WHAT INTERVENTION LOSS CERTIFIES: SHARP BOUNDS FOR FACTOR SEPARATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A small intervention loss can coexist with substantial factor mixing when the available actions do not test the relevant dependency. We determine what separate upper bounds on these losses certify about a frozen representation. The largest compatible off-factor variation equals a fractional-cover linear program: an orthogonal variance decomposition reduces the exponential interaction problem to one variable per forbidden factor, and an additive construction attains its value. Simultaneous confidence bounds permit data-dependent selection of the certificate weights. Predictable variance adaptation and deterministic fiber envelopes reduce statistical conservatism while retaining the same protected output. On nonlinear neural encoders, predictable concentration lowers a pair-mask certificate from 0.00169 to 0.00130 against oracle leakage 0.00085; with overlapping tests only, it nearly halves the certificate. On the public Pendulum generator, three image-only autoencoder families receive raw-output certificates within mean factors 1.02–1.04 of exact finite-reference errors at $32{,}768$ pairs per action, with exhaustive range-audit costs reported separately. The results distinguish dependence certification from normalized disentanglement scores, full identification and information retention.

1 INTRODUCTION

An intervention loss measures whether an encoder leaves selected outputs unchanged when an action modifies other factors. This suggests a natural route to disentanglement, but the loss can be misleading for three distinct reasons. First, an action family may target every factor without distinguishing every pair. Second, an action that distinguishes a dependency may occur too rarely to meaningfully penalize mixing. Third, even a well-separated action family may be collected only in states where the unwanted dependence is absent. In each case, collecting more data from the same deficient design does not resolve the underlying ambiguity.

Existing identification theory establishes why observational data alone are insufficient and how auxiliary information or sparse mechanisms can restrict representation ambiguities (Locatello et al., 2019; Khemakhem et al., 2020; Lachapelle et al., 2022; 2026b). The quantitative question we address is: after an encoder has been fitted, what can its measured off-target loss actually certify? A valid answer must specify the error being bounded, preserve the output scale, and distinguish the geometry of the interventions from the distribution of collected pairs. It must also remain valid when validation data are used to determine which actions receive the largest certificate weights.

A minimum separation probability provides a first answer. For each ordered pair of factors, it measures how often an action changes the first factor while leaving the second untargeted. Combining this quantity with a reference-to-replay overlap bound converts intervention loss into a bound on conditional off-factor variation. The resulting constant is sharp, but averaging action losses can discard information that is important for a fitted encoder: a rare action may directly test a dependency that the encoder does not contain. Weighting every residual by the frequency of the corresponding action can therefore yield a loose certificate even when all relevant action losses have been measured.

Our central result removes this information loss. Given separate upper bounds on the action losses, we show that the largest compatible off-factor variation is exactly the value of a small linear program (LP). Its primal formulation assigns variance to forbidden factors, while its dual covers those

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.

factors using the tested actions. Although a nonlinear encoder may contain exponentially many interaction terms, each interaction can be assigned to one of its forbidden factors without increasing any intervention energy. This yields the reduction to the LP, while a matching additive construction establishes tightness. Classical block factorization provides the weighted inequalities underlying the analysis (Caputo & Parisi, 2021); our contribution is to characterize the resulting exact upper-information envelope for masked representation loss and to develop a statistically valid implementation and empirical evaluation.

Our evaluation isolates the resulting sources of looseness. Overlapping-mask-only experiments distinguish the benefit of fractional weighting from that of selecting precise actions. Frozen nonlinear encoders expose a separate bottleneck in which range corrections can dominate the measured loss. Variance-adaptive confidence bounds and verified fiber envelopes address this bottleneck without altering the underlying LP. We then evaluate the procedure on a public Pendulum generator with image-only autoencoders and raw posterior means, using IRS, DCI, and MIG as complementary diagnostics. Finally, rare-action, support-hole, and drift controls delineate the scope of the certificate and clarify why a certified dependence bound should not be interpreted as a claim of causal recovery.

2 Related work

CITRIS identifies causal factors from intervened temporal sequences, including multidimensional factors; iCITRIS additionally models instantaneous effects (Lippe et al., 2022; 2023). Their temporal likelihoods do not require every untargeted coordinate to remain constant. Mechanism-sparsity theory characterizes permissible entanglement through graph-preserving transformations (Lachapelle et al., 2022; 2026a;b). Our signature restriction specializes this geometry to paired resampling. Invariance and information preservation remain distinct requirements (Yao et al., 2025). The certificate is a post-training evaluation procedure, not an alternative temporal learner.

Interventional evaluation also predates this work. The Interventional Robustness Score (IRS) measures normalized worst-case deviations under changes to nuisance factors (Suter et al., 2019). DCI separates disentanglement, completeness, and informativeness through predictive feature importance (Eastwood & Williams, 2018), while the mutual information gap (MIG) measures the difference between the two largest factor–coordinate mutual informations (Chen et al., 2018). Our residual instead averages squared nuisance variation under a declared reference law. It is a scale-dependent error with an upper confidence bound, not a replacement normalized score. Section 6.3 and Appendix E.4 evaluate these complementary quantities on the same representations.

The variance decomposition and fractional block-cover inequalities are classical (Efron & Stein, 1981; Caputo & Parisi, 2021; Caputo et al., 2025); the refresh energies are unnormalized total-effect Sobol' indices (Owen, 2012). Our mathematical contribution is the exact supremum compatible with masked energy upper bounds, including a polynomial reduction and an attaining construction, rather than a new general variance inequality. Statistical implementation uses existing empirical Bernstein (Maurer & Pontil, 2009) and predictable plug-in bounds (Waudby-Smith & Ramdas, 2024). Intervention-cost optimization also has established precedents (Addanki et al., 2020). Unlike controlled-world-model identification through Gaussian dynamics and excitation (Zhang et al., 2026), our protected error is factor leakage under a product refresh experiment.

3 Intervention geometry and the protected error

Let $C \in \Omega = \prod_{i=1}^{K} \Omega_i$, with $K \geq 2$, and let $S = h(C)$ be an observation. An encoder induces $F = f \circ h : \Omega \to \mathbb{R}^K$. For the qualitative identification theorem, each Ω_i is an open interval and h is a smooth embedding. The finite-error results require only a product reference law and square integrability; they also apply to discrete factors and noninvertible encoders. A known action mask $T_a \subseteq [K]$ is assigned to the corresponding output coordinates. Under an exact intervention, $C'_{-T_a} = C_{-T_a}$. For a replay law $B(dc, a, dc')$, define

$$\mathcal{L}_{\mathrm{int}}(F; B) = \mathbb{E}_B \sum_{j \notin T_A} \big(F_j(C') - F_j(C)\big)^2. \tag{1}$$

The analysis concerns this penalty, not the optimizer of an arbitrary combination of prediction, reconstruction, and isotropy objectives. In particular, constant outputs have zero loss. Information preservation therefore requires a separate assumption or measurement.

Fix a finite used action family and reference probabilities $p_a > 0$. Write $S_i = \{a : i \in T_a\}$ for factor i's signature and define

$$q_{ij} = \sum_a p_a \mathbf{1}\{i \in T_a, j \notin T_a\}, \qquad q_j = \min_{i \neq j} q_{ij}, \qquad q_* = \min_j q_j. \tag{2}$$

The design separates factors exactly when $q_* > 0$. Equivalently, the signatures are pairwise incomparable under inclusion. Equal signatures fail this condition, even if the tied factors are frequently intervened on.

Assumption 1 (Full fiber support)**.** *F is a C^1 diffeomorphism onto its image. Replay has a positive density on Ω, every used action has positive propensity almost everywhere, and its conditional kernel has positive density throughout the targeted fiber while fixing the untargeted factors.*

Theorem 1 (Signature characterization)**.** *Under Assumption 1, $\mathcal{L}_{\mathrm{int}}(F; B) = 0$ if and only if*

$$\partial_i F_j(c) = 0 \quad \textit{whenever } S_i \not\subseteq S_j. \tag{3}$$

Thus $q_* > 0$ implies $F_j(c) = \psi_j(c_j)$ with invertible scalar ψ_j. Conversely, if $S_i \subseteq S_j$ for $i \neq j$, the invertible shear $F_j(c) = c_j + \varepsilon c_i$, with other coordinates unchanged, has zero loss for every $\varepsilon \neq 0$.

The proof turns almost-sure invariance into constancy on each fiber, then differentiates along its coordinates. The inclusion pattern is the action-only form of the graph-preserving restriction in Lachapelle et al. (2026b, Proposition 3 and Lemma 1). Its operational consequence is concrete: no amount of optimization can make the penalty test a dependency excluded by all available masks.

For finite-error analysis, let $Q = \bigotimes_i Q_i$ be a declared product reference law. Action a independently refreshes C_{T_a} from Q_{T_a} and leaves the complement fixed; denote its pair law by Ξ_a. For $F_j \in L^2(Q)$, define

$$\mathcal{R}_j = \mathbb{E}_Q\big(F_j(C) - \mathbb{E}_Q[F_j(C) \mid C_j]\big)^2. \tag{4}$$

This is the squared distance to functions of the assigned factor. Loss and variation must be evaluated for the same frozen map, including any coordinatewise normalization. An MCC assignment optimized separately from the masks does not change that requirement.

4 From aggregate loss to an optimal action-resolved certificate

4.1 Reference energy, overlap, and partial separation

Write $B_a(dc, dc') = B(dc, A = a, dc')$. A sufficient replay condition is

$$p_a \Xi_a \leq \kappa B_a \quad \text{for every used action,} \qquad \kappa < \infty. \tag{5}$$

This domination includes states, action propensities, and intervention outcomes. Positive action counts alone do not establish it. Under exact reference collection, $B_a = p_a \Xi_a$ and $\kappa = 1$.

Theorem 2 (Aggregate block-factorization bound)**.** *Under Equation* (5)*,*

$$\sum_j q_j \mathcal{R}_j \leq \frac{\kappa}{2} \mathcal{L}_{\mathrm{int}}(F; B). \tag{6}$$

If $q_* > 0$, then $\sum_j \mathcal{R}_j \leq \kappa \mathcal{L}_{\mathrm{int}}/(2q_*)$. For any deployment law $\nu \leq wQ$, the same fixed projections $\psi_j = \mathbb{E}_Q[F_j \mid C_j]$ satisfy

$$\sum_j \mathbb{E}_\nu (F_j - \psi_j)^2 \leq \frac{w\kappa}{2q_*} \mathcal{L}_{\mathrm{int}}(F; B). \tag{7}$$

The factor $1/(2q_*)$ is sharp, including among invertible encoders.

To see the sharpness, take independent unit-variance factors, $F_1 = C_1$, $F_2 = C_2 + \varepsilon C_1$, and singleton refresh probabilities p and $1 - p$, where $p \leq 1/2$. Then $\mathcal{R}_2 = \varepsilon^2$, $\mathcal{L}_{\mathrm{int}} = 2p\varepsilon^2$, and $q_* = p$. Rare separation therefore permits substantial mixing at small average loss.

The same argument describes incomplete separation. Let $D_j = \{i : S_i \subseteq S_j\}$ and $I_j = [K] \setminus D_j$. The appropriate residual is

$$\mathcal{R}_j^{\text{allow}} = \mathbb{E}_Q\big(F_j - \mathbb{E}_Q[F_j \mid C_{D_j}]\big)^2. \tag{8}$$

It vanishes precisely on the dependencies left admissible by the masks. If $I_j \neq \varnothing$, set $r_j = \min_{i \in I_j} q_{ij}$; otherwise set $r_j = 0$. Then $\sum_j r_j \mathcal{R}_j^{\text{allow}} \leq \kappa \mathcal{L}_{\text{int}}/2$. A bound on $\mathcal{R}_j^{\text{allow}}$ is not a bound on the additional mixing inside D_j.

The underlying identity will also yield a stronger certificate. For the orthogonal product-space decomposition $F_j = \sum_{U \subseteq [K]} F_{j,U}$, put $v_{j,U} = \|F_{j,U}\|^2_{L^2(Q)}$. Each action supplies the half-energy

$$e_{aj} = \frac{1}{2}\mathbb{E}_{\Xi_a}(F_j(C') - F_j(C))^2 = \sum_{U : U \cap T_a \neq \varnothing} v_{j,U}, \qquad j \notin T_a. \tag{9}$$

Meanwhile, $\mathcal{R}_j^{\text{allow}} = \sum_{U \not\subseteq D_j} v_{j,U}$. The aggregate bound replaces every nonzero interaction weight by its smallest possible value. Retaining the individual e_{aj} avoids that replacement.

4.2 The exact upper-information envelope

Let $\mathcal{A}_j = \{a : j \notin T_a\}$ and suppose finite nonnegative upper bounds $u_{aj} \geq e_{aj}$ are available. Define the packing program and its fractional-cover dual:

$$\begin{aligned} \mathcal{C}_j(u) &= \max_{x_i \geq 0,\ i \in I_j} \left\{ \sum_{i \in I_j} x_i : \sum_{i \in I_j \cap T_a} x_i \leq u_{aj} \ \ (a \in \mathcal{A}_j) \right\} \\ &= \min_{y_a \geq 0} \left\{ \sum_{a \in \mathcal{A}_j} y_a u_{aj} : \sum_{a \in \mathcal{A}_j : i \in T_a} y_a \geq 1 \ \ (i \in I_j) \right\}. \end{aligned} \tag{10}$$

The value is zero when I_j is empty. With only a subset of tested actions, an uncovered forbidden factor gives an infinite value rather than a zero certificate.

Theorem 3 (Optimal certificate from action-energy upper bounds)**.** *Assume Q is a product of nondegenerate factor laws. Then $\mathcal{R}_j^{\text{allow}} \leq \mathcal{C}_j(u)$. Moreover,*

$$\mathcal{C}_j(u) = \sup\left\{\mathcal{R}_j^{\text{allow}}(G; Q) : G \in L^2(Q),\ e_{aj}(G) \leq u_{aj}\ \forall a \in \mathcal{A}_j\right\}. \tag{11}$$

Thus no smaller uniform bound follows from these upper bounds alone. If a valid variance bound $\mathrm{Var}_Q(G) \leq v_j$ is also imposed, the exact supremum is $\min\{v_j, \mathcal{C}_j(u)\}$. Each coordinate's uncapped optimum is realizable by an invertible linear shear when the factors have finite positive variances.

The key reduction transfers interaction variance to forbidden singleton components. For each forbidden interaction U, choose one index $i(U) \in U \cap I_j$ and move its variance to $x_{i(U)}$. Whenever an action touches $i(U)$, it also touches U. Every action energy can only decrease, while the total forbidden variance is unchanged. Conversely, centered unit-variance functions of independent coordinates realize any feasible x additively. This proves that the exponential interaction optimization has exactly the polynomial-size value in Equation (10). The variance cap follows by adding $\sum_i x_i \leq v_j$; an allowed component can supply any remaining variance when an exact variance is specified. Appendix A, specifically Appendix A.4, gives the complete argument.

The theorem concerns upper-bound information, not exact equality constraints or every known structural property of an encoder. Additional reconstruction, variance, smoothness, or inverse-conditioning constraints can tighten the feasible class. For the same reason, the invertible construction proves coordinatewise sharpness; it does not assert that all coordinates' independent optima are attained simultaneously by one invertible map.

Under full separation, $y_a = p_a/q_j$ is dual-feasible. Hence

$$\mathcal{C}_j(e) \leq \frac{\sum_{a \in \mathcal{A}_j} p_a e_{aj}}{q_j}, \tag{12}$$

which is the coordinatewise aggregate certificate. If singleton actions are observed, $\mathcal{C}_j(e) \leq \sum_{i \neq j} e_{\{i\},j}$ also follows. This singleton bound can already be sharp for additive encoders. The same aggregate comparison uses r_j for the partial residual. The benefit of the LP comes from the combination of overlapping masks and heterogeneous energy or uncertainty, not from assuming that aggregate bounds are the only alternatives.

An analytic example isolates the difference. Use three singleton actions with probabilities $(0.49, 0.01, 0.50)$ and $F_3 = C_3 + \varepsilon C_1$. The coordinatewise aggregate bound is $49\varepsilon^2$, whereas the mask-resolved bound equals the true error ε^2. The rarely tested second factor carries no residual dependence. A variance cap does not remove this gap when ε^2 is sufficiently small. Replacing 0.01 by a smaller positive frequency makes the uncapped ratio arbitrarily large.

Overlapping masks give a complementary example with no singleton measurements. For output 4, let the available tests refresh $\{1, 2\}$, $\{1, 3\}$, and $\{2, 3\}$, with respective energy upper bounds a, b, c. The exact certificate is

$$\mathcal{C}_4(a, b, c) = \min\left\{a + b,\ a + c,\ b + c,\ \frac{a + b + c}{2}\right\}. \tag{13}$$

The three two-action covers and the half-weight three-action cover are the extreme dual choices. If $c \geq a + b$, the certificate is $a + b$: the first two tests already cover every forbidden factor, so a loose third upper bound need not weaken the conclusion. Under equal action probabilities, the coordinatewise aggregate bound is instead $(a + b + c)/2$. This example also shows how statistical uncertainty enters the decision. The LP uses an uncertain test only when its upper bound improves the cover, while simultaneous validity protects that data-dependent choice.

4.3 Replay transfer and imperfect interventions

Action-resolved certification can use heterogeneous overlap. Write $B^a = B(\cdot \mid A = a)$ and suppose $\Xi_a \leq \kappa_a B^a$. If $b_{aj} = \frac{1}{2}\mathbb{E}_{B^a}(F_j(C') - F_j(C))^2$, then $u_{aj} = \kappa_a b_{aj}$ is valid. The κ_a here concerns conditional pair distributions, whereas κ in Equation (5) also includes the reference-to-behavior action probabilities. Neither is estimated by counting actions alone.

Exact domination fails when untargeted coordinates drift continuously away from the reference equality set. Let $\widetilde{\Xi}_a \leq \kappa_a B^a$ admit a coupling with Ξ_a, and define δ_{aj}^2 as one half the expected squared difference between the two paired increments of F_j. Minkowski's inequality gives

$$e_{aj} \leq \left(\sqrt{\kappa_a b_{aj}} + \delta_{aj}\right)^2. \tag{14}$$

These corrected upper bounds can be inserted directly into the LP. The coupling and discrepancy bounds are additional premises; a fitted point estimate of drift does not automatically verify them. The pooled counterpart,

$$\sum_j q_j \mathcal{R}_j \leq \frac{1}{2}\left(\sqrt{\kappa \mathcal{L}_{\text{int}}} + \delta_F\right)^2, \tag{15}$$

and its partial-separation version are proved in Appendix A.6. State overlap, outcome overlap, and intervention error remain separate from target-mask geometry.

5 Finite-sample validation

Freeze the encoder, its output assignment, and its normalization before validation. For each tested action, collect $n_a \geq 2$ independent pairs from a declared conditional law. Let $Y_{aj} = \frac{1}{2}(F_j(C') - F_j(C))^2 \in [0, M_{aj}]$, where M_{aj} is a justified bound, not the largest observed value. Denote the sample mean and unbiased variance by $\bar{Y}_{aj}$ and $\widehat{v}_{aj}$. For $m \geq 1$ tested action-output cells, the empirical Bernstein (EB) bound is

$$U_{aj} = \min\left\{M_{aj}, \bar{Y}_{aj} + \sqrt{\frac{2\widehat{v}_{aj}\log(2m/\delta)}{n_a}} + \frac{7M_{aj}\log(2m/\delta)}{3(n_a - 1)}\right\}. \tag{16}$$

Corollary 4 (Simultaneous validation of the optimized certificate)**.** *With probability at least* $1-\delta$*, all conditional population half-energies are bounded by* U_{aj}*. Consequently, the LP evaluated at* U*, or at the overlap- and coupling-corrected values from Section 4.3, simultaneously bounds the corresponding* $\mathcal{R}_j^{\mathrm{allow}}$*. Its dual weights may be selected using these same upper bounds.*

The result follows from the empirical Bernstein inequality and a union bound (Maurer & Pontil, 2009). Dependence between output coordinates within a pair causes no problem. Independence of validation units and freezing the encoder do matter. Selecting an encoder or an estimated normalization using these same observations requires an additional simultaneous argument. A known reference variance cap may be applied afterward; normalizing by an empirical variance does not make that population cap known.

5.1 Variance adaptation without changing the protected output

The M_{aj}/n_a term can dominate a small neural loss even when its empirical variance is tiny. We therefore also use the predictable plug-in empirical Bernstein construction of Waudby-Smith & Ramdas (2024). For one cell, suppress its indices, set $X_t = Y_t/M$, fix $N = n_a$ before sampling, and let $\eta = \delta/m$. Define

$$\widehat{\mu}_t = \frac{1/2 + \sum_{s\le t} X_s}{t+1}, \qquad \widehat{\sigma}_t^2 = \frac{1/4 + \sum_{s\le t}(X_s - \widehat{\mu}_s)^2}{t+1}, \qquad \lambda_t = \min\left\{0.9, \sqrt{\frac{2\log(1/\eta)}{N\widehat{\sigma}_{t-1}^2}}\right\}. \tag{17}$$

Only previous observations determine λ_t. With $\psi(\lambda) = -\log(1-\lambda) - \lambda$, the one-sided upper bound is

$$U^{\mathrm{PrPl}} = M \min\left\{1, \frac{\sum_{t=1}^N \lambda_t X_t + \log(1/\eta) + \sum_{t=1}^N \psi(\lambda_t)(X_t - \widehat{\mu}_{t-1})^2}{\sum_{t=1}^N \lambda_t}\right\}. \tag{18}$$

The reflected empirical Bernstein supermartingale bounds the cell's mean with probability at least $1-\eta$. A union bound therefore preserves Corollary 4 with U^{PrPl} in place of U (Appendix B.2). We compare the two procedures at the same confidence level; selecting their minimum would require a joint error allocation. Known reference-variance caps remain applicable after transferring the energies to the reference law. A variance cap on the output is not, by itself, a tighter bound on every squared increment.

A deterministic envelope can improve both procedures without a nonlinear readout. Let $\mathrm{osc}_{T_a} F_j(c_{-T_a})$ be the range of F_j when only targeted factors vary. Then

$$M_{aj}^{\mathrm{fib}} = \tfrac{1}{2} \sup_{c_{-T_a}} \left[\mathrm{osc}_{T_a} F_j(c_{-T_a})\right]^2 \tag{19}$$

is valid because the two endpoints share c_{-T_a}. On a finite product grid, complete enumeration verifies this envelope and the reference variance exactly for the evaluated map. The computation is charged separately from sampled validation. Neither an observed validation maximum nor a fitted variance may silently replace a support bound. An energy-only lower bound in Appendix B.4 shows that a range-dependent rare-event correction remains necessary in general.

For an unbounded neural output, one can fix $G_j = \tanh(F_j)$ before validation and certify G instead. Then $M_{aj} = 2$ and $\mathrm{Var}_Q(G_j) \le 1$ are valid without estimating a range. The injective coordinatewise readout preserves exact dependence on allowed factor subsets, but changes the numerical residual: its certificate is not automatically a bound for the raw output. The complete procedure, including unavailable-cover decisions and numerical dual feasibility, appears in Appendix B.7. Independent pilot selection can reduce validation cost; the design and allocation formulas appear in Appendix B.

6 Experiments

The evaluation separates certificate geometry, statistical uncertainty, and the training procedure. Every comparison uses one frozen output map and the same validation pairs. Reported uncertainty is

Table 1: Frozen neural encoders, mean ± SE across five training seeds. All error and certificate entries are in 10^{-3} on the same bounded readout. Bounds protect the allowed-class residual, which equals full leakage only under separation. Each tested action has 262,144 pairs; EB and PrPl use identical samples and $\delta = 0.05$.

Design	Full leakage	Allowed residual	EB + LP	PrPl + LP
Singleton	0.502 ± 0.047	0.502 ± 0.047	2.013 ± 0.062	1.246 ± 0.061
Pair masks	0.848 ± 0.088	0.849 ± 0.088	1.689 ± 0.093	1.296 ± 0.091
Nested	1005.6 ± 36.4	0.626 ± 0.111	0.930 ± 0.112	0.751 ± 0.110
Identical	542.7 ± 157.3	0.366 ± 0.024	1.281 ± 0.027	0.801 ± 0.026
No penalty	1050.1 ± 93.8	1050.1 ± 93.5	1103.1 ± 97.3	1101.6 ± 97.3

standard error across independent training seeds or validation repetitions, as specified. The signature study uses 25 neural checkpoints; a public Pendulum-generator study adds nine autoencoder checkpoints trained without intervention losses. Neither experiment claims improved learning performance over CITRIS, iCITRIS, or mechanism-sparsity methods.

6.1 Certificates for nonlinear neural encoders

Five seeds per condition train a GELU encoder $8 \to 64 \to 64 \to 4$ on four independent $U[-1, 1]$ factors observed through an invertible nonlinear mixing map. Only observations and target masks enter optimization. Prediction, sliced isotropy, and off-target invariance are combined as in Equation (40), using 4000 Adam updates of batch size 512. Singleton, null-balanced pair, nested, and identical-signature designs share the architecture and training budget; a singleton no-penalty arm isolates the invariance term. The singleton and pair designs also match marginal target frequencies and expected target count. Appendix D.1 gives all settings.

After training, the fixed map $G_j = \tanh(F_j)$ is used for both action energies and oracle variation. Its half squared increments lie in $[0, 2]$, and each output variance is at most one. These are support bounds, not fitted sample maxima or assumed unit variances. Independent validation applies Corollary 4 with $\delta = 0.05$ simultaneously over all tested action-output cells. The oracle estimate of $\mathcal{R}_j$ refreshes every factor except j; its partial counterpart instead holds C_{D_j} fixed. Neither oracle enters the certificate or the choice of weights.

The separating encoders receive nonvacuous same-scale bounds (Table 1). For pair-mask training, the oracle leakage is $(0.848 \pm 0.088) \times 10^{-3}$. Changing only the concentration bound reduces the certificate from $(1.689 \pm 0.093) \times 10^{-3}$ to $(1.296 \pm 0.091) \times 10^{-3}$, with mean per-seed certificate/error ratio 1.55 ± 0.05. The singleton certificate decreases from $(2.013 \pm 0.062) \times 10^{-3}$ to $(1.246 \pm 0.061) \times 10^{-3}$, but its ratio remains 2.54 ± 0.16. At balanced counts the LP and aggregate bound coincide for these designs; the improvement is statistical, not an additional geometric advantage. Additional visual results are provided in Appendix G.

Nonseparating designs require a different conclusion. Nested encoders retain full leakage 1.006 ± 0.036 but have allowed-class residual $(0.626 \pm 0.111) \times 10^{-3}$, bounded by EB at $(0.930 \pm 0.112) \times 10^{-3}$ and by PrPl at $(0.751 \pm 0.110) \times 10^{-3}$. The routine reports full separation as unavailable instead of relabeling the small partial bound as full recovery. Identical signatures exhibit the same distinction. Conversely, the no-penalty arm has leakage 1.050 ± 0.094 and PrPl bound 1.102 ± 0.097: the procedure does not conceal a mixed representation behind a small certificate. Mean output variances range from 0.345 to 0.458 across conditions, and the singleton and pair-mask arms have mean held-out linear-probe R^2 of 0.977 and 0.970. These diagnostics do not prove injectivity, but rule out constant outputs in the evaluated runs.

6.2 When only overlapping masks can be tested

The next benchmark removes singleton measurements entirely. For $K \in \{4, 10, 25\}$, partition the $K - 1$ forbidden factors of the last output into triads. Each triad supplies its three pair-refresh masks. A normalized nonlinear triangular shear has linear and pair-interaction terms with exact reference residual and variance (Appendix D.4). We compare the fixed equal-weight block bound, the best integral cover, and the fractional LP using identical empirical Bernstein upper bounds. The integral

Table 2: Overlapping-mask-only benchmark: certificate divided by exact residual, mean $\pm$ SE over 100 validation repetitions. Each independent triad supplies three pair masks and no singleton tests. Two masks receive 32,768 pairs; the third receives n_{third}. Every method uses the same simultaneous EB intervals and unit variance cap. The best integral cover is solved exactly.

K	n_{third}	Fixed-weight bound	Best integral cover	Fractional LP
4	32,768	1.0531 ± 0.0004	1.3991 ± 0.0006	1.0531 ± 0.0004
4	32	4.5133 ± 0.0178	1.4032 ± 0.0008	1.4032 ± 0.0008
10	32,768	1.0585 ± 0.0003	1.4065 ± 0.0004	1.0585 ± 0.0003
10	32	5.2696 ± 0.0089	1.4112 ± 0.0005	1.4112 ± 0.0005
25	32,768	1.0629 ± 0.0002	1.4125 ± 0.0002	1.0629 ± 0.0002
25	32	5.9563 ± 0.0057	1.4172 ± 0.0003	1.4172 ± 0.0003

comparator selects the best subset of actions whose union covers every forbidden factor, so it already exploits heterogeneous test quality.

The balanced and scarce-test regimes distinguish two sources of improvement (Table 2). With balanced counts at $K = 25$, the fractional certificate/error ratio is 1.063, versus 1.412 for the best integral cover; fractional weights avoid counting shared factors twice. When one test per triad has only 32 pairs, the fixed-weight ratio rises to 5.956, whereas the optimized cover remains at 1.417. In this regime the best integral and fractional covers agree because both omit the imprecise test. All corrected procedures contain the exact residual in all 100 repetitions per setting. The uncorrected plug-in LP contains it in only $57/100$ repetitions in the four-factor scarce-test case. The guarantees come from simultaneous concentration, not from these empirical containment counts.

The five trained pair-mask encoders show the same distinction on preassigned output four. With only the three triangle masks, balanced counts of 32,768 pairs give EB and PrPl fractional certificates of 0.001270 ± 0.000049 and 0.000661 ± 0.000047, against oracle leakage 0.000207 ± 0.000042. The PrPl integral cover is 0.000854 ± 0.000057. With only 32 pairs for the third action, the optimized PrPl cover omits it and gives 0.000874 ± 0.000051, versus EB's 0.001686 ± 0.000054. These values average 20 validation repetitions within each training seed before computing the five-seed uncertainty. The range correction remains material, but improving concentration reduces the balanced certificate by 48% without changing the readout Appendix D (Appendix D.5).

6.3 Published visual generator and independent learners

The Pendulum generator released with CausalVAE provides a visual evaluation outside our signature simulator (Yang et al., 2021). Its pendulum and illumination angles are independent roots; shadow position and length are their descendants. We regenerate the public angle grid and evaluate the uniform product of its 84×87 root configurations, rendering the shadows after each root change. Images are downsampled from 96×96 to 32×32. This is a two-root product-reference task, not independent intervention on all four endogenous concepts.

An autoencoder (AE), a variational autoencoder (VAE), and a β-VAE with $\beta = 4$ (Kingma & Welling, 2014; Higgins et al., 2017) are each fitted with three seeds, identical hidden widths, and 2000 updates. They receive images only. A separate pilot fixes a one-to-one factor assignment, after which certification uses the raw deterministic encoder means, without $\tanh$ or variance normalization. Complete enumeration of the 7308 frozen outputs supplies the exact reference errors, output-variance caps, and deterministic fiber envelopes. This audit costs 7308 encoder evaluations per model in addition to 4096 pilot lookups and the reported validation-pair lookups. It establishes applicability on a published visual generator, not savings over exhaustive evaluation or use of an externally pretrained causal model.

At 2048 pairs per action, predictable concentration and fiber envelopes reduce the mean ratios from 1.251 to 1.080 for AE, 1.668 to 1.182 for VAE, and 8.344 to 1.139 for β-VAE (Table 3). The two changes are complementary. At 32,768 pairs, the corresponding ratios are 1.021, 1.033, and 1.031. Both output errors are contained in $300/300$ repetitions for each learner at the smaller budget; at the larger budget, PrPl with fiber envelopes contains $297/300$ for VAE and $300/300$ for the other learners. These are repeated finite-sample outcomes, not a guarantee of perfect coverage. With two roots,

Table 3: Pendulum root-grid evaluation. Certificate/exact-error ratios use raw encoder means and known finite-reference variance caps. Entries are mean $\pm$ SE across three training seeds, after averaging 100 validation repetitions per seed. Each action receives n_a pairs; total sampled pairs are $2n_a$. Global and fiber columns differ only in the deterministic envelope. Enumeration cost is reported separately in the text.

Learner	n_a	EB, global	PrPl, global	EB, fiber	PrPl, fiber
AE	$2,048$	1.251 ± 0.061	1.136 ± 0.030	1.101 ± 0.024	1.080 ± 0.015
AE	$32,768$	1.034 ± 0.003	1.025 ± 0.002	1.024 ± 0.002	1.021 ± 0.001
VAE	$2,048$	1.668 ± 0.400	1.310 ± 0.143	1.241 ± 0.092	1.182 ± 0.070
VAE	$32,768$	1.067 ± 0.032	1.047 ± 0.020	1.041 ± 0.013	1.033 ± 0.010
β-VAE	$2,048$	8.344 ± 5.425	3.987 ± 2.189	1.188 ± 0.024	1.139 ± 0.017
β-VAE	$32,768$	1.483 ± 0.341	1.198 ± 0.133	1.036 ± 0.004	1.031 ± 0.004

the LP reduces to the singleton certificate: this experiment tests transfer across learners, raw-output validation, and concentration, rather than a fractional-cover gain. IRS, DCI, MIG, reconstruction error, and every seed's raw leakage are reported together in Appendix E (Appendix E.4); certificate tightness is not a ranking of representation quality.

6.4 Rare separation and failures of the sampling assumptions

A learned orthogonal-after-inverse-nonlinearity control isolates rare actions without uncertainty about information retention. Every output has unit variance and the inverse condition number is one. Across 12 paired seeds, reducing the separation margin from 0.10 to 0.01 increases exact leakage from 0.023 ± 0.012 to 2.814 ± 0.176, although every factor keeps the same marginal intervention frequency. The exact identity $\mathcal{L}_{\mathrm{int}} = 2q_* \sum_j \mathcal{R}_j$ holds for each fitted encoder (Appendix C.4). Separate singleton-versus-LP and pilot-allocation comparisons are retained in Appendix C, specifically in Appendix C.5; they distinguish geometric tightness from validation efficiency.

A localized invertible shear tests state overlap. Changing only the starting-state law gives the known ratio $\kappa = 1/s$; at the support hole $s = 0$, replay loss is exactly zero despite reference leakage 0.015748. The multiplicative transfer certificate is unavailable. A separate drift-cancellation control keeps true leakage at $1/2$ while driving the naive resampling bound to zero; its coupling-corrected population bound remains $1/2$ (Appendices C.7 and C.8). Neither action counts nor low loss verifies these sampling premises. The retained RGB failures delimit visual-learning claims separately in Appendix F.

7 Discussion and conclusion

Intervention losses become interpretable only after specifying the tested dependencies, reference pairs, and output scale. The exact upper-information envelope separates geometric ambiguity from finite-sample uncertainty: a fractional cover extracts everything implied by the stated energy upper bounds, while predictable concentration and verified fiber ranges can tighten those bounds. The neural experiments demonstrate both mechanisms, and the Pendulum root-grid study extends evaluation to image-only learners without a saturating readout. These results do not make the certificate a substitute for IRS, DCI, MIG, reconstruction, or inverse conditioning. Known masks, a product reference experiment, independent validation, and justified transfer premises remain essential; exhaustive finite-grid range verification also carries a cost that does not scale automatically to large simulators. Neither the two-root benchmark nor the retained RGB failures establish identification of general causally dependent visual factors. The operational conclusion is to report a confidence bound for the stated ambiguity class, preserve an unavailable full-separation outcome when necessary, and evaluate information retention separately.

Reproducibility statement

Appendices A–B give the proofs and complete statistical rules. The supplementary package contains the 25 signature checkpoints, nine Pendulum checkpoints, public-generator adaptation, finite-grid

outputs, independent validation records, and seed-level summaries. Appendices D.1 and E specify every training and validation configuration. The RGB scope measurements in Appendix F remain author-provided records; their original checkpoints are not included.

AI USE STATEMENT

Generative AI tools assisted with literature retrieval, development and auditing of mathematical claims and proofs, experimental design, synthetic-data generation, implementation and execution of the controlled and neural validation experiments, analysis of their outputs, and manuscript and reference preparation. The supplementary records expose the executed code and numerical checks. The retained RGB measurements were provided by the authors. The authors retain responsibility for verifying the final submission, its empirical records, and all scientific claims.

REFERENCES

Raghavendra Addanki, Shiva Kasiviswanathan, Andrew McGregor, and Cameron Musco. Efficient intervention design for causal discovery with latents. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 63–73. PMLR, 2020. URL https://proceedings.mlr.press/v119/addanki20a.html.

Pietro Caputo and Daniel Parisi. Block factorization of the relative entropy via spatial mixing. *Communications in Mathematical Physics*, 388:793–818, 2021. doi: 10.1007/s00220-021-04237-1. URL https://arxiv.org/abs/2004.10574.

Pietro Caputo, Zongchen Chen, and Daniel Parisi. Factorizations of relative entropy using stochastic localization. arXiv preprint arXiv:2503.19419, 2025. URL https://arxiv.org/abs/2503.19419.

Ricky T. Q. Chen, Xuechen Li, Roger B. Grosse, and David K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/1ee3dfcd8a0645a25a35977997223d22-Abstract.html.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=By-7dz-AZ.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596, 1981. doi: 10.1214/aos/1176345462. URL https://projecteuclid.org/journals/annals-of-statistics/volume-9/issue-3/The-Jackknife-Estimate-of-Variance/10.1214/aos/1176345462.full.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830.

Huawei Technologies Co., Ltd. CausalVAE pendulum data generator. GitHub repository, 2021. URL https://github.com/huawei-noah/trustworthyAI/blob/master/research/CausalVAE/causal_data/pendulum.py. Source file: causal_data/pendulum.py; MIT License.

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *AISTATS*, 2020.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177, pp. 428–484. PMLR, 2022. URL https://proceedings.mlr.press/v177/lachapelle22a.html.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *Journal of Machine Learning Research*, 27(71):1–90, 2026a. URL https://jmlr.org/papers/v27/24-0771.html.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *Journal of Machine Learning Research*, 27(71):1–90, 2026b. URL https://jmlr.org/papers/v27/24-0771.html.

P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *ICML*, 2022.

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*, 2023. URL https://arxiv.org/abs/2206.06169.

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://arxiv.org/abs/0907.3740.

Art B. Owen. Variance components and generalized Sobol' indices. arXiv preprint arXiv:1205.1774, 2012. URL https://arxiv.org/abs/1205.1774.

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

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009. URL https://doi.org/10.1093/jrsssb/qkad009.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yang_CausalVAE_Disentangled_Representation_Learning_via_Neural_Structural_Causal_Models_CVPR_2021_paper.html.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/85381f4549b5ddf1d48e2e287d7d3d15-Abstract-Conference.html.

Xiangteng Zhang, Yang Guan, Bo Zhang, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models. arXiv:2607.22430v1, 2026. URL https://arxiv.org/abs/2607.22430v1.

A COMPLETE PROOFS

The population results concern the same output assignment and the same frozen map in the loss and the protected residual. Exact identification uses the smoothness and support assumptions in Section 3. The finite-error inequalities require only square integrability under the declared product reference law, together with the relevant overlap or coupling premise.

A.1 SIGNATURE CHARACTERIZATION

Proof of Theorem 1. The intervention loss is an expectation of nonnegative terms. If it vanishes, then for every used action a and every $j \notin T_a$, $F_j(c') = F_j(c)$ almost surely under the corresponding pair law. Positive state density, positive action propensity, and a positive conditional density throughout the target fiber give full support for (c, c'_{T_a}) on $\Omega \times \prod_{i \in T_a} \Omega_i$. The difference is continuous. If it were nonzero at any point, it would remain bounded away from zero on a neighborhood of positive pair probability, contradicting zero expected loss. Therefore F_j is constant on every such fiber.

Differentiating along $i \in T_a$ yields $\partial_i F_j = 0$ whenever some action targets i but not j. Such an action exists exactly when $S_i \not\subseteq S_j$. Conversely, Equation (3) sets all these derivatives to zero. Any two points on a target fiber can be joined by changing its coordinates one at a time, because every factor support is an interval. Integrating the zero derivatives along those segments gives invariance on the fiber and hence zero loss.

Under full separation, every off-diagonal derivative vanishes. Connectedness implies $F_j(c) = \psi_j(c_j)$. Since F is a diffeomorphism, its diagonal Jacobian has no zero diagonal entry. The continuous derivative ψ'_j has a constant nonzero sign on its interval, so ψ_j is injective and has a C^1 inverse on its image.

Finally, if $S_i \subseteq S_j$, an action leaving j untargeted must also leave i untargeted. Thus the shear $F_j = c_j + \varepsilon c_i$, with the other coordinates unchanged, satisfies every off-target constraint. Its determinant is one and its inverse replaces c_j by $z_j - \varepsilon z_i$. It is a diffeomorphism onto its image but is not coordinatewise when $\varepsilon \neq 0$. □

A.2 PRODUCT DECOMPOSITION AND THE COMPLETE REFRESH SPECTRUM

For each coordinate, let P_i integrate over Q_i while leaving the other coordinates fixed. The P_i are commuting orthogonal projections on $L^2(Q)$. For $U \subseteq [K]$, define

$$F_{j,U} = \left[\prod_{i \in U} (I - P_i)\right] \left[\prod_{i \notin U} P_i\right] F_j. \tag{20}$$

Expanding $\prod_i [P_i + (I - P_i)]$ gives $F_j = \sum_U F_{j,U}$. Distinct components are orthogonal, and each nonconstant component integrates to zero in any of its participating coordinates. Consequently,

$$\mathcal{R}_j = \sum_{U \not\subseteq \{j\}} \|F_{j,U}\|_2^2, \qquad \mathcal{R}_j^{\text{allow}} = \sum_{U \not\subseteq D_j} \|F_{j,U}\|_2^2. \tag{21}$$

Independent refresh on T_a is the projection $P_{T_a} = \prod_{i \in T_a} P_i$. It preserves $F_{j,U}$ if $U \cap T_a = \varnothing$ and annihilates it otherwise. Both marginals of Ξ_a equal Q, so

$$\begin{aligned} \frac{1}{2}\mathbb{E}_{\Xi_a}(F_j(C') - F_j(C))^2 &= \langle F_j, (I - P_{T_a}) F_j \rangle_{L^2(Q)} \\ &= \sum_{U : U \cap T_a \neq \varnothing} \|F_{j,U}\|_2^2. \end{aligned} \tag{22}$$

This proves Equation (9). It is the product-space variance decomposition underlying the weighted block-factorization inequality (Efron & Stein, 1981; Caputo & Parisi, 2021).

Define the masked refresh operator and its component weights by

$$A_j = \sum_{a : j \notin T_a} p_a (I - P_{T_a}), \qquad s_{j,U} = \sum_{a : j \notin T_a} p_a \mathbf{1}\{U \cap T_a \neq \varnothing\}. \tag{23}$$

Then $A_j F_{j,U} = s_{j,U} F_{j,U}$. Since used action probabilities are positive, $s_{j,U} = 0$ exactly when every $i \in U$ has $S_i \subseteq S_j$, or equivalently $U \subseteq D_j$. The null space therefore consists precisely of functions of C_{D_j}. If $I_j \neq \varnothing$, every forbidden component contains some $i \in I_j$ and has $s_{j,U} \geq q_{ij} \geq r_j$. A centered nonconstant function of a factor attaining r_j realizes the eigenvalue r_j. Thus r_j is the exact smallest positive eigenvalue for nondegenerate reference factors.

A.3 Aggregate recovery and partial identification

Proof of Theorem 2. For $U \not\subseteq \{j\}$, choose $i \in U \setminus \{j\}$. Equation (23) gives $s_{j,U} \geq q_{ij} \geq q_j$. Orthogonality therefore implies

$$2q_j \mathcal{R}_j \leq \sum_{a: j \notin T_a} p_a \mathbb{E}_{\Xi_a}(F_j(C') - F_j(C))^2. \tag{24}$$

The squared increments are nonnegative. Applying $p_a \Xi_a \leq \kappa B_a$, then summing over j, bounds the right-hand side by $\kappa \mathcal{L}_{\mathrm{int}}(F; B)$. This proves Equation (6); replacing each q_j by q_* gives the unweighted bound.

For deployment, fix $\psi_j(c_j) = \mathbb{E}_Q[F_j \mid C_j = c_j]$. Since $\nu \leq wQ$, integration of the same nonnegative residual gives

$$\sum_j \mathbb{E}_\nu(F_j - \psi_j)^2 \leq w \sum_j \mathbb{E}_Q(F_j - \psi_j)^2.$$

The theorem's deployment statement follows. It does not replace ψ_j by a separately fitted function or a conditional expectation under ν.

For sharpness, take $K = 2$, independent centered unit-variance factors, singleton probabilities p and $1-p$ with $p \leq 1/2$, and $F_1 = C_1$, $F_2 = C_2 + \varepsilon C_1$. The only nonzero off-factor residual is $\mathcal{R}_2 = \varepsilon^2$. Refreshing factor one changes F_2 by $\varepsilon(C_1' - C_1)$, with expected square $2\varepsilon^2$. Thus $\mathcal{L}_{\mathrm{int}} = 2p\varepsilon^2$ and $q_* = p$, attaining equality. Independent standard Gaussian factors supply the smooth diffeomorphic example stated in the Gaussian formulation. □

Proposition 5 (Stability of the admissible ambiguity class)**.** *With D_j, I_j, and r_j as defined in Section 4, Equation* (5) *implies*

$$\sum_j r_j \mathcal{R}_j^{\mathrm{allow}} \leq \frac{\kappa}{2} \mathcal{L}_{\mathrm{int}}(F; B). \tag{25}$$

For $I_j = \varnothing$, $\mathcal{R}_j^{\mathrm{allow}} = 0$ and the convention $r_j = 0$ applies.

Proof. Every component contributing to $\mathcal{R}_j^{\mathrm{allow}}$ contains an index $i \notin D_j$, so its reference eigenvalue is at least $q_{ij} \geq r_j$. Sum its variance contribution as in Equation (24), then apply pair-law domination. Under full separation $D_j = \{j\}$ and the statement reduces to Theorem 2. □

A.4 Optimality of the action-resolved certificate

Proof of Theorem 3. Fix output j and suppress that index. If I_j is empty, the residual is identically zero, and all claims follow. Otherwise write $v_U = \|G_U\|_2^2$. For each $U \not\subseteq D_j$, choose one representative $i(U) \in U \cap I_j$ and define

$$x_i = \sum_{U \not\subseteq D_j : i(U) = i} v_U.$$

The total mass is $\sum_i x_i = \mathcal{R}_j^{\mathrm{allow}}(G)$. For every $a \in \mathcal{A}_j$,

$$\sum_{i \in I_j \cap T_a} x_i = \sum_{U \not\subseteq D_j} \mathbf{1}\{i(U) \in T_a\} v_U \leq \sum_{U : U \cap T_a \neq \varnothing} v_U = e_{aj}(G) \leq u_{aj}. \tag{26}$$

Thus x is feasible for the packing program, proving the upper bound without any restriction on interaction order. Components indexed by subsets of D_j contribute no energy under actions leaving j untargeted, because such actions cannot touch a coordinate in D_j.

Conversely, for each nondegenerate Q_i choose a centered function $g_i \in L^2(Q_i)$ with $\mathbb{E}g_i^2 = 1$. For any feasible packing vector x, let

$$G(C) = \sum_{i \in I_j} \sqrt{x_i}\, g_i(C_i).$$

Independence gives $\mathcal{R}_j^{\mathrm{allow}}(G) = \sum_i x_i$ and $e_{aj}(G) = \sum_{i \in I_j \cap T_a} x_i \le u_{aj}$. Therefore every feasible packing value is attained by a function in the stated information class. Each forbidden factor is covered by a used action, so finite upper bounds make the packing feasible and bounded. Linear-programming duality gives the cover formulation in Equation (10), and the packing maximum is attained. If only a subset of actions is available and a forbidden index is uncovered, its mass can increase without bound; both the information supremum and the uncapped certificate are infinite.

If $\mathrm{Var}_Q(G) \le v_j$ is additionally imposed, then $\mathcal{R}_j^{\mathrm{allow}}(G) \le v_j$. Take an optimal packing vector and scale it down, when necessary, to have total mass $\min\{v_j, \mathcal{C}_j(u)\}$. The additive construction attains this value with that same variance, proving the capped statement. If exact variance v_j rather than an upper bound is desired, an independent centered allowed component of the remaining variance can be added without changing any off-target energy.

For the coordinatewise invertible construction, assume the factors have finite positive variances and set $g_i(c_i) = (c_i - \mathbb{E}C_i)/\sqrt{\mathrm{Var}(C_i)}$. Replace output j of the identity map by

$$F_j(c) = c_j + \sum_{i \in I_j} \sqrt{x_i}\, g_i(c_i).$$

Since $j \in D_j$, the summation excludes j. This is an affine shear with determinant one and an explicit inverse. It realizes the uncapped optimum for output j. Other output coordinates remain unchanged. No assertion of simultaneous attainability of separately maximized coordinates is needed. □

The result does not identify the actual ANOVA spectrum. Different interaction spectra can satisfy the same energy inequalities, and the extremizer in the proof need not resemble the fitted encoder. The certificate is optimal for the stated information class. When all available energies are given as exact equalities, rather than upper bounds, their additional consistency constraints can lead to a smaller maximum; that is a different optimization problem.

For a numerical solver, a primal packing value is a lower bound on the worst compatible error, not by itself a certified upper bound. A feasible dual vector supplies the upper certificate. Our code checks dual coverage. More generally, if a nonnegative approximate dual vector has minimum coverage $\gamma > 0$, replacing it by $y/\min\{1, \gamma\}$ restores exact coverage in real arithmetic. Validated numerical rounding or an explicit feasibility tolerance must be included when machine-level rigor is required. A solver's success flag alone is not a statistical or numerical certificate.

A.5 What action counts cannot establish

Proposition 6 (Separation and pair coverage are distinct)**.** *Positive marginal intervention frequency for every factor implies neither $q_* > 0$ nor finite pair-law domination. A reference-positive, replay-null pair region rules out a uniform finite multiplicative transfer inequality for unrestricted nonnegative pair residuals.*

Proof. An all-target action intervenes on every factor but leaves no output untargeted, so all q_{ij} vanish. Separately, a behavior policy can never execute a distinguishing action in a reference-positive state region, even though its marginal frequency is positive elsewhere. The corresponding reference pair region has positive mass and its replay counterpart has zero mass. Its indicator has zero replay expectation and positive reference expectation, disproving a uniform multiplicative bound for arbitrary nonnegative residuals. This last argument concerns the full residual class; the localized encoder construction in Appendix C.7 shows a concrete off-target representation loss with the same failure. □

A disconnected-support example isolates intervention connectivity. Let $C_1 \sim \mathcal{N}(0, 1)$ and independently let C_2 have equal mass on $(-2, -1)$ and $(1, 2)$. One singleton action refreshes C_1; another

refreshes C_2 only within its current interval. For a smooth χ equal to -1 on the first interval and 1 on the second, the invertible map $F_1 = C_1 + \varepsilon\chi(C_2)$, $F_2 = C_2$ has zero observed off-target loss and reference variation $\mathcal{R}_1 = \varepsilon^2$. Independent reference refreshes cross the intervals, whereas the collected pairs never do. Positive action counts therefore do not imply the intervention support needed by the theorem.

Action-specific subprobability overlap can be stated without a common κ. If $B_a \geq \beta_a \Xi_a$, define

$$\widetilde{q}_j = \min_{i \neq j} \sum_a \beta_a \mathbf{1}\{i \in T_a, j \notin T_a\}.$$

The same decomposition gives

$$\sum_j \widetilde{q}_j \mathcal{R}_j \leq \frac{1}{2}\mathcal{L}_{\mathrm{int}}(F; B). \tag{27}$$

The weights need not sum to one. Taking $\beta_a = p_a/\kappa$ recovers the aggregate theorem. Reweighting an existing buffer changes an estimator; it cannot create physical pairs absent from the buffer's support.

A.6 Stability under coupled intervention error

Proposition 7 (Pooled coupling bound)**.** *Let $R(dc, a, dc') = p_a \Xi_a(dc, dc')$ and $D_F(c, a, c') = \Pi_{-T_a}[F(c') - F(c)]$. Suppose a probability law $\widetilde{R} \leq \kappa B$ admits an action-preserving coupling Γ with R. Define*

$$\delta_F^2 = \mathbb{E}_\Gamma \left\| D_F(C, A, C') - D_F(\widetilde{C}, A, \widetilde{C}') \right\|^2. \tag{28}$$

Then Equation (15) *holds, as does its version with $q_j\mathcal{R}_j$ replaced by $r_j\mathcal{R}_j^{\mathrm{allow}}$. If every $\Pi_{-T_a}F$ is L-Lipschitz, then $\delta_F \leq L\Delta$, where*

$$\Delta^2 = \mathbb{E}_\Gamma\big(\left\|C - \widetilde{C}\right\| + \left\|C' - \widetilde{C}'\right\|\big)^2.$$

Proof. Before transferring to replay, the reference inequality gives $2\sum_j q_j\mathcal{R}_j \leq \mathbb{E}_R \|D_F\|^2$. Minkowski's inequality in $L^2(\Gamma)$ and domination give

$$\big(\mathbb{E}_R \|D_F\|^2\big)^{1/2} \leq \big(\mathbb{E}_{\widetilde{R}} \|D_F\|^2\big)^{1/2} + \delta_F \leq \sqrt{\kappa\mathcal{L}_{\mathrm{int}}(F; B)} + \delta_F.$$

Squaring proves the result. The reference partial-identification inequality gives the same proof for $\mathcal{R}_j^{\mathrm{allow}}$. The action-preserving coupling and the projected Lipschitz bound imply pointwise

$$\|D_F(c, a, c') - D_F(\widetilde{c}, a, \widetilde{c}')\| \leq L\big(\|c - \widetilde{c}\| + \|c' - \widetilde{c}'\|\big).$$

Taking squared expectations proves the discrepancy bound. Restricting the argument to one action and one output, with the half-energy convention, proves Equation (14). □

If $q_* > 0$, the pooled result gives

$$\sum_j \mathcal{R}_j \leq \frac{(\sqrt{\kappa\mathcal{L}_{\mathrm{int}}} + \delta_F)^2}{2q_*}.$$

A deployment law $\nu \leq wQ$ multiplies the bound on the same fixed projection residuals by w. Nonzero coupling discrepancy leaves a nonzero error floor even at zero observed loss.

For an illustrative autoregressive drift protocol, couple initial states and targeted draws exactly. If an untargeted factor evolves as $\widetilde{C}_i' = \rho_i C_i + \sigma_i\epsilon_i$, while the reference keeps $C_i' = C_i$, and the noises are independent, centered, and unit variance, then

$$\Delta^2 = \sum_a p_a \sum_{i \notin T_a} \big[(\rho_i - 1)^2 \mathbb{E}_Q C_i^2 + \sigma_i^2\big].$$

For four unit-second-moment factors, singleton targets, $\rho_i = 0.95$, and $\sigma_i = 0.08$, this equals 0.0267, giving $\Delta \approx 0.1634$. This is a protocol-level calculation. Auxiliary-law overlap and a representation Lipschitz bound are still required before using it as an encoder-error guarantee.

B FINITE-SAMPLE GUARANTEES AND DESIGN PROOFS

B.1 INDEPENDENT VALIDATION

Proof of Corollary 4. Condition on the fitted encoder, assignment, normalization, and any independent design data. For a tested cell (a, j), the bounded empirical Bernstein inequality of Maurer & Pontil (2009), scaled to $[0, M_{aj}]$, gives

$$\mathbb{E}Y_{aj} \le \bar{Y}_{aj} + \sqrt{\frac{2\widehat{v}_{aj}\log(2m/\delta)}{n_a}} + \frac{7M_{aj}\log(2m/\delta)}{3(n_a-1)}$$

with failure probability at most δ/m. The population mean is also at most M_{aj}, so clipping the upper bound at that envelope preserves validity. A union bound over all tested cells gives simultaneous validity with probability at least $1-\delta$; it does not require independence between different cells.

On this event, every energy upper bound supplied to the LP is valid. Multiplication by a justified overlap factor and the monotone map in Equation (14) preserve the event. Theorem 3 then holds for every feasible dual weight vector, including one selected by optimizing these same upper bounds. No additional union bound over weight vectors is needed. This statement does not extend to an encoder fitted using the validation outcomes, because the cellwise concentration step conditioned on an independently frozen encoder. □

The unbiased variance estimator is

$$\widehat{v}_{aj} = \frac{1}{n_a-1}\sum_{r=1}^{n_a}(Y_{aj,r} - \bar{Y}_{aj})^2.$$

An independent simulator can collect a fixed number of pairs for each action. With a fixed iid replay sample and random action counts, the argument may instead condition on the counts when the conditional pairs remain iid; actions with fewer than two samples need a valid envelope-only fallback. Time-correlated trajectory transitions are not iid validation units without further assumptions.

Corollary 8 (Pooled Hoeffding comparison)**.** *For a frozen encoder and n independent validation pairs from B, suppose the full per-pair off-target loss lies in $[0, M]$. If $q_* > 0$ and κ is a valid overlap bound, then with probability at least $1-\delta$,*

$$\sum_j \mathcal{R}_j \le \frac{\kappa}{2q_*}\left[\widehat{\mathcal{L}_{\text{int}}} + M\sqrt{\frac{\log(1/\delta)}{2n}}\right]. \tag{29}$$

Proof. Hoeffding's inequality (Hoeffding, 1963) upper-bounds the population loss by the bracketed term. Apply Theorem 2. The same fixed residual functions transfer to $\nu \le wQ$ with an additional factor w. □

For the bounded two-factor shear $F_2 = (C_2 + C_1)/\sqrt{2}$ with independent uniform factors on $[-\sqrt{3}, \sqrt{3}]$, the full per-pair envelope is $M = 6$. At $n = 200{,}000$ and $\delta = 0.05$, the additive loss correction is approximately 0.01642, before division by $2q_*$. This illustrates why pooled worst-range corrections can dominate a small measured loss. A learned encoder must supply its own valid envelope.

B.2 PREDICTABLE VARIANCE-ADAPTIVE VALIDATION

Proposition 9 (Predictable upper bounds and optimized covers)**.** *Fix the encoder, assignment, support envelopes, and target counts independently of validation. Under the conditional iid sampling assumptions of Corollary 4, Equation* (18) *has simultaneous upper coverage at least $1-\delta$ over the m tested cells. Substituting these bounds into Equation* (10)*, with any justified reference-transfer correction and reference-variance cap, preserves that confidence level for all protected output residuals.*

Proof. For one cell with $M > 0$, put $\mu = \mathbb{E}X_t$. Predictability follows from Equation (17): $\widehat{\mu}_0 = 1/2$, $\widehat{\sigma}_0^2 = 1/4$, and $\lambda_t \in (0, 1)$ uses no current or future observation. Apply the predictable empirical Bernstein supermartingale of Waudby-Smith & Ramdas (2024, Theorem 2) to $1 - X_t$ with predictor $1 - \widehat{\mu}_{t-1}$. In the notation of that result, the factor four in its squared prediction error cancels the factor $1/4$ in its cumulant bound. Thus

$$Z_t(\mu) = \exp\left\{\sum_{s=1}^{t} \lambda_s(\mu - X_s) - \sum_{s=1}^{t} \psi(\lambda_s)(X_s - \widehat{\mu}_{s-1})^2\right\} \tag{30}$$

is a nonnegative supermartingale starting at one. Consequently $\Pr\{Z_N(\mu) > 1/\eta\} \leq \eta$. If μ exceeds the untruncated ratio in Equation (18), this event occurs. Since $\mu \leq 1$, truncation at one also preserves validity. Multiplication by M returns the cell's original half-energy units. A zero envelope implies a known zero energy and needs no samples.

A union bound with $\eta = \delta/m$ makes all cell bounds valid together, regardless of dependence between output coordinates evaluated on the same pair. On that event, every feasible fractional cover bounds the residual. Selection of the cover using these bounds is therefore permitted. Known overlap and coupling corrections are monotone, and a valid reference-variance cap bounds the same residual independently, completing the claim. □

The target N is fixed before sampling. The construction is valid at any time for this one predictable rule, but retuning N separately at several reported budgets does not by itself provide simultaneous coverage over those different rules. We report a separate fixed-budget guarantee in each row. The constant 0.9 is fixed for every learner, seed, action, and sample size; it is not selected on the validation errors. Original EB and PrPl are alternative procedures at the same nominal confidence level. Their observed minimum is not reported as a confidence bound. A procedure that selects between them can allocate $\delta/2$ to each before taking the minimum, but that is not the numerical comparison here.

B.3 Deterministic fiber envelopes and finite supports

For a fixed untargeted state, two independently refreshed endpoints both lie in the same target fiber. Their absolute output difference cannot exceed that fiber's oscillation, proving Equation (19). Replacing the fiber oscillation by the full output range gives the coarser envelope $M_j^{\mathrm{glob}} = \frac{1}{2}(\sup F_j - \inf F_j)^2$. This replacement can be loose because changes in the output's own assigned factor contribute to the global range but cancel in an off-target pair. Neither envelope uses a distributional tail estimate.

For a finite product grid, every fiber minimum and maximum can be obtained from the complete frozen output table. This gives an envelope for the actual deterministic numerical map on that table, not a bound inferred from the most extreme sampled validation increment. The reference mean, variance, and conditional variances follow from finite sums. Smoothness and observation injectivity are unnecessary for these finite-error conclusions. In larger spaces, exhaustive certification is not automatic: a verified analytic or numerical upper bound on the fiber oscillation would be needed instead. The experiments do not claim that a sampled maximum provides that replacement.

B.4 Why a range term cannot disappear from energy-only inference

The small neural losses make a rare-event limitation relevant. The following elementary construction concerns information in the observed half-energies, rather than procedures that inspect the entire encoder or reference support.

Proposition 10 (Zero-observation lower bound)**.** *Let $M > 0$ and $\eta \in (0, 1)$. Any deterministic upper-confidence rule $V(Y_1, \ldots, Y_n; M)$ based only on iid half-energy observations and the common information $0 \leq Y \leq M$, $\mathrm{Var}_Q(F_j) \leq M/2$, and valid with probability at least $1 - \eta$ for every bounded product-refresh encoder in this information class, must satisfy*

$$V(0, \ldots, 0; M) \geq M \min\left\{\tfrac{1}{2},\, 1 - \eta^{1/n}\right\}. \tag{31}$$

Proof. Take independent $C_1, C_2 \sim U[0,1]$, assign the second output to C_2, and set $F_2(C) = \sqrt{2M}(\mathbf{1}\{C_1 \le \rho\} - 1/2)$, with $0 \le \rho \le 1/2$. Under an independent refresh of C_1, its half squared increment equals M with probability $p = 2\rho(1-\rho)$ and zero otherwise. Every $p \in [0, 1/2]$ is attainable. Both its mean half-energy and off-factor variation equal Mp, while its variance is at most $M/2$ and its output range is the same for every $0 < \rho \le 1/2$.

Suppose the left side of Equation (31) were smaller than its right side. Choose p strictly between that value divided by M and $\min\{1/2, 1 - \eta^{1/n}\}$. The event of n zero observations has probability $(1-p)^n > \eta$, and the stated rule undercovers Mp on that event, contradicting uniform validity. □

For fixed η and large n, $1 - \eta^{1/n} \sim \log(1/\eta)/n$. Thus even a known output-variance cap does not eliminate all range-dependent uncertainty. The construction belongs to the square-integrable class of the finite-error theorems; it does not invoke the smooth-invertibility assumptions of exact identification. It motivates tighter justified envelopes and adaptive concentration, rather than removing a necessary correction by assumption.

B.5 Learning-design margin

Collection for learning and collection for validation have different objectives. With available masks and costs c_a, maximizing the worst separation margin solves

$$\max_{p,t} \; t \quad \text{subject to} \quad \sum_a p_a \mathbf{1}\{i \in T_a, j \notin T_a\} \ge t \; (i \ne j), \quad p \in \Delta, \quad \sum_a p_a c_a \le B_0. \tag{32}$$

For all masks of size at most $r \le \lfloor K/2 \rfloor$, the exact unconstrained-cost optimum is $r(K-r)/(K(K-1))$, attained by uniform size-r masks. This optimizes the reference margin, not a learned model's average error or the replay state distribution.

Validation can instead focus on an informative cover. If a dual vector y is fixed using an independent pilot, a Hoeffding certificate has uncertainty proportional to $\sum_a y_a M_{aj}/\sqrt{n_a}$. At fixed total validation count, its continuous minimizer satisfies

$$n_a \;\propto\; (y_a M_{aj})^{2/3}. \tag{33}$$

The selected positive-weight actions must still cover every forbidden factor. Fresh validation on this cover retains validity, and the pilot must be charged to the collection budget. The proofs below give the allocation formulas and integer-rounding rule. This is a validation-efficiency result, not an exploration or control-return guarantee.

Proposition 11 (Optimal reference separation design)**.** *For a fixed mask family, Equation* (32) *maximizes the worst ordered-pair margin under its declared expected-cost budget. If all masks of size at most r are available, where $1 \le r \le \lfloor K/2 \rfloor$, and no additional cost constraint is imposed, then*

$$q_*^{\mathrm{opt}} = \frac{r(K-r)}{K(K-1)}. \tag{34}$$

Uniform sampling over all size-r masks attains this value.

Proof. Introducing t gives the hypograph formulation of $\max_p \min_{i \ne j} q_{ij}$. All constraints, including expected cost, are linear. Moreover,

$$K(K-1)q_* \le \sum_{i \ne j} q_{ij} = \mathbb{E}_p[|T_A|(K - |T_A|)] \le r(K-r),$$

since $s(K-s)$ is increasing on $0 \le s \le K/2$. A uniform size-r mask contains i and excludes j with probability $\frac{r}{K}\frac{K-r}{K-1}$ for every ordered pair. This attains the upper bound. The comparison is per transition, not per factor intervened on; singleton and larger-block designs have different costs unless explicitly matched. □

B.6 Validation allocation and sparse covers

For a fixed output and independently chosen feasible y, let $w_a = y_a M_{aj}\sqrt{\log(m/\delta)/2}$. Conditional Hoeffding bounds give a certificate with uncertainty at most $\sum_a w_a n_a^{-1/2}$. Under a continuous budget $\sum_a c_a n_a = N$, positive weights and costs yield

$$n_a = \frac{N(w_a/c_a)^{2/3}}{\sum_b w_b^{2/3} c_b^{1/3}}, \qquad \min \sum_a \frac{w_a}{\sqrt{n_a}} = \frac{\left(\sum_a w_a^{2/3} c_a^{1/3}\right)^{3/2}}{\sqrt{N}}. \tag{35}$$

Differentiating the strictly convex objective in the positive n_a gives $-w_a n_a^{-3/2}/2 + \lambda c_a = 0$, from which both formulas follow. Zero-weight actions need not be sampled for that fixed certificate. If an output or encoder is selected using a pilot, the fresh validation guarantee is conditional on that pilot and the selection must not consume the validation outcomes.

In the numerical comparison, every selected action receives at least two samples. After reserving these samples, the remaining budget is divided in proportion to $(y_a M_{aj})^{2/3}$. Integer parts are allocated first; the remaining samples go to the largest fractional remainders, with action order resolving ties. This rounding is feasible and reproducible; exact optimality in Equation (35) refers to the continuous allocation, not to every integer implementation.

An optimal uncapped cover may be chosen with at most $|I_j|$ positive action weights. To see this, introduce nonnegative slacks for the $|I_j|$ coverage inequalities. The resulting feasible polyhedron is pointed, and a finite optimum has a basic optimal solution with at most $|I_j|$ basic variables. Its positive action-weight count cannot exceed that number. This explains why pilot selection can reduce validation costs even when the original mask family contains many overlapping actions. A cover selected on noisy pilot energies need not be population-optimal; fresh validation protects its validity, not its optimality.

B.7 A checkable certificate for a frozen encoder

The procedure returns a bound together with its ambiguity class and sampling assumptions. Construct D_j from the declared masks, retain the factor-to-output assignment, and decide whether the target is full separation or only distance to functions of C_{D_j}. An allowed mixed dependency is not an estimation failure. It remains outside the protected error even with arbitrarily many validation pairs.

Next, specify a product reference experiment and the conditional laws from which validation pairs are drawn. Direct simulator refreshes give $\kappa_a = 1$. With replay, each action instead needs a justified domination or coupling correction. Actions for which no finite correction is available can be omitted, but the remaining actions must still cover the forbidden coordinates. A missing cover returns an unavailable multiplicative certificate, not a small numerical bound.

Compute the simultaneous energy upper bounds, solve the fractional-cover dual, and report $\sum_j \min\{v_j, \mathcal{C}_j(U)\}$ when valid variance caps are known. The reported weights identify which measured interventions support the conclusion. The packing solution separately exhibits an additive residual consistent with those upper bounds, explaining why the certificate cannot be tightened without additional information. A feasible dual is required for an upper certificate; numerical solver tolerances must not be confused with statistical uncertainty.

Information retention is a separate endpoint. A constant encoder has zero leakage and zero intervention loss, and a nearly constant encoder can have an ill-conditioned inverse. Reconstruction, output variance, and inverse conditioning therefore accompany the bound rather than being inferred from it. In the invertible controls below, these quantities are known analytically. In the neural settings, oracle variation and recovery metrics remain diagnostics unless all collection and validation premises are established.

C Controlled computational protocols and additional results

The supplementary implementation uses Python, NumPy, SciPy, and PyTorch on the CPU. The included generators and solvers reproduce the controlled results. The numerical routines and

Table 4: Certificate divided by exact off-factor variation, mean ± SE over 100 independent encoder-design draws per dimension. Every bound is capped by the known output variance. Singleton uses the directly observed singleton energies.

K	Global q_*	Coordinate q_j	Singleton	Mask-resolved LP
4	14.803 ± 1.669	6.004 ± 0.818	1.173 ± 0.018	1.097 ± 0.011
8	12.218 ± 0.839	4.669 ± 0.431	1.132 ± 0.008	1.056 ± 0.004
16	9.666 ± 0.560	3.108 ± 0.236	1.126 ± 0.007	1.057 ± 0.004

seed-level arrays accompany the controlled experiments; neural protocols are described in Appendices D.1 and E. The controlled encoder families isolate certificate properties rather than comparing likelihood-based temporal learners.

The loss convention is explicit throughout. The training loss and the aggregate theorem use full squared increments. The action-resolved LP and its confidence intervals use half squared increments. For exact reference pairs, $\mathcal{L}_{\mathrm{int}} = 2\sum_a p_a \sum_{j \notin T_a} e_{aj}$. Every paired comparison uses one frozen output scaling; no bound is formed from raw loss and separately standardized variation.

C.1 Population certificate comparisons

We first use invertible nonlinear encoders with known population variances. For independent $C_i \sim U[-\sqrt{3}, \sqrt{3}]$, the first $K-1$ outputs equal their factors and the last is a normalized triangular shear,

$$F_K(C) = \frac{C_K + \sum_{i<K} b_i C_i + \sum_{\{i,\ell\} \in \mathcal{P}} d_{i\ell} C_i C_\ell}{\sqrt{1 + \sum_i b_i^2 + \sum_{\{i,\ell\} \in \mathcal{P}} d_{i\ell}^2}}. \tag{36}$$

The inverse and all product-decomposition energies are explicit. For each dimension, 100 independently seeded encoders and heterogeneous action designs combine singleton and overlapping block refreshes. The coefficient scale is fixed before evaluation. Table 4 compares each certificate with the exact error; a ratio of one is tight. All bounds use the same action-energy information, except that the singleton baseline deliberately ignores the block measurements.

The LP is substantially tighter than the frequency-based bounds and improves more modestly over the stronger singleton comparison. The singleton comparison isolates the additional benefit of overlapping masks after removing action-frequency imbalance. The exponential interaction LP and the reduced packing LP also agree to numerical precision on 200 cases with $K \le 8$. A large-mixing control in Appendix C.2 shows that the variance cap narrows the advantage when the true leakage already approaches the total output variance.

C.2 Nonlinear certificate benchmark

For dimension $K \in \{4, 8, 16\}$ and seed $s \in \{0, \dots, 99\}$, use a NumPy generator initialized at $60000 + 1000K + s$. Draw $K-1$ independent standard Gaussian linear coefficients. Draw a random permutation of singleton log-weights equally spaced between -3 and 1, exponentiate and normalize, and assign total probability 0.6 to these singleton masks. Add K independently sampled blocks of size $\max\{2, \lfloor K/2 \rfloor\}$, each with probability $0.4/K$. Distinct action identifiers can have equal target masks.

Choose up to $K-1$ unordered pairs from $\{1, \dots, K-1\}$ without replacement, using a random permutation of all pairs. Their independent standard Gaussian coefficients are multiplied by 0.35. In the principal comparison, multiply both coefficient groups by $0.25/\sqrt{K-1}$. The factor law and normalized encoder are given in Equation (36). Independent centered unit-variance factors make each linear term and each distinct pair product orthogonal. Therefore the true residual is

$$\mathcal{R}_K = \frac{\sum_i b_i^2 + \sum_{\{i,\ell\} \in \mathcal{P}} d_{i\ell}^2}{1 + \sum_i b_i^2 + \sum_{\{i,\ell\} \in \mathcal{P}} d_{i\ell}^2},$$

and the energy of a tested mask is the sum of the normalized squared coefficients whose index sets intersect it. The other outputs have zero off-factor variation and zero off-target energy. The known variance cap is exactly one.

The global comparison is $\min\{1, \mathcal{L}_{\text{int}}/(2q_*)\}$. The coordinate comparison replaces q_* with q_K. The singleton comparison is $\min\{1, \sum_{i<K} e_{\{i\},K}\}$. The LP uses all tested action energies and is capped at one as well. Results are ratios computed separately for each encoder and then averaged; they are not ratios of pooled averages. A large-mixing control removes the overall coefficient multiplier, leaving Gaussian linear coefficients and pair coefficients at scale 0.35. The large-mixing control results are reported in Table 5.

Table 5: Large-mixing control with the same action families and seeds. Entries are variance-capped certificate/error ratios, mean $\pm$ SE. The variance cap substantially reduces the advantage over coarse bounds when the residual is already large.

K	Global q_*	Coordinate q_j	Singleton	Mask-resolved LP
4	1.596 ± 0.060	1.555 ± 0.054	1.171 ± 0.018	1.097 ± 0.011
8	1.166 ± 0.009	1.166 ± 0.009	1.117 ± 0.007	1.054 ± 0.004
16	1.069 ± 0.002	1.069 ± 0.002	1.067 ± 0.002	1.050 ± 0.003

For $K \leq 8$, the implementation separately solves a packing LP over every nonempty subset of the $K-1$ forbidden factors and compares it with Equation (10). Components additionally containing the allowed factor K have the same energy incidence and do not change that exponential LP's value. Across 200 comparisons, the largest absolute difference is 5.55×10^{-17}. This is a numerical check of the reduction, not a substitute for its proof.

C.3 Rare separation in learned, invertible encoders

The learned control observes $S = \tanh(AC)$, where A is an unknown orthogonal mixing matrix, and fits $f_W(S) = W \operatorname{artanh}(S)$ with orthogonal W. The encoder receives observation pairs and masks, not latent labels. The known coordinatewise inverse nonlinearity makes this a restricted but exactly invertible family. Consequently $F(C) = MC$, $M = WA$, every output has unit variance, and the inverse condition number is one.

For $K = 4$, give each singleton and each complementary three-factor mask probability $q/2$, and give the all-target and null masks probability $1/2 - 2q$ each. Every factor is targeted with probability $1/2$, the expected target count is two, and every ordered-pair margin equals q. Across 12 paired training seeds, optimization uses 256 steps, batch size 256, and learning rate 0.06. Independent validation uses $50{,}000$ pairs per seed; conditional oracle resampling estimates variation on a disjoint sample.

Rare distinguishing actions leave substantial residual mixing despite identical marginal intervention frequencies. In this additive family, $\mathcal{L}_{\text{int}} = 2q \sum_j \mathcal{R}_j$ exactly, so the experiment checks the sharp scale dependence without conflating it with approximation error in the conditional-variance estimator. It does not establish that nonlinear neural training always approaches equality.

C.4 Learned invertible rare-action control

For each training seed $s \in \{0, \ldots, 11\}$, generate A and the initial W by QR decomposition of independent Gaussian matrices, with the diagonal sign correction that gives a Haar-distributed orthogonal matrix. Their generator seeds are $10000 + s$ and $20000 + s$. At every optimization step, sample a fresh batch of independent uniform factors and independent target refreshes with generator seed $30000 + s$. Use $S = \tanh(AC)$ and $S' = \tanh(AC')$ as the observations. The optimizer evaluates $\operatorname{artanh}(S)$ internally, but it does not receive A, C, or a factor-label regression target.

For a batch of n_b rows, write $\Delta X = \operatorname{artanh}(S') - \operatorname{artanh}(S)$, $\Delta Z = \Delta X W^\top$, and let H be the binary matrix of untargeted outputs. The Euclidean gradient is

$$G_W = \frac{2}{n_b}(H \odot \Delta Z)^\top \Delta X.$$

Update W to the orthogonal polar factor of $W - \eta G_W$ by singular-value decomposition. The fixed settings are $n_b = 256$, $\eta = 0.06$, and 256 updates. The five designs have $q \in \{0.25, 0.10, 0.05, 0.02, 0.01\}$, with the masks and probabilities described in Section C.3. Probability-zero null and all-target actions at $q = 0.25$ are unused.

Table 6: Learned rare-action control. Means ± SE across 12 training seeds. Both endpoints use the same unit-variance encoder. Exact reference values follow from the fitted matrix, not from the optimizer's training loss. A superscript † denotes SE below 0.0005. The corresponding independent Monte Carlo estimates appear in Appendix C.4.

q_*	Exact $\mathcal{L}_{\text{int}}$	Exact $\sum_j \mathcal{R}_j$	MCC
0.25	$(4.27 \pm 2.20) \times 10^{-11}$	$(8.53 \pm 4.40) \times 10^{-11}$	$1.000^{\dagger}$
0.10	0.00452 ± 0.00242	0.023 ± 0.012	0.997 ± 0.002
0.05	0.08331 ± 0.02728	0.833 ± 0.273	0.936 ± 0.016
0.02	0.09535 ± 0.00820	2.384 ± 0.205	0.791 ± 0.026
0.01	0.05628 ± 0.00352	2.814 ± 0.176	0.779 ± 0.028

After fitting, compute $M = WA$. The exact quantities are

$$\sum_j \mathcal{R}_j = \sum_j \sum_{i \neq j} M_{ji}^2, \qquad \mathcal{L}_{\text{int}} = 2 \sum_j \sum_{i \neq j} q_{ij} M_{ji}^2. \tag{37}$$

Unit output variances follow from $MM^\top = I$ and the unit-variance factor law. The Pearson correlation matrix is M, so MCC is the mean absolute entry under an optimal one-to-one assignment, found by the Hungarian algorithm. The diagonal mean absolute correlation is reported separately because the intervention loss uses the designated, not a newly optimized, assignment.

Validation uses generator seed $40000 + s$ and $50{,}000$ independent replay pairs. The variation estimator uses a subsequent disjoint sample for each output, resampling all factors except its assigned factor and averaging half the squared output difference. This is an unbiased oracle-resampling estimator of the same-scale conditional variance. The exact inverse is $C = A^\top W^\top F(C)$. The resulting Monte Carlo checks are reported in Table 7.

Table 7: Independent Monte Carlo checks for the learned rare-action experiment, mean ± SE across 12 training seeds. Assigned correlation uses the fixed mask assignment; MCC in Table 6 allows a separate matching only as a descriptive metric. A superscript † denotes SE below 0.0005.

q	Monte Carlo $\mathcal{L}_{\text{int}}$	Monte Carlo $\sum_j \mathcal{R}_j$	Assigned correlation
0.25	$(4.26 \pm 2.20) \times 10^{-11}$	$(8.56 \pm 4.41) \times 10^{-11}$	$1.000^{\dagger}$
0.10	0.005 ± 0.002	0.023 ± 0.012	0.997 ± 0.002
0.05	0.084 ± 0.027	0.835 ± 0.274	0.851 ± 0.056
0.02	0.097 ± 0.008	2.385 ± 0.205	0.572 ± 0.044
0.01	0.057 ± 0.004	2.815 ± 0.176	0.472 ± 0.040

The near-zero values at the largest margin are retained in scientific notation. They arise from finite-budget convergence in this restricted family, not from a general claim that predictive training reaches an exactly identified optimum. At the smaller margins, high matched correlation can coexist with appreciable variation relative to the designated coordinate assignment.

C.5 Tightness, calibration, and validation cost

When singleton tests are available, they provide a strong comparator. Across 100 nonlinear invertible encoder-design draws per dimension, the singleton and LP certificate/error ratios are 1.173 ± 0.018 and 1.097 ± 0.011 at $K = 4$, and 1.132 ± 0.008 and 1.056 ± 0.004 at $K = 8$ (Table 4). The much larger improvements over frequency-based bounds are not improvements over the singleton baseline. The reduced and exponential interaction LPs agree to numerical precision on 200 small-dimensional cases. A large-mixing control retains the known variance cap for every method (Appendix C.2).

A fixed nonlinear four-factor shear is validated on 100 independent repetitions at each sample size. All methods share the same action-conditional samples. The analytic polynomial ranges give valid loss envelopes, and the encoder is fixed before any validation draw. Table 8 reports the certificate-to-error ratio, not an empirical claim that the nominal coverage equals one. At the smaller budget, the uncorrected plug-in LP covers the true error in only $95/100$ repetitions; its apparent tightness is not a confidence guarantee.

Table 8: Finite-sample certificate/error ratios, mean $\pm$ SE across 100 independent validation repetitions; $\delta = 0.05$. EB denotes empirical Bernstein. The three corrected bounds cover the exact error in all repetitions at both budgets. Budgets count every tested action.

Total pairs	EB + coordinate bound	Hoeffding + LP	EB + LP
8,000	3.210 ± 0.007	1.449 ± 0.002	1.208 ± 0.002
80,000	2.809 ± 0.002	1.164 ± 0.001	1.070 ± 0.001

A separate eight-factor study charges an independent pilot against a total budget of 20,000 pairs. Uniform validation over all tested actions gives a Hoeffding certificate/error ratio of 2.625 ± 0.002. Uniform validation on the pilot-selected cover gives 2.494 ± 0.012; allocating the remaining samples by Equation (33) gives 2.283 ± 0.002. The cover uses 4.09 ± 0.05 actions on average. All three procedures cover in $100/100$ repetitions. Thus choosing which actions to validate and allocating their samples are complementary, even after accounting for the pilot.

C.6 Finite-sample calibration benchmark

Use the nonlinear generator above with $K = 4$, seed 777, and the principal coefficient scale. Fix this encoder and design for all repetitions. There are four nontrivial tested actions for its last output; the remaining off-target outputs are analytically invariant and have known zero energy. For each tested action, let

$$H_a = \sum_{U: U \cap T_a \neq \varnothing} |b_U| (\sqrt{3})^{|U|}, \qquad M_{aK} = \frac{2H_a^2}{1 + \sum_U b_U^2},$$

where b_U includes both the linear and pair coefficients. Untargeted C_K cancels in the increment. Since each changing term has absolute value at most $|b_U|(\sqrt{3})^{|U|}$ at either endpoint, the half squared increment is bounded by M_{aK}. This is an analytic support bound and does not use the validation maximum.

At each action, collect $n_a = 2000$ or 20000 fresh pairs, giving total counts 8000 or 80000. Repetition r uses generator seed $700000 + n_a + r$, for $r = 0, \ldots, 99$. The EB intervals use Equation (16) with $m = 4$ and $\delta = 0.05$. The Hoeffding alternative uses

$$U_{aK}^{\mathrm{H}} = \min\{M_{aK}, \bar{Y}_{aK} + M_{aK}\sqrt{\log(m/\delta)/(2n_a)}\}.$$

The LP is solved separately using each upper-bound vector. The EB coordinate comparison uses $\sum_a p_a U_{aK}/q_K$ and the same variance cap. The uncorrected plug-in comparison solves the LP using $\bar{Y}_{aK}$ without a confidence correction. The resulting calibration results are reported in Table 9.

Table 9: Uncorrected plug-in comparison. The cover count is the number of independent validation repetitions in which the numerical plug-in bound exceeds the known true residual; it is not a guaranteed confidence level.

Total pairs	Plug-in certificate/error	Covers true error
8,000	1.025 ± 0.002	$95/100$
80,000	1.029 ± 0.001	$100/100$

In the validation-allocation study, use $K = 8$ and seed 123. For each of 100 repetitions, the independent pilot collects 100 pairs per tested action and solves the energy LP on those pilot means. The positive dual weights define the cover. Generator seeds are $1100000 + r$ for the pilot and $1200000 + r$ for fresh validation. The uniform-all baseline spends all 20,000 pairs on fresh samples, equally across tested actions. The cover methods subtract the full pilot count from that budget before collecting fresh validation on the selected actions. Uniform-cover sampling divides those remaining pairs equally; allocated-cover sampling uses Equation (33) with the integer procedure in Appendix B. Each selected action is tested with a simultaneous Hoeffding interval at $\delta = 0.05$, conditional on the pilot. The pilot dual vector is kept fixed when summing these intervals. All compared bounds are capped by the known variance.

C.7 LOCALIZED OVERLAP FAILURE

Let C_1, C_2 be independent uniform variables on $[-\sqrt{3}, \sqrt{3}]$, and define

$$\phi(c) = \frac{\max\{-c, 0\}^2}{3}, \qquad F_1(c) = \frac{c_1 + b\phi(c_1)c_2}{\sqrt{1 + b^2/10}}, \qquad F_2(c) = c_2, \tag{38}$$

with $b = 0.4$. The map is C^1 on the open product support. Before the fixed output normalization, its derivative in the first coordinate is at least $1 - 2b = 0.2$. Hence it is injective, with the second coordinate unchanged and the first strictly increasing at every fixed second coordinate. The map is a diffeomorphism onto its image. Its two reference output variances are exactly one.

Since $\mathbb{E}\phi(C_1)^2 = 1/10$, the true residual is

$$\mathcal{R}_1 = \frac{b^2/10}{1 + b^2/10} = 0.015748031\ldots, \qquad \mathcal{R}_2 = 0.$$

Give the two singleton masks probability $q = 0.1$ each, and the joint and null masks probability 0.4 each. Collect initial states from $B_{s,C} = sQ + (1-s)Q(\cdot \mid C_1 > 0)$, with $s \in \{1, 0.5, 0.25, 0.1, 0\}$. Actions and targeted refresh laws remain fixed. For $s > 0$, $\kappa = 1/s$ follows from the exact state-density ratio. Only the second singleton action produces an off-target residual, and its residual is supported in $C_1 < 0$. Therefore

$$\mathcal{L}_{\mathrm{int}}(F; B_s) = 2qs\mathcal{R}_1, \qquad \frac{\kappa\mathcal{L}_{\mathrm{int}}(F; B_s)}{2q} = \mathcal{R}_1 \quad (s > 0). \tag{39}$$

At $s = 0$, the replay loss is zero while the reference residual remains positive. The multiplicative transfer bound is unavailable; the variance cap still gives its trivial upper bound. The localized overlap results are reported in Table 10.

Table 10: Localized overlap control, means ± SE across 12 independent Monte Carlo seeds with 100,000 replay pairs and a separate oracle reference sample per seed. The last column is a plug-in population-formula value, not a finite-sample confidence bound.

s	κ	Replay $\mathcal{L}_{\mathrm{int}}$	Reference $\sum_j \mathcal{R}_j$	Plug-in transfer
1	1	0.003137 ± 0.000016	0.01583 ± 0.00006	0.01569 ± 0.00008
0.5	2	0.001572 ± 0.000022	0.01584 ± 0.00006	0.01572 ± 0.00022
0.25	4	0.000811 ± 0.000014	0.01575 ± 0.00003	0.01622 ± 0.00028
0.1	10	0.000315 ± 0.000013	0.01578 ± 0.00004	0.01573 ± 0.00063
0	∞	0	0.01569 ± 0.00004	Unavailable

Repetition s_0 uses generator seed $900000 + s_0$. Reference variation is estimated by independently resampling C_2 at fixed C_1. The output normalization is the analytic constant in Equation (38); it is not refitted under the skewed replay law.

C.8 CANCELLATION BY UNTARGETED DRIFT

Use the same two-factor unit-variance reference and action probabilities, but set $F_1 = C_1$ and $F_2 = (C_2 + C_1)/\sqrt{2}$. Under action one, collect a transition that also changes the nominally untargeted coordinate according to

$$\widetilde{C}_2' = C_2 - \eta(C_1' - C_1), \qquad 0 \le \eta \le 1.$$

All other actions follow the ideal refresh law. The encoder is defined on $\mathbb{R}^2$, so the perturbed endpoint remains in its domain even when it leaves the reference box. Couple the initial states and all targeted draws. Then

$$\mathcal{R}_2 = \tfrac{1}{2}, \qquad \mathcal{L}_{\mathrm{int}}^{\mathrm{ref}} = q, \qquad \mathcal{L}_{\mathrm{int}}(F; B_\eta) = (1-\eta)^2 q, \qquad \delta_F^2 = \eta^2 q.$$

Thus the naive exact-resampling expression is $(1-\eta)^2/2$, while the corrected population expression is exactly $1/2$. The cancellation is deliberately constructed to test the necessity of the discrepancy term. It does not model or explain the RGB failure. The drift-cancellation results are reported in Table 11.

Table 11: Drift-cancellation control. Population error is $1/2$ in every row. Numerical entries are plug-in expressions, mean $\pm$ SE over 12 Monte Carlo seeds; they are not finite-sample upper confidence bounds. At full cancellation the naive population value is exactly zero.

Cancellation η	Naive expression	Coupling-corrected expression
0	0.49854 ± 0.00110	0.49854 ± 0.00110
0.25	0.28043 ± 0.00062	0.49854 ± 0.00110
0.5	0.12463 ± 0.00028	0.49854 ± 0.00110
0.75	0.03116 ± 0.00007	0.49854 ± 0.00110
1	$< 10^{-30}$	0.49854 ± 0.00110

The grid is $\eta \in \{0, 0.25, 0.5, 0.75, 1\}$, each seed uses $100{,}000$ independently generated pairs, and generator seeds are $800000 + s$ for $s = 0, \ldots, 11$. Common random numbers across grid values make the equality of the corrected estimates a property of the coupling construction, rather than independent evidence from five unrelated experiments.

D Neural validation and overlapping-mask experiments

D.1 Training and the frozen representation

The neural experiment in Section 6.1 evaluates 25 independently trained checkpoints across five mask/objective conditions. Checkpoints, seed-level losses, validation moments, oracle estimates, and feasible certificate weights accompany the specified architecture and observation generator.

For each seed $s \in \{0, 1, 2, 3, 4\}$, draw independent Gaussian matrices and use their QR factors to construct orthogonal $Q_1, Q_2 \in \mathbb{R}^{4\times 4}$ and a column-orthonormal $A_2 \in \mathbb{R}^{8\times 4}$. Set

$$A_1 = Q_1 \operatorname{diag}(0.7, 0.9, 1.1, 1.3) Q_2^\top, \qquad h(c) = A_2 \tanh(A_1 c + b),$$

where b has independent $\mathcal{N}(0, 0.1^2)$ entries. The four factors are independent $U[-1, 1]$ variables. The observation map has the inverse $A_1^{-1}[\operatorname{artanh}(A_2^\top h(c)) - b]$ on its image. The learner is not given this inverse, its matrices, or factor labels.

The encoder has widths $8 \to 64 \to 64 \to 4$ and GELU hidden activations. The predictor has widths $8 \to 64 \to 4$, receives the raw four-dimensional representation concatenated with the binary target mask, and uses the same hidden activation. The training objective is

$$\mathcal{L}_{\text{train}} = \mathbb{E}_B \left\| g(F(C), A) - \operatorname{sg}[F(C')] \right\|^2 + \alpha \mathcal{L}_{\text{iso}} + \lambda \mathcal{L}_{\text{int}}(F; B). \tag{40}$$

Equation (40) is optimized with Adam at learning rate 0.002, batch size 512, and 4000 updates. Prediction and invariance each sum squared residuals over their output coordinates before averaging over examples. Prediction uses stop-gradient on the next-state encoding; the invariance term differentiates through both endpoints. The sliced isotropy loss uses 16 fresh normalized Gaussian directions per update and matches sorted projections of the raw encoder outputs to a fresh standard-Gaussian sample of the same batch size. Its weight is $\alpha = 0.5$, while $\lambda = 3$ except in the no-penalty arm, where $\lambda = 0$. The invariance objective does not use latent labels.

Every update draws fresh independent reference states, target masks, and refresh values. The training count is $2{,}048{,}000$ pairs per model. The observation map and model initialization are shared across designs within each seed, and no hyperparameter or final checkpoint is selected on certificate-validation outcomes. All final models use the fixed 4000-step stopping rule. Uniform singleton masks have probability $1/4$ each. The pair design gives each of the six size-two masks probability $1/12$ and the null mask probability $1/2$. Both designs target each factor with marginal probability $1/4$ and have expected target count one. The nested masks $\{1\}$, $\{1, 2\}$, $\{1, 2, 3\}$, and $\{1, 2, 3, 4\}$ are uniform. The identical-signature masks $\{1, 2\}$, $\{3\}$, and $\{4\}$ are uniform. Their expected target counts are 2.5 and $4/3$, respectively, and both have $q_* = 0$. The no-penalty control uses uniform singletons. A nested design costs 2.5 refreshed factors per training pair and the identical-signature design costs $4/3$, whereas the two separating designs cost one. These are matched-update comparisons, not a claim that every design has the same intervention cost.

The certificate uses $G(c) = \tanh(f(h(c)))$ in every action loss and oracle residual. The readout is fixed before validation and is never rescaled by validation variances. The raw objective is used only for fitting. Boundedness supplies the deterministic half-energy envelope $M_{aj} = 2$ and per-output variance cap one. Neither the observed maximum nor an empirical variance is substituted for these population bounds. The coordinatewise injectivity of $\tanh$ preserves the exact allowed-dependence class, but the magnitude of the protected residual is that of G, not the raw encoder. Small numerical leakage after this transformation does not, by itself, establish a well-conditioned inverse for the raw map.

The implementation uses PyTorch on CPU with float32 network computation and float64 accumulation of validation moments. Each training process uses one computational thread. The observation, initialization, training, validation, and oracle generators have respective seeds $2100000 + s$, $2200000 + s$, $2300000 + s$, $2400000 + s$, and $2500000 + s$. The first and final two are NumPy generators; initialization and training use PyTorch generators. Random-number sharing across designs reduces comparison noise without making different training seeds dependent.

D.2 Independent action validation and oracle uncertainty

For each nontrivial action with at least one untargeted output, validation collects $n_a \in \{2048, 32768, 262144\}$ independent reference pairs. Smaller budgets are prefixes of the largest sample for that action. Thus each fixed-budget guarantee is valid, but the three reported budgets are not independent repetitions. Null-action energies are identically zero and require no sampled pairs; all-target actions test no output. Singleton, pair, nested, and identical designs therefore have respectively 4, 6, 3, and 3 sampled actions and 12, 12, 6, and 8 tested action-output cells. The no-penalty singleton control uses the singleton counts. Each action pair evaluates all four outputs; counting action-output cells as independent collected pairs would overstate the data budget.

The empirical Bernstein intervals use $\delta = 0.05$ and a union bound over the tested cells within one frozen model and one stated budget. There is no claim of simultaneous 95% validity across all training seeds, all budgets, or both experimental suites. Conditional on the independently trained model, each action's validation pairs are iid and have exactly the product reference law, so $\kappa_a = 1$. Corollary 4 covers all four output bounds simultaneously within that model. The LP, aggregate, and singleton comparators use the same corrected upper bounds. The Hoeffding comparison replaces the interval formula, not the samples, masks, or range envelope. Each bound is capped at one per output. When a forbidden factor is uncovered, the routine reports the full-separation cover as unavailable; the trivial variance cap remains valid.

A separate oracle sample holds C_j fixed and refreshes all other factors to estimate $\mathcal{R}_j$ by half the squared paired increment. For $\mathcal{R}_j^{\mathrm{allow}}$, it instead holds C_{D_j} fixed. Each nonzero residual uses $262{,}144$ fresh pairs, with disjoint draws for different outputs and residual definitions. A residual with $D_j = [4]$ is identically zero and is not sampled. Full and partial estimates may therefore differ slightly even when their population quantities coincide. These estimates never enter an energy upper bound, an LP objective, or an encoder-selection decision.

For a fixed checkpoint, the Monte Carlo standard error of a residual is its sample standard deviation divided by the square root of its pair count. For a sum over independently sampled outputs, the reported oracle Monte Carlo error is the square root of the sum of these squared standard errors. The paper's main-table SE instead measures variation across the five independently trained models. Table 12 reports oracle precision and complementary information diagnostics. The finite-sample theorem supplies confidence; exceeding a Monte Carlo oracle estimate in five fitted models is not an empirical proof of 95% coverage of exact unknown neural population errors. The information probe fits a linear map from all four coordinates of G to all four latent factors on $32{,}768$ independent states, then evaluates R^2 on another $32{,}768$ states. Factor labels are used only in this post-training diagnostic. The reported score averages the four factor-specific coefficients of determination. Output variance and the Pearson correlation matrix are evaluated on the independent probe test states. No reconstruction or probe loss enters training. These observations establish nonconstant representations with accessible factor information in the evaluated sample, not a global injectivity or inverse-conditioning theorem.

Table 13 reports how the validation budget changes the certificate. With the global envelope $M_{aj} = 2$, the linear $1/n_a$ correction dominates very small neural losses at the smallest budget.

Table 12: Complementary diagnostics on the same bounded neural representations, mean ± SE over five training seeds. Oracle MCSE is the mean within-checkpoint standard error of the summed allowed residual, in 10^{-6}; its uncertainty across seeds is also reported. MCC uses an optimal matching only as a diagnostic, not to change the certificate assignment. The probe is multivariate ordinary least squares with an intercept; † denotes SE below 0.0005.

Design	Mean variance	MCC	Probe R^2	Oracle MCSE
Singleton	0.446 ± 0.003	$0.988^{\dagger}$	0.977 ± 0.001	1.04 ± 0.09
Pair masks	0.458 ± 0.003	$0.985^{\dagger}$	0.970 ± 0.001	1.83 ± 0.16
Nested	0.345 ± 0.020	0.520 ± 0.072	0.598 ± 0.064	1.87 ± 0.27
Identical	0.434 ± 0.004	0.934 ± 0.029	0.968 ± 0.004	0.78 ± 0.06
No penalty	0.396 ± 0.005	0.759 ± 0.030	0.942 ± 0.003	1349.67 ± 80.89

More observations reduce that uncertainty, but cannot remove mixing permitted by the intervention signatures. The aggregate comparison for incomplete separation uses r_j, not the zero full-separation margin q_*. The stored dual weights are nonnegative and their coverage is checked before forming the upper objective. A positive minimum coverage below one is repaired by rescaling the weights; the implementation also applies a multiplicative $1 + 10^{-12}$ feasibility margin. This is floating-point numerical validation, not interval arithmetic.

Table 13: Neural sample-size sweep, mean ± SE across five training seeds. Bounds concern the summed allowed-class residual on the bounded output scale; every certificate entry is in 10^{-3}. Total pairs count each sampled action once. Data at smaller budgets are prefixes of the largest-budget validation sample.

Design	n_a	Total pairs	EB + aggregate	Hoeffding + LP	EB + LP
Singleton	$2,048$	$8,192$	169.710 ± 0.075	878.590 ± 0.062	169.710 ± 0.075
Singleton	$32,768$	$131,072$	11.267 ± 0.064	220.161 ± 0.061	11.267 ± 0.064
Singleton	$262,144$	$1,048,576$	2.013 ± 0.062	78.279 ± 0.061	2.013 ± 0.062
Pair masks	$2,048$	$12,288$	85.648 ± 0.115	439.976 ± 0.097	85.648 ± 0.115
Pair masks	$32,768$	$196,608$	6.337 ± 0.096	110.755 ± 0.092	6.337 ± 0.096
Pair masks	$262,144$	$1,572,864$	1.689 ± 0.093	39.812 ± 0.091	1.689 ± 0.093
Nested	$2,048$	$6,144$	75.988 ± 0.191	205.753 ± 0.106	38.218 ± 0.124
Nested	$32,768$	$98,304$	5.584 ± 0.171	51.908 ± 0.110	2.996 ± 0.115
Nested	$262,144$	$786,432$	1.462 ± 0.165	18.758 ± 0.110	0.930 ± 0.112
Identical	$2,048$	$6,144$	105.734 ± 0.033	563.654 ± 0.027	105.734 ± 0.033
Identical	$32,768$	$98,304$	7.045 ± 0.028	141.254 ± 0.027	7.045 ± 0.028
Identical	$262,144$	$786,432$	1.281 ± 0.027	50.233 ± 0.026	1.281 ± 0.027
No penalty	$2,048$	$8,192$	1379.033 ± 100.147	1969.702 ± 91.231	1379.033 ± 100.147
No penalty	$32,768$	$131,072$	1132.639 ± 98.931	1311.831 ± 96.611	1132.639 ± 98.931
No penalty	$262,144$	$1,048,576$	1103.109 ± 97.347	1168.884 ± 96.532	1103.109 ± 97.347

The seed-level endpoints are retained in Table 14. They use the largest validation budget and show both full and partial oracle residuals. The two are not interchangeable under nested or identical signatures. No seed is removed for weak correlation, large leakage, or a loose certificate.

D.3 Predictable concentration on the same frozen checkpoints

The PrPl evaluation regenerates the original action-conditional samples with the same random seeds, action order, batch boundaries, and fixed readout. Each stored sample mean agrees with the original validation record to absolute error below 10^{-10}. Oracle residuals and assignments are unchanged. Equation (18) is recomputed for each prespecified budget, using $c = 0.9$ and the same m-cell allocation as EB. Table 15 reports the complete comparison; it does not select the smaller confidence bound after seeing an outcome.

For mask-only validation, the same procedure is applied to the three available tests of output four, on the same 20 repetitions per seed used by the EB comparison. The confidence allocation is $\eta = 0.05/3$. Averaging within a training seed before reporting the five-seed SE keeps validation

Table 14: All neural training seeds. Leakage and certificate values are in 10^{-3}; the bound is EB+LP for the allowed residual with $262{,}144$ pairs per tested action. Oracle point estimates come from independent reference resampling.

Design	Seed	Full leakage	Allowed residual	EB + LP	MCC
Singleton	0	0.5491	0.5518	2.0238	0.9868
Singleton	1	0.5729	0.5728	2.0379	0.9892
Singleton	2	0.4474	0.4454	1.9463	0.9885
Singleton	3	0.5987	0.5948	2.2175	0.9891
Singleton	4	0.3427	0.3431	1.8379	0.9879
Pair masks	0	0.6882	0.6904	1.4575	0.9838
Pair masks	1	1.1565	1.1570	1.9533	0.9846
Pair masks	2	0.7875	0.7882	1.6465	0.9850
Pair masks	3	0.6906	0.6886	1.5417	0.9864
Pair masks	4	0.9178	0.9195	1.8462	0.9844
Nested	0	908.6162	0.2564	0.5547	0.6369
Nested	1	973.5648	0.8281	1.1272	0.6923
Nested	2	1033.8890	0.6603	0.9654	0.3855
Nested	3	1127.3721	0.8617	1.1728	0.3213
Nested	4	984.3178	0.5251	0.8290	0.5645
Identical	0	366.6893	0.4198	1.3259	0.8584
Identical	1	909.4057	0.3806	1.2736	0.9876
Identical	2	873.9883	0.3793	1.3377	0.9864
Identical	3	485.7537	0.3739	1.2831	0.8698
Identical	4	77.5910	0.2770	1.1852	0.9671
No penalty	0	689.9562	691.7624	724.2592	0.7657
No penalty	1	1092.4219	1092.9674	1159.9182	0.6881
No penalty	2	1137.7661	1139.0114	1208.6010	0.6930
No penalty	3	1236.8033	1237.4108	1275.4037	0.8143
No penalty	4	1093.7060	1089.5782	1147.3625	0.8341

Table 15: Variance-adaptive reevaluation of all 25 neural checkpoints. Certificate entries are in 10^{-3}, mean $\pm$ training-seed SE over five seeds. The last column averages per-seed PrPl-certificate/oracle-error ratios, rather than taking a ratio of pooled means. Each row has its own fixed-budget 95% guarantee.

Design	n_a	EB + LP	PrPl + LP	PrPl / oracle
Singleton	$2,048$	169.710 ± 0.075	72.987 ± 0.062	151.75 ± 16.72
Singleton	$32,768$	11.267 ± 0.064	5.192 ± 0.061	10.75 ± 1.07
Singleton	$262,144$	2.013 ± 0.062	1.246 ± 0.061	2.54 ± 0.16
Pair masks	$2,048$	85.648 ± 0.115	37.174 ± 0.097	45.47 ± 4.07
Pair masks	$32,768$	6.337 ± 0.096	3.272 ± 0.092	3.96 ± 0.27
Pair masks	$262,144$	1.689 ± 0.093	1.296 ± 0.091	1.55 ± 0.05
Nested	$2,048$	38.218 ± 0.124	16.439 ± 0.107	31.64 ± 8.05
Nested	$32,768$	2.996 ± 0.115	1.613 ± 0.111	2.91 ± 0.50
Nested	$262,144$	0.930 ± 0.112	0.751 ± 0.110	1.24 ± 0.06
Identical	$2,048$	105.734 ± 0.033	45.122 ± 0.027	125.70 ± 9.51
Identical	$32,768$	7.045 ± 0.028	3.238 ± 0.027	9.00 ± 0.61
Identical	$262,144$	1.281 ± 0.027	0.801 ± 0.026	2.21 ± 0.09
No penalty	$2,048$	1379.033 ± 100.147	1278.403 ± 99.269	1.22 ± 0.02
No penalty	$32,768$	1132.639 ± 98.931	1124.754 ± 98.833	1.07 ± 0.01
No penalty	$262,144$	1103.109 ± 97.347	1101.632 ± 97.288	1.05 ± 0.01

repetition distinct from training replication. Table 16 retains both integral and fractional covers. The global half-energy envelope is still two; the improvement comes from concentration, not an estimated range. The fixed-weight PrPl bound still becomes loose when the third test is scarce. The optimized cover removes it, whereas balanced tests favor half weights on all three. Variance adaptation and mask resolution therefore address different sources of looseness. Neither improvement establishes full factor recovery under a nonseparating design.

Table 16: Mask-only neural reevaluation. Entries are on the actual bounded-output scale, mean ± training-seed SE. Oracle leakage of output four is 0.0002068 ± 0.0000424. Each of the first two tests receives $32{,}768$ pairs.

n_{third}	EB fractional	PrPl fixed	PrPl integral	PrPl fractional
32	0.001686 ± 0.000054	0.162333 ± 0.000047	0.000874 ± 0.000051	0.000874 ± 0.000051
512	0.001686 ± 0.000054	0.009622 ± 0.000047	0.000874 ± 0.000051	0.000874 ± 0.000051
$32,768$	0.001270 ± 0.000049	0.000661 ± 0.000047	0.000854 ± 0.000057	0.000661 ± 0.000047

D.4 A NONLINEAR BENCHMARK WITH NO SINGLETON MEASUREMENTS

Let $g \in \{1, 3, 8\}$, $d = 3g$, and $K = d + 1$. Partition the first d factors into disjoint triads, and let $\mathcal{P}$ contain the three unordered within-triad pairs in every triad. All factors are independent $U[-\sqrt{3}, \sqrt{3}]$ variables. Set $\beta = 0.25/\sqrt{d}$, $\gamma = 0.2$, and

$$F_K(C) = \frac{C_K + \beta \sum_{i=1}^{d} C_i + \gamma\beta \sum_{\{i,\ell\} \in \mathcal{P}} C_i C_\ell}{a}, \qquad a^2 = 1 + d\beta^2 + d\gamma^2\beta^2. \tag{41}$$

All other outputs equal their factors. The map is triangular and invertible, since $C_1, \ldots, C_d$ are retained and C_K is recovered by subtracting the displayed polynomial and multiplying by a. The linear terms and the distinct pair products are orthogonal under Q. Every output has population variance one, and the only off-factor variation is

$$\mathcal{R}_K = \frac{d\beta^2(1 + \gamma^2)}{a^2} = \frac{0.065}{1.065}. \tag{42}$$

This is an exact nonlinear reference error, not an estimated conditional variance.

Each triad provides only its three size-two refresh masks. The study certifies the preassigned last output, whose forbidden set is $[d]$; it does not claim that this restricted family identifies every other output. No singleton energy is measured or inferred as equality data. If the actions receive equal reference weights, the last-output margin is $q_K = 2/d$, so the aggregate dual has weight $1/2$ for every available action. A mask touches two linear terms and all three within-triad pair terms, giving half-energy

$$e_a = \frac{\beta^2(2 + 3\gamma^2)}{a^2}.$$

Changing one triad leaves every term outside it and C_K unchanged. The absolute sum of affected terms at either endpoint is at most $(2\beta\sqrt{3} + 9\gamma\beta)/a$. Therefore

$$M = \frac{2(2\beta\sqrt{3} + 9\gamma\beta)^2}{a^2} \tag{43}$$

is a valid half-squared-increment envelope. The data generator samples the affected triad alone because the remaining terms cancel exactly; it is an exact evaluation of these full-encoder increments, not a replacement model.

For each triad, two actions receive 32768 pairs and the third receives $n_{\text{third}} \in \{32, 512, 32768\}$. This is a controlled test-imbalance experiment: the scarce action is chosen by design, not identified by inspecting a favorable validation outcome. Each configuration has 100 independent validation repetitions. A repetition uses the NumPy seed $3100000 + 10000g + r$, where $r = 0, \ldots, 99$; samples and uncertainty levels are identical across the compared certificates. The different budgets reuse the generator-seed convention and are not claimed to be independent of one another. The confidence allocation uses $m = 3g$ action-output cells and $\delta = 0.05$ separately in each repetition and configuration.

For upper bounds (u_1, u_2, u_3) in one triad, an integral cover must choose at least two masks. Its optimal value is $u_1 + u_2 + u_3 - \max_i u_i$. The fractional optimum is the smaller of that value and $(u_1 + u_2 + u_3)/2$, as in Equation (13). Summing independent triad values solves the full programs exactly. All final values are capped by the same known variance one. This supplies an exact best-integral comparator rather than a greedy cover heuristic. A generic LP solver independently checks the fractional closed form in the first three repetitions of all nine configurations.

At population energies, the fixed-weight and fractional ratios are $(3 + 4.5\gamma^2)/(3 + 3\gamma^2) = 1.01923\ldots$, whereas the best integral ratio is $(4 + 6\gamma^2)/(3 + 3\gamma^2) = 1.35897\ldots$. The fractional bound need not equal the encoder's actual residual, because the upper-information class also permits additive functions with the same upper bounds. Pair interactions are included precisely to avoid making every result an additive equality case. Finite-sample results are given in Table 17. The balanced regime favors fractional weighting over an integral action subset. The highly imbalanced regime instead favors selecting two precise tests and omitting the third, in which case the best integral cover matches the LP. The intermediate budget can use either pattern in different triads. This distinction prevents a large gain over a fixed-weight bound from being misreported as the gain over the strongest alternative.

Table 17: Complete overlapping-only sample-budget sweep. Entries are certificate/error ratios, mean ± SE across 100 independent repetitions per setting. All three corrected methods contain the exact reference error in $100/100$ repetitions in every row. The last column reports plug-in LP containment, not a confidence guarantee.

K	n_{third}	Total pairs	Fixed weight	Best integral	Fractional LP	Plug-in contains
4	32	65,568	4.5133 ± 0.0178	1.4032 ± 0.0008	1.4032 ± 0.0008	$57/100$
4	512	66,048	1.3060 ± 0.0027	1.4032 ± 0.0008	1.3060 ± 0.0027	$76/100$
4	32,768	98,304	1.0531 ± 0.0004	1.3991 ± 0.0006	1.0531 ± 0.0004	$100/100$
10	32	196,704	5.2696 ± 0.0089	1.4112 ± 0.0005	1.4112 ± 0.0005	$66/100$
10	512	198,144	1.3607 ± 0.0016	1.4111 ± 0.0005	1.3602 ± 0.0015	$94/100$
10	32,768	294,912	1.0585 ± 0.0003	1.4065 ± 0.0004	1.0585 ± 0.0003	$100/100$
25	32	524,544	5.9563 ± 0.0057	1.4172 ± 0.0003	1.4172 ± 0.0003	$79/100$
25	512	528,384	1.4067 ± 0.0009	1.4174 ± 0.0003	1.4006 ± 0.0007	$100/100$
25	32,768	786,432	1.0629 ± 0.0002	1.4125 ± 0.0002	1.0629 ± 0.0002	$100/100$

D.5 Overlapping-only validation of learned neural encoders

The final mask-only check uses all five completed pair-mask neural checkpoints from Section 6.1, with no further fitting or checkpoint selection. Output four is fixed before evaluation. The available tests are exactly $\{1, 2\}$, $\{1, 3\}$, and $\{2, 3\}$, so every forbidden factor of that output is covered without singleton observations. The same bounded map G_4 is used for all losses and for the independently computed oracle residual.

Each training seed has 20 fresh validation repetitions with seeds $2700000 + 100s + r$. Within each repetition, the action counts are $(32768, 32768, n_{\mathrm{third}})$, where $n_{\mathrm{third}} \in \{32, 512, 32768\}$. Prefix sampling is used within the third action's budget sweep. The simultaneous EB upper bounds use $m = 3$, $M = 2$, and $\delta = 0.05$. Each reported certificate uses the same half-energy upper bounds and the same variance cap one; the integral and fractional values have the exact triangle formulas above.

Table 18 averages the 20 certificate values within each fixed training seed, then reports mean and SE over the five resulting training-seed means. The 100 validation repetitions are not treated as 100 independently trained encoders. The oracle entry is the mean across five independently evaluated fourth-output residuals. It remains an independent Monte Carlo estimate rather than an exact neural population value.

The large fixed-weight value at $n_{\mathrm{third}} = 32$ is driven by a globally bounded but weakly estimated action, not large measured leakage. Both optimized covers omit that action in every repetition and coincide. At balanced counts the half-weight fractional cover improves on the integral cover in every repetition. The resulting certificate is still conservative relative to the oracle residual: its mean per-training-seed ratio is 7.14 ± 1.25. This remaining gap reflects finite-sample range corrections and is retained rather than hidden behind a large relative gain over a weak comparator.

Table 18: Mask-only validation of the preassigned fourth output of all five pair-mask neural encoders. Bounds use the actual bounded output scale, not a ratio or a 10^{-3} multiplier. Entries are mean $\pm$ training-seed SE after averaging 20 fresh validation repetitions within each seed. Oracle leakage is 0.0002068 ± 0.0000424. The last column counts repetitions in which the optimized cover omits the third action.

n_{third}	Total pairs	Fixed-weight bound	Best integral cover	Fractional LP	Omits third
32	65,568	0.361363 ± 0.000075	0.001686 ± 0.000054	0.001686 ± 0.000054	100/100
512	66,048	0.022810 ± 0.000054	0.001686 ± 0.000054	0.001686 ± 0.000054	100/100
32,768	98,304	0.001270 ± 0.000049	0.001665 ± 0.000059	0.001270 ± 0.000049	0/100

E Public Pendulum-generator evaluation

E.1 Reference grid, observations, and the protected factors

The generator is adapted from the public CausalVAE Pendulum implementation (Yang et al., 2021; Huawei Technologies Co., Ltd., 2021). The original geometry, root-angle grid, patch ordering, axis limits, and 96-pixel raster are retained. The adaptation reuses an Agg canvas, removes file-name and dataframe bookkeeping, and downsamples each raster to 32×32 RGB with Pillow's BOX operator. The emitted image table, source attribution, and image-array digest accompany the code. This is a public-generator replication with an explicit resolution change; it is not a claim of byte identity to every previously distributed image archive.

The pendulum angle is $i\pi/200$ for $i \in \{-40, \dots, 43\}$. The illumination angle is $j\pi/200$ for $j \in \{60, \dots, 147\} \setminus \{100\}$. These give $84 \cdot 87 = 7308$ observations. The pivot is $(10, 10.5)$ and the ball center is $(10 + 8\sin\theta, 10.5 - 8\cos\theta)$. Light projection and shadow geometry follow the released renderer. The two roots are sampled independently and uniformly on their grids. A root intervention refreshes its grid index independently and rerenders all descendants; it does not hold the shadow fixed. CausalVAE's four endogenous concepts are therefore not asserted to form a product reference law. The protected outputs in this evaluation correspond only to the two roots.

The finite-error theory uses this finite product distribution directly. Rasterization need not be a smooth embedding, and neither the renderer nor the learned map is asserted to satisfy Theorem 1. The deterministic posterior means form square-integrable functions on the finite grid, which suffices for the energy identities and certificates. Their exact conditional variances are obtained by averaging along grid rows or columns.

E.2 Image-only training and independent output assignment

The three learners are a deterministic AE, a Gaussian VAE, and a β-VAE with $\beta = 4$ (Kingma & Welling, 2014; Higgins et al., 2017). Each seed randomly partitions the 7308 images into 5846 training and 1462 test images. The partition is shared across learner types within that seed. Images are flattened and scaled from uint8 to $[0, 1]$. The encoder has widths $3072 \to 128 \to 64 \to 2$ for AE, and $3072 \to 128 \to 64 \to 4$ for the variational models, whose final outputs contain two means and two log variances. Hidden activations are ReLU. The decoder is $2 \to 64 \to 128 \to 3072$, with ReLU hidden layers and a sigmoid output.

The optimized objective is mean per-image summed squared pixel error plus β times the analytic Gaussian KL divergence to $\mathcal{N}(0, I_2)$, with $\beta = 0$ for AE and $\beta = 1$ for VAE. Variational training samples a diagonal Gaussian using log variances clipped to $[-12, 8]$. Adam uses learning rate 10^{-3}, batch size 128, and exactly 2000 updates. No factor labels, target masks, invariance terms, validation certificates, or oracle errors enter optimization. The final checkpoint is used without hyperparameter search. At evaluation, only the raw deterministic encoder means are used: there is no posterior sampling, saturating readout, or coordinatewise variance normalization.

Seeds are $s \in \{0, 1, 2\}$. Initialization, the image split, and training sampling use respective seeds $4100000 + s$, $4200000 + s$, and $4300000 + s$. Training uses PyTorch on CPU with one thread per process. The complete output table is evaluated after training. A separate uniform-with-replacement pilot of 4096 grid configurations, seeded at $4500000 + s$, computes a 2×2 mutual-information matrix

using 20 equal-width code bins. Hungarian matching fixes a one-to-one output/root assignment. This pilot supplies evaluation labels only after learning; it does not fit or rescale the encoder. All subsequent oracle and certificate calculations use that same assignment.

E.3 Exact envelopes, budget accounting, and repeated validation

Write the assigned raw output table as z_{abj}, with $a \in [84]$, $b \in [87]$, and $j \in \{1, 2\}$. The exact reference errors are

$$R_1 = \frac{1}{84} \sum_a \mathrm{Var}_b(z_{ab1}), \qquad R_2 = \frac{1}{87} \sum_b \mathrm{Var}_a(z_{ab2}), \tag{44}$$

where each variance uses the population denominator for its uniform finite grid. The output-variance caps are the variances over all 7308 configurations. The global envelopes are half the squared full-table coordinate ranges. The fiber envelopes are

$$M_1^{\mathrm{fib}} = \tfrac{1}{2} \max_a (\max_b z_{ab1} - \min_b z_{ab1})^2, \qquad M_2^{\mathrm{fib}} = \tfrac{1}{2} \max_b (\max_a z_{ab2} - \min_a z_{ab2})^2. \tag{45}$$

They are rounded outward with the next representable floating-point value. All claims concern this evaluated numerical output table; interval arithmetic over the network's real-valued implementation is not claimed.

Each checkpoint has 100 independent validation repetitions, with seeds $4700000 + 1000s + r$, $r = 0, \ldots, 99$. A repetition independently samples a starting grid configuration and refreshes the nuisance root with replacement. The two actions receive 2048 or 32,768 pairs each; the smaller sample is a prefix of the larger. Every comparison uses the same pairs and $\eta = 0.05/2$ for each of the two tested cells. Global/fiber envelopes and EB/PrPl give four predetermined procedures, not an adaptive minimum over four bounds. Each output bound is capped by its exact reference variance. The reported containment event requires both output errors to be below their bounds.

The complete grid audit costs 7308 encoder evaluations per model; pilot and validation then use table lookups. The 4096 pilot lookups and the 4096 or 65,536 validation-pair lookups are listed separately, rather than calling these fresh image forward passes. Exhaustive enumeration already permits exact error computation, so this experiment is a finite-support validity and transfer audit, not a claim that sampled certificates beat enumeration in total cost. It isolates whether a published visual generator, raw output scale, and different training objectives change the certificate's behavior. For larger benchmarks, obtaining a comparably tight verified envelope remains a separate computational problem.

Numerical containment comparisons allow an absolute tolerance of 10^{-12}; support envelopes are rounded upward by one floating-point step. These checks concern the stored deterministic output table, not an interval-arithmetic verification of the real-valued network.

Training-seed means are the independent units for the uncertainties in Table 3. For a ratio, we first divide by that checkpoint's exact error, average its 100 validation outcomes, and then compute mean and SE across three checkpoints. This is not a ratio of pooled errors and bounds. At the larger budget, the three VAE containment failures under PrPl with fiber ranges are retained. No checkpoint or repetition is discarded for collapse, large leakage, or certificate looseness.

E.4 Relation to IRS, DCI, MIG, and information retention

For any nonconstant scalar output, the law of total variance gives

$$\frac{\mathcal{R}_j}{\mathrm{Var}_Q(F_j)} = 1 - \frac{\mathrm{Var}_Q(\mathbb{E}_Q[F_j \mid C_j])}{\mathrm{Var}_Q(F_j)}. \tag{46}$$

The right side is the complement of a first-order variance-explained ratio. Under the product reference, it is also the sum of all normalized variance components not confined to the assigned factor. This relates the target to Sobol' sensitivity analysis (Owen, 2012); it does not make it a new normalized disentanglement metric. The main guarantee bounds the numerator on a fixed output scale. Dividing by an estimated variance without a valid lower bound would not yield a certified normalized score.

IRS explicitly evaluates interventional robustness (Suter et al., 2019). Our implementation uses its worst-case version, rather than a high-quantile approximation: for each fixed factor value, compute the output's conditional mean and its maximum absolute deviation across the other root. Average these maxima over the fixed factor, normalize by the maximum absolute deviation from the global output mean, and take the best factor for each coordinate. The aggregate IRS weights coordinates by those global maximum deviations. Unlike our squared average residual, this score uses absolute worst-case deviations and a normalization; neither numerical quantity determines the other without further assumptions. Factor-wise maximizing IRS also does not change the certificate's independently fixed one-to-one assignment.

DCI uses a supervised predictive-importance matrix to distinguish how exclusively coordinates encode factors, how concentrated each factor's representation is, and how much factor information is retained (Eastwood & Williams, 2018). We use the original framework's regression interpretation with 100-tree random forests, rather than asserting equivalence to a particular library's boosted-classifier implementation. A separate forest predicts each standardized root index from the two raw codes on the image-training indices and is evaluated on the held-out image-test indices. Entropies of impurity-based importances, normalized in base two, give disentanglement and completeness; held-out R^2 measures informativeness. Forests have no imposed maximum depth and use seed $4600000 + s$.

MIG uses 20 equal-width bins for each code and the actual discrete root labels. For each root, subtract the second-largest mutual information from the largest, divide by the root entropy, and average over the two roots (Chen et al., 2018). IRS and MIG are computed on the full finite grid; DCI's prediction error uses held-out images. These are point-estimate evaluation metrics, not upper-confidence certificates. They accompany raw leakage and reconstruction rather than serving as interchangeable baselines whose numerical values should be ranked against a leakage bound. The complementary Pendulum measurements are reported in Table 19.

Table 19: Complementary Pendulum measurements on the same raw representations, mean $\pm$ SE over three training seeds. DCI uses random-forest regression; R^2 is its held-out informativeness endpoint. Pixel MSE uses the decoder at the deterministic encoder mean. Raw leakage and raw variance are scale-dependent and are not cross-learner rankings of disentanglement.

Learner	IRS	DCI disent.	DCI complete	MIG	Held-out R^2
AE	0.720 ± 0.076	0.293 ± 0.064	0.379 ± 0.049	0.144 ± 0.028	0.866 ± 0.064
VAE	0.731 ± 0.081	0.526 ± 0.174	0.531 ± 0.172	0.324 ± 0.065	0.984 ± 0.004
β-VAE	0.831 ± 0.109	0.410 ± 0.254	0.507 ± 0.212	0.222 ± 0.140	0.915 ± 0.070

Learner	Raw leakage	Mean variance	Pixel MSE	Mean $M^{\mathrm{fib}}/M^{\mathrm{glob}}$
AE	21.553 ± 6.764	51.060 ± 9.731	0.00612 ± 0.00206	0.239 ± 0.070
VAE	0.689 ± 0.289	1.741 ± 0.095	0.00237 ± 0.00028	0.238 ± 0.051
β-VAE	0.311 ± 0.301	0.944 ± 0.193	0.00483 ± 0.00171	0.306 ± 0.140

The variation across the three β-VAE runs is substantial. Small leakage can accompany a weakly informative coordinate, so a tight certificate alone is not evidence of a good representation. Table 20 exposes all seeds, output variances, errors, and certified ratios rather than reporting only a favorable aggregate.

The encoders in this section are fitted here with standard image-only objectives. They are not author-released CITRIS, iCITRIS, mechanism-sparsity, or CausalVAE checkpoints, and their results do not reproduce those methods' reported task scores. The public generator and independently trained learner families test the applicability claim under a declared finite product reference; they do not establish scalability to arbitrary visual worlds or resolve the separate convolutional failure in Appendix F.

Table 20: Every Pendulum training seed, with the fixed assignment. v_1, v_2 are exact finite-reference raw variances; R_1, R_2 are exact assigned off-factor errors. The last column is the mean PrPl+fiber certificate/error ratio over 100 validation repetitions at $n_a = 2048$.

Learner	Seed	v_1	v_2	R_1	R_2	Ratio
AE	0	21.07433	42.15051	6.93348	3.81862	1.102
AE	1	85.41150	34.84162	0.91896	33.08865	1.052
AE	2	53.70696	69.17525	6.97013	12.93014	1.084
VAE	0	2.01083	1.84838	0.12408	0.89938	1.119
VAE	1	1.46401	1.85399	0.03492	0.89539	1.105
VAE	2	2.10167	1.16626	0.07621	0.03712	1.320
β-VAE	0	1.10741	1.16905	0.00292	0.01508	1.166
β-VAE	1	1.11691	0.00025	0.00155	0.00024	1.143
β-VAE	2	1.24854	1.02037	0.29978	0.61288	1.108

Table 21: RGB representation recovery. Recovery is Pearson MCC, mean $\pm$ SE. Paired t statistics use matched seed outcomes. Every setting favors the observational baseline.

Setting	Conditioned	Observational	Paired t
Five seeds	0.323 ± 0.035	0.443 ± 0.037	-2.15
Sixteen seeds	0.333 ± 0.016	0.457 ± 0.019	-4.20
Five-seed ablation	0.304 ± 0.013	0.483 ± 0.034	-4.10

F VISUAL SCOPE

The RGB measurements below delimit claims about visual representation learning. They concern a different training pipeline from the controlled certificate experiments and the public Pendulum-generator audit; their reported seed counts and uncertainty are preserved.

Rendered scenes and the unresolved visual failure The procedural scenes vary object position, size, hue, value, and background. A convolutional encoder and a target-conditioned predictor are trained using the stated predictive objective. The five-seed setting uses 32×32 RGB observations, 3000 trajectories of length 16, 300 epochs, and fp32 training. The separately reported sixteen-seed setting and five-seed ablation retain their own measurement records and are not pooled into a single experiment. The RGB representation recovery results are reported in Table 21.

The corresponding full-minus-worst coverage gaps are 0.031, 0.024, and 0.012. In the five-seed ablation, target input without the hard prior reaches MCC 0.426, below the observational 0.483, with paired $t = -1.85$. The best soft-prior arm reaches 0.391. Thus hard-prior misspecification alone does not explain the full performance gap. In the other five-seed setting, the best soft prior reaches 0.350, and transfer R^2 is 0.257 for the conditioned model versus 0.477 for the observational model.

Held-out ridge decoding of the raw rendered pixels gives factor R^2 between approximately 0.27 and 0.65; the recorded minimum in the five-seed setting is 0.265. This demonstrates predictive information in the observations, not global renderer injectivity, an inverse for the learned encoder, or a diagnosis of optimization failure. The numerical results therefore remain a limitation on operational visual recovery. They do not contradict the conditional theorems, and the theorems do not explain them away.

G SUPPLEMENTARY FIGURES

Figure 1 illustrates the frozen neural-encoder results across the five tested designs.

Figure 2 shows the certificate-to-exact-error ratios across learner architectures, sample sizes, and deterministic envelopes.

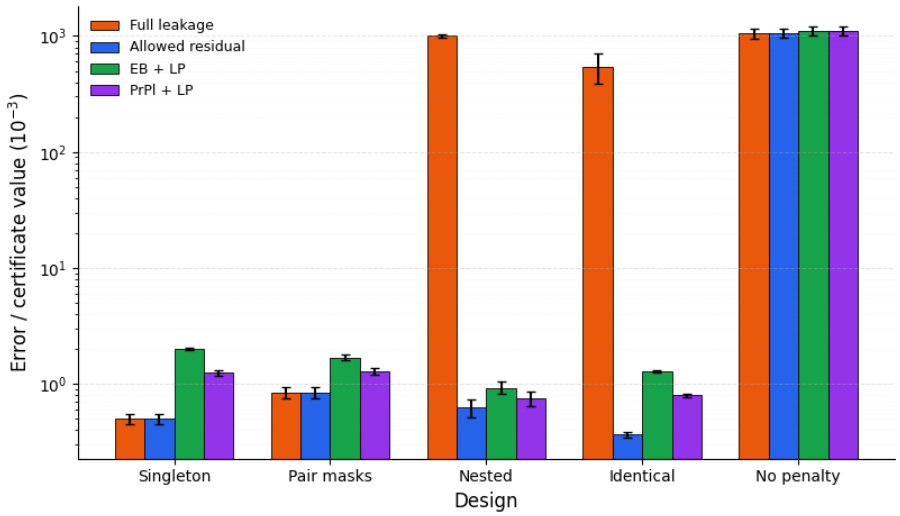


Figure 1: Frozen neural-encoder results across five training seeds. Bars report mean error or certificate values and error bars denote ±SE; all quantities are scaled by 10^{-3}. The logarithmic ordinate preserves visibility across the substantially different leakage scales induced by the tested designs. EB and PrPl use identical samples and $\delta = 0.05$. The separation-dependent distinction between full leakage and allowed residual is explicit in the Nested and Identical designs.

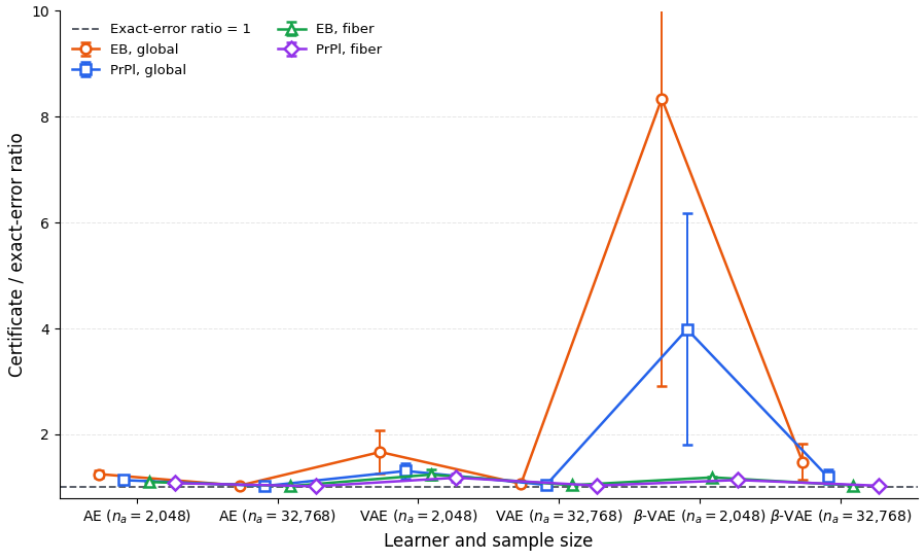


Figure 2: Certificate-to-exact-error ratios for the pendulum root-grid evaluation. Points report means across three training seeds after 100 validation repetitions per seed, with error bars denoting ±SE. The horizontal reference at 1 corresponds to exact agreement between the certificate and exact error. Global and fiber envelopes are shown separately for EB and PrPl, with $n_a \in \{2{,}048, 32{,}768\}$.

Figure 3 characterizes the dependence of certificate-to-exact-variation ratios on the number of factors K across the four certificate constructions.

Figure 4 summarizes the certificate-to-error ratios across increasing factor dimensionality under the large-mixing control.

Figure 5 quantifies how certificate-to-exact-error ratios vary with the third-sample budget across the tested values of K.

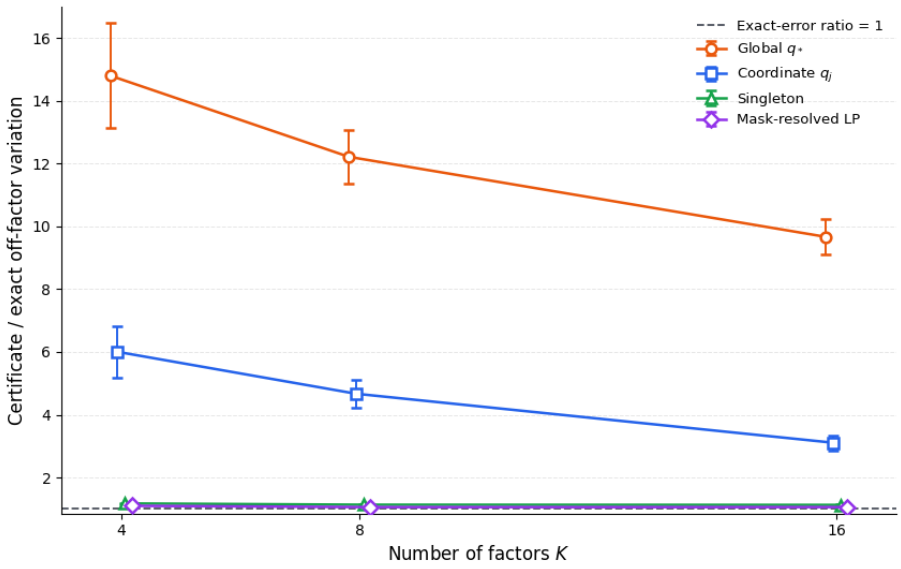


Figure 3: Certificate-to-exact off-factor variation ratios across $K \in \{4, 8, 16\}$. Points report means over 100 independent encoder-design draws per dimension, with error bars denoting $\pm$SE. The horizontal reference at 1 indicates exact agreement between the certificate and the measured off-factor variation. Every certificate is capped by the known output variance, while Singleton uses the directly observed singleton energies.

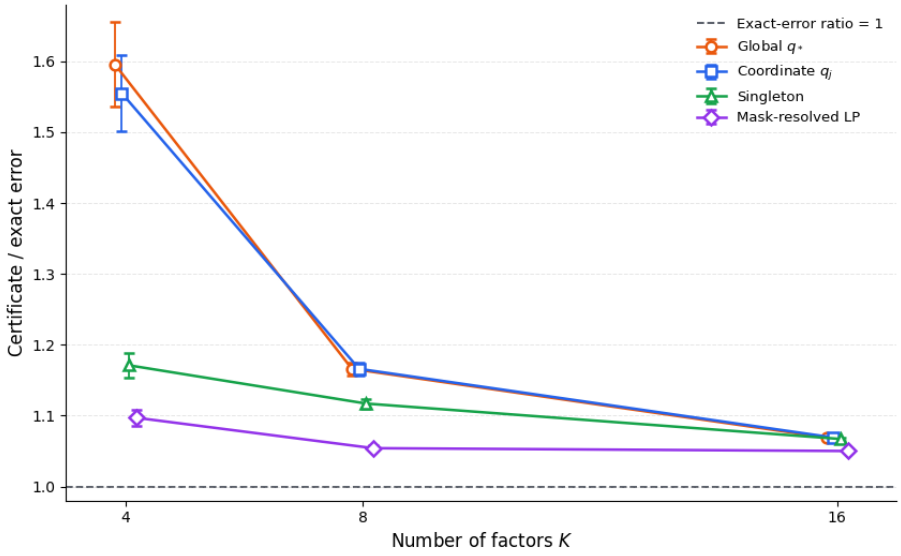


Figure 4: Variance-capped certificate-to-error ratios under the large-mixing control. Points report means with error bars denoting $\pm$SE across the corresponding encoder-design draws. The horizontal reference at 1 denotes exact agreement between the certificate and measured error. All four constructions use the same action families and seeds, isolating the effect of the large-mixing regime as K increases.

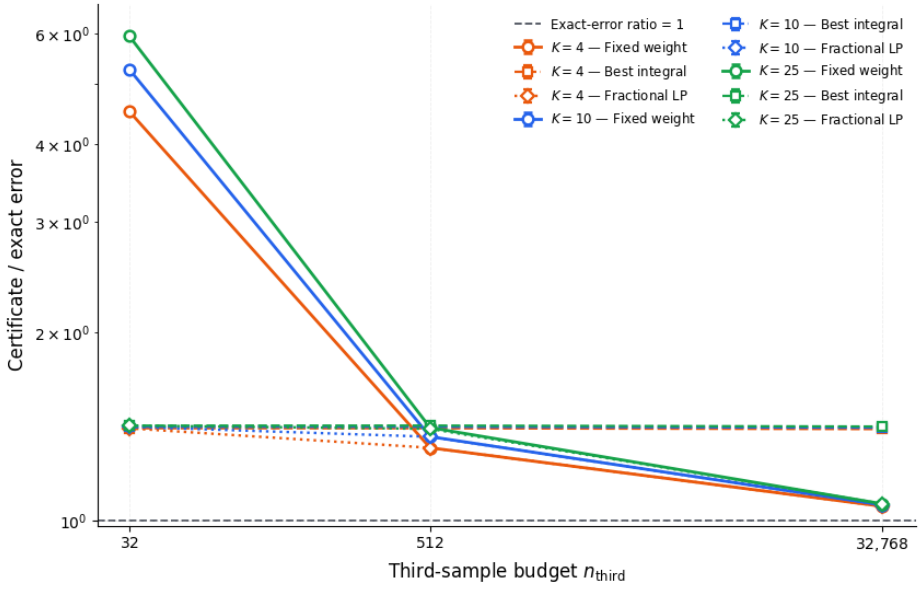


Figure 5: Certificate-to-exact-error ratios across the complete overlapping-only sample-budget sweep. Points report means over 100 independent repetitions, with error bars denoting ±SE. Curves are grouped by $K \in \{4, 10, 25\}$ and distinguish Fixed weight, Best integral, and Fractional LP. The horizontal reference at 1 denotes exact agreement with the reference error. The plug-in containment rates are reported separately in Table 17.