schölkopf, Yarin Gal, and Stefan Bauer. Interventions, where and how? experimental design for causal models at scale. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2203.02016.

Zirui Yan, Emre Acartürk, and Ali Tajer. Reward-oriented causal representation learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1381. URL https://papers.neurips.cc/paper_files/paper/2025/hash/3b04bba0f95caacbd9aa9033e2c913c9-Abstract-Conference.html.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021.

Zhiyu Zhao, Haoxuan Li, Haifeng Zhang, Jun Wang, Francesco Faccio, Jürgen Schmidhuber, and Mengyue Yang. Curious causality-seeking agents in open-ended worlds. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5145. URL https://papers.neurips.cc/paper_files/paper/2025/hash/e21c92a1661709c3caf21fdd958bac99-Abstract-Conference.html.

A Experimental Protocols, Statistical Units, and Environments

A.1 Research questions and protocol index

The experiments distinguish five questions:

Systems Tables 1, 3, and 4 score graphs, population-state means, and RGB using each method's representation and query path. Physical ILCM has different information access.

Orientation Appendices C.2 and C.1 separate the three-seed equal-added-record control from the eight-seed MEC study. Existing calibration interventions mean the former is not a wholly passive-versus-interventional comparison.

Calibration Table 2 fixes each seed's encoder, learned graph, queries, and ID-refitted D', refitting mechanisms in each coordinate system. It reports three-seed means/SDs; Appendix D.4 gives paired differences. Ordinary updating also changes the readout and decoder.

Decoding Appendix E.1 changes only D to D' in a three-seed pairing, leaving causal-state/effect outputs unchanged.

Budget Appendix F.3 compares policies under $B = B_{\text{cal}} + B_{\text{graph}}$ with batch orientation checks, reporting actual calls, recovery, and exhaustion for all three seeds.

All test-time modules, scales, and query rules are frozen. Appendix D.2 uses configuration-specific frozen graphs learned from the same evidence without resampling; Appendix D.3 relearns graphs and mechanisms from a shared acquisition trajectory. Neither substitutes for the main fixed-graph control. Common inputs/decoders do not imply identical mechanism parameters; fixed-graph Gears does not test graph rediscovery.

Comparison and denominator rules are centralized in Appendix A.2; metrics and distributions are in Appendices A.6 and A.8. Gears failures, affine-favoring conditions, and budget exhaustion retain their cohort-specific interpretation.

A.2 Comparison scope and statistical units

The reported system comparisons use a unified data/supervision setting, with learned T for our method. Appendix A.3 distinguishes system comparisons, physical-scene adaptations, and mechanism controls. ICM-VAE is a given-graph generation reference, not a graph-discovery comparator. Physical ILCM has different paired-data permissions and budget and is excluded from matched rankings: $n = 1$, with four targets, two commands, and 32 contexts giving 256/256 queries. Other MEC baseline diagnostics concern the listed seed300 checkpoints ($n = 1$), not every identically named main-table model.

Shared evaluation definitions and comparison permissions. Task comparisons use the stated input and score definitions: factual image plus target/value for counterfactual generation, target/value without factual conditioning for population prediction, and non-target states for consequence errors. Normalization comes from training data; models and query rules remain frozen at test time. Graph scores, state NMSE, and pixel MSE are different outcomes and are not pooled. Equal data and supervision access does not require identical native architectures. Additional graph information, paired-data access, or trainable modules defines a separate reference rather than a matched component ablation. The detailed matched controls below provide stronger component attribution than aggregate system rankings. Equal-total-budget acquisition comparisons are confined to Appendix F.3; task-table errors alone do not establish equal-cost superiority.

Queries, blocked cases, and denominators. Tables 3 and 4 score executed queries; blocked cases receive neither zero error nor arbitrary penalties. Equal task definitions do not establish equal retained-query sets when methods fail on different cases. Table 1's completion fractions count completed/planned runs, not exact-graph recovery, prediction seeds, or executed/planned queries. Main summaries therefore describe conditional error, not end-to-end success or superiority on a common

retained-query population. Table A2 separates these known run counts from prediction denominators. Supplementary counts and valid sets apply only to their own cohorts; no prediction denominator is inferred from the graph table.

Within-seed aggregation and across-seed uncertainty. For Table 2, every configuration uses the same 64 factual contexts, seven targets, and two commands per target per split: 896 CF queries per seed/split, with seeds 5, 6, and 7 all retained. Targets and commands are equally weighted within each seed; the table reports the mean and sample SD of the three seed-level errors, not an SD over 2688 independent images. Appendix D.4 computes paired differences from the same seeds; subtracting two SDs is not a paired SD. Main task-table cells without uncertainty are point estimates whose aggregation weights and per-seed variances cannot be reconstructed from the reported means. They borrow neither this three-seed denominator nor its SD. The separate physical-control aggregate in Table A1 likewise is not used to estimate a paired calibration gain.

Configurations and attribution. Table 2 shares each seed's encoder, data, graph, and D' across three calibration variants ($n = 3$, 896 CF queries per split per seed); ordinary updating also trains readout/decoder parameters. The *no-graph* construction is cohort-specific and may rebuild root models instead of keeping fitted mechanisms fixed (Appendix B.2). Figure 5A shows the full query path schematically, while panel B shows an uncalibrated physical conditional-mean control; Appendix Figure A4 uses learned calibration and factual residuals. Decoder gains and orientation evidence alone do not isolate calibration.

A.3 Evaluation configurations and statistical scope

The task tables report learned T under a unified data and supervision protocol. Table A1 distinguishes system comparisons, physical-scene adaptations, and mechanism controls; counts and configurations are interpreted separately for each cohort.

Table A1: Record links for task-table and supplementary cohorts. Known denominators belong only to the listed runs; method names or equal rounded values do not transfer them.

Result location	Configuration and record	Queries, permissions, and scope
Summary cells in Tables 1, 3, and 4, except the next row	Learned T for our method under a unified data and supervision protocol. Cells without intervals report point estimates.	The lower part of Table 1 counts pipeline completion, not query execution. Prediction errors condition on executed queries; they estimate neither end-to-end success nor superiority on a common query population.
Physical ILCM in Table 4	`run_s42_v1`, seed 42; native implicit mechanisms and its own decoder; epoch 120 selected by ID validation.	1280 training pairs, 256 validation pairs; 32 contexts$\times$4 targets$\times$2 commands, 256/256 queries. Paired GT matches archived seed-42 queries, not necessarily the retained sets of other task-table summaries. Permissions and budget differ (Appendix A.9).
MEC mechanism controls in Appendix C.4	Seeds 300/301, learned T; frozen representation, calibration, and graph mechanisms; a shared decoder trained on 2400 independent ID images.	PnP/Gears retain 2/2 seeds and Balance 1/2; respectively 384/384/320 CF queries per valid seed. Command ranges and 2048-sample outcome evaluation apply only to this cohort.
Uncalibrated physical control (aggregate)	No numerical calibration; aggregate full-image pixel MSE is 0.930 ($\times 10^{-3}$).	Reported separately from the paired control below; its seed count, SD, query count, and training budget are not transferred. This aggregate is not used with the main-table learned-T result to estimate paired calibration gains.
Paired conditional-mean control in Appendix A.9	No numerical calibration or addition of factual residuals to descendants; common completed seeds 42 and 314.	256 common queries per seed; our method completes 2/3 seeds and CausalVAE 3/3, with more concept labels for the latter. Disaggregated errors are reported in Appendix A.9.

Benchmark protocol summary. Table A2 lists the task-table information separately from the independently documented controls below. The common input, metric, freeze, and exclusion rules are defined in Appendix A.2; they do not determine a scene's numerical sampling range or executed-query set.

Table A2: Task-table protocol index. Run completion repeats Table 1, not an exact-graph or prediction-query success rate. Prediction seed counts n, executed queries Q, and calibration-inclusive acquisition totals B are distinct fields.

Benchmark	Task variables	Operation and test grouping	Completed/ planned runs	Prediction n, Q, B
Domino SCM	7 continuous states: pusher, four dominoes, bell, lamp	Single-state setting; reported ID/amplitude groups	8/8	—
Physical Domino	4 binary final toppling events	Fix upright / tilt and release; test sets A/B	7/8	—*
PnP	6 SCM variables	Single-state setting; reported ID/amplitude groups	8/8	—
Balance	5 SCM variables	Single-state setting; reported ID/amplitude groups	7/8	—
Gears	6 SCM variables	Single-state setting; reported ID/amplitude groups	8/8	—
CAUSAL3D Spring	Scored state vector —	Specific targets/ranges —; test sets A/B	8/8	—
CAUSAL3D Seesaw	Scored state vector —	Specific targets/ranges —; test sets A/B	6/8	—
CausalWorld	Specific task and scored state vector —	Specific targets/ranges —; test sets A/B	7/8	—

An em dash means that the aggregate does not specify this field, not zero or a failed run. Exact training/test generators and numerical ranges are not supplied by the reported group labels. A/B alone identifies no OOD axis. *Physical ILCM is a separately documented reference: seed 42, 1280 training pairs, 256 validation pairs, and 256/256 CF queries. Its 3072 pair-generation image calls use different permissions/accounting, not a matched $B_{\text{cal}} + B_{\text{graph}}$ total. These values do not fill other methods' cells.

Independently specified controls. The following records belong to their named controls and are not substituted for task-table protocols.

- **Matched-acquisition Domino (Appendix D.3).** The seven-node scene uses all targets, seeds 5/6/7, 6000 observations, and 896 calibration records. New graph records number 1800/3600/3600, so B is $896 + 1800$ for seed 5 and $896 + 3600$ for seeds 6/7. Independent ID contexts use commands $\mu_j^s \pm 0.75\sigma_j^s$; amplitude tests use $\mu_j^s \pm 2\sigma_j^s$, with training-label statistics. Each seed executes 1792 CF queries across both splits. Its graph/decoder conditions remain distinct from Table 2.
- **Paired physical control (Appendix A.9).** Four binary events are queried by fix/release commands. Training uses 512 labeled images, 128 ID validation images, and 512 independent observations. Seeds 42/314 execute, while 2718 is blocked: $n = 2/3$. Thirty-two held-out initial states give 256 CF queries per completed seed. The known graph-acquisition component is 256 records; a calibration-inclusive total is not inferred. These counts do not describe the separate physical-control aggregate in Table A1.
- **MEC mechanism controls (Appendix C.4).** PnP has arm-x/grasp markers, object x/lift, target x, and lamp; Balance has load difference, tilt, ball x, bin x, and lamp; Gears has drive, two gear angles, pointer, target, and lamp (Appendix A.7). The PnP markers do not represent the live pose of the fixed HOME arm. All scene targets are queried using the training mapping's physical origin $\pm 0.75/\pm 2$ scale units. A shared decoder uses 2400 independent ID images; 32 CF contexts and two commands per target give 384/320/384 queries per executable seed for PnP/Balance/Gears. Seeds 300/301 both execute for PnP/Gears; only 300

executes for Balance. Population tests use 2048 model and simulator samples per command. Full encoder-training distributions and calibration-inclusive totals are not supplied for these controls.

The controlled MEC node definitions, Spring/Seesaw source-image adapters, and CausalWorld Pushing illustration are not substitute definitions of the scored task-table columns.

Configuration distinctions and interpretation. The PnP, Balance, and Gears MEC mechanism controls use learned T, with disaggregated statistics in Appendix G.1; the rigid-body conditional-mean control omits numerical calibration and does not estimate calibration gains. Equal values do not transfer seed denominators or data permissions across cohorts. The cross-environment tables report aggregates rather than complete per-run configurations and per-query records; their point estimates establish neither statistical significance nor superiority on a common query population. Other diagnostic cohorts and the scene illustrations in Figure 4 cannot supply the scored nodes, units, and true graphs for Spring/Seesaw, the specific CausalWorld task and states, or the training/test distributions for these columns. Table 3 retains reported ID/amplitude labels for continuous columns and A/B tests for the others. Appendix A.8 gives distributions and queries for the named controls.

A.4 Causal terminology

Graph and independence. In $G = (V, E)$, $i \to j$ means that j's structural equation directly depends on i; E is the edge set, not encoder E_θ. A DAG has no directed cycle. Parents are $\mathrm{Pa}_G(j) = \{i : i \to j \in E\}$; descendants $\mathrm{De}_G(i)$ are reachable by directed paths and exclude i. Thus $1 \to 2 \to 3$ transmits an effect without requiring $1 \to 3$, and response correlation is not a direct edge.

For disjoint index sets A, B, C, $s_A \perp\!\!\!\perp s_B \mid s_C$ denotes conditional independence. The Markov condition links graphical d-separation to independence; faithfulness excludes extra independences caused by exact cancellation (Colombo & Maathuis, 2014). Zero correlation is not general independence. Removing arrows gives a skeleton; PDAGs may have unoriented edges, while a CPDAG represents an observational equivalence class (Hauser & Bühlmann, 2012). Finite-sample outputs can be incorrect or inconsistent; arbitrary completion is not identification.

Intervention and query. $\mathrm{do}(s_j = u)$ replaces the target mechanism; conditioning $P(s_k \mid s_j = u)$ only selects observations with the original causes intact. Here u is an attained, verified state, not an unchecked command u_{cmd}; whether a physical action implements it is task-specific. Calibration T changes the visual mapping, not the environment.

Causal sufficiency, independent exogenous disturbances, and no selection bias are environment assumptions, not consequences of encoding. A fitted residual η_j is not observed ϵ_j. Factual-image queries preserve estimated context; population queries resample it. Hidden single-frame information can prevent unique individual-counterfactual identification, so paired GT supplies evaluation rather than an identification guarantee. Full model notation and query equations are in Appendices A.5 and B.4.

Labels establish node identities; calibration adjusts their numerical readouts. Training units, $\alpha_j(u)$, $q_j(u)$, and inverse mappings are frozen. Amplitude shifts, new mechanisms, and new views are distinct; static graph evaluation is not temporal dynamics. Budget $B = B_{\text{cal}} + B_{\text{graph}}$ counts returned records, separately from batches and optimization steps. Scoring definitions appear in Appendix A.6.

A.5 Full model notation and fitted residuals

The true SCM comprises the graph, structural equations, and exogenous distribution defined in the main text:

$$\mathcal{S}^\star = (G^\star, \{f_j^\star\}_{j=1}^K, P_\epsilon), \qquad G^\star = (V, \mathcal{E}^\star).$$

The learned objects can be written as

$$\widehat{\mathcal{M}} = \big(E_\theta, T_{\widehat{\phi}}, \widehat{G}, \{\widehat{f}_j\}_{j=1}^K, D_\psi\big), \qquad h_j = \widehat{f}_j(h_{\mathrm{Pa}_{\widehat{G}}(j)}) + \eta_j.$$

E_θ, D_ψ are the encoder and decoder, $T_{\widehat{\phi}}$ the fitted calibration, $\widehat{G}$ the estimated graph, and $\widehat{f}_j$ the local predictive mechanisms. The fitted residual η_j is not recovered true exogenous noise. Observational,

interventional, and factual images enter this model through the same frozen interface. The main method and Appendix B.4 detail its numerical approximations and query procedure.

A.6 Evaluation notation and scene-image provenance

Graph scores are SHD $= N_{\text{add}} + N_{\text{del}} + N_{\text{rev}}$, where additions, deletions, and reversals transform the estimate into the true graph and a reversal counts once, and AUPRC $= \int_0^1 \text{Prec}(\rho)\,\mathrm{d}\rho$, where ρ is recall and Prec precision. AUPRC ranks edge scores; it is distinct from the completion fractions in Table 1.

For MEC partial-graph controls, each unordered pair $i < j$ contributes at most one error: absent/extraneous adjacency, reverse direction, an unoriented true edge, or bidirectional conflict. For an adjacency shared by both graphs, only the correct single arrow scores zero; the total is bounded by $K(K-1)/2$. Wrong directions and unresolved edges are also reported separately; SHD is not a graph-validity or execution gate. System-table scoring scope is given in Appendix G.1. Response correlation is neither a total-effect estimate nor an edge to every affected descendant; its decision rules appear in Appendix B.1.

State means, images, and normalization. Table 3 evaluates the model mean $\widehat{\mu}_k(j,u)$ against the simulator's interventional mean $\mu_k(j,u)$ for each non-target node $k \neq j$ under $a = (j,u)$, with per-node scales fixed from training data before squaring errors. The target is excluded; low error evaluates prediction, not unique recovery of the true mechanism. This is distinct from factual-conditioned $\widehat{x}^{\text{cf}}(x,a)$ and next-frame prediction. Scene distributions are qualified in Appendix A.8.

The ratio $r = E_{\text{A}}/E_{\text{B}}$ uses rounded error summaries, not a mean of per-run ratios or a significance test. For $E_{\text{B}} \geq E_{\text{A}} > 0$, higher r means less relative growth but may also reflect worse first-group error; absolute errors are required. A zero denominator gives no finite ratio. A/B mean ID/amplitude shift only where that shift is defined.

Notation and fixed quantities. $q_\theta(z \mid x)$ is the encoding posterior, while $q_j(u)$ is a model node setting; U_t is an unoriented-edge set, $\mathcal{U}_j$ a feasible value range, and $\mathcal{A}_t \subseteq \mathcal{J}$ the available targets. The coefficient vector a_j differs from operation a. Training labels determine μ_j^s, σ_j^s, and frozen-T training observations determine μ_j^c, σ_j^c. The superscripted μ_j^s, μ_j^c are fixed normalization statistics, distinct from the command-dependent interventional mean $\mu_k(j,u)$. Standard deviations use stabilization 10^{-8} and are not re-estimated on test data. Fitted calibration parameters $\widehat{\phi}$ and residuals η_j are not true environmental parameters/noises. Equation (4) averages permitted training records, with λ balancing target error and observational preservation.

A.7 Environment scenes and generating-graph provenance

For the three MEC scenes in the upper row of Figure 4, the first observational state under fixed sampling seed 20260924 is re-rendered with the original renderer, without image augmentation or model prediction. PnP is displayed at 512×768 and Balance/Gears at 512×512; main-experiment input sizes and samples are unchanged. Causal edges are read directly from each environment's adjacency matrix, and the robot arm retains the fixed HOME pose used in statistical frames. Graph rearrangement changes only node positions, not edges. The figure illustrates controlled SCMs and rendering interfaces; it does not imply that these relations were identified from robot contact dynamics.

PnP-MEC

arm_x $\rightarrow$ obj_x $\leftarrow$ grasp, grasp $\rightarrow$ obj_lift, obj_x $\rightarrow$ led $\leftarrow$ target_x. In the figure, arm x and grasp denote persistent markers of commanded position and grasp opening, not the live end-effector position of the arm fixed at HOME. Object x and lift denote the object's horizontal and vertical states, target x denotes target position, and lamp denotes indicator state.

Balance-MEC

load_delta $\rightarrow$ tilt $\rightarrow$ ball_x $\rightarrow$ led $\leftarrow$ bin_x, corresponding to load difference, beam tilt, ball position, indicator state, and collection-bin position.

Gears-MEC
drive $\rightarrow$ g1 $\rightarrow$ g2 $\rightarrow$ ptr $\rightarrow$ led $\leftarrow$ target, corresponding to drive input, the angular states of two gear stages, pointer position, indicator state, and target-dial position.

These three graphs correspond to independent controlled MEC implementations; their protocols and query results appear in Appendices C.1 and C.4. Other scenes in the main tables use their own node sets and generating relations; the illustrated SCMs are not the true graphs of those environments.

The Spring and Seesaw images in the lower row are the first samples in the already fixed test lists of the official data: Spring:1004 and Seesaw:1001. They use the original 256×256 PNGs directly, without selection by model error or visual quality. CausalWorld uses the native Pushing task in official `causal-world==1.2`, with fixed seed 20260925 and exactly one default reset. The official cameras and geometry are retained, and a 320×240 preview is rendered on CPU with PyBullet DIRECT and TinyRenderer, without action search. These three images illustrate environment appearance only. The CausalWorld preview is not claimed to correspond to the specific task configuration in the main tables, and none of the three images is a GT counterfactual or model prediction.

A.8 Scene-specific test distributions

ID/OOD denotes a relation between training and test generators, not a score. Independent contexts from the training generator are ID; changes in commands, initial states, appearance, or mechanisms define different shifts. Models, normalization, and evaluation seeds are frozen. The following ranges belong to their specified diagnostic cohorts, not every aggregate column.

Domino-SCM
The independent pixel-backbone controls use training-anchor mean ± 0.75 standard deviations for ID and ± 2 for amplitude tests. A separate protocol uses $\pm 1.15 / \pm 1.75$. Gaussian training commands have full support, so these are amplitude-distribution changes rather than strict out-of-support extrapolation.

Other MEC PnP/Balance/Gears independent baseline comparisons use training-concept mean ± 0.75 standard deviations (ID) and ± 2 (amplitude shift). These ranges do not define every main-table run.

Physical Domino
Four final-state events are binary: command 0 fixes the target upright and 1 tilts/releases it. Held-out contexts from the same initial-state generator are paired ID tests; binary commands alone do not define amplitude OOD.

Spring, Seesaw, CausalWorld
Table 3 reports two test datasets, A and B, not method or supervision groups. Their scores do not identify an amplitude, initial-state, appearance, or mechanism shift. Other CausalWorld task protocols cannot supply this column's configuration or establish cross-domain robustness.

The † columns are reported conditional A/B summaries without an identified OOD type or established common retained-query population; the reported labels on continuous columns retain the scope in Appendix A.3. Identifying a shift additionally requires the actual training/test command or context distributions and changed dimensions. No missing ranges are supplied from a different cohort.

A.9 Experimental settings and complete disaggregated results

This appendix provides training data, model selection, sample denominators, and disaggregated errors for the identity-calibration controls and rigid-body queries. Each configuration uses the training and query protocol described below.

Shared coordinate settings. The uncalibrated controls in this appendix use encoder readouts directly and are reported separately from the learned-calibration configurations in the main tables. SCM controls share the frozen visual model and data pool for each seed. Rigid-body controls represent four toppling events and propagate intervention outcomes through conditional means. These results analyze graph propagation, not the gain from nonlinear calibration.

SCM dominoes. The seven nodes are a pusher, four dominoes, a bell, and a lamp; images are generated by kinematic rendering driven by structural equations. RGB64 and RGB128 each use 10 training seeds and 6000 observational training records, with 350 epochs of visual-model training. The subsequent interface collects 14 batches of 64 prespecified ID intervention records; the same record pool fits mechanisms in the corresponding coordinates. Each seed has 64 same-noise paired counterfactual queries per target. RGB64 has 8 downstream-executable seeds, including 1 with a complete but incorrect graph; RGB128 has 9. The remaining seeds retain unoriented edges. Means and standard deviations are computed over valid seeds; both resolutions have a planned denominator of 10. Detailed results appear in Appendix G.1.

Paired rigid-body conditional-mean control. The settings and disaggregated statistics below describe runs with seeds 42, 314, and 2718. They are reported separately from both the uncalibrated physical aggregate in Table A1 and the main-table learned-calibration summary, and do not supply seed denominators, standard deviations, or budgets for either. Four nodes indicate whether each domino ultimately falls; a deviation from vertical of at least 10° is labeled 1. $F_j = 0$ is implemented by fixing the domino upright, and $F_j = 1$ by initially tilting and then releasing it. The generator checks the target's actual settled value. The robot arm, bell, and lamp are fixed scene components. Factual and counterfactual scenes share the physical initial state and other parameters; inputs are RGB128 images and intervention commands. The 32 test initial states, four targets, and two settings form 256 paired queries. This evaluation directly concerns same-initial-state replays under fix/release commands. Checking target attainment alone does not establish modularity of a final-state SCM: physical actions may still affect other variables through paths absent from the binary-event representation. It therefore does not establish that the ideal hard-intervention assumption in Eq. (2) holds in the physical system. Our method uses 512 concept-labeled training images, 128 ID validation images, and 512 independent observational images for skeleton learning. The two complete runs each execute two rounds of 128 intervention records. The visual model is trained for 100 epochs; the encoder is then frozen while the decoder is post-fitted for 100 epochs on the same training images, reconstructing them from the frozen encoder's posterior means. Candidates are selected only by ID validation. The three prespecified seeds are 42, 314, and 2718; the last stops before downstream queries because of partial-graph ambiguity.

CausalVAE. We retain the authors' four-concept model with four dimensions per concept, training losses, DAG module, and latent-intervention path, adapting RGB resolution and label scales; the graph is initialized at zero. The native intervention operator uses different branches for the first two and last two concept slots, and this table reports the task adaptation of that fixed implementation. Inference uses deterministic latent means and takes only the factual image and permitted command as inputs. The 1280-image group contains 512 anchoring, 512 observational, and 256 intervention images, with all concept labels supplied to the model. It therefore uses more labels than our method; both are scored on the same 32-initial-state query set. We report both the original decoder and an additionally ID-refitted decoder; counterfactual comparisons use the stronger original decoder.

Table A3: Disaggregated errors for the paired rigid-body dominoes control, in units of 10^{-3}. Full-image results are means $\pm$ sample standard deviations over common completed seeds 42 and 314; they do not describe the uncalibrated physical aggregate in Table A1 or the main-table learned-calibration summary.

Configuration	Full-image MSE	Keep upright	Make fall
Our method (conditional-mean propagation)	0.880 ± 0.016	1.282	0.478
Factual-residual propagation control	0.924 ± 0.026	0.997	0.850
Target-only latent editing	1.373 ± 0.001	1.064	1.682
CausalVAE, 512 images, original decoder	1.617 ± 0.153	1.305	1.929
CausalVAE, 1280 images, original decoder	1.287 ± 0.039	1.051	1.523
CausalVAE, 512 images, ID refit	1.946 ± 0.077	1.640	2.252
CausalVAE, 1280 images, ID refit	2.272 ± 0.071	1.764	2.780

Interpreting the disaggregated differences. Conditional-mean propagation directly uses local regression means for descendants without adding factual residuals back. Its advantage mainly comes from cascading toppling. For keep-upright queries, CausalVAE's original decoder outperforms this

branch; retaining factual residuals mitigates the difference. The four binary events may omit hidden initial triggers that determine a particular fall, so pixel error and preservation of individual outcomes are reported separately. Decoder refitting improves ID reconstruction but raises CausalVAE's counterfactual error, showing that the two metrics measure different abilities.

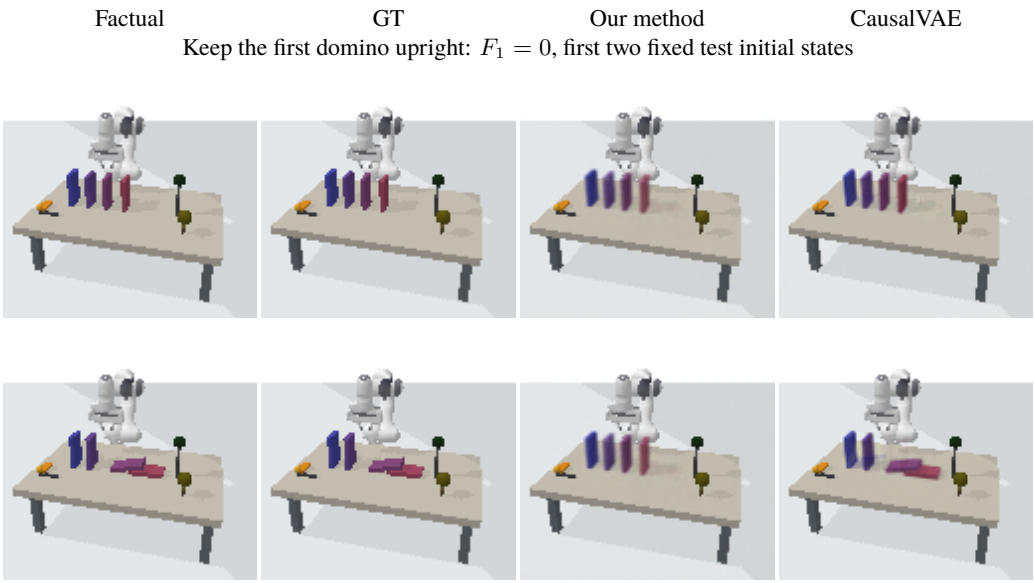


Figure A1: The same seed 42 and factual initial states as in the main text, with the command changed to $F_1 = 0$. The second row shows conditional-mean propagation changing a descendant that fell independently in the factual scene. Predicted pixels retain the original output.

Intervention-distribution evaluation. ID commands are ± 1.15, and amplitude-shift commands are ± 1.75. Model and GT samples are generated separately, and their mean images and means in frozen original-encoder coordinates are compared. Original-coordinate error is visual-readout error, not independent physical-state error. Seeds that fail downstream are not filled with zero error. The two resolutions in Appendix G.1 use different valid-seed sets and are not inserted into the main-text NMSE summary table.

Inputs to graph-backend comparisons. PC receives the shared encoder's observational matrix $Z_{\text{obs}} \in \mathbb{R}^{N \times K}$; intervention-aware graph learners additionally receive the corresponding records and target mask $M \in \{0, 1\}^{N \times K}$. N is the number of records and K the number of nodes. An entry of 1 in M indicates that the corresponding node was directly intervened on in that record. The same encoder, normalization statistics, and data permissions localize differences to the graph-learning backend. This comparison is interpreted separately from the end-to-end comparison with CausalVAE, which trains its own representation from images.

ILCM adaptation to the physical scene in Figure 5B. ILCM uses the same physical environment and RGB128 rendering settings as the paired GT in Figure 5B, training a separate four-dimensional continuous-latent model. Its official image network does not directly support resolution 128, so one native residual scaling layer is added to both the encoder and decoder of the 64-resolution branch. The implicit mechanisms, Gaussian interface, and training objective retain the original implementation. Training uses seed 42, 1280 training image pairs, and 256 independent ID validation pairs. Each pair shares an initial context, and pairs with no post-operation change are retained. Training consists of 20 pretraining epochs and 100 mechanism-training epochs, with Adam learning rate 10^{-4} and batch size 32; the checkpoint is selected by the ID validation objective over the full mechanism stage. Training accesses neither intervention-target labels nor the true graph. After training, only training-set event labels are used to match the latent permutation, and latent means for the two event classes define command endpoints.

For queries, the factual image is first encoded into noise coordinates. The native conditional solution function converts the target value to target noise while retaining all other noise coordinates; ILCM's own decoder then generates the image. Generation calls neither our mechanisms nor our decoder and does not take GT as input. Figure 5B fixes seed 42, context 59000000 (test-initial-state index 0), and the command to topple the first domino, without selecting a case by image quality. The paired data require 3072 simulation calls, a different training budget from the other models; this run's mean over 256 queries appears in the corresponding main-table cell. The four binary toppling events and physical fix/release commands do not fully satisfy the identification assumptions of continuous ILCM, so algebraic intervention checks do not establish identification of causal semantics.

This adaptation produces blurry outputs and weak command responses. Across 32 fixed test contexts, changing the first domino's command from 0 to 1 gives an RGB mean absolute change of 2.28×10^{-5} in ILCM predictions, compared with 9.91×10^{-3} in paired physical GT. This configuration is not an adequately trained counterfactual baseline. Figure 5B shows its actual outputs; the conclusion is limited to this single-seed adaptation and does not characterize ILCM's general capabilities. The effects of representation compression, optimization convergence, and physical-task adaptation cannot yet be separated from these results.

B Algorithm, Query Implementation, and Mechanism Approximation

B.1 Complete procedure and execution conditions

Algorithm 1 fits the visual model and calibration, freezes that interface for graph acquisition, then fits mechanisms. The main procedure learns T by Eq. (4); omitting numerical calibration defines a separate control. All inference parameters are frozen, and budgets count returned environment records.

Algorithm 1: Visual-representation calibration and active graph discovery

Input: Training images and concept labels $\mathcal{D}_{\text{obs}}$; intervenable targets $\mathcal{J}$ and a frozen ID command distribution; environment-call budget $B = B_{\text{cal}} + B_{\text{graph}}$; training, calibration, independence-test, and orientation configurations.

Output: The upstream procedure returns frozen representation h, estimated graph $\widehat{G}$ with unoriented-edge markers, and orientation evidence. When execution conditions are met, downstream processing fits shared local mechanisms and provides counterfactual decoding and outcome-computation interfaces based on $\hat{f}$.

1. Learn E_θ, D_ψ from training images and concept labels, then freeze them. Do not access test states, the true graph, or test results.
2. Call the environment in a prespecified round-robin order to collect B_{cal} records (j, u, x^{int}). A batch of n records counts as n calls, including repeated targets. Optimization steps that replay accumulated records do not add environment calls.
3. Fit T_ϕ using Eq. (4). Freeze T after the prespecified training rounds; compute μ^c, σ^c only from training observations and construct the shared interface h in Eq. (5). The uncalibrated control retains encoder readouts unchanged and is reported separately.
4. Run PC-style adjacency search on $h(\mathcal{D}_{\text{obs}})$, recording separating sets, colliders, and Meek propagation to obtain the initial partial graph $\widehat{G}_0$. If the structure is inconsistent, return a failure record rather than replacing it with the true graph.
5. Initialize the graph-discovery call count $b = 0$ and an empty intervention-record set. Retain the observational orientations and initial unoriented adjacencies of $\widehat{G}_0$.
6. While unoriented edges remain and $b < B_{\text{graph}}$:
   (a) Select a target covering current unoriented edges using Eq. (13). The RGB sequential-update cohort breaks ties by smallest node index and caps each target at six blocks (Appendix B.3). Do not access the true graph.
   (b) Sample commands from the frozen ID distribution, execute the environmental operations, and obtain settled-state images. The last batch must not exceed the remaining budget. After obtaining n records, set $b \leftarrow b + n$.
   (c) Pass the new images through the same frozen h. Append the encoded responses and single-target intervention masks, which identify the directly operated node in each record, to the accumulated evidence. Do not update the encoder, calibration, or normalization statistics online.
   (d) Reconstruct the current partial graph from $\widehat{G}_0$ and accumulated evidence. Test only the initially unoriented candidate edges in the CPDAG, preserving observationally compelled directions. Propose orientations using the response decision table in this appendix.
   (e) Tentatively add each candidate orientation to a graph copy. Reject it if it creates an opposite-direction conflict, a cycle, or an inconsistent DAG extension. If the checks pass, apply Meek orientation propagation until no further direction follows, then commit that update. Record reasons and sources for acceptance, rejection, and propagation.
7. Stop when no unoriented edges remain, target rules permit no further query, or the budget is exhausted. Return the final partial graph and all unresolved directions; do not complete edges by node index.
8. Fit the local mechanisms described in the main text only if the graph is valid, has zero unoriented edges, and is acyclic. Otherwise return a downstream-blocked status. Do not treat an executable DAG completion as a graph identified by the main method.

Full-DAG gate and blocking unit. Here "sufficiently oriented" strictly means zero unoriented edges in the whole graph, not just in a query-relevant subgraph. Appendix D.3 checks validity

and zero unoriented edges for each seed–configuration, then mechanism construction rejects cycles through topological sorting. A failed configuration has no predictor: all its planned CF and population queries are blocked, with no target-specific salvage from the partial graph. This is a seed–configuration gate, not a query-by-query answerability test; a seed execution fraction cannot substitute for query-level answerability. The prespecified consistent-extension control in Appendix C.3 is separate.

Using the trained model. Factual-image queries preserve estimated residuals and decode through Eq. (7); command-only outcomes resample training roots/residuals through Eq. (18). Both use the same fitted $\hat{f}$, without true test states (Appendix B.4). Static graph evaluation is not temporal evolution; temporal prediction would require a trajectory interface, cross-time graph, and transition mechanisms.

Implementation scope and failure reporting. Marginal response correlations on randomized single-target records are unavailable for $n \leq 30$. The rule assumes valid interventions and reliable candidate adjacencies; weak one-sided evidence can reverse an edge and cannot generally distinguish direct from ancestral effects. Fixed-order sequential closure is reproducible, not order-invariant. Thresholds, skeleton errors, visual mixing, and finite samples may misorient edges; Meek propagation cannot correct statistically wrong premises.

Direct-response criterion. Let $c_i = |\widehat{\mathrm{Corr}}(h_i, h_j; \mathcal{I}_i)|$, with c_j defined symmetrically. Each side requires at least 31 single-target records with binary intervention masks to be available. With $\tau = 0.15, \delta = 0.05$, the decisions are as follows; a proposed direction is accepted only after graph-consistency checks.

Evidence	Condition	Proposal
Both sides available	$c_i > \tau,\ c_j < \tau,\ c_i - c_j \geq \delta$	$i \to j$
Both sides available	Swap i, j in the preceding row	$j \to i$
Both sides available	Otherwise (including both strong, both weak, or an insufficient gap)	Unoriented
Only side i	$\lvert c_i - \tau \rvert < \delta$	Unoriented
Only side i	Outside the preceding interval and $c_i > \tau$	$i \to j$
Only side i	Outside the preceding interval and $c_i \leq \tau$	$j \to i$
Only side j	Symmetric treatment	Symmetric
Neither side	No available response score	Unoriented

The implementation records non-numeric correlations caused by constant variables as 0 rather than automatically abstaining. When only that side is available, this can trigger the weak-evidence reverse-direction rule. Graph-consistency checks therefore do not guarantee statistically correct orientations.

Calibration parameterization and inversion. The seven-node nonlinear calibration fixes $L = 5$ and $\kappa = (-2, -1, 0, 1, 2)$, with training parameterization $g_j = 1 + 0.25\tanh(\xi_j)$ and $w_{jl} = 0.1\tanh(\omega_{jl})$; biases b_j are unrestricted. Weights may be positive or negative. These ranges give a derivative lower bound of at least 0.25 and a total of 49 learnable scalars. The query loader checks $g_j > \sum_l |w_{jl}|$. If all weights are zero, inversion uses the analytic expression $(y - b_j)/g_j$. Otherwise, with $W_j = \sum_l |w_{jl}|$, elementwise bisection runs for exactly 60 iterations within $[(y - b_j - W_j)/g_j - 10^{-6}, (y - b_j + W_j)/g_j + 10^{-6}]$. The round-trip error threshold on training observations is 10^{-8}, not a global numerical-accuracy guarantee for every extrapolated input. The fully fitted affine control instead uses weighted least squares with a positive slope lower bound of 10^{-6} and no upper bound, a different parameter range from the bounded affine control.

Budgets and logs. The budget unit is the number of independent records returned by environment calls, not optimization steps or acquisition rounds. The number and training cost of observational samples used to train the encoder are reported separately; calibration and graph discovery together must not exceed B. Implementation logs save at least sample seeds, commands and realized values, stage, cumulative call count, graph states before and after updates, effect scores, rejection reasons, and hashes of frozen parameters, code, and data.

B.2 Confidence orientation backend and empty-graph query operator

The frozen criterion, held-out run records, and query implementation specify two particular controls: eight-seed Domino-MEC confidence orientation and the three-scene empty-graph control in Appendix C.4. These rules do not automatically apply to other aggregate cells with the same method names.

Response statistic and resampling. For an unoriented adjacency $i - j$ in the initial observational CPDAG, retain only records whose mask has exactly one target. Within interventions on i and j, respectively, compute the dimensionless absolute Pearson correlations $\widehat{d}_i = |\widehat{\mathrm{Corr}}(z_i, z_j \mid \mathrm{do}(i))|$ and $\widehat{d}_j = |\widehat{\mathrm{Corr}}(z_i, z_j \mid \mathrm{do}(j))|$. A side is available only with at least 200 records. Resample whole rows within that side with replacement 2000 times, re-centering and recomputing the correlation each time to obtain $D_i^{(b)}$ and $D_j^{(b)}$. Intervals $[L_i, U_i]$ and $[L_j, U_j]$ use the 0.005 and 0.995 quantiles with linear interpolation. Independent deterministic streams serve the two sides. When both exist, $\Delta^{(b)} = D_i^{(b)} - D_j^{(b)}$ gives $[L_\Delta, U_\Delta]$ by the same quantiles; this is not a difference between paired individuals across interventions.

Nonfinite encodings are rejected; fewer than 200 records make a side unavailable. Bootstrap correlations with denominators at most 10^{-12}, point correlations with denominators below 10^{-12}, and nonfinite correlations are set to zero. Constant inputs cannot alone trigger strong-signal orientation, although the weak side of a two-sided decision may be constant. These gates do not establish a true causal direction.

Calibrated noise ceilings and frozen selection. For each calibration edge with both sides available, select the side with the smaller point correlation (side i on a tie) and add its upper confidence endpoint to the multiset $\mathcal{V}_k$ for that intervention target. Let $\mathcal{V} = \bigcup_k \mathcal{V}_k$. The target-specific noise ceiling is

$$b_k = \min\left\{1,\ 0.03 + \begin{cases} \max\{\max \mathcal{V}_k, Q_{0.95}(\mathcal{V})\}, & |\mathcal{V}_k| \geq 4, \\ \max \mathcal{V}, & |\mathcal{V}_k| < 4. \end{cases}\right\}. \tag{8}$$

Here Q_p denotes the empirical p-quantile with linear interpolation. Thus 0.03 is a margin added to a data-derived ceiling, not a zero-effect threshold applied directly to correlation. Under the paper's one-based node indexing, the frozen values are

$$\begin{aligned} b_1 = b_6 &= 0.32561556156709537, \\ b_5 &= 0.32477945538327324, \\ b_2 = b_3 = b_4 = b_7 &= 0.3211601465259132. \end{aligned} \tag{9}$$

Ceiling fitting itself does not read graph truth; an empty $\mathcal{V}$ raises a calibration error. Criterion selection uses leave-one-seed-out predictions on seeds 10–17, seals predictions, then uses the *calibration-set true graphs* to reject candidates introducing wrong arrows or invalid graphs. Remaining candidates are ranked by maximizing complete graphs, maximizing correct edges, minimizing unresolved edges, then fixed priority; ceilings are fitted to all calibration records. The four prespecified (confidence, added margin, signal excess, directional contrast) tuples are $(0.99, 0.03, 0.10, 0.15)$, $(0.95, 0.03, 0.10, 0.15)$, $(0.95, 0.02, 0.05, 0.10)$, and $(0.90, 0.02, 0.05, 0.10)$; the first was frozen. If no candidate passes, a safe mode abstains everywhere. Inference on held-out seeds 20–27 does not read true graphs. This statistical-criterion calibration is distinct from anchor calibration of the numerical map T.

Direction decision and commitment. With only side i available, propose $i \to j$ only if $L_i > b_i + 0.10$; side j is symmetric. Weak one-sided evidence never proposes the reverse direction. With both sides available, use

$$\begin{aligned} i \to j: &\quad \widehat{d}_i > \widehat{d}_j, \quad L_i > b_i + 0.10, \quad U_j \leq b_j, \quad L_\Delta > 0.15; \\ j \to i: &\quad \widehat{d}_j > \widehat{d}_i, \quad L_j > b_j + 0.10, \quad U_i \leq b_i, \quad U_\Delta < -0.15. \end{aligned} \tag{10}$$

With neither side available, equal point estimates, or any failed condition, leave the edge unoriented; an invalid initial graph yields no proposal. Each round recomputes from the frozen initial CPDAG and cumulative records, with *sequential commitment* along its unoriented-edge list. A proposal must

preserve acyclicity and then pass Meek closure with a consistent-extension check before the edge and propagated arrows are committed. Failed updates are discarded, and edges already oriented by closure are skipped. Observationally compelled arrows cannot be overwritten. The 0.99 level is nominal for each interval, not a probability that the whole graph is correct.

The fixed-threshold replacement shares held-out seeds, frozen encoders, initial CPDAGs, the target-by-block-index data cache, and acquisition-budget settings, but uses the original $\tau = 0.15, \delta = 0.05$ rule. Backend-dependent graphs can change subsequent targets, so a shared cache does not imply identical realized action sequences. The 8/8 versus 4/8 result compares final complete recovery, not matched action paths.

This held-out cohort uses 1800 records per block, at most six blocks, and a six-block per-target cap. Acquisition ceases once the graph is valid with no unoriented edges; incomplete runs retain their final states. This runner evaluates graphs only and constructs no downstream query predictor, so 8/8 complete graph recovery is not a query-execution denominator.

Empty-graph population operator and factual branch. This paragraph applies to the MEC mechanism controls with learned T in Appendix C.4. Let w denote the mechanism coordinates produced by their frozen visual interface and $v_j(u)$ its training-unit conversion of the command. The empty-graph model is *rebuilt from the same encoded records and intervention masks* with zero adjacency, making every node a root. Each node stores its observational empirical marginal $\widehat{P}_k^{\text{obs}}$, falling back to records not intervening on that node only if the observational pool is empty. Original parent coefficients and residual pools are not retained, and no intercept regression is fitted for roots.

The command-only population query samples non-target nodes independently with replacement and clamps the target:

$$\widehat{P}_{\varnothing}(w \mid j, u) = \delta_{v_j(u)}(w_j) \prod_{k \neq j} \widehat{P}_k^{\text{obs}}(w_k). \tag{11}$$

Former non-roots therefore use observational marginals, not their original fitted residuals; whole observational rows are not resampled to preserve joint dependence. Both methods receive the same commands and GT means, each generate 2048 model samples per command, and average after conversion to protocol units. For each command, the same seed initializes a fresh random stream for each method, but graph-dependent topological traversal differs, so noises are not paired coordinatewise.

The factual-image branch does not sample these pools: it copies w^{fact}, sets only $w_j = v_j(u)$, retains all other coordinates, and uses the shared decoder. This empty-graph object has no descendants, so its factual operator differs from the population product of marginals. The module ablation that retains fitted mechanisms but disables descendant updates is interpreted separately: identical target-edit outputs would not establish identical population predictors or cohorts. These implementation records bind to the valid diagnostic seeds, 2/2 each for PnP/Gears and 1/2 for Balance. The other Balance seed is excluded for both methods because of unresolved structure; this does not supply operators or query denominators for other main-table aggregates.

B.3 Visual-learning, coverage-targeting, and local-mechanism formulas

This section specifies the visual-model objective, coverage-targeting rule, and linear mechanisms that support the calibration and query procedure in the main text.

Node-aligned variational autoencoder. Encoder E_θ estimates state-aligned latents z, and decoder D_ψ reconstructs images from them. Concept labels $\widetilde{\ell}_i$, standardized by observational training-label means and standard deviations, anchor node identities:

$$\mathcal{L}_{\text{enc}}(x_i, \ell_i) = \mathbb{E}_{z \sim q_\theta(\cdot \mid x_i)} \|D_\psi(z) - x_i\|_2^2 + \beta D_{\text{KL}}\Big(q_\theta(z \mid x_i) \,\|\, \mathcal{N}(\widetilde{\ell}_i, I)\Big). \tag{12}$$

Here q_θ is the encoding posterior, D_{KL} the Kullback–Leibler divergence, β the regularization weight, and I identity covariance; the two terms constrain reconstruction and label matching. After training, the encoder is frozen and its posterior mean $z = E_\theta(x)$ is used. The jointly trained base decoder is denoted by D; with the encoder fixed, it may be post-fitted using ID reconstruction alone to obtain D'. Calibration updates only T, each control shares the same D or D', and all are frozen during queries.

Target selection by unoriented-edge coverage. Let U_t be the set of unoriented node pairs at round t, and $\mathcal{A}_t \subseteq \mathcal{J}$ the intervenable targets with remaining acquisition allowance. The coverage heuristic selects a target adjacent to the most unoriented edges:

$$j_t \in \arg\max_{j \in \mathcal{A}_t} \sum_{\{p,q\} \in U_t} \mathbf{1}[j \in \{p, q\}]. \tag{13}$$

Here t is the acquisition round, $\mathbf{1}[\cdot]$ is the indicator, and arg max returns maximal-coverage targets. The RGB sequential-update cohort (Appendix D.3) breaks ties by smallest node index. Its $\mathcal{A}_t$ contains all targets with fewer than six acquired blocks; each block has 1800 records, at most six blocks are acquired, and an empty set stops acquisition. Repeated targets are allowed with no additional smaller cap; a six-block per-target cap cannot bind before the six-block total horizon. This cohort uses no incident-node fallback and does not relax the acquisition cap. Values come from a predetermined ID command distribution without true-graph or test access, and new images pass through frozen h. Coverage is a heuristic, not a proven budget-optimal policy.

Linear local mechanisms and ridge regression. Local mechanisms are fitted in the same h coordinates as $q_j(u)$ and executed only for valid, acyclic estimated graphs with zero unoriented edges (Appendix B.1). The linear parameterization with an intercept is

$$h_j = \widehat{f}_j(h_{\mathrm{Pa}_{\widehat{G}}(j)}) + \eta_j, \qquad \widehat{f}_j(v) = c_j + a_j^\top v. \tag{14}$$

Here v contains parent readouts, c_j is an intercept, a_j a coefficient vector, $\top$ denotes transpose, and η_j is an additive residual. Ridge regression estimates c_j, a_j, excluding records that directly intervene on j because its mechanism has been replaced. Root observations and fitted residuals form sample pools. Linear mechanisms are predictive approximations in calibrated coordinates, not consequences of invertibility. Nonlinear T can change local relationships and residual–parent dependence, so readouts, propagated outcomes, and counterfactuals are evaluated separately (Appendix B.6).

B.4 Implementation of distribution prediction and individual counterfactuals

The two entry points share a frozen encoder E_θ, calibration T, coordinate statistics, valid DAG $\widehat{G}$, local mechanisms $\widehat{f}_j$, and decoder D_ψ. Conditioning and residual handling depend on *whether an observed factual context is fixed*; no module is retrained. We describe linear mechanisms and the conditional-mean branch for physical simulation below.

A shared mapping specifies a computational interface; it does not automatically ensure equally accurate readouts across direct interventions, interventions on other nodes, and natural observations. Cross-source validation appears in Appendix C.2, and independent image-readout validation in Appendix D.3. Nonlinear calibration also need not preserve the linear additive mechanism family, so reductions in readout MSE and CF error must be tested separately.

Execution conditions and scoring sets. The strict entry point requires a valid, acyclic whole graph with no unoriented edges. Otherwise all factual and population queries for that seed/configuration are blocked; identifiable-looking targets are not salvaged separately. Downstream errors describe executed queries only. Appendices D.2 and D.3 report their own executable/planned seed counts; these do not establish main-table query denominators or common scoring sets (Appendix A.2). A seed-level denominator cannot substitute for a query-level answerability rate. Paired ablations use common valid seeds. The consistent extension in Appendix C.3 is restricted to static graph-backend comparisons and does not remove this blocking condition.

B.4.1 Given a factual image: individual counterfactual decoding

Inputs and preprocessing. Given a factual RGB image x, target j, and verified intervention value u, compute $z^{\mathrm{fact}} = E_\theta(x)$ and $h^{\mathrm{fact}} = h(x)$. Training-anchor standardization α_j and graph-coordinate conversion map u to $q_j(u)$. All sources use the same frozen mapping; means and variances are not re-estimated on the test set.

Factual residuals. For each non-root node, compute

$$\eta_k^{\mathrm{fact}} = h_k^{\mathrm{fact}} - \widehat{f}_k(h^{\mathrm{fact}}_{\mathrm{Pa}_{\widehat{G}}(k)}). \tag{15}$$

These are fitted-mechanism residual approximations, not known true exogenous disturbances. Non-target roots retain factual readouts without new noise. A blocked graph is not completed with the test-time true graph.

Executing the intervention and preserving unrelated variables. Let $\mathcal{D}_j = \mathrm{De}_{\widehat{G}}(j)$ denote the target's descendants. Copy the factual vector, set the target value, and update descendants in topological order:

$$h_k^{\mathrm{cf}} = \begin{cases} q_j(u), & k = j, \\ \widehat{f}_k(h^{\mathrm{cf}}_{\mathrm{Pa}_{\widehat{G}}(k)}) + \eta_k^{\mathrm{fact}}, & k \in \mathcal{D}_j, \\ h_k^{\mathrm{fact}}, & k \notin \mathcal{D}_j \cup \{j\}. \end{cases} \tag{16}$$

The operation cuts the target's original mechanism; it does not set all affected variables to the same u. Non-target non-descendants retain factual values according to the model graph, so omitted or misoriented edges can also make this preservation rule wrong.

Returning to decoder coordinates. First undo graph-interface standardization, then invert calibration:

$$z^{\mathrm{cf}} = T^{-1}(\sigma^c \odot h^{\mathrm{cf}} + \mu^c), \qquad \widehat{x}^{\mathrm{cf}} = D_\psi(z^{\mathrm{cf}}). \tag{17}$$

The decoder receives its training coordinates z, not unconverted h or physical states. Invertibility ensures coordinate compatibility, not accurate decoding of extrapolated points. The model accesses neither GT counterfactual images nor true outcome states, and no paired-outcome prediction head replaces this entry point.

Rigid-body conditional-mean branch. The rigid-body conditional-mean propagation control omits numerical calibration and recomputes descendants without the factual residuals in Eq. (16). Both propagation variants share encoding, graph, and decoding interfaces; disaggregated results appear in Appendix Table A3. Binary events may omit hidden initial triggering information, so conditional-mean propagation can incorrectly change descendants that fell independently in the factual scene.

B.4.2 Given only a command: intervention-distribution prediction

Training mechanisms and sample pools. Fit ridge regressions with intercepts on the parent sets specified by the graph, excluding records directly intervening on node k when fitting that node. The ridge coefficient for the linear mechanisms is 10^{-3}, with no penalty on the intercept. Root nodes retain observational encoded samples; non-root nodes retain fitted residuals. Simulation samples with replacement from each node's corresponding empirical pool. This is a model sampling approximation, not evidence that visual residuals equal independent true physical noise.

Generating samples. The input is only an operation $a = (j, u)$; the estimand is the mean non-target state of its intervention distribution. With no factual x, η^{fact} is not reused. Let M denote the prediction sample count, with $h^{(m)}$ generated for each $m = 1, \ldots, M$. Let $\widetilde{r}_k$ be a readout from root node k's training observational pool and $\widetilde{\eta}_k$ a residual sampled with replacement from its training residual pool. Evaluate nodes in topological order:

$$\widetilde{h}_k = \begin{cases} q_j(u), & k = j, \\ \widetilde{r}_k, & k \neq j,\ \mathrm{Pa}_{\widehat{G}}(k) = \varnothing, \\ \widehat{f}_k(\widetilde{h}_{\mathrm{Pa}_{\widehat{G}}(k)}) + \widetilde{\eta}_k, & k \neq j,\ \mathrm{Pa}_{\widehat{G}}(k) \neq \varnothing. \end{cases} \tag{18}$$

Here $\widetilde{h}_k$ is the predicted node value; repeated sampling yields the settled-state distribution. Sample counts, random seeds, operation ranges, and model configurations are protocol-fixed, not chosen from test performance. Generalization is evaluated under command-distribution changes excluded from model selection only after freezing the graph, representation, and mechanisms. This order is a variable-evaluation order, not physical time. Metric definitions and scene-specific test-distribution qualifications appear in Appendices A.6 and A.8. The empty-graph sampling operator is specified separately in Appendix B.2.

Order of averaging and decoding. State predictions are restored to label units using $\widehat{s}^{(m)} = \mu^s + \sigma^s \odot (\sigma^c \odot h^{(m)} + \mu^c)$, equivalently $\widehat{s}_j = \mu^s_j + \sigma^s_j T_j(z_j)$ componentwise. The sample mean for each non-target node is $\widehat{\mu}_k(j,u) = M^{-1}\sum_{m=1}^{M} \widehat{s}_k^{(m)}$, $k \neq j$; scoring units are fixed by the task protocol. Image generation requires inverse calibration back to the original z coordinates, followed by per-sample decoding and averaging:

$$\overline{x}_{\text{pred}} = \frac{1}{M}\sum_{m=1}^{M} D_\psi\Big(T^{-1}(\sigma^c \odot h^{(m)} + \mu^c)\Big). \tag{19}$$

Because the decoder and calibration are generally nonlinear, Eq. (19) cannot be replaced by averaging latent vectors before decoding. Means in original encoder coordinates are likewise computed after inverse transformation. GT distribution means come from independent sampling of the permitted environment under the same intervention command; model and GT samples need not be paired one-to-one.

Interpretation of the evaluation. The original-encoder-coordinate and mean-image protocols are scored separately from state NMSE. Comparing predicted and GT mean images after averaging generated samples measures population-average appearance, not whether each individual outcome is correct. The original-encoder-coordinate metric compares the means read from two image sets by the same frozen encoder; it is not an independent physical-state error. SCM experiments use ID commands of ± 1.15 and a prespecified amplitude shift to ± 1.75. Because the Gaussian training-command distribution has full support, this is an amplitude-distribution shift, not strict out-of-support extrapolation. This entry point computes what-if outcomes for new commands, but counterfactual examples or reconstruction quality cannot replace an evaluation of OOD accuracy.

B.5 Query interface for operation values and visual readouts

Figure 3 in the main text shows the complete learning procedure. Figure A2 adds the factual-conditioned counterfactual query in the frozen model, showing how visual readouts and verified intervention values enter the same computational coordinates and how predictions return to the decoder's training coordinates.

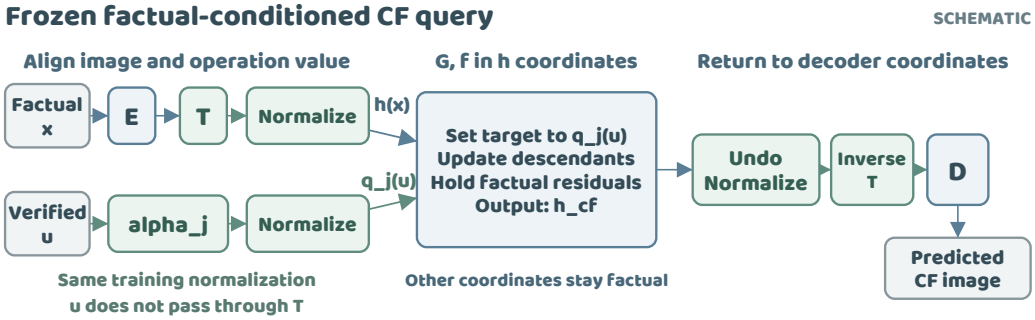


Figure A2: Coordinate transformations for a factual-conditioned counterfactual query (schematic). An image passes through E, T to produce calibrated readouts, while the verified value passes through α_j. Both paths use the same observational normalization to yield $h(x)$ and $q_j(u)$, respectively; the latter does not pass through T again. After setting the target, retaining factual residuals, and updating descendants, normalization and T are inverted before decoding with D. E, D, G, f abbreviate the encoder, decoder, estimated graph, and local mechanisms, respectively. All mappings and statistics are frozen. Population prediction without a factual input is described in Section 4.3.

B.6 Calibrated coordinates and mechanism approximation

Invertible calibration allows queries to return to the decoder's original coordinates; it does not ensure that a specified regression family is preserved by the transformation. To distinguish these properties,

temporarily omit affine standardization and suppose the original readout relation is $z_j = f_j(z_{\mathrm{Pa}(j)}) + e_j$. The exact transformed relation is

$$\widetilde{z}_j = T_j\left[f_j\left(T^{-1}_{\mathrm{Pa}(j)}(\widetilde{z}_{\mathrm{Pa}(j)})\right) + e_j\right].$$

Here $\widetilde{z} = T(z)$, and $T^{-1}_{\mathrm{Pa}(j)}$ acts coordinatewise on the parents. If T_j is nonlinear, the increment $T_j(m+e_j) - T_j(m)$ produced by the same e_j generally depends on $m = f_j(z_{\mathrm{Pa}(j)})$. A linear additive model refitted after transformation is therefore an approximation. Invertibility does not imply that residuals remain independent of parent states or that individual noise has been identified.

For positive-affine $T_j(z_j) = g_j z_j + b_j$, exact moment normalization cancels the gain and bias in image readouts h_j (ignoring numerical stabilization). The command setting $q_j(u)$ still depends on them through μ^c_j, σ^c_j in Eq. (5); unchanged image coordinates do not imply unchanged queries. Nonlinear calibration can bring distorted visual relations closer to the chosen mechanism family, but this is not guaranteed. Readout, mechanism, and end-to-end query accuracy must be evaluated separately.

B.7 Orientation and mechanism diagnostics on frozen records

This analysis uses existing encoders, calibrations, and active-acquisition records for seeds 5, 6, and 7, without training new models. Identity, fully fitted affine, and nonlinear calibration share graph-acquisition records. Prefix-by-prefix replay fixes thresholds and margins and compares ascending, descending, and 30 prespecified random edge orders. Permutations are algorithm-diagnostic conditions, not independent training seeds. We also compare a bilateral-evidence safeguard requiring more than 30 records in both directions and nonzero readout variance. This safeguard is a separate control and does not change the main-table backend.

Mechanism diagnostics use 64 held-out ID factual readouts per seed. For each non-root node, we report frozen-local-mechanism residuals and their relation to parent readouts; bin boundaries for parent readouts are determined only from training observations. Residuals are not errors against ground-truth predictions. Their dependence may arise from visual error, mechanism misspecification, or finite samples and cannot be uniquely attributed to calibration. A pure coordinate diagnostic fixes residual quantiles in original coordinates and computes the standard deviation of the transformed increment as parent states vary; identity and affine transformations provide numerical-zero checks.

Table A4: Frozen diagnostics. Graph columns use the final acquisition prefix; all three coordinate systems give the same results, and each cell is unchanged across 32 traversal orders. SHD counts each unoriented true edge as one unresolved relation and therefore differs from the number of wrong arrows. The last column is the SD of transformed increments for fixed original-coordinate residuals under nonlinear calibration, averaged equally over non-root nodes and three training-residual quantiles, in training-label-standardized coordinates. It is a coordinate diagnostic, not prediction error.

Seed	Original-rule SHD	Bilateral SHD	Bilateral unoriented edges	Nonlinear increment SD
5	0	6	6	0.0343
6	0	0	0	0.0311
7	0	0	0	0.0306

Default replay exactly matches the archived graphs. The final prefix contains 1800 records for seed 5 and 3600 each for seeds 6 and 7, so the table is not an equal-budget comparison across seeds. Changing traversal order does not change final graphs on these records; this does not establish order independence for arbitrary graphs or data. Requiring valid bilateral evidence leaves 6 unoriented edges for seed 5, while the other two seeds remain query-executable. The safeguard rejects constant and weak one-sided evidence but reduces orientation coverage under a finite budget, so it is evaluated separately from the main-table algorithm. Degenerate-input tests also show that the original rule can interpret 31 constant one-sided records as a reverse-direction proposal, demonstrating that numerical validity and graph consistency are different checks.

Fixed-residual increment SDs for identity and affine transformations are both zero within tolerance 10^{-10}; nonlinear results appear in Table A4. Each seed averages 18 SD values from 6 non-root nodes and 3 training-residual quantiles. Each SD is computed across 64 factual contexts, not across

seeds. This gives a concrete example of a changed additive interpretation after transformation, but does not establish that this change is the main cause of downstream error. Local residuals in different coordinates cannot replace outcome evaluation against shared environmental ground truth.

C Environmental Interventions and Graph Orientation

C.1 Multi-environment structure protocols and ablations

The environments share a procedure but differ in state generation, data volumes, and orientation criteria.

Environments and statistical objects. Domino-MEC is a seven-node chain, PnP-MEC a six-variable pick-and-place system, Balance-MEC a five-variable balancing system, and Gears-MEC a six-variable transmission system. Explicit SCMs generate responses; renderers provide settled-state frames after actuators return to the fixed HOME pose. Statistical tests exclude action-transition frames. The arm illustrates operation semantics, not contact-dynamic generation of every SCM outcome. Physical contact instead determines the four-event rigid-body domino environment, scored separately under Appendix A.9.

Training labels establish node-aligned encodings; graph discovery uses frozen encoded observations, intervention images, and known targets, not the true skeleton. Held-out inference uses the true graph only for offline scoring; criterion selection on calibration seeds uses graph references as specified below. Table 1 reports learned-CPDAG mode, which learns the initial partial graph, excluding true-CPDAG/DAG diagnostics. Structure and counterfactual experiments have separately fixed seed lists; Domino structural holdout and mechanism queries also have separate seeds and protocols.

Domino-MEC: independent holdout and orientation criterion. Graph recovery uses held-out seeds 20–27, separate from development seeds 0–9 and criterion-selection/calibration seeds 10–17; these are not numerical-T calibration seeds. Noise ceilings are fitted without graph truth, but criterion selection scores held-out-fold calibration predictions against calibration true graphs. The frozen criterion uses confidence 0.99, 2000 bootstrap resamples, at least 200 records, additive null margin 0.03, signal excess 0.10, and directional contrast 0.15. The margin augments target-specific noise bounds rather than defining a zero-effect interval; one-/two-sided tests and selection rules are given in Appendix B.2. Inputs are 6000 observational images and 1800 per intervention block, with 350 encoder epochs and at most six target-selection rounds. All 8 held-out coverage runs complete within two rounds. Rounds are sampling blocks; calls count returned records.

With the same held-out seeds, frozen encoder, initial observational CPDAG, and target/block-index cache, final complete recovery is 8/8 for unoriented-edge coverage (coverage), random targeting (random), and fixed cycling (round-robin), versus 0/8 for observation-only (passive), which retains 6 unoriented edges in every run. Shared cache entries contain identical pixels, but changed graph states can change later targets and realized trajectories. Each round recomputes from the initial CPDAG using cumulative evidence, accepting directions sequentially with Meek closure. This final-state ablation does not establish equal-budget policy superiority. Replacing coverage's confidence criterion with fixed thresholds $\tau = 0.15, \delta = 0.05$ gives 4/8 complete recovery, retaining errors and unresolved states.

PnP-MEC, Balance-MEC, and Gears-MEC. PnP uses independent held-out seeds 200–207; both initial CPDAG and final DAG recovery are 8/8, with zero final errors and unoriented edges. Balance uses 8 development-configuration seeds. After diagnosing extra adjacencies, the skeleton-effect threshold was adjusted from 0.04 to 0.06 while reusing the frozen encoder. Final recovery is correct in 7/8 runs; the remaining run retains an extra adjacency. Balance results therefore analyze development samples rather than provide evidence of independent held-out generalization. Gears uses seeds 0–7, 4200 training images, 5000 observational images, 900 intervention images per block, and 180 encoder-training epochs. Its observational-skeleton threshold is 0.06, with orientation threshold 0.20 and margin 0.08. All 8 learned-graph runs resolve 3 direction ambiguities in one round with zero wrong arrows; the passive controls each retain 3 unoriented edges at termination.

Relation to the shared interface. The shared sequence is visual encoding, observational partial graph, target selection, intervention evidence, and graph update; criteria are frozen per environment. Results concern this structure-learning procedure, not an additional non-identity calibrator. Appendix C.3 permits protocol-fixed consistent extensions for downstream computation, without counting them as identification success; Appendix B.4's strict query entry point does not. Factor-outcome NMAE, pixel-counterfactual MSE, and distribution-prediction error remain separate metrics.

C.2 On-demand acquisition and frozen-calibration queries

This control uses a seven-node domino scene rendered as 64×64 RGB images and encoder seeds 5, 6, and 7 to evaluate evidence acquisition, readouts, and query outcomes separately. Encoder, decoder, calibration, training statistics, and scoring are frozen before test-data generation, without test-error model selection. Conclusions are limited to this simulation.

C.2.1 What do additional observations and interventions provide at equal record counts?

All groups share 6000 observations, 896 calibration records, and a frozen nonlinear mapping. Additional independent observations re-estimate PC at prefixes of 1800 and 3600. Intervention groups start from the same graph, use coverage or fixed round-robin targeting in batches of 1800 up to 3600, and stop at a valid, sufficiently oriented graph. Each reached intervention prefix is paired with equally many new observations. Both sides pay the same calibration cost, but calibration records are excluded from graph-stage evidence: the comparison isolates *the source of graph-stage evidence*, not a wholly intervention-free procedure.

An inconsistent candidate-direction batch is rejected in full under rules and thresholds fixed before acquisition. This control and Appendices F.1 and F.3 use the same batch-commitment implementation, unlike the sequential commitment in Appendix B.1. The true graph is used only to score fixed trajectories, never for selection, stopping, or graph estimation.

Correspondence among the three acquisition controls. Within each training seed, all three controls reuse the same encoder checkpoint, 6000 observations, and nonlinear T, with identical file hashes in their frozen manifests. Their initial graphs are also identical field by field: a seven-node chain with 6 unoriented edges. A shared training seed does not imply shared graph-stage samples. Batch b for target j uses sampling seed $N + 10007s + 101j + b$, where $s \in \{5, 6, 7\}$ and j, b are zero-based. The equal-record evidence control here uses $N = 142000000$; the three-policy control in Appendix F.1 uses $N = 2600000$; the equal-total-budget control in Appendix F.3 uses $N = 12600000$. All start with an empty graph-intervention pool and retain the same ID command distribution and orientation thresholds. The graph-record cap is 3600 here and 10800 for the nonlinear configurations in the other two controls.

Table A5: Graph-evidence sources under the same frozen visual and calibration interface. Each row compares equal numbers of new records; both sides include the same 896 calibration records and 6000 initial observations. Intervention and additional-observation graphs are scored at matching prefixes; records not acquired after early stopping are not treated as available evidence.

Acquisition policy	Seed	New records	Intervention SHD	Added-observation SHD
Coverage targeting	5	1800	0	6
Fixed round-robin	5	1800	0	6
Coverage targeting	6	3600	0	6
Fixed round-robin	6	1800	0	6
Coverage targeting	7	3600	0	6
Fixed round-robin	7	1800	0	6

Both intervention policies recover the true graph for all 3 seeds; matched additional observations leave 6 unoriented edges, not 6 wrong directions. Round-robin uses 1800 intervention records per run, whereas coverage uses 3600 for two seeds. This supports additional orientation information from interventions, not a coverage-budget advantage or universal inability of observations to orient edges.

Seed 6's coverage trajectory explains the cross-table difference. Here, targets 1 then 0 yield acceptance of $0 \rightarrow 1$ and complete orientation after 3600 graph records. The other two controls retain 0–1 after the same target prefix, then repeatedly sample target 0 until budget exhaustion. Their fixed target-1 samples yield reverse-response scores of 0.15265 and 0.15226, respectively, both above threshold 0.15, so the two-sided rule continues to abstain. The 3/3 here and 2/3 in those controls therefore refer to different specified graph-sampling streams, not different calibration checkpoints or commitment rules. The present 3/3 does not establish stable coverage-policy success across sampling repetitions.

C.2.2 Do calibrated readouts transfer to other response sources?

Per seed, 512 independent exogenous contexts generate natural observations and 7 direct-intervention types with original ID values. Three calibration configurations share frozen encodings; target-state GT comes from the SCM. Each node is evaluated under direct targeting, natural observation, and intervention on other nodes. The last source includes ancestors and non-ancestors; records stratify the ancestor subset rather than treating all responses as indirect causal effects.

Errors cover the full sampled range and common bins, fixed at intervals of 0.5 over $[-2, 2]$ in training-label-standardized coordinates. Retained bins contain at least 8 records per source; weighting is equal across bins, then nodes. All seeds have 7 scorable nodes, with ranges, counts, and exclusions saved. GT stratifies evaluation only; it neither updates mappings nor ensures identical within-bin distributions.

Table A6: Cross-source readout MSE, multiplied by 10^3, reported as mean $\pm$ sample SD over 3 seeds with equal weighting of 7 nodes per seed. Common bins are fixed in training-label units and require at least 8 records per source, node, and bin; common bins then receive equal weight. This controls coarse common support, not full distribution matching. Bold denotes the lowest point estimate in each panel, not statistical significance.

Calibration	Direct intervention	Natural observation	Other-node intervention
Full sampled range			
No calibration	23.636 ± 2.994	36.850 ± 7.400	39.063 ± 6.405
Fully fitted affine	15.224 ± 0.399	$\mathbf{33.258 \pm 3.868}$	$\mathbf{36.365 \pm 2.890}$
Nonlinear calibration	$\mathbf{12.998 \pm 0.233}$	35.144 ± 3.376	38.296 ± 2.306
Fixed common bins			
No calibration	22.953 ± 3.222	30.711 ± 7.441	29.878 ± 5.482
Fully fitted affine	14.321 ± 0.602	18.321 ± 2.813	18.961 ± 1.251
Nonlinear calibration	$\mathbf{12.306 \pm 0.403}$	$\mathbf{14.806 \pm 1.860}$	$\mathbf{16.260 \pm 0.688}$

In Table A6, nonlinear calibration has the lowest mean common-bin error for all three sources, but fully fitted affine calibration is better over the full range for observations and other-node interventions. Readout gains thus depend on state range. Common bins do not fully match distributions, and unequal source errors do not establish elimination of source differences.

C.2.3 From calibrated coordinates to true-state and image queries

Queries use the same-record comparison's frozen graphs and mechanisms for no calibration, fully fitted positive affine calibration, and nonlinear calibration, not the acquisition-policy graphs above. Seed 6 has unoriented edges in all three graphs and remains in the planned denominator but is blocked; seeds 5 and 7 have identical executable graphs.

Each valid seed uses 64 new factual contexts, 7 targets, ID commands at the training-label mean ± 0.75 standard deviations, and amplitude-shift commands at ± 2, giving 896 CF queries per class. GT replays the same exogenous context. Models receive only the factual image, target, and value, retain factual residuals, and inverse-calibrate before decoding. Population queries have no fixed factual image: model and GT means use 4096 samples from empirical pools and the SCM, respectively. Models share test contexts and random seeds; amplitude tests do not adjust models.

Predictions return to environmental units via training-concept scales, are compared with SCM states, and are normalized by those scales. CF-state columns measure individual outcomes; population

columns measure each non-target node's interventional mean $\mu_k(j, u)$, $k \neq j$, from Section 3. Both exclude the set target, so command copying is not predictive performance. Descendant, non-descendant, and factual-relative change errors are saved separately. State metrics are not independent decoder-semantic readouts; image metrics are pixel MSE of raw floating-point decoder outputs.

Table A7: Query results with frozen graphs, mechanisms, and decoders; errors are multiplied by 10^3. State columns use true simulator states as GT and exclude the direct target; the image column reports MSE of each model's own decoder output. Values are means $\pm$ sample SD over common valid seeds 5 and 7 out of 3 planned seeds; seed 6 is blocked because its frozen graph is not sufficiently oriented. Bold denotes the lowest point estimate within each command group; amplitude shift is not strict out-of-support extrapolation.

Calibration	Command	CF-state NMSE	Population-mean NMSE	Image MSE
No calibration	ID	37.303 ± 8.593	3.366 ± 2.889	2.257 ± 0.025
Fully fitted affine	ID	$\mathbf{32.810 \pm 3.066}$	$\mathbf{1.758 \pm 0.025}$	$\mathbf{2.232 \pm 0.023}$
Nonlinear calibration	ID	34.875 ± 1.268	2.045 ± 0.176	2.237 ± 0.019
No calibration	Amplitude shift	38.396 ± 8.949	3.983 ± 3.330	1.699 ± 0.039
Fully fitted affine	Amplitude shift	$\mathbf{35.102 \pm 2.667}$	3.756 ± 0.321	$\mathbf{1.688 \pm 0.043}$
Nonlinear calibration	Amplitude shift	36.703 ± 0.869	$\mathbf{3.492 \pm 0.510}$	1.692 ± 0.048

On the 2 jointly executable seeds, nonlinear calibration reduces mean non-target CF-state NMSE versus no calibration by 6.51% for ID and 4.41% for amplitude shift; population-mean NMSE falls by 39.26% and 12.33%. These gains are unstable: both query errors rise on seed 5 and fall on seed 7. Against fully fitted affine calibration, only amplitude-shift population-mean error is lower; every other column is higher. Mean rankings do not establish significance.

For factual-relative *change*, nonlinear calibration instead raises non-target effect NMSE versus no calibration by 0.88% for ID and 11.77% for amplitude shift. Absolute state, change, population mean, and image errors measure distinct abilities. Together with blocked seed 6, these results verify executable frozen-interface queries and range-limited readout gains, not a stable end-to-end calibration advantage.

C.3 Static mechanism comparisons on four MEC tasks

Fixed static instances of Domino-MEC, PnP, Balance-MEC, and Gears-MEC use native encoder-training protocols and 8 seeds each; PnP uses held-out seeds 200–207. Within each environment, visual representations and mechanism families are fixed, varying only the graph interface. Greedy Interventional Equivalence Search (GIES) (Hauser & Bühlmann, 2012) and passive graph extension receive the same numerical representations as our method. The metric is factor-level counterfactual NMAE: normalized absolute error in variable outcomes, not pixels. These graph-backend/mechanism comparisons are independent of the main-text RGB and rigid-body image benchmarks and do not evaluate image MSE or dynamic prediction.

Table A8: Static factor-level counterfactual NMAE, lower is better. Means and 95% bootstrap intervals come from the corresponding per-seed records. Metrics with different normalization definitions are not ranked together.

Method	Domino	PnP	Balance	Gears
Our method	0.1745 [0.1650, 0.1855]	0.1345 [0.1251, 0.1452]	0.0886 [0.0771, 0.1041]	0.1021 [0.0995, 0.1045]
GIES	0.2716 [0.1750, 0.3939]	0.1319 [0.1235, 0.1413]	0.0894 [0.0781, 0.1049]	0.1018 [0.0998, 0.1037]
Passive graph extension	0.6975 [0.6841, 0.7094]	0.1345 [0.1252, 0.1448]	0.5966 [0.4478, 0.6782]	0.7199 [0.7024, 0.7374]
Shuffled graph	0.4824 [0.4183, 0.5600]	0.8287 [0.7532, 0.9065]	0.6069 [0.5283, 0.6883]	0.6478 [0.5312, 0.7558]
No graph propagation	0.6975 [0.6841, 0.7091]	0.9662 [0.9486, 0.9865]	0.9418 [0.9290, 0.9552]	0.9081 [0.8879, 0.9315]
True graph (structural reference)	0.1745 [0.1650, 0.1858]	0.1345 [0.1253, 0.1448]	0.0886 [0.0770, 0.1043]	0.1021 [0.0995, 0.1045]

Passive graph extension selects a compatible DAG from the observational graph by a prespecified rule; *shuffled graph* changes node correspondence, and *no graph propagation* removes inter-variable transmission. Protocol-fixed consistent extensions of partial graphs enable computation, not intervention-based identification. The supplied true graph is a structural reference with mechanisms fitted on the same visually encoded data. GIES has slightly lower error point estimates on PnP and Gears; our method is lower on Domino and Balance, subject to the reported intervals and task conditions.

Balance-MEC's skeleton-effect threshold was selected on development samples, unlike PnP's independent held-out seeds 200–207. Results retain these environment-specific selection and evaluation scopes.

C.4 What-if and pixel-counterfactual protocols for three scenes

Native PnP, Balance, and Gears MEC implementations each use seeds 300 and 301 with learned T; the encoder, calibration map, and normalization statistics are frozen for graph discovery and queries. Structure learning matches the main-text implementation but uses an independent seed list and denominator. PnP selects among 3 encoder initializations by training-anchor validation error; Balance uses a frozen skeleton threshold of 0.06 and 90 epochs, and Gears uses 180 epochs. Discovery receives pixel encodings, not the true graph. Queries require a fully directed acyclic graph, without offline-SHD seed selection. All six runs remain: PnP and Gears execute 2/2, Balance 1/2, with seed 301 blocked by orientation ambiguity. Reported errors condition on valid runs; blocked runs are not zero errors.

Mechanism prediction. Local ridge mechanisms use ID observations and original target-selected intervention records processed by the same frozen visual interface, excluding each node's directly intervened samples. Roots sample saved observational readouts and non-roots fitted residuals, with replacement. Parameters are fixed before prediction, without ID/OOD test-query selection. Each node has two ID and two OOD commands at the training mapping's physical origin ± 0.75 and ± 2 scale units. Both methods share the original data adapter's training-normalization scale; test outcomes determine neither commands nor denominators.

The no-graph population comparator constructs a new empty-adjacency model from the same encoded records and masks. Every node becomes a root, independently sampled with replacement from its observational marginal pool (falling back to rows not intervening on that node), before clamping the target. It reuses neither parent coefficients nor residual pools. The shared random seed does not pair coordinate-wise noises because traversal order differs; implementation details appear in Appendix B.2.

For target j and command u, let $\overline{s}^{\mathrm{GT}}$ and $\overline{s}^{\mathrm{pred}}$ be 2048-sample simulator and model means in this protocol's scoring units, with training-fixed scale d_k:

$$E(j,u) = \frac{1}{K-1} \sum_{k \neq j} \left(\frac{\overline{s}_k^{\mathrm{pred}} - \overline{s}_k^{\mathrm{GT}}}{d_k} \right)^2 . \tag{20}$$

Targets and positive/negative commands are equally weighted within each valid seed, then summarized by across-seed mean and sample SD; Balance reports its sole valid seed's point value. Excluding the set target avoids rewarding command copying, while retaining non-descendants measures preservation alongside propagation. Lower degradation factor $g = E_{\mathrm{OOD}}/E_{\mathrm{ID}}$ means less relative error growth, opposite to the stability ratio in Table 3: $r = 1/g$ only for identical error summaries. Different run sets prohibit reciprocating values across these protocols. Raw records also retain worst-query and distribution-standard-deviation errors. This tests intervention-amplitude shifts, not strict out-of-support generalization or full-distribution distance.

Image queries. With the same encoder, learned graph, and mechanisms, a decoder fits reconstruction on 2400 independent ID observations and frozen encodings for 200 epochs, batch size 256, and Adam learning rate 0.001, without test factual–counterfactual pairs. Both methods share this decoder. Our method sets the factual target, retains residuals, and propagates to descendants before decoding; the empty-graph comparator sets only the target and preserves all other factual coordinates, without pool sampling. GT replays the same exogenous noise. The 32 fixed contexts, K targets, and two

commands yield $64K$ queries. Generated images are decoder outputs; the simulator supplies only factual images and paired GT.

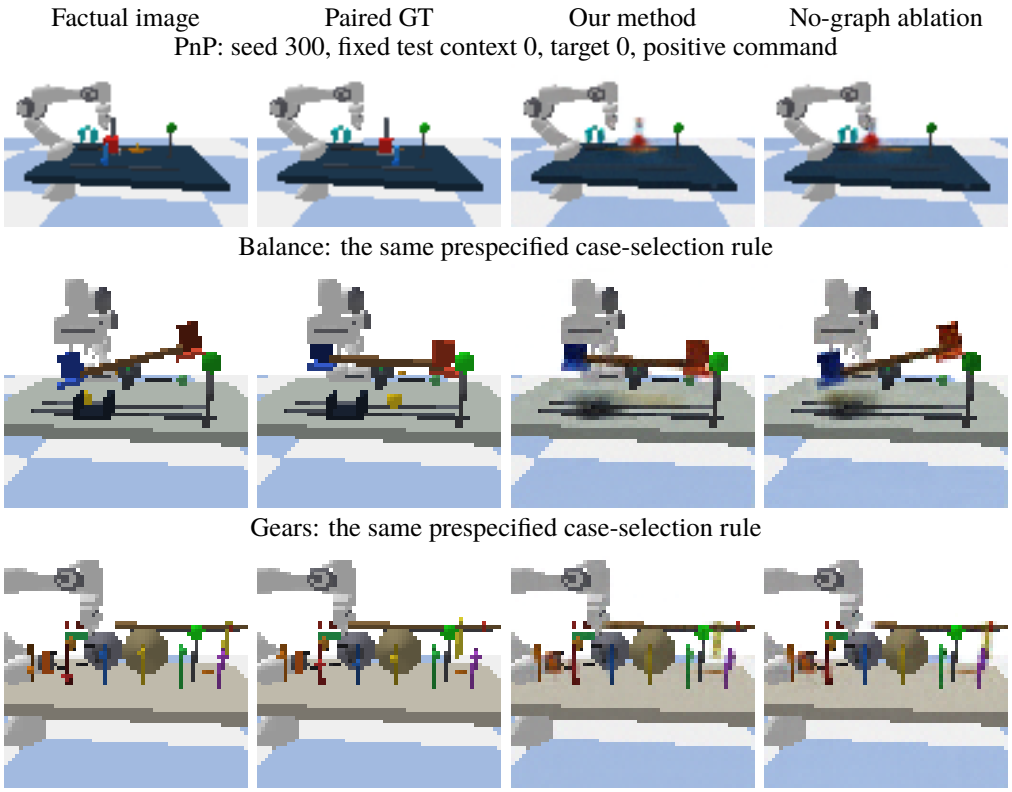


Figure A3: Fixed queries in three controlled MEC scenes. Cases are prespecified by test index, not filtered by prediction quality. PnP uses 64×96 images and Balance/Gears use 64×64; this group is not pooled with rigid-body domino or RGB128 image protocols.

These within-implementation graph comparisons are limited by scene scale, valid seeds, and blocking rates. The no-graph ablation is not an independently published model and does not establish cross-scene superiority over CausalVAE.

D Intervention-Value Calibration and Query Capability

D.1 Query metrics and interpretation of intervention-value calibration

Table 2 uses one three-seed cohort for readout, effect, and RGB comparisons, with the following metric and update permissions.

Counterfactual effect error. For one query, non-target factual-to-counterfactual change error is

$$\frac{1}{K-1}\sum_{k\neq j}\left[\frac{(\widehat{s}_k^{\mathrm{cf}}-\widehat{s}_k^{\mathrm{fact}})-(s_k^{\mathrm{cf}}-s_k^{\mathrm{fact}})}{\sigma_k^s}\right]^2.$$

This is neither absolute-state NMSE nor population-mean NMSE without a factual image. Table 2 reports readout, effect, and RGB errors; population means, graph discovery, and independent image readouts have separate evaluations.

Trainable modules and comparison scope. Identity, fully fitted affine, and nonlinear calibration share an ID-refitted decoder. Ordinary joint updating changes both readout heads and decoder, precluding attribution solely to invertibility, parameter count, or one loss term. Main-table means and

SD aggregate equally weighted targets and commands within seeds, then across seeds; improved-seed counts and paired differences appear in Appendix D.4.

Trade-offs in other controls. Appendix D.3 uses the original decoder and sequential orientation updates, whereas the main cohort uses an ID-refitted decoder. Each cohort fixes its own test contexts and shares a decoder among calibration variants; readout, effect, and RGB statistics are not pooled across cohorts. Appendix E.1 isolates decoder replacement.

D.2 Restricted affine and nonlinear calibration with a frozen decoder

Independent ID validation uses the seven-node domino scene with 64×64 RGB images. It reuses the frozen encoders/decoders for seeds 5, 6, and 7, 6000 observations, 896 target records, and shared coverage-targeting data, without new training or OOD selection. Each configuration has its own normalization, frozen graph, and mechanisms; graphs use the same evidence without resampling.

Restricted positive-affine calibration fixes curvature weights to 0; nonlinear calibration adds strictly monotone terms under the same learning schedule and data access. The uncalibrated control does not fit T or reallocate its saved calibration budget to graph discovery.

All nine mappings T, graphs, and mechanisms are frozen before validation. Each seed contributes 64 independent random ID intervention readouts for each of 7 targets, totaling 1344 records. Graph-valid seeds additionally use 64 independent factual contexts and two commands per target at the training-anchor mean ± 0.75 standard deviations. Shared factual images and commands are paired with GT replay under fixed exogenous context. Configurations retain factual residuals in their own frozen coordinates/mechanisms and apply inverse-T before the frozen decoder; predictors never receive GT states.

Seed 6 retains one unoriented edge in all configurations and is blocked, without arbitrary completion or zero errors. Readout/graph statistics use 3 seeds; CF uses the 2 jointly valid seeds, each with 896 paired queries. The 3264 evaluation calls comprise 1344 readout, 1792 CF GT, and 128 factual records. Another 32 consistency records are excluded; there are no training calls.

In Table A9, nonlinear calibration has the best readouts but unchanged graph recovery, smaller pixel gains, and slightly higher mean CF error than restricted affine. Invertible coordinate-wise maps add no population conditional-independence information; finite-sample and mechanism-fitting changes remain separate. Fully fitted affine and equal-total-budget controls appear in Appendix F.3.

Table A9: Restricted affine and nonlinear calibration with a frozen decoder. ID readout and CF errors are multiplied by 10^3, with mean $\pm$ sample SD across seeds. The graph column lists SHD of each configuration's frozen graphs for seeds 5/6/7. Seed 6 is blocked by the graph gate, so no CF error is entered.

Configuration	Target MSE (3/3)	SHD	CF pixel MSE (2/3)
Uncalibrated	23.090 ± 3.183	0/1/0	2.1894 ± 0.0183
Restricted positive affine	15.036 ± 0.773	0/1/0	$\mathbf{2.1643} \pm 0.0192$
Nonlinear calibration	$\mathbf{12.849} \pm 0.798$	0/1/0	2.1707 ± 0.0204

D.3 Matched-record validation of intervention-value calibration and standard supervised updating

This matched-record control uses the seven-node domino scene with 64×64 RGB images and encoder seeds 5, 6, and 7 to separate readout, numerical-query, and image changes. It jointly evaluates calibration, sequential graph updates, local mechanisms, and both query types; every graph is learned from this experiment's acquisition records.

D.3.1 Shared data and update permissions

Each seed reuses the frozen backbone, 6000 observations, and 896 verification records. Original seeds and commands replay training images, individually checked against saved encodings, without

adding independent data. Uncalibrated, fully fitted positive-affine, and nonlinear configurations share frozen encoder/decoder weights. Nonlinear calibration has 49 coefficients: gain, bias, and five basis weights for each of seven coordinates.

Ordinary joint updating is a same-supervision control, not a published method. With convolutional features frozen, seven linear heads are refitted on 256-dimensional features, using 1799 parameters. New labels come only from verified direct-target values; observation-preservation weight is 0.01. Four-fold cross-validation within the original 896 records selects ridge weight from $\{10^{-4}, 10^{-2}, 1\}$, followed by fitting all records. The shared decoder checkpoint refits only reconstruction of 6000 observation and 896 calibration images for 40 epochs, batch size 128, Adam learning rate 10^{-4}, using the final epoch. Its 1,154,659 parameters receive no test CF pairs. Shared data and labels but different trainable modules/parameter counts prevent separate attribution to heads or decoder.

Table A10 combines readout and query results under these update permissions. Its original decoder and test queries differ from the ID-refitted decoder and queries in Table 2; the cohorts are not pooled.

D.3.2 From newly acquired graphs to frozen queries

After all visual interfaces freeze, nonlinear calibration selects targets by the main-text coverage rule: 1800 new pixel observations per batch, at most six batches. Appendix B.1's sequential tentative additions, consistency checks, and Meek closure replace batch commitment. Other configurations receive exactly the acquired images and commands, learn graphs in their own coordinates, and fit local linear mechanisms. Acquisition records are matched; the four configurations do not optimize separate active policies.

The seeds use 1800, 3600, and 3600 graph records. Every configuration pays for 896 calibration records; observations are reported separately. Stopping is truth-blind and structure is scored after model freezing. All four configurations have SHD 0 and no unoriented edges in all three seeds: every query executes, without evidence of improved recovery from calibration. The batch-commitment 2/3 denominator belongs to a different update procedure and is not directly comparable.

Each seed uses 64 new factual contexts, 7 targets, and four commands per target: training-label mean ± 0.75 standard deviations for ID and ± 2 for amplitude shift. Each seed/split has $64 \times 7 \times 2 = 896$ paired CF queries: 1792 per configuration/seed across splits and 5376 across three seeds, independent of the 896 training verification records. GT replays fixed exogenous context; models receive only factual images and commands. Population evaluation compares non-target means from 4096 model and simulator outcomes each. Graphs, mechanisms, visual modules, and statistics stay frozen.

Nonlinear target-readout error falls by 46.26% versus no calibration and 15.21% versus fully fitted affine, improving all three seeds. Affine nevertheless has lower mean numerical-query error. Nonlinear population-mean error versus no calibration falls by 24.10% for ID with only one seed improving, and 16.64% for amplitude shift with two improving: aggregate reductions are not consistent gains. Against joint updating, its amplitude-shift population-mean error falls by 71.03% and improves all three seeds, but joint updating has lower image error. These evaluations are not interchangeable.

D.3.3 Independent evaluation of visible states in generated images

An independent image readout Q, with no shared weights, evaluates renderer-visible activations $\gamma(s_j) = 1/(1 + \exp(-3.8 s_j))$, not arbitrary-precision recovery of s_j from saturated pixels. Independently sampled node activations break causal-correlation shortcuts: an equal mixture of $\mathcal{U}(0.001, 0.999)$ and $\mathrm{Beta}(0.5, 0.5)$ is clipped to $[10^{-5}, 1 - 10^{-5}]$, inverse-mapped, and rendered. Its 8192 training and 1024 independent validation images train none of the four configurations.

Q has three stride-2 convolutions (32/64/96 channels), a 256-dimensional fully connected layer, and seven sigmoid outputs. Adam uses learning rate 10^{-3}, batch size 128, and 80 epochs, with random 3×3 average blur. Checkpoints minimize the mean clean/blurred independent-validation MSE. All seven nodes pass preset gates of MAE at most 0.08 and R^2 at least 0.5 on both validation types: clean MAE is 0.0066–0.0190, blurred 0.0066–0.0249. This learned measurement is not error-free.

For direct-target, true-descendant, or true-non-descendant sets A, absolute error compares $Q_k(\widehat{x}^{\mathrm{cf}})$ to true $\gamma(s_k^{\mathrm{cf}})$; $Q_k(x^{\mathrm{GT}})$ supplies only a measurement-error reference. Change evaluation uses factual

Table A10: Matched-record controls with sequential direction updates in the seven-node domino scene (64×64 RGB images). Errors are multiplied by 10^3 and reported as mean $\pm$ sample SD across three encoder seeds; all configurations have 3/3 valid seeds and 896 CF queries per seed and split. The top panel reports target-readout MSE in standardized label units. Updated modules are, respectively, none, affine T, nonlinear T, and readout heads plus D; joint updating shares supervision but differs in parameter count. The bottom panel excludes the direct target from state, effect, and population-mean NMSE. Bold marks the lowest comparable point estimate; blue marks nonlinear calibration.

Configuration	Uncalibrated	Fully fitted affine	Nonlinear calibration	Ordinary joint updating
Target-readout MSE	22.683 ± 4.703	14.377 ± 1.253	12.190 ± 1.201	$\mathbf{8.075} \pm 0.860$

Configuration	Command	State NMSE	Effect NMSE	Mean NMSE	Image MSE
Uncalibrated	ID	35.074 ± 4.111	5.748 ± 0.818	2.932 ± 2.157	2.297 ± 0.043
Fully fitted affine	ID	$\mathbf{33.871} \pm 4.674$	$\mathbf{5.250} \pm 0.405$	$\mathbf{1.830} \pm 0.234$	2.278 ± 0.040
Nonlinear calibration	ID	36.884 ± 5.158	5.703 ± 0.248	2.226 ± 0.404	2.285 ± 0.038
Ordinary joint updating	ID	43.699 ± 6.104	10.194 ± 0.856	3.212 ± 0.599	$\mathbf{1.052} \pm 0.021$
Uncalibrated	Amplitude shift	36.469 ± 4.106	7.143 ± 1.173	4.596 ± 2.928	1.747 ± 0.021
Fully fitted affine	Amplitude shift	$\mathbf{35.764} \pm 5.030$	$\mathbf{7.142} \pm 0.706$	$\mathbf{3.784} \pm 0.103$	1.739 ± 0.027
Nonlinear calibration	Amplitude shift	38.431 ± 5.409	7.250 ± 0.576	3.831 ± 0.209	1.744 ± 0.028
Ordinary joint updating	Amplitude shift	54.163 ± 8.155	20.658 ± 2.591	13.225 ± 2.447	$\mathbf{0.850} \pm 0.019$

reconstruction $x^{\text{rec}} = D_\psi(E_\theta(x))$:

$$\big[Q_k(\widehat{x}^{\text{cf}}) - Q_k(x^{\text{rec}})\big] \quad \text{with} \quad \big[\gamma(s_k^{\text{cf}}) - \gamma(s_k^{\text{fact}})\big].$$

Absolute differences average within nonempty A, equally across targets/commands, then across seeds for mean and SD. Empty sets are not zero. True graphs stratify offline evaluation only.

Table A11: Visible-state MAE measured by an independent image readout model, multiplied by 10^3; mean $\pm$ SD across three encoder seeds. Target, descendant, and non-descendant readouts are scored against true simulator activations; changes use each model's factual reconstruction as the reference. GT-image readout errors are reported separately as a measurement-error reference, not as a substitute for ground truth. Bold marks the lowest point estimate within each split and column.

Configuration	Command	Target	Desc.	Non-desc.	Desc. change	Non-desc. change
Uncalibrated	ID	219.4 ± 4.7	156.6 ± 5.1	88.6 ± 2.4	123.9 ± 0.5	42.4 ± 3.1
Fully fitted affine	ID	209.8 ± 4.5	155.0 ± 6.0	92.8 ± 3.1	122.1 ± 1.4	47.0 ± 3.1
Nonlinear calibration	ID	212.0 ± 4.0	157.2 ± 6.2	92.5 ± 3.0	125.5 ± 1.6	46.5 ± 2.8
Ordinary joint updating	ID	$\mathbf{28.7} \pm 2.5$	$\mathbf{61.8} \pm 4.7$	$\mathbf{17.9} \pm 0.4$	$\mathbf{61.3} \pm 2.4$	$\mathbf{6.1} \pm 0.4$
Uncalibrated	Amplitude	77.1 ± 2.9	86.5 ± 2.1	174.2 ± 5.7	149.3 ± 1.9	144.0 ± 6.5
Fully fitted affine	Amplitude	70.2 ± 3.2	82.9 ± 3.9	181.2 ± 3.6	146.8 ± 3.3	151.7 ± 4.8
Nonlinear calibration	Amplitude	65.1 ± 3.9	81.8 ± 3.8	184.9 ± 3.5	147.8 ± 3.3	156.3 ± 4.5
Ordinary joint updating	Amplitude	$\mathbf{16.0} \pm 1.0$	$\mathbf{48.3} \pm 2.9$	$\mathbf{26.4} \pm 1.9$	$\mathbf{57.4} \pm 1.6$	$\mathbf{16.9} \pm 2.6$

Under amplitude shift, nonlinear target image-readout error falls by 15.52% versus no calibration and 7.31% versus affine, improving all three seeds; affine has better ID target readouts. Raw MAEs (not multiplied by 10^3) show non-descendant error 0.18491, above uncalibrated 0.17422 and affine 0.18121, versus GT-image reference 0.01335. Model non-descendant states stay invariant across commands, but decoded images show substantial readout changes, localizing this preservation failure to the frozen decoding path rather than graph recovery or masking. Joint updating alleviates it without separating head updates, decoder refitting, or their interaction.

D.3.4 Per-seed results

Table A12: Per-seed results for matched-record controls with sequential direction updates; errors are multiplied by 10^3. The four configurations within each seed share acquired pixels and test contexts. All graphs have SHD=0, and all three prespecified seeds are scored.

Seed	Configuration	Readout MSE	ID mean NMSE	Amplitude mean NMSE	ID image MSE
5	Uncalibrated	18.898	1.027	1.969	2.260
5	Fully fitted affine	13.975	1.596	3.826	2.239
5	Nonlinear calibration	11.982	1.972	3.803	2.247
5	Ordinary joint updating	7.996	3.446	15.700	1.031
6	Uncalibrated	21.204	2.496	4.065	2.287
6	Fully fitted affine	13.374	2.065	3.860	2.276
6	Nonlinear calibration	11.106	2.691	4.053	2.284
6	Ordinary joint updating	7.257	3.659	13.168	1.053
7	Uncalibrated	27.948	5.274	7.753	2.344
7	Fully fitted affine	15.781	1.828	3.667	2.319
7	Nonlinear calibration	13.481	2.014	3.637	2.323
7	Ordinary joint updating	8.972	2.532	10.807	1.073

Table A12 retains all three prespecified seeds. Percentage improvements are ratios of means.

D.4 Seed-paired comparisons for the main-text controlled experiment

For seeds 5, 6, and 7, no calibration, fully fitted affine, and nonlinear calibration share fixed encoders, learned graphs, fitting records, queries, and a seed-specific ID-refitted decoder; the original decoder is a separate sensitivity condition. Each seed uses 64 factual contexts, 7 targets, and four commands, totaling 1792 CF queries. ID/amplitude commands are training-label mean $\pm 0.75/\pm 2$ standard deviations. All 3 training seeds remain the statistical denominator, not image count.

For configuration a and seed s, $E_{a,s}$ equally weights targets and commands within that seed. Paired difference $d_s = E_{a,s} - E_{b,s}$ favors a when negative. Table A13 reports $\bar{d}$, paired sample SD, and counts with $d_s < 0$. These three seeds descriptively assess repeatability of this frozen cohort, not other tables' counts or variances.

Table A13: Paired differences for the same three-seed cohort as main-text Table 2. Errors and SDs are multiplied by 10^3; parentheses give the number of improved seeds/3. SD is computed from per-seed differences, not by subtracting the two groups' SDs.

Comparison (first minus second)	Query	Non-target effect NMSE difference	RGB MSE difference
Affine − uncalibrated	ID	-0.700 ± 0.930(2/3)	-0.042 ± 0.021(3/3)
Affine − uncalibrated	Amplitude shift	-0.202 ± 1.583(2/3)	-0.018 ± 0.020(2/3)
Nonlinear − uncalibrated	ID	-0.495 ± 1.022(2/3)	-0.020 ± 0.027(2/3)
Nonlinear − uncalibrated	Amplitude shift	-0.343 ± 1.643(2/3)	0.004 ± 0.023(2/3)
Nonlinear − affine	ID	0.204 ± 0.259(1/3)	0.022 ± 0.006(0/3)
Nonlinear − affine	Amplitude shift	-0.141 ± 0.261(2/3)	0.022 ± 0.004(0/3)

Nonlinear calibration lowers mean non-target effect error versus no calibration in both query types, but worsens one seed each; amplitude-shift RGB improves in 2/3 seeds yet has a slightly higher mean. Against fully fitted affine, nonlinear mean amplitude effect error is lower, but ID effect and both RGB errors are higher. Means and improved-seed counts support only a conditional calibration benefit, not universal nonlinear superiority.

Standard supervised updating uses the same direct-target labels but updates heads and decoder. Its state/effect errors are comparable on the same queries and physical units; its image results cannot support attribution to T alone. Main-table point estimates without per-seed records do not borrow this cohort's SDs or denominator.

D.4.1 Gears-MEC transfer with fixed learned graphs

Gears-MEC's six-node $drive \to g_1 \to g_2 \to ptr \to led \leftarrow target$ includes negative coefficients and a collider, differing from domino while remaining a linear additive SCM. Seeds 300 and 301 supply independently trained aligned grayscale CNNs and RGB decoders with 64×64 inputs. Encoders, learned complete graphs, and decoders stay frozen; only calibration and coordinate-specific local linear mechanisms refit. This does not vary front-end training or test new graph-recovery gains.

Each seed shares 5000 original observations and 900 graph-intervention records, adding 768 calibration records in two rounds per node, 64 records per round. Only verified direct-target values supervise calibration. Nonlinear calibration uses Eq. (3)'s five bases, 100 cumulative-replay steps per round, Adam learning rate 0.01, and preservation weight 0.01; affine fully fits the same terms. New records serve calibration only; mechanisms use original records and ridge coefficient 10^{-3}. Inverse calibration returns to original encoder coordinates before original decoder normalization, with maximum round-trip error below 10^{-10}.

After freezing, each seed uses 64 new factual contexts, six targets, and four commands: 1536 paired CFs at training-label mean $\pm 0.75/\pm 2$ standard deviations for ID/amplitude shift. Each target adds 64 ID direct-readout evaluations; each population query uses 4096 model and 4096 independent simulator samples. Training labels fix command units without test selection. This split does not replace main-table runs; its two independent visual models give $n = 2$, with data and queries shared among calibration configurations within seeds.

Table A14: Calibration transfer on Gears-MEC with fixed graphs, seeds 300/301. All errors are multiplied by 10^3 and reported as mean $\pm$ sample SD across seeds; readouts use independent ID verification. All three configurations share D; $n = 2$ permits only descriptive comparison.

Calibration	Query	ID readout MSE	Non-target effect NMSE	Population mean NMSE	RGB MSE
Uncalibrated	ID	7.517 ± 1.327	1.239 ± 0.229	0.455 ± 0.152	0.923 ± 0.011
Uncalibrated	Amplitude shift	7.517 ± 1.327	1.556 ± 0.384	0.845 ± 0.632	1.141 ± 0.007
Affine	ID	7.295 ± 1.599	1.347 ± 0.287	0.615 ± 0.323	0.926 ± 0.017
Affine	Amplitude shift	7.295 ± 1.599	2.001 ± 0.751	1.363 ± 1.135	1.144 ± 0.016
Nonlinear	ID	7.298 ± 1.364	1.396 ± 0.369	0.742 ± 0.471	0.928 ± 0.019
Nonlinear	Amplitude shift	7.298 ± 1.364	2.324 ± 1.153	1.724 ± 1.581	1.159 ± 0.010

Both calibrations slightly lower mean ID target-readout error but raise mean non-target effect, population-mean, and RGB errors. Nonlinear amplitude effect NMSE rises from uncalibrated 1.556×10^{-3} to 2.324×10^{-3}: better direct readouts do not guarantee better propagation. Two seeds establish neither general failure conditions nor single-component attribution.

Shared-decoder RGB error is separate from effect error against simulator states. Without an independent Gears image-semantic readout, these results do not verify generated target/descendant/non-descendant semantics. Both prespecified seeds remain; results reselect neither calibration nor test scope.

E Decoders and Paired Image Evaluation

E.1 Paired decoder diagnostics with fixed causal-state predictions

This post hoc paired diagnostic changes only the decoder, fixing encoder, calibration, learned graph, and mechanisms to separate RGB quality from causal computation. Test contexts are independent; amplitude shift uses the same command distribution without method selection.

Training and testing. For the seven-node domino scene with 64×64 RGB images, seeds 5, 6, and 7 each refit one decoder shared by uncalibrated, fully fitted affine, and nonlinear configurations. Deterministic replays of 6000 observation and 896 verification images must match saved encodings within 2×10^{-4}. No independent training data, GT CFs, test states, or test images are added. Decoder-only training minimizes ID reconstruction squared error for 40 epochs, batch size 128, Adam learning rate 10^{-4}, using the final epoch without test selection. Other parameters and caches remain unchanged.

After freezing all modules, each seed uses 64 independent factual contexts, 7 targets, and standardized commands $\{-0.75, 0.75, -2, 2\}$: 896 paired CFs per seed/split. Factual/GT images share simulator exogenous context; configurations share factual images, commands, GT, and the independent readout. Each configuration has 5376 CFs across three seeds, not 5376 independent seeds. Frozen ordinary joint updating is only a same-query reference, outside decoder-only pairing.

Table A15: Paired decoder diagnostics; errors are multiplied by 10^3, with mean $\pm$ sample SD across three seeds and 896 paired queries per seed/split. The upper panel compares frozen D and ID-refitted D' on the same images with encoder, calibration, graph, mechanisms, and state predictions fixed. Joint updating changes heads and decoder and is only a reference. Visible-activation MAE uses an independent frozen readout against simulator activations; GT-image readout error is only a measurement reference. The lower panel reports $10^3(\text{MSE}_{D'} - \text{MSE}_D)$: negative is better, and parentheses count improved seeds for each pair. Its SD is computed from per-seed differences, not subtracted group SDs.

Configuration	ID image MSE↓	Amplitude shift image MSE↓	Amplitude shift target visible MAE↓	Amplitude shift non-descendant effect MAE↓
Uncalibrated, frozen D	2.259 ± 0.038	1.698 ± 0.069	79.256 ± 3.823	143.379 ± 9.280
Uncalibrated, ID-refitted D	1.082 ± 0.007	0.854 ± 0.012	16.212 ± 1.226	15.145 ± 1.766
Fully fitted affine, frozen D	2.241 ± 0.044	1.690 ± 0.076	72.223 ± 4.315	150.861 ± 8.913
Fully fitted affine, ID-refitted D	1.037 ± 0.021	0.836 ± 0.029	15.810 ± 1.201	16.719 ± 2.287
Nonlinear calibration, frozen D	2.246 ± 0.044	1.695 ± 0.076	66.577 ± 3.543	155.635 ± 8.886
Nonlinear calibration, ID-refitted D	1.058 ± 0.019	0.859 ± 0.028	15.394 ± 1.384	18.058 ± 2.758
Ordinary joint updating (reference)	1.047 ± 0.034	0.828 ± 0.020	16.001 ± 1.145	16.625 ± 2.312

Paired image-MSE difference (improved seeds) Calibration	ID	Amplitude shift
Uncalibrated	-1.177 ± 0.039 (3/3)	-0.844 ± 0.058 (3/3)
Fully fitted affine	-1.204 ± 0.028 (3/3)	-0.854 ± 0.047 (3/3)
Nonlinear calibration	-1.188 ± 0.027 (3/3)	-0.836 ± 0.047 (3/3)

Image gains with unchanged causal states. All three configurations lower RGB error in both splits, with 3/3 seeds improving per pair. Nonlinear ID pixel MSE falls from 2.246×10^{-3} to 1.058×10^{-3} and amplitude MSE from 1.695×10^{-3} to 0.859×10^{-3}, mean reductions of 52.90% and 49.32%. CF latent variables, physical states, and numerical errors stay unchanged: gains arise from decoder adaptation, not graph or causal-state improvement.

Under amplitude shift, the frozen readout measures target visible-activation MAE falling from 0.06658 to 0.01539 and non-descendant change MAE from 0.15564 to 0.01806, not errors against direct physical-state labels. Its GT-image target MAE, approximately 0.02511, is a measurement reference, neither a theoretical lower bound nor evidence that generated images surpass GT. Affine/uncalibrated refitted decoders remain better on several metrics, and joint updating yields similar images; nonlinear calibration does not uniquely obtain these gains.

Pairing and cost. Non-decoder weights are identical; maximum paired CF-latent difference is 1.78×10^{-15}, and state/effect NMSE differences are both 0. Queries average within seeds before reporting three-seed mean, sample SD, and paired differences; paired SD is not a difference of group SDs.

Each seed renders 8752 images: 6896 training replays and 1856 factual/test-CF images. Test rendering is outside the training-intervention budget; every configuration pays calibration cost 896 and graph-discovery cost. Computation totals approximately 0.192 GPU-hours.

E.2 Paired counterfactual images with learned calibration

This shared-decoder comparison fixes factual index and command. Its environment differs from the main-text physical baseline, precluding cross-figure rankings.

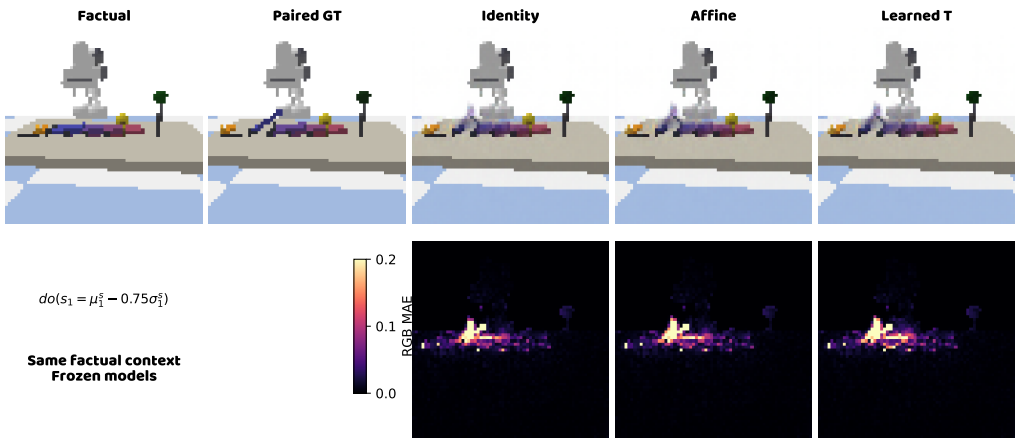


Figure A4: Counterfactual queries in the seven-node domino scene with 64×64 RGB images, at fixed seed 5, factual index 0, target 1, and ID command $u = \mu_1^s - 0.75\sigma_1^s$. All three configurations share factual context/decoder and retain factual residuals in their own coordinates; Learned T is nonlinear calibration. Bottom-row absolute RGB errors share a color scale. This index-prespecified case illustrates an independent frozen diagnostic cohort, not a single-image method ranking.

F Component Ablations and Budget Controls

F.1 Component ablations and intervention-budget diagnostics

Ablations on the seven-node domino scene with 64×64 RGB images use a frozen backbone and prespecified training seeds 5, 6, and 7 to isolate calibration, descendant propagation, and acquisition strategy.

F.1.1 What calibration corrects

With the encoder, decoder, 6000 observations per seed, and 896 target-intervention records fixed, we apply coordinate-wise transformations to the encoded readouts:

$$C_0(z) = z, \qquad C_1(z) = 1.2z + 0.1, \qquad C_2(z) = z + 0.08z^3.$$

Stress conditions were designed post hoc, then formulas, seeds, families, and training schedules were frozen before independent validation, without searching distortion strengths. C_2 is strictly monotone but has non-small tails; the finite bounded affine-plus-tanh family cannot exactly represent its inverse. Each condition compares identity, restricted positive-affine, and monotone nonlinear calibration. All observation/training/validation readouts use the same C; commands and labels retain training-anchor units, with supervision only from direct targets.

Each seed adds 448 independent ID target-validation records, 1344 total, excluded from fitting/selection. Nonlinear target-readout error falls by 44.93%, 88.68%, and 71.06% versus identity (Table A16A), supporting numerical value/readout alignment, not added structural information. Positive-affine calibration leaves standardized image readouts unchanged up to numerical stabilization, but changes command alignment $q_j(u)$ (Appendix B.6). Invertible coordinate-wise maps preserve population conditional independences.

Table A16: Domino component and budget diagnostics. Errors in A/B are in units of 10^{-3}, reported as mean $\pm$ sample standard deviation. A reports target-readout MSE and B reports ID counterfactual RGB MSE; values are not compared across panels. Light blue marks nonlinear calibration, and bold marks the lowest error in each row. No standard deviation is reported for a single seed. C gives records used per strategy, including 896 calibration records. † denotes failure to recover after exhausting the budget, not a successful-recovery cost.

A. Target-calibration error under fixed readout conditions

Condition	Seeds	Uncalibrated	Restricted positive affine T	Nonlinear T
No added distortion	3	23.337 ± 5.417	15.333 ± 2.244	$\mathbf{12.852 \pm 1.690}$
Affine distortion	3	114.354 ± 20.126	15.421 ± 2.135	$\mathbf{12.950 \pm 1.909}$
Cubic distortion	3	61.438 ± 15.002	22.149 ± 3.796	$\mathbf{17.781 \pm 2.835}$

B. Counterfactual image error within each condition
Jointly executable seeds

Condition	Valid/planned	Uncalibrated	Restricted positive affine T	Nonlinear T
No added distortion	2/3	$2.2845 \pm .0321$	$\mathbf{2.2604 \pm .0325}$	$2.2683 \pm .0320$
Affine distortion	2/3	$2.3372 \pm .0430$	$\mathbf{2.2656 \pm .0317}$	$2.2668 \pm .0325$
Cubic distortion	1/3	2.2671	**2.2156**	2.2223

C. Acquisition strategies with calibration and orientation rules fixed

Strategy	Seed 5	Seed 6	Seed 7	Complete graph recovery
Coverage targeting	4496	$11696^{\dagger}$	4496	2/3
Random targeting	9896	6296	2696	3/3
Fixed round-robin	2696	2696	2696	3/3

F.1.2 From readout gains to counterfactual images

All conditions/T variants share each seed's coverage-trajectory encoding records, then learn their own graphs and mechanisms. Inputs are $C(E(x))$ with shared coordinate-wrapped decoder $D'(z') = D(C^{-1}(z'))$; known C^{-1} does not repair identity-group commands. Factual and CF GT images share simulator exogenous context, not model reconstructions.

Each executable seed uses 64 fixed factual contexts and two ID commands for each of 7 targets at training-anchor mean ± 0.75 standard deviations. The first 8 contexts also check implementation without error-based condition selection; the remaining 56 join them for scoring. Only ID queries are tested. Insufficiently oriented units are blocked without substitute predictions or zero errors.

For all three T mappings, seeds 5 and 7 pass under C_0/C_1, and only 5 under C_2: rates 2/3, 2/3, and 1/3. No direction is wrong; blocked graphs retain one unoriented edge. Of 27 planned units, 15 complete image evaluation and 12 are blocked; only seed 5 is common to all conditions/variants. Panel B therefore supports paired comparisons *within each row's condition*, not across unequal valid denominators.

Nonlinear pixel MSE falls by only 0.71%, 3.01%, and 1.98% versus identity, far below readout gains; positive-affine point estimates are slightly lower throughout. This supports neither general nonlinear superiority nor improved graph recovery. With C_0 nonlinear representation, graph, and mechanisms fixed, removing only descendant propagation raises ID CF MSE from 2.2683 ± 0.0320 to 2.9261 ± 0.0060 ($\times 10^{-3}$, common seeds 5 and 7), separating propagation from calibration complexity.

F.1.3 Budget cost and acquisition strategy

Panel C shares backbone, frozen nonlinear T, initial graph, ID command distribution, and graph criteria across three strategies. Each pays $B_{\text{cal}} = 896$, then starts graph acquisition from an empty intervention pool with 1800 records per batch, at most 6 batches: $B_{\max} = 896 + 10800 = 11696$. All use *batch* consistency checks, unlike Algorithm 1's sequential commitment.

This three-policy control uses the graph-sampling stream with $N = 2600000$ defined in Appendix C.2. The equal-record control there and the equal-total-budget control in Appendix F.3 reuse this control's frozen backbone, T, and initial graph, but not its graph samples. For seed 5, the first target-1 batch gives reverse-response score 0.10640 on 0–1, within the one-sided margin 0.05 of threshold 0.15. A target-0 batch is therefore needed, giving 4496 total calls. The equal-total-budget control's first-batch score is 0.05252, which already satisfies the one-sided criterion, giving 2696 total calls. This difference reflects the sampled evidence under the specified streams, not omitted calibration cost.

Coverage exhausts the budget with one unoriented edge on seed 6; 11696 is not successful-recovery cost. Round-robin recovers every seed with 2696 records. This shows no coverage-budget advantage, nor does a fixed-numbered seven-node chain establish general round-robin optimality. Terminal accuracy alone hides cost and failure.

Panel C counts *records actually used by each method*, separately from physical new calls reduced by cached target-specific random streams. The strategies acquire 28800 new graph records, without recounting cached replay. Mapping controls hold data fixed; Table A19 instead fixes total budget and reallocates the uncalibrated configuration's saved calibration budget to graph acquisition. These controls are summarized separately.

Scope and cost. This ID readout-distortion diagnostic uses 1344 independent validation, 1792 CF GT, and 128 factual records. Fitting/graph learning reuse training records without new training calls; thresholds, distortion strengths, and seeds are fixed in advance.

F.2 Module controls for preservation, replay, propagation, and mechanism fitting

Preservation, replay, orientation propagation, factual residuals, and fitting-record controls use the seven-node domino scene with 64×64 RGB images. They share a frozen backbone and planned seeds 5, 6, and 7. Configurations are fixed before evaluation, without OOD selection.

Table A17: Module controls with a fixed visual backbone. Values are mean ± sample SD across training seeds. Bold marks only the lowest point estimate in each error column; light blue marks the full configuration. A uses independent ID validation, while B/C reuse fixed evidence and queries.

A. Calibration objective and replay

3/3 seeds, all values multiplied by 10^3

Configuration	Target-readout MSE↓	Paired difference (ablation − full)
Full configuration	13.327 ± 0.648	0.0000 ± 0.0000
Without observation preservation	$\mathbf{13.320} \pm 0.647$	-0.0071 ± 0.0016
Without cumulative sample replay	15.245 ± 1.377	1.9174 ± 1.3410

B. Post-intervention orientation propagation

3/3 seeds, identical fixed evidence

Configuration	SHD↓	Complete valid graphs
Full configuration	$\mathbf{0.333} \pm 0.577$	2/3
Without post-intervention Meek propagation	4.333 ± 0.577	0/3

C. Residuals and mechanism fitting

Jointly valid 2/3 seeds, all values $\times 10^3$

Configuration	CF pixel MSE↓	Non-target encoding NMSE↓
Full configuration	2.3077 ± 0.0293	$\mathbf{23.9778} \pm 0.7817$
Without descendant factual residuals	2.7120 ± 0.0189	280.1258 ± 0.5836
Non-root mechanisms fitted on all mixed records	$\mathbf{2.3067} \pm 0.0333$	24.8983 ± 0.3083
Non-root mechanisms fitted on observations only	2.3091 ± 0.0292	23.9997 ± 0.8275

A: Preservation term and cumulative replay. With encoder, 896 target records, and 1400 Adam steps fixed, either preservation weight changes from 0.01 to 0 or a single acquisition batch replaces cumulative evidence while retaining model/optimizer states. After all 9 mappings freeze, shared validation adds 448 independent ID records per seed, 1344 total. Removing replay raises mean

target error; removing preservation slightly lowers it but increases training-observation drift. Neither a preservation benefit for target accuracy nor downstream improvement from drift is established.

B: Orientation propagation. Both configurations share initial PC graph, calibration, and full coverage evidence prefixes. Disabling only post-intervention Meek additions retains direct evidence and acyclicity/collider/consistent-extension checks. Final SHDs change from $0, 1, 0$ to $4, 5, 4$, without incorrect directions; complete valid graphs fall from 2/3 to 0/3. This isolates fixed-evidence propagation, not online budget advantage under resampling.

C: Factual residuals and mechanism fitting. With graph, representation, root pool, and decoder fixed, controls separately zero updated descendants' factual residuals, fit non-roots from a mixed pool including self-interventions, or fit observations only. Valid seeds share 64 contexts, 7 targets, and two commands from a fixed ID cache: a retrospective paired control. Seed 6 is insufficiently oriented in all four configurations and blocked without zero imputation; means/SDs use common valid seeds 5 and 7.

Removing factual residuals raises pixel and encoded-readout errors. The mixed pool slightly lowers mean pixel error but raises encoding error, with inconsistent per-seed pixel differences; observation-only fitting stays close to full. Full is therefore not best on every metric. Encoding NMSE is frozen-coordinate error, not independent physical-state error. The mixed pool intervenes only on nodes 0 and 1, so self-target exclusion changes non-root fitting only for node 1. Observation-only fitting also reduces sample size, preventing single-rule attribution.

Cost and interpretation scope. No training-environment calls are added; 1344 independent ID validation records are counted separately. B/C reuse evidence, with $B_{\text{cal}} = 896$ charged per configuration. B's final graph-record counts are $3600, 10800, 3600$.

F.3 Component ablations, fitting tradeoffs, and intervention budgets

These independent paired controls examine affine constraints, nonlinear calibration, and acquisition budgets with separately defined seeds, graph-commitment rules, and evaluation data. They are not pooled as one end-to-end ablation; statistics are descriptive, not significance tests.

Experimental conditions and data access. Appendices C.2, D.3, and F.2 retain graph-stage, interface, and replay/Meek/residual controls. Below, frozen-decoder calibration and equal-budget acquisition controls include calibration cost but do not cover every environment/query in the eight-scenario main tables.

F.3.1 Restricted affine, fully fitted affine, and nonlinear calibration

Restricted affine gains lie in $[0.75, 1.25]$; fully fitted affine retains only $g_j \geq 10^{-6}$ and solves coordinate-wise weighted least squares with the same target squared error and $\lambda = 0.01$ preservation loss. Positivity bounds are inactive in all seeds. Both affine variants and nonlinear calibration share 896 records; uncalibrated does not fit T. Each configuration learns its graph/mechanisms using its own observation normalization, then freezes before independent test generation.

Nonlinear mean target error falls by 43.94% versus no calibration and 13.10% versus fully fitted affine; these are ratios of means, with improvement in all three seeds, not identical per-seed reductions. Yet all four graphs are identical. Mean CF error is only 0.76% below uncalibrated and 0.34% above fully fitted affine, which is lower in both valid seeds. Better state estimates therefore establish neither graph nor image superiority. Similar restricted/fully fitted affine results indicate that the gain interval is not the main constraint here.

This tests direct-target readout correction, not separate image target execution, descendant outcomes, or non-descendant preservation. Affine/nonlinear maps share verified-value supervision; comparison with uncalibrated T does not fully isolate extra supervision from standard updating. Nonlinear changes to local relations/residuals may create tradeoffs with linear mechanisms, but this explanation is not isolated. Cross-source readouts appear in Appendix C.2; standard updating and image semantics in Appendix D.3.

Table A18: Calibration-mapping controls with a frozen decoder; errors are multiplied by 10^3. The upper panel reports mean $\pm$ sample SD and SHD in seed order 5/6/7; the lower panel gives every seed value. State estimation uses 3 seeds, CF the jointly valid 2/3. Seed 6 retains one unoriented edge in every configuration: CF is blocked (–), not zero.

Configuration	Target MSE	CF image MSE	SHD
Uncalibrated	25.232 ± 3.391	2.1892 ± 0.1470	0/1/0
Restricted positive affine	16.278 ± 1.553	$\mathbf{2.1651} \pm 0.1441$	0/1/0
Fully fitted positive affine	16.278 ± 1.553	$\mathbf{2.1651} \pm 0.1441$	0/1/0
Nonlinear calibration	$\mathbf{14.146} \pm 0.850$	2.1726 ± 0.1428	0/1/0

	Seed 5		Seed 6		Seed 7	
Configuration	Target MSE	CF image MSE	Target MSE	CF image MSE	Target MSE	CF image MSE
Uncalibrated	21.971	2.08521	28.740	–	24.984	2.29315
Restricted affine	16.069	2.06321	17.925	–	14.840	2.26705
Fully fitted affine	16.069	2.06321	17.925	–	14.840	2.26705
Nonlinear calibration	13.869	2.07161	15.100	–	13.468	2.27358

Each seed adds 64 independent ID validation records for each of 7 targets, 1344 total. CF uses valid seeds 5 and 7, 64 factual contexts, 7 targets, and two commands per target at ± 0.75 training standard deviations: 1792 GT records and 128 factual images. These 3264 test returns are outside training budgets; predictors receive no GT states. Seed 6 stays in the planned denominator without true-graph completion. Appendix D.2's independent ID validation yields 44.35%/14.54% reductions and is summarized separately.

F.3.2 Acquisition efficiency including calibration cost

Seeds 5, 6, and 7 have budget $B = 11696$ per configuration. Three calibrated configurations pay $B_{\text{cal}} = 896$, leaving 10800 graph records; uncalibrated allocates all 11696 to graph acquisition. Independent random streams pair batches for shared targets; cached reuse still counts against each information budget. Batches contain at most 1800 records with batch consistency checks. Target-value distribution, PC/orientation thresholds are frozen; stopping uses only estimated graph, budget, and available targets, never graph truth.

This equal-total-budget control uses graph-sampling stream $N = 12600000$, defined alongside the cross-control frozen-file correspondence in Appendix C.2. Its nonlinear configurations share training seeds, backbone, T, initial graphs, and batch-orientation code with Table A16C, but use different graph samples; Appendix F.1 explains the different first-batch decisions for seed 5. Total calls for the nonlinear configurations include 896 calibration records, whereas Table A5 reports graph records. Comparing the tables requires matching both the sampling stream and the counting convention.

Table A19: Equal-total-budget controls. Call counts include calibration. An asterisk denotes exhaustion of the budget without full orientation, not the cost of successful recovery. Seed order is 5/6/7 throughout.

Configuration	Complete true-graph recovery	Actual calls	SHD
Nonlinear calibration + coverage targeting	2/3	$2696/11696^*/4496$	0/1/0
Nonlinear calibration + random targeting	3/3	$6296/6296/2696$	0/0/0
Nonlinear calibration + round-robin targeting	3/3	$2696/2696/2696$	0/0/0
Uncalibrated + coverage targeting	2/3	$1800/11696^*/3600$	0/1/0

Coverage has no budget advantage; round-robin recovers all three seeds with 2696 total calls. On seed 6, coverage samples node 1 then repeatedly node 0; final reverse-response score 0.1523 exceeds fixed threshold 0.15, so the two-sided rule retains 0–1. These are zero-based implementation indices. Round-robin samples node 0 first, completing orientation under one-sided evidence. Tra-

jectory, one-/two-sided criteria, and stopping jointly explain this difference, not general inferiority of active design.

Appendix F.2 retains positive and negative preservation, replay, Meek, and fitting results. Divergent target-MSE, observation-drift, and RGB changes do not imply that every module benefits every metric.

G Cross-Environment Results and Model Information Access

G.1 Configurations and statistics for supplementary comparisons

Appendices A.3 and A.2 give the main tables' reported configurations, run links, and statistical conventions. This section reports the MEC mechanism, representation-model, and rigid-body controls separately. -- denotes non-participation in a protocol and exclusion from ranking. Calibration cohorts are scored under their own graph-update rules: frozen-decoder, batch-orientation CF uses common valid seeds 5 and 7; sequential orientation with same-record controls and independent tests completes 3/3 seeds for all four configurations (Appendix D.3).

Physical-counterfactual ILCM. This cell uses seed 42, 1280 training image pairs, and 256 validation image pairs. Four targets, two binary commands, and 32 fixed contexts give 256 queries. Raw floating-point decoded outputs yield pixel MSE 5.707387×10^{-3}; clipping to $[0, 1]$ for display gives 5.696037×10^{-3}, without replacing the prespecified raw metric. Paired GT arrays match archived seed-42 queries for our method and CausalVAE, but the other two main-table methods retain their own summaries; ILCM represents one seed. Its paired-data permissions, label usage, and simulation budget differ, and its decoded command response is weak. Results therefore characterize this adaptation, not ILCM's general capability limit (Appendix A.9).

MEC query groups and training permissions. The MEC mechanism-comparison group contains our method with learned T and the no-graph ablation; both share the frozen visual and calibration interface. Structure learning uses 8 seeds per scene, with a development configuration for Balance. Image and OOD queries instead use seeds 300/301: PnP/Gears complete 2/2 and Balance 1/2, with its second seed blocked by an unoriented graph. Means and SDs use executable seeds only, retaining the failure denominator. The MEC representation-model comparison group uses CausalVAE and ILCM at seed 300 in each scene, with RGB64, 2400 training contexts, and 256 validation contexts. CausalVAE uses observational images and concept labels; ILCM trains on paired images and uses training labels for semantic matching and unit calibration during evaluation. Encoder training, resolution, permitted inputs, and normalized readouts are not fully matched across these groups, which are reported separately and not pooled in main-table rankings.

Graph and counterfactual interfaces. The MEC representation-model group seals graph scores before accessing the true graph, with discretization threshold fixed at 0.1. CausalVAE clamps target values at the input/output of a general K-concept Mask Layer and calls its own decoder, differing from the official four-slot elementwise query. ILCM uses its own conditional implicit mechanisms to convert the set causal-latent coordinate to conditional noise, preserves other noise values, and calls its own decoder; graph extraction passes ID forward/inverse round-trip validation. Each scene uses 32 fixed factual contexts and two ID commands per target: Domino 448 pairs, PnP 384, Balance 320, and Gears 384. GT uses same-exogenous-noise replay; RGB scoring uses the fixed $[0, 1]$ scale without per-image normalization.

OOD units. The MEC representation-model group normalizes physical readouts by true training-concept standard deviations, using 512 reference contexts and 512 independent simulator contexts. Two ID and two OOD commands per target give $4K$ commands. The mechanism group uses frozen training coordinates and scales, with 2048 outcomes per command. Both exclude the directly set target, but their NMSE readouts and protocols differ; multiplication by 10^3 is only a display conversion. Test distributions appear in Appendix A.8.

ILCM quality diagnostics in the representation-model comparison group. PnP ILCM in this group produces nearly identical representations for different ID images, with semantic-readout R^2

near 0 and reconstruction worse than the training-mean image. A two-stage frontend control also fails ID quality requirements and therefore receives no OOD evaluation. Actual scores are retained; collapse in this single-seed adaptation does not characterize the main-table method or other scenes.

Rigid-body conditional-mean propagation control. The paired rigid-body conditional-mean control's training protocol and checkpoints differ from those of the four-scene MEC representation group. Conditional-mean propagation completes 2/3 seeds and CausalVAE 3/3. Over common seeds 42/314, our result is 0.880 ± 0.016 and CausalVAE's is 1.287 ± 0.039 ($\times 10^{-3}$). They share 32 initial states and 256 paired queries, but CausalVAE receives more training concept labels. These statistics belong only to this paired control; they are not per-seed evidence for either the uncalibrated physical aggregate in Table A1 or the main-table learned-calibration result. Appendix A.9 provides this cohort's full settings and the uncalibrated RGB64/128 visual controls.

Table A20: Disaggregated statistics for the MEC mechanism-comparison group, with errors multiplied by 10^3. Two-seed results give mean $\pm$ sample standard deviation; parentheses show executable/planned seed counts.

Task	Method	PnP (2/2)	Balance (1/2)	Gears (2/2)
CF-image MSE	Our method	$1.118 \pm .134$	2.180	$.938 \pm .021$
	No-graph ablation	$2.161 \pm .078$	5.014	$1.859 \pm .093$
ID NMSE	Our method	$1.771 \pm .433$	2.085	$.729 \pm .228$
	No-graph ablation	57.011 ± 2.039	102.146	$93.292 \pm .179$
OOD NMSE	Our method	3.461 ± 1.367	4.101	$1.091 \pm .341$
	No-graph ablation	393.035 ± 10.084	700.012	663.595 ± 5.185

Table A21: Supplementary results for dominoes (RGB images) with our method; errors are multiplied by 10^3. Neither original-coordinate MSE nor mean-image MSE substitutes for physical-readout NMSE in the main text.

Metric	Dominoes (64, 8/10)	Dominoes (128, 9/10)
Counterfactual-image MSE	$2.082 \pm .073$	$2.116 \pm .068$
ID original-coordinate MSE	19.382 ± 11.244	16.208 ± 2.870
OOD original-coordinate MSE	32.297 ± 40.164	18.480 ± 2.970
ID mean-image MSE	$.370 \pm .025$	$.399 \pm .046$
OOD mean-image MSE	$.314 \pm .065$	$.331 \pm .050$

G.2 Baseline ID diagnostics and model selection

System-error differences must be interpreted together with frontend quality and query responses. This section links evaluation checkpoints in the MEC representation-model comparison group to their ID training records, distinguishing training objectives, state readouts, and image responses. Native validation objectives have different units across models and are not used as cross-model RGB reconstruction comparisons.

Model selection and executability. CausalVAE uses an observation-only training set, while ILCM uses paired images. Semantic matching and unit calibration use only ID training labels. Evaluation checkpoints match the SHA of their corresponding training records and are selected at the epoch with the lowest ID validation objective over the full mechanism stage. All four ILCM minima occur at epoch 120; this identifies the selected models but does not establish adequate optimization convergence. Passing no-op and algebraic target-setting checks in each scene also does not mean that generated images correctly respond to commands.

Available ID reconstruction and command responses. PnP ILCM has ID reconstruction MSE 0.007011, worse than the training-mean image's 0.004991, and node-readout R^2 near 0. Representation quality in this single-seed adaptation is inadequate; its error cannot be interpreted as an upper limit on ILCM's general capabilities. For the remaining MEC models, the training records do not

Table A22: Available ID validation and model-selection records for the MEC representation-model comparison group; each row uses seed 300, $n = 1$. R^2 is the range across nodes, not a seed-level confidence interval. Selection uses the native ID validation objective over the full mechanism stage. Diagnostics apply only to these checkpoints; coinciding values do not justify extending them to other main-table runs.

Scene	Method	ID node-R^2 range	Selected/trained epochs
Domino-MEC	CausalVAE	[0.849, 0.984]	86/100
Domino-MEC	ILCM	[0.220, 0.797]	120/120
PnP-MEC	CausalVAE	[0.919, 0.986]	86/100
PnP-MEC	ILCM	[-0.005, 0.044]	120/120
Balance-MEC	CausalVAE	[0.920, 0.995]	98/100
Balance-MEC	ILCM	[0.118, 0.804]	120/120
Gears-MEC	CausalVAE	[0.902, 0.982]	93/100
Gears-MEC	ILCM	[-0.003, 0.826]	120/120

separately provide same-protocol ID reconstruction MSE and visible command responses. Native validation objectives or algebraic checks do not substitute for these quantities.

Physical ILCM uses the seed-42 checkpoint corresponding to Figure 5B, selected at epoch 120 by the full-mechanism-stage ID objective. Across 32 fixed contexts, changing the first domino's command from 0 to 1 yields mean absolute change 2.28×10^{-5} in unclipped decoded outputs, compared with 9.91×10^{-3} in paired physical GT. Executable target setting remains clearly distinct from effective visible response. This physical adaptation has different budgets and supervision permissions; complete settings appear in Appendix A.9.

Scope of the comparison. These diagnostics apply only to the listed MEC and physical runs; they do not automatically cover CAUSAL3D, CausalWorld, or other models with the same names. ILCM already uses interventional paired data, so the performance gap cannot be attributed to the baseline lacking interventions. Our interface comparisons examine how known verified intervention values are used for calibration. The effects of data permissions, representation quality, and update scope on system differences are interpreted separately.

H Assumptions and Scope of Related Methods

H.1 Related-method assumptions and scope

Tables A23 and A24 expand the settings in Figure 2: identifying unknown factors, calibrating known variables, and selecting experiments are distinct tasks. The tables describe original inputs, supervision, modules, and graph outputs, not equal-budget performance or proof of priority.

Visual variables and intervention information. Causal representation learning targets causally meaningful generative factors rather than reconstruction alone (Schölkopf et al., 2021). CausalVAE uses concept supervision (Yang et al., 2021); CITRIS (Causal Identifiability from Temporal Intervened Sequences) uses temporal images and known targets, and iCITRIS includes instantaneous effects (Lippe et al., 2022; 2023b). BISCUIT (Causal Representation Learning from Binary Interactions) learns variable–interaction correspondence (Lippe et al., 2023a). The implicit latent causal model (ILCM) uses paired observations around unknown random interventions and permits subsequent graph extraction (Brehmer et al., 2022). Identification from interventional-distribution geometry can require neither target indices nor actual values in the corresponding settings (Ahuja et al., 2023). These methods use different graph-learning stages; later extraction is not absence of graph capability. Our known target identities and actual values are stronger supervision, used to calibrate an existing visual correspondence while preserving decoder coordinates.

Control, acquisition, and structure. Independently controllable factors (ICF) jointly learns pixel factors and policies using selectivity feedback, without estimating their adjacency graph (Thomas et al., 2018); Causal Curiosity learns behaviors that distinguish environment factors (Sontakke et al.,

2021). Action-controllable factorization (ACF) instead estimates policies to correct given transition-data distributions, not to acquire new experiments (Rodriguez-Sanchez et al., 2025). The visual causal discovery network (V-CDN) learns keypoints and structure from given videos (Li et al., 2020). For observed variables, active strategies (Hauser & Bühlmann, 2014), active intervention targeting (AIT) (Scherrer et al., 2021), causal Bayesian experimental design (CBED) (Tigas et al., 2022), and causal amortized active structure learning (CAASL) (Annadani et al., 2024) choose experiments as detailed in the tables. Our calibration acquisition is fixed round-robin; only subsequent graph discovery selects targets from current ambiguity. Values follow a fixed distribution.

Extrapolation and task-specific structure. Rep4Ex uses action–latent invariance for intervention extrapolation (Saengkyongam et al., 2024); action-informed representations therefore have precedent. Causality-Seeking Agent learns context-relevant subgraphs from structured state–action trajectories (Zhao et al., 2025), reward-oriented causal representation learning recovers task-relevant variables/structure through adaptive exploration (Yan et al., 2025), and Active Causal Experimentalist (ACE) studies mechanism estimation with a known graph (Cooper & Velasquez, 2026). Their specialized outputs should not be interpreted as general capability deficits.

Our scope. We calibrate the magnitude of pixel-derived readouts with known semantics, then connect target settings, graph-conditioned consequences, and decoding. Constraint-based and interventional-equivalence-class methods supply graph-learning tools (Colombo & Maathuis, 2014; Hauser & Bühlmann, 2012); invertible coordinate-wise calibration adds no identifiability guarantee. Empirical gains require matched-data and matched-budget controls. Image counterfactuals and OOD outcomes have distinct query/distribution definitions; a checkmark denotes neither a theorem nor overall superiority.

Table A23: Assumptions and information available to related methods. A "concept anchor" is supervision that associates a visual coordinate with a known physical concept during training. More information does not imply a stronger method; these conditions determine whether a comparison is fair.

Work	Input/samples	Node or variable supervision	Intervention information and key conditions
CausalVAE (2021)	Images	Concept labels	Causal layer and latent-space editing; no active environment acquisition
ICF (2018)	Pixel interaction trajectories	Independently controllable factors	Selectivity feedback jointly trains representations and policies
Causal Curiosity (2021)	Interaction trajectories	Learned environment factors	Executable experimental behaviors; not standard node-setting supervision
CITRIS (2022)	Temporal images	Learned latent variables	Known target labels; actual values not required
iCITRIS (2023)	Temporal images	Learned latent variables	Known targets, including instantaneous effects
BISCUIT (2023)	Interactive temporal images	Learned variable/interaction correspondence	Binary interaction mechanisms
ILCM (2022)	Paired pre/post-intervention observations	Weakly supervised variable learning	Pairing with shared exogenous conditions; unknown targets
Ahuja et al. (2023)	Sets of observational/interventional distributions	Representation identification	Targets and actual values need not be known in the corresponding settings
ACF (2025, preprint)	Image-transition triplets	Controllable-factor identification	Given data; policy estimation corrects the acquisition distribution
V-CDN (2020)	Video images	Keypoint and structure learning	Given data from different configurations; no active experiment selection
Rep4Ex (2024)	High-dimensional features, actions, outcomes	Representation identification	Assumptions including linear structure from actions to latent variables
Hauser–Bühlmann (2014)	Observed variables	Given nodes	Experimental design over equivalence classes of candidate graphs
AIT (2021)	Observed variables	Given nodes	Given intervention-target space and generator
CBED (2022)	Observed variables	Given nodes	Causal-model posterior and executable interventions; active value selection
CAASL (2024)	Variables/experiment history	Given nodes	Simulator pretraining with true graphs in the reward; active value selection
Causality-Seeking Agent (2025)	Structured states–actions	Given state variables	Learned context and its relevant subgraph
Reward-oriented CRL (2025)	High-dimensional mixed observations/rewards	Learned task-relevant variables	Linear mixing, causal, and reward assumptions
ACE (2026, preprint)	Variables/experiment history	Given graph and nodes	Learned mechanisms and experimental policy; active value selection
This work	**RGB images and environment experiments**	**Known, concept-anchored nodes**	**Known hard-intervention targets and setting values; frozen after calibration**

Interpretation: Using intervention data does not imply identical supervision. Our assumptions of known nodes and setting values are stronger than those of several weakly supervised representation methods, but do not require their particular temporal or factual–interventional pairing formats. High-dimensional inputs in Rep4Ex and reward-oriented methods are not equated here with image experiments; context-dependent subgraphs are also distinct from our static post-settlement graph. Sources for all rows appear in Appendix H.1.

Table A24: Capabilities and representation constraints of related methods. ✓: explicitly included in the original method; △: a different form or a subsequent module; —: not a component of the method, without ruling out extensions. The last column states what is constrained, rather than treating different tasks as a single type of calibration.

Work	Repr. learn.	Active exp.	Graph output/task	Representation constraints
CausalVAE	✓	—	Static causal graph	Concept supervision and causal layer
ICF	△	✓	No explicit graph-recovery task	Independent controllability and intervention selectivity
Causal Curiosity	△	✓	No general explicit graph recovery	Experiments distinguish environment factors
CITRIS	✓	—	△ Subsequent temporal graph	Intervention-target constraints
iCITRIS	✓	—	Instantaneous and temporal graphs	Intervention-target constraints
BISCUIT	✓	—	△ Subsequent temporal graph	Binary interaction-mechanism correspondence
ILCM	✓	—	△ Post-training graph extraction	Pre/post-intervention pairing
Ahuja et al.	✓	—	Not the main output	Geometry/constancy of interventional distributions
ACF	✓	—	No explicit graph-recovery task	Action effects relative to natural transitions
V-CDN	✓	—	Structure among keypoints	Structure and dynamics in given videos
Rep4Ex	✓	—	Not a latent-graph discovery task	Action–representation invariance
Hauser–Bühlmann	—	✓	Graph equivalence classes/orientation	—
AIT	—	✓	Interfaces with graph-learning backends	—
CBED	—	✓	Graph and mechanism posterior	—
CAASL	—	✓	Interfaces with graph-learning backends	—
Causality-Seeking Agent	△	✓	Context-dependent subgraph	Latent context state
Reward-oriented CRL	✓	✓	Task-relevant coarse structure	Reward relevance
ACE	—	✓	Mechanism estimation for a known graph	—
This work	△	✓	**Static graph estimates/unoriented edges**	**Invertible coordinate-wise calibration to actual setting values**

Scope: Active value selection means selecting or optimizing the numerical hard-intervention value of a known node using feedback; it cannot be inferred merely from a continuous-action policy. Our method actively selects only targets, sampling values from a prespecified distribution without optimizing them. Our △ denotes coordinate learning and readout calibration with existing concept identities, not disentanglement of unknown factors. Active experiments occur only during graph discovery; calibration acquisition uses fixed round-robin. Factor learning in ICF and Causal Curiosity is not a general latent-structure identification guarantee, and the latent context learned by the Causality-Seeking Agent is not node recovery from pixels. Subsequent graph extraction does not imply an absence of graph-discovery capability. Checkmarks imply neither theoretical guarantees nor empirical superiority under equal budgets.