BEYOND TASK SUCCESS: A THEORY OF REPRESENTATION REUSE

Anonymous authors
Paper under double-blind review

ABSTRACT

A learned representation is useful beyond its source task only if later tasks can make use of what the encoder has already learned. This matters most when the new task has limited supervision. Source-task performance, however, gives no guarantee that the representation itself will remain useful once the original head is removed. We study representation reuse with respect to a specified family of future tasks. For such a family, we define the task-relevant state as the coarsest deterministic state sufficient for all of its tasks. A larger task family can require a finer state, and losing a required distinction makes at least one future task unrecoverable on a target support containing that conflict. Exact reuse also depends on how the retained information is presented to the downstream rule. We separate this into structural accessibility, which concerns whether the required distinctions remain recoverable, and coordinate consistency, which concerns whether those distinctions can be interpreted by the same downstream rule across samples. Both failure modes appear separately in the experiments. Reuse improves as task-relevant distinctions become more accessible. Reversible sample-dependent coordinate changes can sharply reduce reuse even when the representation remains recoverable, and inverting the applicable coordinate change restores performance. The same accessibility–reuse ordering appears in prospectively specified CIFAR-100 experiments with both ResNet-18 and DenseNet-121. Finally, preserving a candidate state that contains the distinctions needed by the anticipated task family improves reuse on compatible held-out tasks, with the clearest benefit when target supervision is limited. Representation reuse is therefore relative to the tasks that follow source learning: a useful representation must retain the distinctions those tasks need and make them available in a form that a shared downstream rule can use.

1 INTRODUCTION

A learned representation may be reused after its original task changes. Once the source head is replaced, later tasks depend on distinctions retained and made usable by the encoder, especially when target supervision is limited. Source-task performance only shows that the encoder and source head work well together; it does not reveal which distinctions remain available or can be interpreted consistently across samples. Two models can therefore match on the source task yet differ substantially in reuse.

Consider a synthetic-aperture radar (SAR) encoder trained for target recognition and later reused, with few labeled examples, for change detection or correspondence. Two encoders can match in classification accuracy yet expose different information about local geometry, spatial arrangement, or scattering structure. Source accuracy therefore cannot determine which representation will support the new task. This motivates our general question: which distinctions must remain usable when the task changes?

Learned representations are commonly evaluated by frozen probes, feature transfer, transferability, and sample-efficient adaptation (Alain & Bengio, 2017; Yosinski et al., 2014; Kornblith et al., 2019b; Nguyen et al., 2020; Chen et al., 2024). Other work asks what pretraining transfers and how source-induced feature spaces interact with target tasks and supervision (Neyshabur et al., 2020; Tahir et al., 2025; Lauditi et al., 2026), or studies sufficiency, minimality, invariance, equivariance, and

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.

identifiability of learned representations (Achille & Soatto, 2018; Tian et al., 2020; Wang et al., 2022; Gupta et al., 2024; Khemakhem et al., 2020; Roeder et al., 2021; Reizinger et al., 2025; Nelson et al., 2026). Our question begins after the source decision rule is removed. Different future tasks can require different distinctions, so reuse cannot in general be described by a single task-independent score. We therefore define reuse with respect to a specified family of future tasks.

For a future-task family $\mathcal{H}$, we define the task-relevant state $s^\star_{\mathcal{H}}$ as the coarsest deterministic state sufficient for every task in the family. It captures the distinctions that must remain available to support all tasks in $\mathcal{H}$. If $\mathcal{H}_1 \subseteq \mathcal{H}_2$, then $s^\star_{\mathcal{H}_2}$ is at least as fine as $s^\star_{\mathcal{H}_1}$. Collapsing a distinction required by $s^\star_{\mathcal{H}}$ makes at least one task in $\mathcal{H}$ unrecoverable on a target support containing that conflict. Reuse is therefore defined by the distinctions required by the specified future-task family, not by the total amount of information retained by the representation.

Recovering the required distinctions is not sufficient for exact reuse. Their representation must also admit a common downstream interpretation. *Structural accessibility* concerns whether the required distinctions are recoverable from the representation. *Coordinate consistency* concerns whether the same downstream rule can interpret them across samples. A single globally shared reversible reparameterization can be absorbed by the decoder under unrestricted decoding. Sample-dependent conventions may prevent a common decoder even when the representation remains recoverable.

Source-matched models can still differ sharply in reuse. On CIFAR-100, nine matched representation pairs have a mean absolute source-performance gap of only 0.0036, while C1 exceeds C0 by 0.1718 in low-shot reuse. Reversible coordinate changes expose a separate failure: reuse falls from 0.4096 under a common convention to 0.0154 under sample-wise conventions even though the representation remains recoverable; inverting the applicable convention restores reuse to 0.4096. In prospectively specified CIFAR-100 experiments, held-out reuse increases from Q1 through Q5 for both ResNet-18 and DenseNet-121. On compatible held-out tasks, preserving a state that contains the distinctions required by the anticipated task family improves low-shot reuse by 0.1715; the full-support gain is 0.0241 and its confidence interval includes zero. The practical effect is clearest when target supervision is limited, where the measured reuse advantage is largest.

Contributions. We separate representation reuse from source-task success and show that source-matched representations can differ in reuse and can change ordering across future-task families. We formalize this dependence by defining the coarsest deterministic state sufficient for a specified family, deriving refinement and insufficient-state impossibility results, and characterizing exact reuse through structural accessibility and coordinate consistency. Intervention and rescue experiments test these two failure modes, including reversible coordinate changes that preserve recoverability while reducing blind reuse. We then connect the theory to source learning: preserving a candidate state that refines the anticipated task requirements improves low-shot reuse on compatible held-out tasks, without implying a task-independent advantage for unrelated families.

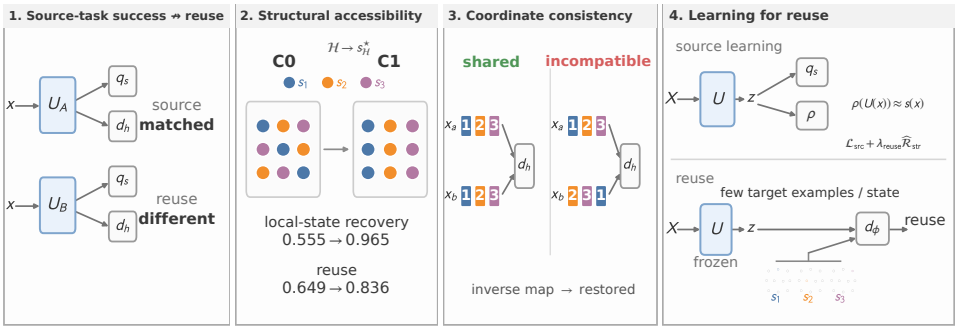

Figure 1: **Representation reuse.** **(1)** Matched source performance need not imply matched reuse. **(2)** Future-task distinctions must remain structurally accessible. **(3)** Retained distinctions must admit a shared interpretation. **(4)** Structure-preserving source learning promotes recoverability of an anticipated candidate state.

2 A TASK-RELATIVE THEORY OF REPRESENTATION REUSE

2.1 FROM SOURCE-TASK LEARNING TO REPRESENTATION REUSE

Let $U_\theta : \mathcal{X} \to \mathcal{Z}$ be an encoder and $q_\psi : \mathcal{Z} \to \mathcal{Y}_s$ a source head. Source learning minimizes

$$\mathcal{R}_{\mathrm{src}}(\theta, \psi) = \mathbb{E}_{(x,y_s)\sim P_s}\left[\ell_s\left(q_\psi(U_\theta(x)), y_s\right)\right]. \tag{1}$$

Let (θ^*, ψ^*) be a minimizer and write $U = U_{\theta^*}$. Low source-task risk concerns the composite predictor $q_{\psi^*} \circ U$ and does not determine which targets remain recoverable once the source head is removed.

For a target $h : \mathcal{X} \to \mathcal{Y}_t$, target distribution P_t, and downstream class $\mathcal{D} \subseteq \{d : \mathcal{Z} \to \mathcal{Y}_t\}$, define

$$\mathcal{R}_{\mathrm{reuse}}\left(U; h, \mathcal{D}\right) = \inf_{d\in\mathcal{D}} \mathbb{E}_{x\sim P_t}\left[\ell_t\left(d(U(x)), h(x)\right)\right]. \tag{2}$$

This population quantity is distinct from estimating a downstream rule from finite target supervision.

For deterministic h, exact reuse under unrestricted deterministic decoding means that there exist $d_h : \mathcal{Z} \to \mathcal{Y}_t$ and $\Omega_t \subseteq \mathrm{supp}(P_t)$ with $P_t(\Omega_t) = 1$ such that

$$h(x) = d_h(U(x)), \qquad \forall x \in \Omega_t. \tag{3}$$

Population-risk maps are assumed measurable. The exact existence results below are deterministic statements; the corresponding measurability qualification is given in the Appendix.

In our primary experiments, source learning uses pairwise relations. Let $q^{\mathrm{rel}}_\psi : \mathcal{Z} \times \mathcal{Z} \to \mathcal{R}$ be the relational source head, with population objective

$$\mathcal{L}_{\mathrm{rel}}(\theta, \psi) = \mathbb{E}_{(x_i,x_j,r_{ij})}\left[\ell_{\mathrm{rel}}\left(q^{\mathrm{rel}}_\psi\left(U_\theta(x_i), U_\theta(x_j)\right), r_{ij}\right)\right]. \tag{4}$$

We write $\widehat{\mathcal{L}}_{\mathrm{src}}$ for the empirical source-learning loss actually optimized in the corresponding experimental setting.

2.2 THE TASK-RELEVANT STATE OF A FUTURE-TASK FAMILY

Learning representations that support multiple related tasks has a long history in inductive-bias learning and multitask representation learning (Baxter, 2000; Maurer et al., 2016). We instead treat the future-task family as specified and ask for the coarsest deterministic state sufficient for every task in that family.

Let $\mathcal{H} = \{h_1, \ldots, h_M\}$ be a finite family of deterministic future tasks. Define

$$x \sim_{\mathcal{H}} x' \quad \Longleftrightarrow \quad h(x) = h(x') \ \ \forall h \in \mathcal{H}, \tag{5}$$

and

$$s^\star_{\mathcal{H}}(x) = [x]_{\mathcal{H}}, \qquad \mathcal{S}^\star_{\mathcal{H}} = \mathcal{X}/\!\sim_{\mathcal{H}}. \tag{6}$$

A deterministic state $s : \mathcal{X} \to \mathcal{S}$ is sufficient for $\mathcal{H}$ if every $h \in \mathcal{H}$ factors through s. By construction, every task factors through $s^\star_{\mathcal{H}}$, and every sufficient s satisfies

$$s^\star_{\mathcal{H}} = r \circ s \tag{7}$$

for some deterministic $r : \mathcal{S} \to \mathcal{S}^\star_{\mathcal{H}}$. Thus, up to relabeling, $s^\star_{\mathcal{H}}$ is the coarsest deterministic state sufficient for the family. This construction is related to prior notions of sufficient and minimal representations (Achille & Soatto, 2018; Tian et al., 2020; Wang et al., 2022), but here sufficiency is defined relative to an explicit family of future tasks.

Lemma 1 (Task-Family Refinement). If $\mathcal{H}_1 \subseteq \mathcal{H}_2$, then there exists a deterministic map r_{21} such that

$$s^\star_{\mathcal{H}_1} = r_{21} \circ s^\star_{\mathcal{H}_2}. \tag{8}$$

Hence enlarging the future-task family cannot make its required state strictly coarser.

Corollary 1 (Insufficient-State Impossibility). If $s^\star_{\mathcal{H}}$ does not factor through a state $s : \mathcal{X} \to \mathcal{S}$, then there exist $x_a, x_b \in \mathcal{X}$ and $h \in \mathcal{H}$ such that

$$s(x_a) = s(x_b), \qquad h(x_a) \neq h(x_b). \tag{9}$$

No deterministic decoder acting only through s can recover that task exactly on a support containing both inputs. A strictly positive distributional error lower bound additionally requires non-degenerate conditional target labels on a set of positive state measure; see the Appendix.

2.3 Exact Reuse Through an Effective Interface

The task-relevant state specifies which distinctions must remain available. Exact reuse also depends on whether those distinctions admit a consistent interpretation through the representation exposed to a downstream rule. Throughout this subsection, decoding is unrestricted and deterministic.

Prior identifiability results study when learned representations are recoverable or aligned only up to shared transformation classes (Khemakhem et al., 2020; Roeder et al., 2021; Reizinger et al., 2025; Nelson et al., 2026). Related work on representation comparison likewise makes explicit that the notion of equivalence depends on which coordinate transformations are treated as immaterial (Kornblith et al., 2019a). Given the representation interface exposed to reuse, we ask whether all admissible realizations admit one common downstream interpretation.

For each $x \in \mathcal{X}$, let $\mathcal{G}(x) \subseteq \mathrm{Bij}(\mathcal{Z})$ be a nonempty family of admissible reversible coordinate transformations with $\mathrm{id} \in \mathcal{G}(x)$. The admissible family is fixed by the reuse setting independently of observed outcomes. Define the *effective interface*

$$\mathcal{E}^{\mathcal{G}}_U(x) = \{g(U(x)) : g \in \mathcal{G}(x)\}. \tag{10}$$

A target h is exactly reusable over this interface if, for some full-measure $\Omega_t \subseteq \mathrm{supp}(P_t)$, one decoder d satisfies

$$d(z) = h(x), \qquad \forall x \in \Omega_t,\ \forall z \in \mathcal{E}^{\mathcal{G}}_U(x). \tag{11}$$

Theorem 1 (Effective-Interface Reuse Criterion). A deterministic target h is exactly reusable over the effective interface if and only if there exists a full-measure $\Omega_t \subseteq \mathrm{supp}(P_t)$ such that

$$\mathcal{E}^{\mathcal{G}}_U(x_a) \cap \mathcal{E}^{\mathcal{G}}_U(x_b) \neq \varnothing \quad \Longrightarrow \quad h(x_a) = h(x_b), \qquad \forall x_a, x_b \in \Omega_t. \tag{12}$$

Failure occurs exactly when two inputs requiring different target outputs can present a common effective representation to the shared decoder. The proof and measurability qualification are given in Appendix A.2.

Structural accessibility and coordinate consistency. Define

$$\mathcal{A}_{U,h}(\Omega_t) = \left\{(x_a, x_b) \in \Omega_t^2 : U(x_a) = U(x_b),\ h(x_a) \neq h(x_b)\right\}, \tag{13}$$

$$\mathcal{K}^{\mathcal{G}}_{U,h}(\Omega_t) = \left\{(x_a, x_b) \in \Omega_t^2 : U(x_a) \neq U(x_b),\ \mathcal{E}^{\mathcal{G}}_U(x_a) \cap \mathcal{E}^{\mathcal{G}}_U(x_b) \neq \varnothing,\ h(x_a) \neq h(x_b)\right\}. \tag{14}$$

Structural accessibility holds when $\mathcal{A}_{U,h}(\Omega_t) = \varnothing$, and coordinate consistency when $\mathcal{K}^{\mathcal{G}}_{U,h}(\Omega_t) = \varnothing$. Because $\mathrm{id} \in \mathcal{G}(x)$, these distinguish the two ways the criterion can fail: a required distinction may already be collapsed by U, or distinct base representations may become incompatible through their effective interfaces.

For the identity-interface special case $\mathcal{G}(x) = \{\mathrm{id}\}$, coordinate conflicts vanish and structural accessibility alone is sufficient for unrestricted exact reuse. The graded conditional-accessibility quantity used in Section 3 is an empirical intervention diagnostic, not this binary exact condition; no universal monotonicity with restricted reuse follows from Theorem 1.

Corollary 2 (Family Reuse as Task-State Recovery). For finite $\mathcal{H}$ under a common target distribution, exact reuse of every task over a common full-measure support $\Omega_{\mathcal{H}}$ is equivalent to recovery of the family state:

$$\begin{aligned} &\forall h \in \mathcal{H},\ \exists d_h :\ d_h(z) = h(x) \\ &\qquad \forall x \in \Omega_{\mathcal{H}},\ \forall z \in \mathcal{E}^{\mathcal{G}}_U(x) \\ &\iff \exists \rho :\ \rho(z) = s^\star_{\mathcal{H}}(x) \quad \forall x \in \Omega_{\mathcal{H}},\ \forall z \in \mathcal{E}^{\mathcal{G}}_U(x). \end{aligned} \tag{15}$$

Proposition 1 (Shared-Coordinate Invariance). For any bijection $\tau : \mathcal{Z} \to \mathcal{Z}$, let

$$\widetilde{U} = \tau \circ U, \qquad \widetilde{\mathcal{G}}(x) = \{\tau \circ g \circ \tau^{-1} : g \in \mathcal{G}(x)\}.$$

Then

$$\mathcal{E}^{\widetilde{\mathcal{G}}}_{\widetilde{U}}(x) = \tau\big(\mathcal{E}^{\mathcal{G}}_{U}(x)\big)\,,$$

and $d \circ \tau^{-1}$ decodes the transformed representation whenever d decodes the original one. Hence a globally shared bijective reparameterization preserves unrestricted exact reuse.

For a restricted decoder class $\mathcal{D}$, define

$$\widetilde{\mathcal{D}} = \{d \circ \tau^{-1} : d \in \mathcal{D}\}.$$

Comparing $(U, \mathcal{D})$ with $(\widetilde{U}, \widetilde{\mathcal{D}})$ preserves the corresponding reuse risk. If the same fixed decoder class is used before and after reparameterization, closure under the corresponding forward and inverse compositions is a sufficient condition for the same invariance.

For $x_{1:m} = (x_1, \ldots, x_m)$, define

$$U^{(m)}(x_{1:m}) = (U(x_1), \ldots, U(x_m))\,.$$

Let $\mathcal{G}^{(m)}(x_{1:m})$ act componentwise:

$$\mathcal{G}^{(m)}(x_{1:m}) = \{(g_1, \ldots, g_m) : g_j \in \mathcal{G}(x_j)\}\,,$$

with

$$(g_1, \ldots, g_m)\,(U(x_1), \ldots, U(x_m)) = (g_1(U(x_1)), \ldots, g_m(U(x_m)))\,.$$

The joint effective interface is therefore

$$\mathcal{E}^{\mathcal{G}^{(m)}}_{U^{(m)}}(x_{1:m}) = \prod_{j=1}^{m} \mathcal{E}^{\mathcal{G}}_{U}(x_j). \tag{16}$$

2.4 Learning for Reuse Under Candidate-State and Decoder Constraints

Multitask learning commonly uses related auxiliary objectives to shape a shared representation (Caruana, 1997; Maurer et al., 2016). Here the candidate state is tied explicitly to an anticipated future-task family and is related to the corresponding task-relevant state through Eq. (18).

Let $s : \mathcal{X} \to \mathcal{S}$ be a candidate state. For an m-input downstream setting, let

$$\Phi_m \subseteq \{\phi : \mathcal{S}^m \to \mathcal{Y}_t\}$$

be a class of composition rules and define

$$\mathcal{H}^{(m)}_S = \{h_\phi : h_\phi(x_{1:m}) = \phi(s(x_1), \ldots, s(x_m))\,,\ \phi \in \Phi_m\}\,. \tag{17}$$

Writing

$$s^{(m)}(x_{1:m}) = (s(x_1), \ldots, s(x_m))\,,$$

any finite anticipated family $\mathcal{H} \subseteq \mathcal{H}^{(m)}_S$ satisfies

$$s^\star_{\mathcal{H}} = r \circ s^{(m)} \tag{18}$$

for some $r : \mathcal{S}^m \to \mathcal{S}^\star_{\mathcal{H}}$. Thus the tuple extension of the candidate state refines the task-relevant state required by the anticipated family.

Let $\mathcal{D}_{\mathrm{str}}$ be a class of readouts $\rho : \mathcal{Z} \to \mathcal{S}$. Targeting candidate-state recoverability during source learning corresponds to

$$\mathcal{R}_{\mathrm{str}}\,(U; s, \mathcal{D}_{\mathrm{str}}) = \inf_{\rho \in \mathcal{D}_{\mathrm{str}}} \mathbb{E}_{X \sim P^X_s}\left[\ell_{\mathrm{str}}\,(\rho(U(X)), s(X))\right]. \tag{19}$$

For the discrete states used experimentally, we optimize

$$\widehat{\mathcal{R}}_{\mathrm{str}}(\theta, \omega) = -\frac{1}{N}\sum_{i=1}^{N} \log p_\omega\,(s_i \mid U_\theta(x_i)) \tag{20}$$

and train with

$$\mathcal{L}_{\mathrm{SPR}}(\theta, \psi, \omega) = \widehat{\mathcal{L}}_{\mathrm{src}}(\theta, \psi) + \lambda_{\mathrm{reuse}}\widehat{\mathcal{R}}_{\mathrm{str}}(\theta, \omega). \tag{21}$$

SPR targets the structural-accessibility side of the theory. It does not directly enforce coordinate consistency, and the structural readout is discarded before reuse evaluation.

Proposition 2 (Candidate-State Recovery Suffices for Reuse). Let $P_t^{(m)}$ be a downstream distribution on $\mathcal{X}^m$. Suppose there exist $\Omega_t \subseteq \mathcal{X}$ and $\rho : \mathcal{Z} \to \mathcal{S}$ such that

$$P_t^{(m)}(\Omega_t^m) = 1 \tag{22}$$

and

$$\rho(z) = s(x), \qquad \forall x \in \Omega_t,\ \forall z \in \mathcal{E}_U^{\mathcal{G}}(x). \tag{23}$$

Then every $h_\phi \in \mathcal{H}_S^{(m)}$ is exactly reusable over the joint effective interface via

$$d_\phi(z_1, \ldots, z_m) = \phi(\rho(z_1), \ldots, \rho(z_m)). \tag{24}$$

The proposition assumes candidate-state recovery on the target effective interface. The source-domain SPR objective promotes such recoverability during training but does not itself establish this target-domain condition.

Restricted reuse under decoder constraints. For restricted reuse, we fix the realized coordinates of U and separate decoder-class restriction from coordinate ambiguity. Let

$$\mathcal{Q}_s \subseteq \{a : \mathcal{Z} \to \mathcal{S}\}$$

be a nonempty class of target-domain candidate-state readouts, and let

$$\mathcal{D}_m \subseteq \{d : \mathcal{Z}^m \to \mathcal{Y}_t\}$$

be a nonempty downstream decoder class.

For $X_{1:m} \sim P_t^{(m)}$, define under 0–1 loss

$$\epsilon_s(a) = \frac{1}{m} \sum_{j=1}^{m} \Pr\left[a(U(X_j)) \neq s(X_j)\right], \tag{25}$$

$$\eta_\phi(a) = \inf_{d \in \mathcal{D}_m} \Pr\left[d(U^{(m)}(X_{1:m})) \neq \phi(a(U(X_1)), \ldots, a(U(X_m)))\right], \tag{26}$$

$$\alpha_\phi = \inf_{d \in \mathcal{D}_m} \Pr\left[d(U^{(m)}(X_{1:m})) \neq h_\phi(X_{1:m})\right]. \tag{27}$$

For a fixed a, write

$$G_{\phi,a} = \phi(a(U(X_1)), \ldots, a(U(X_m))).$$

Proposition 3 (Restricted-Reuse Sandwich). For every $a \in \mathcal{Q}_s$,

$$|\alpha_\phi - \eta_\phi(a)| \leq \Pr\left[G_{\phi,a} \neq h_\phi(X_{1:m})\right] \leq m\,\epsilon_s(a). \tag{28}$$

Consequently,

$$\sup_{a \in \mathcal{Q}_s} \left[\eta_\phi(a) - m\epsilon_s(a)\right]_+ \leq \alpha_\phi \leq \inf_{a \in \mathcal{Q}_s} \left[\eta_\phi(a) + m\epsilon_s(a)\right]. \tag{29}$$

The same readout couples recoverability and restricted composition, so better candidate-state recovery need not monotonically improve restricted reuse; Appendix A.10 gives a counterexample. For finite $\mathcal{D}_m$, fitting from n i.i.d. target tuples adds the standard $O(\sqrt{\log|\mathcal{D}_m|/n})$ estimation term (Appendix A.9). SPR optimizes a source-domain surrogate and directly optimizes neither η_ϕ, this finite-sample term, nor coordinate consistency; Section 3 evaluates the corresponding empirical behavior.

3 Experiments

The experiments follow the three questions raised by the theory. First, can representations with matched source-task performance nevertheless support substantially different future tasks after the source head is removed? Second, when reuse differs, can the failure be separated into loss of task-relevant distinctions and loss of a shared downstream interpretation? Third, if the future-task family is anticipated, can source learning preserve the required distinctions so that later tasks reuse the representation more effectively under limited target supervision?

C0 (*Source-only*) uses only the source objective. C1 (*SPR*) adds Eq. (21). C4 (*Globally relabeled SPR*) uses the same objective after one fixed bijective relabeling of candidate-state labels; representation coordinates are unchanged. Unless stated otherwise, reuse evaluation discards the source head, freezes the encoder, and fits a new downstream rule. Implementation details, admissibility criteria, and statistical procedures are given in the Appendix.

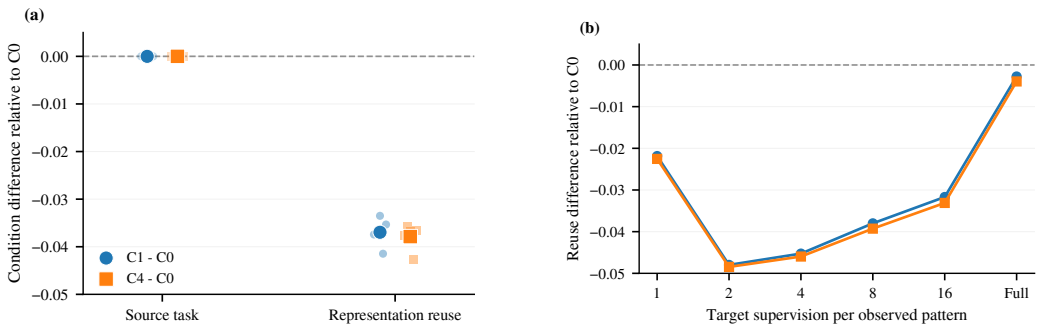

Figure 2: **Source-task performance does not determine reuse. (a)** Paired reuse differences under exact source matching. **(b)** Mean reuse differences as target supervision increases; the dashed line denotes zero.

3.1 Source-Task Success Does Not Determine Representation Reuse

This section tests whether source-task success can rank learned representations for future reuse; it does not assume that SPR should improve every replacement task family. Under a task-relative account, representations with matched source performance can differ in reuse, and their ordering need not be preserved across future-task families that require different distinctions.

Exact source matching and target supervision. On Shapes3D (Kim & Mnih, 2018), five seeds for each of C0, C1, and C4 all achieve exactly 1.0000 source validation accuracy. We use *low-shot reuse* for the predefined aggregate over the five limited-supervision budgets $K \in \{1, 2, 4, 8, 16\}$. Across 1004 admissible novel-rule evaluations, mean low-shot reuse is 0.6410, 0.6040, and 0.6031, respectively; the paired C1–C0 and C4–C0 differences are -0.0370 and -0.0379, with the same sign for all five seeds (Figure 2(a)). With full target support, these differences contract to -0.0028 and -0.0040 (Figure 2(b)). The source-matched representations differ in reuse, with a larger separation under limited target supervision.

Extension across data and source-learning settings. On CIFAR-100 (Krizhevsky, 2009), nine matched representation pairs have a mean absolute source-performance gap of 0.0036, while C1 exceeds C0 in low-shot reuse by 0.1718. Under an ordinary supervised-classification source task, source accuracy is similarly close (0.7535 versus 0.7602), while the held-out-rule reuse difference is $+0.0768$. Source-matched representations therefore need not have matched reuse, and their ordering can change across future-task families that require different distinctions. In our experiments, preserving a candidate state helps tasks that depend on that state but does not make the resulting representation uniformly better for unrelated replacement families. This family dependence is consistent with the task-relative account.

3.2 Structural Accessibility and Coordinate Consistency

We test finite-supervision counterparts of the two exact-reuse requirements in Section 2.3. These measurements are empirical diagnostics, not the exact conflict sets defined by the theory.

Candidate-state recoverability. Across five matched seeds, C1 raises aligned held-out state macro-F1 from 0.5547 ± 0.1807 to 0.9646 ± 0.0052, a paired gain of 0.4099 (95% CI $[0.2719, 0.5479]$). Low-shot reuse rises from 0.6491 ± 0.1130 to 0.8356 ± 0.0274, with a paired gain of 0.1866 ($[0.1012, 0.2875]$), positive in all five seeds (Figure 3(a)). C4 reaches similar recoverability (0.9572 ± 0.0109) and differs from C1 in reuse by -0.0072 ($[-0.0375, 0.0203]$), consistent with insensitivity to a global relabeling of candidate-state symbols. Random-auxiliary and Shuffled-state controls have lower recoverability and reuse; C1 exceeds them in reuse by 0.5845 and 0.6576, respectively, with the same sign in all five seeds. The aligned macro-F1 is diagnostic only and is never supplied to the downstream reuse evaluator.

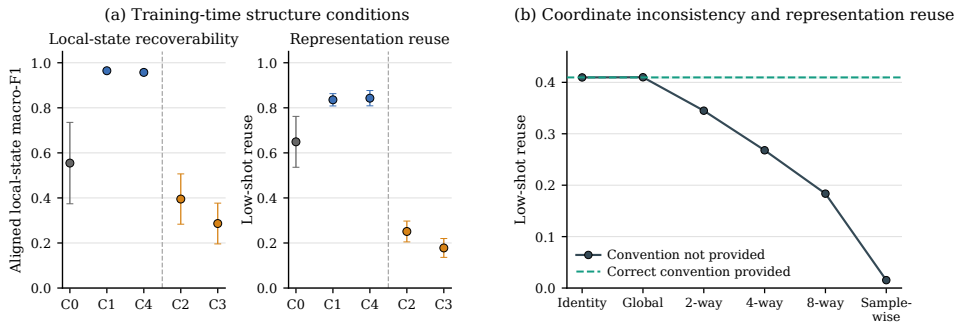

Figure 3: **Accessibility and coordinate tests. (a)** State recoverability and low-shot reuse for C0, C1, C4, Random auxiliary (C2), and Shuffled-state (C3). **(b)** Reuse under reversible coordinate interventions; increasingly heterogeneous sample-dependent conventions reduce blind reuse, while convention-aware inversion returns reuse to the Identity condition. Error bars show one standard deviation.

Recoverable information can still be unusable. We apply reversible coordinate transformations to the same frozen representation. Reuse is 0.4096 under Identity and 0.4099 under one globally shared reparameterization, but falls to 0.3449, 0.2679, 0.1836, and 0.0154 under 2-way, 4-way, 8-way, and sample-wise conventions, respectively (Figure 3(b)). The same ordering holds for all three source-training seeds. Every transformation is invertible. Supplying the applicable convention and applying its inverse restores reuse to 0.4096 in every condition, with zero reconstruction difference from the common-reference representation. The information needed for reuse therefore remains recoverable, yet a shared downstream rule fails when different samples expose that information through incompatible coordinate conventions.

Prospective natural-image replication. We test five protocol-defined CIFAR-100 conditions Q1–Q5 with normalized conditional-accessibility values approximately 0, 0.4182, 0.6555, 0.8277, and 1.0000. These are predefined intervention levels, not post hoc estimates of information retained by U; Q1$=0$ refers only to the controlled conditional component. Construction and normalization details are given in the Appendix.

Held-out reuse increases from Q1 through Q5 for both architectures. ResNet-18 rises from 0.2670 to 0.3213 ($\Delta_{5-1} = +0.0542$), and DenseNet-121 from 0.2811 to 0.3261 ($\Delta_{5-1} = +0.0450$). Full condition-wise values are reported in Table 5 of Appendix C.6. At the representation-seed level, all five accessibility–reuse slopes are positive for each architecture. Mean slopes are 0.0516 for ResNet-18 (95% exact-bootstrap CI $[0.0485,\ 0.0547]$) and 0.0460 for DenseNet-121 ($[0.0440,\ 0.0477]$). The seed-level estimand and resampling procedure are given in the Appendix.

The result is specific to the predefined accessibility hierarchy, future-task family, and intervention construction. Proposition 3 does not imply a universal monotonic relationship between accessibility and restricted reuse.

3.3 Structure-Preserving Learning Improves Reuse under Limited Target Supervision

On held-out tasks whose task-relevant states are refined by the candidate state promoted during source learning, the reuse gain is concentrated under limited target supervision. Across $K \in \{1, 2, 4, 8, 16\}$ examples per state, the paired C1–C0 low-shot gain is 0.1715 (95% CI $[0.0986,\ 0.2591]$), and the C4–C0 gain is 0.1771 ($[0.1172,\ 0.2593]$); both are positive in all five matched seeds (Figure 4). Under full target support, the corresponding gains are 0.0241 ($[-0.0260,\ 0.0992]$) and 0.0300 ($[-0.0201,\ 0.0990]$).

These tasks instantiate $\mathcal{H}_S^{(m)}$ from Eq. (17): they share the candidate state s and vary the composition rule ϕ, so their task-relevant states are refined by $s^{(m)}$ as in Eq. (18). The primary paired analysis retains 21 of 24 registered family–rule evaluations after a representation-independent admissibility screen; registry construction and class support are given in the Appendix.

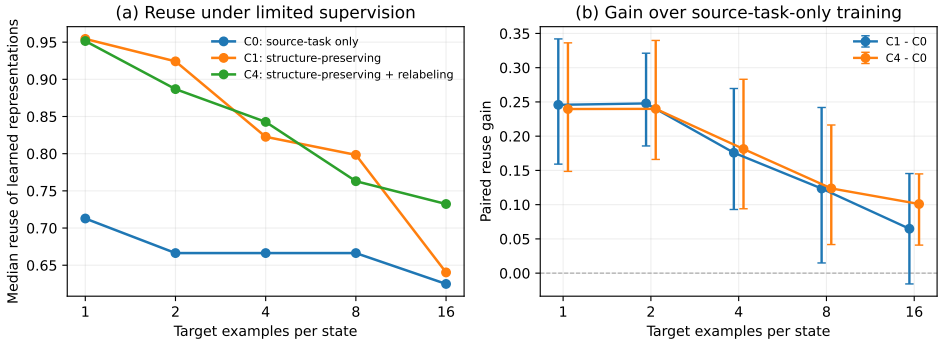

Figure 4: **Structure preservation improves reuse primarily under limited target supervision.** **(a)** Reuse across the predefined target-supervision conditions for C0, C1, and C4. **(b)** Paired seed-level gains over C0: C1 improves low-shot reuse by 0.1715, whereas the full-support gain is 0.0241 and its confidence interval includes zero. C4 shows the same supervision-localized pattern. Panel (a) is descriptive; panel (b) reports the paired estimand with bootstrap 95% confidence intervals.

Let $\Delta_{\mathrm{LS}}(C)$ and $\Delta_{\mathrm{full}}(C)$ denote paired gains under the predefined limited-supervision and full-support regimes, respectively. Define

$$\Gamma(C) = \Delta_{\mathrm{LS}}(C) - \Delta_{\mathrm{full}}(C). \tag{30}$$

We obtain $\Gamma(\mathrm{C1}) = 0.1474$ (95% CI $[0.1044, 0.1905]$) and $\Gamma(\mathrm{C4}) = 0.1471$ ($[0.1193, 0.1749]$). Both Γ values are positive, while both full-support intervals include zero. The advantage of C1 and C4 over C0 is concentrated under limited target supervision. In this setting, distinctions already accessible from the frozen representation matter most when the downstream learner has few labeled target examples.

The direct C1–C4 low-shot contrast is -0.0056 (95% CI $[-0.0284, 0.0173]$), with no consistent direction across seeds, consistent with insensitivity to a global relabeling of candidate-state symbols. The gain reported here is specific to the predefined held-out tasks that factor through s. As Section 3.1 shows, families requiring different distinctions can induce a different ordering of the same source-matched representations.

4 Conclusion

A representation can solve its source task well and still be difficult to reuse once the source head is removed. Our framework separates two failures: distinctions required by later tasks may no longer be recoverable, or they may remain recoverable but be exposed through coordinates that a shared downstream rule cannot interpret consistently. The reversible coordinate experiment isolates the second case: reuse falls under sample-dependent conventions and returns when the applicable convention is inverted.

The practical consequence is clearest when target supervision is limited. The predefined accessibility ordering repeats with both ResNet-18 and DenseNet-121, and preserving a candidate state that refines the held-out task requirements improves low-shot reuse while the full-support intervals include zero. The benefit is family-specific: future tasks requiring different distinctions can change the ordering of source-matched representations. Representation reuse should therefore be evaluated relative to the tasks that follow source learning—whether the distinctions they require remain available in a form that a shared downstream rule can use.

AI use statement. Generative AI tools were used to assist with literature organization, translation, language editing, and limited implementation support during the preparation of this work. All scientific claims, theoretical statements, experimental designs, implementations, numerical results, and final manuscript content were reviewed and verified by the authors. The authors take full responsibility for the content of this paper.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9, 2018. doi: 10.1109/ITA.2018.8503149.

Guillaume Alain and Yoshua Bengio. Understanding Intermediate Layers Using Linear Classifier Probes. In *International Conference on Learning Representations*, 2017. URL https://mlanthology.org/iclr/2017/alain2017iclr-understanding/.

Jonathan Baxter. A model of inductive bias learning. *J. Artif. Int. Res.*, 12(1):149–198, March 2000. ISSN 1076-9757.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, Jul 1997. doi: 10.1023/a:1007379606734. URL http://dx.doi.org/10.1023/A:1007379606734.

Annie Chen, Yoonho Lee, Amrith Setlur, Sergey Levine, and Chelsea Finn. Project and probe: Sample-efficient adaptation by interpolating orthogonal features. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 11788–11811, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/3209cf3312b2cbb68e33644362ddc2bd-Paper-Conference.pdf.

Sharut Gupta, Joshua Robinson, Derek Lim, Soledad Villar, and Stefanie Jegelka. Structuring representation geometry with rotationally equivariant contrastive learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lgaFMvZHSJ.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

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, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019a. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2656–2666, 2019b. doi: 10.1109/CVPR.2019.00277.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Clarissa Lauditi, Blake Bordelon, and Cengiz Pehlevan. Transfer learning in infinite width feature learning networks. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 53982–54028, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/58ab5d6ea13c87550a2eb51c4aa1420f-Paper-Conference.pdf.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *J. Mach. Learn. Res.*, 17(1):2853–2884, January 2016. ISSN 1532-4435.

Walter Nelson, Marco Fumero, THEOFANIS KARALETSOS, and Francesco Locatello. Statistical and structural identifiability in representation learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 151985–152020, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f67e5f99b23b108a3a6665f410034bcd-Paper-Conference.pdf.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 512–523. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/0607f4c705595b911a4f3e7a127b44e0-Paper.pdf.

Cuong V. Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: a new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Patrik Reizinger, Alice Bizeul, Attila Juhos, Julia E Vogt, Randall Balestriero, Wieland Brendel, and David Klindt. Cross-entropy is all you need to invert the data generating process. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 38186–38216, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5ee7ed60a7e8169012224dec5fe0d27f-Paper-Conference.pdf.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9030–9039. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/roeder21a.html.

Javan Tahir, Surya Ganguli, and Grant M. Rotskoff. Features are fate: a theory of transfer learning in high-dimensional regression. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58142–58168. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/tahir25a.html.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6827–6839. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c2e5eaae9152079b9e95845750bb9ab-Paper.pdf.

Haoqing Wang, Xun Guo, ZhiHong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16020–16029, 2022. doi: 10.1109/CVPR52688.2022.01557.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/532a2f85b6977104bc93f8580abbb330-Paper.pdf.

A Formal Details and Proofs

A.1 Minimal State for a Finite Future-Task Family

Let $\mathcal{H}$ be a finite family of deterministic target tasks on $\mathcal{X}$. Define

$$x \sim_{\mathcal{H}} x' \quad \Longleftrightarrow \quad h(x) = h(x') \quad \forall h \in \mathcal{H}.$$

The corresponding task-relevant state is

$$s^\star_{\mathcal{H}}(x) = [x]_{\mathcal{H}}.$$

For every $h \in \mathcal{H}$, $h(x)$ is constant on each equivalence class. Hence there exists a function ϕ_h such that

$$h = \phi_h \circ s^\star_{\mathcal{H}}.$$

Thus $s^\star_{\mathcal{H}}$ is sufficient for every task in $\mathcal{H}$.

It is also the coarsest deterministic state with this property. Suppose

$$s : \mathcal{X} \to \mathcal{S}$$

supports every task in $\mathcal{H}$, so that for each $h \in \mathcal{H}$ there exists $\widetilde{\phi}_h$ satisfying

$$h = \widetilde{\phi}_h \circ s.$$

If

$$s(x) = s(x'),$$

then

$$h(x) = \widetilde{\phi}_h(s(x)) = \widetilde{\phi}_h(s(x')) = h(x') \qquad \forall h \in \mathcal{H}.$$

Therefore

$$x \sim_{\mathcal{H}} x',$$

so $s^\star_{\mathcal{H}}$ is constant on every fiber of s. Consequently, there exists a deterministic map r such that

$$s^\star_{\mathcal{H}} = r \circ s.$$

Hence every deterministic state sufficient for the entire family refines $s^\star_{\mathcal{H}}$. Any other coarsest deterministic sufficient state induces the same partition of $\mathcal{X}$ and differs from $s^\star_{\mathcal{H}}$ only by a bijective relabeling of its state values.

Task-family refinement. Let

$$\mathcal{H}_1 \subseteq \mathcal{H}_2.$$

If

$$s^\star_{\mathcal{H}_2}(x) = s^\star_{\mathcal{H}_2}(x'),$$

then every task in $\mathcal{H}_2$ takes the same value on x and x'. Because $\mathcal{H}_1 \subseteq \mathcal{H}_2$, the same is true for every task in $\mathcal{H}_1$, and therefore

$$s^\star_{\mathcal{H}_1}(x) = s^\star_{\mathcal{H}_1}(x').$$

Thus $s^\star_{\mathcal{H}_1}$ is constant on every fiber of $s^\star_{\mathcal{H}_2}$, so there exists a deterministic map r_{21} satisfying

$$s^\star_{\mathcal{H}_1} = r_{21} \circ s^\star_{\mathcal{H}_2}.$$

Enlarging the future-task family therefore cannot make its required state strictly coarser.

Insufficient-state impossibility. Let

$$s : \mathcal{X} \to \mathcal{S}.$$

Suppose $s^\star_{\mathcal{H}}$ does not factor through s. Then $s^\star_{\mathcal{H}}$ is not constant on at least one fiber of s, so there exist $x_a, x_b \in \mathcal{X}$ such that

$$s(x_a) = s(x_b), \qquad s^\star_{\mathcal{H}}(x_a) \neq s^\star_{\mathcal{H}}(x_b).$$

By definition of $s^\star_{\mathcal{H}}$, there exists some $h \in \mathcal{H}$ for which

$$h(x_a) \neq h(x_b).$$

Any deterministic decoder that observes only $s(x)$ must assign the same output to x_a and x_b, and therefore cannot recover that task exactly on a target support containing both inputs.

This conclusion is pointwise. A strictly positive population-risk lower bound requires the conflicting target labels to remain indistinguishable on a set with positive probability under the target distribution.

Distributional Bayes obstruction. Let

$$S = s(X), \qquad Y = h(X),$$

where $\mathcal{Y}_t$ is finite and the relevant regular conditional distribution exists. Define

$$B_s(h) = \mathbb{E}_S \left[1 - \max_{y \in \mathcal{Y}_t} \Pr(Y = y \mid S) \right]. \tag{31}$$

This is the Bayes error of predicting $h(X)$ from $s(X)$. For every measurable decoder $\delta : \mathcal{S} \to \mathcal{Y}_t$,

$$\Pr[\delta(S) \neq Y] \geq B_s(h).$$

Moreover, if

$$U = f \circ s$$

for a deterministic measurable map f, every measurable decoder acting through U is also a measurable function of S. Hence

$$\inf_d \Pr[d(U(X)) \neq h(X)] \geq B_s(h).$$

The quantity $B_s(h)$ is strictly positive precisely when

$$1 - \max_y \Pr(Y = y \mid S) > 0$$

on a set of state values having positive P_S-measure. Thus positive distributional error requires non-degenerate conditional target labels on a set of positive measure; it does not require any individual state value to have positive probability.

A.2 Proof of Theorem 1

Let $\Omega \subseteq \mathrm{supp}(P_t)$ satisfy

$$P_t(\Omega) = 1.$$

For every $x \in \Omega$, let

$$\mathcal{E}_U^{\mathcal{G}}(x) = \{g(U(x)) : g \in \mathcal{G}(x)\}$$

denote the effective interface.

We prove that exact reuse over the effective interface is possible if and only if

$$\mathcal{E}_U^{\mathcal{G}}(x_a) \cap \mathcal{E}_U^{\mathcal{G}}(x_b) \neq \varnothing \quad \Longrightarrow \quad h(x_a) = h(x_b) \tag{32}$$

for all $x_a, x_b \in \Omega$.

Necessity. Suppose there exists one deterministic decoder d satisfying

$$d(z) = h(x), \qquad \forall x \in \Omega, \quad \forall z \in \mathcal{E}_U^{\mathcal{G}}(x).$$

Take any $x_a, x_b \in \Omega$ such that

$$z \in \mathcal{E}_U^{\mathcal{G}}(x_a) \cap \mathcal{E}_U^{\mathcal{G}}(x_b).$$

Then

$$h(x_a) = d(z) = h(x_b),$$

which establishes Eq. (32).

Sufficiency. Assume Eq. (32). Define

$$\mathcal{Z}_\Omega = \bigcup_{x \in \Omega} \mathcal{E}_U^{\mathcal{G}}(x).$$

For $z \in \mathcal{Z}_\Omega$, choose any $x \in \Omega$ satisfying

$$z \in \mathcal{E}_U^{\mathcal{G}}(x)$$

and define

$$d(z) = h(x).$$

This definition is well posed. If the same z belongs to the effective interfaces of both x_a and x_b, then Eq. (32) gives

$$h(x_a) = h(x_b).$$

Extend d arbitrarily from $\mathcal{Z}_\Omega$ to the remainder of $\mathcal{Z}$. The resulting decoder satisfies

$$d(z) = h(x), \qquad \forall x \in \Omega, \quad \forall z \in \mathcal{E}_U^{\mathcal{G}}(x).$$

Therefore exact reuse holds on a full-measure support.

Measurability qualification. The argument above establishes existence of an unrestricted deterministic decoder as a set-theoretic exact-decoding statement. The overlap condition alone does not, under arbitrary measurable-space structures, guarantee that the induced decoder or an arbitrary extension is measurable.

Accordingly, Theorem 1 is interpreted first as a deterministic exact-decoding criterion on the relevant full-measure support. Whenever the result is translated into a zero population-risk statement involving an expectation, the induced decoder is additionally required to be measurable. This regularity condition is separate from restriction to a particular parametric decoder class.

A.3 Accessibility and Coordinate Consistency

Recall

$$\mathcal{A}_{U,h}(\Omega) = \left\{(x_a, x_b) \in \Omega^2 : h(x_a) \neq h(x_b),\ U(x_a) = U(x_b)\right\},$$

and

$$\mathcal{K}^{\mathcal{G}}_{U,h}(\Omega) = \left\{ (x_a, x_b) \in \Omega^2 : \begin{array}{l} h(x_a) \neq h(x_b), \\ U(x_a) \neq U(x_b), \\ \mathcal{E}^{\mathcal{G}}_U(x_a) \cap \mathcal{E}^{\mathcal{G}}_U(x_b) \neq \varnothing \end{array} \right\}.$$

Consider any target-inconsistent pair whose effective interfaces overlap. If

$$U(x_a) = U(x_b),$$

the pair belongs to $\mathcal{A}_{U,h}(\Omega)$. Otherwise it belongs to $\mathcal{K}^{\mathcal{G}}_{U,h}(\Omega)$.

Conversely, because $\mathrm{id} \in \mathcal{G}(x)$,

$$U(x_a) = U(x_b) \quad \Longrightarrow \quad \mathcal{E}^{\mathcal{G}}_U(x_a) \cap \mathcal{E}^{\mathcal{G}}_U(x_b) \neq \varnothing.$$

Hence

$$\left\{ (x_a, x_b) \in \Omega^2 : \begin{array}{l} h(x_a) \neq h(x_b), \\ \mathcal{E}^{\mathcal{G}}_U(x_a) \cap \mathcal{E}^{\mathcal{G}}_U(x_b) \neq \varnothing \end{array} \right\} = \mathcal{A}_{U,h}(\Omega) \,\dot\cup\, \mathcal{K}^{\mathcal{G}}_{U,h}(\Omega).$$

Theorem 1 is therefore equivalent to requiring both conflict sets to be empty on the relevant full-measure support. Structural accessibility asks whether a target-relevant distinction has already been collapsed by the representation. Coordinate consistency asks whether distinct retained representations remain jointly interpretable by one shared decoder over the admissible interface.

A.4 Task-Family Reuse and Recovery of the Task-Relevant State

Let $\mathcal{H}$ be finite and suppose all tasks are evaluated under the same target distribution P_t.

If every task $h \in \mathcal{H}$ admits a full-measure support Ω_h on which exact reuse holds, then

$$\Omega_{\mathcal{H}} = \bigcap_{h \in \mathcal{H}} \Omega_h$$

also satisfies

$$P_t(\Omega_{\mathcal{H}}) = 1.$$

First consider the identity interface. If there exists

$$\rho : \mathcal{Z} \to \mathrm{range}(s^\star_{\mathcal{H}})$$

such that

$$\rho(U(x)) = s^\star_{\mathcal{H}}(x)$$

on $\Omega_{\mathcal{H}}$, then for each $h \in \mathcal{H}$,

$$h = \phi_h \circ s^\star_{\mathcal{H}},$$

and therefore

$$d_h = \phi_h \circ \rho$$

exactly reuses the representation.

Conversely, suppose every task in $\mathcal{H}$ is exactly reusable. If

$$U(x) = U(x'),$$

then every task-specific decoder must return the same value on x and x'. Thus

$$h(x) = h(x') \qquad \forall h \in \mathcal{H},$$

so

$$x \sim_{\mathcal{H}} x'.$$

Hence $s^{\star}_{\mathcal{H}}(x)$ is determined by $U(x)$, and a shared state readout exists on the common support.

The same argument applies to the effective interface. Exact reuse of every task in the finite family is equivalent to the existence of one state readout satisfying

$$\rho(z) = s^{\star}_{\mathcal{H}}(x), \qquad \forall x \in \Omega_{\mathcal{H}}, \quad \forall z \in \mathcal{E}^{\mathcal{G}}_{U}(x).$$

The quantifier structure is

$$\forall h \in \mathcal{H},\ \exists d_h,$$

rather than the existence of one decoder implementing all distinct tasks simultaneously.

When these deterministic constructions are interpreted through population risks, the same measurability convention as in Appendix A.2 applies.

A.5 Proof of Proposition 1

Let

$$\tau : \mathcal{Z} \to \mathcal{Z}$$

be a globally shared bijection and define

$$\widetilde{U} = \tau \circ U.$$

For each x, define

$$\widetilde{\mathcal{G}}(x) = \left\{\tau \circ g \circ \tau^{-1} : g \in \mathcal{G}(x)\right\}.$$

Then

$$\begin{aligned}
\mathcal{E}^{\widetilde{\mathcal{G}}}_{\widetilde{U}}(x) &= \left\{(\tau \circ g \circ \tau^{-1})\big(\tau(U(x))\big) : g \in \mathcal{G}(x)\right\} \\
&= \{\tau(g(U(x))) : g \in \mathcal{G}(x)\} \\
&= \tau\left(\mathcal{E}^{\mathcal{G}}_{U}(x)\right).
\end{aligned}$$

Suppose d exactly decodes the original effective interface. Define

$$\widetilde{d} = d \circ \tau^{-1}.$$

For any

$$\widetilde{z} \in \mathcal{E}^{\widetilde{\mathcal{G}}}_{\widetilde{U}}(x),$$

there exists $z \in \mathcal{E}^{\mathcal{G}}_{U}(x)$ with $\widetilde{z} = \tau(z)$. Thus

$$\widetilde{d}(\widetilde{z}) = d(z) = h(x).$$

The converse follows by applying the same argument with τ^{-1}. Therefore a globally shared bijective reparameterization preserves exact reuse when the admissible ambiguity family is transformed consistently.

Restricted decoder classes. For a restricted decoder class $\mathcal{D}$, define

$$\widetilde{\mathcal{D}} = \left\{ d \circ \tau^{-1} : d \in \mathcal{D} \right\}.$$

For the same target distribution and loss,

$$\mathcal{R}_{\text{reuse}}\left(\widetilde{U}; h, \widetilde{\mathcal{D}}\right) = \mathcal{R}_{\text{reuse}}\left(U; h, \mathcal{D}\right).$$

Thus restricted reuse is invariant when the decoder class is transported together with the coordinate change.

If the same fixed decoder class is used before and after reparameterization, closure under the corresponding forward and inverse compositions is a sufficient condition for the same invariance. It is not asserted to be necessary for a particular representation–task pair.

When these identities are interpreted as population-risk statements, τ, τ^{-1}, and the corresponding transported decoders are additionally required to be measurable so that the induced random variables and expectations are well defined. This is the same population-risk regularity convention used in Appendix A.2.

A.6 Joint Effective Interfaces

For an m-input downstream task, let

$$x_{1:m} = (x_1, \ldots, x_m) \in \mathcal{X}^m$$

and let $P_t^{(m)}$ be the target distribution on $\mathcal{X}^m$. No product or independence assumption is required.

Define

$$U^{(m)}(x_{1:m}) = (U(x_1), \ldots, U(x_m)).$$

Let $\mathcal{G}^{(m)}(x_{1:m})$ contain componentwise transformations

$$g_{1:m}(z_1, \ldots, z_m) = (g_1(z_1), \ldots, g_m(z_m)), \qquad g_j \in \mathcal{G}(x_j).$$

Then

$$\mathcal{E}^{\mathcal{G}^{(m)}}_{U^{(m)}}(x_{1:m}) = \left\{ g_{1:m}\big(U(x_1), \ldots, U(x_m)\big) : g_{1:m} \in \mathcal{G}^{(m)}(x_{1:m}) \right\}.$$

Because the transformation family acts componentwise,

$$\mathcal{E}^{\mathcal{G}^{(m)}}_{U^{(m)}}(x_{1:m}) = \prod_{j=1}^{m} \mathcal{E}^{\mathcal{G}}_{U}(x_j).$$

For a candidate state

$$s : \mathcal{X} \to \mathcal{S},$$

the corresponding tuple state is

$$s^{(m)}(x_{1:m}) = (s(x_1), \ldots, s(x_m)).$$

For $\Phi_m \subseteq \{\phi : \mathcal{S}^m \to \mathcal{Y}_t\}$, the induced family is

$$\mathcal{H}_S^{(m)} = \left\{ h_\phi : h_\phi(x_{1:m}) = \phi(s^{(m)}(x_{1:m})), \quad \phi \in \Phi_m \right\}.$$

A.7 Proof of Proposition 2

Suppose there exist $\Omega_t \subseteq \mathcal{X}$ with

$$P_t^{(m)}(\Omega_t^m) = 1$$

and a readout

$$\rho : \mathcal{Z} \to \mathcal{S}$$

such that

$$\rho(z) = s(x), \qquad \forall x \in \Omega_t, \quad \forall z \in \mathcal{E}^{\mathcal{G}}_U(x).$$

Take any $h_\phi \in \mathcal{H}_S^{(m)}$ and define

$$d_\phi(z_1, \ldots, z_m) = \phi(\rho(z_1), \ldots, \rho(z_m)).$$

For any

$$x_{1:m} \in \Omega_t^m$$

and

$$z_{1:m} \in \mathcal{E}_{U^{(m)}}^{\mathcal{G}^{(m)}}(x_{1:m}),$$

the componentwise product identity gives

$$z_j \in \mathcal{E}_U^{\mathcal{G}}(x_j), \qquad j = 1, \ldots, m.$$

Hence

$$\rho(z_j) = s(x_j)$$

for every component and therefore

$$d_\phi(z_{1:m}) = \phi(s(x_1), \ldots, s(x_m)) = h_\phi(x_{1:m}).$$

Thus every task in $\mathcal{H}_S^{(m)}$ is exactly reusable over the joint effective interface.

The proposition assumes candidate-state recovery on the target effective interface. The SPR objective in Eq. (21) promotes recoverability during source learning but does not itself imply this target-domain condition.

When the deterministic decoder above is interpreted through population risk, the measurability convention in Appendix A.2 applies.

A.8 Proof of Proposition 3: Restricted-Reuse Sandwich

Let

$$X_{1:m} = (X_1, \ldots, X_m) \sim P_t^{(m)}$$

and fix $a \in \mathcal{Q}_s$. Write

$$Y = h_\phi(X_{1:m})$$

and

$$G = G_{\phi,a} = \phi\left(a(U(X_1)), \ldots, a(U(X_m))\right).$$

For every fixed $d \in \mathcal{D}_m$,

$$\left|\mathbf{1}\{d(U^{(m)}(X_{1:m})) \neq Y\} - \mathbf{1}\{d(U^{(m)}(X_{1:m})) \neq G\}\right| \leq \mathbf{1}\{Y \neq G\}.$$

Taking expectations gives

$$|R_Y(d) - R_G(d)| \leq \Pr[Y \neq G]. \tag{33}$$

By definition,

$$\alpha_\phi = \inf_{d \in \mathcal{D}_m} R_Y(d), \qquad \eta_\phi(a) = \inf_{d \in \mathcal{D}_m} R_G(d).$$

A uniform inequality $|R_Y(d) - R_G(d)| \leq c$ for every d implies

$$\left|\inf_d R_Y(d) - \inf_d R_G(d)\right| \leq c.$$

Therefore

$$|\alpha_\phi - \eta_\phi(a)| \leq \Pr[G_{\phi,a} \neq h_\phi(X_{1:m})]. \tag{34}$$

Define

$$\widehat{S}_j = a(U(X_j)), \qquad S_j = s(X_j).$$

If all $\widehat{S}_j = S_j$, then

$$G_{\phi,a} = h_\phi(X_{1:m}).$$

Thus

$$\{G_{\phi,a} \neq h_\phi(X_{1:m})\} \subseteq \bigcup_{j=1}^{m} \{\widehat{S}_j \neq S_j\}.$$

By the union bound,

$$\Pr[G_{\phi,a} \neq h_\phi(X_{1:m})] \leq \sum_{j=1}^{m} \Pr[a(U(X_j)) \neq s(X_j)] \tag{35}$$

$$= m\,\epsilon_s(a). \tag{36}$$

No independence among the components of $X_{1:m}$ is required.

Combining Eqs. (34) and (36) gives

$$\eta_\phi(a) - m\epsilon_s(a) \leq \alpha_\phi \leq \eta_\phi(a) + m\epsilon_s(a).$$

Since $\alpha_\phi \geq 0$,

$$[\eta_\phi(a) - m\epsilon_s(a)]_+ \leq \alpha_\phi.$$

Taking the supremum of the lower bounds and the infimum of the upper bounds over $a \in \mathcal{Q}_s$ gives

$$\sup_{a\in\mathcal{Q}_s} [\eta_\phi(a) - m\epsilon_s(a)]_+ \leq \alpha_\phi \leq \inf_{a\in\mathcal{Q}_s} [\eta_\phi(a) + m\epsilon_s(a)].$$

The same readout a determines both terms. Candidate-state recoverability and restricted composition therefore cannot in general be optimized independently, and the result does not imply a monotone ordering of representations by state-recovery error alone.

A.9 Finite-Sample Extension for a Finite Decoder Class

For $d \in \mathcal{D}_m$, define

$$\mathcal{R}_\phi(d) = \Pr[d(U^{(m)}(X_{1:m})) \neq h_\phi(X_{1:m})].$$

Then

$$\alpha_\phi = \inf_{d\in\mathcal{D}_m} \mathcal{R}_\phi(d).$$

Let

$$X_{1:m}^{(1)}, \ldots, X_{1:m}^{(n)} \overset{\text{i.i.d.}}{\sim} P_t^{(m)}$$

and define

$$\widehat{\mathcal{R}}_{\phi,n}(d) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{d(U^{(m)}(X_{1:m}^{(i)})) \neq h_\phi(X_{1:m}^{(i)})\right\}.$$

Let

$$\widehat{d}_n \in \arg\min_{d\in\mathcal{D}_m} \widehat{\mathcal{R}}_{\phi,n}(d).$$

If $\mathcal{D}_m$ is finite, Hoeffding's inequality and a union bound imply that with probability at least $1-\delta$,

$$\sup_{d\in\mathcal{D}_m} \left|\mathcal{R}_\phi(d) - \widehat{\mathcal{R}}_{\phi,n}(d)\right| \leq \Delta_n, \qquad \Delta_n = \sqrt{\frac{\log(2|\mathcal{D}_m|/\delta)}{2n}}. \tag{37}$$

On this event,

$$\mathcal{R}_\phi(\widehat{d}_n) \leq \alpha_\phi + 2\Delta_n.$$

Using the upper side of Proposition 3 gives

$$\mathcal{R}_\phi(\widehat{d}_n) \leq \inf_{a\in\mathcal{Q}_s} [\eta_\phi(a) + m\epsilon_s(a)] + \sqrt{\frac{2\log(2|\mathcal{D}_m|/\delta)}{n}}. \tag{38}$$

The three terms have distinct roles: candidate-state recovery, restricted composition, and finite-sample estimation. The bound does not imply that the difference between two learned representations must increase as n decreases.

The experimental target-supervision protocols use held-out patterns and do not directly instantiate the i.i.d. quantity n above. Eq. (38) should therefore be read as a conceptual finite-sample decomposition; it does not predict the supervision-localized effect in Section 3.3.

A.10 Better Candidate-State Recovery Need Not Improve Restricted Reuse

Let

$$S = (S_1, S_2) \sim \mathrm{Unif}(\{0,1\}^2)$$

and let the candidate state be

$$s(S) = S.$$

Consider the target

$$h(S) = S_1 \oplus S_2$$

and two representations

$$U_0(S) = (S_1 \oplus S_2, 0), \qquad U_1(S) = (S_1, S_2).$$

Define the optimal full-state recovery error

$$\epsilon_s^\star(U) = \inf_a \Pr[a(U(S)) \neq s(S)],$$

where the infimum ranges over deterministic state readouts.

The full state is exactly recoverable from U_1, so

$$\epsilon_s^\star(U_1) = 0.$$

From U_0, only parity is available. Conditional on either parity value, two full candidate states are equally possible, so

$$\epsilon_s^\star(U_0) = \frac{1}{2}.$$

Now restrict downstream decoders to affine threshold functions

$$\mathcal{D}_{\mathrm{aff}} = \left\{ z \mapsto \mathbf{1}\{w^\top z + b \geq 0\} : w \in \mathbb{R}^2,\ b \in \mathbb{R} \right\}.$$

For U_0, the target is the first representation coordinate, so

$$\mathcal{R}_{\mathrm{reuse}}(U_0; h, \mathcal{D}_{\mathrm{aff}}) = 0.$$

For U_1, XOR on the Boolean square is not linearly separable. The minimum affine-threshold error is

$$\mathcal{R}_{\mathrm{reuse}}(U_1; h, \mathcal{D}_{\mathrm{aff}}) = \frac{1}{4}.$$

Consequently,

$$\epsilon_s^\star(U_1) < \epsilon_s^\star(U_0),$$

while

$$\mathcal{R}_{\mathrm{reuse}}(U_1; h, \mathcal{D}_{\mathrm{aff}}) > \mathcal{R}_{\mathrm{reuse}}(U_0; h, \mathcal{D}_{\mathrm{aff}}).$$

For the singleton family $\mathcal{H} = \{h\}$, the minimal task-relevant state is, up to relabeling, parity itself. Thus U_0 preserves exactly the distinction needed for the task, whereas U_1 preserves a larger candidate state whose required composition is not realizable by the restricted decoder class.

The counterexample concerns the interaction between candidate-state recovery and decoder-class restriction. Here U_1 recovers the full candidate state exactly, yet the required XOR composition is not realizable by $\mathcal{D}_{\mathrm{aff}}$. Candidate-state recovery alone therefore does not order representations by restricted reuse.

B Experimental Setting and Evaluation Protocols

This section specifies the experimental conditions, frozen-representation evaluation protocols, admissibility rules, diagnostic measurements, and statistical estimands used in Section 3.

Table 1: Source-learning conditions. C2 and C3 are structural controls.

ID	Name	Definition
C0	Source-only	Optimizes the source-task objective without structural supervision.
C1	SPR	Adds the structure-preserving objective in Eq. (21) using candidate-state labels.
C2	Random auxiliary	Uses random three-state auxiliary targets in place of the candidate state.
C3	Shuffled-state	Permutes candidate-state targets across observations, breaking their observation–state correspondence.
C4	Globally relabeled SPR	Uses the same structural supervision as C1 after one fixed, globally shared bijective relabeling of candidate-state labels.

B.1 Experimental Conditions

The experiments use five source-learning conditions.

C4 changes only the symbolic state naming,

$$s(x) \longmapsto \pi(s(x)), \qquad \pi : \mathcal{S} \to \mathcal{S}$$

with π bijective. It does not transform representation coordinates and is therefore distinct from Proposition 1 and from the coordinate intervention below.

B.2 Candidate State and Held-Out Task Construction

In the controlled relational setting, the candidate state has three values,

$$\mathcal{S} = \{\text{MATCH}, \text{MISMATCH}, \text{UNKNOWN}\},$$

abbreviated as

$$M, \qquad D, \qquad U.$$

For two inputs,

$$s^{(2)}(x_1, x_2) = (s(x_1), s(x_2)) \in \mathcal{S}^2.$$

Ordered configurations that differ only by permutation are grouped into six composition families,

$$MM, \quad MD, \quad MU, \quad DD, \quad DU, \quad UU.$$

For example, MD contains both (M, D) and (D, M). The downstream rules used here are symmetric, so this grouping does not alter their target values.

The four replacement rules are

ANY_MATCH, ANY_MISMATCH, ANY_UNKNOWN, HOMOGENEOUS.

For $s_1, s_2 \in \mathcal{S}$,

$$\begin{aligned}
\phi_{\text{AM}}(s_1, s_2) &= \mathbf{1}\{M \in \{s_1, s_2\}\}, \\
\phi_{\text{AD}}(s_1, s_2) &= \mathbf{1}\{D \in \{s_1, s_2\}\}, \\
\phi_{\text{AU}}(s_1, s_2) &= \mathbf{1}\{U \in \{s_1, s_2\}\}, \\
\phi_{\text{H}}(s_1, s_2) &= \mathbf{1}\{s_1 = s_2\}.
\end{aligned}$$

Thus the candidate-state mapping remains fixed while the downstream composition rule changes, directly instantiating $\mathcal{H}_S^{(2)}$.

Leave-one-composition-out construction. Each fold withholds one composition family from source learning and downstream readout adaptation. The six folds respectively hold out

$$MM, \quad MD, \quad MU, \quad DD, \quad DU, \quad UU.$$

For the representation-independent oracle audit, a state pair $(z_1, z_2) \in \mathcal{S}^2$ is represented by the six-dimensional semantic interface

$$\psi(z_1, z_2) = \text{concat}\left(e(z_1) + e(z_2),\ e(z_1) \odot e(z_2)\right), \tag{39}$$

where $e(z) \in \{0, 1\}^3$ is the one-hot encoding of the candidate state.

Oracle ridge readout. For one replacement rule and one held-out composition, let

$$\{(\psi_i, y_i)\}_{i=1}^{n}, \qquad n = 5,$$

denote the five non-held-out oracle composition prototypes and their binary targets. Targets are encoded as

$$t_i = 2y_i - 1 \in \{-1, +1\}.$$

The oracle evaluator fits

$$(\widehat{w}, \widehat{b}) = \arg\min_{w,b} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(t_i - w^\top \psi_i - b\right)^2 + \lambda_{\text{ridge}} \|w\|_2^2 \right\}, \qquad \lambda_{\text{ridge}} = 10^{-3}. \tag{40}$$

The intercept $\widehat{b}$ is not regularized. The semantic feature vectors in Eq. (39) are used directly without per-coordinate centering or standardization; "normalized" refers to the $1/n$ normalization of the squared-error objective, equivalently of the Gram and right-hand-side terms. For the held-out prototype ψ_*, the binary prediction is

$$\widehat{y}_* = \mathbf{1}\left\{ \widehat{w}^\top \psi_* + \widehat{b} \geq 0 \right\}. \tag{41}$$

This leave-one-composition-out calculation determines whether a family–rule cell lies within the extrapolation capability of the specified downstream readout when the true candidate states are available. It is therefore a property of the oracle state interface and fixed readout protocol, not of any learned representation.

Crossing six held-out composition families with four replacement rules gives

$$6 \times 4 = 24$$

candidate cells. The oracle-solvability criterion retains

$$21$$

of them.

This $24 \to 21$ restriction is not a generic two-class-support filter. The target on an individual held-out composition prototype may be constant. The relevant question is whether a readout fitted on the five non-held-out oracle prototypes extrapolates the correct target to the held-out prototype.

B.3 Frozen-Representation Reuse Protocol

All reuse experiments separate source learning from downstream adaptation. After source learning,

1. the source-task head is discarded;
2. the encoder is frozen;
3. replacement-task gradients do not update the encoder; and
4. a new downstream rule is fitted using target-task supervision.

For the primary held-out-family analysis, five independently trained representations are compared under matched source-training seeds.

Target supervision varies over

$$K \in \{1, 2, 4, 8, 16\}$$

examples per state together with a full-support condition. Adaptation measurements are paired across compared representation conditions according to the pre-specified evaluator.

The auxiliary structural readout used during source learning is discarded and is never supplied to the downstream reuse evaluator.

Table 2: Leave-one-composition-out oracle disposition of the 24 candidate family–rule cells. "Included" means that the fixed downstream readout extrapolates the replacement rule correctly from the five non-held-out oracle composition prototypes to the held-out prototype.

Held-out family	ANY_MATCH	ANY_MISMATCH	ANY_UNKNOWN	HOMOGENEOUS
MM	Included	Included	Included	Excluded
MD	Included	Included	Included	Included
MU	Included	Included	Included	Included
DD	Included	Included	Included	Excluded
DU	Included	Included	Included	Included
UU	Included	Included	Included	Excluded

B.4 Source-Matched Novel-Rule Evaluation

The Shapes3D experiment in Section 3.1 tests whether identical source-task performance determines reuse.

The pre-specified candidate registry contains

$$2032$$

rule–evaluation cells. A representation-independent audit partitions these into

$$1008 \text{ oracle-solvable}, \qquad 1008 \text{ oracle-unsolvable}, \qquad 16 \text{ single-class}$$

cells. Novelty and admissibility screening retains

$$1004$$

of the oracle-solvable cells for the source-matched analysis.

All five source-training seeds under C0, C1, and C4 attain source validation accuracy

$$1.0000,$$

yielding 15 exactly source-matched models.

Registry inclusion is determined independently of the measured reuse performance of the learned representations.

B.5 Representation-Independent Admissibility Criteria

Admissibility is determined without using the performance of C0, C1, C4, or any other learned representation.

Source-matched novel-rule protocol. Candidate cells are screened by the representation-independent oracle, novelty, and registry criteria described above. The resulting primary registry contains

$$1004$$

cells.

Held-out candidate-state protocol. For each of the 24 family–rule cells, candidate states are converted to the semantic interface in Eq. (39). The oracle ridge readout in Eq. (40) is fitted using the five non-held-out composition prototypes and evaluated on the sixth using Eq. (41). A cell is designated *oracle-solvable* when this prediction equals the registered target on the held-out composition.

The resulting disposition is:

The three excluded cells are

$$MM \times \text{HOMOGENEOUS}, \qquad DD \times \text{HOMOGENEOUS}, \qquad UU \times \text{HOMOGENEOUS}.$$

For each of these cells, the held-out target is 1, whereas the leave-one-composition-out oracle readout predicts 0.

When all six oracle composition prototypes are simultaneously available, all four replacement rules are realizable by the same oracle interface and readout family. The three exclusions are therefore specific to leave-one-composition-out readout extrapolation. They are neither single-class exclusions nor evidence that a learned representation has lost the corresponding candidate state.

The same 21-cell registry is used for every source-learning condition, representation seed, target-supervision level, and paired comparison.

B.6 Aligned State-Recoverability Diagnostic

Section 3.2 uses state macro-F1 as a diagnostic of candidate-state recoverability from the frozen representation.

Let Π_3 denote the 3! permutations of the three state labels. If $\widehat{s}(x)$ is a diagnostic state prediction, define

$$\mathrm{F1}_{\mathrm{aligned}} = \max_{\pi \in \Pi_3} \mathrm{MacroF1}\left(\pi(\widehat{s}(x)), s(x)\right).$$

The global permutation removes arbitrary naming differences among the three states. It does not modify the frozen representation and is not supplied to the downstream reuse evaluator. The metric is therefore a diagnostic of whether the three-state structure is recoverable up to one global relabeling, not the reuse score itself.

B.7 Coordinate-Intervention and Rescue Protocol

Coordinate interventions are applied after representation learning to the same frozen representation. Each transformation is bijective, so the original representation remains exactly recoverable when the applicable convention is known.

The intervention conditions are:

- **Identity**: no coordinate transformation;
- **Global**: one reversible transformation shared by all samples;
- **2-way**: samples use one of two reversible conventions;
- **4-way**: samples use one of four conventions;
- **8-way**: samples use one of eight conventions;
- **Sample-wise**: convention identity varies at the sample level.

The blind downstream evaluator receives the transformed representation but not the applicable convention identity. The multi-convention conditions therefore increase the heterogeneity of sample-dependent coordinate conventions while withholding the information needed to invert them.

These finite-supervision interventions provide an empirical diagnostic of shared interpretability. They do not correspond one-to-one to the exact coordinate-conflict set $\mathcal{K}^{\mathcal{G}}_{U,h}$.

For the rescue condition, the convention identity is supplied and the exact inverse transformation is applied before downstream decoding. The reconstruction audit gives

$$\max_i \|\widehat{z}_i - z_i\| = 0,$$

and the convention-aware evaluator returns to Identity-level reuse.

B.8 Prospective Conditional-Accessibility Protocol

A separate CIFAR-100 protocol varies task-relevant conditional distinctions during source learning and evaluates frozen-representation reuse across the resulting intervention levels.

Let C denote the CIFAR-100 coarse class and F the corresponding fine class. Each coarse class contains five fine classes. For intervention level

$$q \in \{1, 2, 3, 4, 5\},$$

the source-learning protocol uses a deterministic auxiliary target

$$A_q = m_q(C, F). \tag{42}$$

The mappings m_q are fixed before representation training and remain the same across representation seeds and across the two tested encoder architectures.

Five-state partition construction. Within each coarse class, the five fine states are grouped by the auxiliary mapping. Under the balanced five-state calibration distribution, the five intervention levels have the following block-size profiles:

$$\begin{array}{lcl} Q1 & : & (5), \\ Q2 & : & (3,2), \\ Q3 & : & (2,2,1), \\ Q4 & : & (2,1,1,1), \\ Q5 & : & (1,1,1,1,1). \end{array}$$

Thus Q1 assigns all five fine states to one auxiliary value, whereas Q5 distinguishes all five. The intermediate conditions have progressively lower calibrated residual fine-state uncertainty, but this block-size ordering alone is not used to assert a nested-refinement relation between their identity-level partitions.

For each coarse class c, let

$$\mathcal{F}_c = \{f_{c,1}, \ldots, f_{c,5}\}$$

denote its five fine classes. The actual experimental auxiliary target is determined by the fixed lookup

$$m_q(c, f_{c,j}), \qquad q = 1, \ldots, 5, \quad j = 1, \ldots, 5.$$

The identity-level assignments are frozen before representation training and are identical across representation seeds and encoder architectures. The anonymous protocol configuration records these lookup tables explicitly. The information-theoretic calibration below depends on the induced partition sizes; the learned-representation analysis uses the actual frozen lookup rather than only the block-size profile.

Information-theoretic calibration. Let

$$H_q = H_{\mathrm{cal}}\left(F \mid C, A_q\right)$$

denote the conditional entropy under the fixed balanced calibration distribution. For a partition $\mathcal{P}_q$ of the five fine states,

$$H_q = \frac{1}{5} \sum_{B \in \mathcal{P}_q} |B| \log_2 |B|.$$

The Q1 baseline is

$$H_1 = \log_2 5 = 2.3219280949 \quad \text{bits}.$$

We define the normalized conditional-accessibility coordinate

$$a_q = 1 - \frac{H_{\mathrm{cal}}(F \mid C, A_q)}{H_1}. \tag{43}$$

The resulting intervention levels are:

The intervention level is therefore a property of the predefined auxiliary target, not a post-hoc estimate of how much fine-state information the learned representation actually retains.

Where the intervention acts. The variable A_q is used only as an auxiliary source-training target. It is not concatenated to the frozen representation and is not supplied to the downstream readout.

For a fixed architecture and matched seed identifier, Q1–Q5 correspond to five separately trained representations,

$$U_{s,1}, \ldots, U_{s,5}.$$

Table 3: Conditional-accessibility calibration. The quantities are determined by the predefined auxiliary-target construction, not by learned representation performance.

Level	$H_{\mathrm{cal}}(F \mid C, A_q)$	a_q	Oracle fine-state Bayes accuracy
Q1	2.3219281	0.0000000	0.2
Q2	1.3509775	0.4181657	0.4
Q3	0.8000000	0.6554588	0.6
Q4	0.4000000	0.8277294	0.8
Q5	0.0000000	1.0000000	1.0

The five runs use the same source-task protocol but differ in the predefined auxiliary target. After source learning, the auxiliary head is discarded, the encoder is frozen, and held-out future-rule reuse is evaluated under a common downstream protocol.

The Q1–Q5 curve therefore compares representations produced by different predefined source-learning interventions. It is not obtained by progressively revealing side information to a single frozen encoder.

Endpoint interpretation. The value

$$a_1 = 0$$

means that Q1 provides no reduction in the calibrated residual uncertainty about F relative to the Q1 baseline. It does not mean that the learned representation contains no information useful for downstream tasks.

Similarly,

$$a_5 = 1$$

means that the Q5 auxiliary construction fully resolves the controlled fine-state component under the calibration distribution. It does not imply that every downstream distinction is perfectly represented.

Source matching and replication. For each architecture, source-task matching is evaluated separately for

$$5 \text{ matched seed blocks} \times 5 \text{ Q levels} = 25$$

trained representations. The accessibility construction, future-task protocol, data split, and source-matching rule remain fixed across Q levels.

The empirical claim is restricted to this predefined source-learning intervention: under source matching, condition-level mean held-out reuse increases across Q1–Q5 for both tested encoder architectures. The experiment does not establish universal monotonicity for arbitrary representations, auxiliary objectives, task families, or decoder classes.

B.9 Statistical Estimands and Replication Units

Primary held-out-family analysis. Let $A_{s,C}(K)$ denote aggregate reuse accuracy for source-training seed s, condition C, and target-supervision level K, evaluated over the fixed 21-cell registry.

Define

$$A^{\mathrm{LS}}_{s,C} = \frac{1}{5} \sum_{K \in \{1,2,4,8,16\}} A_{s,C}(K).$$

For a comparison condition C against Source-only C0, define

$$\delta^{\mathrm{LS}}_{s,C} = A^{\mathrm{LS}}_{s,C} - A^{\mathrm{LS}}_{s,\mathrm{C0}},$$

and

$$\Delta_{\mathrm{LS}}(C) = \frac{1}{S} \sum_{s=1}^{S} \delta^{\mathrm{LS}}_{s,C}, \qquad S = 5.$$

Analogously,

$$\delta^{\mathrm{full}}_{s,C} = A_{s,C}(\mathrm{Full}) - A_{s,\mathrm{C0}}(\mathrm{Full}),$$

and

$$\Delta_{\text{full}}(C) = \frac{1}{S}\sum_{s=1}^{S} \delta_{s,C}^{\text{full}}.$$

The supervision-localized contrast is

$$\Gamma(C) = \Delta_{\text{LS}}(C) - \Delta_{\text{full}}(C).$$

The independently trained source representation is the primary replication unit. Repeated task cells, adaptation draws, and supervision levels contribute to a seed-level aggregate and are not counted as independent representation replicates.

Confidence intervals for paired representation comparisons resample at the representation-seed level, preserving all within-seed task, adaptation, and supervision measurements as one block.

Conditional-accessibility analysis. For architecture A, matched seed block s, and intervention level q, let

$$R_{A,s,q}$$

denote the frozen held-out reuse aggregate of the separately trained Q-specific representation.

The condition-level mean is

$$\overline{R}_{A,q} = \frac{1}{5}\sum_{s=1}^{5} R_{A,s,q}.$$

The endpoint contrast is

$$\Delta_{5-1}(A) = \overline{R}_{A,5} - \overline{R}_{A,1},$$

computed before display rounding.

For each matched seed block, the accessibility–reuse trend is summarized by the ordinary least-squares slope

$$\widehat{\beta}_{A,s} = \frac{\sum_{q=1}^{5}(a_q - \overline{a})(R_{A,s,q} - \overline{R}_{A,s})}{\sum_{q=1}^{5}(a_q - \overline{a})^2}, \tag{44}$$

where

$$\overline{a} = \frac{1}{5}\sum_{q=1}^{5} a_q, \qquad \overline{R}_{A,s} = \frac{1}{5}\sum_{q=1}^{5} R_{A,s,q}.$$

The architecture-level slope is

$$\overline{\beta}_A = \frac{1}{5}\sum_{s=1}^{5} \widehat{\beta}_{A,s}.$$

The five Q-specific models belonging to one seed identifier form one matched block. They are separately trained representations, but their Q1–Q5 trajectory is kept together for inference.

For the reported bootstrap interval, complete seed blocks are resampled with replacement. With five seed blocks, exhaustive ordered resampling contains

$$5^5 = 3125$$

bootstrap samples. For each resample, the architecture-level statistic is the mean of the selected seed-block slopes. The reported 95% interval is the equal-tailed percentile interval of this exhaustively enumerated empirical bootstrap distribution. Here "exact bootstrap" refers to exact enumeration of the bootstrap resampling space, not to exact finite-sample confidence coverage.

Two empirical statements are kept separate:

1. the five condition-level means $\overline{R}_{A,1}, \ldots, \overline{R}_{A,5}$ are strictly ordered; and
2. all five seed-block slopes $\widehat{\beta}_{A,s}$ are positive.

A positive seed-block slope does not imply that every adjacent Q transition is positive within that block.

B.10 Relationship Between Experimental Protocols

The experiments address different parts of the claim–evidence chain, and their absolute accuracies are not treated as directly interchangeable.

The Shapes3D source-matched novel-rule registry tests the empirical separation

$$\text{source-task success} \not\Rightarrow \text{representation reuse}.$$

The structural intervention tests whether preserving a candidate state is accompanied by increased candidate-state recoverability and reuse on tasks using that state.

The coordinate intervention instead manipulates the availability of a shared coordinate convention while keeping the original frozen representation recoverable through inversion.

The 21-cell held-out-family protocol tests the constructive case in which anticipated tasks factor through the candidate state. It is therefore the empirical counterpart of the candidate-state construction in Section 2.4, not a claim that SPR improves every replacement-task family.

Finally, the CIFAR-100 conditional-accessibility experiment varies a predefined auxiliary source-learning intervention under source matching and tests whether the resulting intervention–reuse relationship is reproduced with two encoder architectures. The calibrated coordinate a_q characterizes the auxiliary source-learning condition and is not a direct measurement of structural accessibility in the trained representation.

Scientific comparisons are therefore made within protocol using paired or pre-specified contrasts; absolute accuracies from these distinct protocols are not compared as scientific contrasts.

C Supporting Experimental Results

C.1 Source-Matched Reuse Separation

In the Shapes3D protocol, all five seeds under C0, C1, and C4 achieve source validation accuracy

$$1.0000.$$

Across the 1004 admissible novel-rule evaluations, mean low-shot reuse is

$$0.6410 \quad \text{for C0},$$

$$0.6040 \quad \text{for C1},$$

and

$$0.6031 \quad \text{for C4}.$$

The paired low-shot differences relative to C0 are

$$-0.0370 \quad \text{for C1}, \qquad -0.0379 \quad \text{for C4},$$

with the same direction across all five paired source seeds.

Under full target support, the corresponding differences contract to

$$-0.0028 \qquad \text{and} \qquad -0.0040.$$

Thus identical source-task validation accuracy does not determine frozen-representation reuse, and the observed separation is substantially larger under limited target supervision.

Figure 5 exposes the corresponding seed-level replication structure and the heterogeneity across the fixed admissible novel-rule registry.

C.2 Candidate-State Recoverability and Structural Controls

Across five matched seeds, aligned held-out state macro-F1 changes from

$$0.5547 \pm 0.1807$$

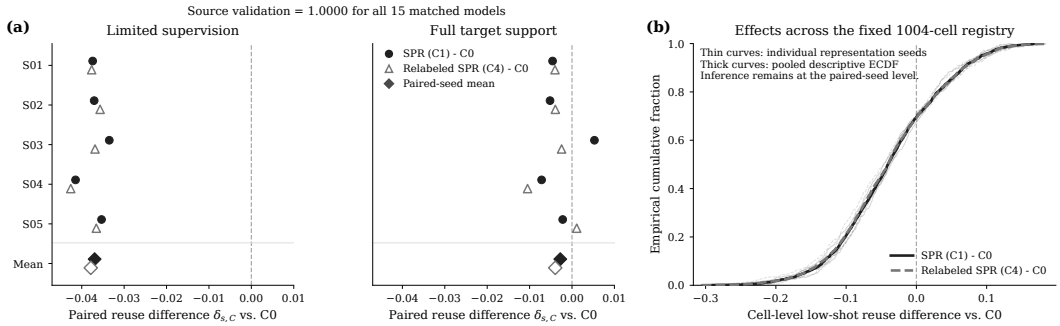

Figure 5: **Paired audit of source-matched reuse separation.** All C0, C1, and C4 representations achieve identical source-task validation accuracy (1.0000 across all 15 trained models). **(a)** Paired reuse differences relative to the corresponding C0 representation for each independently trained source seed. Under limited target supervision, C1 and C4 show the same direction of separation from C0 across all five seeds, whereas the differences contract substantially under full target support. **(b)** Descriptive distributions of cell-level low-shot paired effects across the fixed 1004-cell admissible novel-rule registry. Thin curves show individual representation seeds and thick curves show pooled descriptive empirical cumulative distribution functions. Registry cells expose heterogeneity across replacement tasks but are not treated as independent statistical replicates; inference is performed at the paired representation-seed level.

under C0 to

$$0.9646 \pm 0.0052$$

under C1. The paired change is

$$+0.4099, \qquad 95\%\ \mathrm{CI} = [0.2719,\ 0.5479].$$

Low-shot reuse changes from

$$0.6491 \pm 0.1130$$

to

$$0.8356 \pm 0.0274,$$

with paired difference

$$+0.1866, \qquad 95\%\ \mathrm{CI} = [0.1012,\ 0.2875].$$

The reuse contrast is positive for all five paired seeds.

C4 retains similar state recoverability:

$$0.9572 \pm 0.0109$$

versus

$$0.9646 \pm 0.0052$$

for C1. The direct C1–C4 reuse contrast is

$$-0.0072, \qquad 95\%\ \mathrm{CI} = [-0.0375,\ 0.0203].$$

Random auxiliary and Shuffled-state controls do not reproduce the C1 reuse level. The paired C1 advantages are

$$+0.5845$$

and

$$+0.6576,$$

respectively, with the same direction across all five matched seeds.

Within this protocol, task-aligned candidate-state preservation is associated with higher recoverability and reuse. This association does not establish a general monotone relationship between arbitrary approximate state recovery and restricted reuse.

Figure 6 summarizes the condition-level relationship between held-out candidate-state recoverability and low-shot frozen reuse, and shows the paired C1–C0 gains for each independently trained source seed.

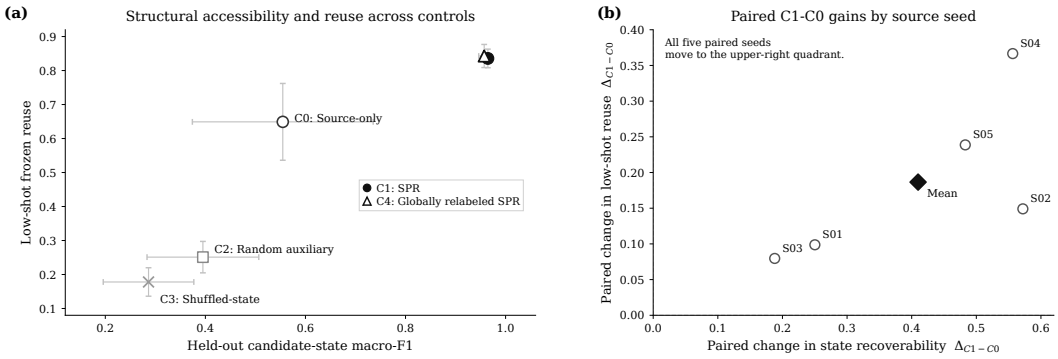

Figure 6: **Structural accessibility and low-shot reuse under matched controls. (a)** Condition-level relationship between held-out candidate-state recoverability and low-shot frozen reuse across Source-only (C0), SPR (C1), Random auxiliary (C2), Shuffled-state (C3), and Globally relabeled SPR (C4). Markers show condition means and error bars show standard deviations across five independently trained source seeds. Improving candidate-state recoverability is accompanied by improved low-shot reuse; auxiliary training without task-aligned state structure (C2) or with broken observation–state correspondence (C3) does not reproduce the effect, whereas a globally shared relabeling of the preserved state space (C4) remains close to C1. **(b)** Paired C1–C0 gains for each source seed in held-out candidate-state recoverability and low-shot reuse. All five paired seeds lie in the upper-right quadrant; both gains are positive for every source seed.

C.3 Coordinate-Convention Intervention and Rescue

The same frozen representation is evaluated after reversible coordinate interventions.

Mean low-shot reuse is

$$0.4096$$

under Identity and

$$0.4099$$

under one globally shared transformation. Under increasingly heterogeneous sample-dependent conventions, reuse is

$$0.3449, \quad 0.2679, \quad 0.1836, \quad 0.0154$$

for 2-way, 4-way, 8-way, and sample-wise conditions, respectively.

The ordering is reproduced for each of the three source-training seeds, and the Identity-to-sample-wise difference is

$$0.3942.$$

When convention identity is supplied and the corresponding inverse is applied, reuse returns to

$$0.4096$$

in every condition. The reconstruction audit gives

$$\max_i \|\widehat{z}_i - z_i\| = 0.$$

Within this finite-supervision intervention, the degradation is associated with unavailable sample-specific coordinate conventions; convention-aware inversion exactly recovers the common-coordinate representation. This intervention is not a direct numerical measurement of the exact coordinate-conflict set in Theorem 1.

Figure 7 summarizes the coordinate-consistency continuum and the corresponding gauge-aware rescue, and exposes the blind-decoder penalty separately for each independently trained source representation.

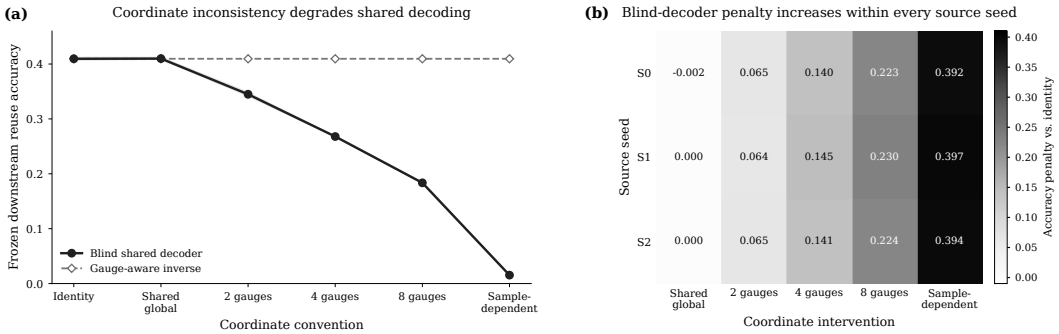

Figure 7: **Coordinate inconsistency degrades shared decoding and is exactly rescued by convention-aware inversion.** **(a)** Frozen downstream reuse under increasingly heterogeneous reversible coordinate conventions. A single globally shared reparameterization leaves blind reuse essentially unchanged relative to the identity convention, whereas reuse decreases progressively as the number of unresolved conventions increases, reaching its minimum under sample-dependent conventions. Supplying the applicable inverse convention restores the identity-level reuse accuracy for every intervention condition; the frozen reconstruction audit reports zero residual relative to the common-coordinate representation. **(b)** Blind-decoder accuracy penalty relative to the identity convention for each independently trained source representation. For every source seed, the penalty increases from the shared-global control through the 2-, 4-, and 8-convention interventions to the sample-dependent condition.

Table 4: Paired held-out-family reuse effects relative to Source-only. The inferential unit is the matched representation-training seed.

Condition	Δ_{LS}	Δ_{full}	Γ
SPR	$+0.1715$	$+0.0241$	$+0.1474$
95% CI	$[0.0986,\ 0.2591]$	$[-0.0260,\ 0.0992]$	$[0.1044,\ 0.1905]$
Globally relabeled SPR	$+0.1771$	$+0.0300$	$+0.1471$
95% CI	$[0.1172,\ 0.2593]$	$[-0.0201,\ 0.0990]$	$[0.1193,\ 0.1749]$

C.4 Reuse Across Held-Out Candidate-State Task Families

The primary compatible-family analysis uses the fixed 21-cell admissible registry and

$$K \in \{1, 2, 4, 8, 16\}$$

for the limited-supervision regime.

For C1 relative to C0,

$$\Delta_{\mathrm{LS}} = +0.1715, \qquad 95\%\ \mathrm{CI} = [0.0986,\ 0.2591],$$

whereas

$$\Delta_{\mathrm{full}} = +0.0241, \qquad 95\%\ \mathrm{CI} = [-0.0260,\ 0.0992].$$

Hence

$$\Gamma = +0.1474, \qquad 95\%\ \mathrm{CI} = [0.1044,\ 0.1905].$$

For C4 relative to C0,

$$\Delta_{\mathrm{LS}} = +0.1771, \qquad 95\%\ \mathrm{CI} = [0.1172,\ 0.2593],$$

$$\Delta_{\mathrm{full}} = +0.0300, \qquad 95\%\ \mathrm{CI} = [-0.0201,\ 0.0990],$$

and

$$\Gamma = +0.1471, \qquad 95\%\ \mathrm{CI} = [0.1193,\ 0.1749].$$

The direct C1–C4 limited-supervision contrast is

$$-0.0056, \qquad 95\%\ \mathrm{CI} = [-0.0284,\ 0.0173],$$

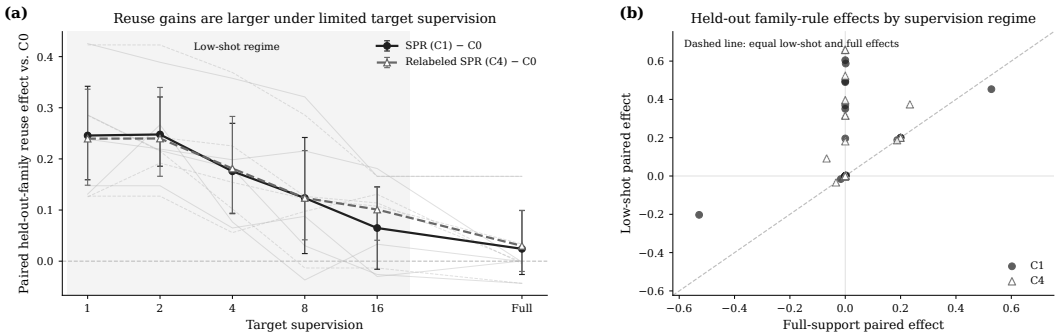

Figure 8: **Held-out-family reuse gains are concentrated under limited target supervision. (a)** Paired reuse effects relative to Source-only across target-supervision levels $K \in \{1, 2, 4, 8, 16\}$ and the full-support condition. Solid circles denote SPR (C1) and open triangles denote Globally relabeled SPR (C4); thick curves show paired means across the five representation-training seeds, error bars show bootstrap 95% confidence intervals, and faint curves show the corresponding seed-level paired effects. For both conditions, paired reuse gains are larger at each predefined limited-supervision level than under full target support. **(b)** Descriptive low-shot and full-support paired effects for the 21 admissible held-out family–rule evaluations. Points above the diagonal have larger low-shot than full-support effects. The cell-level distribution shows heterogeneity across replacement tasks and is descriptive; inferential comparisons remain defined at the paired representation-training-seed level.

with no consistent direction across seeds.

The positive Γ contrasts indicate larger reuse gains under limited target supervision. The full-support effects for C1 and C4 remain unresolved because both confidence intervals include zero.

Figure 8 reports the corresponding supervision profile and the variation of paired effects across the admissible held-out family–rule evaluations.

C.5 Evidence Beyond the Primary Controlled Setting

Natural-image source-matched evidence. Across nine matched representation pairs on CIFAR-100, the mean absolute within-pair source-performance gap is

$$0.0036,$$

while the low-shot SPR–Source-only reuse difference is

$$+0.1718.$$

Alternative source-learning formulation. Under ordinary supervised classification, source accuracy is

$$0.7535$$

for Source-only and

$$0.7602$$

for SPR. After freezing the representations, the held-out-rule reuse difference is

$$+0.0768$$

in favor of SPR.

The source-success/reuse separation is also observed beyond the primary relational construction. This evidence does not establish universal superiority of one source objective across arbitrary future-task families.

Table 5: Condition-level mean held-out reuse in the conditional-accessibility replications. Q5–Q1 is computed from the unrounded condition means.

Encoder	Q1	Q2	Q3	Q4	Q5	Q5–Q1
ResNet-18	0.2670415	0.2916385	0.3000700	0.3075740	0.3212570	0.0542155
DenseNet-121	0.2810590	0.3001825	0.3080730	0.3216465	0.3260565	0.0449975

C.6 Prospective Natural-Image Accessibility Replications

The CIFAR-100 experiments vary the predefined conditional-accessibility source-learning intervention while enforcing source-task matching. For each architecture, a matched seed block contains five separately trained Q-specific representations. Q1–Q5 are not different amounts of side information supplied to one fixed encoder.

For both encoders, the condition-level means satisfy

$$Q1 < Q2 < Q3 < Q4 < Q5.$$

This statement concerns means across the five matched seed blocks and does not assert strict adjacent monotonicity within every individual seed block.

ResNet-18. The five seed-level intervention–reuse slopes are

$$0.0474311, \quad 0.0533076, \quad 0.0473905, \quad 0.0529626, \quad 0.0569875.$$

All five are positive. Their mean is

$$0.0516159,$$

with exhaustive seed-block bootstrap 95% interval

$$[0.0485212,\ 0.0547106].$$

All

$$25/25$$

Q-specific source-training runs satisfy the pre-specified source-matching criterion.

DenseNet-121. The five seed-level intervention–reuse slopes are

$$0.0425421, \quad 0.0456553, \quad 0.0467809, \quad 0.0490245, \quad 0.0458580.$$

Again all five are positive. Their mean is

$$0.0459722,$$

with exhaustive seed-block bootstrap 95% interval

$$[0.0440125,\ 0.0477280].$$

All

$$25/25$$

Q-specific DenseNet-121 runs satisfy the same pre-specified source-matching criterion.

The monotone condition-level ordering is therefore an ordering across predefined auxiliary source-learning interventions. It should not be interpreted as obtaining five reuse points by progressively revealing side information to one fixed encoder, nor as directly measuring the amount of task-relevant information retained by each learned representation.

The positive intervention–reuse relationship is reproduced across the two tested encoder architectures under the same CIFAR-100 hierarchy and future-task construction. They do not establish architecture-independent universality or a general theorem that increasing a source-side auxiliary accessibility coordinate must monotonically improve restricted reuse.

Figure 9 complements the reuse results above with a source-task matching audit across the Q1–Q5 interventions for both encoder replications.

Figure 10 visualizes the condition-level Q1–Q5 trajectories and the corresponding seed-level intervention–reuse slopes for the two architecture replications.

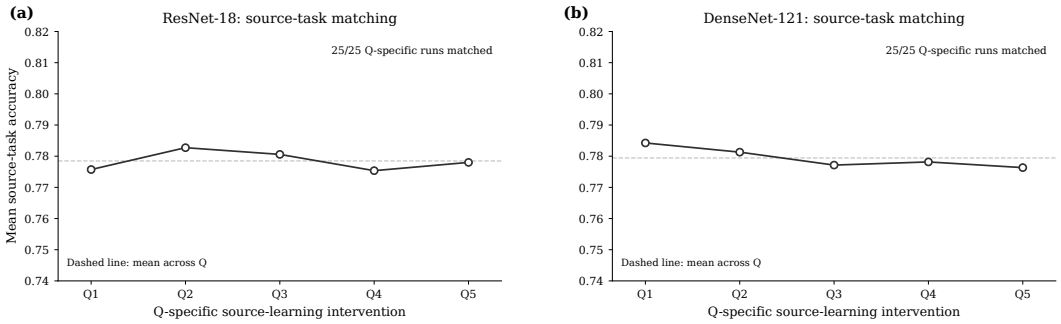

Figure 9: **Source-task matching across Q-specific natural-image replications. (a)** Mean source-task accuracy across the Q1–Q5 source-learning interventions for the ResNet-18 replication. Each Q level corresponds to a separately trained representation, and all 25 architecture–seed–Q source-training runs satisfy the pre-specified source-matching criterion. The dashed horizontal line denotes the mean source-task accuracy across the five Q conditions. **(b)** The corresponding source-task matching audit for the DenseNet-121 replication. All 25 Q-specific DenseNet-121 runs likewise satisfy the same pre-specified matching criterion. Together with Figure 10, this audit separates the observed intervention–reuse relationship from uncontrolled failure of the source-task matching protocol; it does not require source-task accuracy to be numerically identical across Q conditions.

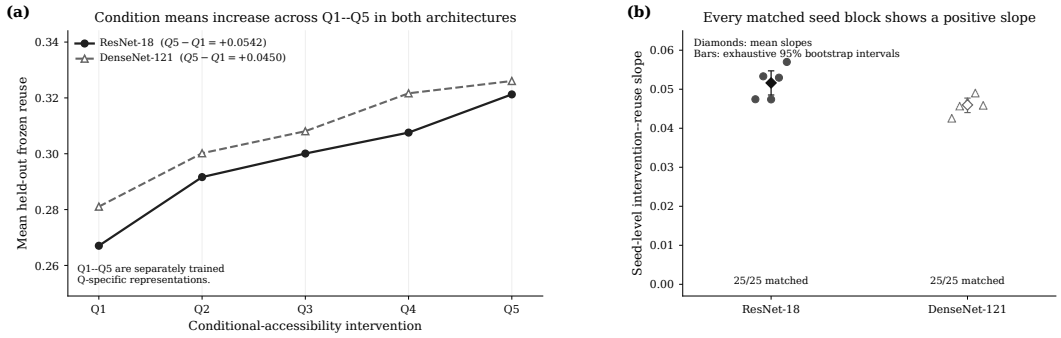

Figure 10: **Conditional-accessibility intervention–reuse relationship across two encoder architectures. (a)** Held-out frozen reuse across the predefined Q1–Q5 source-learning interventions for ResNet-18 and DenseNet-121. Solid circles denote ResNet-18 and dashed open triangles denote DenseNet-121. For each architecture, the legend reports the corresponding endpoint contrast $Q5 - Q1$. Condition means increase from Q1 to Q5 for both architectures. Q1–Q5 denote separately trained intervention-specific representations and should not be interpreted as progressively revealing additional side information to one fixed encoder. **(b)** Seed-level intervention–reuse slopes for the two architecture replications. Individual markers denote matched seed blocks, diamonds denote mean slopes, and vertical bars denote exhaustive seed-block bootstrap 95% intervals. All seed-level slopes are positive for both architectures. All 25 Q-specific source-training runs for each architecture satisfy the pre-specified source-matching criterion.

D Implementation and Reproducibility Details

This section records implementation properties required to interpret and reproduce the claim-supporting experiments. Protocol definitions and statistical estimands are given in Appendix B; numerical results are given in Appendix C.

D.1 Datasets and Split Discipline

The primary controlled experiments use Shapes3D. Natural-image extensions and the conditional-accessibility replications use CIFAR-100.

Source learning and reuse evaluation are separated at both the model and protocol levels. The source model is trained before target adaptation begins. The source head is subsequently removed and the encoder is frozen. Target labels used to fit replacement rules never update the encoder.

In the held-out-composition benchmark, each evaluation fold withholds one of the six composition families from source training. The six held-out families are

$$MM, \quad MD, \quad MU, \quad DD, \quad DU, \quad UU.$$

The natural-image confirmatory experiments use fixed source, adaptation, and evaluation partitions. The conditional-accessibility replications use the same split construction and future-task protocol across Q levels and across the two encoder architectures. The official TEST partition is not used by these confirmatory evaluations.

Target-adaptation draws, task cells, and Q conditions are never counted as additional independently trained representation replicates.

D.2 Encoder and Source-Head Architectures

Within each controlled comparison, the encoder architecture is fixed across the source-learning conditions being compared. Source-task and auxiliary heads are training components only and are removed before reuse evaluation.

The ResNet replication uses a CIFAR-adapted ResNet-18 encoder (He et al., 2016) followed by a 128-dimensional representation projection.

The cross-architecture replication uses DenseNet-121 (Huang et al., 2017) initialized without pretrained weights. For CIFAR-sized inputs, its initial convolution is replaced by a

$$3 \to 64$$

convolution with kernel size 3, stride 1, padding 1, and no bias, and the initial pooling operation is removed. The original classifier is removed and the backbone output is projected to 128 dimensions.

The two architecture replications use the same conditional-accessibility construction, future-task protocol, representation dimension, and downstream evaluation logic. Q1–Q5 are not redefined for the second architecture.

D.3 Source-Training Objectives and Matching

For the primary relational experiments, all compared representation conditions optimize the same relational source objective. C1 additionally optimizes the empirical structural loss,

$$\mathcal{L}_{\mathrm{SPR}} = \widehat{\mathcal{L}}_{\mathrm{src}} + \lambda_{\mathrm{reuse}} \widehat{\mathcal{R}}_{\mathrm{str}}.$$

C4 uses the same structural objective after one fixed global bijection of candidate-state labels. C2 and C3 retain the auxiliary-training scaffold while replacing the task-aligned structural target by random or shuffled targets.

In the exact Shapes3D comparison, all C0, C1, and C4 models attain source validation accuracy

$$1.0000.$$

Conditional-accessibility source learning. The Q1–Q5 experiment trains a separate representation for each architecture–seed–Q combination. For level q, the auxiliary target is

$$A_q = m_q(C, F).$$

The auxiliary-loss coefficient is fixed at

$$\lambda_{\mathrm{aux}} = 1.0.$$

The auxiliary target is used only during source representation learning. It is neither concatenated to the representation nor supplied to the downstream task.

For a matched seed block, all five Q runs share the same source-task, optimization, split, and evaluation protocol and differ only in the predefined Q-specific auxiliary target.

For ResNet-18, the five matched seed identifiers are

$$101, \quad 202, \quad 303, \quad 404, \quad 505,$$

and all 25 architecture–seed–Q runs satisfy the source-matching interval

$$[0.77,\ 0.79].$$

For DenseNet-121, the five matched seed identifiers are

$$1101, \quad 1202, \quad 1303, \quad 1404, \quad 1505,$$

and all 25 runs satisfy the same source-matching criterion.

D.4 Downstream Reuse Readout

Every reported reuse evaluation operates on frozen encoder outputs. The source head and training-time structural head are unavailable to the downstream evaluator.

For limited-supervision experiments, the downstream rule is fitted only from the registered target-adaptation observations for the corresponding budget. The candidate-state task-family experiment evaluates

$$K \in \{1, 2, 4, 8, 16\}$$

examples per state and a full-support condition.

Adaptation indices are paired across representation conditions whenever a paired model contrast is reported. Thus C1–C0, C4–C0, and related paired effects use common evaluation tasks and matched adaptation realizations.

Readout class, optimization, and regularization are fixed within each evaluation protocol and are not tuned independently for each representation condition. Protocols using distinct evaluators are analyzed separately; their absolute accuracies are not interpreted as measurements on one common scale.

D.5 State-Recoverability Diagnostic

The recoverability experiment fits a diagnostic state readout to the frozen representation and reports macro-F1 after one global three-state label alignment. The alignment searches over the six permutations of $\{M, D, U\}$ and selects a single global relabeling for the diagnostic metric.

This is permutation-invariant metric alignment. It does not modify the encoder and is not supplied as side information to the downstream reuse evaluator.

The same diagnostic is used for C0, C1, C4, Random auxiliary, and Shuffled-state.

D.6 Coordinate-Transformation Implementation

All coordinate interventions are applied after representation learning to the same frozen representation.

The transformation bank consists of reversible coordinate maps. Identity applies no transformation; Global applies one transformation shared by all samples; 2-way, 4-way, and 8-way use increasingly heterogeneous sample-dependent convention sets; and the sample-wise condition varies the convention at the sample level. Convention identity is withheld from the blind downstream decoder.

For the rescue evaluation, convention identity is supplied and its exact inverse is applied before downstream evaluation. The reconstruction audit records

$$\max_i \|\widehat{z}_i - z_i\| = 0,$$

and reuse returns to

$$0.4096$$

in every rescued condition.

The rescue removes the imposed coordinate intervention; it does not retrain the representation or provide additional task information.

D.7 Task Registry and Admissibility Implementation

All primary registry decisions are representation independent.

For the Shapes3D novel-rule analysis, the executable registry reproduces the partition

$$2032 = 1008 + 1008 + 16$$

into oracle-solvable, oracle-unsolvable, and single-class candidate cells, and the final novelty/admissibility criteria retain 1004 cells.

For the held-out candidate-state analysis, the executable oracle audit applies the semantic interface in Eq. (39), the ridge objective in Eq. (40), the zero-threshold binary rule in Eq. (41), and the leave-one-composition-out procedure defined in Appendix B.5. It reproduces the 21 included and three excluded cells in Table 2.

The resulting registries are fixed before learned-representation comparisons and reused unchanged across source-learning conditions, seeds, and supervision levels.

D.8 Statistical Aggregation and Confidence Intervals

The primary inferential unit for representation comparisons is the independently trained representation seed.

For the held-out candidate-state family, repeated target-adaptation measurements are first aggregated within each seed–condition–supervision cell. The fixed admissible task registry is then aggregated to form $A_{s,C}(K)$. Limited-supervision effects, full-support effects, and Γ are computed from the seed-level paired estimands defined in Appendix B.9.

Confidence intervals for paired representation comparisons resample complete representation-seed blocks so that repeated task and adaptation measurements from the same trained representation remain grouped.

For the conditional-accessibility experiment, one matched seed block contains five separately trained Q-specific models. The Q1–Q5 trajectory remains intact during resampling. The seed-level trend coefficient is Eq. (44), and the architecture-level trend is the arithmetic mean of the five seed-level coefficients.

Exact bootstrap inference enumerates the

$$5^5 = 3125$$

ordered resamples of the five matched seed blocks. The reported 95% interval is the equal-tailed percentile interval of the resulting distribution of the mean trend coefficient. Here "exact" means that the complete finite bootstrap resampling space is enumerated; it does not denote an exact finite-sample coverage guarantee.

Condition-level monotonicity and positivity of the five seed-level slopes are reported separately because they are logically distinct empirical properties.

D.9 Anonymous Reproducibility Manifest and Prospective Protocol

The anonymous reproducibility materials separate four objects: the scientific protocol, executable implementation, per-run outcomes, and statistical aggregation. Machine-local paths and internal experiment-management names are not part of the public release interface.

Prospective freeze chronology. For each prospective natural-image replication, the scientific design is fixed before the corresponding confirmatory reuse outcomes are generated. The frozen design includes the Q1–Q5 auxiliary-target construction, identity-level lookup tables $m_q(C, F)$, source and

Table 6: Public-facing organization of the anonymous reproducibility materials.

Component	Contents
Protocol specification	Task construction, split definitions, source-matching criteria, exact Q1–Q5 identity-level mappings, hyperparameters, seed sets, and statistical estimands.
Task registries	Source-matched novel-rule registry and the 24-cell leave-one-composition-out oracle disposition.
Training records	Per-seed source-training summaries for all reported representation conditions and all Q levels, including source-performance matching.
Evaluation records	Frozen-representation reuse measurements, recoverability diagnostics, and coordinate-intervention and rescue results.
Statistical code and summaries	Seed-level paired contrasts, supervision-localized effects, conditional-accessibility trend coefficients, exhaustive bootstrap enumeration, and manuscript-level aggregate tables.
Integrity manifest	Release-relative file identifiers together with version or checksum information linking protocol definitions, executable code, per-run outputs, and aggregate manuscript statistics.

evaluation splits, representation-seed blocks, future-task construction, source-matching criterion, downstream evaluation rule, and primary statistical analysis.

Implementation-only corrections after this freeze are restricted to protocol-preserving execution changes: they do not redefine the scientific conditions, seed sets, task family, source-matching criterion, primary estimand, or statistical analysis.

Random-state identifiers. The ResNet-18 replication uses the matched representation seed identifiers

$$101, \quad 202, \quad 303, \quad 404, \quad 505,$$

and the DenseNet-121 replication uses

$$1101, \quad 1202, \quad 1303, \quad 1404, \quad 1505.$$

The future-task construction uses the fixed random seed

$$2026092401.$$

Each Q-specific representation has an individual source-training record and an individual frozen-reuse evaluation record. The aggregate statistics reported in this Appendix are computed from these per-run records rather than from manually entered manuscript values.

Anonymous release organization. The supplementary material is organized by scientific role:

The release provides a traceable path from each manuscript-level claim to its scientific protocol, executable evaluator, per-run measurements, and aggregate statistic without exposing internal development-stage naming or machine-specific filesystem structure.