TASK-COMPLETED IDENTIFIABILITY: RECOVERING TASK-RELEVANT FACTORS FOR SELECTIVE USE

Anonymous authors
Paper under double-blind review

ABSTRACT

Representations can preserve all information needed for prediction while still mixing the latent factors that a task should use separately. A particularly important case arises when environmental variation alone cannot distinguish task-relevant factors, even though task responses provide complementary constraints on the same latent structure. We propose *Task-Completed Identifiability* (TCI), which formalizes when these two information sources jointly remove such relevant-factor ambiguities within a shared generative model. We derive the Task-Relevant Rank Condition (TRC), a projected rank condition on the first- and second-order derivatives of environmental and task-response contrasts. Under the stated assumptions, TRC identifies task-relevant factors and the common task–factor correspondence up to permutation and factor-wise invertible transformations, without requiring individual identification of factors irrelevant to all tasks. We further construct settings in which either information source alone admits genuine mixing among relevant factors, whereas their combination resolves it. To estimate the shared model, we introduce TC-iVAE, which couples a response-free inference branch with task-conditioned variational branches through shared latent coordinates. A composite evidence lower bound (ELBO) decomposition links exact observable-law fit and posterior approximation to population identification. In complementary settings, joint environmental and task constraints resolve relevant-factor mixtures left unresolved by environmental variation while preserving task-relevant information. The resulting representations support more selective factor-wise editing, with substantially reduced changes in non-target factors. Together, these results distinguish preserving task information from identifying the factors needed to use that information selectively.

1 INTRODUCTION

A representation may preserve all information needed for prediction while still mixing latent factors that should be used separately. For example, if position and color are encoded jointly, either attribute may remain predictable even though selecting only position, or changing position while preserving color, is difficult. We therefore study *selective use*: identifying which latent factors are required by each task and using or editing those factors without unnecessary changes to others (Bengio et al., 2013; Locatello et al., 2019; 2020).

The obstacle is observational ambiguity: different mixtures of latent factors can induce the same observed distribution. Prior work reduces such ambiguity using distributional variation across environments (Hyvarinen et al., 2019; Khemakhem et al., 2020a), task supervision and sparse task supports (Lachapelle et al., 2023a; Zheng et al., 2026), generative structure (Lachapelle et al., 2023b; 2026), or complementary observations (Gresele et al., 2020; Yao et al., 2024). A central question is what happens when one information source leaves mixing among task-relevant factors unresolved, but another source constrains different directions.

We propose *Task-Completed Identifiability* (TCI), which combines environmental variation and task responses in a shared latent model (Figure 1). An observationally equivalent alternative model must explain both sets of observable laws simultaneously. We derive the *Task-Relevant Rank Condition* (TRC) on the first- and second-order derivatives of the combined contrasts. TRC requires relevant factor directions to remain distinguishable after quotienting out the column space of factors irrele-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

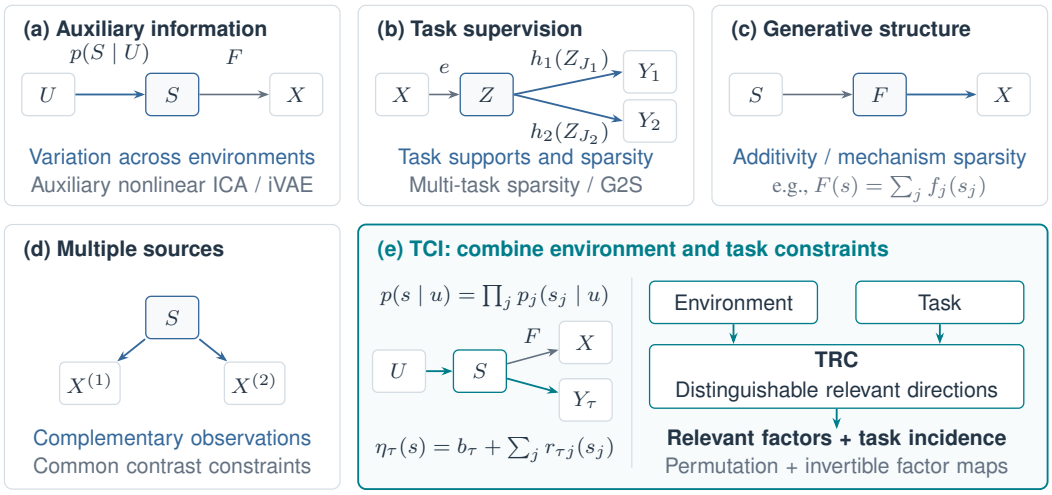


Figure 1: **Identifiability cues that constrain latent mixing and their combination in TCI.** (a) Environment-specific distributional variation, (b) task dependence and sparsity, (c) structure in the generative function or mechanisms, and (d) complementary multiple observations provide distinct identifiability cues. (e) TCI jointly enforces environmental and task-response constraints to restrict mixing among task-relevant factors. Under the stated model assumptions and TRC, task-relevant factors and the task–factor correspondence are identified up to a common permutation and factor-wise invertible transformations, without requiring individual identification of factors irrelevant to all tasks. Here, J_τ denotes the subset of representation factors used for task τ. Arrows in (b) denote predictive pathways; those in (a,c,d) and on the left of (e) denote generative relations, while the right of (e) denotes the combination of constraints in the analysis. The research directions may overlap.

vant to all tasks. Under the stated assumptions, this identifies the relevant factors and the task–factor correspondence up to one common permutation and factor-wise invertible transformations, while allowing nuisance factors to remain individually unidentified. Appendix C.2 gives an admissible model in which either source alone admits genuine relevant-factor mixing but their combination removes it.

To estimate the shared model, we introduce TC-iVAE, a task-conditional variational estimator. A response-free base branch and response-conditioned task branches share the prior, decoder, and latent coordinates; task likelihoods therefore constrain the same representation inferred from (X, U) at test time. A composite evidence lower bound (ELBO) decomposition separates observable-law fit from posterior approximation error and links a population optimum to TCI identification. Experiments with scalar and vector-valued factors evaluate the distinction directly: methods can preserve task-relevant or even full latent information while differing substantially in factor recovery, task-specific selection, and editing interference.

Contributions. **(i)** We give TRC, a sufficient condition for identifying task-relevant factors and their task–factor correspondence without requiring individual identification of factors irrelevant to all tasks (Section 3). **(ii)** We introduce TC-iVAE and connect its composite ELBO to population identification through observable-law fit and posterior approximation error (Section 4). **(iii)** We jointly evaluate information preservation, factor recovery, selection, and editing under controlled complementary information and a learned vector-factor setting (Section 5).

2 Problem Setup

2.1 Generative Model and Observations

Let $S = (S_1, \ldots, S_d) \in \mathcal{S} = \prod_j \mathcal{S}_j$ be latent factors, U the observed environment, and Y_τ the response for task $\tau \in [T]$. Each $\mathcal{S}_j$ is a connected, boundaryless m-dimensional C^2 manifold ($m = 1$ gives an open interval or circle), and true and candidate models use the same d and m.

The environment changes a conditionally factorized latent distribution, while observations share one generative function:

$$p(s \mid u) = \prod_{j=1}^{d} p_j(s_j \mid u), \qquad V = F(S), \qquad X = V + \varepsilon.$$

Coordinates within a factor may be dependent, and factors need not remain independent after marginalizing U (Hyvarinen et al., 2019; Khemakhem et al., 2020a).

We assume $Y_{1:T} \perp U \mid S$ and model each task by a fixed minimal regular exponential family whose natural parameter is additive across factors:

$$p_\tau(y \mid s) = h_\tau(y) \exp\{t_\tau(y)^\top \eta_\tau(s) - A_\tau(\eta_\tau(s))\}, \tag{1}$$

$$\eta_\tau(s) = b_\tau + \sum_{j=1}^{d} r_{\tau j}(s_j). \tag{2}$$

Additivity is imposed on the task natural parameter, not on F; $r_{\tau j}$ may be nonlinear and may couple coordinates within a vector-valued factor. The analysis also assumes a C^2 embedding F, a shared known independent noise law, positive smooth latent densities, suitable common support and contrast values, and admissible candidate models with the same factor dimensions and response families. Precise geometric, regularity, and support conditions are stated in Assumptions 1–4 of Appendix A.

2.2 Task Relevance and Observational Equivalence

A factor is relevant to task τ exactly when its contribution is nonconstant:

$$I_\tau := \{j : r_{\tau j} \text{ is nonconstant}\}, \qquad I_* := \bigcup_{\tau=1}^{T} I_\tau, \qquad N_* := [d] \setminus I_*.$$

The task–factor incidence matrix is $M_{\tau j} = \mathbf{1}\{j \in I_\tau\}$. Here, I_τ^c contains factors irrelevant to a particular task, whereas N_* contains only factors irrelevant to every task; a factor may therefore be irrelevant to one task but required by another. This distinction matters because TCI seeks one factor correspondence shared across tasks rather than a separate latent alignment for each task. We observe the environment marginal and each task marginal,

$$\mathcal{O}_J(\mathcal{M}) := \left(P_{\mathcal{M}}(X, U), \{P_{\mathcal{M}}(X, U, Y_\tau)\}_{\tau=1}^{T}\right), \tag{3}$$

without requiring all task responses to be observed jointly. Two admissible models are *observationally equivalent* when they have the same $\mathcal{O}_J$. Appendix A.2 shows that any such pair shares a smooth comparison map $G = F^{-1} \circ \widehat{F}$ from candidate to true latent states.

Our target is selective identification: for every observationally equivalent candidate, there should exist an injective $\pi : I_* \to [d]$ and factor-wise C^2 diffeomorphisms ρ_j such that

$$G_j(\widehat{s}) = \rho_j(\widehat{s}_{\pi(j)}) \quad (j \in I_*), \qquad \widehat{I}_\tau = \pi(I_\tau) \quad (\tau \in [T]). \tag{4}$$

Thus, the relevant factors and their common task–factor correspondence are identified up to a common permutation and factor-wise invertible transformations; ambiguity may remain inside N_*. In learning, $Z = e_\phi(X, U) = (Z_1, \ldots, Z_d)$ denotes the Euclidean representation and J_τ the learned factor subset selected for task τ; these are distinct from the alternative model state $\widehat{S}$ and the true relevance set I_τ.

3 Task-Relevant Identification

3.1 Combined Contrasts and TRC

For reference environments and responses from Assumption 3, define

$$a_r(s) = \log \frac{p(s \mid u_r)}{p(s \mid u_0)}, \qquad \ell_{\tau r}(s) = \log \frac{p_\tau(y_{\tau r} \mid s)}{p_\tau(y_{\tau 0} \mid s)}.$$

The latent factorization makes a_r additive, while the exponential-family log-partition cancels from $\ell_{\tau r}$. With $\Delta t_{\tau r} := t_\tau(y_{\tau r}) - t_\tau(y_{\tau 0})$, $\ell_{\tau r}(s) = c_{\tau r} + \sum_j \Delta t_{\tau r}^\top r_{\tau j}(s_j)$. Hence each factor contributes the joint signature

$$\psi_j(s_j) = \begin{bmatrix} (\log p_j(s_j \mid u_r) - \log p_j(s_j \mid u_0))_r \\ (\Delta t_{\tau r}^\top r_{\tau j}(s_j))_{\tau,r} \end{bmatrix},$$

and stacking all environmental and task contrasts gives

$$\Psi(s) = \begin{bmatrix} \Psi_A(s) \\ \Psi_T(s) \end{bmatrix} = c + \sum_{j=1}^{d} \psi_j(s_j), \qquad \widehat{\Psi}(\widehat{s}) = \Psi(G(\widehat{s})). \tag{5}$$

The second equality follows from observational equivalence (Appendix A.2); the same map also satisfies $\widehat{\eta}_\tau(\widehat{s}) = \eta_\tau(G(\widehat{s}))$.

Mixed derivatives across two distinct candidate factors vanish on the left of Eq. (5). The chain rule therefore gives a linear constraint in the first- and second-order derivatives of the factor signatures. In local coordinates, let $h = m + m(m+1)/2$ and collect these columns as

$$W_j(s_j) = \left[(\partial_\alpha \psi_j)_{\alpha=1}^m, (\partial^2_{\alpha\beta}\psi_j)_{1\le\alpha\le\beta\le m}\right], \qquad W = [W_1, \dots, W_d] \in \mathbb{R}^{q \times dh}. \tag{6}$$

Here, q is the total number of stacked scalar contrast components. The *Task-Relevant Rank Condition* (TRC) is

$$\operatorname{rank} W(s) - \operatorname{rank} W_{N_*}(s) = h|I_*|. \tag{7}$$

Equivalently, $Wc = 0 \Rightarrow c_{I_*} = 0$: relevant derivative coefficients cannot cancel either one another or the irrelevant column space, while arbitrary dependence within irrelevant columns is allowed. Coordinate-invariant and kernel formulations are given in Appendix B.1.

Theorem 1 (Identification of task-relevant factors and task–factor correspondence)**.** *Consider the true model and any $\mathcal{O}_J$-observationally equivalent alternative satisfying Assumptions 1–4. If TRC holds on a dense subset of $\mathcal{S}$, then there exist an injective map $\pi : I_* \to [d]$ and factor-wise C^2 diffeomorphisms ρ_j such that, for all latent states,*

$$G_j(\widehat{s}) = \rho_j(\widehat{s}_{\pi(j)}) \quad (j \in I_*), \qquad \widehat{I}_\tau = \pi(I_\tau) \quad (\tau \in [T]). \tag{8}$$

Extending π to a permutation $\widetilde{\pi}$ gives $\widehat{M}_{\tau,\widetilde{\pi}(j)} = M_{\tau j}$.

Why TRC removes mixing. Additivity of the candidate contrasts makes every cross-factor mixed derivative zero. After transport through G, this yields $Wc = 0$ for coefficients built from first and second derivatives of G. TRC forces the coefficients attached to relevant factors to vanish; in particular, a relevant true factor cannot depend differentially on two distinct candidate factors. Jacobian invertibility, connectedness, and the common factor dimension then globalize this local statement, while natural-parameter transport fixes the task–factor correspondence. Appendix B.2 gives the complete proof.

3.2 Complementarity and Selective Use

Writing W_A, W_T for environmental and task rows gives $\ker W = \ker W_A \cap \ker W_T$. Let $K_A := \ker W_A$ and $K_A^0 := K_A \cap \ker P_{I_*}$, where P_{I_*} extracts coefficients belonging to relevant factors. Because task rows vanish on N_*, TRC is equivalent to

$$\ker(W_T|_{K_A}) = K_A^0. \tag{9}$$

Thus task information need only eliminate the environment-permitted directions that still involve relevant coefficients; ambiguity confined entirely to nuisance coefficients may remain. This quotient view is the sense in which tasks *complete* the environmental information rather than identify every factor. Appendix B.4 gives the full statement. Appendix C.2 constructs genuine relevant-factor mixing under either source alone that disappears jointly, while Appendix C.4 shows that environmentally modulated nuisance factors can remain genuinely mixed after the relevant factor is identified.

Corollary 1 (Task-specific representation)**.** *Under Theorem 1, with $S = G(\widehat{S})$,*

$$Z_\tau^* = (\widehat{S}_{\pi(j)})_{j \in I_\tau} = H_\tau(S_{I_\tau}), \qquad H_\tau(s_{I_\tau}) = (\rho_j^{-1}(s_j))_{j \in I_\tau}. \tag{10}$$

Thus the selected factors retain the task-relevant state up to an invertible map and are conditionally independent of $S_{I_\tau^c}$ given U.

Corollary 2 (Selective editing under full factor identification)**.** *If $I_* = [d]$, changing only candidate factor $\pi(j)$ changes only true factor j.*

Under partial identification, other identified relevant factors are still preserved, but unidentified nuisance factors may change (Appendix C.5). These are population correspondence results; finite-model recovery and editing are evaluated in Section 5.

4 Estimation with Task-Conditional Variational Inference

TC-iVAE does not optimize TRC directly. Instead, it fits the environmental and task observable laws with one latent generative model, so the two information sources act on the same coordinates. A response-free base branch approximates $p(S \mid X, U)$ and each task branch approximates $p(S \mid X, U, Y_\tau)$; after training, only the base branch is needed to compute the representation.

4.1 Shared Model and Composite ELBO

A single parameter θ determines the environment-conditioned prior, decoder, and task heads:

$$p_{\theta,0}(x, s \mid u) = p_\varepsilon(x - F_\theta(s)) \prod_j p_{\theta,j}(s_j \mid u), \tag{11}$$

$$p_{\theta,\tau}(x, y, s \mid u) = p_{\theta,0}(x, s \mid u) p_{\theta,\tau}(y \mid s), \tag{12}$$

$$\eta_{\theta,\tau}(s) = b_{\theta,\tau} + \sum_j r_{\theta,\tau j}(s_j). \tag{13}$$

Thus $p_\theta(S \mid U)$ supplies the environmental constraint and $p_{\theta,\tau}(Y_\tau \mid S)$ supplies the task constraint within the same model. Conditioning on a task response may induce posterior dependence even though the prior factorizes across factors.

Let $q_0 = q_{\theta,\phi,0}(S \mid X, U)$ and $q_\tau = q_{\theta,\phi,\tau}(S \mid X, U, Y_\tau)$. Their branch ELBOs are

$$\begin{aligned} \mathcal{E}_0 &= \mathbb{E}_{q_0} \log p_\theta(X \mid S) - \mathrm{KL}(q_0 \| p_\theta(S \mid U)), \\ \mathcal{E}_\tau &= \mathbb{E}_{q_\tau}[\log p_\theta(X \mid S) + \log p_{\theta,\tau}(Y_\tau \mid S)] - \mathrm{KL}(q_\tau \| p_\theta(S \mid U)). \end{aligned} \tag{14}$$

The additional response likelihood is where task information constrains the shared latent coordinates during training. These task branches are not separate task-specific representations: they are variational distributions used to fit the same generative parameter θ. Thus responses may shape the common coordinates during training even though the deployed representation is computed without responses.

With positive branch weights w_k and $O_0 = X$, $O_\tau = (X, Y_\tau)$, define the composite observable log-likelihood $\mathcal{Q}(\theta) := \sum_{k=0}^{T} w_k \mathbb{E}[\log p_{\theta,k}(O_k \mid U)]$. The composite ELBO satisfies

$$\mathcal{J}(\theta, \phi) = \sum_{k=0}^{T} w_k \mathbb{E}[\mathcal{E}_k] = \mathcal{Q}(\theta) - \Delta(\theta, \phi), \tag{15}$$

$$\Delta(\theta, \phi) = \sum_{k=0}^{T} w_k \mathbb{E}\, \mathrm{KL}(q_{\theta,\phi,k}(\cdot \mid O_k, U) \| p_{\theta,k}(\cdot \mid O_k, U)). \tag{16}$$

Each term uses one observable marginal, so task responses need not be observed jointly. Task-specific posterior branches are used because adding Y_τ can change the posterior for the same (X, U); the base branch provides response-free inference.

4.2 Connection to Population Identification

Let $\mathcal{Q}_*$ be the composite score of the true observable densities and define

$$\mathcal{L}(\theta) = \sum_k w_k \mathbb{E}_U\, \mathrm{KL}(P_{*,k}(\cdot \mid U) \| P_{\theta,k}(\cdot \mid U)). \tag{17}$$

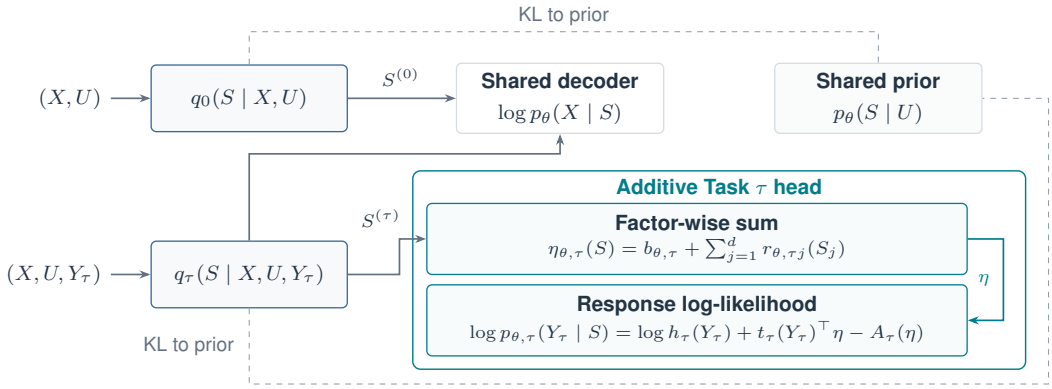


Figure 2: **Training paths of TC-iVAE.** The base and task branches share the same prior and decoder. The task head constructs the natural parameter $\eta = \eta_{\theta,\tau}(S)$ by summing factor-wise functions and evaluates the log-likelihood of the observed Y_τ under the corresponding response distribution. Solid arrows denote the flow of inputs, latent samples, and natural parameters, while dashed lines denote KL terms between the posteriors and the prior. A task branch is instantiated for each $\tau = 1, \ldots, T$ and does not require all responses to be observed jointly. After training, the representation is the mean of the base posterior, which takes only (X, U) as input.

Then the standard ELBO identity gives

$$\mathcal{J}(\theta, \phi) = \mathcal{Q}_* - \mathcal{L}(\theta) - \Delta(\theta, \phi), \tag{18}$$

separating observable-law fit from posterior approximation error. Because every $w_k > 0$, $\mathcal{L}(\theta) = 0$ means that each observed marginal in $\mathcal{O}_J$ is matched individually, not merely that their weighted sum is small.

Proposition 1 (Population identification through composite ELBO)**.** *Assume the density and integrability conditions of Appendix D. If the model class contains a θ_* matching all true observable marginals and a ϕ_* matching all corresponding branch posteriors, every global maximizer $(\widehat{\theta}, \widehat{\phi})$ satisfies $\mathcal{L}(\widehat{\theta}) = \Delta(\widehat{\theta}, \widehat{\phi}) = 0$. If Theorem 1 and TRC also hold, the relevant factors and task–factor correspondence are identified as in Eq.* (8)*.*

Hence TC-iVAE needs no separate identifiability penalty: at an exact population optimum, its shared model reproduces the same observable laws to which TCI applies. Realizability is required at one true parameter/posterior pair; exact posterior representation at every candidate parameter is not assumed. The result still requires global population optimization, and finite variational families, sampling, and optimization introduce additional errors analyzed in Appendix D. Observable-law closeness alone also does not control contrast derivatives without further stability assumptions.

4.3 Task-Specific Use in Practice

The learned representation is the base-posterior mean

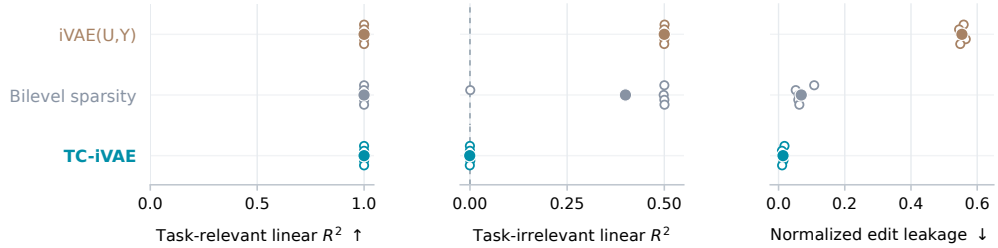
$$Z = e_{\widehat{\theta},\widehat{\phi}}(X, U) := \mathbb{E}_{q_{\widehat{\theta},\widehat{\phi},0}(\cdot \mid X,U)}[S], \qquad Z_\tau = Z_{J_\tau}, \tag{19}$$

where J_τ is selected from validation data using a common task probe; true factor labels and I_τ are not used for selection, and all m coordinates of a factor are taken together. Thus task responses shape the shared coordinates during training, while inference uses (X, U) alone and a subsequent task-specific subset.

The scalar experiment isolates the effect of complementary information using a fixed nonlinear observation transform. The vector experiment learns the encoder and decoder and uses

$$\widehat{\mathcal{L}}_{\text{train}} = \mathcal{R} + \beta\mathcal{K} + \lambda_y\mathcal{C} + \lambda_g\Omega_g + \lambda_{\text{loc}}\Omega_{\text{loc}} + \lambda_J\Omega_J, \tag{20}$$

Here, $\mathcal{R}$, $\mathcal{K}$, and $\mathcal{C}$ denote the reconstruction loss, prior KL, and response negative log-likelihood, respectively. The additional finite-sample regularizers control gate magnitude, branch locality, and task Jacobians; they are implementation choices rather than replacements for TRC. Exact coefficients and the unregularized composite-ELBO comparison are given in Appendix E.2.



(a) Relevant factor R^2 (b) Irrelevant factor R^2 (c) Coordinate editing

Figure 3: **Preservation versus selective use in Collider.** Relevant/irrelevant linear R^2 of the validation-selected coordinates and normalized non-target RMS change under factor editing. *Shared-posterior* denotes the ablation in which all branches use the same response-free variational posterior $q(S \mid X, U)$, i.e., the task-dependent posterior corrections are removed. Large markers are five-seed means; open markers are individual seeds.

Table 1: Collider control experiment (mean ± SD, five seeds). Edit leak is the non-target RMS change; lower irrelevant R^2 and leakage indicate greater selectivity.

Method	MCC ↑	Rel. R^2 ↑	Irrel. $R^2 \approx 0$	Edit leak ↓
iVAE(U)	0.9451 ± 0.0810	0.9987 ± 0.0026	0.3011 ± 0.2722	0.2462 ± 0.2553
iVAE(U, Y)	0.8476 ± 0.0031	1.0000 ± 0.0000	0.4999 ± 0.0007	0.5538 ± 0.0082
CI-iVAE(U, Y)	0.8463 ± 0.0024	1.0000 ± 0.0000	0.4999 ± 0.0007	0.5574 ± 0.0057
Mechanism-sparse	0.9412 ± 0.0731	0.9998 ± 0.0001	0.3994 ± 0.2248	0.2615 ± 0.2427
Task-additive	0.9123 ± 0.0413	0.9999 ± 0.0001	0.4997 ± 0.0008	0.3827 ± 0.1106
Bilevel sparsity	0.9967 ± 0.0026	0.9997 ± 0.0005	0.4000 ± 0.2232	0.0686 ± 0.0222
G2S-Jacobian	0.8199 ± 0.0218	1.0000 ± 0.0000	1.0000 ± 0.0000	0.6760 ± 0.0560
TC-iVAE	0.9999 ± 0.0001	0.9999 ± 0.0000	-0.0006 ± 0.0015	0.0135 ± 0.0031

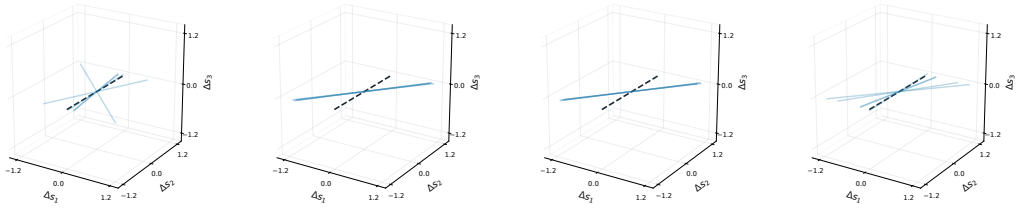
5 Experiments

We test the distinction motivating TCI: a representation may preserve task-relevant information yet fail to separate the factors needed for task-specific selection or editing. We jointly evaluate factor recovery, relevant/irrelevant information in the selected representation, and non-target changes under factor-wise editing. The same validation-based selection rule is applied to all methods, and true factors are used only for post-training calibration and evaluation. We compare adaptations of iVAE, CI-iVAE, mechanism sparsity, task-additive models, bilevel sparsity, and G2S (Khemakhem et al., 2020a; Kim et al., 2023; Lachapelle et al., 2026; 2023b;a; Zheng et al., 2026). Results are mean ± SD over five seeds; complete protocols and results are in Appendix E.

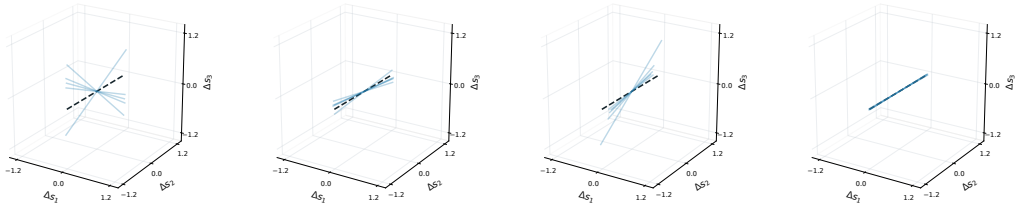
5.1 Controlled Complementary Information

The scalar setting has three factors, two four-category tasks, and a fixed nonlinear observation transform. In the representative Collider condition, task 1 depends on S_1, S_2 and task 2 on S_2, S_3; conditioning on their common-effect responses can break factor independence in $p(S \mid U, Y)$. The learner does not receive the true mixing directions. MCC measures one-to-one factor recovery, while relevant and irrelevant R^2 evaluate what the validation-selected task representation retains.

Table 1 isolates preservation from selectivity. Both iVAE(U, Y) and TC-iVAE retain essentially all relevant information, but their irrelevant R^2 is about 0.500 versus 0, respectively. TC-iVAE selects the two required coordinates for every task in every seed (incidence F1 $= 1$), reaches MCC 0.9999, and reduces mean editing leakage by about 97.6% relative to iVAE(U, Y); its target-edit error is 0.0008 ± 0.0001. Thus the relevant information can be present while the factorization needed for selective use remains different.



(a) iVAE(U) (b) iVAE(U, Y) (c) CI-iVAE(U, Y) (d) Mechanism-sparse



(e) Task-additive (f) Bilevel sparsity (g) G2S-Jacobian (h) TC-iVAE

Figure 4: **Factor-change trajectories under vertical-position editing in the Collider setting.** The black dashed line is the ground-truth trajectory, and the translucent blue lines show the results from five seeds. $(\Delta s_1, \Delta s_2, \Delta s_3)$ denote changes in horizontal position, vertical position, and color relative to the pre-edit reconstruction. The experimental setup and results for horizontal-position and color editing are provided in Appendix E.1.8.

Figure 4 makes the editing criterion explicit. For visualization, S_1, S_2, S_3 denote horizontal position, vertical position, and color. A ground-truth edit of vertical position changes only S_2. The five TC-iVAE trajectories remain close to that path. Several comparison methods instead show visible non-target changes in S_1 or S_3.

The plot visualizes the synthetic factor changes rather than samples from a learned image generator; horizontal-position and color edits are reported in Appendix E.1.8.

Across six controlled information conditions, the result follows the available constraints rather than the method name alone. In Collider and Complementary the environmental, task, and combined derivative ranks are $3, 5, 6$, so the combined source supplies missing directions. When one source is already sufficient, methods using that source can also recover the factors; in Redundant, where both sources allow the same mixing, TC-iVAE still has editing leakage 0.225. These controls show both the benefit and the limit of complementarity. Full condition tables, regularizer sensitivity, and additional editing trajectories are in Appendix E.1.

5.2 Learned Vector-Valued Factors

The second setting learns both encoder and decoder with three two-dimensional factors, four tasks, and 24-dimensional observations. S_1, S_2 are task-relevant and S_3 is irrelevant to all tasks but responds to the environment. The true observation-generating map and its inverse are not available during training, and all methods use the same factor-additive neural decoder family. Because factor-wise invertible transformations are allowed, recovery uses nonlinear R^2 between matched two-dimensional factors; selection and editing operate on entire factors. Changes in the other relevant factor are the main non-target editing metric, while nuisance changes are reported separately in the appendix.

All methods have full-representation nonlinear $R^2 \gtrsim 0.999$, yet their matched factor recovery and selective-use metrics differ. TC-iVAE preserves relevant information ($R^2 \approx 0.999$), has mean relevant-factor recovery 0.740, irrelevant $R^2 = 0.422$, and target-edit error 0.496. Other-relevant interference is below 0.008 in four seeds but 0.985 in one, so the learned model remains imperfect. The combination of near-equal full-information preservation with different factor-wise recovery and editing supports the paper's central distinction; complete variability, nuisance editing, and the composite-ELBO implementation are reported in Appendix E.2.

Table 2: Learned vector-factor experiment (mean $\pm$ SD, five training seeds). The full representation preserves nearly all latent information for every method; the table tests factor correspondence and selective use.

Method	Factor $R^2 \uparrow$	Rel. $R^2 \uparrow$	Irrel. $R^2 \approx 0$	Target Δ err. $\downarrow$	Other leak $\downarrow$
iVAE(U)	0.5829 ± 0.1432	0.9989 ± 0.0001	0.5984 ± 0.2024	0.7961 ± 0.2676	0.3988 ± 0.4097
iVAE(U, Y)	0.5829 ± 0.1433	0.9955 ± 0.0005	0.5980 ± 0.2026	0.7959 ± 0.2672	0.3993 ± 0.4100
CI-iVAE(U, Y)	0.5828 ± 0.1426	0.9949 ± 0.0012	0.6971 ± 0.0704	0.7947 ± 0.2687	0.2943 ± 0.2643
Mechanism-sparse	0.5829 ± 0.1431	0.9990 ± 0.0001	0.5985 ± 0.2025	0.7962 ± 0.2676	0.3988 ± 0.4096
Task-additive	0.5872 ± 0.1385	0.9956 ± 0.0006	0.6951 ± 0.0702	0.8012 ± 0.2725	0.2855 ± 0.2547
Bilevel sparsity	0.6370 ± 0.1411	0.9975 ± 0.0006	0.6709 ± 0.0698	0.7068 ± 0.2625	0.2046 ± 0.2721
G2S-Jacobian	0.5870 ± 0.1384	0.9957 ± 0.0006	0.6951 ± 0.0702	0.8013 ± 0.2722	0.2858 ± 0.2542
TC-iVAE	0.7398 ± 0.1837	0.9987 ± 0.0003	0.4223 ± 0.2850	0.4963 ± 0.3450	0.2023 ± 0.4376

6 Discussion and Conclusion

TCI separates two goals that need not coincide: preserving information for prediction and identifying the factors needed for selective use. By combining environmental and task-response contrasts, TRC rules out relevant-factor mixing modulo the nuisance column space, yielding relevant factors and a common task–factor correspondence up to permutation and factor-wise invertible transformations. TC-iVAE fits the corresponding observable laws within one shared latent model, and the experiments show that similar information preservation can coexist with markedly different factor recovery, selection, and editing.

Limitations. The theorem is a population sufficient condition with equal factor dimensions, a shared known noise law, and the stated generative structure; partial identification does not protect unidentified nuisance factors under editing. The variational bridge additionally requires realizability and global optimization, and the experiments are synthetic. Each m-dimensional relevant factor contributes $h = m + m(m+1)/2$ first- and second-order derivative columns to TRC, so scaling depends not only on the number of tasks but also on whether their contrasts provide sufficiently distinct directions. The current validation procedure additionally compares all factor subsets, which is not scalable to many factors.

Future work. Natural next steps are to relax the equal-dimension and known-noise assumptions, develop stability and finite-sample theory for approximate TRC, and replace exhaustive subset selection with structured or sparse procedures. Evaluating TCI with missing task responses, larger factor/task systems, broader decoder and posterior families, and real data will be important for determining when complementary task information yields reliable selective use in practice.

AI use statement

Generative AI assisted with manuscript drafting, translation, literature search, research ideation, mathematical and code review, and experimental code. The authors reviewed and, where necessary, independently verified all outputs and take full responsibility for the claims, citations, proofs, and experimental results.

References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50. URL https://doi.org/10.1109/TPAMI.2013.50.

Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ICA. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 217–227. PMLR, 2020. URL https://proceedings.mlr.press/v115/gresele20a.html.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868. PMLR, 2019. URL https://proceedings.mlr.press/v89/hyvarinen19a.html.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020a. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. ICE-BeeM: Identifiable conditional energy-based deep models based on nonlinear ICA. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12768–12778. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/962e56a8a0b0420d87272a682bfd1e53-Abstract.html.

Young-Geun Kim, Ying Liu, and Xue-Xin Wei. Covariate-informed representation learning to prevent posterior collapse of iVAE. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 2641–2660. PMLR, 2023. URL https://proceedings.mlr.press/v206/kim23c.html.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114. Preprint.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11455–11472. PMLR, 2022. URL https://proceedings.mlr.press/v162/kong22a.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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18171–18206. PMLR, 2023a. URL https://proceedings.mlr.press/v202/lachapelle23a.html.

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. In *Advances in Neural Information Processing Systems*, volume 36, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4ef594af0d9a519db8fb292452c461fa-Abstract-Conference.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, 2026. URL https://jmlr.org/papers/v27/24-0771.html.

Zijian Li, Shunxing Fan, Yujia Zheng, Ignavier Ng, Shaoan Xie, Guangyi Chen, Xinshuai Dong, Ruichu Cai, and Kun Zhang. Synergy between sufficient changes and sparse mixing procedure for disentangled representation learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=G1r2rBkUdu.

Yuhang Liu, Zhen Zhang, Dong Gong, Erdun Gao, Biwei Huang, Mingming Gong, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Towards identifiable latent additive noise models, 2024. URL https://arxiv.org/abs/2403.15711. Version 2, revised January 2026.

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 *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Francesco Locatello, Ben Poole, Gunnar Raetsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6348–6359. PMLR, 2020. URL https://proceedings.mlr.press/v119/locatello20a.html.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9030–9039. PMLR, 2021. URL https://proceedings.mlr.press/v139/roeder21a.html.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN). In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygeHgSFDH.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011. URL https://www3.stat.sinica.edu.tw/statistica/oldpdf/A21n11.pdf.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=OGtnhKQJms.

Soheun Yi, Yizhou Lu, Chandler Squires, and Pradeep Ravikumar. Concept modulation models: A unified framework for identifiability and extrapolation, 2026. URL https://arxiv.org/abs/2606.18509. Preprint.

Yujia Zheng, Fan Feng, Yuke Li, Shaoan Xie, Kevin Murphy, and Kun Zhang. From generalist to specialist representation, 2026. URL https://arxiv.org/abs/2605.12733. Author-submitted version; the arXiv record lists ICML 2026.

A Assumptions and Observational Equivalence

This appendix specifies the model class introduced in Section 2 and proves how a common comparison map between latent coordinates can be obtained from the observable laws. We then show how combining the environmental observation with task-specific observations restricts the set of observationally equivalent models. These results provide the basic facts used in the subsequent identification analysis; identification of the relevant factors itself requires additional conditions.

A.1 Model Class and Assumptions

We write $[d] = \{1, \dots, d\}$ and $[T] = \{1, \dots, T\}$. The probability law induced by a model $\mathcal{M}$ is denoted by $P_{\mathcal{M}}$, and the subscript is omitted when the model being compared is clear. For a measurable map H and measure μ, $H_{\#}\mu$ denotes the pushforward measure defined by $H_{\#}\mu(B) = \mu(H^{-1}(B))$. We write S for the latent state of the true model and $\widehat{S}$ for that of an alternative model, and distinguish both from the learned encoder output Z.

Assumption 1 (Latent geometry and observation noise)**.** The latent space is a product of connected, boundaryless, m-dimensional C^2 manifolds,

$$\mathcal{S} = \prod_{j=1}^{d} \mathcal{S}_j. \tag{21}$$

The number of factors is d, the common factor dimension is $m \geq 1$, and the total dimension is $D = dm$. Each $\mathcal{S}_j$ is Hausdorff and second countable, and carries a fixed volume measure λ_j with a positive C^1 density in local charts. We write the product measure as $\lambda_{\mathcal{S}} = \bigotimes_{j=1}^{d} \lambda_j$. Alternative models also use a product of the same number d of m-dimensional factor spaces. When $m = 1$, each connected factor space is diffeomorphic to an open interval or a circle. The observation function $F : \mathcal{S} \to \mathbb{R}^p$ is a C^2 embedding. We assume that the latent space of each model is compact or that its observation function is proper, so that its image is closed.

The observation is generated as

$$V = F(S), \qquad X = V + \varepsilon, \qquad \varepsilon \perp (S, U, Y_{1:T}). \tag{22}$$

In the noiseless case, we set $P_\varepsilon = \delta_0$. In the noisy case, all candidate models share the same known law P_ε, and the set of points at which its characteristic function

$$\varphi_\varepsilon(t) = \mathbb{E}[\exp(it^\top \varepsilon)] \tag{23}$$

is nonzero is dense in $\mathbb{R}^p$. All random-variable spaces are assumed to be standard Borel spaces.

The embedding and closed image guarantee a global comparison map over the common signal space. Gaussian noise satisfies the noise condition because its characteristic function is nowhere zero. Candidate models are not allowed to vary the noise law freely.

Assumption 2 (Environmental modulation and task responses)**.** For each environment u, the conditional latent density with respect to $\lambda_{\mathcal{S}}$ satisfies

$$p(s \mid u) = \prod_{j=1}^{d} p_j(s_j \mid u) > 0, \tag{24}$$

and each density is C^2 in its corresponding factor. The factorization is imposed only across factors; dependence among coordinates within a factor is allowed. The environment and task responses satisfy

$$Y_{1:T} \perp U \mid S. \tag{25}$$

For task τ, the conditional response density belongs to a fixed minimal regular exponential family with respect to a fixed dominating measure ν_τ. Specifically,

$$p_\tau(y \mid s) = h_\tau(y) \exp\left\{ t_\tau(y)^\top \eta_\tau(s) - A_\tau(\eta_\tau(s)) \right\}, \tag{26}$$

$$\eta_\tau(s) = b_\tau + \sum_{j=1}^{d} r_{\tau j}(s_j), \qquad r_{\tau j} \in C^2(\mathcal{S}_j, \mathbb{R}^{k_\tau}). \tag{27}$$

Each $r_{\tau j}$ may jointly depend on all coordinates within the factor. For every s, $\eta_\tau(s)$ lies in the open natural-parameter space of the corresponding family. Because the sufficient statistic t_τ is minimal, for any coefficient vector $a \in \mathbb{R}^{k_\tau}$ of the same dimension, if $a^\top t_\tau(y)$ is constant $h_\tau\, d\nu_\tau$-almost everywhere, then $a = 0$. The model includes a joint law $P(Y_{1:T} \mid S)$ with these task-specific marginals, and conditional independence among tasks is not required.

Positivity in Eq. (24) ensures that every environment has full support on the same latent space. Equation (25) means that the environment changes the latent distribution while the response mechanism remains unchanged when the latent state is fixed. Additivity is imposed on the natural parameter of the response and does not require additivity of the observation function F. Minimality removes redundant representations of the natural parameter. Indeed, if $p_\eta = p_{\eta'}$ within the same family, then $(\eta - \eta')^\top t(y) = A(\eta) - A(\eta')$ almost everywhere, and minimality therefore implies $\eta = \eta'$.

Assumption 3 (Support and regularity at contrast values)**.** Choose a reference environment u_0 and finitely many environment values $u_1, \ldots, u_{q_A}$. For a discrete environment, each selected value has positive mass under P_U. For a continuous environment, let $\mathcal{U}$ be a Polish space and assume that P_U is a Borel probability measure with full support on $\mathcal{U}$. In this case, each $p_j(s_j \mid u)$ of both the true and alternative models is positive and jointly continuous in (s_j, u), and the selected values belong to $\mathcal{U}$.

For task τ, choose response values $y_{\tau 0}, \ldots, y_{\tau q_\tau}$. For a discrete response, each selected point has positive probability. For a continuous response, use a common support and a fixed density version in neighborhoods of the selected points. On these neighborhoods, h_τ is finite and positive, $p_\tau(y \mid s)$ is jointly continuous in (y, s), and ν_τ has full support with respect to the relative topology. We further require

$$\operatorname{span}\{t_\tau(y_{\tau r}) - t_\tau(y_{\tau 0}) : r \in [q_\tau]\} = \mathbb{R}^{k_\tau}. \tag{28}$$

All candidate models share these conditions and the same selected values.

This assumption prevents the result from depending on arbitrarily chosen versions at conditioning values of probability zero. Continuity in the environment is used to extend equalities obtained for almost every u to the selected environment values. The span condition on the response values ensures that the selected response contrasts distinguish every direction of the natural parameter. The combination of contrasts and their derivative conditions are treated in the identification theory.

Assumption 4 (Admissible alternatives and relevance)**.** The true model and all alternative models satisfy the preceding three assumptions and use the same observation space, number of factors d, common factor dimension m, environment space, task-specific response families, and noise law. The coordinate systems within factors, relevance sets of alternative models, and factor correspondence are not fixed to their true-model values. The task-specific relevance sets and their union are defined as

$$I_\tau = \{j \in [d] : r_{\tau j} \text{ is nonconstant}\}, \qquad I_* = \bigcup_{\tau=1}^{T} I_\tau, \qquad N_* = [d] \setminus I_*. \tag{29}$$

For an alternative model, the same definition gives $\widehat{I}_\tau$, $\widehat{I}_*$, and $\widehat{N}_*$.

On a connected factor space, $r_{\tau j}$ being nonconstant is equivalent to $Dr_{\tau j} \not\equiv 0$. This notion of relevance is unaffected by the additive decomposition's freedom to shift constants. If the same natural parameter is represented by two additive decompositions, fixing all other factors implies that the two functions associated with each factor can differ only by a constant. Thus, the relevance sets are determined without an additional centering convention. Coordinates in N_* are irrelevant to the entire collection of tasks, but their distributions may still vary with the environment. A coordinate with $j \notin I_\tau$ may also be relevant to another task.

A.2 Observational Equivalence and a Common Latent Map

Noise separation and composition of the generative functions are also used as preparatory steps in the iVAE identification proof (Khemakhem et al., 2020a, Supplement, Sections B.2.2 and B.3.2). The following result adapts that argument to the present joint-law, manifold, and support conditions.

Lemma 1 (Observation transport and common comparison map)**.** *The following three transfer steps hold.*

(i) Let C be a standard Borel variable and $X = V + \varepsilon$, with $\varepsilon \perp (V, C)$. If two models use the same noise law from Assumption 1, then $P(X, C) = \widehat{P}(X, C)$ implies $P(V, C) = \widehat{P}(V, C)$.

(ii) Under Assumptions 1–2, two models generating the same $P(X, U)$ have the same image of the observation function, and

$$G = F^{-1} \circ \widehat{F} : \widehat{\mathcal{S}} \to \mathcal{S}, \qquad \widehat{F} = F \circ G \tag{30}$$

is a C^2 diffeomorphism. For this single map,

$$G_{\#}\widehat{P}_{\widehat{S}|U=u} = P_{S|U=u} \quad \textit{for } P_U\textit{-almost every } u. \tag{31}$$

(iii) If, in addition, Assumptions 3–4 hold and $\mathcal{O}_J(\widehat{\mathcal{M}}) = \mathcal{O}_J(\mathcal{M})$, then for the two models constructed using the same contrast values,

$$\widehat{\eta}_\tau = \eta_\tau \circ G, \qquad \widehat{\Psi} = \Psi \circ G \tag{32}$$

holds at every latent point.

Justification. (i) For each measurable set B, define the finite measure $\mu_B(A) = P(V \in A, C \in B)$. Then

$$\mathbb{E}[e^{it^\top X}\mathbf{1}_{\{C\in B\}}] = \varphi_\varepsilon(t) \int e^{it^\top v}\, d\mu_B(v).$$

Observational equivalence makes the Fourier transforms of the two μ_B agree on the dense set where $\varphi_\varepsilon \neq 0$. Continuity and uniqueness of the Fourier transform for finite measures imply $\mu_B = \widehat{\mu}_B$, and agreement on all rectangles gives equality of the joint laws. This applies to $C = U, Y_\tau, (U, Y_\tau), Y_{1:T}$ and does not require C to have a density.

(ii) Positive latent densities give P_S full support, and compactness or properness makes the image of the embedding closed, so

$$F(\mathcal{S}) = \operatorname{supp} P_V = \operatorname{supp} \widehat{P}_V = \widehat{F}(\widehat{\mathcal{S}}). \tag{33}$$

Composing the two C^2 diffeomorphisms onto the common embedded manifold gives G. Applying (F^{-1}, Id) to the common (V, U) law obtained in (i) and using uniqueness of regular conditional distributions gives Eq. (31).

(iii) The positive Jacobian J_G with respect to the fixed volume measures satisfies

$$\int_{G(A)} f\, d\lambda_{\mathcal{S}} = \int_A (f \circ G) J_G\, d\lambda_{\widehat{\mathcal{S}}}, \qquad J_G = \frac{w \circ G}{\widehat{w}} |\det DG|, \tag{34}$$

where the second expression is written in local charts whose volume densities are w and $\widehat{w}$. Hence

$$\widehat{p}(\widehat{s} \mid u) = p(G(\widehat{s}) \mid u) J_G(\widehat{s}) \tag{35}$$

initially holds almost everywhere with respect to the product measure. Continuity in the latent coordinates and full support extend it to every $\widehat{s}$ for almost every u, and the positive-mass or joint-continuity conditions of Assumption 3 extend it to every selected u_r. The same J_G cancels in environmental density ratios. Meanwhile, applying (i) to the task observation laws yields a common (V, Y_τ) law, which identifies the task kernels at $\widehat{P}_{\widehat{S}}$-almost every point. Minimality of the fixed exponential family gives equality of the natural parameters, and continuity and full support extend this equality to every latent point. Under the fixed density version,

$$\ell_{\tau r}(s) = \log \frac{h_\tau(y_{\tau r})}{h_\tau(y_{\tau 0})} + \Delta t_{\tau r}^\top \eta_\tau(s),$$

so the task contrasts are transferred by the same G. The fact that the contrasts of both models are factor-wise additive is used in the subsequent mixed-derivative argument. This preparatory result does not itself assert a factor-wise decomposition of G.

A.3 Observation Fibers and the Role of Task Coupling

We write the observation operators from Section 2.2 as

$$\mathcal{O}_A(\mathcal{M}) = P_{\mathcal{M}}(X, U), \tag{36}$$

$$\mathcal{O}_{T,m}(\mathcal{M}) = \{P_{\mathcal{M}}(X, Y_\tau)\}_{\tau=1}^{T}, \tag{37}$$

$$\mathcal{O}_J(\mathcal{M}) = \left(P_{\mathcal{M}}(X, U), \{P_{\mathcal{M}}(X, U, Y_\tau)\}_{\tau=1}^{T}\right), \tag{38}$$

$$\mathcal{O}_{T,j}(\mathcal{M}) = P_{\mathcal{M}}(X, Y_{1:T}). \tag{39}$$

The subscript m denotes the collection of task-wise marginals, while j denotes the joint law of all tasks. When $T \geq 1$, $P(X, U)$ can be obtained from $P(X, U, Y_\tau)$, but we retain it in $\mathcal{O}_J$ to make the environmental channel explicit. For a fixed admissible model class $\mathfrak{M}$, define the observation fiber by

$$\mathfrak{F}_O(\mathcal{M}) = \{\widehat{\mathcal{M}} \in \mathfrak{M} : \mathcal{O}(\widehat{\mathcal{M}}) = \mathcal{O}(\mathcal{M})\}. \tag{40}$$

Proposition 2 (Intersection of environmental and task constraints)**.** *Under Assumptions 1–2,*

$$\mathfrak{F}_J(\mathcal{M}) = \mathfrak{F}_A(\mathcal{M}) \cap \mathfrak{F}_{T,m}(\mathcal{M}). \tag{41}$$

If the joint law of all tasks is used, then

$$\mathfrak{F}_A(\mathcal{M}) \cap \mathfrak{F}_{T,j}(\mathcal{M}) \subseteq \mathfrak{F}_J(\mathcal{M}), \tag{42}$$

and this inclusion can be strict.

Proof. One inclusion follows by marginalization. Conversely, a candidate in $\mathfrak{F}_A \cap \mathfrak{F}_{T,m}$ has the same $P(V, U)$ and $P(V, Y_\tau)$ by Lemma 1(i). Because the embedding implies that S and V generate the same σ-field, $Y_\tau \perp U \mid S$ gives $Y_\tau \perp U \mid V$, and

$$P(dv, du, dy_\tau) = P_V(dv)P(du \mid v)P(dy_\tau \mid v) \tag{43}$$

has kernels determined almost everywhere by the common marginal laws. Adding back the same independent noise then gives the same $P(X, U, Y_\tau)$. Equality of the joint law of all tasks implies equality of each task marginal, yielding the second inclusion. □

This does not claim that two arbitrary marginals determine a joint law; the result uses noise separation and the conditional independence above.

For an example of strict inclusion, let $S \mid U = u \sim \mathcal{N}(u, 1)$, $X = S$, and consider two Bernoulli tasks with common success probability $a(s) = (1 + e^{-s})^{-1}$. Let $U \sim \mathrm{Bernoulli}(1/2)$ and keep the two task marginals fixed while varying the coupling as

$$P_\rho(Y_1 = 1, Y_2 = 1 \mid s) = a(s)^2 + \rho a(s)(1 - a(s)), \qquad 0 \leq \rho < 1.$$

The remaining probabilities are determined by the two marginals. All ρ yield the same $\mathcal{O}_J$ and relevance sets $I_1 = I_2 = \{1\}$, while $P(Y_1 = Y_2 \mid s)$ differs.

A.4 Response Families and Interpretation of the Target

For a Bernoulli response, $t(y) = y$ and the natural parameter is the logit. For a Gaussian response with fixed variance σ^2, $t(y) = y$ and $\eta(s) = \mu(s)/\sigma^2$. The C-category responses used in the experiments take category 0 as the reference and are represented as

$$t(y) = (\mathbf{1}\{y = c\})_{c=1}^{C-1}, \qquad \eta_c(s) = \log \frac{p(Y = c \mid s)}{p(Y = 0 \mid s)}.$$

The $C - 1$ log-odds are the natural parameters, and the differences between the reference category and the remaining categories satisfy the response-span condition. Additivity is imposed on these natural parameters. The identification target is the correspondence between factors of observationally equivalent generative models; it does not mean exact recovery of the latent state of an individual noisy sample.

B Proofs of Identification and Selectivity Results

This appendix presents the arguments from the identification criterion in Section 3 to global factor correspondence and selective use. The derivation of a common comparison map and natural-parameter correspondence from observational equivalence was proved in Appendix A. Throughout this appendix, $G = F^{-1} \circ \widehat{F}$ denotes that comparison map, and equalities between latent coordinates refer to the deterministic correspondence induced by this map.

B.1 Equivalent Forms and Coordinate Invariance of the Criterion

In Eq. (6), each factor contributes $h = m + m(m+1)/2$ columns. For a nonempty set $I \subseteq [d]$, let $N = [d] \setminus I$, and let W_I, W_N denote the matrices collecting the derivative columns of the corresponding factors. Let $P_I : \mathbb{R}^{dh} \to \mathbb{R}^{h|I|}$ extract the coefficients associated with the factors in I.

Lemma 2 (Kernel, rank, and projected singular value)**.** *At a fixed latent point and in fixed local product coordinates, the following three conditions are equivalent:*

$$\ker W \subseteq \ker P_I, \tag{44}$$

$$\operatorname{rank} W - \operatorname{rank} W_N = h|I|, \tag{45}$$

$$\sigma_{\min}\big((I_q - \Pi_N)W_I\big) > 0. \tag{46}$$

Here, Π_N is the orthogonal projection onto $\operatorname{col} W_N$, with $\Pi_N = 0$ when $N = \varnothing$. We define $\sigma_{\min}(B) = \inf_{\|v\|_2=1} \|Bv\|_2$.

Proof. Let $B_I = (I_q - \Pi_N)W_I$. If $W_I c_I + W_N c_N = 0$, then $B_I c_I = 0$. Conversely, if $B_I c_I = 0$, then $W_I c_I \in \operatorname{col} W_N$, so there exists c_N that cancels it. Thus, the kernel condition is equivalent to injectivity of B_I, which is equivalent to $\sigma_{\min}(B_I) > 0$. Moreover, from $W_I = \Pi_N W_I + B_I$ we obtain the orthogonal direct sum

$$\operatorname{col} W = \operatorname{col} W_N \oplus \operatorname{col} B_I. \tag{47}$$

Hence, the increase in rank is $\operatorname{rank} B_I$, and requiring this value to equal the number of columns $h|I|$ is again equivalent to injectivity. □

Lemma 3 (Invariance under product-coordinate changes)**.** *The three conditions in Lemma 2 have the same truth value after changing the local coordinates of each latent factor by a C^2 diffeomorphism $s_j = \chi_j(t_j)$.*

Proof. Under the coordinate change $s_j = \chi_j(t_j)$ for factor j, the first- and second-order chain rules give

$$\widetilde{W}_j = W_j T_j, \qquad T_j = \begin{bmatrix} D\chi_j & C_j \\ 0 & \operatorname{Sym}^2(D\chi_j) \end{bmatrix}. \tag{48}$$

Here, C_j is the coefficient matrix formed from the second derivatives of χ_j, and $\operatorname{Sym}^2(D\chi_j)$ is the induced map on the space of symmetric second-order tensors. This notation follows the ordering of symmetric derivative columns in Eq. (6). Both $D\chi_j$ and its symmetric square are invertible, and therefore so is T_j. When $m = 1$, $T_j = \left[\begin{smallmatrix} \chi_j' & \chi_j'' \\ 0 & (\chi_j')^2 \end{smallmatrix}\right]$. Collecting these matrices into $T = \operatorname{diag}(T_1, \ldots, T_d)$ gives $\widetilde{W} = WT$, and T preserves the coefficient subspaces associated with I and N separately. Thus, the total rank and the irrelevant-factor column space are preserved, and so is the rank condition. The remaining two conditions are equivalent by Lemma 2. □

This result is independent of the choice of local charts for the factors. The numerical value of the minimum singular value depends on the scaling of coordinates and contrasts, so coordinates and normalization must be fixed when discussing quantitative stability.

B.2 Proof of Task-Relevant Factor Identification

Proof of Theorem 1. Let $\mathcal{D}_0 \subseteq \mathcal{S}$ be the dense set on which TRC holds. By Lemma 1, G is a global C^2 diffeomorphism.

Step 1: Mixed derivatives between different candidate factors. In local product coordinates, choose tangent directions v, w associated with distinct candidate factors $a \neq b$. Factor-wise additivity of the candidate contrasts gives $D^2\widehat{\Psi}[v, w] = 0$. Applying the chain rule to Lemma 1 yields

$$0 = \sum_{j=1}^{d} \left\{ D\psi_j D^2 G_j[v, w] + D^2\psi_j[DG_j v, DG_j w] \right\} = W(G(\widehat{s}))c_{v,w}. \tag{49}$$

The derivatives of ψ_j are evaluated at $G_j(\widehat{s})$. Writing $u_j = DG_j v$ and $z_j = DG_j w$, the coefficients for factor j are

$$\begin{aligned} (c_{v,w})_j &= \left(D^2 G_j[v, w],\ (q_{j,\alpha\beta})_{\alpha \leq \beta} \right), \\ q_{j,\alpha\beta} &= \begin{cases} (u_j)_\alpha (z_j)_\alpha, & \alpha = \beta, \\ (u_j)_\alpha (z_j)_\beta + (u_j)_\beta (z_j)_\alpha, & \alpha < \beta. \end{cases} \end{aligned} \tag{50}$$

We collect these coefficients into a single column vector following the column ordering in Eq. (6). In particular, a diagonal second-derivative column is multiplied by a single product, whereas an off-diagonal column is multiplied by the sum of two products. If $G(\widehat{s}) \in \mathcal{D}_0$, the kernel form of TRC gives

$$D^2 G_j[v, w] = 0, \qquad u_j z_j^\top + z_j u_j^\top = 0, \qquad j \in I_*. \tag{51}$$

The set $G^{-1}(\mathcal{D}_0)$ is also dense, and the derivative coefficients are continuous in each local coordinate chart. We therefore first extend the equalities for pairs of coordinate-basis directions to all points, and then obtain them for arbitrary v, w by bilinearity. By Lemma 3, this conclusion is independent of the chart.

Step 2: A unique candidate factor for each relevant factor. Let $A_{ja} = D_a G_j$. This is a linear map from the m-dimensional tangent space of candidate factor a to the m-dimensional tangent space of true factor j. If, at some point, both A_{ja} and A_{jb} are nonzero for distinct a, b, then we can choose directions such that $A_{ja} v \neq 0$ and $A_{jb} w \neq 0$. However, for real vectors,

$$\|uz^\top + zu^\top\|_F^2 = 2\|u\|^2\|z\|^2 + 2(u^\top z)^2,$$

so Eq. (51) forces $u = 0$ or $z = 0$, a contradiction. Hence, for a relevant factor j, the corresponding block of Jacobian rows contains at most one nonzero $m \times m$ block. Because DG is invertible, these m rows are linearly independent. Therefore, exactly one such block is nonzero, and its rank is m.

For each $a \in [d]$, define

$$\mathcal{A}_{ja} = \{\widehat{s} : D_a G_j(\widehat{s}) \neq 0\}. \tag{52}$$

By continuity, these sets are open, and for a fixed relevant j they are pairwise disjoint and cover the connected space $\widehat{\mathcal{S}}$. Therefore, exactly one of them equals the entire space. Let its index be $\pi(j)$. Then

$$D_{\pi(j)} G_j \text{ is invertible at every point}, \qquad D_a G_j = 0 \quad (a \neq \pi(j)). \tag{53}$$

If distinct relevant factors j, k used the same candidate factor, then the corresponding $2m$ rows of DG would be supported on the same m columns and hence would be linearly dependent. This contradicts invertibility, so $\pi : I_* \to [d]$ is injective. Here we use the assumption that true and candidate factors have the same dimension.

Step 3: Dependence on one entire factor. Fixing $\widehat{s}_{\pi(j)} = t$ defines a fiber equal to the product of the remaining connected factor spaces. A connected manifold can be joined by piecewise smooth paths, and Eq. (53) implies that the derivative of G_j vanishes along every tangent direction of this fiber. Thus, G_j is constant on the fiber and can be written as

$$G_j(\widehat{s}) = \rho_j(\widehat{s}_{\pi(j)}), \qquad j \in I_*. \tag{54}$$

Restricting G_j to a smooth section obtained by fixing the other factors at reference values gives $\rho_j \in C^2$. Since $D\rho_j$ is invertible, ρ_j is a local diffeomorphism, and since G is surjective, ρ_j is also surjective.

Step 4: Global invertibility of the factor maps. For any $y \in \mathcal{S}_j$,

$$B_y = \{s \in \mathcal{S} : s_j = y\} \tag{55}$$

is connected because it is a product of the remaining connected factor spaces. Since G is a homeomorphism, the preimage $G^{-1}(B_y)$ is also connected. By Eq. (54),

$$G^{-1}(B_y) \cong \rho_j^{-1}(\{y\}) \times \prod_{a \neq \pi(j)} \widehat{\mathcal{S}}_a. \tag{56}$$

Local invertibility implies that $\rho_j^{-1}(\{y\})$ is discrete. Projecting the connected space on the left onto candidate factor $\pi(j)$ gives this entire discrete fiber. Because the continuous image of a connected space is connected, this fiber consists of a single point. Thus, ρ_j is bijective, and its local C^2 inverse maps combine into a global inverse.

Step 5: Common task–factor correspondence. Absorbing constant $r_{\tau j}$ terms into the intercept, natural-parameter transfer gives

$$\widehat{\eta}_\tau(\widehat{s}) = \bar{b}_\tau + \sum_{j \in I_\tau} r_{\tau j}(\rho_j(\widehat{s}_{\pi(j)})). \tag{57}$$

The right-hand side does not depend on factors outside $\pi(I_\tau)$. For $a = \pi(j)$ with $j \in I_\tau$,

$$D\widehat{r}_{\tau a} = (Dr_{\tau j} \circ \rho_j) D\rho_j. \tag{58}$$

Because $r_{\tau j}$ is nonconstant, ρ_j is surjective, and $D\rho_j$ is invertible, this derivative cannot vanish identically. An additive decomposition on a product space is unique up to factor-wise additive constants, so $\widehat{I}_\tau = \pi(I_\tau)$. Extending π to a permutation $\widetilde{\pi}$ of all indices gives

$$\widehat{M}_{\tau, \widetilde{\pi}(j)} = M_{\tau j}, \qquad \tau \in [T],\ j \in [d]. \tag{59}$$

This extension also aligns the zero columns of the incidence matrix; it does not claim factor-wise correspondence of the states of irrelevant factors. □

Almost-everywhere conditions. If the condition holds almost everywhere with respect to a full-support volume measure, then its validity set intersects every nonempty open set and is therefore dense. Hence, the main theorem also applies in this case. This argument does not turn a positive minimum singular value on a dense set into a uniformly positive global lower bound.

B.3 Special Cases and the Dimension Assumption

Full identification. If $I_* = [d]$, the irrelevant rank is zero, so TRC reduces to $\operatorname{rank} W = dh$. The injective map in Theorem 1 is then a permutation of all factors, and all factors are identified.

One task and the environment. Restrict the observations to $P(X, U)$ and $P(X, U, Y_\tau)$, and define $W^{(\tau)} = [W_A^\top, W_{T,\tau}^\top]^\top$. If

$$\operatorname{rank} W^{(\tau)} - \operatorname{rank} W^{(\tau)}_{I_\tau^c} = h|I_\tau| \tag{60}$$

holds on a dense set, then the same theorem applies and identifies I_τ. If $I_\tau = \varnothing$, the claim about relevant factors is vacuous, and parameter transfer implies that the candidate task is also constant. When the same candidate model is considered across multiple tasks, the resulting correspondences agree through the unique active Jacobian blocks; this does not imply automatic alignment of coordinate indices across separately trained models.

When $m = 1$, we recover the scalar case with $h = 2$. The assumption that true and candidate factors have the same dimension is needed in Step 2. If only the total dimension is fixed, a single larger candidate factor may combine multiple true factors, so TRC should not be interpreted as estimating the factor dimension partition itself.

B.4 Completion of the Environmental Deficit

At a fixed latent point, define $K_A = \ker W_A$ and $K_A^0 = K_A \cap \ker P_{I_*}$, and let

$$\nu_{A \to I_*} = \dim K_A - \dim K_A^0. \tag{61}$$

The space K_A consists of coefficient directions allowed by the environmental mixed-derivative equations, and K_A^0 consists of those directions that do not contribute to the relevant coefficients.

Proposition 3 (Quotient completion)**.** *Since $W_{T,N_*} = 0$, W_T induces a linear map on the quotient space K_A/K_A^0,*

$$\overline{W}_T : K_A/K_A^0 \longrightarrow \mathbb{R}^{q_T}, \qquad \overline{W}_T([c]) = W_T c. \tag{62}$$

The kernel condition for the relevant factors is equivalent to injectivity of this map, which is in turn equivalent to

$$\ker(W_T|_{K_A}) = K_A^0. \tag{63}$$

Under this condition,

$$\operatorname{rank}(W_T|_{K_A}) = \nu_{A \to I_*} \le q_T. \tag{64}$$

Proof. Since $W_{T,N_*} = 0$, we have $K_A^0 \subseteq \ker(W_T|_{K_A})$, so the quotient map is well defined. Moreover, $\ker W = \ker W_A \cap \ker W_T = \ker(W_T|_{K_A})$, so the kernel condition of TRC is precisely the reverse inclusion, namely Eq. (63). The kernel of the quotient map is $\ker(W_T|_{K_A})/K_A^0$, making injectivity equivalent to the same condition. Rank–nullity then gives $\operatorname{rank}(W_T|_{K_A}) = \dim K_A - \dim K_A^0 = \nu_{A \to I_*}$, and the codomain has dimension q_T. □

The quantity $\nu_{A \to I_*}$ is the dimension of the relevant coefficient directions that remain after imposing the environmental constraints in this mixed-derivative linear system. The task contrasts must therefore provide enough independent constraints to distinguish these directions. Here, q_T is the number of selected scalar response-contrast rows, not the number of tasks. A single vector-valued task may contribute multiple rows.

Exactness within the relaxed coefficient system. The kernel condition is, by definition, equivalent to

$$\big[Wc = 0 \implies P_{I_*} c = 0 \text{ for every } c \in \mathbb{R}^{dh}\big]. \tag{65}$$

In this sense, the condition is necessary and sufficient. However, coefficients arising from an actual comparison map have the form in Eq. (50) and, for every pair of directions from distinct candidate factors, must arise from derivatives of the same C^2 invertible map. The preceding argument does not include an integrability result showing that an arbitrary kernel element can be realized by derivatives of such a map. Therefore, failure of the rank condition alone does not imply non-identifiability of the observational model. To establish actual non-identifiability under a single observation source, we instead construct a separate admissible alternative model that preserves the observable law.

B.5 Selective Representation and Task Information Preservation

Proof of Corollary 1. Under the coupling $S = G(\widehat{S})$, factor-wise invertibility gives

$$Z_\tau^* = H_\tau(S_{I_\tau}), \qquad H_\tau(s_{I_\tau}) = (\rho_j^{-1}(s_j))_{j \in I_\tau}. \tag{66}$$

The product over an empty relevance set is interpreted as a single point. Because the response kernel depends only on the relevant state s_{I_τ}, define $K_\tau(\cdot \mid z)$ by substituting $H_\tau^{-1}(z)$ into that kernel. Then

$$P(Y_\tau \in \cdot \mid S, U) = K_\tau(\cdot \mid Z_\tau^*), \qquad Y_\tau \perp (S_{I_\tau^c}, U) \mid Z_\tau^*. \tag{67}$$

Moreover, the prior factorization gives $S_{I_\tau} \perp S_{I_\tau^c} \mid U$, and independence is preserved under the measurable map H_τ, yielding $Z_\tau^* \perp S_{I_\tau^c} \mid U$. Under an independent evaluation distribution $Q_S = Q_{I_\tau} \otimes Q_{I_\tau^c}$, the same deterministic correspondence gives unconditional independence. □

The relation between population readouts and finite-sample evaluation metrics is discussed in Appendix C.6.

B.6 Selective Editing under Full Factor Correspondence

Proof of Corollary 2. If $I_* = [d]$, then π is a permutation. Changing only candidate factor $\pi(j)$ gives

$$s'_k = \rho_k(\widehat{s}'_{\pi(k)}) = \rho_k(\widehat{s}_{\pi(k)}) = s_k \quad (k \neq j). \tag{68}$$

Any target value $s_j^\dagger$ can be obtained by setting $\widehat{s}'_{\pi(j)} = \rho_j^{-1}(s_j^\dagger)$, and because $\widehat{F} = F \circ G$, the edited signal also satisfies $\widehat{F}(\widehat{s}') = F(s')$. Under partial identification, the same equality applies only to $k \in I_* \setminus \{j\}$. An example in which an unidentified irrelevant factor changes is given in Appendix C.5. □

This result concerns exact state correspondence between observationally equivalent models. Target editing error and non-target changes produced by an actual encoder–decoder are evaluated separately in the experiments.

C Comparison with Prior Results and Further Analysis

We first compare our conditions with prior results and then present constructions supporting complementarity and residual nuisance ambiguity. We next summarize the editing limitations under partial identification and the interpretation of the evaluation metrics.

C.1 Relation to Prior Identifiability Results

The comparison axes are the information source, the function on which structure is imposed, and the identification target. Because the conditions are compared across different candidate-model classes, the table does not imply a uniform ordering of the overall strength of the results.

Approach	Main constraint	Relation to TCI
Auxiliary nonlinear ICA, iVAE (Hyvarinen et al., 2019; Khemakhem et al., 2020a)	Sufficient variation of conditional distributions or sufficient statistics	TCI requires separation of the relevant columns after excluding the irrelevant column space in the combined contrasts.
IMCA and partial identification (Khemakhem et al., 2020b; Kong et al., 2022; Sorrenson et al., 2020)	Nonfactorized base measure or identification of modulated components	Partial identification is common to both settings; TCI defines the target through task relevance and also allows modulation of irrelevant factors.
Sparse multi-task, G2S (Lachapelle et al., 2023a; Zheng et al., 2026)	Sparse and diverse predictor supports or relevant subgroup structure	TCI applies combined environmental and task constraints to a conditionally factorized latent distribution and an additive natural parameter.
Approaches based on generative structure (Lachapelle et al., 2023b; 2026; Liu et al., 2024; Li et al., 2025)	Decoder additivity, sparsity of mechanisms or mixing, or changes in causal mechanisms	TCI imposes additivity on the response natural parameter and does not require additivity of the observation function or edge minimization.
Multi-view and common contrasts (Gresele et al., 2020; Yao et al., 2024; Yi et al., 2026)	Complementary observations, common latent transformations, and contrast preservation	Within this framework, TCI specifies a projected rank condition that yields relevant factors and a common task correspondence.

TCI uses existing identifiability analyses based on common transformations and contrasts, and specifies the projected rank condition required for relevant factors and task correspondence. In particular, CMM provides an analysis based on common transformations and contrasts (Yi et al., 2026, Theorems 1–2 and Appendix E.1). By Bayes' rule, a task response can be written as

$$\log \frac{p(s \mid u, Y_\tau = y)}{p(s \mid u, Y_\tau = y_0)} = \log \frac{p_\tau(y \mid s)}{p_\tau(y_0 \mid s)} - \log \frac{p(Y_\tau = y \mid u)}{p(Y_\tau = y_0 \mid u)} \tag{69}$$

Therefore, the main distinction lies not in the direction itself, but in the admissible model class and the required rank condition. The conditions of TCI and actual complementarity are compared through the admissible alternative models and rank calculations in Appendices C.2, C.4. Unlike general linear correspondence (Roeder et al., 2021), the conclusion is a factor-wise invertible correspondence that excludes mixing among relevant factors.

C.2 A Construction Showing Complementary Constraints

This construction shows that combining the observations can remove actual mixing ambiguities allowed by either observation alone. Rather than concluding non-identifiability from rank deficiency, we provide alternative models that preserve the observable laws.

Let $\mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})$ carry the normalized Haar measure $d\lambda(s) = ds/(2\pi)$, and set $\mathcal{S} = \mathbb{T}^4$, $U \sim \mathrm{Unif}(\mathbb{T}^2)$, and $\kappa > 0$.

$$p(s \mid u) = B_\kappa^{-2} e^{\kappa\cos(s_1-u_1)+\kappa\cos(s_2-u_2)}, \qquad B_\kappa = \int_{\mathbb{T}} e^{\kappa\cos t}\, d\lambda(t). \tag{70}$$

S_3, S_4 are independent uniform factors. The logits of two conditionally independent Bernoulli responses and the noiseless observation function are

$$\eta_1(s) = \cos s_1 + \cos s_3, \qquad \eta_2(s) = \sin s_2 + \sin s_3, \qquad F(s) = (\cos s_j, \sin s_j)_{j=1}^4. \tag{71}$$

Thus, $I_1 = \{1,3\}$, $I_2 = \{2,3\}$, and $I_* = \{1,2,3\}$. The positive smooth product density, compact embedding, additive logits, and positive probability of every response value satisfy Assumptions 1–4.

Proposition 4 (Complementarity with residual nuisance ambiguity)**.** *In this model, the relevant factors are not identified if only $P(X,U)$ or only $P(X,Y_1,Y_2)$ is fixed. If $P(X,U)$ and both $P(X,U,Y_\tau)$ are fixed jointly, the relevant factors and task correspondence are identified, while individual identification of S_4 is not guaranteed.*

Proof. All transformations below are interpreted modulo 2π, and in each alternative model we use $\widehat{F} = F \circ H^{-1}$.

Fixing only the environmental observation. For $\alpha \neq 0$,

$$H_A(s) = (s_1, s_2, s_3 + \alpha \sin s_1 \sin s_2, s_4) \tag{72}$$

is a diffeomorphism that translates the conditionally uniform factor S_3. It therefore preserves every $p(s \mid u)$ and $P(X,U)$. Assigning the unobserved responses the same additive logits as in Eq. (71) in the candidate coordinates gives another admissible model. However, $G_3(\widehat{s}) = \widehat{s}_3 - \alpha \sin \widehat{s}_1 \sin \widehat{s}_2$ depends on multiple coordinates on an open set, so the relevant factor S_3 is not identified.

Fixing only the task observation. After integrating out U, the law of S is $\mathrm{Unif}(\mathbb{T}^4)$. Consider the flow Φ_t of the vector field

$$\mathcal{V}(s) = (\sin s_3 \cos s_2,\ \sin s_1 \cos s_3,\ -\sin s_1 \cos s_2,\ 0). \tag{73}$$

Direct differentiation gives $\mathcal{V}^\top \nabla\eta_1 = \mathcal{V}^\top \nabla\eta_2 = 0$ and $\operatorname{div} \mathcal{V} = 0$. A smooth flow on a compact manifold exists globally and satisfies $\Phi_t^{-1} = \Phi_{-t}$; the chain rule and the Jacobian flow equation then show that the flow preserves both logits and Haar measure. Using $H = \Phi_t$ and the same logits therefore preserves the entire $P(X, Y_1, Y_2)$. If the unobserved U of the alternative model is assigned the same uniform marginal law but made independent of $\widehat{S}$, then $\widehat{p}(\widehat{s} \mid u) = 1$, yielding an admissible model. Since $\mathcal{V}(0) = 0$ and $\partial_3 \mathcal{V}_1(0) = 1$, for small $t \neq 0$, $(D\Phi_{-t}(0))_{13} = -t + O(t^2) \neq 0$ and $(D\Phi_{-t}(0))_{11} = 1 + O(t^2) \neq 0$. Thus, the first factor of the comparison map is genuinely mixed.

Combined observations. The environmental contrasts between $u_0 = (0,0)$ and $(\pi,0), (\pi/2,0), (0,\pi), (0,\pi/2)$ reduce, under an invertible constant row transformation, to $\cos s_1, \sin s_1, \cos s_2, \sin s_2$. Adding the Bernoulli response contrasts η_1, η_2 also yields $\cos s_3, \sin s_3$. Hence, after a constant row transformation, the relevant derivative columns form the three blocks

$$D_j(s_j) = \begin{pmatrix} -\sin s_j & -\cos s_j \\ \cos s_j & -\sin s_j \end{pmatrix}, \qquad \det D_j = 1, \qquad j = 1,2,3. \tag{74}$$

The irrelevant derivative columns are zero, so $\operatorname{rank} W = 6$ and $\operatorname{rank} W_{\{4\}} = 0$ at every point. By TRC and Theorem 1, the three relevant factors and the common task correspondence are identified.

Residual ambiguity in the irrelevant factor. For $\beta \neq 0$,

$$H_N(s) = (s_1, s_2, s_3, s_4 + \beta \sin s_1) \tag{75}$$

translates only the conditionally uniform factor S_4 and preserves all conditional latent laws and both logits. Thus, the same response mechanisms together with $\widehat{F} = F \circ H_N^{-1}$ preserve the entire $P(X, U, Y_1, Y_2)$, while $G_4(\widehat{s}) = \widehat{s}_4 - \beta \sin \widehat{s}_1$ is not factor-wise. □

C.3 A Conditional Representation of the Task Marginals

Under Assumptions 1–4, let $K_\tau(dy_\tau \mid v, u)$ be regular conditional kernels for the noise-separated observable marginals. Then we may define

$$P^\dagger(dv, du, dy_{1:T}) = P_{VU}(dv, du) \prod_{\tau=1}^{T} K_\tau(dy_\tau \mid v, u). \tag{76}$$

For finitely many kernels, differences between versions occur only on P_{VU}-null sets, and integrating the remaining kernels gives one, so each (V, U, Y_τ) marginal equals the observed law. Thus, observationally equivalent models induce the same $P^\dagger$, although this does not recover the actual task coupling.

Suppose additionally that the conditionally exponential-family prior has the form $p_j(s_j \mid u) = Q_j(s_j) \exp\{\lambda_j(u)^\top T_j(s_j) - B_j(u)\}$. Since, in latent coordinates, $p^\dagger(s, y_{1:T} \mid u) = p(s \mid u) \prod_\tau p_\tau(y_\tau \mid s)$, for conditioning values at which the normalizing constant is positive and finite,

$$p^\dagger(s \mid u, y_{1:T}) \propto \mu(s) \exp\left\{ \sum_j \lambda_j(u)^\top T_j(s_j) + \sum_{\tau, j} t_\tau(y_\tau)^\top r_{\tau j}(s_j) \right\}, \tag{77}$$

$$\mu(s) = \left(\prod_j Q_j(s_j) \right) \exp\left\{ - \sum_\tau A_\tau(\eta_\tau(s)) \right\}. \tag{78}$$

This expression is obtained by substituting the response exponential families and absorbing terms independent of s into the normalizing constant. It is related to the IMCA form that allows a nonfactorized base measure (Khemakhem et al., 2020b), but the existence of this representation alone does not guarantee additional conditions of identifiability theorems, such as independent modulation of sufficient statistics. No step assumes conditional independence among the actual tasks.

C.4 Identifying a Relevant Factor with Modulated, Ambiguous Nuisance

All factors in this subsection are scalar, so $m = 1$.

Even when all derivative columns are not independent, the two columns of a relevant coordinate may still be separated from the irrelevant column space. We show this in a normalizable generative model and also verify the ambiguity of the irrelevant factors through an actual observationally equivalent transformation. This example is an analytical construction illustrating the scope of Theorem 1.

Generative model. Let the latent and environment spaces be

$$\mathcal{S} = \mathbb{R} \times (1, 2) \times \mathbb{R}^2, \qquad \mathcal{U} = \{0, e_1, e_2, e_3, e_4\}, \qquad P(U = u) = \frac{1}{5}, \tag{79}$$

where e_ℓ is the ℓth standard basis vector of $\mathbb{R}^4$. We use the product Lebesgue measure on the latent space. The components are conditionally independent given U, with

$$S_1 \mid U = u \sim \mathcal{N}\big(0, (1 + 2u_1)^{-1}\big), \tag{80}$$

$$p_2(t \mid u) = \frac{\exp\{\sum_{\ell=1}^4 u_\ell t^\ell\}}{Z_2(u)}, \qquad t \in (1, 2), \qquad Z_2(u) = \int_1^2 \exp\left\{ \sum_{\ell=1}^4 u_\ell t^\ell \right\} dt, \tag{81}$$

$$S_3 \mid U = u \sim \mathcal{N}(u_4, 1), \qquad S_4 \mid U = u \sim \mathcal{N}(0, 1). \tag{82}$$

The response and observation are

$$Y \mid S = s \sim \mathcal{N}(s_1, 1), \qquad I_* = \{1\}, \quad N_* = \{2, 3, 4\}, \tag{83}$$

$$X = F(S), \qquad F(s) = A \begin{pmatrix} s_1 \\ \log \dfrac{s_2 - 1}{2 - s_2} \\ s_3 \\ s_4 \end{pmatrix}, \qquad A = I_4 + \mathbf{1}\mathbf{1}^\top. \tag{84}$$

Here, $\mathbf{1} = (1, 1, 1, 1)^\top$. The diagonal entries of A are 2, the off-diagonal entries are 1, and its eigenvalues are $5, 1, 1, 1$. Thus, A is invertible and each observation coordinate depends on every latent coordinate. The logit transformation in the second coordinate is a C^∞ diffeomorphism from $(1, 2)$ to $\mathbb{R}$, so $F : \mathcal{S} \to \mathbb{R}^4$ is also a surjective diffeomorphism and is proper. All latent densities are positive and smooth in the interior, and $0 < Z_2(u) < \infty$. The Gaussian response is a minimal regular exponential family, and its natural parameter $\eta(s) = s_1$ is additive. Therefore, this model satisfies the noiseless case of Assumptions 1–4. The same argument applies after noise separation if the same known independent Gaussian observation noise is added. For alternative models, the task support is not fixed to the true support; we allow $\widehat{\eta}(\widehat{s}) = \widehat{b} + \sum_{j=1}^{4} \widehat{r}_j(\widehat{s}_j)$.

Contrasts and projected rank. For the four environmental contrasts relative to the reference environment 0 and the response contrast between $y = 1$ and $y = 0$, omitting constants independent of the state gives

$$\Psi_A(s) = \begin{pmatrix} -s_1^2 + s_2 \\ s_2^2 \\ s_2^3 \\ s_2^4 + s_3 \end{pmatrix}, \qquad \Psi_T(s) = s_1. \tag{85}$$

Dropping constants does not affect the derivative matrix. Writing $t = s_2 \in (1, 2)$ and ordering the columns by the first and second derivatives of each component gives

$$W(s) = \begin{pmatrix} -2s_1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2t & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3t^2 & 6t & 0 & 0 & 0 & 0 \\ 0 & 0 & 4t^3 & 12t^2 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{86}$$

The irrelevant column space is generated by

$$n_1 = (1, 2t, 3t^2, 4t^3, 0)^\top, \qquad n_2 = (0, 2, 6t, 12t^2, 0)^\top, \qquad n_3 = (0, 0, 0, 1, 0)^\top. \tag{87}$$

The minor formed by the second and third rows of n_1, n_2 is

$$\det \begin{pmatrix} 2t & 2 \\ 3t^2 & 6t \end{pmatrix} = 6t^2 > 0, \tag{88}$$

so n_1, n_2, n_3 are linearly independent. Now let $f_1, \ldots, f_5$ be the standard basis of $\mathbb{R}^5$. If $af_1 + bf_5 = c_1 n_1 + c_2 n_2 + c_3 n_3$, the fifth row gives $b = 0$, and the second and third rows together with Eq. (88) give $c_1 = c_2 = 0$. The first and fourth rows then give $a = c_3 = 0$. Hence, $\operatorname{span}\{f_1, f_5\}$ intersects the irrelevant column space only at $\{0\}$. The two relevant columns are $-2s_1 f_1 + f_5$ and $-2f_1$, so

$$\operatorname{rank} W(s) = 5, \qquad \operatorname{rank} W_{N_*}(s) = 3, \qquad 5 - 3 = 2 = 2|I_*|, \qquad 5 < 2d = 8. \tag{89}$$

Thus, the full-column-rank condition fails, while the condition for the relevant coordinate holds at every point. By Theorem 1, S_1 and the task–factor correspondence are identified.

Actual nuisance ambiguity. The ambiguity of the components excluded from the rank calculation can also be realized by an actual alternative model. Using the rotation matrix

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \tag{90}$$

define

$$\widehat{S}_1 = S_1, \qquad \widehat{S}_2 = S_2, \qquad (\widehat{S}_3, \widehat{S}_4)^\top = R_\theta (S_3, S_4)^\top. \tag{91}$$

Call this transformation H_θ. Since the conditional mean of $(S_3, S_4) \mid U = u$ is $(u_4, 0)^\top$ and its covariance is I_2, the rotated mean is $R_\theta(u_4, 0)^\top$ and the covariance remains I_2. Thus, the two rotated Gaussian components remain conditionally independent given U, and their conditional independence from (S_1, S_2) is also preserved. Define the alternative decoder and response by

$$\widehat{F} = F \circ H_\theta^{-1}, \qquad \widehat{Y} \mid \widehat{S} = \widehat{s} \sim \mathcal{N}(\widehat{s}_1, 1). \tag{92}$$

Then all observable laws are unchanged. Because H_θ is a diffeomorphism with determinant one and $\widehat{F}$ is also a surjective diffeomorphism, the geometric conditions for the candidate model are satisfied. If $\sin\theta\cos\theta \neq 0$, the two irrelevant components are mixed in a non-factorwise manner. Thus, actual non-identifiability remains in the irrelevant part, including a component that responds to the environment, while the relevant coordinate is identified.

No nontrivial constant environmental contrast removes the nuisance. For $a \in \mathbb{R}^4$, the irrelevant part of the environmental contrast $a^\top \Psi_A$ is

$$a_1 s_2 + a_2 s_2^2 + a_3 s_2^3 + a_4(s_2^4 + s_3). \tag{93}$$

If this function is constant in (s_2, s_3, s_4), differentiation with respect to s_3 gives $a_4 = 0$, and comparing polynomial coefficients on $(1, 2)$ gives $a_1 = a_2 = a_3 = 0$. Therefore, there is no nontrivial constant linear environmental contrast that removes the variation of the irrelevant distribution. In particular, the irrelevant part cannot have the same law under two distinct observed environments. If the environment is fixed to a single value, the remaining response contrast is the monotone statistic s_1.

Conditional representation and tied modulation. Applying the conditional representation of Appendix C.3 to this example, the nonconstant modulated sufficient statistics in the original coordinates may be taken as

$$T_1 = (s_1, s_1^2), \qquad T_2 = (s_2, s_2^2, s_2^3, s_2^4), \qquad T_3 = s_3. \tag{94}$$

There is no modulated statistic for s_4. These seven functions are linearly independent up to constants, but their coefficient vectors vary only along the following five directions:

$$(y, -u_1, u_1, u_2, u_3, u_4, u_4)^\top. \tag{95}$$

Fixed coefficients are absorbed into the base measure. The coefficients of s_1^2 and s_2 are tied to the same u_1, while those of s_2^4 and s_3 are tied to the same u_4. Hence, in this representation, the coefficient-variation matrix for the seven statistics has rank at most five, so a condition requiring full rank of this matrix is not satisfied.

This difference does not disappear automatically under an invertible reparameterization of only the irrelevant coordinates while leaving the relevant coordinate unchanged. Indeed, the irrelevant functions

$$s_2, \quad s_2^2, \quad s_2^3, \quad s_2^4 + s_3 \tag{96}$$

are linearly independent up to constants, and composition with an invertible reparameterization of the irrelevant coordinates preserves their linear independence. The relevant coordinate requires the two independent functions s_1, s_1^2. Therefore, if there exists a factor-wise statistic representation in which the relevant and irrelevant coordinates are separated, the total dimension of the modulated statistics must be at least six. By contrast, the function space of conditional log-density differences is the five-dimensional space generated by the five functions in Eq. (85), and its dimension is preserved under a fixed invertible reparameterization. Thus, within such reparameterizations, independent coefficient variation over the full set of statistics cannot be obtained.

This example shows that TRC can identify a relevant factor under weaker requirements than independence of all derivative columns and full statistic variation in a representation that separates relevant and irrelevant factors. This comparison is within the model class above and does not imply an overall ordering of different identifiability theorems or non-identifiability under each individual information source. The latter is established by the separate construction in Appendix C.2.

C.5 Partial Identification and Nuisance Changes under Editing

Reuse the observationally equivalent nuisance shear from Appendix C.2, $H_N(s) = (s_1, s_2, s_3, s_4 + \beta \sin s_1)$. Its inverse satisfies $G_4(\widehat{s}) = \widehat{s}_4 - \beta \sin \widehat{s}_1$ and $G_j(\widehat{s}) = \widehat{s}_j$ for $j = 1, 2, 3$. If $\widehat{s}_1$ is edited from a to b while the other candidate coordinates are held fixed, then

$$s_2' = s_2, \quad s_3' = s_3, \qquad s_4' - s_4 = -\beta(\sin b - \sin a) \pmod{2\pi}. \tag{97}$$

For a, b such that the right-hand side is not in $2\pi\mathbb{Z}$, the irrelevant factor changes. Because this model satisfies TRC and exact correspondence of the relevant factors, it provides a counterexample showing that partial identification preserves other identified factors, but not necessarily unidentified nuisance factors.

C.6 Factor Correspondence and Readout Metrics

Invertibility versus linear prediction. If $S \sim \mathcal{N}(0, 1)$ and $T = S + S^3$, then $dT/dS = 1 + 3S^2 > 0$, so the transformation is invertible. However, $\mathrm{Cov}(S, T) = 4$ and $\mathrm{Var}(T) = 22$, giving

$$\mathrm{Corr}(S, T) = 4/\sqrt{22}, \qquad R^2_{\mathrm{linear}}(S \mid T) = 8/11 < 1. \tag{98}$$

The nonlinear inverse permits exact recovery. Therefore, a linear readout or Pearson-based MCC below one does not by itself imply failure of factor-wise identification.

Independent evaluation and nuisance readout. Consider $Z^*_\tau = H_\tau(S_{I_\tau})$ from Corollary 1 under the independent evaluation distribution $Q_{I_\tau} \otimes Q_{I^c_\tau}$. If $N_\tau = g(S_{I^c_\tau})$ is a real-valued readout with finite positive variance, then $\mathbb{E}_Q[N_\tau \mid Z^*_\tau] = \mathbb{E}_Q[N_\tau]$. Since the conditional expectation is the optimal squared-error predictor,

$$\inf_f \mathbb{E}_Q[(N_\tau - f(Z^*_\tau))^2] = \mathrm{Var}_Q(N_\tau), \qquad R^2_{Q,\mathrm{opt}} = 0. \tag{99}$$

The infimum is over all square-integrable predictors. This is a statement about the population-optimal predictor; a finite-sample linear probe measures empirical performance within a restricted function class.

Correlations induced through the environment may remain under the natural observational distribution. Let $U \in \{-1, 1\}$ be uniform and, for independent $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, define $S_i = U + \epsilon_i$ for $i = 1, 2$ and $T = S_1$. Then $S_1 \perp S_2 \mid U$, but $R^2_{\mathrm{linear}}(S_2 \mid T) = 1/(1 + \sigma^2)^2 > 0$. Accordingly, in the experiments we distinguish the structure of the representation from the evaluation distribution, and report readouts under independent clean factors separately from prediction performance under the noisy observational distribution.

D Estimation and Variational Approximation

This appendix connects the objective used to fit the observable information in Section 2 with the identification result in Section 3. We first give an exact decomposition of the conditional composite log-likelihood and ELBO, and then establish the relation among generative-model fit error, variational approximation error, sampling error, and optimization error. Each S_j is an m-dimensional latent factor, and the total dimension is $D = dm$. The density and KL identities below are independent of the factor dimension. These results specify the conditions under which the identification theorem can be applied to the estimation procedure.

D.1 Dominated Models and Shared Marginals

We treat the law P_U of the environment as a fixed observational design distribution. The shared generative parameter θ includes the observation function, the conditional prior, and all task response functions. For the likelihood and ELBO calculations in this appendix, we restrict the admissible models of Appendix A to the following subclass.

1. We fix a common volume measure $\lambda_{\mathcal{S}}$ on the latent space used for estimation. For every candidate model, the observation density $p_\theta(x \mid s)$ exists with respect to a common σ-finite measure μ_X. Task responses use the same minimal regular exponential families and dominating measures ν_τ specified in Assumption 2.

2. Let $O_0 = X$ and $O_\tau = (X, Y_\tau)$, and define $\mu_0 = \mu_X$ and $\mu_\tau = \mu_X \otimes \nu_\tau$. The true conditional law $P_{*,k}(\cdot \mid u)$ and every candidate conditional law $P_{\theta,k}(\cdot \mid u)$ have densities with respect to μ_k. The candidate evidence $p_{\theta,k}(O_k \mid U)$ is positive and finite almost surely under the true observable law.
3. The variational distribution $q_{\theta,\phi,k}(\cdot \mid o_k, u)$ has a density with respect to $\lambda_{\mathcal{S}}$ and is absolutely continuous with respect to the candidate posterior $p_{\theta,k}(\cdot \mid o_k, u)$. The log-density expectations, ELBOs, and posterior KL divergences compared below are well defined and finite. In particular, $\mathcal{Q}_*$ in Eq. (103) is finite.
4. Candidate generative models satisfy the constraints of Section 2 and Assumptions 1–4. All branches use the same generative parameter θ, and the fixed weights $w_0, w_1, \ldots, w_T$ are all positive.

The scalar control experiment with positive Gaussian observation noise provides a density-based model of this form. By contrast, a noiseless embedding allowed in Appendix A may induce a singular joint law of observations and latent states, or a singular posterior. In that case, the domination and variational-distribution conditions above must be verified separately.

The branch-specific joint densities of the shared generative model are

$$p_{\theta,0}(x, s \mid u) = p_\theta(x \mid s) \prod_{j=1}^{d} p_{\theta j}(s_j \mid u), \tag{100}$$

$$p_{\theta,\tau}(x, y, s \mid u) = p_{\theta,0}(x, s \mid u) p_{\theta,\tau}(y \mid s), \qquad \tau \in [T], \tag{101}$$

$$p_{\theta,k}(o_k \mid u) = \int_{\mathcal{S}} p_{\theta,k}(o_k, s \mid u)\, d\lambda_{\mathcal{S}}(s). \tag{102}$$

For an additive-noise model with density p_ε, $p_\theta(x \mid s) = p_\varepsilon(x - F_\theta(s))$. The response natural parameter is $\eta_{\theta,\tau}(s) = b_{\theta,\tau} + \sum_j r_{\theta,\tau j}(s_j)$, and this additive constraint is distinct from the structure of the observation function.

Allowing $q_{\theta,\phi,k}$ to depend on θ includes implementations in which a variational distribution in computational coordinates is pushed forward through a learned generative coordinate transformation. An encoder independent of the generative parameters is a special case. The vector ϕ collects the encoder parameters of all branches, together with any sharing constraints among branches.

D.2 Composite Decomposition and Population Identification

A weighted sum of marginal log-likelihoods is a composite log-likelihood (Varin et al., 2011). Because each term is averaged under its corresponding true observable marginal law, neither conditional independence among task responses nor their simultaneous observation is required. Using $\mathcal{Q}, \mathcal{E}_k, \mathcal{J}, \Delta$ from Section 4.1, define

$$\mathcal{Q}_* := \sum_{k=0}^{T} w_k \mathbb{E}_{P_{*,k} P_U}[\log p_{*,k}(O_k \mid U)]. \tag{103}$$

Applying the definition of KL divergence and the standard ELBO identity (Kingma & Welling, 2013) to each branch gives

$$\mathcal{Q}_* - \mathcal{Q}(\theta) = \sum_{k=0}^{T} w_k \mathbb{E}_U \operatorname{KL}(P_{*,k}(\cdot \mid U) \| P_{\theta,k}(\cdot \mid U)) = \mathcal{L}(\theta), \tag{104}$$

$$\mathcal{E}_k = \log p_{\theta,k}(O_k \mid U) - \operatorname{KL}(q_{\theta,\phi,k}(\cdot \mid O_k, U) \| p_{\theta,k}(\cdot \mid O_k, U)), \tag{105}$$

$$\mathcal{J}(\theta, \phi) = \mathcal{Q}_* - \mathcal{L}(\theta) - \Delta(\theta, \phi). \tag{106}$$

Here, Δ is the weighted sum of posterior KL divergences in Eq. (16). Both the observable-fit error $\mathcal{L}$ and the approximation error Δ to the current candidate posterior are nonnegative. These identities hold at every fixed parameter pair even when q depends on θ.

Proof of Proposition 1. If a single pair (θ_*, ϕ_*) simultaneously represents all true marginal laws and all corresponding branch posteriors, then $\mathcal{L}(\theta_*) = \Delta(\theta_*, \phi_*) = 0$. Since the upper bound $\mathcal{Q}_*$ in

Eq. (106) is attained, every global maximizer also satisfies $\mathcal{L}(\widehat{\theta}) = \Delta(\widehat{\theta}, \widehat{\phi}) = 0$. Because every $w_k > 0$, each observable KL divergence vanishes P_U-almost everywhere, and combining with the common P_U yields $\mathcal{O}_J(\mathcal{M}_{\widehat{\theta}}) = \mathcal{O}_J(\mathcal{M}^*)$. Under TRC and the assumptions defining the admissible model class, Theorem 1 then gives the result. □

The principle of linking variational estimation and identification through posterior realizability is also used in the estimation analysis of iVAE (Khemakhem et al., 2020a). Here, simultaneous realizability of all branches is required only at one true parameter pair; exact posterior representation at every candidate parameter or convergence of practical training is not assumed.

The observation branch. Integrating the normalized task kernel gives $\int p_{\theta,\tau}(x, y \mid u) d\nu_\tau(y) = p_{\theta,0}(x \mid u)$. Thus, exact fitting of a task marginal also determines the base observation law, while the base branch directly learns response-free inference and allows the use of samples without responses.

The branch-wise expansions used in computation are given in Eq. (14). The prior and decoder are shared across all branches, and each task branch adds its corresponding response term.

D.3 Task-Conditioned and Shared Posterior Families

Task responses are included in the posterior input because adding Y_τ to the same observation (X, U) can change the posterior of the candidate generative model. For a fixed θ, write the two posteriors as

$$p_0^\theta := p_{\theta,0}(S \mid X, U), \qquad p_\tau^\theta := p_{\theta,\tau}(S \mid X, Y_\tau, U). \tag{107}$$

They condition on different inputs but are distributions on the same latent space. Let $\Delta_{\mathrm{sh}}(\theta, \phi)$ denote Eq. (16) when every branch uses a single response-free variational distribution $q^{\mathrm{sh}} = q^{\mathrm{sh}}_{\theta,\phi}(S \mid X, U)$.

Proposition 5 (An unavoidable gap for a shared posterior)**.** *For task τ, define*

$$D_\tau(\theta) := \mathbb{E}_{(X,Y_\tau,U)\sim P_{*,\tau}P_U}[\mathrm{TV}(p_0^\theta, p_\tau^\theta)^2]. \tag{108}$$

Then every admissible shared variational posterior satisfies

$$\Delta_{\mathrm{sh}}(\theta, \phi) \geq \frac{2w_0 w_\tau}{w_0 + w_\tau} D_\tau(\theta). \tag{109}$$

If $\theta = \theta_*$ represents the true marginal laws, then $D_\tau(\theta_*) = 0$ is equivalent to $S \perp Y_\tau \mid (X, U)$ under the generative model $\mathcal{M}_{\theta_*}$.

Proof. At each (x, y_τ, u), the triangle inequality for total variation and the weighted Cauchy–Schwarz inequality give

$$\begin{aligned}
\mathrm{TV}(p_0^\theta, p_\tau^\theta)^2 &\leq \left[\mathrm{TV}(p_0^\theta, q^{\mathrm{sh}}) + \mathrm{TV}(q^{\mathrm{sh}}, p_\tau^\theta)\right]^2 \\
&\leq \left(\frac{1}{w_0} + \frac{1}{w_\tau}\right)\left[w_0\,\mathrm{TV}(p_0^\theta, q^{\mathrm{sh}})^2 + w_\tau\,\mathrm{TV}(q^{\mathrm{sh}}, p_\tau^\theta)^2\right] \\
&\leq \frac{w_0 + w_\tau}{2w_0 w_\tau}\left[w_0\,\mathrm{KL}(q^{\mathrm{sh}} \| p_0^\theta) + w_\tau\,\mathrm{KL}(q^{\mathrm{sh}} \| p_\tau^\theta)\right].
\end{aligned}$$

The last line uses Pinsker's inequality under the normalization $\mathrm{TV} = \frac{1}{2}\int |p - q|$. The integrand of the first KL term does not depend on Y_τ, so averaging it under the true task observation law gives the same expectation as under the base observation law. The second term is the posterior gap for task τ. Because the gaps from the remaining branches are nonnegative, Eq. (109) follows. At θ_*, the true task observation law agrees with the candidate observation law. Thus, $D_\tau(\theta_*) = 0$ is equivalent to $p_{\theta_*}(S \mid X, U, Y_\tau) = p_{\theta_*}(S \mid X, U)$ almost surely under the corresponding observation law, which is precisely the stated conditional independence. □

This result quantifies the case in which increasing the expressive power of a shared encoder still cannot exactly represent two distinct posteriors with a single distribution for the same input. If $D_\tau(\theta) > 0$, the aggregate gap for a shared posterior has a positive lower bound. This bound applies separately to each τ; task-specific lower bounds are not summed, because doing so would reuse the

same base term. Moreover, $D_\tau(\theta)$ depends on the generative parameter, so the proposition compares posterior families for a fixed generative model. If the response does not alter the posterior already determined by (X, U), this source of a positive lower bound disappears.

For a fixed θ, let $\mathcal{Q}_{\mathrm{sh}}(\theta)$ denote the shared posterior family and $\mathcal{Q}_{\mathrm{TC}}(\theta)$ the task-conditioned family. Each element specifies the variational distributions for all branches. If the task-conditioned encoder can recover the shared encoder by setting the response-dependent correction to zero, then $\mathcal{Q}_{\mathrm{sh}}(\theta) \subseteq \mathcal{Q}_{\mathrm{TC}}(\theta)$, and hence

$$\inf_{q \in \mathcal{Q}_{\mathrm{TC}}(\theta)} \Delta(\theta, q) \leq \inf_{q \in \mathcal{Q}_{\mathrm{sh}}(\theta)} \Delta(\theta, q). \tag{110}$$

This follows directly from inclusion of the distribution families. For fixed θ, the composite log-likelihood $\mathcal{Q}(\theta)$ is the same for both variational families, so the optimized population ELBO satisfies the reverse inequality. This comparison concerns the representational capacity of the variational families; the performance ordering of parameters obtained from different training runs is evaluated empirically.

D.4 Approximation Error and the Scope of the Guarantee

The empirical composite ELBO is computed from branch-specific samples representative of the corresponding marginal laws as

$$\widehat{\mathcal{J}}_{\mathrm{data}} = \sum_k \frac{w_k}{n_k} \sum_i \mathcal{E}_k(\theta, \phi; o_{ki}, u_{ki}).$$

In practical optimization, the expectations within each ELBO may also be approximated by Monte Carlo samples. Below, $\widehat{\mathcal{J}}$ denotes the actual optimization objective including these sample and Monte Carlo approximations. If the availability of responses is subject to selection bias, sampling or weighting must be corrected accordingly.

A conditional error comparison. Suppose the following two bounds hold over the same admissible parameter set $\mathcal{A}$:

$$\sup_{\mathcal{A}} |\widehat{\mathcal{J}} - \mathcal{J}| \leq \epsilon_{\mathrm{stat}}, \tag{111}$$

$$\widehat{\mathcal{J}}(\widehat{\theta}, \widehat{\phi}) \geq \sup_{\mathcal{A}} \widehat{\mathcal{J}} - \epsilon_{\mathrm{opt}}. \tag{112}$$

For any comparison point $a \in \mathcal{A}$,

$$\mathcal{J}(\widehat{\theta}, \widehat{\phi}) \geq \widehat{\mathcal{J}}(\widehat{\theta}, \widehat{\phi}) - \epsilon_{\mathrm{stat}} \geq \widehat{\mathcal{J}}(a) - \epsilon_{\mathrm{stat}} - \epsilon_{\mathrm{opt}} \geq \mathcal{J}(a) - 2\epsilon_{\mathrm{stat}} - \epsilon_{\mathrm{opt}}.$$

Substituting Eq. (106) and taking the infimum over comparison points gives

$$\mathcal{L}(\widehat{\theta}) + \Delta(\widehat{\theta}, \widehat{\phi}) \leq \inf_{\mathcal{A}} (\mathcal{L} + \Delta) + 2\epsilon_{\mathrm{stat}} + \epsilon_{\mathrm{opt}} =: B. \tag{113}$$

The infimum need not be attained. If a θ_* representing the true marginal laws is admissible, then

$$\mathcal{L}(\widehat{\theta}) \leq \inf_{\phi:(\theta_*,\phi)\in\mathcal{A}} \Delta(\theta_*, \phi) + 2\epsilon_{\mathrm{stat}} + \epsilon_{\mathrm{opt}}. \tag{114}$$

By nonnegativity of each KL term, Pinsker's inequality, and Jensen's inequality, if $B < \infty$ then

$$\mathbb{E}_U \, \mathrm{KL}(P_{*,k} \| P_{\widehat{\theta},k}) \leq B/w_k, \tag{115}$$

$$\mathbb{E}_U \, \mathrm{TV}(P_{*,k}, P_{\widehat{\theta},k}) \leq \sqrt{B/(2w_k)}. \tag{116}$$

Both laws are conditional on U, and $\mathrm{TV} = \frac{1}{2} \int |p - q|$. If sampling error and Monte Carlo error are each controlled uniformly, their sum can be absorbed into ϵ_{stat}. This is a deterministic comparison under the assumed errors; it does not establish finite-sample convergence rates for neural networks or global convergence of the optimizer. It applies to the objective with fixed weights after warm-up.

Observable fit versus factor error. Small KL divergence does not imply a small error in derivatives of the contrasts. On the circle, for $n \in \mathbb{N}$, let $p^*(t) = 1/(2\pi)$ and $p_n(t) \propto \exp(n^{-1} \cos nt)$. A Taylor expansion gives $\mathrm{KL}(P^* \| P_n) = O(n^{-2})$, while $\|(\log(p_n/p^*))''\|_\infty = n$. Thus, converting observable-law error into factor error requires additional assumptions on derivative control and stability of the inverse problem.

D.5 Implementation with Scalar or Vector Factors

Write $z = (z_1, \dots, z_d)$ with $z_j \in \mathbb{R}^m$, and output the distribution parameters of each branch from a common encoder feature. The prior factorizes across factors while allowing within-factor covariance:

$$p_\theta(z \mid u) = \prod_j \mathcal{N}(z_j; \mu_{\theta j}(u), \Sigma_{\theta j}(u)), \qquad \Sigma_{\theta j}(u) \succ 0. \tag{117}$$

This factorization of the prior does not imply factor independence in the posterior. The base branch takes (X, U) as input, the task branch takes (X, U, Y_τ), and Eq. (14) is evaluated with reparameterized samples in the same latent space. Choosing a Gaussian variational family does not generally guarantee exact posterior representation.

For the responses, we impose the factor-wise additivity in Eq. (13). The theory does not require decoder additivity, but it does require the embedding and closed-image conditions; these conditions are not guaranteed merely by using a general neural network. The decomposition in Eq. (106) does not apply unchanged to losses that reweight response terms or add separate regularizers.

At inference time, we use the base-posterior mean and validation-based factor selection from Section 4.3, and selection and editing operate on all m coordinates of a factor together. The scalar rotation posterior and the vector covariance and loss coefficients are specified once in Appendices E.1 and E.2, respectively.

E Experimental Evaluation

This appendix presents the scalar and vector settings from Section 5, together with the training and evaluation procedures and complete results. The two settings address the same recovery, selection, and editing questions, but differ in the observation-generating process and the unit of editing. In particular, scalar editing is normalized by the target change magnitude, whereas vector editing is normalized by the calibration standard deviation, so numerical comparisons are made within each setting.

E.1 Controlled Experiments with Scalar Factors

This setting examines under controlled conditions whether factor-wise selection and editing can differ even when relevant information is preserved. We learn the directions of three factors under a nonlinear observation transform of known form, and vary the sufficiency, complementarity, and redundancy of the information sources to compare the relation among recovery, selection, and editing. The six information conditions are summarized in Table 3, training and selection in Appendix E.1.4.

E.1.1 Data-generating Process and Information Conditions

Latent variables and observations. We use three latent factors $S = (S_1, S_2, S_3)$ and three observed environment variables $U \sim \mathrm{Unif}([-1, 1]^3)$. Conditioned on the environment, the latent factors are independent Gaussians.

$$S_j \mid U = u \sim \mathcal{N}\big(\mu_j(u), v_j(u)\big), \qquad j \in \{1, 2, 3\}. \tag{118}$$

In the default environment, with $a = u_1$,

$$\mu(u) = \big(0, 0, 0.9\sin(2\pi a)\big), \tag{119}$$

$$v(u) = \big(0.25 + 0.9\sigma(3a),\ 0.25 + 0.9\sigma(3a),\ 0.25 + 0.9\sigma(-3a)\big), \tag{120}$$

where $\sigma(t) = (1 + e^{-t})^{-1}$. Thus, the conditional distribution of (S_1, S_2) is spherical, and U_2, U_3 are observed environment variables that do not change this distribution. Only in the Independent modulation condition do we instead use

$$\mu_j(u) = 0.9\sin(\pi u_j), \qquad v_j(u) = 0.25 + 0.9\sigma(3u_j) \tag{121}$$

so that each environment variable independently modulates the corresponding factor.

Table 3: Information conditions in the scalar experiment. I_1, I_2 are the true relevance sets. Rank denotes the rank of the environmental coefficient matrix (A) for the common Gaussian sufficient statistics, the task-logit coefficient matrix (T), and their row-wise stack (J). All three factors are relevant, so the total number of derivative columns is 6.

Condition	Modified component and evaluation question	I_1/I_2	Rank (A,T,J)
Collider	Y_1 is a common effect of S_1, S_2. Effect of factor dependence induced by response conditioning.	$\{1,2\}/\{2,3\}$	$(3,5,6)$
Complementary	Default prior and tasks. Effect of complementary constraints provided by the contrasts from the two information sources.	$\{1\}/\{2,3\}$	$(3,5,6)$
Independent modulation	Each U_j independently modulates S_j. Condition in which the environment alone provides the required rank.	$\{1\}/\{2,3\}$	$(6,5,6)$
Task-rich	Modified task logits. Condition in which the tasks alone provide the required rank.	$\{1,3\}/\{2,3\}$	$(3,6,6)$
Factor-additive observation	$Q_c = I_3$ under the default information condition. Relation between an observation structure additive in the true factors and an additive decoder.	$\{1\}/\{2,3\}$	$(3,5,6)$
Redundant	Both information sources allow rotations of (S_1, S_2). Difference between preserving subset information and recovering individual factors.	$\{1,2\}/\{3\}$	$(3,3,3)$

The input coordinates $r = Q_c s$ to the observation transform and the observation are generated as

$$g(r) = \begin{bmatrix} r \\ 0.4(r \odot r - \mathbf{1}) \end{bmatrix}, \qquad X = g(Q_c S) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 0.03^2 I_6) \tag{122}$$

where $Q_c \in \mathrm{SO}(3)$ is the mixing matrix for condition c. The Factor-additive observation condition uses $Q_c = I_3$; all other conditions use the same matrix obtained by QR decomposition of a 3×3 standard-normal matrix from NumPy `default_rng(91831)`, with the sign of the first column flipped if the determinant is negative.

The learning algorithms are given the nonlinear observation transform g and latent dimension 3, but not the mixing matrix Q_c used to generate the data or the sample-wise values of S. Because the first three observation coordinates are r itself, this setting simplifies inversion of the observation transform relative to learning a general image generator.

Categorical tasks. We use two tasks $Y_1, Y_2 \in \{0,1,2,3\}$, with the reference-category logit fixed to zero for each task. Let $\varphi_j = (s_j^2 - 1)/2$, and denote the three non-reference logits by $\ell_\tau(s) \in \mathbb{R}^3$.

$$p_\tau(y \mid s) = \frac{\exp\{\widetilde{\ell}_{\tau,y}(s)\}}{\sum_{k=0}^{3} \exp\{\widetilde{\ell}_{\tau,k}(s)\}}, \qquad \widetilde{\ell}_\tau = (0, \ell_\tau^{\mathsf{T}})^{\mathsf{T}}. \tag{123}$$

Table 3 summarizes the component varied in each of the six conditions and the corresponding evaluation question. We specify the prior and logits for each condition below.

The default task logits are

$$\ell_1^{\mathrm{base}}(s) = \big(2s_1,\, 2\varphi_1,\, -2s_1\big), \tag{124}$$

$$\ell_2^{\mathrm{base}}(s) = \big(-2(\varphi_2+\varphi_3) + 0.3s_2,\, 2(\varphi_2+\varphi_3) + 0.3s_3,\, 4(\varphi_2+\varphi_3) + 0.3(s_2+s_3)\big). \tag{125}$$

The default tasks are used in Complementary, Independent modulation, and Factor-additive observation. In Collider, we keep the second task unchanged and replace the first by

$$\ell_1^{\mathrm{col}}(s) = \big(6(s_1+s_2),\, 4\varphi_1,\, -6(s_1+s_2)\big) \tag{126}$$

Because Y_1 becomes a common effect of S_1, S_2, the latent distribution conditioned on U, Y_1 is generally not factorial Gaussian. This condition examines the difference between modeling the response as a downstream likelihood and treating it as an additional conditioning variable of the upstream prior. In Task-rich, we use

$$\ell_1^{\mathrm{rich}} = (2s_1, 2\varphi_1, 2s_3), \qquad \ell_2^{\mathrm{rich}} = (2s_2, 2\varphi_2, 2\varphi_3) \tag{127}$$

and in Redundant we use

$$\ell_1^{\mathrm{red}} = \big(-2(\varphi_1+\varphi_2), 2(\varphi_1+\varphi_2), 4(\varphi_1+\varphi_2)\big), \qquad \ell_2^{\mathrm{red}} = (2s_3, 2\varphi_3, -2s_3) \tag{128}$$

The implementation samples the two categorical responses separately given S. Unlike this sampling choice, the marginal composite-likelihood theory in the main text does not require the actual tasks to have a conditionally factorized joint distribution.

For the rank calculation, we order the statistics as $(S_1, S_1^2/2, S_2, S_2^2/2, S_3, S_3^2/2)$ and use the environmental natural parameters $\big(\mu_j/v_j, -1/v_j\big)_{j=1}^{3}$. The contrast row space of the common environment is

$$\operatorname{span}\big\{(0,1,0,1,0,0),\ (0,0,0,0,1,0),\ (0,0,0,0,0,1)\big\} \tag{129}$$

whereas the independent environment varies all six statistic directions. Applying exact elimination to this row-space basis and the rational coefficients in Eqs. (125)–(128) yields the ranks in Table 3. That the row space of the common environment has dimension exactly three can be verified using finitely many environmental contrasts. Taking $a = 0$ as the reference environment gives $v_1 = v_3 = v_0 = 0.7$, and at $a = \pm 1/2$, $\mu_3 = 0$ and the two variances v_+, v_- are interchanged. Since $v_+ \neq v_-$ and $v_+ + v_- = 2v_0$, for $A = -1/v_+ + 1/v_0$ and $B = -1/v_- + 1/v_0$,

$$\det \begin{bmatrix} A & B \\ B & A \end{bmatrix} = (A-B)(A+B) \neq 0. \tag{130}$$

the first factor is nonzero because the two variances differ, while the second factor is negative by strict convexity of $t \mapsto 1/t$. Thus, we obtain two variance directions, and $\mu_3/v_3 \neq 0$ at $a = 1/4$ provides the third direction. For the independent environment, we vary one u_j at a time. The values $u_j = 0, 1$ have zero mean but different variances, whereas $u_j = 1/2$ has nonzero mean, so the two natural-parameter contrasts are independent. The boundary point 1 can be replaced by a sufficiently nearby interior point by continuity. Applying this argument to all three coordinates gives a six-dimensional environmental row space. In a separate numerical check of the implementation, we form contrasts by subtracting the first environment from 2,048 environments generated with fixed seed 4431, and stack the six coefficient rows of the non-reference category logits. The implementation check for Table 3 uses singular-value tolerance 10^{-8}. This numerical check verifies the information structure of the true generative model and does not certify an identification condition for the learned network from data. For quadratic sufficient statistics, this coefficient rank is directly related to the derivative rank in the main text. Writing the contrast for factor j as $\psi_j(s_j) = a_j s_j + b_j s_j^2/2 + c_j$ gives

$$[\psi_j'(s_j), \psi_j''(s_j)] = [a_j, b_j] \begin{bmatrix} 1 & 0 \\ s_j & 1 \end{bmatrix}. \tag{131}$$

Since the determinant of the right 2×2 matrix is one, the rank of the coefficient matrix obtained by stacking the three blocks equals the rank of the corresponding derivative matrix for every s. Because $I_* = [3]$ in all six conditions, conditions with joint rank 6 satisfy the derivative-rank condition for all relevant factors in the true generative model. The exact coefficient ranks and numerical ranks from fixed environmental samples agree in all six conditions. The Redundant condition retains the rotational symmetry of the spherical (S_1, S_2) block. Although the condition for all three factors does not hold in this case, the S_3 direction of the second task and the block information for (S_1, S_2) in the first task can still be recovered separately. Thus, high recovery of block information and low alignment of individual coordinates can occur simultaneously.

E.1.2 Scalar Estimator

Scalar estimator. Let z denote the learned coordinates and r the input coordinates of the known observation transform.

$$R_\theta = \exp(A - A^\top) R_{\text{init}} \in \mathrm{SO}(3), \qquad F_\theta(z) = g(R_\theta^\top z), \tag{132}$$

$$p_\theta(x \mid z) = \mathcal{N}(x; F_\theta(z), \sigma^2 I_6), \qquad p_\theta(z \mid u) = \prod_{j=1}^{3} \mathcal{N}(z_j; m_{\theta j}(u), v_{\theta j}(u)). \tag{133}$$

Here, $g(r) = [r; 0.4(r \odot r - \mathbf{1})]$, $\sigma = 0.03$, and $v_{\theta j} > 0$. Because g contains the linear coordinates and $\|g(r)\| \geq \|r\|$, every F_θ is a proper embedding. The learned object is the rotation direction of the known-form nonlinear observation transform; the true factors and actual mixing matrix are not

provided as inputs. The two tasks are defined by softmax distributions with reference-category logit zero and three non-reference logits

$$\eta_{\theta,\tau,c}(z) = b_{\theta,\tau,c} + \sum_{j=1}^{3} \left(h_{\theta,\tau,j,1,c} z_j + h_{\theta,\tau,j,2,c}(z_j^2 - 1)/2 \right), \quad c = 1, 2, 3 \tag{134}$$

This scalar implementation uses no head sparsity or gate penalty.

The base network $B_\phi : (x, u) \mapsto b_0$ and task correction $b_\tau = b_0 + D_{\phi,\tau}(x, u, \mathrm{onehot}(y_\tau))$ are tanh MLPs of sizes $9 \to 32 \to 6$ and $13 \to 32 \to 6$, respectively. Starting from the first three observation coordinates, we compute $\bar{r}(x)$ using five damped Newton updates, and use a learnable Gaussian around it:

$$\bar{v}_j = \sigma^2/(1 + 0.64\bar{r}_j^2), \quad \mu_{\phi,k,j} = \bar{r}_j + \sigma \tanh(b_{k,j}),$$

$$\ell_{\phi,k,j} = \tfrac{1}{2} \log \max(\bar{v}_j, 10^{-6}) + \mathrm{clip}(b_{k,j+3}, -2, 2), \tag{135}$$

$$q^r_{\phi,k} = \mathcal{N}(\mu_{\phi,k}, \mathrm{diag}(e^{2\ell_{\phi,k}})), \quad q^z_{\theta,\phi,k} = \mathcal{N}(R_\theta \mu_{\phi,k}, R_\theta \,\mathrm{diag}(e^{2\ell_{\phi,k}}) R_\theta^\top). \tag{136}$$

Setting all $D_{\phi,\tau} = 0$ gives the response-free shared posterior.

Objective and approximation. In each branch, we draw one $r \sim q^r$ and evaluate the prior and head at $z = R_\theta r$. Because the rotation preserves entropy, we use $\log q^z(z) = \log q^r(r)$. After removing the Gaussian observation constant $c_\sigma = 3 \log(2\pi\sigma^2)$, let the negative ELBO estimator be $\widehat{B}_k$; the implemented loss is

$$\widehat{\mathcal{L}}_{\mathrm{scalar}}(t) = \frac{\widehat{B}_0 + a_t \widehat{B}_1 + a_t \widehat{B}_2}{6(1 + 2a_t)}, \qquad a_t = \alpha \min\{1, t/\max(1, 0.1 N_{\mathrm{step}})\}. \tag{137}$$

After warm-up, this differs only by an additive constant and overall scaling from a composite ELBO with $w_0 = 1/(1 + 2\alpha)$ and $w_1 = w_2 = \alpha/(1 + 2\alpha)$. The prior KL coefficient is one. However, the log density of the exact base posterior contains $-0.08 \sum_j r_j^4/\sigma^2$, so it is not Gaussian; we therefore do not directly apply the exact-posterior condition of the population proposition to this finite Gaussian implementation. For active tasks and a full-support posterior, the response changes the posterior, so a shared posterior can retain the positive lower bound in Eq. (109). Approximation error may also remain for the task-specific Gaussians. Inference uses the base-posterior mean.

E.1.3 Baseline Information and Implementations

Variational and deterministic comparison methods use the common inverse of the observation transform and the rotated coordinates. Variational models use the Gaussian in Eq. (136), whereas deterministic models use rotations of the calibrated coordinates. Only Additive-AE learns a decoder of the form $Bz + C(z \odot z - \mathbf{1}) + b$. Table 4 summarizes information access, and the specific differences in the adaptations are described below.

Auxiliary conditioning and the composite estimator. iVAE(U) uses an environment-conditioned factorial Gaussian prior. For iVAE(U, Y), the prior context is an 11-dimensional vector obtained by concatenating the three-dimensional environment with two four-dimensional response one-hot vectors. Its encoder takes X, U as input and does not take the response one-hot vectors as input. TC-iVAE combines a conditional prior with additive heads and uses $q_\tau(s \mid x, u, y_\tau)$ in the task branches. Implementation details for the generative model and estimation objective are given in Appendix E.1.2.

CI-iVAE adaptation. Following the objective of Kim et al. (2023), we use a mixture of $q_e(z \mid x)$ and the normalized Gaussian product $q_p(z \mid x, u) \propto q_e(z \mid x) p(z \mid u)$. The latter is computed while preserving the full covariance of the common rotated posterior. For $q_a = a q_e + (1 - a) q_p$, the negative ELBO is

$$L_{\mathrm{mix}}(a) = a L_e + (1 - a) L_p - \left[a D_{\mathrm{KL}}(q_e \| q_a) + (1 - a) D_{\mathrm{KL}}(q_p \| q_a) \right]. \tag{138}$$

For each sample, we choose the branch with the smallest value among $a \in \{0, 0.05, \ldots, 1\}$. The Gaussian KL and expected reconstruction term for the quadratic observation transform are computed

Table 4: Information access and scalar implementations by method. "Training" denotes information used by the loss and inference during training, whereas "Inference" denotes information used to compute the final representation. Test responses are not provided to any method.

Method	Training	Inference	Implementation scope
VAE	X	X	VAE with standard Gaussian prior and fixed observation transform
iVAE(U)	X, U	X, U	Common observation-transform family applied to the public Gaussian-prior/ELBO core
iVAE(U, Y)	X, U, Y	X, U	Concatenates U with one-hot encodings of the two responses in the prior
CI-iVAE(U)	X, U	X	Adapter reimplementing the mixture ELBO of the original paper
CI-iVAE(U, Y)	X, U, Y	X	Extends the prior context to (U, Y) under the same objective
Mechanism-sparse	X, U	X, U	Public mask/prior core with environment–latent edge constraints
Additive-AE	X	X	Adaptation using a sum of univariate quadratic decoders
Task-additive	X, Y	X	Deterministic representation with an additive quadratic categorical head
Task-VAE	X, Y	X	Standard prior and X-conditioned posterior with an additive response term
Bilevel sparsity	X, Y	X	Categorical adaptation of the sparse bilevel principle
G2S-Jacobian	X, Y	X	Adaptation with full-quadratic Jacobian sparsification related to G2S
TC-iVAE	X, U, Y	X, U	Composite ELBO with a shared prior and task-specific posteriors

analytically, and the mixture-entropy term is estimated using 16 Monte Carlo samples per Gaussian. We use fixed observation noise and $\beta = 1$. This is an adaptation of the original objective with a fixed observation transform and should not be interpreted as reproducing all hyperparameters or posterior-collapse experiments of the public example trainer.

Mechanism-sparse. Using the public implementation of Lachapelle et al. (2026), we use the public hard-Gumbel mask and Gaussian-prior core for the nine candidate edges from the three environment variables to the three latent variables. The budget on the expected number of edges decreases from 9 to 3 over the first half of the 8,000 training steps. Primal variables are updated with Adam and nonnegative dual variables with a projected update at learning rate 0.003; the dual variable is reset to zero when the constraint is satisfied. A checkpoint is considered only when the expected number of edges is at most $3 + 0.15$. Because the first two factors under the common environment have the same environmental parent, graph sparsity alone cannot determine the true factor directions within that spherical block.

Bilevel sparsity adaptation. Following the sparse bilevel principle of Lachapelle et al. (2023a), we group the linear and quadratic features for each of the three latent coordinates and fit the categorical softmax heads for the two tasks as a sparse inner problem. The objective is the sum of cross-entropy, a group-wise Frobenius-norm penalty, and ridge $10^{-3}\|W\|_F^2/2$, with no penalty on the intercept. The group penalty is the selected ratio multiplied by $\lambda_{\max}$ computed from the current features. The inner problem is solved on the first half of each minibatch, and the rotation is updated using the response NLL on the second half. Backpropagation uses implicit differentiation through the active-group Hessian. The default maximum number of FISTA iterations is 100, with KKT tolerance 10^{-4}. Immediately before validation, the head is refit for up to 1,500 iterations using the first 4,096 training examples. This categorical adapter is distinct from the regression and group-SVM experiments in the original paper.

G2S-Jacobian. As a comparison method related to the task-specific subgroup perspective of Zheng et al. (2026), the first half of the total budget trains an X-only VAE, the task term is ramped up over the next 10%, after which task cross-entropy and a task-Jacobian group penalty are added to the same reconstruction objective. The head contains all three linear, three quadratic, and three cross features. To make the penalty invariant to a common shift of the softmax logits, we subtract the mean of the four logits to form centered logits and differentiate them with respect to z. For the resulting Jacobian, we take the Euclidean norm of the four-category gradient vector for each coordinate, sum over coordinates, and then average over tasks and samples. This row is reported as a static adaptation implementing a G2S-related principle, not as a reproduction of the official estimator or image benchmark.

Table 5: Fixed execution settings for the scalar controls.

Item	Value
Training seeds b	$43, 71, 101, 131, 173$
Training/validation/test samples	$32{,}768/4{,}096/4{,}096$
Generation seed by split	$100000 + b/200000 + b/500000 + b$
Training steps / batch size	$8{,}000/256$
Optimizer / learning rate	Adam / 0.003
Gradient norm limit	20
Network width	32
TC task weight / warm-up	$\alpha = 1$ / first 800 steps
Validation interval / batch	200 steps / 512
Common-probe fitting samples	first 4,096 examples of the training split
Evaluation calibration / final independent-factor samples	2,048 each

Additive-decoder adaptation. As a small comparison corresponding to the additive-decoder principle of Lachapelle et al. (2023b), we use a sum of univariate quadratic decoders. In the Factor-additive observation condition, $Q_c = I_3$, so in the true factor coordinates this decoder family can represent the true observation transform. The remaining conditions use a generic rotational mixing matrix and therefore do not satisfy the same assumption; this comparison isolates the role of additive observation structure.

E.1.4 Training, validation, and selected hyperparameters

Methods with the same seed share the training, validation, and test samples and the initial rotation. For each method, we select the checkpoint with the lowest validation objective among steps 4,000–8,000. The TC-iVAE objective follows Eq. (137).

For the two sparsity-related adaptations, penalty candidates $\{0, 0.03, 0.1\}$ are each trained for 1,200 steps per condition using a separate development seed 17. The training/validation/test sizes are $16{,}384/2{,}048/2{,}048$, and candidate selection uses only the validation response NLL produced by each method's task head. Among candidates within one standard error of the minimum NLL, we select the largest penalty coefficient and fix it for the five evaluation seeds. The selected values for Bilevel sparsity are 0 for Collider and 0.03 for Complementary/Task-rich/Redundant, whereas for G2S-Jacobian they are 0 for Collider/Complementary/Task-rich. Thus, no positive penalty is applied to either method in Collider or to G2S-Jacobian in the remaining conditions. Sensitivity to positive Jacobian penalties is reported in Appendix E.1.7.

For Collider and Complementary, we report results for all 12 methods in Table 4. The positive/negative-control conditions include methods corresponding to the relevant assumptions: Independent modulation compares 4 methods, Task-rich 5 methods, Factor-additive observation 3 methods, and Redundant 4 methods.

E.1.5 Independent-factor evaluation and coordinate alignment

After model training and all selections are complete, we generate 2,048 independent calibration examples and 2,048 final evaluation examples from a separate evaluation distribution Q_{eval}.

$$S \sim \mathcal{N}(0, 0.8^2 I_3), \qquad U = 0, \qquad X = g(Q_c S). \tag{139}$$

No observation noise is added to either split. The independent-factor distribution removes the effect of inter-factor correlations that may be present in the observational distribution, and the noise-free observations are used to examine the geometry of coordinate mixing. These metrics are therefore interpreted separately from noisy-test risk under Eq. (118). True factor values are used only in this post-hoc evaluation stage and are not fed back into selection of the network, penalty coefficient, checkpoint, or task-specific factor set.

We denote the calibration representations by z_n^{cal} and the final representations by z_n. On the calibration split, we determine a one-to-one correspondence by maximizing the absolute Pearson-

correlation matrix using the Hungarian algorithm.

$$\widehat{\pi} = \arg\max_{\pi \in \mathfrak{S}_3} \sum_{j=1}^{3} \left|\mathrm{Corr}_{\mathrm{cal}}(S_j, Z_{\pi(j)})\right|, \qquad \mathrm{MCC} = \frac{1}{3}\sum_{j=1}^{3} \left|\mathrm{Corr}_{\mathrm{eval}}(S_j, Z_{\widehat{\pi}(j)})\right|. \tag{140}$$

This correspondence handles sign and ordering but does not post-hoc rotate the full representation or apply multivariate affine alignment.

Full-factor readout. Full linear R^2 is evaluated on test data after fitting an affine map from all of Z to S on calibration data. This map is an evaluation readout of information preservation and does not alter the representation used for MCC or editing. Thus, a rotated representation can have high Full R^2 while having low MCC or editing selectivity for individual factors.

Task-selected groups and nuisance predictability. For each task, we compare the seven nonempty subsets of the three factors. For each candidate J, we fit a quadratic logistic probe with linear terms, squared terms, cross terms, and an intercept. Training features are divided by $\max\{\mathrm{SD}, 0.1\}$, and a ridge penalty $10^{-3}\|W\|_F^2/2$ is added to all coefficients except the intercept. The same probe is used for every method, and the selected set is not extracted from the native TC-iVAE head. Let $\overline{\ell}_{\tau,J}$ denote the mean validation NLL and $s_{\tau,J}$ the standard error, computed as the sample SD of per-example NLL divided by $\sqrt{n_{\mathrm{val}}}$. For the candidate J_τ^{best} with minimum validation NLL, define

$$J_\tau = \underset{J:\, \overline{\ell}_{\tau,J} \le \overline{\ell}_{\tau,J_\tau^{\mathrm{best}}} + s_{\tau,J_\tau^{\mathrm{best}}}}{\arg\min^{\mathrm{lex}}} \quad (|J|, \overline{\ell}_{\tau,J}, J) \tag{141}$$

Thus, among admissible candidates, selection is lexicographic by size, NLL, and index order. True factors are not used for selection; after selection, predictability of relevant and irrelevant factors from the fixed Z_{J_τ} is evaluated on the independent test set.

For $K = I_\tau$ or $K = I_\tau^c$, on the calibration split we fit

$$(\widehat{B}_{\tau,K}, \widehat{b}_{\tau,K}) = \arg\min_{B,b} \sum_{n=1}^{N_{\mathrm{cal}}} \left\| s_{n,K}^{\mathrm{cal}} - B z_{n,J_\tau}^{\mathrm{cal}} - b \right\|_2^2 \tag{142}$$

by least squares. On the final evaluation set, we compute

$$R^2_{\tau,K} = 1 - \frac{\sum_{n=1}^{N} \left\| s_{n,K} - \widehat{B}_{\tau,K} z_{n,J_\tau} - \widehat{b}_{\tau,K} \right\|_2^2}{\max\left\{ \sum_{n=1}^{N} \|s_{n,K} - \overline{s}_K\|_2^2, 10^{-12} \right\}} \tag{143}$$

Here, $K = I_\tau$ gives the relevant-factor R^2 and $K = I_\tau^c$ the irrelevant-factor R^2. Because the numerator and denominator are aggregated over both samples and target factors, this quantity generally differs from the simple average of coordinate-wise R^2. The final group metric is averaged over the two tasks. Finite-sample R^2 may be negative and is not clipped at zero.

The selection probe is quadratic, whereas the evaluation readout is affine linear. A low irrelevant-factor R^2 is not interpreted as absence of all nonlinear dependence; factor-wise invertibility and the relation between the evaluation distribution and metrics follow Appendix C.6.

We report the selected dimension as $\frac{1}{2}\sum_\tau |J_\tau|$. To define an evaluation index set for incidence F1, we use the Hungarian correspondence to form $\widetilde{I}_\tau = \{j : \widehat{\pi}(j) \in J_\tau\}$ and then compute

$$F_{1,\tau} = \frac{2|I_\tau \cap \widetilde{I}_\tau|}{|I_\tau| + |J_\tau|} \tag{144}$$

and average over the two tasks. The true relevance set is used only to compute performance and does not reselect J_τ.

Coordinate editing and leakage. Single-coordinate editing is performed separately for each of the three true factors and independently of the subset-selection probe. For learned coordinate $k =$

$\widehat{\pi}(j)$ matched to factor j, we fit the univariate affine least-squares model $S_j = a_j Z_k + b_j$ on the calibration split and adjust the slope as

$$\widetilde{a}_j = \mathrm{sgn}_+(a_j)\max\{|a_j|, 0.05\}, \qquad \mathrm{sgn}_+(0) = 1 \tag{145}$$

The target change is $\delta = 0.8$, and

$$z_n^{(j)} = z_n + \frac{\delta}{\widetilde{a}_j} e_{\widehat{\pi}(j)} \tag{146}$$

changes only that learned coordinate. This sign and univariate-scale calibration is used to compare the same editing magnitude in units of the true factor and does not allow a rotation that moves other learned coordinates simultaneously.

Applying the known evaluation inverse mixing to the first three observation coordinates of the decoder output, define

$$\widetilde{s}_n^{(0)} = Q_c^{\mathsf{T}}[D_\theta(z_n)]_{1:3}, \qquad \widetilde{s}_n^{(j)} = Q_c^{\mathsf{T}}[D_\theta(z_n^{(j)})]_{1:3}, \qquad d_n^{(j)} = \widetilde{s}_n^{(j)} - \widetilde{s}_n^{(0)} \tag{147}$$

This is a post-hoc evaluation in units of the true factors provided by the simulator. Per-factor edit leakage and target-change error are

$$L_j^{\mathrm{edit}} = \frac{\sqrt{N^{-1}\sum_n \|d_{n,-j}^{(j)}\|_2^2}}{\delta}, \qquad L^{\mathrm{edit}} = \frac{1}{3}\sum_{j=1}^{3} L_j^{\mathrm{edit}}, \tag{148}$$

$$E_j^{\mathrm{target}} = \frac{\sqrt{N^{-1}\sum_n (d_{n,j}^{(j)} - \delta)^2}}{\delta}, \qquad E^{\mathrm{target}} = \frac{1}{3}\sum_{j=1}^{3} E_j^{\mathrm{target}}. \tag{149}$$

Leakage sums the squared changes of the other two factors before taking the square root of the sample mean. We do not divide again by the number of the two non-target factors, and rather than computing a single RMS across all three factors, we average the per-factor values at the end. Low leakage measures preservation of non-target factors, so whether the target change was actually achieved is checked jointly with E^{target}.

Observational prediction and reporting conventions. On the noisy test split, we evaluate reconstruction MSE against the observed input and reconstruction MSE against the noise-free observation separately. The two errors are the sample averages of $\|D_\theta(e_{\theta,\phi}(x,u)) - x\|_2^2/6$ and $\|D_\theta(e_{\theta,\phi}(x,u)) - g(Q_c s)\|_2^2/6$, respectively. For methods with a native task head, we average the KL between the true simulator categorical distribution $p_\tau(\cdot \mid s)$ and the categorical distribution of the learned head over test examples and the two tasks. Native task KL is not defined for VAE, auxiliary-only, mechanism-sparsity, or Additive-AE methods without a task head. The validation NLL of the common probe is a separate validation metric and is not relabeled as a test response-prediction score.

Results are reported as the mean and sample SD over the five training/data seeds for each condition–method pair.

E.1.6 Complete Results across the Six Conditions

For each condition, we jointly report factor recovery, selection, and editing results together with overall information preservation and observation reconstruction.

In Collider, the full linear R^2 is nearly one for several comparison methods. Nevertheless, the irrelevant-factor R^2 of the selected subrepresentation and editing leakage differ substantially. For TC-iVAE, high relevant-factor R^2, irrelevant-factor R^2 near zero, F1$= 1$, and low editing leakage show that preservation of relevant information and selective use can hold simultaneously. In Complementary, iVAE(U, Y) and CI-iVAE(U, Y) also achieve high coordinate recovery. In Independent modulation, Task-rich, and Factor-additive observation, respectively, comparison methods whose environmental information, task information, or observation structure is sufficient also succeed.

In Redundant, information in the selected task block is preserved, but rotations remain within the block, increasing the variability of MCC and individual-factor editing. Thus, low task-irrelevant R^2 and low individual editing leakage measure different aspects and need not decrease together in every condition.

Table 6: Factor recovery and selectivity in Collider and Complementary. Mean ± SD over five seeds. Representations selected by the same probe are evaluated.

Method	MCC ↑	Rel. R^2 ↑	Irrel. $R^2 \approx 0$	$\lvert J_\tau \rvert$	F1 ↑	Edit leak ↓
Collider						
VAE	0.8565±0.0607	1.0000±0.0000	1.0000±0.0000	3.0000±0.0000	0.8000±0.0000	0.5914±0.1619
iVAE(U)	0.9451±0.0810	0.9987±0.0026	0.3011±0.2722	2.3000±0.2700	0.9400±0.0500	0.2462±0.2553
iVAE(U, Y)	0.8476±0.0031	1.0000±0.0000	0.4999±0.0007	2.5000±0.0000	0.9000±0.0000	0.5538±0.0082
CI-iVAE(U)	0.9473±0.0816	0.9993±0.0009	0.2003±0.2736	2.2000±0.2700	0.9600±0.0500	0.2233±0.2700
CI-iVAE(U, Y)	0.8463±0.0024	1.0000±0.0000	0.4999±0.0007	2.5000±0.0000	0.9000±0.0000	0.5574±0.0057
Mechanism-sparse	0.9412±0.0731	0.9998±0.0001	0.3994±0.2248	2.4000±0.2200	0.9200±0.0400	0.2615±0.2427
Additive-AE	0.7592±0.0023	1.0000±0.0000	1.0000±0.0000	3.0000±0.0000	0.8000±0.0000	0.8642±0.0057
Task-additive	0.9123±0.0413	0.9999±0.0001	0.4997±0.0008	2.5000±0.0000	0.9000±0.0000	0.3827±0.1106
Task-VAE	0.8473±0.0183	0.9999±0.0001	0.4996±0.0008	2.5000±0.0000	0.9000±0.0000	0.5557±0.0480
Bilevel sparsity	0.9967±0.0026	0.9997±0.0005	0.4000±0.2232	2.4000±0.2200	0.9200±0.0400	0.0686±0.0222
G2S-Jacobian	0.8199±0.0218	1.0000±0.0000	1.0000±0.0000	3.0000±0.0000	0.8000±0.0000	0.6760±0.0560
TC-iVAE	0.9999±0.0001	0.9999±0.0000	−0.0006±0.0015	2.0000±0.0000	1.0000±0.0000	0.0135±0.0031
Complementary						
VAE	0.8565±0.0607	0.9995±0.0011	0.9004±0.1364	2.8000±0.2700	0.6800±0.0500	0.5914±0.1619
iVAE(U)	0.9451±0.0810	0.9941±0.0079	0.4051±0.3777	2.0000±0.5000	0.8700±0.1300	0.2462±0.2553
iVAE(U, Y)	1.0000±0.0000	0.9999±0.0000	−0.0008±0.0012	1.5000±0.0000	1.0000±0.0000	0.0077±0.0020
CI-iVAE(U)	0.9473±0.0816	0.9982±0.0026	0.3026±0.4120	1.9000±0.5500	0.8900±0.1500	0.2233±0.2700
CI-iVAE(U, Y)	1.0000±0.0000	1.0000±0.0000	−0.0008±0.0012	1.5000±0.0000	1.0000±0.0000	0.0088±0.0032
Mechanism-sparse	0.9412±0.0731	0.9998±0.0002	0.6013±0.3376	2.3000±0.4500	0.7900±0.1200	0.2615±0.2427
Additive-AE	0.7592±0.0023	1.0000±0.0000	1.0000±0.0000	3.0000±0.0000	0.6500±0.0000	0.8642±0.0057
Task-additive	0.9828±0.0342	0.9999±0.0001	−0.0008±0.0011	1.5000±0.0000	1.0000±0.0000	0.1014±0.1484
Task-VAE	0.9635±0.0426	0.9999±0.0001	−0.0008±0.0012	1.5000±0.0000	1.0000±0.0000	0.1953±0.1650
Bilevel sparsity	0.8264±0.0315	0.9999±0.0001	−0.0007±0.0011	1.5000±0.0000	1.0000±0.0000	0.6103±0.0840
G2S-Jacobian	0.7165±0.0166	1.0000±0.0000	1.0000±0.0000	3.0000±0.0000	0.6500±0.0000	0.9878±0.0390
TC-iVAE	0.9999±0.0001	0.9999±0.0001	−0.0006±0.0013	1.5000±0.0000	1.0000±0.0000	0.0118±0.0040

Table 7: Overall information preservation and target editing in Collider and Complementary. Mean ± SD over five seeds. The two MSE columns are in units of 10^{-4}; lower MSE, KL, and Target error are better. — indicates methods without a native head.

Method	Full R^2 ↑	Clean MSE	Observed MSE	Native KL	Target error
Collider					
VAE	1.0000±0.0000	4.5110±0.0150	4.4920±0.0380	—	0.0128±0.0120
iVAE(U)	1.0000±0.0000	4.5120±0.0140	4.4930±0.0380	—	0.0064±0.0108
iVAE(U, Y)	1.0000±0.0000	4.5110±0.0130	4.4930±0.0390	—	0.0126±0.0111
CI-iVAE(U)	1.0000±0.0000	4.5100±0.0140	4.4890±0.0390	—	0.0062±0.0107
CI-iVAE(U, Y)	1.0000±0.0000	4.5100±0.0150	4.4890±0.0390	—	0.0127±0.0112
Mechanism-sparse	1.0000±0.0000	4.5120±0.0150	4.4930±0.0380	—	0.0065±0.0091
Additive-AE	1.0000±0.0000	4.5390±0.0220	4.5120±0.0470	—	0.0213±0.0133
Task-additive	1.0000±0.0000	4.5130±0.0150	4.4890±0.0390	0.0034±0.0005	0.0051±0.0027
Task-VAE	1.0000±0.0000	4.5120±0.0150	4.4930±0.0380	0.0032±0.0003	0.0080±0.0025
Bilevel sparsity	1.0000±0.0000	4.5130±0.0150	4.4890±0.0390	0.0123±0.0006	0.0009±0.0004
G2S-Jacobian	1.0000±0.0000	4.5110±0.0140	4.4920±0.0380	0.0038±0.0004	0.0168±0.0111
TC-iVAE	1.0000±0.0000	4.5090±0.0130	4.4940±0.0390	0.0033±0.0003	0.0008±0.0001
Complementary					
VAE	1.0000±0.0000	4.5110±0.0150	4.4920±0.0380	—	0.0128±0.0120
iVAE(U)	1.0000±0.0000	4.5120±0.0140	4.4930±0.0380	—	0.0064±0.0108
iVAE(U, Y)	1.0000±0.0000	4.5120±0.0130	4.4940±0.0390	—	0.0006±0.0002
CI-iVAE(U)	1.0000±0.0000	4.5100±0.0140	4.4890±0.0390	—	0.0062±0.0107
CI-iVAE(U, Y)	1.0000±0.0000	4.5100±0.0140	4.4890±0.0390	—	0.0003±0.0000
Mechanism-sparse	1.0000±0.0000	4.5120±0.0150	4.4930±0.0380	—	0.0065±0.0091
Additive-AE	1.0000±0.0000	4.5390±0.0220	4.5120±0.0470	—	0.0213±0.0133
Task-additive	1.0000±0.0000	4.5130±0.0150	4.4890±0.0390	0.0015±0.0002	0.0010±0.0014
Task-VAE	1.0000±0.0000	4.5120±0.0150	4.4920±0.0370	0.0015±0.0001	0.0025±0.0020
Bilevel sparsity	1.0000±0.0000	4.5130±0.0150	4.4890±0.0390	0.0094±0.0010	0.0086±0.0034
G2S-Jacobian	1.0000±0.0000	4.5110±0.0140	4.4930±0.0380	0.0016±0.0001	0.0227±0.0162
TC-iVAE	1.0000±0.0000	4.5090±0.0130	4.4930±0.0390	0.0015±0.0001	0.0007±0.0000

Table 8: Factor recovery and selectivity under single-source, observation-structure, and redundant conditions. Mean ± SD over five seeds. Representations selected by the same probe are evaluated.

Method	MCC ↑	Rel. R^2 ↑	Irrel. $R^2 \approx 0$	$\lvert J_\tau \rvert$	F1 ↑	Edit leak ↓
Independent modulation						
iVAE(U)	0.9999 ± 0.0002	0.9997 ± 0.0004	-0.0005 ± 0.0014	1.5000 ± 0.0000	1.0000 ± 0.0000	0.0144 ± 0.0083
CI-iVAE(U)	0.9999 ± 0.0001	0.9999 ± 0.0001	-0.0006 ± 0.0012	1.5000 ± 0.0000	1.0000 ± 0.0000	0.0124 ± 0.0033
Mechanism-sparse	0.9999 ± 0.0001	0.9998 ± 0.0003	-0.0006 ± 0.0012	1.5000 ± 0.0000	1.0000 ± 0.0000	0.0118 ± 0.0067
TC-iVAE	0.9999 ± 0.0000	0.9999 ± 0.0000	-0.0007 ± 0.0012	1.5000 ± 0.0000	1.0000 ± 0.0000	0.0102 ± 0.0028
Task-rich						
Task-additive	0.9997 ± 0.0001	0.9997 ± 0.0002	-0.0004 ± 0.0016	2.0000 ± 0.0000	1.0000 ± 0.0000	0.0217 ± 0.0056
Task-VAE	0.9998 ± 0.0001	0.9998 ± 0.0001	-0.0007 ± 0.0014	2.0000 ± 0.0000	1.0000 ± 0.0000	0.0186 ± 0.0064
Bilevel sparsity	0.9996 ± 0.0004	0.9996 ± 0.0004	-0.0001 ± 0.0015	2.0000 ± 0.0000	1.0000 ± 0.0000	0.0241 ± 0.0159
G2S-Jacobian	0.6779 ± 0.0265	1.0000 ± 0.0000	1.0000 ± 0.0000	3.0000 ± 0.0000	0.8000 ± 0.0000	1.0976 ± 0.0833
TC-iVAE	0.9999 ± 0.0001	0.9999 ± 0.0001	-0.0004 ± 0.0014	2.0000 ± 0.0000	1.0000 ± 0.0000	0.0148 ± 0.0074
Factor-additive observation						
VAE	0.8509 ± 0.0408	0.9989 ± 0.0013	0.8020 ± 0.1107	2.6000 ± 0.2200	0.7200 ± 0.0400	0.6056 ± 0.1051
Additive-AE	1.0000 ± 0.0000	1.0000 ± 0.0000	-0.0008 ± 0.0012	1.5000 ± 0.0000	1.0000 ± 0.0000	0.0013 ± 0.0004
TC-iVAE	0.9998 ± 0.0002	0.9997 ± 0.0004	-0.0005 ± 0.0013	1.5000 ± 0.0000	1.0000 ± 0.0000	0.0161 ± 0.0073
Redundant						
iVAE(U)	0.9451 ± 0.0810	0.9998 ± 0.0002	-0.0006 ± 0.0009	1.5000 ± 0.0000	1.0000 ± 0.0000	0.2462 ± 0.2553
Task-additive	0.9778 ± 0.0196	0.9999 ± 0.0001	-0.0006 ± 0.0007	1.5000 ± 0.0000	1.0000 ± 0.0000	0.1636 ± 0.0919
Bilevel sparsity	0.8989 ± 0.0635	0.9998 ± 0.0001	-0.0005 ± 0.0008	1.5000 ± 0.0000	1.0000 ± 0.0000	0.4141 ± 0.1760
TC-iVAE	0.9532 ± 0.0636	0.9998 ± 0.0001	-0.0006 ± 0.0009	1.5000 ± 0.0000	1.0000 ± 0.0000	0.2255 ± 0.2134

Table 9: Overall information preservation and target editing under single-source, observation-structure, and redundant conditions. Mean ± SD over five seeds. The two MSE columns are in units of 10^{-4}; lower MSE, KL, and Target error are better. — indicates methods without a native head.

Method	Full R^2 ↑	Clean MSE	Observed MSE	Native KL	Target error
Independent modulation					
iVAE(U)	1.0000 ± 0.0000	4.5000 ± 0.0330	4.5040 ± 0.0550	—	0.0005 ± 0.0001
CI-iVAE(U)	1.0000 ± 0.0000	4.5000 ± 0.0330	4.5010 ± 0.0560	—	0.0003 ± 0.0001
Mechanism-sparse	1.0000 ± 0.0000	4.5010 ± 0.0330	4.5040 ± 0.0550	—	0.0005 ± 0.0001
TC-iVAE	1.0000 ± 0.0000	4.5000 ± 0.0320	4.5030 ± 0.0560	0.0014 ± 0.0002	0.0006 ± 0.0001
Task-rich					
Task-additive	1.0000 ± 0.0000	4.5130 ± 0.0150	4.4890 ± 0.0390	0.0012 ± 0.0003	0.0004 ± 0.0003
Task-VAE	1.0000 ± 0.0000	4.5110 ± 0.0150	4.4920 ± 0.0380	0.0012 ± 0.0002	0.0005 ± 0.0002
Bilevel sparsity	1.0000 ± 0.0000	4.5130 ± 0.0150	4.4890 ± 0.0390	0.0056 ± 0.0015	0.0004 ± 0.0004
G2S-Jacobian	1.0000 ± 0.0000	4.5130 ± 0.0150	4.4930 ± 0.0380	0.0014 ± 0.0001	0.0196 ± 0.0094
TC-iVAE	1.0000 ± 0.0000	4.5110 ± 0.0130	4.4940 ± 0.0380	0.0011 ± 0.0003	0.0007 ± 0.0001
Factor-additive observation					
VAE	1.0000 ± 0.0000	4.5020 ± 0.0610	4.5020 ± 0.0740	—	0.0067 ± 0.0049
Additive-AE	1.0000 ± 0.0000	4.5320 ± 0.0610	4.5260 ± 0.0700	—	0.0009 ± 0.0004
TC-iVAE	1.0000 ± 0.0000	4.5010 ± 0.0610	4.5050 ± 0.0730	0.0017 ± 0.0003	0.0008 ± 0.0001
Redundant					
iVAE(U)	1.0000 ± 0.0000	4.5120 ± 0.0140	4.4930 ± 0.0380	—	0.0064 ± 0.0108
Task-additive	1.0000 ± 0.0000	4.5130 ± 0.0150	4.4890 ± 0.0390	0.0013 ± 0.0001	0.0028 ± 0.0024
Bilevel sparsity	1.0000 ± 0.0000	4.5130 ± 0.0150	4.4890 ± 0.0390	0.0097 ± 0.0010	0.0066 ± 0.0088
TC-iVAE	1.0000 ± 0.0000	4.5100 ± 0.0120	4.4940 ± 0.0380	0.0012 ± 0.0001	0.0057 ± 0.0091

E.1.7 Post-Hoc Regularization Sensitivity

Selection by the validation response NLL of the G2S-Jacobian task head yielded $\lambda = 0$ in all three conditions. Table 10 reports sensitivity to $\lambda = 0.03, 0.1$ using the same five seeds. The selection rule of the main comparison is retained.

In Complementary and Task-rich, positive penalties substantially improve coordinate alignment and editing selectivity. The positive-penalty results in Collider show substantial variation across seeds. Therefore, failure of the default setting selected by validation NLL alone does not imply that the corresponding structural principle cannot separate coordinates. This sensitivity analysis is not a procedure for replacing the original setting post hoc; it is a diagnostic showing that observed task fit and coordinate-identification metrics provide different selection criteria.

E.1.8 Factor Edits and Visualization

Table 11 breaks down Collider editing leakage by true factor. The largest differences occur in the entanglement of the position factors, whereas the color factor already shows small non-target changes for several methods. We report the target-change error together with the overall average across the three factors to verify that an edit is actually performed.

Seed-wise editing trajectories. Together with the vertical-position edit in Figure 4, Figures 5 and 6 show the horizontal-position and color edits, respectively. All three trajectory figures use the same eight methods as the main comparison and show all five seeds (43, 71, 101, 131, 173) for each method. Within each seed, all methods edit the same reference scene. Coordinate correspondence and scalar calibration use the values fixed by the existing evaluation-calibration procedure.

At 25 equally spaced target changes $\delta \in [-1.2, 1.2]$, let $\widetilde{\mathbf{s}}_f(\delta)$ denote the reconstructed true factors when factor f is edited; the trajectory

$$\mathbf{d}_f(\delta) = \widetilde{\mathbf{s}}_f(\delta) - \widetilde{\mathbf{s}}_f(0)$$

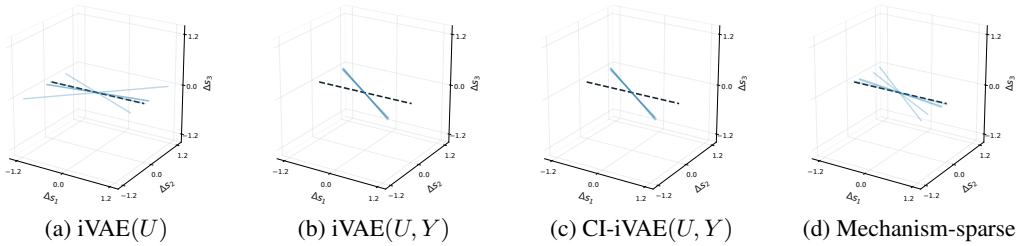
is defined as follows. Each run is centered at its pre-edit reconstruction, and the trajectory is not additionally rotated or normalized in length. The ideal ground-truth trajectory is $\delta \mathbf{e}_f$. The black dashed line denotes the ground-truth trajectory, and the blue lines are the five seed trajectories overlaid with equal width and transparency ($\alpha = 0.30$). The three axes are changes in horizontal position, vertical position, and color, and all panels use the same axis limits and viewpoints. By comparing trajectory overlap and agreement with the ground truth, we assess consistency across seeds and editing interference on non-target factors.

Table 10: Post-hoc fixed-grid sensitivity for G2S-Jacobian. Values are mean ± SD over five seeds for each condition and coefficient. We do not replace the $\lambda = 0$ setting selected in the main comparison post hoc; the table evaluates how positive penalties affect selectivity.

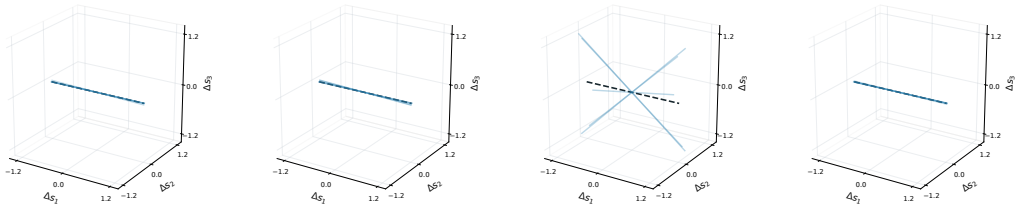
Case	λ	MCC ↑	R^2_{rel} ↑	$R^2_{\mathrm{irr}} \approx 0$	Edit leak ↓
Collider	0.03	0.9262 ± 0.1010	1.0000 ± 0.0001	0.1995 ± 0.2742	0.2624 ± 0.3439
Collider	0.10	0.8596 ± 0.0782	0.9998 ± 0.0002	0.3999 ± 0.2234	0.5003 ± 0.2626
Complementary	0.03	0.9997 ± 0.0003	0.9999 ± 0.0001	-0.0008 ± 0.0012	0.0199 ± 0.0103
Complementary	0.10	0.9996 ± 0.0003	0.9998 ± 0.0002	-0.0006 ± 0.0012	0.0266 ± 0.0082
Task-rich	0.03	0.9998 ± 0.0002	0.9999 ± 0.0001	-0.0009 ± 0.0015	0.0164 ± 0.0092
Task-rich	0.10	0.9991 ± 0.0007	0.9993 ± 0.0005	0.0003 ± 0.0015	0.0383 ± 0.0147

Table 11: Non-target RMS change by edit target and overall target-editing error in Collider. All values are normalized by target magnitude 0.8; mean ± sample SD over five seeds is reported.

Method	S_1 edit leak	S_2 edit leak	S_3 edit leak	Target error
iVAE(U, Y)	0.8151 ± 0.0185	0.8341 ± 0.0263	0.0123 ± 0.0062	0.0126 ± 0.0111
CI-iVAE(U, Y)	0.8204 ± 0.0158	0.8397 ± 0.0267	0.0119 ± 0.0049	0.0127 ± 0.0112
Bilevel sparsity	0.0185 ± 0.0131	0.0938 ± 0.0355	0.0934 ± 0.0357	0.0009 ± 0.0004
TC-iVAE	0.0150 ± 0.0051	0.0138 ± 0.0071	0.0115 ± 0.0055	0.0008 ± 0.0001

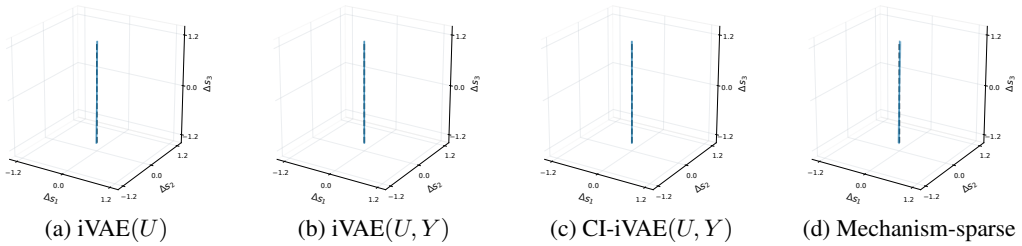


(a) iVAE(U) (b) iVAE(U, Y) (c) CI-iVAE(U, Y) (d) Mechanism-sparse

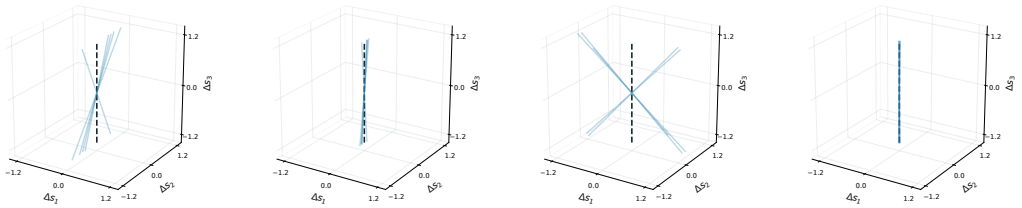


(e) Task-additive (f) Bilevel sparsity (g) G2S-Jacobian (h) TC-iVAE

Figure 5: **Factor-change trajectories under horizontal-position editing in Collider.** The black dashed line denotes the ground-truth trajectory, and the translucent blue lines show the five seeds. $(\Delta s_1, \Delta s_2, \Delta s_3)$ are the changes in horizontal position, vertical position, and color relative to the pre-edit reconstruction.



(a) iVAE(U) (b) iVAE(U, Y) (c) CI-iVAE(U, Y) (d) Mechanism-sparse



(e) Task-additive (f) Bilevel sparsity (g) G2S-Jacobian (h) TC-iVAE

Figure 6: **Factor-change trajectories under color editing in Collider.** The black dashed line denotes the ground-truth trajectory, and the translucent blue lines show the five seeds. $(\Delta s_1, \Delta s_2, \Delta s_3)$ are the changes in horizontal position, vertical position, and color relative to the pre-edit reconstruction.

E.2 Evaluation with Vector-Valued Factors

We apply the same recovery, selection, and editing evaluations as in the scalar setting to vector-valued factors with a learned encoder and decoder. The environment and task responses are constructed to provide complementary information, allowing us to examine whether information preservation and individual correspondence of the relevant factors improve together. All results use five training seeds with one fixed data-generating model and dataset. The main text reports the final TC-iVAE and seven comparison methods, while results for the composite-ELBO implementation without the additional regularization terms are reported below.

E.2.1 Data-generating process

Each factor is $S_j = (a_j, b_j) \in \mathbb{R}^2$, and $U \sim \mathrm{Unif}([-1,1]^3)$. Let σ denote the sigmoid function and $R(\alpha)$ a two-dimensional rotation. We generate

$$v(u) = 0.25 + 0.9\sigma(3u_1), \tag{150}$$

$$A(u) = R(0.65u_3)\,\mathrm{diag}(1.35, 0.65)R(0.65u_3)^\top, \tag{151}$$

$$\Sigma_1(u) = \Sigma_2(u) = v(u)A(u), \tag{152}$$

$$\Sigma_3(u) = (0.35 + 0.8\sigma(3u_2))I_2, \qquad p(s \mid u) = \prod_{j=1}^{3} \mathcal{N}(s_j; 0, \Sigma_j(u)) \tag{153}$$

The four tasks have four-category responses, with the reference-category logit fixed to zero. The non-reference logit vectors of the first three tasks are

$$\eta_1(s) = 3(a_1, b_1, (a_1^2 - b_1^2)/2), \tag{154}$$

$$\eta_2(s) = 3(a_2, b_2, (a_2^2 - b_2^2)/2), \tag{155}$$

$$\eta_3(s) = 3(a_1 b_1, a_2 b_2, (a_1^2 + b_1^2 - a_2^2 - b_2^2)/2) \tag{156}$$

We stack the preceding three logit vectors as $e(s) = [\eta_1(s)^\top, \eta_2(s)^\top, \eta_3(s)^\top]^\top \in \mathbb{R}^9$ and define the fourth task by $\eta_4(s) = L_4 e(s)$ using a fixed linear map $L_4 \in \mathbb{R}^{3\times 9}$ in the data-generating model. This linear combination is fixed to depend nontrivially on both relevant factors. Thus, $I_1 = \{1\}$, $I_2 = \{2\}$, $I_3 = I_4 = \{1, 2\}$, and $N_* = \{3\}$.

After applying a fixed rotation to each true factor, we map it through

$$\ell(a, b) = (a, b, 0.30 \sin a, 0.30 \sin b, 0.20(a^2 - 1), 0.20(b^2 - 1), 0.15ab, 0.15 \sin(ab)).$$

We concatenate the three vectors and apply a fixed 24-dimensional orthogonal transformation to obtain the clean observation. Independent Gaussian observation noise with standard deviation 0.05 is added for training and validation. Because the map contains linear components, the simulator can be inverted; this inverse is used only for post-training factor-editing evaluation.

Each two-dimensional factor contributes two first-order and three symmetric second-order derivative columns. At eight simulator evaluation points, the projected ranks for the environmental, task, and combined contrasts are $3, 9, 10$, respectively. The corresponding total ranks are $4, 9, 11$, and the irrelevant-column ranks are $1, 0, 1$. The two relevant factors require $2 \times 5 = 10$ columns. This is a finite-point numerical diagnostic of complementarity between the two information sources; the dense-set condition of Theorem 1 remains a separate requirement.

E.2.2 Models and objectives

The six-dimensional latent space is partitioned into three two-dimensional factors. All comparison methods use the same factor-additive neural decoder $F_\theta(z) = \sum_{j=1}^{3} f_{\theta j}(z_j)$. Each $f_{\theta j}$ is a $2 \to 128 \to 128 \to 128 \to 24$ SiLU MLP, giving 109,512 trainable decoder parameters in total. The encoder and heads also use width-128, depth-3 SiLU MLPs by default. This architecture separates the factors in the output of an encoder that processes the full input. The true factor values, task supports, and inverse of the true observation-generating function are not provided during training. Observation centering and scaling are computed from the training split only.

The TC-iVAE prior factorizes across factors and uses a full-covariance Gaussian within each factor. The base branch and four task branches each output a Gaussian over the same six-dimensional latent

Table 12: Training information and implementation for the vector-factor experiment. All methods use the same decoder family and response-free representation at evaluation.

Method	Inputs and training objective
iVAE(U)	$q_0(X, U)$, $p(Z \mid U)$; observation ELBO.
iVAE(U, Y)	$q_0(X, U)$, $p(Z \mid U, Y_{1:T})$; responses are used as auxiliary prior inputs.
CI-iVAE(U, Y)	Mixture objective using the Gaussian product of $q_0(X)$ and $p(Z \mid U, Y_{1:T})$.
Mechanism-sparse	$q_0(X, U)$, coordinate-wise gated environmental prior and gate penalty.
Task-additive	$q_0(X)$, reconstruction and factor-additive response loss.
Bilevel sparsity	$q_0(X)$, reconstruction and inner optimization of a sparse response predictor.
G2S-Jacobian	$q_0(X)$, reconstruction and task-Jacobian penalty for a dense response head.
TC-iVAE (ELBO)	$q_0(X, U)$ and $q_\tau(X, U, Y_\tau)$, full posterior covariance, composite ELBO.
TC-iVAE	Same branch inputs, factor-wise posterior covariance and gated response head, Eq. (20).

space. The final implementation sets posterior covariance across factors to zero while learning a 2×2 covariance within each factor. The response head is $\eta_\tau(z) = b_\tau + \sum_j g_{\tau j} r_{\tau j}(z_j)$, where $g_{\tau j} \in (0, 1)$ is a learned sigmoid gate. Because the exact posterior need not belong to these Gaussian families, they are variational approximations.

Let $\mathcal{R}, \mathcal{K}$ denote the average reconstruction loss and prior KL over the $K = T + 1 = 5$ branches, and let $\mathcal{C}$ denote the average response NLL over the four tasks. The negative composite ELBO with equal branch weights is

$$-\widehat{\mathcal{J}} = \mathcal{R} + \mathcal{K} + \tfrac{4}{5}\mathcal{C}. \tag{157}$$

Final training uses Eq. (20) with $\lambda_y = 2$, $\lambda_g = 0.05$, $\lambda_{\mathrm{loc}} = 0.50$, and $\lambda_J = 0.03$. The terms involving the gates and branch means μ_τ, μ_0 are

$$\Omega_g = \frac{1}{Td} \sum_{\tau,j} g_{\tau j}, \tag{158}$$

$$\Omega_{\mathrm{loc}} = \frac{1}{TD} \sum_{\tau,j} (1 - g_{\tau j})\, \mathbb{E}_{\mathrm{mb}} \|\mu_{\tau,j} - \mu_{0,j}\|_2^2, \tag{159}$$

$$\Omega_J = \frac{1}{T(C-1)d} \sum_{\tau=1}^{T} \sum_{c=1}^{C-1} \sum_{j=1}^{d} \sqrt{\mathbb{E}_{\mathrm{mb}} \|\nabla_{z_j} \eta_{\tau,c}(\mu_0)\|_2^2 / m + 10^{-12}} \tag{160}$$

Here, $C = 4$, $D = 6$, and $\mathbb{E}_{\mathrm{mb}}$ denotes the minibatch average. A locality cost is incurred when a response branch produces a large mean shift for a factor with a small gate. These terms do not use the true support. At $\beta = 1$, the final objective adds $1.2\mathcal{C}$ and the three regularization terms to Eq. (157). Because reweighting only the response term differs from changing all branch weights, we do not directly apply Proposition 1 to the optimum of the final regularized objective.

TC-iVAE (ELBO) and TC-iVAE differ jointly in posterior covariance, response weighting, and additional training terms, so their difference should be interpreted as a comparison of the overall implementations. The remaining comparison methods use full posterior covariance. All methods use the same data and training budget, although computational cost is not identical because the number of branches and active parameters differs.

The regularization coefficients of the comparison methods are fixed. Bilevel sparsity uses a group penalty of 0.03 over grouped first- and second-order features for each coordinate and ridge coefficient 10^{-3}. The inner predictor is fit on the first half of each minibatch, and the encoder is trained using the response loss on the second half. The task-wise coordinate Jacobian penalty coefficient for G2S-Jacobian is also 0.03. Mechanism-sparse multiplies by 0.03 the positive squared excess of the sum of the environment gates over the total latent dimension. The two endpoint-mixture weights of CI-iVAE are each $1/2$.

Table 13: Vector-factor recovery, information preservation in the full representation, and predictability of irrelevant factors from the selected representation. Values are mean ± SD over five seeds. Full nonlinear R^2 is the nonlinear readout score for predicting the complete true latent state from the complete learned latent representation.

Method	Forward factor R^2	Reverse factor R^2	Full nonlinear R^2	Irrel. linear R^2	Irrel. nonlinear R^2
iVAE(U)	0.5829 ± 0.1432	0.6044 ± 0.1499	0.9998 ± 0.0000	0.5984 ± 0.2024	0.5934 ± 0.2057
iVAE(U, Y)	0.5829 ± 0.1433	0.6045 ± 0.1450	0.9996 ± 0.0000	0.5980 ± 0.2026	0.5933 ± 0.2058
CI-iVAE(U, Y)	0.5828 ± 0.1426	0.6259 ± 0.1184	0.9996 ± 0.0001	0.6971 ± 0.0704	0.6939 ± 0.0708
Mechanism-sparse	0.5829 ± 0.1431	0.6192 ± 0.1354	0.9998 ± 0.0000	0.5985 ± 0.2025	0.5934 ± 0.2058
Task-additive	0.5872 ± 0.1385	0.6067 ± 0.1363	0.9994 ± 0.0001	0.6951 ± 0.0702	0.6938 ± 0.0709
Bilevel sparsity	0.6370 ± 0.1411	0.7137 ± 0.1722	0.9997 ± 0.0000	0.6709 ± 0.0698	0.6680 ± 0.0705
G2S-Jacobian	0.5870 ± 0.1384	0.6343 ± 0.1336	0.9994 ± 0.0001	0.6951 ± 0.0702	0.6937 ± 0.0709
TC-iVAE (ELBO)	0.6928 ± 0.2131	0.7161 ± 0.1972	0.9998 ± 0.0000	0.5477 ± 0.3139	0.5428 ± 0.3175
TC-iVAE	0.7398 ± 0.1837	0.7628 ± 0.1898	0.9998 ± 0.0000	0.4223 ± 0.2850	0.4155 ± 0.2888

E.2.3 Training and evaluation

The training and validation sample sizes are $32{,}768/4{,}096$. We train for 8,000 steps with Adam(3×10^{-4}) and batch size 256. The gradient norm is clipped at 10, and β increases from 0 to 1 over the first 1,600 steps. Every 250 steps, the validation objective of the corresponding method is evaluated with four posterior samples, and the checkpoint with the minimum value is selected. The validation score for the final TC-iVAE includes the response reweighting and regularization terms. The data-generating-model seed is 91831, the data seed is 20260917, and the training seeds are 43, 71, 101, 131, and 173. All seeds use the same generative model and data splits. Thus, the reported standard deviation reflects variation due to training randomness in this fixed data-generating environment.

For task-specific factor selection, a quadratic polynomial logistic probe is fit on up to 4,000 training examples for each of the seven nonempty factor subsets. Polynomial features are standardized, and logistic regression uses $C = 1$ with at most 500 L-BFGS iterations. The smallest subset whose validation NLL on 2,048 examples lies within one standard error of the minimum is then selected. Ties are broken by NLL and then index order. Calibration and test each use 2,048 examples with independent factors $S \sim \mathcal{N}(0, 0.8^2 I_6)$, $U = 0$, and clean observations. The representation is the base-posterior mean computed without responses.

Factor correspondence uses bidirectional RBF kernel-ridge R^2 between each true and learned factor. Kernel settings are selected on an internal calibration split, the Hungarian correspondence is fixed using a separate calibration evaluation set, and the score is computed on test data. Factor recovery in the main text is the mean forward nonlinear R^2 over $I_* = \{1, 2\}$. The reverse score is also reported. Relevant and irrelevant R^2 for the task-specific selected representation are computed on test data using linear readouts fit on calibration data, and negative values are retained. The nonlinear readout is used as a finite-sample recovery metric and does not by itself certify invertibility.

Editing uses 128 fixed test samples and donor samples obtained by a cyclic shift. One matched learned factor is replaced by the corresponding donor value, and the decoder output is mapped back through the simulator inverse. Let $\Delta\widetilde{S}$ denote the change from the pre-edit reconstruction, and define the target change as the difference between the donor and source true factors. RMS errors are computed after normalizing by the calibration standard deviation of each coordinate. We report the target-factor change error, the change in the other relevant factor, and the change in N_* separately. The main text averages over the two relevant targets.

E.2.4 Complete results and variability

The mean relevant-factor recovery of TC-iVAE changes from 0.693 for the composite-ELBO implementation to 0.740 for the final implementation, while the irrelevant linear R^2 of the selected representation changes from 0.548 to 0.422. The target change error decreases from 0.619 to 0.496, whereas the mean interference with the other relevant factor increases from 0.191 to 0.202. Thus, the effect of the additional terms depends on the metric.

For the final TC-iVAE, interference with the other relevant factor is below 0.008 in four seeds, while it is approximately 0.985 for seed 173. The Full nonlinear R^2 in Table 13 is approximately

Table 14: Complete results for editing relevant factors. Lower values are better for all metrics. Nuisance leak measures changes in factors irrelevant to all tasks and is reported separately from changes in the other relevant factor.

Method	Target Δ error	Other-relevant leak	Nuisance leak
iVAE(U)	0.7961 ± 0.2676	0.3988 ± 0.4097	0.3781 ± 0.2100
iVAE(U, Y)	0.7959 ± 0.2672	0.3993 ± 0.4100	0.3780 ± 0.2099
CI-iVAE(U, Y)	0.7947 ± 0.2687	0.2943 ± 0.2643	0.4602 ± 0.0397
Mechanism-sparse	0.7962 ± 0.2676	0.3988 ± 0.4096	0.3782 ± 0.2099
Task-additive	0.8012 ± 0.2725	0.2855 ± 0.2547	0.4448 ± 0.0412
Bilevel sparsity	0.7068 ± 0.2625	0.2046 ± 0.2721	0.4606 ± 0.0418
G2S-Jacobian	0.8013 ± 0.2722	0.2858 ± 0.2542	0.4601 ± 0.0413
TC-iVAE (ELBO)	0.6194 ± 0.4085	0.1914 ± 0.2535	0.3794 ± 0.2122
TC-iVAE	0.4963 ± 0.3450	0.2023 ± 0.4376	0.2964 ± 0.2661

0.999 or higher for all methods, whereas individual factor correspondence and the predictability of irrelevant factors from the selected representation differ substantially. This distinguishes preserving information in the full representation from making that information usable factor by factor.