WHICH TASKS IDENTIFY A REPRESENTATION? SPECIFICITY BOUNDS AND TASK ACQUISITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Task structure can rule out latent mixing without making a representation easy to estimate. We study this distinction for finitely many outcomes with known overlapping supports and injective conditional signatures. Building on existing support-separation and partial-view principles, we characterize allowed mixing and derive a fractional-cover bound on block specificity. The bound combines task conditioning with signature error, avoids repeatedly charging the same excluded directions, and is optimal given its local variance constraints. A smooth signature example attains its constant and improves the coordinatewise bound by a factor $2(B-1)$ for B blocks. We then evaluate acquisition at a fixed sampled-label budget. In a Gaussian-signature comparison, a pilot-conditioned policy improves shifted accuracy over uniform candidate acquisition, and an information-matched inner-Lasso adaptation nearly matches the proposed spectral estimator. An invertible nonlinear model gives a more cautious result: a frozen 20-seed follow-up estimates a 2.52-point shifted-accuracy gain, with a 95% interval spanning zero. A separate 240-run composite-image benchmark exposes further failures outside the theorem. Together, the results distinguish structural separation, conditioning, and finite-budget utility, with released implementations, complete negative results, and explicit limits on practical certification.

1 INTRODUCTION

A representation can predict its training outcomes while mixing the factors those outcomes are meant to distinguish. This matters when a scientist wants to acquire another measurement, interpret a block, or reuse it after unrelated factors change. Adding a task may remove a population ambiguity yet provide little usable information at the available sample size. Without appropriate inductive biases, disentangled recovery is impossible (Locatello et al., 2019); even predictive function-space identifiability can leave a linear ambiguity (Roeder et al., 2021). We ask: *how do support overlap and task conditioning jointly constrain block specificity, and when does acquiring a separating task help at a fixed label budget?*

We consider known task supports. A sufficiently rich outcome distribution acts as a partial view of its supported latent coordinates. Their intersections constrain equivalent representations, but exact separation is only the first step. Near-insensitive task directions amplify estimation error, and overlapping tasks can provide redundant control of the same nuisance variation. A useful quantitative analysis should account for both effects.

Our main technical result aggregates task-level inverse-conditioned errors through a fractional-cover program. Its objective bounds variance in a learned block attributable to other factors; its LP dual gives a matching construction for the local-variance relaxation. A singleton-task example attains the signature-level constant and shows a dimension-dependent improvement over coordinatewise error accumulation. Its dual assigns worst-case variance to nuisance coordinates; its primal identifies which overlapping task constraints control that variance. This provides an auditable sensitivity calculation rather than a label-only confidence certificate. We then test task acquisition with fixed training-label budgets and compare an adapted sparse multitask objective, including a control with identical support information. Controlled nonlinear and image experiments retain failures that separate identifiability from useful finite-sample estimation.

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.

What is inherited, and what is added. Sparse multitask learning already identifies representations under task variability and sparsity (Lachapelle et al., 2023). Its Assumption 3.6 is equivalent to the antichain support condition used here. Partial-view identification supplies the underlying intersection principle (Yao et al., 2024); multimodal identifiability studies shared content under related view assumptions (Daunhawer et al., 2023; von Kügelgen et al., 2021). Diverse dictionary learning identifies latent set operations and dependency structure under sufficient nonlinearity and a sparsity constraint (Zheng et al., 2026). Near-identifiability theory quantifies ambiguity using bi-Lipschitz structure (Nelson et al., 2026). We do not claim these principles as new. Our focused addition is the task-exclusion specificity program, its optimal variance relaxation and sharp signature example, plus budget-matched acquisition evidence. The variance decomposition, LP duality, Sperner bound and greedy set cover are classical tools (Efron & Stein, 1981; Madiman & Barron, 2007; Sperner, 1928; Chvátal, 1979). The distinction is quantitative: given task errors and inverse constants, which overlapping exclusions control residual nuisance variance most tightly? Unlike conditional-prior nonlinear ICA (Khemakhem et al., 2020), our assumptions concern outcome signatures on prescribed supports. We neither discover those supports nor identify semantics from arbitrary labels.

2 Setup: probability signatures and equivalence

Let $Z = (Z_1, \dots, Z_B)$, with $B \geq 2$ and $Z_b \in \Omega_b \subseteq \mathbb{R}^{d_b}$. The support $\Omega = \prod_b \Omega_b$ is a product of connected open sets with positive density throughout. Dependence between blocks is allowed. Observations satisfy $X = f(Z)$ for a diffeomorphism onto its image. This includes an already recovered invertible content representation; it does not include arbitrary observation noise or a guarantee that raw multimodal encoders recover such content.

Task t has K_t classes and a known nonempty support $S_t \subseteq [B]$:

$$q_t(z) = \mathbb{P}(Y_t = \cdot \mid Z = z) = g_t(z_{S_t}). \tag{1}$$

Both g_t and the fitted map $\widetilde{g}_t$ are smooth embeddings: they are globally injective, have full column rank, and have smooth inverses on their images. An equivalent representation is a diffeomorphism $\widetilde{Z} = h(Z)$ with the same named block dimensions and supports such that

$$\widetilde{g}_t(h_{S_t}(z)) = g_t(z_{S_t}) \qquad \text{for every } t, z. \tag{2}$$

Block identification means every admissible equivalent h has $h_b(z) = \phi_b(z_b)$ for within-block diffeomorphisms ϕ_b. The equivalence concerns all task conditional distributions. Where a joint-label construction is required, tasks may be sampled conditionally independently given Z, yielding the same complete observable law in the witnesses below.

What a finite label supplies. The theorem concerns the conditional distribution, not one observed label. Estimating it from sampled outcomes is a separate statistical problem. Write $d(S) = \sum_{b \in S} d_b$. The necessary dimension budget is $K_t - 1 \geq d(S_t)$; this is not sufficient for embedding. A proportional-odds head factors through one scalar score and has Jacobian rank at most one, regardless of its number of categories. Thus a two-dimensional supported block cannot be identified through such a head under our theorem.

Other conditional signatures. The proofs use an injective signature of the conditional distribution, not categorical outcomes specifically. For Gaussian outcomes with fixed covariance and an invertible supported mean map, the mean is such a signature. Our linear comparison uses this extension and counts each scalar outcome in its budget; the finite-class dimension condition does not apply to that continuous setting.

3 The ambiguity allowed by a finite task pattern

Define block b's incidence signature and its allowed dependency set by

$$C_b = \{t : b \in S_t\}, \qquad A_b = \{c : C_b \subseteq C_c\}. \tag{3}$$

If C_b is nonempty, $A_b = \bigcap_{t \in C_b} S_t$. If b is never observed by a task, $A_b = [B]$. Two distinct signatures need not isolate their blocks: for example $\{1\} \subset \{1, 2\}$ still permits mixing in one direction.

Theorem 1 (Allowed dependence and sharp witnesses)**.** *Under Section 2, every equivalent representation satisfies*

$$h_b(z) = H_b(z_{A_b}). \tag{4}$$

Conversely, on $\Omega_b = \mathbb{R}^{d_b}$, every permitted pair $c \in A_b \setminus \{b\}$ admits a support-preserving non-block equivalence. In particular, for any nonzero linear map $L : \mathbb{R}^{d_c} \to \mathbb{R}^{d_b}$, the shear

$$h_b(z) = z_b + Lz_c, \qquad h_j(z) = z_j \quad (j \neq b) \tag{5}$$

is an admissible alternative after transforming the observation decoder, task heads, and latent density.

Proof. For each $t \in C_b$, invert the fitted embedding in Eq. (2): $h_{S_t}(z) = \widetilde{g}_t^{-1} \circ g_t(z_{S_t})$. Thus h_b depends only on every such S_t. Any excluded coordinate has zero derivative; the connected product fibers imply dependence only on their intersection. For an uncovered block the conclusion is vacuous. For the converse, every task containing b also contains c. The shear therefore restricts to an invertible map on every task support. Set $\widetilde{g}_t = g_t \circ (h_{S_t})^{-1}$, replace f by $f \circ h^{-1}$, and push forward the density. Embeddings, full Euclidean support, and all observable probabilities are preserved, while h_b depends on z_c. □

Corollary 2 (Universal identification)**.** *A finite task pattern guarantees block identification for all models in Section 2 exactly when every C_b is nonempty and*

$$C_b \not\subseteq C_c \qquad \textit{for all } b \neq c. \tag{6}$$

Sufficiency holds on the stated product domains. Necessity for a universal guarantee already holds within the full-Euclidean subclass with unrestricted positive latent densities.

Proof. Under Eq. (6), $A_b = \{b\}$, so h is blockwise. Its nonsingular Jacobian and global injectivity imply that each ϕ_b is a diffeomorphism onto its image. If coverage or Eq. (6) fails, Theorem 1 supplies a mixed equivalent model; an uncovered block can absorb any other block. □

The characterization is universal over the stated model class. Restrictions on a particular model can eliminate further ambiguities; the necessary derivative pattern does not make every such map globally admissible.

Two related outcomes. For $S_D = \{s, d\}$ and $S_P = \{s, p\}$, $C_d \subset C_s$ and $C_p \subset C_s$. Rich heads isolate the shared block but permit $\widetilde{d} = d + as$ and $\widetilde{p} = p + bs$. The additional support $\{d, p\}$ produces three incomparable signatures and identifies all blocks. If the two original heads are scalar ordinal models, the richer theorem does not apply even to the shared block. Supplement Section S2 gives a stronger Gaussian counterexample preserving independent sources and both ordinal scores.

4 Approximate signatures and conditioning

Exact probability matching idealizes statistical estimation. To quantify deviations, let r_t be an injective signature of q_t, and $\widetilde{r}_t$ the corresponding fitted signature. This may be the probability vector itself or, for positive probabilities, the vector of reference-class log odds. Assume the inverse of $\widetilde{r}_t$ on its image is L_t-Lipschitz. Define

$$e_t(z) = \widetilde{r}_t(h_{S_t}(z)) - r_t(z_{S_t}), \qquad T_{bc} = \{t : b \in S_t,\ c \notin S_t\}. \tag{7}$$

In this section h need not satisfy exact equivalence; we retain the supports and controlled fitted inverses. The quantities below are population signature errors, not directly observed label residuals.

Theorem 3 (Approximate block specificity)**.** *Suppose Eq.* (6) *holds. If $\sup_z \|e_t(z)\| \leq \delta_t$, then for any fixed anchor $a \in \Omega$ and $\phi_b(z_b) = h_b(z_b, a_{-b})$,*

$$\sup_z \|h_b(z) - \phi_b(z_b)\| \leq 2 \sum_{c \neq b} \min_{t \in T_{bc}} L_t \delta_t. \tag{8}$$

If instead the blocks are independent, $h_b(Z)$ is square integrable, and $\epsilon_t^2 = \mathbb{E}\|e_t(Z)\|^2$, then

$$\mathbb{E}\|h_b(Z) - \mathbb{E}[h_b(Z) \mid Z_b]\|^2 \leq 2 \sum_{c \neq b} \min_{t \in T_{bc}} L_t^2 \epsilon_t^2. \tag{9}$$

The approximating block functions need not be invertible; these are specificity bounds, not approximate diffeomorphism guarantees.

Proof sketch. If z, z' differ only in block c, choose $t \in T_{bc}$. Their true task signatures agree. Inverse Lipschitzness bounds $\|h_b(z) - h_b(z')\|$ by $L_t(\|e_t(z)\| + \|e_t(z')\|)$. Changing other blocks one at a time proves Eq. (8). For Eq. (9), independently resample each Z_c and apply the conditional Efron–Stein inequality (Efron & Stein, 1981). Full details and a conditioning example are in Supplement Section S3. □

Why accurate prediction is insufficient. A small signature error only controls mixing when the inverse constants are controlled. A head may become nearly insensitive to a supported direction, making its inverse unstable. Furthermore, held-out cross-entropy is not itself an estimate of error relative to the unknown true conditional probabilities. Supplement Section S3 connects excess log loss to Eq. (9) under an explicit probability floor. We do not turn an observed point estimate or an empirical eigen-gap into a distribution-free identification certificate.

4.1 A SHARP VARIANCE RELAXATION FOR OVERLAPPING TASKS

The coordinatewise bound above can charge the same task many times. We instead exploit every set of coordinates that a task excludes. For a target block b, let $\mathcal{T}_b = \{t : b \in S_t\}$ and $a_t = L_t^2 \epsilon_t^2$. Define

$$\mathcal{C}_b(a) = \min_{w \geq 0} \left\{ \sum_{t \in \mathcal{T}_b} w_t a_t : \sum_{t \in T_{bc}} w_t \geq 1 \text{ for each } c \neq b \right\}, \tag{10}$$

with value $+\infty$ when infeasible. These weights cover excluded coordinates, not observed variables; one task may cover several coordinates simultaneously.

Theorem 4 (Fractional-cover specificity)**.** *Under the independent-block, square-integrability and inverse-Lipschitz assumptions of Theorem 3,*

$$\mathbb{E}\|h_b - \mathbb{E}[h_b \mid Z_b]\|^2 \leq \mathcal{C}_b(a). \tag{11}$$

Furthermore, $\mathcal{C}_b(a)$ is the smallest universal upper bound obtainable from only the inequalities $\mathbb{E}\operatorname{Var}(H \mid Z_{S_t}) \leq a_t$, for scalar square-integrable H on independent standard-normal blocks. For these variance constraints, equality is attained by an additive H. This optimality claim concerns the variance relaxation, not attainability of every error vector by a common fitted task model.

Proof sketch. Conditionally resample the complement of S_t. True signatures agree, so the inverse bound gives $\mathbb{E}\operatorname{Var}(h_b \mid Z_{S_t}) \leq L_t^2 \mathbb{E}\operatorname{Var}(e_t \mid Z_{S_t}) \leq a_t$. The orthogonal functional-ANOVA decomposition assigns nonnegative variance to each coordinate subset. Every component involving a coordinate outside b is charged at least once by a feasible fractional cover, proving the bound. LP duality gives $\mathcal{C}_b(a) = \max_{u \geq 0}\{\sum_{c \neq b} u_c : \sum_{c \notin S_t} u_c \leq a_t\}$. For independent standard normals, $H = Z_b + \sum_{c \neq b} \sqrt{u_c} Z_c$ attains these local variances and total nuisance variance. Supplement S12 supplies the full proof. □

A singleton task $S_t = \{b\}$ gives bound a_t, while Eq. (9) gives $2(B-1)a_t$. The constant one is sharp even for smooth invertible representations: use the additive shear above, true signature βZ_b, and fitted signature βh_b. This yields $L_t = 1/\beta$ and equality. Thus the refinement can matter increasingly with dimension; it does not remove the need to estimate signature error and conditioning.

Functional-ANOVA variance decompositions and fractional-cover tools are classical (Efron & Stein, 1981; Madiman & Barron, 2007). Our result applies them to task-conditioned representation specificity, includes the optimal variance relaxation and explicit signature-level equality case, and exposes a design quantity beyond binary separability. It is not a claim to invent fractional covers. A label-only finite-sample certificate remains outside the result.

Why fractional aggregation matters. Consider three tasks containing target b, each excluding a different pair among three nuisance coordinates, with all $a_t = 1$. Half weight on each task covers every nuisance coordinate, giving $\mathcal{C}_b = 3/2$; no integral cover costs below two. The coordinatewise bound in Eq. (9) is six. Additional tasks may isolate the other blocks without changing this target calculation. The dual assigns $u_c = 1/2$ to each nuisance coordinate, certifying optimality. This separates the improvement from merely dropping the earlier factor two: task overlap itself reduces the worst-case variance bound. The released LP solver returns both the task weights and nuisance-coordinate dual weights, allowing the calculation to be audited for any supplied support pattern.

4.2 A CONTROLLED EXTENSION BEYOND INDEPENDENT FACTORS

Independence matters for aggregating local constraints. Let P be the latent law and Q a product reference law on the same support, with $0 < m \le dP/dQ \le M < \infty$. Define $a_t^P = L_t^2 \mathbb{E}_P \|e_t\|^2$.

Corollary 5 (Bounded density-ratio dependence)**.** *Under the pointwise signature assumptions and square integrability,*

$$\mathbb{E}_P \|h_b - \mathbb{E}_P[h_b \mid Z_b]\|^2 \le (M/m)\mathcal{C}_b(a^P). \tag{12}$$

Proof. The conditional mean minimizes squared error, so the left side is at most $\mathbb{E}_P \|h_b - \mathbb{E}_Q[h_b \mid Z_b]\|^2 \le M \mathbb{E}_Q \|h_b - \mathbb{E}_Q[h_b \mid Z_b]\|^2$. Apply Theorem 4 under Q, use $\mathbb{E}_Q\|e_t\|^2 \le \mathbb{E}_P\|e_t\|^2/m$, and use positive homogeneity of the covering LP. □

This is a standard change-of-measure extension, not a claim that arbitrary dependence is harmless. The ratio must be justified; a correlated unbounded Gaussian generally has no such finite global sandwich against its product marginals.

A failure witness for ignoring dependence. On $(-1, 1)^3$, let Q be uniform product and $dP/dQ = 1 + \rho \sin(\pi z_1)\sin(\pi z_2)$, $0 < \rho < 1$. Use the three pair supports and the invertible shear $h_0 = z_0 + \alpha z_1$, $h_1 = z_1$, $h_2 = z_2$, with $\alpha = .01$. The $(0, 1)$ and $(1, 2)$ fitted signatures undo this shear exactly. The $(0, 2)$ fitted signature is $(h_0 - \alpha\rho \sin(\pi h_2)/\pi, h_2)$, while its true signature is (z_0, z_2). Its inverse constant is $L = (\sqrt{4 + (\alpha\rho)^2} + \alpha\rho)/2$. The true target specificity is $\alpha^2/3$, whereas applying the independent formula would give $L^2\alpha^2[1/3 - \rho^2/(2\pi^2)]$. At $\rho = .9$ this is only 0.885 times the actual variance and therefore invalid. The density-ratio correction uses $(M/m) = (1 + \rho)/(1 - \rho)$ and restores validity, conservatively. All signatures are smooth embeddings and can be realized as three-class log odds. Supplement S12 verifies the conditional moments. This example isolates dependence as a genuine failure mode even when supports separate and fitted heads are invertible.

5 DESIGNING AND ESTIMATING AN IDENTIFIABLE REPRESENTATION

5.1 TASK DESIGN UNDER SUPPORT AND RICHNESS CONSTRAINTS

For unrestricted rich tasks, Sperner's theorem (Sperner, 1928) gives

$$T_{\min}(B) = \min\{T : \tbinom{T}{\lfloor T/2 \rfloor} \ge B\}. \tag{13}$$

Assign distinct middle-layer incidence signatures to blocks. This minimizes task count, not annotation cost: each chosen support must still satisfy its outcome-dimension budget. Supplement S16 derives the count and the different tradeoff when every support has at most two scalar blocks.

For an existing pattern, define unresolved ordered pairs $U = \{(b, c) : b \ne c,\ C_b \subseteq C_c\}$. Add a coverage token for each block with $C_b = \varnothing$. A candidate support S covers (b, c) exactly when $b \in S, c \notin S$, and covers the token for b when $b \in S$. Discard candidates that fail the dimension budget. Covering all remaining tokens is equivalent to repairing the structural condition. With fixed positive costs, selecting maximum uncovered tokens per cost is classical weighted greedy set cover. If a feasible cover exists, its cost is at most $H_{|U|}$ times optimum, counting coverage tokens in $|U|$ (Chvátal, 1979). This inherited guarantee concerns known supports and rich tasks, not discovery of semantic dependencies.

5.2 PATI ESTIMATORS

Linear signatures. For $X = AZ$ and multinomial-logit task heads, centered logit coefficients have row space equal to the span of the corresponding rows of A^{-1} when each supported head has full rank. Let P_t project onto that space. For block b, the kernel of

$$M_b = \sum_{t \in C_b} (I - P_t) \tag{14}$$

is the intersection of the task subspaces. Under Eq. (6), its dimension is d_b. PATI estimates the heads from training labels and takes the bottom d_b eigenvectors of $\widehat{M}_b$. Finite-sample rank selection requires additional care when $K_t - 1 > d(S_t)$; the reported experiments use equality. Supplement Section S5 gives the elementary deterministic perturbation bound.

Nonlinear signatures. PATI-Flow uses an invertible matrix-exponential layer and affine coupling layers. A head on $u \in \mathbb{R}^{d(S_t)}$ has reference logit zero and core logits $\exp(B_t)u + b_t$. Their log odds recover u uniquely. The inverse log-odds constant is $\| \exp(-B_t) \|_{\mathrm{op}}$, making the conditioning term in Section 4 explicit for this fitted family. Training minimizes summed task cross-entropy. This enforces invertibility and rich fitted heads, but not correct supports, true-head richness, or convergence to a population optimum. The implementation takes square content vectors; it is not a raw-interview encoder.

6 EXPERIMENTS: BUDGET, INFORMATION, AND UTILITY

We distinguish theorem-aligned experiments from stress tests outside its assumptions. New categorical acquisition runs use the actual invertible PATI-Flow and injective fitted heads. Gaussian-signature runs allow a close regression-method comparison. Neither is a clinical experiment. All selection uses training/pilot/validation data; test factors are reserved for diagnostics. Supplement S13 records all settings, budgets and adaptation details.

6.1 FIXED-BUDGET GAUSSIAN ACQUISITION AND A CLOSE-METHOD COMPARISON

Three independent standard-normal factors are mixed by an orthogonal matrix. A two-output task has mean $\beta R Z_S$, with fixed invertible rotation R, plus independent Gaussian noise of SD .5. Start with supports $\{0, 1\}, \{0, 2\}$. Candidate measurements repeat $\{0, 1\}$, separate via $\{1, 2\}$ with $\beta = .1$, or separate with $\beta = 1$. Repeated measurements have fresh noise: they can improve estimation but do not repair the support pattern. Unlike deterministic image-pair labels, each new outcome is genuinely sampled.

Every allocation uses 1,200 or 4,800 scalar training labels and 600 validation labels, divided equally among its tasks. Acquiring a third task redistributes the budget rather than adding supervision. PATI estimates task row spaces by least squares and intersects them spectrally. On non-separating patterns its scalar choice within an unresolved subspace has no identification guarantee; these rows remain as controls.

We adapt the published inner-Lasso objective and active-support hypergradient (Lachapelle et al., 2023) to the same finite linear task family. The inner objective, relative Lasso penalty, stopped penalty gradient and no-inner/outer-split setting follow the reference code. Three penalty ratios and checkpoints are selected by validation MSE. A second version receives the same support masks as PATI. Independent Lasso-solver and finite-difference gradient checks pass; maximum recorded selected-fit KKT residual is below 5×10^{-15}. This is an objective-level adaptation, not a reproduction of the original architecture or benchmark. Its unknown-support population theorem's task-variability conditions are not asserted here. With strong separating tasks, support-informed inner-Lasso nearly matches PATI ($R^2 = 0.999$ each at 4,800 labels). Weak tasks retain the identifying pattern but worsen recovery; Figure 1 includes every allocation.

Acquiring from pilot labels. A common 600-scalar-label pilot budget measures all three candidates. Reject the non-separating candidate, then choose the largest estimated smallest singular value of its mean coefficient matrix. This simple inverse-conditioning proxy selects the strong candidate

Table 1: Gaussian-signature comparison, six seeds; mean ± SD. Training budgets count scalar labels. Inner-Lasso is an objective/solver adaptation, not the original benchmark reproduction. The masked variant supplies the same support masks as PATI.

Method	Supplied supports	Strong-task recovery R^2 1,200	Strong-task recovery R^2 4,800	Weak-task R^2 4,800	Strong shift 4,800
PATI spectral	Yes	0.998 ± 0.001	0.999 ± 0.000	0.963 ± 0.023	0.979 ± 0.004
Inner-Lasso	No	0.677 ± 0.078	0.678 ± 0.090	0.696 ± 0.060	0.650 ± 0.069
Inner-Lasso + masks	Yes	0.998 ± 0.002	0.999 ± 0.001	0.819 ± 0.193	0.973 ± 0.008

Table 2: Acquisition-policy shifted accuracy, six-seed mean ± SD. Total cost includes pilot, fitting and validation; common diagnostic probe supervision is additional. Uniform policies integrate candidate choices exactly.

Total cost	Estimator	Pilot-conditioned	Uniform separating	Uniform all
2400	PATI	0.972 ± 0.016	0.914 ± 0.052	0.864 ± 0.038
2400	Masked inner-Lasso	0.972 ± 0.012	0.882 ± 0.044	0.848 ± 0.050
6000	PATI	0.979 ± 0.004	0.939 ± 0.025	0.880 ± 0.032
6000	Masked inner-Lasso	0.973 ± 0.008	0.888 ± 0.050	0.855 ± 0.053

in all six seeds; it is not an estimated certificate for Theorem 4. Including pilot and validation, acquisition/fitting/validation costs are 2,400 or 6,000 scalars for every policy; an identical 500-example post-hoc factor-probe set is additional. At the larger budget, shifted accuracy is 0.979 for pilot-conditioned PATI acquisition, 0.939 for uniform choice between separating candidates, and 0.880 for uniform choice among all three. Uniform-policy values average all candidate outcomes per seed, exactly integrating the randomized choice. The support-informed Lasso shows the same direction: 0.973, 0.888 and 0.855. These results establish usefulness in this controlled measurement family, not optimal acquisition on arbitrary tasks.

6.2 Nonlinear acquisition with the actual identifiable architecture

We next use the invertible generator and PATI-Flow from Section 5. Three-class tasks have logits $(0, \beta Z_i, \beta Z_j)$. Each training event supplies exactly one sampled outcome. Compare two-task allocation, a fresh repeated-support task, weak separation ($\beta = .1$), and strong separation ($\beta = 1$). All configurations use the same total training budget, 600 validation labels, encoder initialization and minibatch seeds. Fitted heads remain injective; conditioning and convergence are not assumed favorable merely because the structure separates.

After six exploratory seeds at two budgets, we froze the protocol for 20 additional seeds at 4,800 labels. The primary comparison was strong separation versus repeated-support acquisition on shifted block-probe accuracy. Each frozen model receives 500 independent probe examples; shifts change only non-target latent means. Spline recovery complements linear probes because within-block equivalence may be nonlinear. Conditional-variance estimates directly measure residual nuisance dependence. The primary paired gain is 2.52 percentage points, with 95% Student-t interval $[-1.00, 6.03]$. It does not establish a reliable nonlinear benefit. Mean specificity improves from 0.129 to 0.078, but variability remains substantial. These follow-up results are kept separate from development seeds; no additional stopping or tuning followed the interval. Population identification does not guarantee successful finite-budget optimization or transfer.

6.3 Broader checks and the real-image boundary

Earlier controlled experiments recover nonlinear factors at mean $R^2 = 0.986$ with three tasks, but used more labels than their two-task comparator; they illustrate recovery, not budget-matched acquisition. Linear tests expose weak-signal, noise and support-misspecification failures. Supplement S6 and S14 retain these results and their original protocols.

The separate 240-fit MNIST/Fashion composite benchmark (LeCun et al., 1998; Xiao et al., 2017) tests ordinary support-masked MLPs outside the theorem. At 8,000 correlated training composites,

Table 3: Frozen-protocol nonlinear follow-up: 20 new seeds, 4,800 sampled training labels and 600 validation labels per fit. Mean ± SD. Specificity is nuisance conditional variance divided by total coordinate variance; lower is better.

Allocation	Spline recovery R^2	Specificity fraction	Shift accuracy
Two tasks	0.895 ± 0.113	0.104 ± 0.113	0.774 ± 0.096
Repeated support	0.870 ± 0.119	0.129 ± 0.120	0.761 ± 0.086
Weak separating	0.880 ± 0.112	0.121 ± 0.115	0.764 ± 0.079
Strong separating	0.920 ± 0.100	0.078 ± 0.093	0.787 ± 0.098

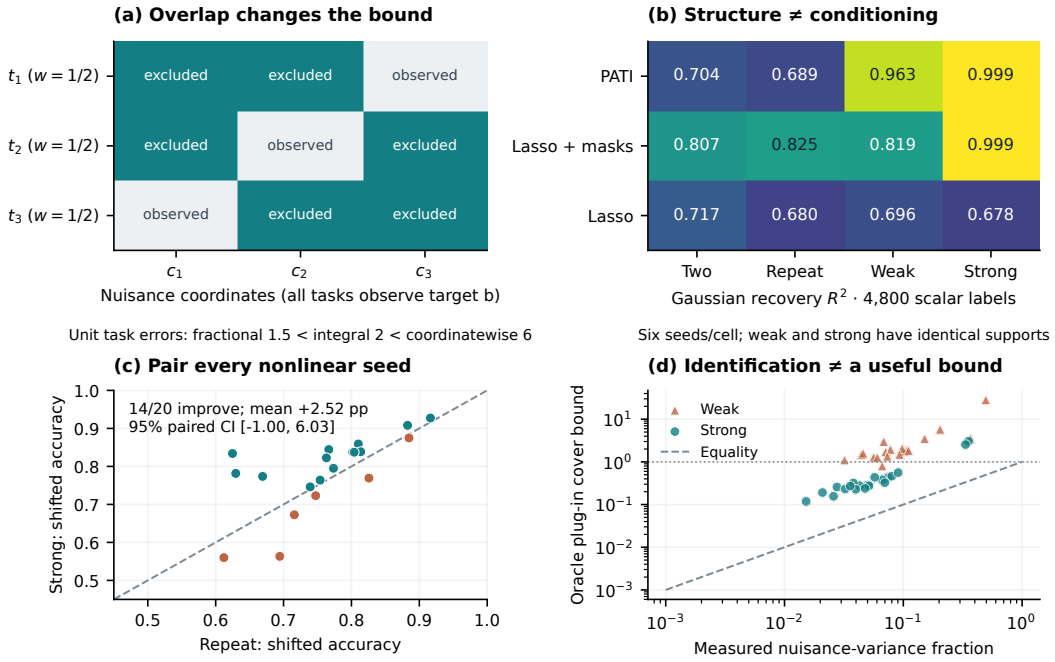


Figure 1: **From task overlap to finite-budget utility.** (a) A fractional exclusion cover shares the cost of controlling nuisance directions (Theorem 4). (b) Gaussian recovery across all allocations and support-information controls; each cell averages six seeds. (c) All 20 paired nonlinear follow-up seeds: points below the diagonal deteriorate after strong-task acquisition. (d) All 40 weak/strong fits: normalized oracle plug-in bounds versus measured nuisance variance, averaged across blocks. The horizontal line is the trivial variance bound; neither Monte Carlo quantity is a finite-sample certificate.

recombined matched-block accuracy is 0.604/0.589, versus dense heads' 0.587/0.575, with paired intervals spanning zero. An informed crop reference reaches 0.797/0.744; unadapted MNIST-to-UCI transfer is near chance. Crucially, its third pair label is reconstructible from the first two. It tests an extra readout constraint, not new measurement information. The released source-disjoint splits and evaluator remain useful as a boundary test; they do not validate continuous-factor identification. Table 4 preserves the central comparison; full results are in Supplement S10.

6.4 What controls the observed gains?

Figure 1 separates support geometry, conditioning and optimization. In the Gaussian strong-task comparison, PATI exceeds unknown-support inner-Lasso by 32.93 percentage points in shifted accuracy (paired 95% interval $[25.82, 40.03]$), but only by 0.63 points when both receive the supports ($[-0.20, 1.45]$). These six-seed contrasts show why an information-matched comparator changes the interpretation: the large gap cannot be attributed to the spectral optimizer alone.

Table 4: Constructed real-image benchmark, 8,000 training composites with correlation .9. Five-seed mean accuracy; Supplement S10 reports variability and all budgets. Matched blocks use a probe-validation assignment. External transfer is MNIST to UCI digits; missing input removes non-target tiles.

Source	Encoder	IID matched	Recombined matched	Recombined full	Missing named
Mnist	three	0.646	0.604	0.741	0.637
Mnist	two	0.631	0.598	0.730	0.604
Mnist	dense	0.612	0.587	0.730	0.562
Mnist	wrong	0.618	0.583	0.723	0.598
Fashion	three	0.587	0.589	0.664	0.621
Fashion	two	0.584	0.583	0.655	0.613
Fashion	dense	0.580	0.575	0.657	0.601
Fashion	wrong	0.586	0.595	0.667	0.602

Table 5: Original clinical results, not PATI. Weighted F1 is source-reported; fusion macro F1 and highest-band recall are recalculated from aggregate confusion counts. No paired significance claim is made.

Task	Text $F1_w$	Fusion $F1_w$	Fusion $F1_{macro}$	Highest-band recall
Depression	.860	.850	.832	10/15 (.667)
PTSD	.861	.854	.851	125/140 (.893)

The nonlinear follow-up improves in 14 of 20 paired seeds and deteriorates in six. To diagnose conditioning, we evaluate the covering program using true synthetic log odds, fitted inverse constants and Monte Carlo signature errors. Each block's bound is divided by its learned variance before averaging. For weak versus strong tasks, median bounds are 1.769 versus 0.282; 19/20 versus 2/20 exceed the trivial bound of one. Median bound-to-measured-variance ratios are 21.88 versus 6.31. Thus identical separating supports can yield markedly different numerical control, while even strong tasks leave substantial slack. These are post-hoc diagnostics, not new training runs or simultaneous confidence guarantees; the acquisition policy did not use oracle errors. Supplement S15 specifies the calculations and released seed-level records.

7 Original clinical evidence and scope

The original DAIC-family models provide the clinical motivation: text weighted F1 was .860/.861 for depression/PTSD versus fusion .850/.854. Table 5 retains their severe-case error context. These are original-model results, not PATI. Cross-release uniqueness of the 405 records and the 1,170 repeated clinician ratings cannot be reanalysed from aggregate files; Supplement S8–S9 preserves the evidence and unresolved EATD protocol issues.

Limitations. The bound requires correct supports, controlled fitted inverses, population signature errors, and independence or a justified density-ratio correction. Its sharpness concerns local variance constraints. The acquisition proxy has only small-family linear validation; the comparator is adapted, and nonlinear/real-image utility remains unresolved. No clinical identification or safety guarantee follows.

8 Conclusion

The fractional-cover program quantifies how overlapping task exclusions control nuisance variance. Equal-budget, equal-information comparisons support linear acquisition benefits; nonlinear and image failures expose the gap between population identification and usable estimation. Released primal/dual calculations and seed-level diagnostics make that distinction testable.

REPRODUCIBILITY STATEMENT

The release includes proofs, the fractional-cover LP, fixed-budget acquisition code, a documented inner-Lasso adaptation, raw results, source-image manifests and tests. Development and frozen follow-up runs are separate. No restricted interview content is redistributed; original clinical runs cannot be reconstructed from aggregate artifacts alone.

AI-USE DISCLOSURE

Generative AI was used for manuscript language preparation. Numerical results were produced by released programs.

REFERENCES

Vašek Chvátal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979. doi: 10.1287/moor.4.3.233.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E. Vogt. Identifiability results for multimodal contrastive learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/pdf?id=U_2kuqoTcB.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3): 586–596, 1981. doi: 10.1214/aos/1176345462.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217, 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Sebastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *International Conference on Machine Learning*, pp. 18171–18206, 2023. URL https://proceedings.mlr.press/v202/lachapelle23a.html.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pp. 4114–4124, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Mokshay Madiman and Andrew Barron. Generalized entropy power inequalities and monotonicity properties of information. *IEEE Transactions on Information Theory*, 53(7):2317–2329, 2007.

Walter Nelson, Marco Fumero, Theofanis Karaletsos, and Francesco Locatello. Statistical and structural identifiability in representation learning. In *International Conference on Learning Representations*, 2026.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pp. 9030–9039, 2021. URL https://proceedings.mlr.press/v139/roeder21a.html.

Emanuel Sperner. Ein satz über untermengen einer endlichen menge. *Mathematische Zeitschrift*, 27:544–548, 1928. doi: 10.1007/BF01171114.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.nips.cc/paper_files/paper/2021/hash/8929c70f8d710e412d38da624b21c3c8-Abstract.html.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/956e5427549a82a7472e02adc88360e9-Abstract-Conference.html.

Yujia Zheng, Zijian Li, Shunxing Fan, Andrew Gordon Wilson, and Kun Zhang. Diverse dictionary learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6796bfdbc8f435a60d5b795eb8930959-Abstract-Conference.html.