CONCEPT ATTRIBUTION ANALYSIS: A PRINCIPLED APPROACH TO ATTRIBUTING CHANGES IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) encode rich semantic information in their internal representations, motivating growing interest in extracting human-interpretable concepts for model understanding and analysis. Existing approaches, including linear probing, sparse autoencoders, and concept component analysis, primarily adopt a single-context perspective, asking what concepts are encoded in the representation associated with *a single input context*. Such a perspective is inherently underdetermined: multiple concept-level explanations may provide equally plausible interpretations of the same representation, making it difficult to determine which one faithfully reflects the underlying concept structure. In this work, we move from single contexts to *pairs of related contexts* and exploit the structured variation in their corresponding LLM representations. Building on the recent theoretical characterization of LLM representations through concept log-posteriors, we formulate the representation change between paired contexts as a sum of concept-specific change contributions. We then establish conditions under which these contributions are identifiable up to a global permutation, even when multiple concepts change simultaneously within a given context pair. Guided by this analysis, we develop a practical method, termed *Concept Attribution Analysis*, that attributes an observed representation change to individual concept changes from paired LLM representations. Experiments on controlled simulations validate the predicted identifiability behavior and demonstrate robustness under departures from the idealized assumptions, while experiments on real LLM representations demonstrate the practical utility of concept-level change attribution.

1 INTRODUCTION

Large language models (LLMs) have demonstrated increasingly significant capabilities across a wide range of tasks (Brown et al., 2020; Dubey et al., 2024), yet the internal mechanisms underlying their behaviors remain difficult to interpret (Sharkey et al., 2025; Rai et al., 2024). Understanding what information is encoded and processed internally by LLMs has therefore become a fundamental problem for understanding how these models work. More importantly, such understanding provides an operational handle for more precisely exploiting and evaluating LLM capabilities. By revealing which internal information is relevant to a particular task or behavior, it may enable task-specific information to be identified and utilized more selectively, rather than relying on the representation as an undifferentiated whole (Zou et al., 2023; Li et al., 2023). At the same time, a mechanistic understanding can provide principled criteria for evaluating whether a method acts on the intended internal information, offering a foundation for diagnosing, comparing, and ultimately improving downstream methods (Geiger et al., 2021; Meng et al., 2022; Marks & Tegmark, 2024).

A promising direction is to extract human-interpretable concepts from LLM internal representations, thereby providing a human-aligned semantic lens through which the internal behavior of LLMs can be understood. Most existing approaches adopt a largely *single-context* perspective, asking what concepts are encoded in the representation induced by one input context. For example, motivated by the linear representation hypothesis (Park et al., 2024; Alain & Bengio, 2016), linear probing has been widely used to extract concepts associated with a given input context from LLM representations using supervised labels (Belinkov, 2022; Marks & Tegmark, 2024). Building further on the superposition hypothesis (Elhage et al., 2022), sparse autoencoders (SAEs) aim to recover such concepts in an unsupervised manner through sparsity regularization (Bricken et al., 2023; Huben et al.,

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.

2024). Beyond these two empirical hypotheses, a recent theoretical finding, showing that LLM representations can be approximated as linear mixtures of concept log-posteriors (Liu et al., 2026b), motivates concept component analysis (ConCA), which aims to recover the underlying concept-posterior components from LLM representations (Liu et al., 2026a).

Challenges. Despite the progress above, these approaches share a common focus: they primarily characterize *what concepts are encoded at a given input context*, providing a largely single-context view of LLM internals. Such a view is inherently underdetermined: a single input context induces only a single LLM representation, which may admit multiple concept-level explanations that are equally consistent with that representation, making it unclear which explanation faithfully reflects the underlying concept structure. This underdetermination directly limits existing concept-extraction methods. Linear probes may exploit spurious correlations (Hewitt & Liang, 2019; Ravichander et al., 2021), while SAEs may admit multiple sparse decompositions despite accurate reconstruction (Nelson et al., 2026; Cui et al., 2026). Even ConCA, although grounded in a theoretical characterization of concept posteriors in LLM representations, remains an ill-posed inverse problem, with the identifiability of the extracted concepts left unresolved (Liu et al., 2026a).

Key insight. In this work, we move beyond the static, single-context perspective by considering pairs of related contexts and the corresponding changes in their LLM representations. Our key insight is that such paired changes provide structured variation that is unavailable from an isolated representation: while a single representation may admit multiple plausible concept-level explanations, how representations vary across related contexts imposes additional constraints on these explanations, thereby reducing the underdetermination inherent in single-context analysis.

Contributions. Beyond the above insight, we make the following contributions. We first formulate concept extraction through the lens of paired input contexts, extending the recent theoretical characterization of concept posteriors in LLM representations (Liu et al., 2026b) in Sec. 2. This formulation decomposes the representation change between two paired contexts into changes attributable to individual latent concepts. We then establish identifiability guarantees for the proposed formulation in Sec. 3, showing that, under certain conditions, individual concept change contributions are identifiable up to a trivial permutation, even when multiple concepts change simultaneously for a given pair. We further translate this theoretical insight into a practical concept attribution method, termed concept attribution analysis (ConAn), which exploits the subspace geometry and sparsity of concept changes to recover individual concept change contributions from paired LLM representations in Sec. 4. We empirically validate the theoretical findings and the proposed method in Sec. 5. Controlled simulations verify the predicted identifiability behavior and demonstrate graceful degradation under violations of the idealized assumptions. We also evaluate our method on real LLMs, showing its ability to recover and attribute concept-level changes in practical settings.

Compared with concept extraction from a single context, which primarily reveals *what concepts are represented*, the paired-context perspective in this work provides a complementary view of *how concept information changes*. This distinction is important in many settings where the goal is not merely to characterize a model state, but to understand what internal changes accompany a transition in input, reasoning, or behavior. By attributing such transitions to individual concepts, this work provides a basis for analyzing how LLMs respond and evolve across related contexts.

2 Preliminaries and Problem Setup

In this section, we first review the theoretical characterization of LLM representations as linear mixtures of latent concept posteriors, which provides the foundation for our analysis. Building on this characterization, we then introduce the paired-context setting, formulate representation changes in terms of concept-specific change contributions, and define the corresponding identifiability.

2.1 Preliminaries: LLM Representations as a Mixture of Concept Posteriors

To relate LLM representations to the human-interpretable concepts underlying text, Liu et al. (2026b) consider a latent-variable generative model for text. Let $\mathbf{z} = (z_1, \ldots, z_K)$ denote a collection of K discrete latent concepts, where each z_k represents an individual human-interpretable

semantic factor, such as topic or sentiment, and takes values in a concept-specific space $\mathcal{Z}_k$. The joint latent space is $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_K$. An observed input context $\mathbf{x}$ and its next token y are assumed to be generated conditionally on the latent concept configuration $\mathbf{z}$. Marginalizing over the latent concepts gives the observed joint distribution, as follows:

$$p(\mathbf{x}, y) = \sum\nolimits_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x} \mid \mathbf{z})p(y \mid \mathbf{z})p(\mathbf{z}), \tag{1}$$

where $p(\mathbf{z})$ denotes the joint prior over the latent concepts, while $p(\mathbf{x} \mid \mathbf{z})$ and $p(y \mid \mathbf{z})$ characterize the conditional generation of the input context and next token, respectively. Here, the latent concepts are not required to be mutually independent, and their dependencies can be captured by $p(\mathbf{z})$.

We use the individual-concept representation relation of Liu et al. (2026a, Theorem 2.2). Under their output-embedding diversity, informational sufficiency, and regularity conditions, LLM representations learned through next-token prediction are approximately linear mixtures of individual concept log-posteriors. The earlier result of Liu et al. (2026b) concerns log-posteriors over joint latent configurations. Specifically, the individual-concept relation used here is

$$\mathbf{f}(\mathbf{x}) \approx \mathbf{A}\left[[\log p(z_1 \mid \mathbf{x})]_{z_1}; \ldots; [\log p(z_K \mid \mathbf{x})]_{z_K}\right] + \mathbf{b}, \tag{2}$$

where $\mathbf{f}(\mathbf{x})$ denotes the last-layer LLM representation of the input context $\mathbf{x}$, $\mathbf{A}$ is a linear mixing matrix, and $\mathbf{b}$ is a constant bias vector. Here, $[\log p(z_k \mid \mathbf{x})]_{z_k}$ denotes the vector obtained by evaluating the log-posterior over all possible states of the k-th latent concept.

2.2 Problem Setup: Concept Change Contributions from Paired Contexts

In this work, we move beyond the single-context setting and consider pairs of related input contexts $(\mathbf{x}, \mathbf{x}') \sim \Pi$, where Π denotes a distribution over context pairs. We emphasize that the two contexts may differ in one or multiple underlying concepts, without requiring prior knowledge of which concepts change within each pair. To expose the concept-wise structure of Eq. (2), we partition the mixing matrix as $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_K]$ and define the contribution of concept z_k under context $\mathbf{x}$ as $\mathbf{C}_k(\mathbf{x}) := \mathbf{A}_k[\log p(z_k \mid \mathbf{x})]_{z_k}$. Given these definitions, for a context pair $(\mathbf{x}, \mathbf{x}')$, taking the difference between their representations gives:

$$\Delta\mathbf{f}(\mathbf{x}, \mathbf{x}') \approx \sum\nolimits_{k=1}^{K} \mathbf{G}_k(\mathbf{x}, \mathbf{x}'), \tag{3}$$

where $\Delta\mathbf{f}(\mathbf{x}, \mathbf{x}') := \mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}')$, $\mathbf{G}_k(\mathbf{x}, \mathbf{x}') := \mathbf{C}_k(\mathbf{x}) - \mathbf{C}_k(\mathbf{x}')$. We refer to $\mathbf{G}_k(\mathbf{x}, \mathbf{x}')$ as the *concept change contribution*, representing the representation change associated with concept z_k. We abbreviate the observed representation change as $\mathbf{d}(\mathbf{x}, \mathbf{x}') := \Delta\mathbf{f}(\mathbf{x}, \mathbf{x}')$, writing $\mathbf{d}$ when the context pair is clear.

Our central problem is to determine whether the individual concept change contributions $\mathbf{G}_k(\mathbf{x}, \mathbf{x}')$ can be uniquely recovered from the collection of paired representation changes $\Delta\mathbf{f}(\mathbf{x}, \mathbf{x}')$. Such recovery would provide a complementary, change-centric view of LLM internals, moving beyond the predominantly static perspective of existing concept-extraction methods. To formalize this question, we introduce the following notion of identifiability.

Definition 2.1 (Contribution identifiability)**.** Let $\mathbf{G} = (\mathbf{G}_1, \ldots, \mathbf{G}_K)$ denote the true concept change decomposition of the representation changes $\Delta\mathbf{f}(\mathbf{x}, \mathbf{x}')$, and let $\mathcal{M}$ be the class of candidate decompositions. We say that $\mathbf{G}$ is contribution-identified if every $\widetilde{\mathbf{G}} \in \mathcal{M}$ that induces the same $\mathbf{d}$ coincides with $\mathbf{G}$ up to a permutation, i.e., there exists a permutation σ of $[K] = \{1, \ldots, K\}$ such that $\widetilde{\mathbf{G}}_{\sigma(k)}(\mathbf{x}, \mathbf{x}') = \mathbf{G}_k(\mathbf{x}, \mathbf{x}')$.

3 Theory: Contribution Identifiability of Concept Change

We now study when the concept change contributions $\mathbf{G}_k(\mathbf{x}, \mathbf{x}')$ in Eq. (3) can be uniquely identified from paired representation changes $\Delta\mathbf{f}(\mathbf{x}, \mathbf{x}')$. The main challenge is that the same observed representation change may admit multiple plausible concept decompositions, making the underlying concept contributions generally non-identifiable without additional assumptions. Our key observation is that variation across context pairs provides additional geometric structure for identifiability, as detailed in Sec. 3.1. By exploiting this structure, we show that the concept change contributions $\mathbf{G}_k(\mathbf{x}, \mathbf{x}')$ can be uniquely identified under suitable conditions, as formalized in Sec. 3.2.

3.1 Geometric Intuition for Contribution Identifiability

Before presenting the formal identifiability result, we first explain the key geometric intuition behind why paired representation changes may make individual concept change contributions identifiable. To this end, for each concept k, let $\mathcal{U}_k^\Delta$ denote the smallest linear subspace containing $\mathbf{G}_k(\mathbf{x}, \mathbf{x}')$ almost surely, and let $r_k := \dim(\mathcal{U}_k^\Delta) > 0$. Thus, $\mathcal{U}_k^\Delta$ characterizes all possible directions along which concept k can contribute to a representation change. For any subset of concepts $T \subseteq [K]$, define their combined change space as $\mathcal{V}_T^\Delta := \sum_{k \in T} \mathcal{U}_k^\Delta$. We further define the change support of a context pair as $S^\Delta(\mathbf{x}, \mathbf{x}') := \{k \in [K] : \mathbf{G}_k(\mathbf{x}, \mathbf{x}') \neq \mathbf{0}\}$. If $S^\Delta(\mathbf{x}, \mathbf{x}') = T$, then $\Delta\mathbf{f}(\mathbf{x}, \mathbf{x}') = \sum_{k \in T} \mathbf{G}_k(\mathbf{x}, \mathbf{x}')$ and hence $\Delta\mathbf{f}(\mathbf{x}, \mathbf{x}') \in \mathcal{V}_T^\Delta$. Therefore, different sets of changing concepts correspond to different combined change spaces.

The key observation is that these combined spaces inherit the overlap structure of their underlying concepts. When two change supports share a concept, both combined spaces contain the change space of that shared concept. Under suitable geometric conditions, their intersection contains no additional directions and thus isolates the concept space, as illustrated by the following example.

Example 3.1. Consider three concepts with possible change supports $\{1, 2\}$, $\{1, 3\}$, and $\{2, 3\}$. Their corresponding representation changes lie in $\mathcal{V}_{\{1,2\}}^\Delta$, $\mathcal{V}_{\{1,3\}}^\Delta$, and $\mathcal{V}_{\{2,3\}}^\Delta$, respectively. The key observation is that the first two supports share only concept 1. Under suitable geometric conditions, the intersection of their combined change spaces therefore isolates the change space of concept 1: $\mathcal{V}_{\{1,2\}}^\Delta \cap \mathcal{V}_{\{1,3\}}^\Delta = \mathcal{U}_1^\Delta$. Similarly, $\mathcal{V}_{\{1,2\}}^\Delta \cap \mathcal{V}_{\{2,3\}}^\Delta = \mathcal{U}_2^\Delta$ and $\mathcal{V}_{\{1,3\}}^\Delta \cap \mathcal{V}_{\{2,3\}}^\Delta = \mathcal{U}_3^\Delta$.

This example illustrates the key mechanism behind contribution identifiability: variation across context pairs provides different combinations of concept changes, while intersections of their corresponding combined change spaces can isolate the change space of each individual concept. Importantly, this does not require context pairs in which only one concept changes. We next formalize the conditions under which this geometric intuition guarantees contribution identifiability.

3.2 Identifiability Analysis

The intuition above suggests three requirements for contribution identifiability. First, different concept change spaces should be sufficiently separated so that their contributions can be uniquely decomposed. Second, for each observed change pattern, the representation changes should vary sufficiently to reveal the corresponding combined change space. Third, the change patterns across pairs should overlap in a way that allows each individual concept space to be isolated.

To formalize these conditions, among all nonempty change supports that occur with positive probability, let $\mathcal{S}_{\max}$ denote those that are maximal under set inclusion. Intuitively, each $T \in \mathcal{S}_{\max}$ corresponds to a maximal combination of concepts that can change together, with associated combined change space $\mathcal{V}_T^\Delta$.

Assumption 3.2 (Local directness)**.** For any $T, T' \in \mathcal{S}_{\max}$, $\mathcal{V}_{T \cup T'}^\Delta = \bigoplus_{k \in T \cup T'} \mathcal{U}_k^\Delta$.

This assumption requires the concept change spaces involved in any two maximal change patterns to be linearly independent. As a result, a representation change within these spaces admits a unique decomposition into concept-specific contributions. Orthogonality is not required.

Assumption 3.3 (Conditional richness)**.** For every $T \in \mathcal{S}_{\max}$, the conditional distribution of $\mathbf{d}$ given $S^\Delta = T$ is not concentrated on any finite union of proper linear subspaces of $\mathcal{V}_T^\Delta$.

This assumption ensures that, when a change pattern T occurs, the corresponding representation changes exhibit sufficient variation to reveal the full combined space $\mathcal{V}_T^\Delta$, rather than only a lower-dimensional part of it.

Assumption 3.4 (Support intersection)**.** For every $k \in [K]$, there exists a nonempty collection $\mathcal{G}_k \subseteq \mathcal{S}_{\max}$ such that $\bigcap_{T \in \mathcal{G}_k} T = \{k\}$.

This assumption ensures that each concept can be isolated through the overlap of multiple change patterns. In particular, it does not require observing a pair in which only concept k changes; it is sufficient that several multi-concept change patterns share only concept k.

Together, these three assumptions respectively ensure that the concept change spaces are separable, the combined change spaces are observable from variation across pairs, and each individual concept

space can be isolated through their overlap. Under these conditions, we can establish contribution identifiability, as below.

Theorem 3.5 (Contribution Identifiability Result)**.** *Suppose that the paired representation changes are generated by the true concept change decomposition* $\Delta\mathbf{f}(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^{K} \mathbf{G}_k(\mathbf{x}, \mathbf{x}')$*, and that* $\mathbf{G} = (\mathbf{G}_1, \ldots, \mathbf{G}_K)$ *satisfies Assumptions 3.2-3.4. Then for any candidate decomposition* $\widetilde{\mathbf{G}} = (\widetilde{\mathbf{G}}_1, \ldots, \widetilde{\mathbf{G}}_{\widetilde{K}})$ *that satisfies the same assumptions and induces the same observed representation changes, i.e.,* $\sum_{j=1}^{\widetilde{K}} \widetilde{\mathbf{G}}_j(\mathbf{x}, \mathbf{x}') = \Delta\mathbf{f}(\mathbf{x}, \mathbf{x}')$*, the true decomposition* $\mathbf{G}$ *is contribution-identified.*

Proof sketch. The proof proceeds in four steps. First, conditional richness ensures that each maximal support reveals its full combined change space $\mathcal{V}_T^{\Delta}$. Since the true and candidate decompositions induce the same observed representation changes, their maximal combined change spaces must coincide. Second, local directness makes intersections of these spaces reflect intersections of their underlying supports. Together with Assumption 3.4, this isolates each individual concept change space $\mathcal{U}_k^{\Delta}$. Third, the recovered concept spaces determine a consistent global correspondence between the true and candidate decompositions. Finally, local directness guarantees a unique decomposition of each observed representation change over the corresponding concept spaces, which yields $\widetilde{\mathbf{G}}_{\sigma(k)} = \mathbf{G}_k$ almost surely. The complete proof is provided in Appendix C.3.

Selection by group sparsity. Theorem 3.5 establishes contribution identifiability when both the true and candidate decompositions satisfy Support Intersection (Assumption 3.4). We now consider a broader candidate class in which this condition is imposed only on the true decomposition. Without candidate Support Intersection, exact reconstruction alone may admit refinements that split a true concept into multiple candidate blocks, even when Local Directness and Conditional Richness are retained. We show that minimizing the expected sum of contribution norms rules out such nontrivial refinements and selects the true contributions within this broader class. To ensure that the selection objective is finite for the true decomposition, we impose the following mild integrability condition.

Assumption 3.6 (Contribution integrability)**.** For every $k \in [K]$, $\mathbb{E}_{(\mathbf{x},\mathbf{x}')\sim\Pi}\|\mathbf{G}_k(\mathbf{x}, \mathbf{x}')\|_2 < \infty$.

Then, let $\mathcal{M}_{\text{sel}}$ denote the class of finite, measurable candidate decompositions that reconstruct $\Delta\mathbf{f}$ exactly, have each contribution nonzero with positive Π-probability, and satisfy Assumptions 3.2 and 3.3, but not necessarily Assumption 3.4. Further, we define the group-sparsity criterion as $\mathcal{N}(\widetilde{\mathbf{G}}) := \mathbb{E}_{\Pi} \sum_{j=1}^{\widetilde{K}} \|\widetilde{\mathbf{G}}_j\|_2$. Given these, we obtain:

Theorem 3.7 (Selection by group sparsity)**.** *Suppose the true decomposition* $\Delta\mathbf{f} = \sum_{k=1}^{K} \mathbf{G}_k$ *satisfies Assumptions 3.2–3.4 and 3.6. Then, for every* $\widetilde{\mathbf{G}} \in \mathcal{M}_{\text{sel}}$*,* $\mathcal{N}(\widetilde{\mathbf{G}}) \geq \mathcal{N}(\mathbf{G})$*, with equality if and only if* $\widetilde{K} = K$ *and there exists a permutation* σ *of* $[K]$ *such that* $\widetilde{\mathbf{G}}_{\sigma(k)} = \mathbf{G}_k$ Π*-almost surely for every* $k \in [K]$*.*

Proof sketch. The maximal-space and intersection arguments imply that every admissible candidate is an internal direct-sum refinement of the true contributions. The Euclidean triangle inequality is strict on a positive-probability maximal-support event whenever a true block is split, so only the original contributions minimize the expected sum of norms; see Appendix C.4.

Quantitative stability. Appendix C.6 extends contribution identifiability to decompositions whose sums differ by a small L^2 error. Under the quantitative directness, richness, and support conditions of Assumption C.9 on both decompositions, Theorem C.10 bounds the errors in the recovered change spaces and contribution vectors. Appendix C.7 incorporates representation residuals and observation noise through explicit error budgets.

Discussion: Comparison with Prior Work. Our identifiability results exploit structured variation across paired LLM representations, providing information beyond the single-context representations analyzed by SAEs (Bricken et al., 2023) and ConCA (Liu et al., 2026a). This shifts the focus to how latent concepts contribute to representation changes: overlapping combinations of changes help resolve concept attribution ambiguities, without requiring single-concept interventions.

Relative to other approaches that also exploit paired changes (Joshi et al., 2026), our analysis starts from the posterior representation framework of Liu et al. (2026b) and ConCA (Liu et al., 2026a).

This framework derives approximate posterior linearity from latent-variable models of next-token prediction under stated conditions, rather than postulating that representations linearly encode latent concept values. Furthermore, by imposing local geometric and distributional conditions on the true and candidate decompositions, Theorem 3.5 identifies multidimensional contributions through intersections of combined change spaces. This bypasses global injectivity over varying concepts and a globally invertible latent coordinate transformation, allowing globally overcomplete change spaces and competing decompositions with different block counts and dimensions. Theorem 3.7 further establishes identifiability through norm-based selection: minimizing the expected sum of Euclidean contribution norms among exact candidates satisfying local directness and conditional richness recovers the true contributions without requiring candidate support intersection. This provides a population-level theoretical basis for contribution regularization in the subsequent method.

Discussion: Identifiability without Single-Concept Changes. We emphasize, again, that our identifiability result does not require context pairs in which only a single concept changes. Individual concept contributions can instead be identified from overlapping multi-concept changes, provided that the observed change patterns exhibit sufficient variation and overlap in a way that isolates each concept. This substantially relaxes the need for carefully constructed single-concept interventions, which may be difficult to obtain in realistic LLM settings.

4 Concept Attribution Analysis

Our theoretical results motivate two complementary design choices: exploiting the geometry of paired changes to initialize concept spaces (Theorem 3.5), and using group sparsity to guide contribution recovery (Theorem 3.7). These insights inform Concept Attribution Analysis (ConAn), a practical method that combines graph-based initialization with representation-change reconstruction, regularized by the sum of contribution norms as specified in Eqs. (4)–(5).

Reconstructing Representation Changes with Sparsity. We parameterize each candidate concept contribution within a low-dimensional change space. Let $\mathbf{B}_j \in \mathbb{R}^{D\times r}$ denote the basis of the j-th candidate concept change space, and let $\mathbf{u}_{\boldsymbol{\theta},j}(\Delta\mathbf{f}) \in \mathbb{R}^r$ denote its pair-specific coefficients. The estimated contribution is $\widehat{\mathbf{G}}_j(\mathbf{x},\mathbf{x}') = \mathbf{B}_j\mathbf{u}_{\boldsymbol{\theta},j}(\Delta\mathbf{f}(\mathbf{x},\mathbf{x}'))$, where m denotes the number of candidate concept spaces. Given the above, we minimize the following two terms:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}\left\|\Delta\mathbf{f} - \sum\nolimits_{j=1}^{m}\widehat{\mathbf{G}}_j\right\|_2^2, \qquad \mathcal{R}_{\text{sp}} = \mathbb{E}\sum\nolimits_{j=1}^{m}\|\widehat{\mathbf{G}}_j\|_2, \tag{4}$$

where the first term encourages accurate reconstruction of the observed representation change, and the second encourages block-sparse concept contributions. Our training objective is

$$\min_{\boldsymbol{\theta},\mathbf{B}} \quad \mathcal{L}_{\text{rec}} + \lambda\mathcal{R}_{\text{sp}}, \tag{5}$$

where λ balances reconstruction accuracy and group sparsity.

Geometry-Guided Concept Space Initialization. The reconstruction loss encourages the estimated contributions to explain the observed representation change, while the group-sparsity penalty encourages a small number of active concept blocks for each pair. To further align the estimator with Assumptions 3.2-3.4, we use the following geometry-guided initialization. See Sec. D for details.

Assumption 3.3 implies that sufficiently varied changes sharing a common change pattern reveal its combined change space. Accordingly, we construct a similarity graph over the observed representation changes, obtain overlapping soft memberships through graph factorization, and estimate candidate change directions from the corresponding second-order moments.

Assumption 3.4 motivates identifying change directions that recur across different multi-concept patterns. We therefore apply sparse subspace clustering to the candidate directions and group the resulting recurring subspace structure into candidate concept blocks $\{\mathbf{B}_j\}_{j=1}^m$.

Finally, Assumption 3.2 characterizes when contributions over the recovered concept spaces admit a unique decomposition. Our estimator therefore performs block-wise sparse reconstruction over the initialized spaces. During optimization, each decoder block is maintained as a full-rank orthonormal basis for numerical stability, while no orthogonality is imposed across different blocks.

5 EXPERIMENTS

5.1 SYNTHETIC EXPERIMENTS

Data Generation. We generate representations from dependent Gaussian concepts with known multidimensional contributions. For each context pair, we jointly resample multiple concepts conditional on the unchanged ones, ensuring that no pair isolates a single concept. The baseline configurations satisfy Assumptions 3.2–3.4 almost surely. We vary the observation dimension and maximum support size to examine different levels of overcompleteness and concept coactivity. To study the role of each structural condition, we additionally introduce controlled violations through shared concept directions, restricted joint change variation, and always-coactive concept pairs. Full details of the data generation and experimental configurations are provided in Appendix E.

Evaluation Protocol. Since the synthetic data provide the underlying concept, we can directly evaluate both aggregate reconstruction and individual concept attribution. Specifically, for test pairs indexed by n, let $\mathbf{d}_n$ denote the observed representation change, $\mathbf{G}_{n,k}$ the true contribution of concept k, and $\widehat{\mathbf{G}}_{n,j}$ its predicted counterpart. The reconstructed change is $\widehat{\mathbf{d}}_n = \sum_{j=1}^{m} \widehat{\mathbf{G}}_{n,j}$. We define

$$\text{Delta FVE} = 1 - \frac{\sum_n \|\widehat{\mathbf{d}}_n - \mathbf{d}_n\|_2^2}{\sum_n \|\mathbf{d}_n\|_2^2}, \quad \text{Block FVE} = 1 - \frac{\sum_{n,k} \|\widehat{\mathbf{G}}_{n,\sigma(k)} - \mathbf{G}_{n,k}\|_2^2}{\sum_{n,k} \|\mathbf{G}_{n,k}\|_2^2}, \tag{6}$$

where σ is a single global permutation obtained by matching predicted and true contributions on a separate alignment set and then held fixed for test evaluation. Delta FVE measures reconstruction of the aggregate representation change, whereas Block FVE measures recovery of the individual concept contributions, including their magnitudes and predictions for inactive concepts.

To distinguish individual recovery from recovery of a combined contribution, we additionally evaluate a preselected concept pair (i, j) using

$$\text{Pair FVE}_{i,j} = 1 - \frac{\sum_n \left\| \widehat{\mathbf{G}}_{n,\sigma(i)} + \widehat{\mathbf{G}}_{n,\sigma(j)} - \mathbf{G}_{n,i} - \mathbf{G}_{n,j} \right\|_2^2}{\sum_n \|\mathbf{G}_{n,i} + \mathbf{G}_{n,j}\|_2^2}. \tag{7}$$

This metric assesses whether the combined contribution remains recoverable when its individual components are difficult to separate.

We compare two training protocols, *graph* and *random*, using geometry-guided and random initialization, respectively. Both use the same architecture, learning objective, and total update budget. All main-text results use identity inputs, direct predictions from the final checkpoint, and a common $\lambda = 0.1$ (see Appendix F.2 for the penalty ablation). We additionally conduct penalty sweeps to examine the effect of group sparsity on contribution recovery. We average two training seeds within each of three independently generated worlds and report the sample standard deviation across world averages. Further implementation details and complete experimental results are provided in Appendices D–F.

Recovery under the assumptions. Figure 1(a) shows that the geometry-guided protocol recovers individual contributions with little variation across the sampled worlds, despite observing only simultaneous multi-concept changes. The random-initialization protocol also reconstructs the aggregate change accurately, but its contribution recovery is substantially weaker. Recovery under the graph protocol improves as the observation space becomes less overcomplete, and becomes less accurate as more concepts change together.

Recovery under assumption violations. We separately introduce shared directions between two concept spaces, restrict joint change variation, and bind pairs of concepts to change together. Limited subspace overlap has a modest effect on recovery, whereas complete overlap sharply reduces recovery of the individual contributions (Figure 1(b)). Binding a small number of pairs leaves overall recovery relatively stable, but already limits separation of the concepts within those pairs. As more pairs are bound, overall Block FVE deteriorates while their combined contributions remain well recovered (Figure 1(d)). Thus, stable aggregate recovery can conceal ambiguity at the level of

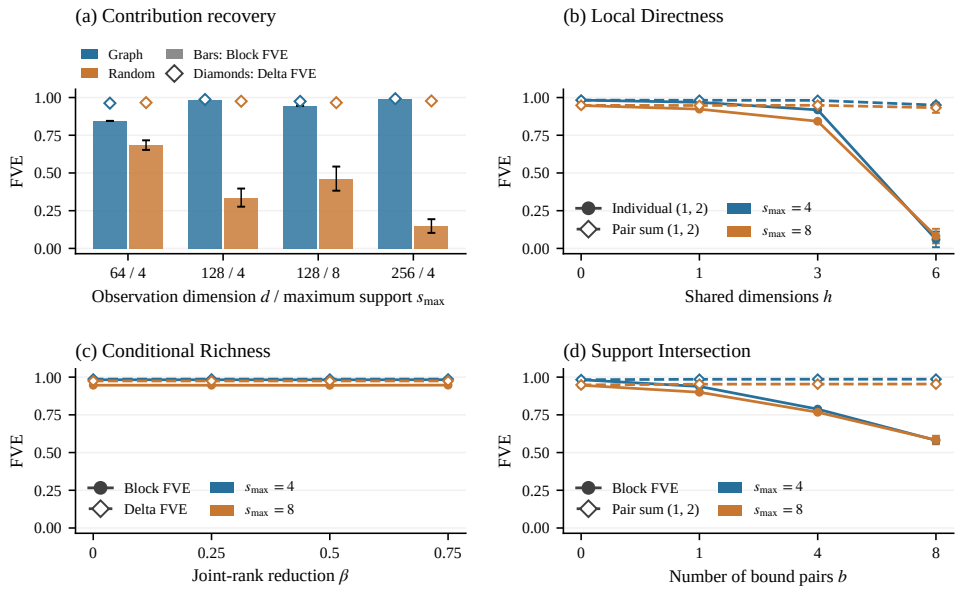

Figure 1: **Contribution recovery under the assumptions and their violations.** (a) Bars show Block FVE; diamonds show Delta FVE. (b)–(d) use the graph protocol at $d = 128$, with color distinguishing the maximum support size $s_{\max}$. In (b), individual recovery averages the per-block FVE of concepts 1 and 2; in (c)–(d), Block FVE includes all concepts. Dashed curves show recovery of the sum of the same pair $(1, 2)$ in (b),(d), and Delta FVE in (c). Intervention levels are displayed at equal spacing; zero denotes the matched baseline.

individual concepts: for always-coactive blocks, their joint contribution is a more reliable target than either block separately. Shared-direction transfers and pairwise rotations provide explicit alternative decompositions that preserve the observed sum.

Conditional Richness is an exception to this deterioration. Restricting the joint change to a proper subspace, while preserving each individual concept's change space, does not systematically reduce recovery over the tested severities (Figure 1(c)). These experiments therefore show relative stability under limited structural departures and poorer individual recovery under stronger overlap or coactivity, rather than a universal failure threshold for every assumption violation.

5.2 Real-World Experiments

We evaluate the proposed ConAn on real LLM representations to examine whether the recovered contributions capture meaningful concept changes, and more importantly, can influence model predictions. Since ground-truth contributions are unavailable in real data, we evaluate semantic readout on StylePTB (Lyu et al., 2021) and interventions on counterfactual text. We compare ConAn with SSAE (Joshi et al., 2026), Top-K SAE (Gao et al., 2025), $L1$ SAE (Huben et al., 2024) and ConCA (Liu et al., 2026a). Experimental details are provided in Appendix E.6.

Semantic Readout. We first examine whether the recovered contributions capture semantic changes between paired contexts. To this end, we evaluate how well the recovered components predict which semantic transformations occur between paired texts, using labels at three levels of granularity: 17 fine-grained factors, eight factor clusters, and four semantic classes. Fig. 2 (a) shows that ConAn achieves higher macro average precision than the baselines across all three levels. Fig. 2 (b) further shows improvements across multiple semantic categories, although their magnitude varies. These results indicate that the recovered contributions contain information about the semantic changes between contexts.

Reconstruction and Intervention. We next examine whether the recovered contributions explain observed representation changes and influence model predictions. Using counterfactual text pairs with changes in object identity, color, material, or size, we evaluate reconstruction error and whether adding a selected contribution shifts the model's prediction toward the target answer. Interventions

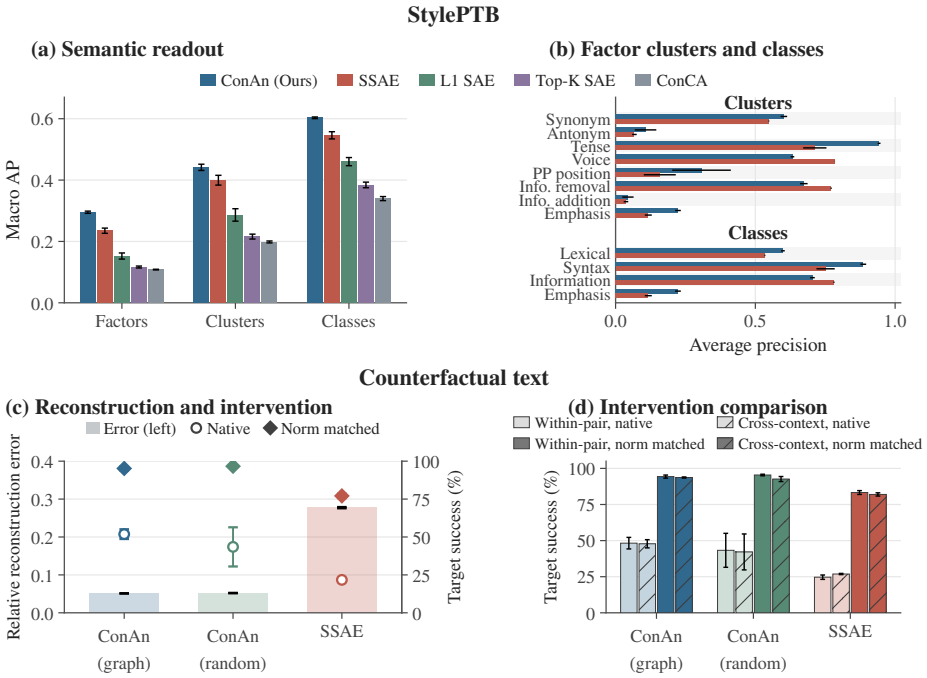

Figure 2: **Semantic readout and counterfactual intervention.** (a) StylePTB macro average precision (AP); (b) cluster- and class-level AP. (c) Counterfactual relative reconstruction error (left axis) and target success at native amplitudes and after rescaling the selected contribution to match the norm of the original representation difference (right axis). (d) Within-pair intervention and cross-context transfer on identical donor-covered recipients, a subset of (c). Error bars show sample standard deviations over three training seeds.

are evaluated at their original magnitudes and after rescaling to match the norm of the observed representation difference. Fig. 2 (c) shows that the proposed ConAn achieves lower reconstruction error and higher intervention success than SSAE. Success increases substantially after norm matching, indicating that the extracted directions can influence target predictions even when their original magnitudes are insufficient. Graph initialization also improves intervention success at native amplitudes, while the two initializations perform similarly after norm matching.

Cross-Context Intervention. Finally, we examine whether contributions extracted from one context pair can produce the same semantic change across contexts. We replace the recipient-pair contribution with one extracted from a different pair expressing the same directed change. Fig. 2 (d) shows that the proposed ConAn largely maintains intervention success after transfer and outperforms SSAE in both within-pair and cross-context settings. These results suggest that the recovered contributions can be reused across contexts rather than being useful only for their source pairs.

6 Discussion

In this work, we introduced a change-centric perspective for analyzing LLM representations, i.e., how concept-level information changes across pairs of related contexts. Our main theoretical result shows that concept change contributions can be identifiable across context pairs. We further developed a theory-guided estimator for recovering concept-specific change contributions from paired LLM representations, which provides a natural tool for understanding how internal LLM representations evolve across changes in input, reasoning or behavior. Beyond interpretation, the recovered concept change spaces may also support downstream applications such as diagnosing representation shifts, evaluating whether an intervention modifies the intended concepts, and developing more targeted concept-level steering methods. Several directions remain open. Our current theory focuses on identifiability, while reliable recovery from finite and noisy LLM representations requires further study. More broadly, validating concept attribution across diverse models and tasks, and investigating whether the recovered concept directions support behaviorally selective interventions, are important directions for future work.

AI USE STATEMENT

Generative AI tools were used only to assist with language editing and presentation. All technical content, theoretical results, experimental design, analysis, and conclusions were developed and verified by the authors.

ETHICS STATEMENT

This work focuses on the analysis and interpretation of large language model representations. We do not identify any specific ethical concerns beyond those generally associated with the analysis and manipulation of model representations.

REPRODUCIBILITY STATEMENT

We provide the theoretical assumptions and proofs, detailed algorithmic procedures, optimization and inference details, and experimental settings in the main paper and appendices. Code and configurations required to reproduce the experiments will be released upon publication.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL https://arxiv.org/abs/1610.01644.

Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Conference on Learning Theory*, pp. 779–806. PMLR, 2014.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying functionally important features with end-to-end sparse dictionary learning. *Advances in Neural Information Processing Systems*, 37:107286–107325, 2024.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL

https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024.

Jingyi Cui, Qi Zhang, Yifei Wang, and Yisen Wang. On the limits of sparse autoencoders: A theoretical framework and reweighted remedy. In *International Conference on Learning Representations*, volume 2026, pp. 156036–156055, 2026.

Seyed Arshan Dalili and Mehrdad Mahdavi. Subspace-aware sparse autoencoders for effective mechanistic interpretability. *arXiv preprint arXiv:2606.06333*, 2026. URL https://arxiv.org/abs/2606.06333v1.

David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5): 2197–2202, 2003.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783v1. Version 1.

Yonina C. Eldar and Moshe Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009. doi: 10.1109/TIT.2009.2030471. URL https://arxiv.org/abs/0807.4581.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013. doi: 10.1109/TPAMI.2013.57. URL https://arxiv.org/abs/1203.1005.

Thomas Fel, Matthew Kowal, Mozes Jacobs, Dron Hazra, Usha Bhalla, Lee Sharkey, Lucius Bushnaq, Satchel Grant, Tal Haklay, Thomas Icard, et al. Structuring sparsity: Block-sparse featurizers capture visual concept manifolds. *arXiv preprint arXiv:2606.25234*, 2026.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

Charles J Garfinkle and Christopher J Hillar. On the uniqueness and stability of dictionaries for sparse representation of noisy signals. *IEEE Transactions on Signal Processing*, 67(23):5884–5892, 2019.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021. URL https://papers.nips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

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

Rémi Gribonval and Karin Schnass. Dictionary identification—sparse matrix-factorization via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

Christopher J Hillar and Friedrich T Sommer. When can dictionary learning uniquely recover sparse data from subsamples? *IEEE Transactions on Information Theory*, 61(11):6290–6297, 2015.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Shruti Joshi, Andrea Dittadi, Sébastien Lachapelle, and Dhanya Sridhar. Sparse shift autoencoders for identifying concepts from large language model activations. *arXiv preprint arXiv:2502.12179*, 2026. URL https://arxiv.org/abs/2502.12179v2. Version 2.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing. *arXiv preprint arXiv:2502.16681*, 2025.

Adam Karvonen, Benjamin Wright, Can Rager, Rico Angell, Jannik Brinkmann, Logan Smith, Claudio Mayrink Verdun, David Bau, and Samuel Marks. Measuring progress in dictionary learning for language model interpretability with board game models. *Advances in Neural Information Processing Systems*, 37:83091–83118, 2024.

Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *International Conference on Machine Learning*, pp. 18171–18206. PMLR, 2023.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.03341.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024.

Yuhang Liu, Erdun Gao, Dong Gong, Anton van den Hengel, and Javen Qinfeng Shi. Concept component analysis: A principled approach for concept extraction in LLMs. *arXiv preprint arXiv:2601.20420*, 2026a. URL https://arxiv.org/abs/2601.20420v2. Version 2.

Yuhang Liu, Dong Gong, Yichao Cai, Erdun Gao, Zhen Zhang, Biwei Huang, Mingming Gong, Anton Hengel, and Javen Qinfeng Shi. I predict therefore i am: Is next token prediction enough to learn human-interpretable concepts from data? In *International Conference on Learning Representations*, 2026b.

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

Yiwei Lyu, Paul Pu Liang, Hai Pham, Eduard Hovy, Barnabás Póczos, Ruslan Salakhutdinov, and Louis-Philippe Morency. StylePTB: A compositional benchmark for fine-grained controllable text style transfer. In *Proceedings of NAACL-HLT*, pp. 2116–2138. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.171. URL https://aclanthology.org/2021.naacl-main.171/.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in LLM representations of true/false datasets. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2310.06824v3.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Anish Mudide, Josh Engels, Eric Michaud, Max Tegmark, and Christian Schroeder de Witt. Efficient dictionary learning with switch sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 101830–101844, 2025.

Walter Nelson, Theofanis Karaletsos, and Francesco Locatello. Toward identifiable sparse autoencoders. *arXiv preprint arXiv:2605.31245*, 2026. URL https://arxiv.org/abs/2605.31245v1.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024. URL https://arxiv.org/abs/2407.02646v4. Version 4, revised 2025.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, Janos Kramar, Rohin Shah, and Neel Nanda. Improving sparse decomposition of language model activations with gated sparse autoencoders. *Advances in Neural Information Processing Systems*, 37: 775–818, 2024a.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024b.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025. URL https://arxiv.org/abs/2501.16496.

Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Conference on Learning Theory*, volume 23, pp. 37.1–37.18. JMLR Workshop and Conference Proceedings, 2012.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. doi: 10.1093/biomet/asv008. URL https://arxiv.org/abs/1405.0680.

Yujia Zheng, Zijian Li, Shunxing Fan, Andrew Gordon Wilson, and Kun Zhang. Diverse dictionary learning. *arXiv preprint arXiv:2604.17568*, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

Appendix

Table of Contents

A Related Work **16**

B Limitations **17**

C Proofs **18**
C.1 Notation and Formal Setup 19
C.2 Lemmas 20
C.3 Proof of Contribution Identifiability 22
C.4 Proof of Contribution Selection 22
C.5 Endpoint Contributions and Anchoring 24
C.6 Quantitative Stability 25
C.7 Representation Error, Noise, and Contamination 28
C.8 Quantitative Recovery of Endpoint Contributions 29

D Method and Implementation Details **30**
D.1 Geometry-Guided Subspace Initialization 30
D.2 Statistical Analysis of Subspace Initialization 33
D.3 Learning Concept Change Contributions 35
D.4 Penalized Training and Schedule 36
D.5 Learning Objective and Optimization 37
D.6 Prediction and Native Sparsity 39
D.7 Sources of Recovery Error and Experimental Evidence 40
D.8 Numerical Environment and Computational Budgets 41

E Data and Evaluation Protocols **41**
E.1 Gaussian DAG Pairs 41
E.2 Configurations, Splits, and Replication 43
E.3 Comparisons and Experimental Matrix 43
E.4 Interventions on the Generating Assumptions 43
E.5 Recovery Metrics 44
E.6 Real-World Experimental Protocols 45

F Additional Experimental Results **47**
F.1 Reference Configurations and Input Features 47
F.2 Penalty Sensitivity 48
F.3 Assumption Interventions and Additional Diagnostics 49
F.4 Numerical Diagnostics and Reproducibility 51

A Related Work

Concept Extraction from LLM Representations. Extracting interpretable concepts from learned representations has been a central objective of representation learning. In language models, linear probing extracts concept-related information through supervised prediction from internal representations (Alain & Bengio, 2016; Belinkov, 2022; Marks & Tegmark, 2024), while sparse autoencoders (SAEs) seek to discover interpretable features through sparse reconstruction (Rajamanoharan et al., 2024a; Gao et al., 2025; Rajamanoharan et al., 2024b; Bricken et al., 2023; Braun et al., 2024; Huben et al., 2024; Mudide et al., 2025; Lieberum et al., 2024; Karvonen et al., 2024; Bussmann et al., 2024). However, high probing accuracy does not necessarily imply faithful recovery of the underlying concept, as probes may instead exploit features that are correlated with the target concept (Hewitt & Liang, 2019; Ravichander et al., 2021). SAEs, on the other hand, draw on sparse dictionary learning to extract interpretable features, where identifiability guarantees have been established under suitable sparsity and dictionary assumptions. However, whether and under what conditions these classical identifiability results extend to SAEs remains insufficiently understood. Consistent with this gap, recent empirical studies have found that SAE-based representations do not consistently outperform simpler non-SAE baselines on downstream probing tasks (Kantamneni et al., 2025). Recent studies have begun to address this theoretical gap. Cui et al. (2026) investigate the conditions under which SAEs recover ground-truth monosemantic features, revealing limitations of standard SAEs under general sparsity conditions. Nelson et al. (2026) further study the instability of learned SAE dictionaries and sparse codes, establishing near-identifiability results under suitable geometric conditions. Moving beyond individual activations, Joshi et al. (2026) introduce Sparse Shift Autoencoders (SSAEs), which exploit sparse changes between paired LLM representations to identify latent concept shifts, even when multiple concepts change simultaneously. Complementary to these sparsity-based approaches, Liu et al. (2026b) establish a theoretical connection between next-token prediction and latent concept log-posteriors, motivating Concept Component Analysis (ConCA), which seeks to recover concept-posterior components from LLM representations (Liu et al., 2026a). However, despite its theoretical grounding, ConCA remains an ill-posed inverse problem, with the identifiability of the extracted concept components left unresolved. Building on this posterior representation framework, our work shifts the identification target from concept-posterior components to their contributions to representation changes, establishing conditions under which individual concept change contributions can be uniquely recovered from paired contexts.

Identifiability from Paired Observations. Paired observations have been widely exploited to improve identifiability in representation learning by revealing shared and varying latent factors across observations. Locatello et al. (2020) establish identifiability of disentangled representations from pairs of observations sharing subsets of latent factors, without requiring knowledge of which factors remain unchanged. Related multi-view identifiability results exploit shared latent sources across observations (Gresele et al., 2020), while content-preserving augmentations enable identification of an invariant content block up to an invertible transformation (Von Kügelgen et al., 2021). Closely related to our setting, Joshi et al. (2026) introduce Sparse Shift Autoencoders (SSAEs), which exploit sparse activation differences between paired LLM contexts to identify latent concept shifts, even when multiple unknown concepts change simultaneously. Unlike these approaches, our work builds on the concept log-posterior characterization of LLM representations and investigates the identifiability of individual concept change contributions. By exploiting structured variation across paired contexts, we establish conditions under which these contributions can be uniquely identified, even when multiple concepts change simultaneously.

Identifiability in Sparse Dictionary Learning. Identifiability has also been extensively studied in dictionary learning and sparse representation learning, where recovery depends on the geometry of dictionary atoms and the diversity of their activation patterns (Donoho & Elad, 2003; Gribonval & Schnass, 2010; Spielman et al., 2012; Arora et al., 2014; Hillar & Sommer, 2015; Garfinkle & Hillar, 2019). Structured sparsity and task diversity have been exploited to establish identifiability of latent representations in multi-task learning (Lachapelle et al., 2023). More recently, Diverse Dictionary Learning investigates the recovery of latent subsets through set operations and establishes element-wise identifiability under sufficient structural diversity (Zheng et al., 2026). Related developments in sparse autoencoders also investigate the identifiability of learned dictionaries and the representation of multidimensional features through subspaces or blocks (Nelson et al., 2026; Dalili & Mahdavi,

2026; Fel et al., 2026). Unlike these approaches, our work investigates the identifiability of concept-specific contributions to representation changes by exploiting structured sparsity and variation across paired contexts.

B LIMITATIONS

Our identifiability results rely on several structural and distributional assumptions that characterize when individual concept change contributions can be uniquely recovered. Local Directness requires concept change spaces to be linearly independent within the union of any two maximal change supports. This condition does not require global independence or orthogonality across all concept spaces, allowing globally overcomplete representations. Nevertheless, substantial overlap between concept spaces may prevent their individual contributions from being distinguished. Conditional Richness requires sufficient variation within each change pattern to reveal its corresponding combined change space. While this condition formalizes the information needed for subspace recovery, it may be restrictive for real LLM representations with limited or highly structured context variations. Support Intersection requires sufficiently diverse change patterns to isolate each concept through their overlaps, without requiring single-concept changes. When certain concepts consistently change together, their individual contributions may remain ambiguous, even if their combined contribution can be recovered.

Our main identifiability results further assume an exact concept-wise decomposition of representation changes, whereas the underlying concept log-posterior characterization of LLM representations generally holds only approximately. Although our stability analysis extends the theoretical guarantees to small perturbations under quantitative conditions, finite-sample recovery and optimization guarantees for the proposed estimator remain open. These limitations motivate further investigation of concept attribution under realistic LLM representation errors and more general patterns of concept changes. Nevertheless, these assumptions provide sufficient conditions for identifiability rather than strict requirements for successful recovery in practice. Our experiments demonstrate a degree of robustness to violations of these assumptions, with concept contributions remaining recoverable under moderate structural deviations, although more severe violations may compromise individual attribution.

C PROOFS

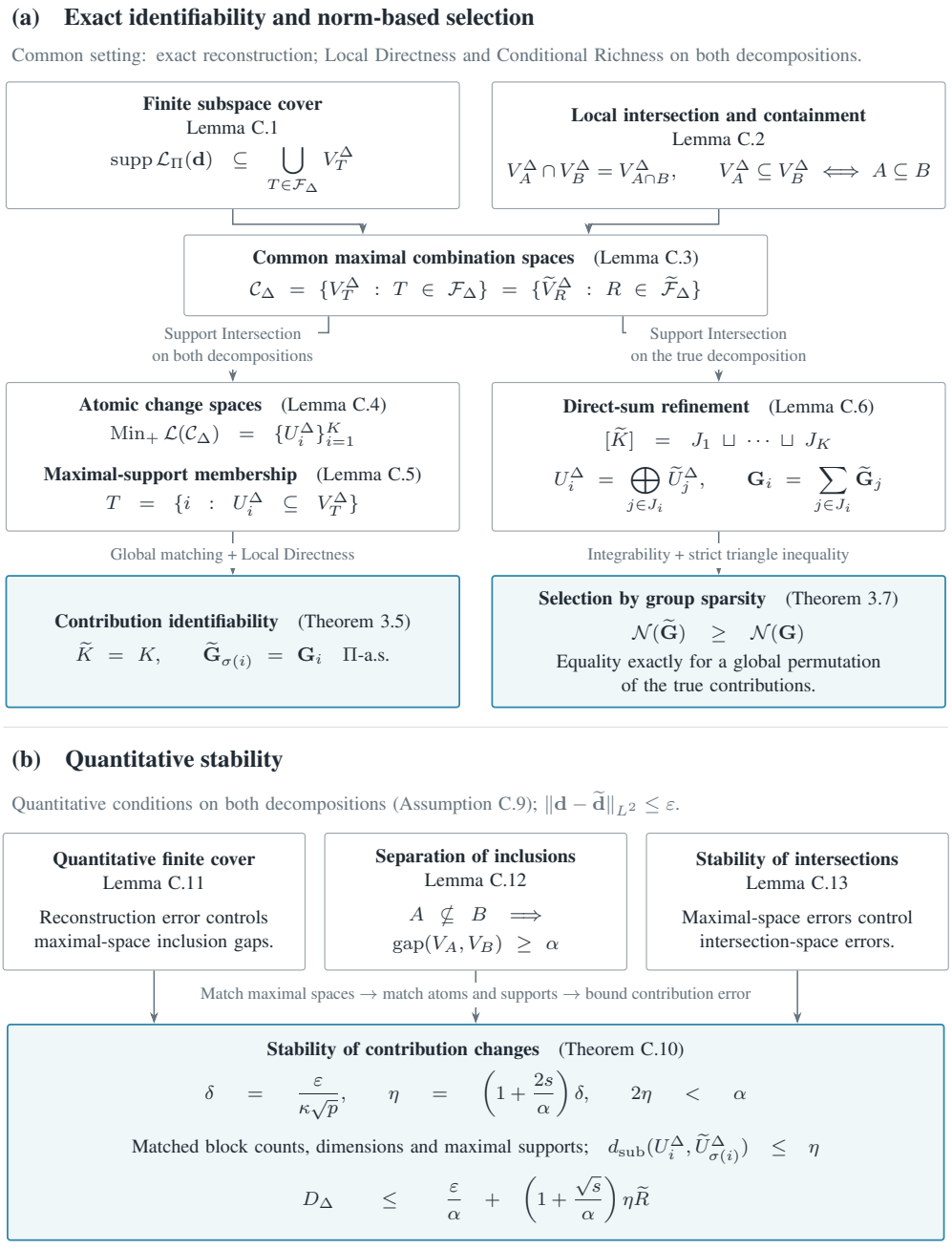

Figure 3: **Proof roadmap.** The geometric lemmas establish contribution identifiability and norm-based selection in (a), and their quantitative counterparts yield stability bounds in (b). Arrows indicate the main logical dependencies.

Table 1: Notation for the exact contribution-identifiability proofs.

Notation	Meaning
Observed representations and changes	
$\mathbf{f}, \mathbf{C}_i$	Representation and endpoint contribution of concept i.
$\Pi, \bar{\mu}$	Pairing law and averaged endpoint marginal.
$\omega, \mathbf{d}$	Context pair and its observed representation change.
$\mathbf{G}_i$	True change contribution of concept i.
Change spaces and support patterns	
U_i^Δ, r_i	Change space of concept i and its dimension.
V_T^Δ	Combined change space for concept subset T.
S_Δ	Active concept indices for a context pair.
$\mathscr{S}_\Delta, \mathcal{F}_\Delta$	Positive-probability nonempty supports and their inclusion-maximal members.
Competing decomposition	
$\widetilde{K}, \widetilde{\mathbf{G}}_j$	Candidate block count and contribution.
$\widetilde{U}_j^\Delta, \widetilde{V}_R^\Delta$	Candidate change space and combined change space.
$\widetilde{S}_\Delta, \widetilde{\mathscr{S}}_\Delta, \widetilde{\mathcal{F}}_\Delta$	Candidate support and support families.

C.1 Notation and Formal Setup

We use Euclidean norms for vectors and operator norms for matrices unless otherwise indicated. The notation $\|\cdot\|_F$ denotes the Frobenius norm.

Let $K \geq 1$, let $(\mathcal{X}, \mu)$ be the input probability space, and $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$ a measurable representation. For each $i \in [K]$, write $n_i = |\mathcal{Z}_i|$ and $\mathbf{p}_i(\mathbf{x}) := [p(z_i \mid \mathbf{x})]_{z_i \in \mathcal{Z}_i}$. These measurable posterior maps take values in $\Delta^\circ_{n_i-1} = \{\mathbf{q} \in \mathbb{R}^{n_i}_{>0} : \mathbf{1}^\top \mathbf{q} = 1\}$. With $\mathbf{A}_i \in \mathbb{R}^{d \times n_i}$ and $\mathbf{b} \in \mathbb{R}^d$, the exact model used in Section 2.2 is

$$\mathbf{f}(\mathbf{x}) = \mathbf{b} + \sum_{i=1}^{K} \mathbf{C}_i(\mathbf{x}), \qquad \mathbf{C}_i(\mathbf{x}) = \mathbf{A}_i \log \mathbf{p}_i(\mathbf{x}) \quad \mu\text{-a.e.}, \tag{8}$$

where the logarithm acts componentwise. This is the zero-residual specialization of Eq. (2). The underlying ConCA result uses a softmax predictor that matches the conditional token law, output-embedding diversity, vanishing conditional entropy $H(\mathbf{z} \mid \mathbf{x})$, and a total-variation regularity condition on $p(\mathbf{z} \mid y)$ (Liu et al., 2026a, Theorem 2.2).

The pairing law Π is a probability measure on $\mathcal{X} \times \mathcal{X}$, with marginals $\Pi_1, \Pi_2 \ll \mu$. The absolute-continuity condition transfers the exact representation model to the paired-context distribution. Put $\bar{\mu} = (\Pi_1 + \Pi_2)/2$ and $\omega = (\mathbf{x}, \mathbf{x}')$. Eq. (3) gives $\mathbf{d} = \sum_i \mathbf{G}_i$ on this pair space. More generally, the geometric proof below applies to any finite family $\mathbf{G}_i \in L^0(\Pi; \mathbb{R}^d)$ satisfying that identity, with $\Pi(\mathbf{G}_i \neq \mathbf{0}) > 0$ for every i. Here L^0 denotes measurable functions modulo almost-sure equality.

For $\mathcal{L}_\Pi$ the distribution under Π and supp its topological support, define

$$U_i^\Delta = \operatorname{span} \operatorname{supp} \mathcal{L}_\Pi(\mathbf{G}_i), \qquad r_i = \dim U_i^\Delta > 0, \qquad S_\Delta(\omega) = \{i \in [K] : \mathbf{G}_i(\omega) \neq \mathbf{0}\}. \tag{9}$$

This is the smallest linear subspace containing $\mathbf{G}_i$ almost surely: any such subspace is closed and contains the support of its law. Write $V_T^\Delta = \sum_{i \in T} U_i^\Delta$, with $V_\varnothing^\Delta = \{\mathbf{0}\}$, and set

$$\mathscr{S}_\Delta = \{T \subseteq [K] : T \neq \varnothing,\ \Pi(S_\Delta = T) > 0\}, \qquad \mathcal{F}_\Delta = \operatorname{Max}_\subseteq \mathscr{S}_\Delta. \tag{10}$$

Empty change supports are allowed. The competing family has $\widetilde{K} \geq 1$, $\widetilde{\mathbf{G}}_j \in L^0(\Pi; \mathbb{R}^d)$, $\mathbf{d} = \sum_j \widetilde{\mathbf{G}}_j$ almost surely, and $\Pi(\widetilde{\mathbf{G}}_j \neq \mathbf{0}) > 0$ for every j. Its objects $\widetilde{U}_j^\Delta$, $\widetilde{S}_\Delta$, $\widetilde{\mathscr{S}}_\Delta$, $\widetilde{\mathcal{F}}_\Delta$, and $\widetilde{V}_R^\Delta$ are defined in the same way.

The main notation is summarized in Table 1. The table is a guide to the objects defined above; the definitions in Equations (9) and (10) are used throughout the proofs.

For use in the proofs, put $Y_T^\Delta = \operatorname{supp} \mathcal{L}_\Pi(\mathbf{d} \mid S_\Delta = T)$ for $T \in \mathcal{F}_\Delta$. Assumption 3.3 is equivalent to the following finite-cover implication, for every $q \geq 1$ and linear subspaces $W_1, \ldots, W_q \subseteq \mathbb{R}^d$:

$$Y_T^\Delta \subseteq \bigcup_{a=1}^{q} W_a \quad \Longrightarrow \quad \exists a \in [q] : \; V_T^\Delta \subseteq W_a. \tag{11}$$

Indeed, $Y_T^\Delta \subseteq V_T^\Delta$, so if no W_a contains V_T^Δ, the intersections $W_a \cap V_T^\Delta$ are proper subspaces covering Y_T^Δ. Conversely, a finite cover by proper subspaces of V_T^Δ violates the implication. A finite union of linear subspaces is closed, so it contains the conditional support exactly when it has conditional probability one. The same statements apply to $\widetilde{Y}_R^\Delta$ and $\widetilde{V}_R^\Delta$.

C.2 Lemmas

Throughout the exact-reconstruction arguments, $\mathbf{G}$ and $\widetilde{\mathbf{G}}$ are two decompositions of the same $\mathbf{d}$ as defined in Section 2.2. Every contribution belongs to its actual change space almost surely. The nonzero-space condition gives $\Pi(\mathbf{G}_i \neq \mathbf{0}) > 0$ for every i. Since the support takes finitely many values, each index belongs to a member of $\mathscr{S}_\Delta$ and hence to a member of $\mathcal{F}_\Delta$. In particular, $\mathcal{F}_\Delta \neq \varnothing$. The same facts hold for the competing decomposition.

The proof proceeds from the observed changes to progressively finer geometric structure. Lemmas C.1 and C.2 relate the observations and support patterns to their combined change spaces. Lemma C.3 then shows that admissible decompositions share the same maximal combined spaces. Finally, Lemmas C.4 and C.5 recover the individual concept change spaces and their maximal-support memberships, which are used to prove contribution identifiability in Appendix C.3. See Figure 3 for a schematic overview of this proof structure.

Lemma C.1 (Finite subspace cover)**.** *The observation law satisfies*

$$\operatorname{supp} \mathcal{L}_\Pi(\mathbf{d}) \subseteq \bigcup_{T \in \mathcal{F}_\Delta} V_T^\Delta, \qquad \operatorname{supp} \mathcal{L}_\Pi(\mathbf{d}) \subseteq \bigcup_{R \in \widetilde{\mathcal{F}}_\Delta} \widetilde{V}_R^\Delta. \tag{12}$$

Each inclusion also holds for the conditional support of $\mathbf{d}$ *given any event of positive probability.*

Proof. Outside a null set, either $S_\Delta = \varnothing$ and $\mathbf{d} = \mathbf{0}$, or $S_\Delta = A \in \mathscr{S}_\Delta$ and

$$\mathbf{d} = \sum_{i \in A} \mathbf{G}_i \in V_A^\Delta \subseteq V_T^\Delta \quad \text{for some } T \in \mathcal{F}_\Delta.$$

Thus $\mathbf{d}$ lies almost surely in $\mathcal{U} = \bigcup_{T \in \mathcal{F}_\Delta} V_T^\Delta$. This finite union of closed subspaces contains the support of the law of $\mathbf{d}$. For any event E with $\Pi(E) > 0$, $\Pi(\mathbf{d} \in \mathcal{U} \mid E) = 1$, giving the conditional assertion. Apply the same argument to $\widetilde{\mathbf{G}}$ for the second inclusion. □

Lemma C.2 (Local intersection and containment)**.** *Under Assumption 3.2, for $A \subseteq T$ and $B \subseteq T'$ with $T, T' \in \mathcal{F}_\Delta$,*

$$V_A^\Delta \subseteq V_B^\Delta \iff A \subseteq B, \qquad V_A^\Delta \cap V_B^\Delta = V_{A \cap B}^\Delta. \tag{13}$$

Consequently, for every nonempty $\mathcal{G} \subseteq \mathcal{F}_\Delta$,

$$\bigcap_{T \in \mathcal{G}} V_T^\Delta = V_{\bigcap_{T \in \mathcal{G}} T}^\Delta. \tag{14}$$

Both statements hold for the competing decomposition.

Proof. The family indexed by $A \cup B \subseteq T \cup T'$ forms a direct sum. If $A \subseteq B$, then $V_A^\Delta \subseteq V_B^\Delta$. Conversely, suppose $V_A^\Delta \subseteq V_B^\Delta$ and choose $i \in A \setminus B$. A nonzero $\mathbf{u} \in U_i^\Delta$ then satisfies

$$\mathbf{u} = \sum_{j \in B} \mathbf{v}_j, \qquad \mathbf{v}_j \in U_j^\Delta,$$

contradicting directness on $A \cup B$. Hence $A \subseteq B$.

If $\mathbf{v} \in V_A^\Delta \cap V_B^\Delta$, write $\mathbf{v} = \sum_{i\in A} \mathbf{u}_i = \sum_{i\in B} \mathbf{v}_i$ with $\mathbf{u}_i, \mathbf{v}_i \in U_i^\Delta$. Directness implies $\mathbf{u}_i = \mathbf{0}$ on $A \setminus B$, $\mathbf{v}_i = \mathbf{0}$ on $B \setminus A$, and $\mathbf{u}_i = \mathbf{v}_i$ on $A \cap B$. Thus $\mathbf{v} \in V_{A\cap B}^\Delta$; the reverse inclusion follows from the definition of the spaces.

For Equation (14), enumerate $\mathcal{G} = \{T_1, \dots, T_m\}$ and apply the intersection identity successively. Since $T_1 \cap \cdots \cap T_q \subseteq T_1$, the next intersection uses only directness on $T_1 \cup T_{q+1}$. □

Lemma C.3 (Common maximal combination spaces)**.** *Suppose both the true and competing decompositions satisfy Assumptions 3.2 and 3.3. Then*

$$\mathcal{C}_\Delta := \{V_T^\Delta : T \in \mathcal{F}_\Delta\} = \{\widetilde{V}_R^\Delta : R \in \widetilde{\mathcal{F}}_\Delta\}. \tag{15}$$

Each space in this family has a unique maximal support on each side.

Proof. For $T \in \mathcal{F}_\Delta$, Lemma C.1 gives

$$Y_T^\Delta \subseteq \bigcup_{R\in\widetilde{\mathcal{F}}_\Delta} \widetilde{V}_R^\Delta.$$

Assumption 3.3 yields $R \in \widetilde{\mathcal{F}}_\Delta$ with $V_T^\Delta \subseteq \widetilde{V}_R^\Delta$. Applying the same argument to R yields $T' \in \mathcal{F}_\Delta$ with

$$V_T^\Delta \subseteq \widetilde{V}_R^\Delta \subseteq V_{T'}^\Delta.$$

Lemma C.2 gives $T \subseteq T'$. Both sets are maximal in $\mathscr{S}_\Delta$, so $T = T'$ and $V_T^\Delta = \widetilde{V}_R^\Delta$. Interchanging the decompositions proves Equation (15). Finally, Equation (13) implies that two maximal supports generating the same space are equal. □

For a finite family $\mathcal{C}$ of subspaces, define

$$\mathcal{L}(\mathcal{C}) := \left\{ \bigcap_{V\in\mathcal{H}} V : \varnothing \neq \mathcal{H} \subseteq \mathcal{C} \right\}.$$

Write $\mathrm{Min}_+ \mathcal{L}(\mathcal{C})$ for its minimal nonzero elements under subspace inclusion.

Lemma C.4 (Recovery of atomic change spaces)**.** *Under Assumptions 3.2 and 3.4,*

$$\mathrm{Min}_+ \mathcal{L}\big(\{V_T^\Delta : T \in \mathcal{F}_\Delta\}\big) = \{U_i^\Delta : i \in [K]\}. \tag{16}$$

The corresponding identity holds for the competing decomposition.

Proof. For every i, Assumption 3.4 and Lemma C.2 give a nonempty family $\mathcal{G}_i \subseteq \mathcal{F}_\Delta$ with

$$\bigcap_{T\in\mathcal{G}_i} V_T^\Delta = V_{\bigcap_{T\in\mathcal{G}_i} T}^\Delta = V_{\{i\}}^\Delta = U_i^\Delta.$$

Thus every atomic space is a nonzero member of the intersection family.

Every member W of that family has the form $W = V_A^\Delta$, where A is an intersection of a nonempty family of maximal supports. If $W \neq \{\mathbf{0}\}$, then $A \neq \varnothing$. When $|A| > 1$, take $i \in A$ and a maximal support containing A. Directness within this support gives $\{\mathbf{0}\} \subsetneq U_i^\Delta \subsetneq W$; hence W is not minimal. If $A = \{i\}$ and another nonzero intersection V_B^Δ lies in U_i^Δ, Equation (13) gives $B \subseteq \{i\}$. The set B is nonempty, so $B = \{i\}$ and $V_B^\Delta = U_i^\Delta$. Therefore U_i^Δ is minimal. □

Lemma C.5 (Recovery of maximal support membership)**.** *Under Assumption 3.2,*

$$T = \{i \in [K] : U_i^\Delta \subseteq V_T^\Delta\} \qquad (T \in \mathcal{F}_\Delta). \tag{17}$$

The same identity holds for the competing decomposition.

Proof. For each i, choose $T_i \in \mathcal{F}_\Delta$ containing i. Apply Equation (13) to $\{i\} \subseteq T_i$ and $T \subseteq T$ to obtain $U_i^\Delta \subseteq V_T^\Delta \iff i \in T$. □

C.3 Proof of Contribution Identifiability

We now combine the preceding lemmas to identify the individual change contributions from the common maximal change spaces.

Proof of Theorem 3.5. Step 1: maximal combination spaces. Lemma C.3 identifies the common family

$$\mathcal{C}_\Delta = \{V_T^\Delta : T \in \mathcal{F}_\Delta\} = \{\widetilde{V}_R^\Delta : R \in \widetilde{\mathcal{F}}_\Delta\}.$$

Step 2: atomic change spaces. Applying Lemma C.4 to both decompositions gives

$$\{U_i^\Delta : i \in [K]\} = \mathrm{Min}_+ \mathcal{L}(\mathcal{C}_\Delta) = \{\widetilde{U}_j^\Delta : j \in [\widetilde{K}]\}.$$

Within either family the spaces are distinct. Indeed, for $i \neq k$, choose maximal supports T_i, T_k containing the respective indices. Directness on $T_i \cup T_k$ gives $U_i^\Delta \cap U_k^\Delta = \{\mathbf{0}\}$, and both spaces are nonzero. It follows that $\widetilde{K} = K$ and there is a permutation $\sigma : [K] \to [K]$ such that $\widetilde{U}_{\sigma(i)}^\Delta = U_i^\Delta$ for every i.

Step 3: maximal supports. For $R \in \widetilde{\mathcal{F}}_\Delta$, let $T \in \mathcal{F}_\Delta$ be the unique support with $\widetilde{V}_R^\Delta = V_T^\Delta$. Lemma C.5 gives

$$i \in T \iff U_i^\Delta \subseteq V_T^\Delta \iff \widetilde{U}_{\sigma(i)}^\Delta \subseteq \widetilde{V}_R^\Delta \iff \sigma(i) \in R.$$

Hence $R = \sigma(T)$ and $\widetilde{\mathcal{F}}_\Delta = \{\sigma(T) : T \in \mathcal{F}_\Delta\}$.

Step 4: individual changes. Fix a pair ω in the common full-probability set where reconstruction holds, all contributions belong to their actual spaces, and every nonempty support belongs to its positive-probability support family. Set

$$A = S_\Delta(\omega), \qquad B = \sigma^{-1}\big(\widetilde{S}_\Delta(\omega)\big).$$

There are $T, T' \in \mathcal{F}_\Delta$ with $A \subseteq T$ and $B \subseteq T'$: use maximality for the true support and Step 3 for the competing support. An empty support is contained in any maximal support. Exact reconstruction now gives

$$\sum_{i \in A \cup B} \big(\mathbf{G}_i(\omega) - \widetilde{\mathbf{G}}_{\sigma(i)}(\omega)\big) = \mathbf{0}, \qquad \mathbf{G}_i(\omega) - \widetilde{\mathbf{G}}_{\sigma(i)}(\omega) \in U_i^\Delta.$$

Since $A \cup B \subseteq T \cup T'$, local directness makes every summand zero. Contributions outside $A \cup B$ vanish by definition. Thus $\widetilde{\mathbf{G}}_{\sigma(i)} = \mathbf{G}_i$ for every i almost surely. □

C.4 Proof of Contribution Selection

Unlike Theorem 3.5, the selection theorem assumes support intersection only for the true decomposition. Lemmas C.1, C.2, and C.3 do not require candidate support intersection and therefore remain applicable. The key step is to show that every admissible candidate is a direct-sum refinement of the true decomposition; the expected contribution norm then distinguishes the true blocks from nontrivial refinements.

Lemma C.6 (Refinement without candidate support intersection)**.** *Suppose the true decomposition satisfies Assumptions 3.2–3.4, and let $\widetilde{\mathbf{G}} \in \mathcal{M}_{\mathrm{sel}}$. There is a partition into nonempty sets $[\widetilde{K}] = J_1 \sqcup \cdots \sqcup J_K$ such that*

$$U_i^\Delta = \bigoplus_{j \in J_i} \widetilde{U}_j^\Delta, \qquad \mathbf{G}_i = \sum_{j \in J_i} \widetilde{\mathbf{G}}_j \quad \Pi\text{-almost surely}. \tag{18}$$

Moreover, the maximal supports admit a bijection $\rho : \mathcal{F}_\Delta \to \widetilde{\mathcal{F}}_\Delta$ with

$$V_T^\Delta = \widetilde{V}_{\rho(T)}^\Delta, \qquad \rho(T) = \bigsqcup_{i \in T} J_i \quad (T \in \mathcal{F}_\Delta). \tag{19}$$

The partition is uniquely determined by $J_i = \{j : \widetilde{U}_j^\Delta \subseteq U_i^\Delta\}$.

Proof. Step 1: matching maximal combination spaces. Both decompositions satisfy local directness and conditional richness and have the same sum. Lemma C.3 therefore gives a bijection $\rho : \mathcal{F}_\Delta \to \widetilde{\mathcal{F}}_\Delta$ such that $V_T^\Delta = \widetilde{V}_{\rho(T)}^\Delta$ for every $T \in \mathcal{F}_\Delta$.

Step 2: recovering true spaces as sums of candidate spaces. For each $i \in [K]$, choose a nonempty witness family $\mathcal{G}_i \subseteq \mathcal{F}_\Delta$ with $\bigcap_{T\in\mathcal{G}_i} T = \{i\}$, as supplied by true support intersection, and put $J_i = \bigcap_{T\in\mathcal{G}_i} \rho(T)$. Applying Lemma C.2 on both sides yields

$$U_i^\Delta = \bigcap_{T\in\mathcal{G}_i} V_T^\Delta = \bigcap_{T\in\mathcal{G}_i} \widetilde{V}_{\rho(T)}^\Delta = \widetilde{V}_{J_i}^\Delta. \tag{20}$$

Since $U_i^\Delta \neq \{\mathbf{0}\}$, the set J_i is nonempty. It is contained in a candidate maximal support, so candidate local directness makes the sum in Equation (20) direct. For distinct i, k, choose true maximal supports containing the respective indices. True local directness on their union gives $U_i^\Delta \cap U_k^\Delta = \{\mathbf{0}\}$. Every candidate space is nonzero, so the sets J_i are pairwise disjoint.

Step 3: covering all candidate blocks. Fix $T \in \mathcal{F}_\Delta$ and $i \in T$. Then $\widetilde{V}_{J_i}^\Delta = U_i^\Delta \subseteq V_T^\Delta = \widetilde{V}_{\rho(T)}^\Delta$. The candidate containment identity in Lemma C.2 applies because J_i is contained in a candidate maximal support; hence $J_i \subseteq \rho(T)$. Consequently, for $L_T = \bigcup_{i\in T} J_i$,

$$L_T \subseteq \rho(T), \qquad \widetilde{V}_{L_T}^\Delta = \sum_{i\in T} U_i^\Delta = V_T^\Delta = \widetilde{V}_{\rho(T)}^\Delta.$$

The same containment identity gives $L_T = \rho(T)$, proving the second identity in Equation (19). Every candidate block belongs to a candidate maximal support, and ρ is onto, so the disjoint sets J_i cover $[\widetilde{K}]$. If $\widetilde{U}_j^\Delta \subseteq U_i^\Delta$ and $j \in J_k$, then $\widetilde{U}_j^\Delta \subseteq U_i^\Delta \cap U_k^\Delta$; nonzero candidate spaces force $i = k$. This proves the stated intrinsic description of J_i and the uniqueness of the partition.

Step 4: recovering the grouped contributions. Set $\mathbf{H}_i = \sum_{j\in J_i} \widetilde{\mathbf{G}}_j \in U_i^\Delta$. Work on the common full-probability set where exact reconstruction holds, all contributions lie in their actual spaces, and every nonempty actual support has positive probability. For each such pair, choose $T, T' \in \mathcal{F}_\Delta$ with $S_\Delta \subseteq T$ and $\widetilde{S}_\Delta \subseteq \rho(T')$. These choices exist by maximality and the bijection ρ; an empty support can be included in any maximal support. Equation (19) implies $\mathbf{H}_i = \mathbf{0}$ for $i \notin T'$. Exact reconstruction gives

$$\sum_{i\in T\cup T'} (\mathbf{G}_i - \mathbf{H}_i) = \mathbf{0}, \qquad \mathbf{G}_i - \mathbf{H}_i \in U_i^\Delta.$$

True local directness makes every summand zero. Outside $T \cup T'$, both contributions vanish. This proves the contribution identity in Equation (18). □

Proof of Theorem 3.7. Fix any $\widetilde{\mathbf{G}} \in \mathcal{M}_{\text{sel}}$ and use the partition supplied by Lemma C.6. For $j \in J_i$, let $P_{ij} : U_i^\Delta \to \widetilde{U}_j^\Delta$ be the linear coordinate projection associated with the direct sum in Equation (18). Uniqueness of coordinates in that direct sum gives $\widetilde{\mathbf{G}}_j = P_{ij}\mathbf{G}_i$ almost surely. The spaces are finite-dimensional, so each P_{ij} is bounded and

$$\mathbb{E}_\Pi\|\widetilde{\mathbf{G}}_j\|_2 \leq \|P_{ij}\|_{\text{op}}\, \mathbb{E}_\Pi\|\mathbf{G}_i\|_2 < \infty.$$

Thus every admissible candidate has finite expected contribution norm; no separate candidate moment assumption is needed.

For each i, the Euclidean triangle inequality gives

$$\|\mathbf{G}_i\|_2 = \left\|\sum_{j\in J_i} \widetilde{\mathbf{G}}_j\right\|_2 \leq \sum_{j\in J_i} \|\widetilde{\mathbf{G}}_j\|_2 \quad \Pi\text{-almost surely.} \tag{21}$$

Summing over i and taking expectations yields $\mathcal{N}(\mathbf{G}) \leq \mathcal{N}(\widetilde{\mathbf{G}})$.

Suppose some J_i has at least two elements. Choose $T \in \mathcal{F}_\Delta$ containing i. By Equation (19), $\rho(T)$ is a candidate maximal support containing J_i, so the event $E = \{\widetilde{S}_\Delta = \rho(T)\}$ has positive Π-probability. On E, every $\widetilde{\mathbf{G}}_j$ with $j \in J_i$ is nonzero. Two such vectors cannot be collinear: they

belong to distinct summands of a direct sum, whose intersection is $\{\mathbf{0}\}$. Equality in the Euclidean triangle inequality requires all nonzero summands to be nonnegative multiples of a common vector. Hence Equation (21) is strict on E for this i. Its gap is nonnegative everywhere and integrable, and is positive on an event of positive probability. Therefore $\mathcal{N}(\widetilde{\mathbf{G}}) > \mathcal{N}(\mathbf{G})$.

It follows that equality of the expected norms forces every J_i to be a singleton. Since the J_i partition $[\widetilde{K}]$, we obtain $\widetilde{K} = K$ and a permutation σ with $J_i = \{\sigma(i)\}$; Equation (18) then gives $\widetilde{\mathbf{G}}_{\sigma(i)} = \mathbf{G}_i$ almost surely. Conversely, a global permutation leaves $\mathcal{N}$ unchanged. The true decomposition belongs to $\mathcal{M}_{\mathrm{sel}}$ and has finite norm, so it attains the lower bound. This proves both existence of a minimum and the characterization of all minimizers. □

C.5 Endpoint Contributions and Anchoring

For the endpoint measure $\bar{\mu}$ defined in Appendix C.1, let

$$D_\Pi : L^0(\bar{\mu}; \mathbb{R}^d) \to L^0(\Pi; \mathbb{R}^d), \qquad (D_\Pi \mathbf{h})(\mathbf{x}, \mathbf{x}') = \mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{x}'). \tag{22}$$

The same notation denotes the scalar difference operator. Since $\Pi_1, \Pi_2 \ll \bar{\mu}$, this map is well-defined on almost-everywhere equivalence classes. We call the pairing law Π *essentially connected* if

$$\ker D_\Pi = \{[\mathbf{x} \mapsto \mathbf{a}]_{\bar{\mu}} : \mathbf{a} \in \mathbb{R}^d\}. \tag{23}$$

For finite support, a function with zero change is constant on each connected component of the graph of positive-probability pairs. Thus this condition is equivalent to connectivity of that graph.

Corollary C.7 (Recovery of single-context contributions)**.** *Under the conditions of Theorem 3.5, suppose the pairing distribution is essentially connected and*

$$\widetilde{\mathbf{G}}_j(\mathbf{x}, \mathbf{x}') = \widetilde{\mathbf{C}}_j(\mathbf{x}) - \widetilde{\mathbf{C}}_j(\mathbf{x}').$$

Then, under the same global permutation σ, there exist constants $\mathbf{a}_k \in \mathbb{R}^d$ such that

$$\widetilde{\mathbf{C}}_{\sigma(k)}(\mathbf{x}) = \mathbf{C}_k(\mathbf{x}) + \mathbf{a}_k, \qquad k \in [K], \tag{24}$$

almost everywhere under the endpoint distributions.

Proof of Corollary C.7. For the permutation from Theorem 3.5, put $\mathbf{h}_i = \widetilde{\mathbf{C}}_{\sigma(i)} - \mathbf{C}_i$. Equality of the changes gives

$$(D_\Pi \mathbf{h}_i)(\mathbf{x}, \mathbf{x}') = \widetilde{\mathbf{G}}_{\sigma(i)}(\mathbf{x}, \mathbf{x}') - \mathbf{G}_i(\mathbf{x}, \mathbf{x}') = \mathbf{0} \quad \Pi\text{-almost surely.}$$

Equation (23) yields $\mathbf{h}_i = \mathbf{a}_i$ $\bar{\mu}$-almost everywhere for some $\mathbf{a}_i \in \mathbb{R}^d$. If both endpoint models reconstruct $\mathbf{f}$ with constant biases, substitution into $\mathbf{b} + \sum_i \mathbf{C}_i = \widetilde{\mathbf{b}} + \sum_i \widetilde{\mathbf{C}}_{\sigma(i)}$ also gives $\widetilde{\mathbf{b}} = \mathbf{b} - \sum_i \mathbf{a}_i$. □

A general centering rule acts on a translation-closed subset $\mathcal{H} \subseteq L^0(\bar{\mu}; \mathbb{R}^d)$ through a map $\boldsymbol{\alpha} : \mathcal{H} \to \mathbb{R}^d$ satisfying $\boldsymbol{\alpha}(\mathbf{h} + \mathbf{a}) = \boldsymbol{\alpha}(\mathbf{h}) + \mathbf{a}$ for every $\mathbf{h} \in \mathcal{H}$ and $\mathbf{a} \in \mathbb{R}^d$.

Corollary C.8 (Recovery after centering)**.** *Under Corollary C.7, if $\mathbf{C}_i, \widetilde{\mathbf{C}}_{\sigma(i)} \in \mathcal{H}$, then*

$$\widetilde{\mathbf{C}}_{\sigma(i)} - \boldsymbol{\alpha}(\widetilde{\mathbf{C}}_{\sigma(i)}) = \mathbf{C}_i - \boldsymbol{\alpha}(\mathbf{C}_i) \quad \bar{\mu}\text{-a.e.}, \qquad i \in [K]. \tag{25}$$

Proof of Corollary C.8. The map $\boldsymbol{\alpha}$ acts on $\bar{\mu}$-almost-everywhere equivalence classes. Corollary C.7 and translation equivariance give

$$\boldsymbol{\alpha}(\widetilde{\mathbf{C}}_{\sigma(i)}) = \boldsymbol{\alpha}(\mathbf{C}_i + \mathbf{a}_i) = \boldsymbol{\alpha}(\mathbf{C}_i) + \mathbf{a}_i.$$

Subtracting this equality from $\widetilde{\mathbf{C}}_{\sigma(i)} = \mathbf{C}_i + \mathbf{a}_i$ yields

$$\widetilde{\mathbf{C}}_{\sigma(i)} - \boldsymbol{\alpha}(\widetilde{\mathbf{C}}_{\sigma(i)}) = \mathbf{C}_i - \boldsymbol{\alpha}(\mathbf{C}_i) \quad \bar{\mu}\text{-almost everywhere.}$$

□

For integrable contributions, $\boldsymbol{\alpha}(\mathbf{H}) = \mathbb{E}_{\bar{\mu}}\mathbf{H}$ is a finite, translation-equivariant anchor. A conditional mean on a fixed event of positive $\bar{\mu}$-probability is another example when the conditional first moment exists.

C.6 Quantitative Stability

Theorem 3.5 compares decompositions with equal sums. We now allow a small discrepancy between their sums and quantify the resulting error in the recovered contributions. All spaces and supports retain the definitions in Appendix C.1. On the same pair probability space, let

$$\mathbf{d} = \sum_{i=1}^{K} \mathbf{G}_i, \qquad \widetilde{\mathbf{d}} = \sum_{j=1}^{\widetilde{K}} \widetilde{\mathbf{G}}_j, \qquad \mathbf{d}, \widetilde{\mathbf{d}} \in L^2(\Pi; \mathbb{R}^d), \tag{26}$$

where every contribution is measurable and nonzero with positive probability. The spaces U_i^{Δ} and $\widetilde{U}_j^{\Delta}$ describe these structural contributions. Observation noise is introduced separately in Appendix C.7.

C.6.1 Quantitative Conditions and Recovery Bound

For a linear subspace V, let $\mathbf{P}_V$ be its orthogonal projector and $\mathbb{S}(V) = \{\mathbf{v} \in V : \|\mathbf{v}\| = 1\}$. Define

$$\operatorname{gap}(V, W) = \|(\mathbf{I} - \mathbf{P}_W)\mathbf{P}_V\|_{\mathrm{op}}, \qquad d_{\mathrm{sub}}(V, W) = \|\mathbf{P}_V - \mathbf{P}_W\|_{\mathrm{op}}. \tag{27}$$

The directed gap measures approximate inclusion and accommodates unequal dimensions. For a nonzero space V and $\mathbf{Y} \in L^2(V)$, define the finite-cover richness constant

$$\kappa_{\mathbf{Y}}(M) = \inf_{\mathbf{a}_1, \ldots, \mathbf{a}_M \in \mathbb{S}(V)} \left(\mathbb{E} \min_{m \in [M]} |\langle \mathbf{a}_m, \mathbf{Y} \rangle|^2 \right)^{1/2}. \tag{28}$$

Write $\kappa_T(M)$ for this constant under $\mathcal{L}_\Pi(\mathbf{d} \mid S_\Delta = T)$, and $\widetilde{\kappa}_R(M)$ for its analogue under $\mathcal{L}_\Pi(\widetilde{\mathbf{d}} \mid \widetilde{S}_\Delta = R)$.

Assumption C.9 (Quantitative identifiability conditions)**.** Fix integers $M, s \geq 1$ and constants $0 < \alpha \leq 1$ and $p, \kappa > 0$. Both decompositions satisfy the following conditions.

1. The maximal supports obey
$$|\mathcal{F}_\Delta| \leq M, \qquad |T| \leq s, \qquad \Pi(S_\Delta = T) \geq p \quad (T \in \mathcal{F}_\Delta). \tag{29}$$
2. For $T, T' \in \mathcal{F}_\Delta$ and $\mathbf{u}_i \in U_i^{\Delta}$,
$$\Big\| \sum_{i \in T \cup T'} \mathbf{u}_i \Big\|^2 \geq \alpha^2 \sum_{i \in T \cup T'} \|\mathbf{u}_i\|^2. \tag{30}$$
3. $\kappa_T(M) \geq \kappa$ for every $T \in \mathcal{F}_\Delta$.
4. The support intersection property in Assumption 3.4 holds.

The same inequalities and support intersection property hold with tildes, using the same constants and the competing sum $\widetilde{\mathbf{d}}$.

For fixed M, the infimum in Equation (28) is attained: the product of unit spheres is compact, and dominated convergence applies with dominating function $\|\mathbf{Y}\|^2$. The minimum is zero exactly when $\mathbf{Y}$ is concentrated on a union of at most M hyperplanes in V. Thus Assumption 3.3 gives a positive constant for each fixed conditional law and finite M. A common lower bound κ quantifies this condition across the comparison class.

Theorem C.10 (Stability of contribution changes)**.** *Suppose Equation (26) and Assumption C.9 hold, and* $\|\mathbf{d} - \widetilde{\mathbf{d}}\|_{L^2(\Pi)} \leq \varepsilon$. *Put*

$$c_\cap = 1 + \frac{2s}{\alpha}, \qquad \delta = \frac{\varepsilon}{\kappa\sqrt{p}}, \qquad \eta = c_\cap \delta. \tag{31}$$

If $2\eta < \alpha$, *then* $\widetilde{K} = K$ *and there is a permutation* σ *of* $[K]$ *such that*

$$\begin{aligned} \widetilde{\mathcal{F}}_\Delta &= \{\sigma(T) : T \in \mathcal{F}_\Delta\}, \qquad d_{\mathrm{sub}}(V_T^{\Delta}, \widetilde{V}_{\sigma(T)}^{\Delta}) \leq \delta \quad (T \in \mathcal{F}_\Delta), \\ \dim \widetilde{U}_{\sigma(i)}^{\Delta} &= \dim U_i^{\Delta}, \qquad d_{\mathrm{sub}}(U_i^{\Delta}, \widetilde{U}_{\sigma(i)}^{\Delta}) \leq \eta \quad (i \in [K]). \end{aligned} \tag{32}$$

The maximal-space matching with distance at most δ is unique. For the same permutation,

$$D_\Delta := \left(\sum_i \|\mathbf{G}_i - \widetilde{\mathbf{G}}_{\sigma(i)}\|^2_{L^2(\Pi)}\right)^{1/2} \le \frac{\varepsilon}{\alpha} + \left(1 + \frac{\sqrt{s}}{\alpha}\right)\eta\widetilde{R},$$
$$\widetilde{R} := \left(\sum_j \|\widetilde{\mathbf{G}}_j\|^2_{L^2(\Pi)}\right)^{1/2} \le \frac{\|\widetilde{\mathbf{d}}\|_{L^2(\Pi)}}{\alpha}. \tag{33}$$

Intuition. Small reconstruction error forces each maximal combination close to one on the other side. The directness bound separates distinct combinations and controls their intersections, so the atom and support argument of Theorem 3.5 survives the perturbation. The final bound combines reconstruction error with the error from projecting competing contributions onto their matched spaces.

C.6.2 Geometric Lemmas

In the proof, abbreviate $U_i = U_i^\Delta$, $V_A = V_A^\Delta$, $\mathcal{F} = \mathcal{F}_\Delta$, and use tildes for the competing family. The projector identities needed below are

$$\begin{aligned} \operatorname{gap}(V, Z) &\le \operatorname{gap}(V, W) + \operatorname{gap}(W, Z), \\ d_{\mathrm{sub}}(V, W) &= \max\{\operatorname{gap}(V, W), \operatorname{gap}(W, V)\}. \end{aligned} \tag{34}$$

For the first inequality, project $\mathbf{v} \in V$ onto W and use $\operatorname{dist}(\mathbf{v}, Z) \le \|\mathbf{v} - \mathbf{P}_W\mathbf{v}\| + \operatorname{dist}(\mathbf{P}_W\mathbf{v}, Z)$. For the second, write $\mathbf{P} = \mathbf{P}_V, \mathbf{Q} = \mathbf{P}_W$. The operator

$$(\mathbf{P} - \mathbf{Q})^2 = \mathbf{P}(\mathbf{I} - \mathbf{Q})\mathbf{P} + (\mathbf{I} - \mathbf{P})\mathbf{Q}(\mathbf{I} - \mathbf{P})$$

is block diagonal on $V \oplus V^\perp$. Its blocks have norms $\|(\mathbf{I} - \mathbf{Q})\mathbf{P}\|^2$ and $\|\mathbf{Q}(\mathbf{I} - \mathbf{P})\|^2$, respectively; the latter equals $\|(\mathbf{I} - \mathbf{P})\mathbf{Q}\|^2$ by taking adjoints. This proves the identity. If both directed gaps are less than one, then $\mathbf{P}_W|_V$ and $\mathbf{P}_V|_W$ are injective, and hence $\dim V = \dim W$.

Lemma C.11 (Quantitative finite cover)**.** *Let $\mathbf{Y} \in L^2(V)$ with $V \ne \{\mathbf{0}\}$ and $\kappa_{\mathbf{Y}}(M) \ge \kappa$. For $1 \le q \le M$ subspaces $W_1, \dots, W_q$ of $\mathbb{R}^d$,*

$$\left\|\min_{j\in[q]} \operatorname{dist}(\mathbf{Y}, W_j)\right\|_{L^2} \ge \kappa \min_{j\in[q]} \operatorname{gap}(V, W_j). \tag{35}$$

Proof. Let $b_j = \|(\mathbf{I} - \mathbf{P}_{W_j})|_V\| = \operatorname{gap}(V, W_j)$. If any $b_j = 0$, the conclusion follows by nonnegativity. Otherwise, choose unit right and left singular vectors $\mathbf{a}_j \in V$ and $\mathbf{v}_j \in W_j^\perp$ for the largest singular value b_j. For every $\mathbf{u} \in V$,

$$\operatorname{dist}(\mathbf{u}, W_j) \ge |\langle \mathbf{v}_j, \mathbf{u}\rangle| = b_j|\langle \mathbf{a}_j, \mathbf{u}\rangle|.$$

Take the minimum over j and then the L^2 norm. Repeating directions when $q < M$ leaves the minimum unchanged, so Equation (28) gives the result. □

Lemma C.12 (Separation of support inclusions)**.** *Under Equation (30), for $A \subseteq T$, $B \subseteq T'$ and $T, T' \in \mathcal{F}$,*

$$A \not\subseteq B \quad \Longrightarrow \quad \operatorname{gap}(V_A, V_B) \ge \alpha. \tag{36}$$

The containment and intersection identities of Lemma C.2 also hold.

Proof. Choose $i \in A \setminus B$ and a unit vector $\mathbf{u} \in U_i$. For any $\mathbf{v} = \sum_{j\in B} \mathbf{v}_j \in V_B$, with $\mathbf{v}_j \in U_j$, Equation (30) gives

$$\|\mathbf{u} - \mathbf{v}\|^2 \ge \alpha^2 \left(1 + \sum_{j\in B} \|\mathbf{v}_j\|^2\right) \ge \alpha^2.$$

Taking the infimum over $\mathbf{v}$ proves the gap bound. The same inequality implies local directness, so Lemma C.2 applies. □

Lemma C.13 (Stability of all combination intersections). *Under Equation (30) and $|T| \le s$, for every nonempty $\mathcal{H} \subseteq \mathcal{F}$, set $A = \bigcap_{T\in\mathcal{H}} T$. Then*

$$\bigcap_{T\in\mathcal{H}} V_T = V_A, \qquad \operatorname{dist}(\mathbf{v}, V_A) \le c_\cap \max_{T\in\mathcal{H}} \operatorname{dist}(\mathbf{v}, V_T) \quad (\mathbf{v}\in\mathbb{R}^d). \tag{37}$$

Proof. The intersection identity follows from Lemma C.2. Put $b = \max_{T\in\mathcal{H}} \operatorname{dist}(\mathbf{v}, V_T)$, choose $T_0 \in \mathcal{H}$, and write $\mathbf{v}_0 = \mathbf{P}_{V_{T_0}}\mathbf{v} = \sum_{i\in T_0} \mathbf{u}_i$, with $\mathbf{u}_i \in U_i$. For each $i \in T_0 \setminus A$, choose $T_i \in \mathcal{H}$ excluding i. Then

$$\operatorname{dist}(\mathbf{v}_0, V_{T_i}) \le \|\mathbf{v}_0 - \mathbf{v}\| + \operatorname{dist}(\mathbf{v}, V_{T_i}) \le 2b.$$

Apply Equation (30) on $T_0 \cup T_i$ to $\mathbf{v}_0 - \mathbf{P}_{V_{T_i}}\mathbf{v}_0$. Its ith component is $\mathbf{u}_i$, so $\alpha\|\mathbf{u}_i\| \le 2b$. Therefore

$$\operatorname{dist}(\mathbf{v}, V_A) \le \Big\|\mathbf{v} - \sum_{i\in A} \mathbf{u}_i\Big\| \le b + \sum_{i\in T_0\setminus A} \|\mathbf{u}_i\| \le (1 + 2s/\alpha)b.$$

This argument also applies when $A = \varnothing$. □

C.6.3 Proof of Theorem C.10

Proof. Step 1: maximal combination spaces. The finite-cover argument of Lemma C.1, applied separately to each sum, gives $\widetilde{\mathbf{d}} \in \bigcup_R \widetilde{V}_R$ almost surely. Consequently, for every $T \in \mathcal{F}$,

$$\mathbb{E}_\Pi\Big[\min_{R\in\widetilde{\mathcal{F}}} \operatorname{dist}(\mathbf{d}, \widetilde{V}_R)^2 \mid S_\Delta = T\Big] \le \frac{\|\mathbf{d} - \widetilde{\mathbf{d}}\|^2_{L^2(\Pi)}}{\Pi(S_\Delta = T)} \le \frac{\varepsilon^2}{p}. \tag{38}$$

Lemma C.11 yields a map $\phi : \mathcal{F} \to \widetilde{\mathcal{F}}$ with $\operatorname{gap}(V_T, \widetilde{V}_{\phi(T)}) \le \delta$. Interchanging the two decompositions gives $\psi : \widetilde{\mathcal{F}} \to \mathcal{F}$ with the reverse directed bound. Thus

$$\operatorname{gap}(V_T, V_{\psi\phi(T)}) \le 2\delta < \alpha.$$

Lemma C.12 gives $T \subseteq \psi\phi(T)$; maximality gives equality. The same argument on the competing side shows $\phi\psi = \mathrm{id}$. Hence ϕ and ψ are inverse bijections, with $d_{\mathrm{sub}}(V_T, \widetilde{V}_{\phi(T)}) \le \delta$. If two candidate maxima were both within this distance of V_T, their directed gap would be at most $2\delta < \alpha$, forcing their supports to coincide. The matching is therefore unique.

Step 2: atomic spaces, counts, and dimensions. For each nonempty $\mathcal{H} \subseteq \mathcal{F}$, let

$$W_\mathcal{H} = \bigcap_{T\in\mathcal{H}} V_T, \qquad \widetilde{W}_\mathcal{H} = \bigcap_{T\in\mathcal{H}} \widetilde{V}_{\phi(T)}.$$

If $\mathbf{v} \in \widetilde{W}_\mathcal{H}$, then $\operatorname{dist}(\mathbf{v}, V_T) \le \delta\|\mathbf{v}\|$ for all $T \in \mathcal{H}$. Lemma C.13 gives $\operatorname{gap}(\widetilde{W}_\mathcal{H}, W_\mathcal{H}) \le c_\cap\delta = \eta$. Applying the lemma on the competing side gives the reverse bound. Hence

$$d_{\mathrm{sub}}(W_\mathcal{H}, \widetilde{W}_\mathcal{H}) \le \eta < 1, \qquad \dim W_\mathcal{H} = \dim \widetilde{W}_\mathcal{H}. \tag{39}$$

In particular, zero intersections correspond to zero intersections. If $W_\mathcal{H} \subseteq W_\mathcal{J}$, then

$$\operatorname{gap}(\widetilde{W}_\mathcal{H}, \widetilde{W}_\mathcal{J}) \le 2\eta < \alpha.$$

Each competing intersection has the form $\widetilde{V}_A$, with A contained in a competing maximal support. Lemma C.12 therefore gives $\widetilde{W}_\mathcal{H} \subseteq \widetilde{W}_\mathcal{J}$. The reverse implication follows by symmetry. Equality and inclusion of intersections are thus preserved, so the correspondence is well-defined on the two families of distinct intersection spaces. Lemma C.4 identifies their minimal nonzero elements as the atomic spaces. These elements correspond bijectively and have the same dimensions, giving $\widetilde{K} = K$ and a permutation σ with $d_{\mathrm{sub}}(U_i, \widetilde{U}_{\sigma(i)}) \le \eta$.

Step 3: maximal supports. The inclusion correspondence from Step 2 and Lemma C.5 give

$$i \in T \iff U_i \subseteq V_T \iff \widetilde{U}_{\sigma(i)} \subseteq \widetilde{V}_{\phi(T)} \iff \sigma(i) \in \phi(T).$$

Therefore $\phi(T) = \sigma(T)$.

Step 4: contribution error. First, Equation (30) on a maximal support containing the actual support gives, almost surely,

$$\alpha^2 \sum_i \|\mathbf{G}_i\|^2 \le \|\mathbf{d}\|^2, \qquad \alpha^2 \sum_j \|\widetilde{\mathbf{G}}_j\|^2 \le \|\widetilde{\mathbf{d}}\|^2. \tag{40}$$

Thus every contribution is in L^2, and the bound on $\widetilde{R}$ in Equation (33) follows. Set

$$\mathbf{H}_i = \mathbf{P}_{U_i} \widetilde{\mathbf{G}}_{\sigma(i)}, \qquad \mathbf{e}_i = \widetilde{\mathbf{G}}_{\sigma(i)} - \mathbf{H}_i, \qquad \mathbf{a}_i = \mathbf{G}_i - \mathbf{H}_i.$$

Write $\|\mathbf{e}\|_{2,\mathrm{blk}} = (\sum_i \|\mathbf{e}_i\|^2_{L^2(\Pi)})^{1/2}$, and similarly for $\mathbf{a}$. The space bound gives $\|\mathbf{e}\|_{2,\mathrm{blk}} \le \eta \widetilde{R}$. For almost every pair, the nonzero $\mathbf{a}_i$ are indexed by the union of two true maximal supports, using Step 3 for the competing support. Equation (30) therefore yields

$$\left(\sum_i \|\mathbf{a}_i\|^2 \right)^{1/2} \le \alpha^{-1} \left\| \mathbf{d} - \widetilde{\mathbf{d}} + \sum_i \mathbf{e}_i \right\|.$$

At most s of the $\mathbf{e}_i$ are nonzero on any pair, so $\|\sum_i \mathbf{e}_i\| \le \sqrt{s} (\sum_i \|\mathbf{e}_i\|^2)^{1/2}$. Minkowski's inequality now gives

$$\|\mathbf{a}\|_{2,\mathrm{blk}} \le \alpha^{-1}(\varepsilon + \sqrt{s}\,\eta \widetilde{R}), \qquad D_\Delta \le \|\mathbf{a}\|_{2,\mathrm{blk}} + \|\mathbf{e}\|_{2,\mathrm{blk}},$$

which proves Equation (33). □

Dependence on the conditions. With bounded signal energy and fixed positive α, p, κ, the contribution bound is linear in ε. Small directness or richness constants, and rare maximal supports, reduce the admissible error threshold and enlarge the bound. The support intersection property is a discrete condition. For example, adding a $\{1,2,3\}$ event of arbitrarily small positive probability to the support family $\{\{1,2\},\{1,3\},\{2,3\}\}$ leaves only $\{1,2,3\}$ maximal and destroys that property. Such departures can be treated by restricting to a high-probability event, as below.

C.7 Representation Error, Noise, and Contamination

Approximate ConCA representations. Suppose the representation has residual $\mathbf{e}$, $\mathbf{f} = \mathbf{b} + \sum_i \mathbf{C}_i + \mathbf{e}$, and the observed difference is $\mathbf{d}_{\mathrm{obs}} = \mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}') + \boldsymbol{\nu}$. With $\mathbf{G}_i = \Delta \mathbf{C}_i$, $\mathbf{d} = \sum_i \mathbf{G}_i$, and $\bar{\mu} = (\Pi_1 + \Pi_2)/2$, assume

$$\|\mathbf{e}\|_{L^2(\bar{\mu})} \le \varepsilon_{\mathrm{rep}}, \qquad \|\boldsymbol{\nu}\|_{L^2(\Pi)} \le \varepsilon_{\mathrm{noise}}, \qquad \|\mathbf{d}_{\mathrm{obs}} - \widetilde{\mathbf{d}}\|_{L^2(\Pi)} \le \varepsilon_{\mathrm{fit}}.$$

Then Theorem C.10 applies whenever its structural conditions and error threshold hold with

$$\varepsilon = 2\varepsilon_{\mathrm{rep}} + \varepsilon_{\mathrm{noise}} + \varepsilon_{\mathrm{fit}}. \tag{41}$$

Indeed,

$$\mathbb{E}_\Pi \|\mathbf{e}(\mathbf{x}) - \mathbf{e}(\mathbf{x}')\|^2 \le 2\mathbb{E}_{\Pi_1}\|\mathbf{e}\|^2 + 2\mathbb{E}_{\Pi_2}\|\mathbf{e}\|^2 = 4\|\mathbf{e}\|^2_{L^2(\bar{\mu})},$$

and $\mathbf{d} - \widetilde{\mathbf{d}} = (\mathbf{d} - \mathbf{d}_{\mathrm{obs}}) + (\mathbf{d}_{\mathrm{obs}} - \widetilde{\mathbf{d}})$. All three errors are population quantities; the joint law may include dependence between the endpoints, residual, and noise.

Approximately sparse contributions. Let $\mathbf{F}_i = \mathbf{G}_i + \mathbf{L}_i$ and $\widetilde{\mathbf{F}}_j = \widetilde{\mathbf{G}}_j + \widetilde{\mathbf{L}}_j$, with square-integrable tails. The cores $\mathbf{G}, \widetilde{\mathbf{G}}$ satisfy Assumption C.9. Define

$$\tau = \left(\sum_i \|\mathbf{L}_i\|^2_{L^2} \right)^{1/2}, \qquad \rho_\Sigma = \left\| \sum_i \mathbf{L}_i \right\|_{L^2},$$

with $\widetilde{\tau}, \widetilde{\rho}_\Sigma$ defined analogously. If $\mathbf{d}_{\mathrm{obs}} = \sum_i \mathbf{F}_i + \mathbf{r}$ and $\|\mathbf{d}_{\mathrm{obs}} - \sum_j \widetilde{\mathbf{F}}_j\|_{L^2} \le \varepsilon_{\mathrm{fit}}$, the core discrepancy and full contribution error satisfy

$$\begin{aligned} \|\mathbf{d} - \widetilde{\mathbf{d}}\|_{L^2} &\le \|\mathbf{r}\|_{L^2} + \varepsilon_{\mathrm{fit}} + \rho_\Sigma + \widetilde{\rho}_\Sigma, \\ \left(\sum_i \|\mathbf{F}_i - \widetilde{\mathbf{F}}_{\sigma(i)}\|^2_{L^2} \right)^{1/2} &\le D_\Delta + \tau + \widetilde{\tau}, \end{aligned} \tag{42}$$

where the second line uses the permutation from Theorem C.10 applied with the first line as its error budget. Both inequalities follow by subtraction and the triangle inequality. In general $\rho_\Sigma \le \sqrt{K}\,\tau$; cancellation can make ρ_Σ small while τ remains large. The recovered dimensions belong to the structural cores. A full-dimensional tail of arbitrarily small norm can enlarge the actual spaces of the complete functions $\mathbf{F}_i$.

Corollary C.14 (Clean risk under mixture contamination)**.** *Let $Q = (1-\zeta)P + \zeta H$, where $0 \le \zeta < 1$ and P, H are joint laws of paired inputs and observed differences $\mathbf{d}_{\mathrm{obs}}$. Suppose Equation (26) and Assumption C.9 hold under the clean pair marginal of P, and*

$$\|\mathbf{d}_{\mathrm{obs}} - \mathbf{d}\|_{L^2(P)} \le n_0, \qquad \|\mathbf{d}_{\mathrm{obs}} - \widetilde{\mathbf{d}}\|_{L^2(Q)} \le f_Q.$$

Then Theorem C.10 applies on the clean distribution with error budget

$$\varepsilon_P = n_0 + \frac{f_Q}{\sqrt{1-\zeta}}, \tag{43}$$

provided its threshold holds for ε_P.

Proof. For every nonnegative loss L, $\mathbb{E}_Q L \ge (1-\zeta)\mathbb{E}_P L$. Apply this to $L = \|\mathbf{d}_{\mathrm{obs}} - \widetilde{\mathbf{d}}\|^2$ and use the triangle inequality under P. □

This corollary controls clean population error conditional on f_Q. The contamination fraction alone gives no upper bound on f_Q: arbitrary observations under H may have unbounded second moments.

Corollary C.15 (Structural conditions on a high-probability event)**.** *Let E be an event with $\Pi(E) = 1 - \zeta > 0$. Suppose Theorem C.10 applies under $P = \Pi(\cdot \mid E)$ to the same fixed contribution indices, each nonzero with positive P-probability. Let σ be the resulting permutation and D_g its contribution error bound under P. If, almost surely on E^c,*

$$\left(\sum_i \|\mathbf{G}_i\|^2\right)^{1/2} \le R_\infty, \qquad \left(\sum_j \|\widetilde{\mathbf{G}}_j\|^2\right)^{1/2} \le R_\infty,$$

then the same permutation satisfies

$$\left(\sum_i \|\mathbf{G}_i - \widetilde{\mathbf{G}}_{\sigma(i)}\|^2_{L^2(\Pi)}\right)^{1/2} \le \sqrt{(1-\zeta)D_g^2 + 4R_\infty^2\zeta}. \tag{44}$$

Proof. The squared block error integrated over E is at most $(1-\zeta)D_g^2$. On E^c, the triangle inequality bounds the block error by $2R_\infty$. Add the two integrals. □

All spaces, support probabilities, and richness constants used on the good event are defined under P. Contributions occurring only on E^c are excluded by the nonzero condition, and their number is not determined by the conditional theorem. Without a bound on the bad-event error, an error of magnitude A on mass ζ contributes $A\sqrt{\zeta}$ to the overall L^2 error.

If the structural conditions hold under the original Π, a simpler error budget follows from $\mathbb{E}_\Pi[\|\mathbf{d} - \widetilde{\mathbf{d}}\|^2 \mathbf{1}_E] \le \varepsilon_g^2$ and $\|\mathbf{d}\|, \|\widetilde{\mathbf{d}}\| \le B_\infty$ almost surely:

$$\|\mathbf{d} - \widetilde{\mathbf{d}}\|_{L^2(\Pi)} \le \sqrt{\varepsilon_g^2 + 4B_\infty^2\Pi(E^c)}. \tag{45}$$

This follows from the same decomposition of the squared error over E and E^c.

C.8 Quantitative Recovery of Endpoint Contributions

For stable endpoint recovery, impose a Poincaré inequality on the pairing law: for some $\lambda_\Pi > 0$,

$$\|D_\Pi h\|^2_{L^2(\Pi)} \ge \lambda_\Pi \inf_{a \in \mathbb{R}} \|h - a\|^2_{L^2(\bar{\mu})} \qquad (h \in L^2(\bar{\mu}; \mathbb{R})). \tag{46}$$

This strengthens the essential connectivity condition in Equation (23) by a positive lower bound on the energy of nonconstant functions.

Corollary C.16 (Endpoint stability after mean centering)**.** *Under Theorem C.10 and Equation (46), suppose* $\mathbf{G}_i = D_\Pi \mathbf{C}_i$ *and* $\widetilde{\mathbf{G}}_j = D_\Pi \widetilde{\mathbf{C}}_j$*, with* $\mathbf{C}_i, \widetilde{\mathbf{C}}_j \in L^2(\bar{\mu}; \mathbb{R}^d)$*. For* $\mathbf{C}_i^\circ = \mathbf{C}_i - \mathbb{E}_{\bar{\mu}} \mathbf{C}_i$ *and analogously* $\widetilde{\mathbf{C}}_j^\circ$*, the same permutation obeys*

$$\left(\sum_i \|\mathbf{C}_i^\circ - \widetilde{\mathbf{C}}_{\sigma(i)}^\circ\|_{L^2(\bar{\mu})}^2 \right)^{1/2} \leq \frac{D_\Delta}{\sqrt{\lambda_\Pi}}. \tag{47}$$

Proof. Apply Equation (46) to each coordinate of $\mathbf{h}_i = \widetilde{\mathbf{C}}_{\sigma(i)} - \mathbf{C}_i$ and sum over coordinates and concepts. The minimizing constant is $\mathbb{E}_{\bar{\mu}} \mathbf{h}_i$, and $D_\Pi \mathbf{h}_i = \widetilde{\mathbf{G}}_{\sigma(i)} - \mathbf{G}_i$. □

For example, let $0 < q \leq 1$. On two inputs with a uniform endpoint marginal, give each directed cross-pair probability $q/2$ and each self-pair probability $(1-q)/2$. Then $\lambda_\Pi = 4q$. Each such law is connected, while the endpoint error factor diverges as $q \to 0$. The normalization in Equation (46) fixes the factor in Equation (47).

D Method and Implementation Details

This appendix describes the estimator used in the synthetic experiments of Section 5.1. Given observed representation changes $\mathbf{d} = \Delta\mathbf{f}(\mathbf{x}, \mathbf{x}') \in \mathbb{R}^d$, we estimate concept-specific contributions of the form

$$\widehat{\mathbf{G}}_j(\mathbf{d}) = \mathbf{B}_j \mathbf{u}_{\boldsymbol{\theta},j}(\mathbf{d}),$$

where $\mathbf{B}_j$ spans a candidate concept change space and $\mathbf{u}_{\boldsymbol{\theta},j}(\mathbf{d})$ gives its pair-specific coefficients.

The estimator is given the number of concept blocks $m = K$, their common width r, and the generating maximum support size $s = s_{\max}$. It does not observe individual change supports, true concept identities, or true contributions. We use

$$E_0 = \frac{1}{N_{\text{train}}} \sum_{n=1}^{N_{\text{train}}} \|\mathbf{d}_n\|_2^2$$

to denote the mean squared norm of the training differences. This quantity is fixed throughout fitting and prediction.

Our implementation has two stages. First, guided by Conditional Richness and Support Intersection (Assumptions 3.3 and 3.4), we initialize candidate concept change spaces from observed representation changes. We estimate overlapping change patterns, extract directions associated with these patterns, and group recurring directions into concept blocks. Second, motivated by Local Directness (Assumption 3.2) and the group-sparsity selection result (Theorem 3.7), we learn pair-specific contributions through block-sparse reconstruction, starting from these estimated spaces.

D.1 Geometry-Guided Subspace Initialization

The first stage estimates concept change spaces from observed differences. The observation graph relates similar changes; graph factorization estimates overlapping change-pattern memberships; moments extract their directions; and subspace grouping constructs concept blocks. We select a dictionary using held-out differences, without individual support labels or contributions.

Initialization uses the first $n_0 = 4096$ training differences. For each training seed, we construct four candidate dictionaries by combining two NMF capacities with two coordinate systems for direction grouping. The NMF capacities are 18 and 24 for the small configuration, 192 and 256 for the large configuration, and 384 and 512 for the rank-six configuration. These capacities control the number of candidate change patterns used during initialization.

Observation graph. We begin by constructing a graph that connects representation changes with similar directions. We whiten the discovery observations to reduce the influence of coordinates with large marginal variance. We use squared cosine similarity so that directions differing only in sign receive the same similarity.

Let $\widehat{\boldsymbol{\mu}}_{\mathrm{disc}}$ and

$$\widehat{\boldsymbol{\Sigma}} = \mathbf{V}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_d)\mathbf{V}^\top$$

be the sample mean and covariance of the discovery observations, where $\lambda_1 \le \cdots \le \lambda_d$ and the covariance is divided by $n_0 - 1$. Define

$$\mathbf{W} = \mathbf{V}\,\mathrm{diag}\Big(\max\{\lambda_j, 10^{-4}\lambda_d\}^{-1/2}\Big), \qquad \mathbf{q}_n = \frac{\mathbf{W}^\top(\mathbf{d}_n - \widehat{\boldsymbol{\mu}}_{\mathrm{disc}})}{\|\mathbf{W}^\top(\mathbf{d}_n - \widehat{\boldsymbol{\mu}}_{\mathrm{disc}})\|_2}.$$

For each observation, we retain its 250 largest non-self similarities $(\mathbf{q}_n^\top \mathbf{q}_{n'})^2$. Equal boundary scores are resolved in favor of the smaller observation index. We take the maximum of the two directed weights for each observation pair, remove zero-weight edges, and divide the remaining weights by their mean.

Observations with zero whitened norm are rejected. The nearest-neighbor search is exact and computed in tiles to reduce storage; its arithmetic cost remains quadratic in the number of discovery observations.

Graph factorization. The observation graph indicates which changes are geometrically similar, but it does not identify their underlying change supports. We therefore factorize the graph to estimate soft memberships in candidate change patterns. An observation may belong to multiple patterns, and these memberships are not treated as true support indicators.

For graph matrix $\mathbf{K}$ and NMF capacity q, we fit nonnegative, untied factors $\mathbf{L}, \mathbf{R} \in \mathbb{R}_+^{n_0 \times q}$ by minimizing

$$\|\mathbf{K} - \mathbf{L}\mathbf{R}^\top\|_F^2.$$

Starting from one NNDSVDar initialization, we perform 300 iterations of the sequential multiplicative updates

$$\mathbf{L} \leftarrow \mathbf{L} \odot \frac{\mathbf{K}\mathbf{R}}{\mathbf{L}(\mathbf{R}^\top\mathbf{R})}, \qquad \mathbf{R} \leftarrow \mathbf{R} \odot \frac{\mathbf{K}\mathbf{L}}{\mathbf{R}(\mathbf{L}^\top\mathbf{L})}.$$

The multiplication $\odot$ and both matrix quotients are entrywise. Each denominator is floored at 10^{-12} times its largest entry, with a machine-tiny lower bound. Missing graph edges remain part of the Frobenius objective.

We combine the two factors into the membership matrix $\mathbf{H}$:

$$\mathbf{H}_{:b} = \frac{1}{2}\left(\frac{\mathbf{L}_{:b}}{\|\mathbf{L}_{:b}\|_2} + \frac{\mathbf{R}_{:b}}{\|\mathbf{R}_{:b}\|_2}\right),$$

with norm denominators floored at 10^{-12}.

For world identifier w and training seed $\nu \in \{1, 2\}$, the NMF seed is

$$(1 + 1009w + 104729(\nu - 1)) \bmod 2^{32}.$$

The same seed is used at both NMF capacities.

Moment-based direction estimation. The memberships describe candidate change patterns at the observation level, whereas the decoder requires directions in representation space. Motivated by the variation underlying Conditional Richness, we construct membership-dependent second moments to extract directions associated with each candidate pattern.

For any nonnegative membership matrix $\mathbf{H}$, define

$$P_{nb} = \frac{sH_{nb}^2}{\sum_c H_{nc}^2}, \qquad \mathbf{T} = \mathbf{P}(\mathbf{P}^\top\mathbf{P} + \eta\mathbf{I})^{-1}, \qquad \eta = 10^{-3}\frac{\mathrm{tr}(\mathbf{P}^\top\mathbf{P})}{q}. \tag{48}$$

The powered memberships form the design matrix $\mathbf{P}$. Ridge-regression weights $\mathbf{T}$ are then used to account for overlapping membership signals. These weights may be negative and are not interpreted as conditional probabilities.

The row denominator and ridge scale are floored at 10^{-12}. A membership column whose total squared membership is at most 10^{-8} is rejected.

For each candidate pattern b, we form the signed second moment

$$\mathbf{M}_b = \sum_{n=1}^{n_0} T_{nb} \mathbf{d}_n \mathbf{d}_n^\top. \tag{49}$$

These moments use the original, uncentered representation differences. Since the regression weights can be negative, $\mathbf{M}_b$ need not be positive semidefinite.

At this stage, we retain the leading algebraic eigenvector of each $\mathbf{M}_b$, yielding one candidate direction per pattern. These q directions are used in the subsequent grouping step. After grouping, we repeat the moment estimation and retain r directions per concept block.

Selected eigenvalues and solver gaps are recorded as numerical diagnostics. We do not apply random completion or replace negative eigenvalues by their positive parts.

Grouping directions into concept blocks. A candidate pattern may involve several concepts, while the same concept may recur across patterns. Its leading direction therefore need not identify a single concept space.

Motivated by Support Intersection, we group directions that share subspace structure across the estimated patterns. This grouping uses recurring geometric structure as an approximation; it does not explicitly compute the subspace intersections in the proof.

We use sparse subspace clustering (SSC) (Elhamifar & Vidal, 2013). Let $\mathbf{X} \in \mathbb{R}^{d \times q}$ contain the unit-normalized candidate directions. For each NMF fit, we construct $\mathbf{X}$ once in the original coordinates and once after applying the discovery covariance whitening transform. We solve

$$\min_{\mathbf{C}} \quad \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + 0.02\|\mathbf{C}\|_1 \qquad \text{subject to} \quad \operatorname{diag}(\mathbf{C}) = \mathbf{0}. \tag{50}$$

Here, $\|\mathbf{C}\|_1 = \sum_{a,b} |C_{ab}|$ is the entrywise ℓ_1 norm.

The solver uses zero initialization and FISTA (Beck & Teboulle, 2009), with step size $0.99/\|\mathbf{X}\|_{\mathrm{op}}^2$ and at most 4096 iterations. The diagonal is set to zero after each soft-thresholding step. Every 64 iterations, we check the relative primal–dual gap against 10^{-7}.

For residual $\mathbf{R}_\mathbf{X} = \mathbf{X} - \mathbf{X}\mathbf{C}$, each column is scaled by the smaller of 1 and 0.02 divided by its maximum absolute off-diagonal correlation with $\mathbf{X}$. The resulting residuals are dual feasible. The gap is normalized by $\frac{1}{2}\|\mathbf{X}\|_F^2$.

We construct the symmetric affinity

$$\mathbf{F} = |\mathbf{C}| + |\mathbf{C}|^\top$$

and its degree matrix $\mathbf{D}_\mathbf{F}$. We then form the normalized Laplacian

$$\mathbf{I} - \mathbf{D}_\mathbf{F}^{-1/2} \mathbf{F} \mathbf{D}_\mathbf{F}^{-1/2}.$$

Inverse degree scales use a 10^{-30} floor. We retain the m eigenvectors corresponding to the smallest eigenvalues, normalize their rows, and run k-means with m clusters, 20 initializations, and seed ν.

Let $\mathcal{I}_j$ contain the candidate directions assigned to cluster j. We combine their original pattern memberships:

$$H_{nj}^{\mathrm{merged}} = \left(\sum_{b \in \mathcal{I}_j} H_{nb}^2 \right)^{1/2}.$$

This merging preserves the sum of squared memberships within each group.

We apply Equations (48)– (49) to the merged membership matrix, replacing q by m, and retain the r leading algebraic eigenvectors for each group. Finally, we apply a blockwise polar projection to obtain the initialized bases $\{\mathbf{B}_j\}_{j=1}^m$. The columns are orthonormal within each block; orthogonality between different blocks is not imposed.

Observed-data dictionary selection. The two NMF capacities and two grouping coordinate systems yield four candidate dictionaries for each training seed. We select among them using validation representation differences, without true concept contributions or support labels.

For each fixed candidate dictionary $\mathbf{B}$, we code the first 1024 validation differences after dividing them by $\sqrt{E_0}$:

$$\min_{\mathbf{v}_1,\dots,\mathbf{v}_m} \quad \frac{1}{2}\left\|\frac{\mathbf{d}}{\sqrt{E_0}} - \sum_{j=1}^{m}\mathbf{B}_j\mathbf{v}_j\right\|_2^2 + 0.02\sum_{j=1}^{m}\|\mathbf{v}_j\|_2. \tag{51}$$

This is convex coefficient inference with the dictionary held fixed.

We solve it using FISTA from zero, with step size

$$\frac{0.99}{\|[\mathbf{B}_1 \ \cdots \ \mathbf{B}_m]\|_2^2}.$$

The iteration budget is 2048 for the reference configurations and 8192 for rank-six concepts. A primal–dual certificate is checked every 64 iterations.

For residual $\mathbf{e}_n$ and normalized observation $\bar{\mathbf{d}}_n$, define

$$\mathbf{t}_n = \mathbf{e}_n \min\left\{1, \frac{0.02}{\max_j \|\mathbf{B}_j^\top \mathbf{e}_n\|_2}\right\}.$$

The corresponding dual objective is

$$\bar{\mathbf{d}}_n^\top \mathbf{t}_n - \frac{1}{2}\|\mathbf{t}_n\|_2^2.$$

A candidate dictionary is eligible when its mean primal–dual gap, divided by

$$\frac{1}{2}\,\mathrm{mean}_n\,\|\bar{\mathbf{d}}_n\|_2^2,$$

is at most 10^{-5}. Zero denominators use the implementation's 10^{-30} floor. Among eligible candidates, we select the dictionary with the lowest mean primal objective.

The remaining 7168 validation observations are used for a separate coding diagnostic. If no candidate is eligible, the initializer reports failure; it does not substitute the separately computed raw-graph diagnostic dictionary.

The SSC stopping tolerance is used to assess optimization accuracy, not to reject an otherwise eligible dictionary. SSC fits that exhaust their iteration budget retain their final iterates. The block-sparse coding eligibility check remains mandatory. Appendix F.4 reports the recorded convergence warnings for the current experiments.

D.2 Statistical Analysis of Subspace Initialization

The identifiability proof recovers individual concept change spaces from intersections of population combined change spaces. Directly estimating every change combination may require substantial data. For example, the $K = 32$, $s_{\max} = 4$ benchmark has $\binom{32}{4} = 35{,}960$ maximal supports, whereas initialization uses only 4096 discovery observations.

We examine which features of the initialization are supported by simple statistical arguments and which would require additional assumptions beyond Theorem 3.5.

Whitening and angular similarity. Whitening reduces the influence of representation directions with large marginal variance when constructing the observation graph. An invertible linear transformation $\mathbf{W}^\top$ preserves subspace intersections:

$$\mathbf{W}^\top(U \cap V) = (\mathbf{W}^\top U) \cap (\mathbf{W}^\top V).$$

It also preserves directness, although it changes the angles between subspaces and therefore changes their neighborhood relationships.

Squared cosine similarity treats opposite normalized directions equally. Its relation to subspace overlap can be seen in a simple reference model. Suppose $\mathbf{q}$ and $\mathbf{q}'$ are independent and uniformly distributed on the unit spheres of nonzero subspaces U and V, respectively. Then

$$\mathbb{E}\left[(\mathbf{q}^\top \mathbf{q}')^2\right] = \frac{\operatorname{tr}(\mathbf{P}_U \mathbf{P}_V)}{\dim U \ \dim V}, \tag{52}$$

where $\mathbf{P}_U$ and $\mathbf{P}_V$ are their orthogonal projections.

Indeed, $\mathbb{E}[\mathbf{q}\mathbf{q}^\top] = \mathbf{P}_U / \dim U$ and $\mathbb{E}[\mathbf{q}'\mathbf{q}'^\top] = \mathbf{P}_V / \dim V$. Independence gives the trace identity. The expectation equals $1/\dim U$ when $U = V$ and is zero when the subspaces are orthogonal.

Equation (52) provides intuition for using angular similarity. In practice, representation changes need not be isotropic and may lie in overlapping, composite change spaces. As a result, the observation graph may capture change patterns shared across concepts, as well as those specific to individual concepts.

Overlapping memberships. NMF allows an observation to have nonzero membership in several graph factors, consistent with the possibility that multiple concepts change within one context pair.

The capacity of the NMF model determines the number of candidate patterns that can be identified, with each factor potentially corresponding to a single concept, a combination of multiple concepts, or a region comprising many similar observations, while the resulting membership weights are used to weight the observations when estimating candidate directions of change.

In particular, the powered memberships have a common row total, whereas the actual number of changing concepts may vary across pairs. The purpose of direction grouping is to use information from several estimated patterns to construct each concept space.

Moment estimation and its error. The membership-dependent moments are obtained by ridge regression. Let $\mathbf{Y}_2 \in \mathbb{R}^{n_0 \times d^2}$ have rows

$$\operatorname{vec}(\mathbf{d}_n \mathbf{d}_n^\top)^\top.$$

For membership design $\mathbf{P}$ and ridge parameter $\eta > 0$, the moments in Equation (49), stacked as rows, satisfy

$$\widehat{\mathbf{M}}_{\text{stack}} = (\mathbf{P}^\top \mathbf{P} + \eta \mathbf{I})^{-1} \mathbf{P}^\top \mathbf{Y}_2.$$

For any proposed target moments $\mathbf{M}^\star_{\text{stack}}$, define

$$\mathbf{\Xi} = \mathbf{Y}_2 - \mathbf{P}\mathbf{M}^\star_{\text{stack}}.$$

Then

$$\begin{aligned} \widehat{\mathbf{M}}_{\text{stack}} - \mathbf{M}^\star_{\text{stack}} = & -\eta(\mathbf{P}^\top \mathbf{P} + \eta \mathbf{I})^{-1} \mathbf{M}^\star_{\text{stack}} \\ & + (\mathbf{P}^\top \mathbf{P} + \eta \mathbf{I})^{-1} \mathbf{P}^\top \mathbf{\Xi}. \end{aligned} \tag{53}$$

This identity separates the bias introduced by ridge regularization from the residual propagated through the membership design. Strongly correlated memberships make the estimate sensitive to perturbations. Increasing η reduces this sensitivity at the cost of additional bias.

The membership design is itself estimated from the observations. Consequently, treating $\mathbf{\Xi}$ as conditionally mean-zero would require a separate statistical argument. The regression weights may also be negative, so the estimated moment matrices need not be positive semidefinite.

The choice of target moments matters. Whenever $\mathbb{E}\sum_k \|\mathbf{G}_k\|_2^2 < \infty$, for any positive-probability support T,

$$\mathbb{E}[\mathbf{d}\mathbf{d}^\top \mid S_\Delta = T] = \sum_{k,j \in T} \mathbb{E}[\mathbf{G}_k \mathbf{G}_j^\top \mid S_\Delta = T]. \tag{54}$$

The next lemma describes the eigenspace error once an appropriate target moment has been specified.

Lemma D.1 (Conditional subspace perturbation). *Let* $\mathbf{M}^\star = \mathbf{Q}_\star \mathbf{\Sigma}_\star \mathbf{Q}_\star^\top$, *where* $\mathbf{Q}_\star^\top \mathbf{Q}_\star = \mathbf{I}_r$, $1 \le r < d$, *and* $\mathbf{\Sigma}_\star \succ 0$. *Set* $\delta = \lambda_{\min}(\mathbf{\Sigma}_\star)$.

For a symmetric estimate $\widehat{\mathbf{M}}$, *let* $\widehat{U}$ *be a leading* r*-dimensional eigenspace and let* $U^\star = \mathrm{range}(\mathbf{Q}_\star)$. *Then*

$$\|\sin \mathbf{\Theta}(\widehat{U}, U^\star)\|_F \le \frac{2\sqrt{r}\|\widehat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{op}}}{\delta}, \tag{55}$$

where $\mathbf{\Theta}$ *contains the principal angles between the two subspaces.*

Proof. The gap between the rth and $(r+1)$st eigenvalues of $\mathbf{M}^\star$ is δ. Apply the population-gap Davis–Kahan bound of Yu et al. (2015, Theorem 2) and bound its numerator by $2\sqrt{r}\|\widehat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{op}}$. □

For this lemma to describe concept recovery, the proposed target moment must have $U^\star = U_k^\Delta$. The bound then makes two further requirements explicit: the moment estimate must be accurate, and the weakest concept direction must have sufficient signal. Qualitative Conditional Richness alone does not provide a lower bound on δ or on the probabilities of individual change supports.

Direction grouping and dictionary selection. SSC connects candidate directions through sparse self-expression, following Elhamifar & Vidal (2013). To obtain a recovery guarantee for the present initializer, one would need additional conditions on the accuracy of the estimated directions, their coverage of each concept space, and the connectivity of the resulting graph.

Directions estimated from mixed moments may connect different concepts. Conversely, weak coverage may split directions from one concept into separate groups. A small self-expression objective or primal–dual gap certifies neither the correctness of the grouping nor the recovery of individual concept spaces.

The held-out block-sparse coding step selects dictionaries that give accurate reconstruction with a low contribution-norm penalty on new differences. Its primal–dual certificate controls numerical error in this comparison.

However, similar coding scores do not necessarily imply similar concept contributions. Relating held-out dictionary selection to concept recovery would require further conditions.

D.3 Learning Concept Change Contributions

Given the initialized bases $\{\mathbf{B}_j\}_{j=1}^m$, the second stage learns how each observed representation change is distributed across the concept blocks. For an input difference $\mathbf{d}$, the encoder predicts block-specific coefficients $\mathbf{u}_{\boldsymbol{\theta},j}(\mathbf{d})$, which are mapped back to representation space through

$$\widehat{\mathbf{G}}_j(\mathbf{d}) = \mathbf{B}_j \mathbf{u}_{\boldsymbol{\theta},j}(\mathbf{d}).$$

The predicted contributions are trained to reconstruct the observed change, while a group-sparsity penalty encourages fewer active blocks for each observation.

Encoder and block thresholding. Reversing a context pair changes $\mathbf{d}$ to $-\mathbf{d}$. We therefore construct the encoder so that its predicted contributions change sign under pair reversal.

Let

$$a = \sqrt{E_0/d}$$

be a fixed scale computed from the training differences. We use one GELU hidden layer of width $4mr$ and a bias-free linear output of width mr:

$$\boldsymbol{\psi}_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{W}_2\, \mathrm{GELU}(\mathbf{W}_1 \mathbf{x} + \mathbf{b}).$$

The encoder output is divided into m blocks of width r. Its pre-threshold coefficients, thresholded coefficients, and predicted contributions are

$$\begin{aligned}\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{d}) &= \frac{a}{2}\left[\boldsymbol{\psi}_{\boldsymbol{\theta}}\big(\phi(\mathbf{d}/a)\big) - \boldsymbol{\psi}_{\boldsymbol{\theta}}\big(\phi(-\mathbf{d}/a)\big)\right], \\ \mathbf{u}_{\boldsymbol{\theta},j}(\mathbf{d}) &= \left(1 - \frac{\tau_j}{\|\mathbf{v}_{\boldsymbol{\theta},j}(\mathbf{d})\|_2}\right)_+ \mathbf{v}_{\boldsymbol{\theta},j}(\mathbf{d}), \qquad \tau_j = a\,\mathrm{softplus}(t_j), \\ \widehat{\mathbf{G}}_j(\mathbf{d}) &= \mathbf{B}_j \mathbf{u}_{\boldsymbol{\theta},j}(\mathbf{d}).\end{aligned} \tag{56}$$

The thresholding operation returns zero whenever $\|\mathbf{v}_{\boldsymbol{\theta},j}(\mathbf{d})\|_2 \le \tau_j$, including when the input block is zero. Each threshold parameter t_j is learned, with τ_j/a initialized to 0.1. The same thresholded coefficients are used during training and prediction.

The difference between the encoder outputs at $\mathbf{d}$ and $-\mathbf{d}$ makes the resulting map odd:

$$\widehat{\mathbf{G}}_j(-\mathbf{d}) = -\widehat{\mathbf{G}}_j(\mathbf{d}).$$

The hidden bias allows a nonlinear odd map. Without this bias, the identity $\mathrm{GELU}(z) - \mathrm{GELU}(-z) = z$ would make the one-layer encoder with identity input features linear before block thresholding.

The main experiments use $\phi(\mathbf{x}) = \mathbf{x}$. The input-feature ablation instead uses the fixed odd map

$$\phi(\mathbf{x}) = [\mathbf{x}; \sin(\mathbf{x}); \sin(2\mathbf{x}); \sin(4\mathbf{x})].$$

The frequencies act coordinatewise in radians, without a 2π factor. The hidden and output widths remain unchanged. This ablation adds $3d(4mr)$ input-layer weights and is not a parameter-count-matched comparison.

Concept bases and block sparsity. Each concept basis $\mathbf{B}_j \in \mathbb{R}^{d\times r}$ has orthonormal columns. After each joint parameter update, we apply the blockwise polar projection

$$\mathbf{B}_j \leftarrow \mathbf{B}_j(\mathbf{B}_j^\top \mathbf{B}_j)^{-1/2}.$$

When $d \ge 32$ and $\|\mathbf{B}_j^\top \mathbf{B}_j - \mathbf{I}\|_\infty < 0.1$ for every block, the implementation uses this Gram-matrix form. Otherwise, it uses the thin-SVD form $\mathbf{U}_j\mathbf{V}_j^\top$. Rank-deficient blocks are rejected by the SVD guard. Both are numerical implementations of the same polar projection.

Orthonormality is enforced within each block, but not between different blocks. This allows different concept change spaces to have overlapping directions.

The sparsity penalty is evaluated on contributions in representation space, $\|\widehat{\mathbf{G}}_j(\mathbf{d})\|_2$, as in Theorem 3.7. For orthonormal blocks, $\|\mathbf{B}_j\mathbf{u}_j\|_2 = \|\mathbf{u}_j\|_2$. We retain the ambient contribution expression in the computational graph rather than substituting coefficient norms. On the orthonormality constraint, the penalty has no variation along feasible changes of basis with fixed coefficients; its ambient derivative nevertheless differs from that of the coefficient-only expression.

Zero contribution norms use the zero subgradient, without a positive smoothing constant. All m blocks enter the reconstruction and sparsity penalty; we do not impose a hard top-s support constraint.

D.4 Penalized Training and Schedule

The initialization provides starting concept spaces. We now learn the contributions of individual observations while allowing these spaces to change during joint training. We train the model using the reconstruction objective and group-sparsity regularizer in Equation (5).

For minibatch $\mathcal{B}$, we normalize the reconstruction error by the mean squared norm of the training differences and the contribution penalty by its square root:

$$\begin{aligned} R_{\mathcal{B}} &= \frac{1}{|\mathcal{B}|E_0} \sum_{n\in\mathcal{B}} \left\| \mathbf{d}_n - \sum_{j=1}^{m} \widehat{\mathbf{G}}_j(\mathbf{d}_n) \right\|_2^2, \\ S_{\mathcal{B}} &= \frac{1}{|\mathcal{B}|\sqrt{E_0}} \sum_{n\in\mathcal{B}} \sum_{j=1}^{m} \|\widehat{\mathbf{G}}_j(\mathbf{d}_n)\|_2, \\ L_{\mathcal{B}} &= R_{\mathcal{B}} + \lambda S_{\mathcal{B}}. \end{aligned} \tag{57}$$

The coefficient λ is fixed throughout each training run. We evaluate

$$\lambda \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$$

and use $\lambda = 0.1$ uniformly in the main figure. Neither the penalty coefficient nor the reported checkpoint is selected using test results.

The training procedure does not use dual updates, a calibration subset, or an enforced reconstruction constraint. The stored threshold $\epsilon_{\text{rec}} = 10^{-3}$ is used only as a reconstruction diagnostic.

Optimization settings. Every model receives $10{,}000$ Adam updates, with moment coefficients $(0.9, 0.999)$, numerical epsilon 10^{-8}, zero weight decay, and gradient-norm clipping at 5.

The initial encoder and decoder learning rates are 10^{-3} and 3×10^{-4}, respectively. After 70% of the updates, cosine decay reduces both rates to 0.2 times their initial values.

Training observations are sampled uniformly with replacement. The batch size is 1024 for the small and large reference configurations, and 10240 for rank-six concepts. Each fit therefore uses $10{,}240{,}000$ and $102{,}400{,}000$ sampled observations, respectively.

Initialization and training protocols. We compare graph-initialized staged training with random-initialized joint training.

In the graph-initialized staged protocol, the decoder begins at the selected dictionary and remains fixed for the first 5000 updates while the encoder is trained. The encoder and decoder are then trained jointly for another 5000 updates.

In the random-initialized joint protocol, the decoder begins from random initialization and both components are updated throughout all $10{,}000$ steps.

The protocols use the same architecture, training objective, total number of updates, and minibatch stream for a fixed network seed. For $\nu \in \{1, 2\}$, the random decoder uses seed $\nu + 3{,}000{,}000$ and the minibatch sampler uses seed $\nu + 1{,}000{,}000$. The graph-initialized dictionary is reused across penalty coefficients and input-feature maps for the same world and seed.

These protocols differ in both decoder initialization and the schedule for updating the decoder.

D.5 Learning Objective and Optimization

Contribution geometry and model capacity. The recovery target is the contribution in observed representation space, not its coordinates in a particular concept basis.

For any invertible $\mathbf{R}_j$, the change of coordinates

$$(\mathbf{B}_j, \mathbf{u}_j) \longmapsto (\mathbf{B}_j\mathbf{R}_j, \mathbf{R}_j^{-1}\mathbf{u}_j)$$

preserves $\mathbf{B}_j\mathbf{u}_j$. It therefore preserves each contribution, its norm, and the reconstructed difference.

Our estimator uses orthonormal columns within each block to fix a normalized coordinate system:

$$\|\mathbf{B}_j\mathbf{u}_j\|_2 = \|\mathbf{u}_j\|_2.$$

Within this parameterization, orthogonal changes of basis preserve the blockwise orthonormality constraint. No orthogonality is imposed between different concept blocks.

The prescribed block count and width (m, r) restrict the estimator. Too little capacity can prevent the model from representing the true contributions, while additional capacity may allow redundant or split representations.

What group sparsity enforces. The group-sparsity penalty acts on each contribution as a whole. It allows multiple coordinates to represent one changing concept while encouraging entire contribution blocks to become zero.

The selection result in Theorem 3.7 shows how this criterion distinguishes admissible population decompositions under exact reconstruction and the stated structural conditions. The penalty alone does not enforce those conditions.

For example, if the parameterization permits two blocks to represent the same contribution, then for $0 < a < 1$,

$$\mathbf{G} = a\mathbf{G} + (1-a)\mathbf{G}, \qquad \|a\mathbf{G}\|_2 + \|(1-a)\mathbf{G}\|_2 = \|\mathbf{G}\|_2. \tag{58}$$

Thus the norm penalty alone cannot distinguish a contribution from this duplicated representation. The pieces are parallel and have identical supports, so this split violates candidate Local Directness and does not provide the distinct support patterns needed for Support Intersection. The example therefore illustrates the importance of the structural conditions in Theorem 3.7, rather than contradicting its conclusion.

Even with a fixed dictionary, exact coefficient recovery by group-norm minimization requires conditions on the dictionary geometry. Mixed group-norm recovery under suitable conditions is studied by Eldar & Mishali (2009).

The following standard sufficient condition makes this distinction explicit.

Write $\mathbf{Q} = [\mathbf{Q}_1 \ \cdots \ \mathbf{Q}_m]$, where $\mathbf{Q}_j^\top \mathbf{Q}_j = \mathbf{I}_r$, and define

$$J(\mathbf{v}) = \sum_{j=1}^{m} \|\mathbf{v}_j\|_2.$$

Proposition D.2 (A sufficient condition for exact group coding)**.** *Fix $S \subseteq [m]$. Suppose every nonzero $\mathbf{h} \in \ker \mathbf{Q}$ satisfies*

$$\sum_{j \in S} \|\mathbf{h}_j\|_2 < \sum_{j \notin S} \|\mathbf{h}_j\|_2. \tag{59}$$

Then, for every $\mathbf{v}^\star$ supported on S, $\mathbf{v}^\star$ is the unique minimizer of $J(\mathbf{v})$ subject to

$$\mathbf{Q}\mathbf{v} = \mathbf{Q}\mathbf{v}^\star.$$

Proof. Every other feasible vector has the form $\mathbf{v}^\star + \mathbf{h}$ for a nonzero $\mathbf{h} \in \ker \mathbf{Q}$. The triangle inequality gives

$$\begin{aligned} J(\mathbf{v}^\star + \mathbf{h}) - J(\mathbf{v}^\star) &\geq \sum_{j \notin S} \|\mathbf{h}_j\|_2 - \sum_{j \in S} \|\mathbf{h}_j\|_2 \\ &> 0. \end{aligned}$$

□

Condition (59) controls all directions in the kernel of the dictionary. Local Directness on the observed support unions does not by itself establish this objective comparison.

The proposition concerns exact coding with a fixed dictionary. Our estimator also learns the dictionary, uses a shared encoder, and balances reconstruction error against a fixed penalty.

Pair reversal and shared inference. The odd encoder guarantees

$$\widehat{\mathbf{G}}_j(-\mathbf{d}) = -\widehat{\mathbf{G}}_j(\mathbf{d}).$$

This builds pair-reversal symmetry into the estimated contributions.

Recovering a single-input concept function from these pairwise predictions would require additional structure. For example, writing

$$\widehat{\mathbf{G}}_j(\mathbf{x}, \mathbf{x}') = \widehat{\mathbf{G}}_j(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}')),$$

the predicted contributions would need to satisfy

$$\sum_{a=1}^{p} \widehat{\mathbf{G}}_j(\mathbf{x}_a, \mathbf{x}_{a+1}) = \mathbf{0}, \qquad \mathbf{x}_{p+1} = \mathbf{x}_1,$$

on cycles where the representation is imposed. Oddness alone does not imply this identity.

The encoder also provides a shared coefficient inference rule across observations. To describe the difference between shared prediction and independent coefficient optimization, consider a fixed dictionary $\mathbf{Q}$ and define

$$\ell_{\mathbf{Q}}(\mathbf{d}, \mathbf{v}) = \frac{1}{2} \left\| \frac{\mathbf{d}}{\sqrt{E_0}} - \mathbf{Q}\mathbf{v} \right\|_2^2 + \lambda_{\mathrm{c}} J(\mathbf{v}),$$

where $\lambda_{\mathrm{c}} > 0$.

When the expectations are finite, the amortization gap is

$$\mathcal{A}_{\mathbf{Q}}(\boldsymbol{\theta}) = \mathbb{E}\Big[\ell_{\mathbf{Q}}\left(\mathbf{d}, \frac{\mathbf{u}_{\boldsymbol{\theta}}(\mathbf{d})}{\sqrt{E_0}}\right) - \min_{\mathbf{v}} \ell_{\mathbf{Q}}(\mathbf{d}, \mathbf{v})\Big] \geq 0. \tag{60}$$

For fixed $\mathbf{Q}$ and $\lambda_{\mathrm{c}} = \lambda/2$, this coding loss is one half of the normalized per-observation training objective.

The gap measures the difference between the shared encoder's prediction and independent coefficient optimization at the same dictionary.

Why the encoder is trained first. The staged protocol begins by fitting the encoder while holding the initialized dictionary fixed.

For the reconstruction term, with encoder output fixed,

$$\nabla_{\mathbf{B}_j} \frac{1}{2} \|\widehat{\mathbf{d}} - \mathbf{d}\|_2^2 = (\widehat{\mathbf{d}} - \mathbf{d}) \mathbf{u}_{\boldsymbol{\theta},j}(\mathbf{d})^\top.$$

Early dictionary updates therefore depend on coefficients produced by an encoder that has not yet adapted to the initialized spaces. Freezing the decoder allows the encoder to learn coefficients in these spaces before the spaces themselves are updated.

Joint training can then adjust both the concept bases and their coefficient map. The fixed penalty balances reconstruction error and contribution norm.

D.6 Prediction and Native Sparsity

All reported predictions use the final checkpoint after 10,000 updates. For each observed difference, the model predicts contributions directly using the thresholded encoder in Equation (56) and the learned concept bases.

Prediction requires one encoder evaluation followed by blockwise decoding. We do not perform proximal refinement, least-squares amplitude refitting, or any other iterative coefficient inference at prediction time. No true-support mask or sample-dependent support budget is used.

Reference fits are validated every 500 updates and saved every 1000 updates. Rank-six fits use validation and saving intervals of 50 and 100 updates, respectively. Intermediate checkpoints do not determine the reported predictor.

Formal alignment and test prediction batches contain at most 4096 observations. Predictions are fixed before ground-truth contribution scoring.

What native sparsity guarantees. The block soft-threshold in Equation (56) is the proximal map of $\tau_j\|\cdot\|_2$ applied to the encoder's pre-threshold block. It produces exact block sparsity: block j is active precisely when $\|\mathbf{v}_{\boldsymbol{\theta},j}(\mathbf{d})\|_2 > \tau_j$, and its contribution is exactly zero otherwise. Since each decoder block has orthonormal columns, the decoded contribution satisfies

$$\|\widehat{\mathbf{G}}_j(\mathbf{d})\|_2 = \big(\|\mathbf{v}_{\boldsymbol{\theta},j}(\mathbf{d})\|_2 - \tau_j\big)_+.$$

For an active block, thresholding preserves the direction of $\mathbf{B}_j\mathbf{v}_{\boldsymbol{\theta},j}(\mathbf{d})$ and reduces its magnitude by τ_j.

The learned block-specific thresholds therefore determine a pair-specific active set, while retaining continuous variation in the magnitude and direction of active contributions. The same thresholded coefficients enter reconstruction and regularization during training and produce the reported contributions at prediction time. Sparsity is thus built directly into the learned contribution map.

D.7 Sources of Recovery Error and Experimental Evidence

Model and statistical error. If the observed representation is only approximately additive,

$$\mathbf{f}(\mathbf{x}) = \mathbf{b} + \sum_k \mathbf{C}_k(\mathbf{x}) + \mathbf{e}_{\mathrm{rep}}(\mathbf{x}),$$

then each representation difference contains

$$\mathbf{e}_{\mathrm{rep}}(\mathbf{x}) - \mathbf{e}_{\mathrm{rep}}(\mathbf{x}').$$

Exact contribution identifiability does not bound this residual or its effect on estimation. The current synthetic benchmarks use exact mixtures, but finite sample size can still make rare supports and weak change directions difficult to estimate.

The prescribed block count and width can also introduce approximation error when they cannot represent the generating contributions.

Equations (53) and (55) describe moment-estimation and eigenspace errors conditional on appropriate target moments.

Error from an estimated dictionary. Errors in the estimated concept spaces limit the accuracy of individual contributions.

For $m = K$ and a fixed permutation σ, let

$$\widehat{U}_k = \mathrm{range}(\mathbf{B}_{\sigma(k)}).$$

Every prediction in these spaces satisfies

$$\sum_k \|\widehat{\mathbf{G}}_{\sigma(k)} - \mathbf{G}_k\|_2^2 \geq \sum_k \|(\mathbf{I} - \mathbf{P}_{\widehat{U}_k})\mathbf{G}_k\|_2^2. \tag{61}$$

To see this, decompose each prediction error into its projection onto $\widehat{U}_k$ and the orthogonal component

$$-(\mathbf{I} - \mathbf{P}_{\widehat{U}_k})\mathbf{G}_k.$$

The squared norms of these components add.

Thus, when the dictionary is fixed, coefficient optimization cannot remove error caused by components of the true contributions outside the estimated spaces.

When reconstruction controls contribution error. Now suppose matched predictions and true contributions belong to the same spaces U_k^{Δ}. Write

$$\mathbf{e}_k = \widehat{\mathbf{G}}_{\sigma(k)} - \mathbf{G}_k$$

and let $A = \{k : \mathbf{e}_k \neq \mathbf{0}\}$.

Suppose A is nonempty and is contained in $T \cup T'$ for some $T, T' \in \mathcal{F}_\Delta$ satisfying Local Directness. Choose orthonormal bases $\mathbf{Z}_k$ of U_k^Δ, and concatenate the bases for $k \in A$ into $\mathbf{Z}_A$. Then

$$\sum_k \|\mathbf{e}_k\|_2^2 \le \frac{\|\widehat{\mathbf{d}} - \mathbf{d}\|_2^2}{\sigma_{\min}(\mathbf{Z}_A)^2}. \tag{62}$$

Indeed, write $\mathbf{e}_k = \mathbf{Z}_k \mathbf{a}_k$. Block orthonormality gives

$$\sum_k \|\mathbf{e}_k\|_2^2 = \|\mathbf{a}_A\|_2^2,$$

while

$$\widehat{\mathbf{d}} - \mathbf{d} = \mathbf{Z}_A \mathbf{a}_A.$$

Local Directness makes $\mathbf{Z}_A$ injective, yielding the bound. When A is empty, all contribution errors are zero.

This result separates uniqueness from stability. Even with the correct concept spaces, nearly dependent directions can make contribution recovery sensitive to small reconstruction errors.

Interpretation of the current experiments. Figure 1 reports both aggregate reconstruction and individual contribution recovery under the native-prediction protocol.

The Directness and Support Intersection interventions introduce specific ambiguity mechanisms, while selected sums of contributions can remain accurately reconstructed. The Conditional Richness intervention shows that recovery can also remain accurate when this sufficient population condition is violated.

The penalty sweeps in Appendix F show the difference between dense solutions at small penalties and shrinkage or collapse at large penalties.

D.8 Numerical Environment and Computational Budgets

The experiments used two NVIDIA GeForce RTX 3090 GPUs with 24 GB memory each, CUDA 12.6, Python 3.10.12, PyTorch 2.11.0, NumPy 2.2.6, SciPy 1.15.3, and scikit-learn 1.7.2. TF32 was disabled.

Training and prediction use float32. SSC self-expression and spectral grouping use CPU float64. Graph construction, NMF, and moment estimation use float32, with float64 accumulation of the NMF objective. Each worker uses two CPU threads.

Independent fits are batched in groups of at most six, without sharing model parameters or optimizer state.

Each initializer uses 4096 discovery observations, two 300-step NMF fits, four SSC self-expression problems with at most 4096 iterations each, and four validation coding problems with the iteration budgets described above.

The observation graph is shared across network seeds within each world and condition. The selected dictionary is reused across input-feature maps and penalty coefficients for that world and seed.

Exact nearest-neighbor search requires $O(n_0^2 d)$ arithmetic. The search is computed in tiles and the graph is stored sparsely.

The experiment manifests retain the commands, software versions, random seeds, and source hashes needed to reproduce the runs.

E Data and Evaluation Protocols

E.1 Gaussian DAG Pairs

Following the dependent-latent simulations in ConCA (Liu et al., 2026a, Appendix M), we generate representations with known multidimensional contributions. Let $\mathbf{u}_i \in \mathbb{R}^{r_i}$ and $\mathbf{u} =$

$(\mathbf{u}_1^\top, \dots, \mathbf{u}_K^\top)^\top \in \mathbb{R}^q$, with $q = \sum_{i=1}^K r_i$. A linear Gaussian directed acyclic graph specifies

$$\begin{aligned} &\mathbf{u} = \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q), \\ &\boldsymbol{\Sigma} = (\mathbf{I}_q - \mathbf{W})^{-1}(\mathbf{I}_q - \mathbf{W})^{-\top}, \qquad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = (\mathbf{I}_q - \mathbf{W})^\top(\mathbf{I}_q - \mathbf{W}). \end{aligned} \tag{63}$$

We sample a topological order uniformly. At each node, the number of parents is the smaller of the number of preceding nodes and $1 + \mathrm{Bernoulli}(1/2)$; parents are selected uniformly without replacement. For a node with $k_i > 0$ parents, each edge matrix is an independent Gaussian matrix rescaled to spectral norm $a/\sqrt{k_i}$. We use $a = 0.5$ for dependent nodes and $a = 0$ for the independent control. Node coordinates are not standardized. Separately, we draw independent Gaussian $d \times r_i$ matrices (using the paired row-prefix construction below in the rank-six experiments) and obtain orthonormal block decoders $\mathbf{B}_i^\star$ by thin QR decomposition. We define $\mathbf{f}(\mathbf{u}) = \mathbf{b} + \sum_i \mathbf{B}_i^\star \mathbf{u}_i$, where the entries of $\mathbf{b}$ are independent $\mathcal{N}(0, 0.01)$ variables.

Paired observations. The baseline configurations use

$$\Pr(|S| = s) = \frac{\mathbf{1}\{2 \le s \le s_{\max}\}}{s_{\max} - 1}, \qquad \Pr(S = T \mid |S| = s) = \binom{K}{s}^{-1} \quad (|T| = s).$$

For each pair, sample S independently of the latent draw and then draw $\mathbf{u}' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Preserve $\mathbf{u}_{-S} = \mathbf{u}'_{-S}$ and draw $\mathbf{u}_S$ independently from the conditional distribution of $\mathbf{u}'_S$ given $\mathbf{u}'_{-S}$:

$$\mathbf{u}_S \mid \mathbf{u}'_{-S}, S \sim \mathcal{N}\left(-\boldsymbol{\Lambda}_{SS}^{-1}\boldsymbol{\Lambda}_{S,-S}\mathbf{u}'_{-S}, \boldsymbol{\Lambda}_{SS}^{-1}\right). \tag{64}$$

Both endpoints have marginal law $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Their observed difference and generating contributions are

$$\mathbf{d} = \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{u}') = \sum_i \mathbf{G}_i, \qquad \mathbf{G}_i = \mathbf{B}_i^\star(\mathbf{u}_i - \mathbf{u}'_i), \qquad (\mathbf{u} - \mathbf{u}')_S \mid S \sim \mathcal{N}(\mathbf{0}, 2\boldsymbol{\Lambda}_{SS}^{-1}). \tag{65}$$

Conditional resampling preserves the unselected latent coordinates; the selected coordinates are drawn jointly using the conditional covariance in Equation (64). The observations contain no additional noise.

Proposition E.1 (Generating assumptions)**.** *Suppose the support size has positive probability on each integer $s_{\min}, \dots, s_{\max}$, with $1 \le s_{\min} \le s_{\max} < K$, and the support is uniform conditional on its size. If the sum of the largest $\min(2s_{\max}, K)$ block dimensions is at most d, then, with probability one over the block decoders, the contributions in Equation (65) satisfy the true clause of Assumption 3.2, support separation (Assumption 3.4), and conditional richness (Assumption 3.3). Their actual spaces are $U_i^\Delta = \mathrm{range}(\mathbf{B}_i^\star)$ and they have finite second moments.*

Proof. Acyclicity makes $\mathbf{I}_q - \mathbf{W}$ invertible, so $\boldsymbol{\Lambda}$ and all of its principal submatrices are positive definite. Conditional on $\mathbf{u}'_{-S}$ and S, the two selected endpoint blocks are independent Gaussians with the same conditional mean and covariance. Subtraction therefore gives Equation (65). Every edited block is nonzero almost surely and has full support in its coordinate space, hence the actual support equals S almost surely and $U_i^\Delta = \mathrm{range}(\mathbf{B}_i^\star)$.

For every fixed union of at most $2s_{\max}$ blocks, independent Gaussian matrices have full column rank almost surely under the stated dimension bound. QR changes the within-block bases without changing their spans. There are finitely many such unions, so all required directness statements hold simultaneously almost surely. On each actual support, the injective map $\mathbf{B}_S^\star$ sends a nondegenerate Gaussian to a distribution with positive density on V_S^Δ; every proper linear subspace has probability zero. Finally, the maximal supports are all $s_{\max}$-subsets. For any $j \ne i$, there is such a support containing i and excluding j, which proves support separation. Finite Gaussian moments prove the last claim. □

Connection to ConCA coordinates. For $\mathbf{H}_i \in \mathbb{R}^{(r_i+1)\times r_i}$ satisfying $\mathbf{H}_i^\top \mathbf{H}_i = \mathbf{I}$ and $\mathbf{H}_i^\top \mathbf{1} = \mathbf{0}$, set $\mathbf{p}_i(\mathbf{u}) = \mathrm{softmax}(\mathbf{H}_i \mathbf{u}_i)$ and $\mathbf{A}_i = \mathbf{B}_i^\star \mathbf{H}_i^\top$. Then $\mathbf{A}_i \log \mathbf{p}_i(\mathbf{u}) = \mathbf{B}_i^\star \mathbf{u}_i$ exactly.

E.2 Configurations, Splits, and Replication

Configurations. The main experiments use $K = 32$ and $r = 6$, with $(d, s_{\max}) = (64, 4), (128, 4), (128, 8), (256, 4)$ and DAG strength $a = 0.5$. The local-directness dimension bounds are respectively $48 \le 64$, $48 \le 128$, $96 \le 128$, and $48 \le 256$. Thus Proposition E.1 applies to all four. Only the $d = 256$ configuration has a full-column-rank complete decoder ($Kr = 192$); the others are overcomplete. The support family has $\binom{32}{4} = 35{,}960$ or $\binom{32}{8} = 10{,}518{,}300$ maximal members, so finite discovery data need not cover all population change patterns.

Additional reference configurations use $(K, r, d, s_{\max}) = (6, 2, 10, 2)$ and $(32, 3, 64, 4)$, each with dependent ($a = 0.5$) and independent ($a = 0$) concepts. These also satisfy the proposition's dimension bound. The reference configurations change several factors together and are not an isolated intervention on concept count or dimension.

Splits and paired random draws. Each generated world contains 131,072 training, 8,192 validation, 4,096 alignment, and 8,192 test pairs. Discovery uses the first 4,096 training observations; dictionary selection uses the first 1,024 validation observations, leaving 7,168 for a separate initialization diagnostic. No alignment or test contributions enter model fitting or dictionary selection.

The reference worlds have seeds 2026091401, 2026091402, and 2026091403; rank-six worlds use 2026091411, 2026091412, and 2026091413. Each world has two network seeds, 1 and 2. Within a reference scale, dependent and independent conditions share topology, decoder draws, support draws, and Gaussian random-number streams. For rank-six concepts, a common 256×6 Gaussian draw per concept supplies the first d rows before blockwise QR. All four settings share DAG parameters; the three $s_{\max} = 4$ settings also share latent endpoints and supports. The $s_{\max} = 8$ setting uses its specified support law with the same random-number streams. No world is discarded based on its recovery score.

E.3 Comparisons and Experimental Matrix

The baseline/input study crosses eight configurations, two input maps (identity and Fourier), seven fixed penalty weights, two training protocols, three worlds, and two network seeds: 1,344 fitted predictors in 224 cells. The two protocols are graph and random, defined in Appendix D.4; there is no random-staged or initial-dictionary-only evaluation in this current matrix. One graph initializer per configuration/world/seed is reused across its fourteen input/penalty arms. Random decoder seeds are independent of encoder parameter draws, so widening the input projection does not change the random dictionary. The wider Fourier projection nevertheless changes encoder capacity and its initialization, as well as the input features.

The robustness study crosses two configurations ($d = 128$, $s_{\max} = 4, 8$), ten conditions, seven penalties, two protocols, three worlds, and two seeds: 1,680 fitted predictors in 280 cells. All use identity inputs. The ten conditions comprise a matched baseline and three levels of each intervention below. Each condition has its own graph initializer, shared across penalties, so a change in generator can affect both estimated geometry and subsequent coefficient learning. The main figure fixes $\lambda = 0.1$ across every setting and intervention, rather than reporting a separate best test coefficient for each. Appendices F.2 and F.3 show the full grid.

E.4 Interventions on the Generating Assumptions

Interventions share the original DAG, decoder draws, support-size stream, and Gaussian random arrays within a world. Auxiliary randomness for the interventions uses separate deterministic streams. All conditions retain exact additive contributions and finite Gaussian moments. Each intervention changes one population condition of the generating decomposition; it can nevertheless affect several numerical aspects of the estimator.

Local Directness: shared concept directions. Write $(\mathbf{B}_1^\star)^\top \mathbf{B}_2^\star = \mathbf{U} \operatorname{diag}(\cos\theta_1, \dots, \cos\theta_6) \mathbf{V}^\top$, and set $\mathbf{A} = \mathbf{B}_1^\star \mathbf{U}$, $\mathbf{D} = \mathbf{B}_2^\star \mathbf{V}$. For $h \in \{1, 3, 6\}$, replace the first h columns of $\mathbf{D}$ by those of $\mathbf{A}$, then set the modified second block to $\mathbf{D}_h \mathbf{V}^\top$. The baseline $h = 0$ leaves the original decoder unchanged. Each block still has rank six, while the concatenation of the first two blocks has rank $12 - h$. The support distribution and conditional redraw remain

unchanged. Gaussian changes are rich on each actual combined span, and Support Intersection still holds, while Local Directness fails.

Conditional Richness: a proper joint change subspace. Keep the decoder and support law fixed. For support S, let $q_S = 6|S|$, $\boldsymbol{\mu} = -\boldsymbol{\Lambda}_{SS}^{-1}\boldsymbol{\Lambda}_{S,-S}\mathbf{u}'_{-S}$, and $\mathbf{L} = \mathrm{chol}(\boldsymbol{\Lambda}_{SS})^{-\top}$, so $\mathbf{L}\mathbf{L}^\top = \boldsymbol{\Lambda}_{SS}^{-1}$. Write $\mathbf{a} = \mathbf{L}^{-1}(\mathbf{u}'_S - \boldsymbol{\mu})$, draw independent $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{q_S})$, and define

$$\begin{aligned} \ell_S &= \max\{6, \lceil (1-\beta) q_S \rceil\}, \qquad & \mathbf{Q}_S^\top \mathbf{Q}_S &= \mathbf{I}_{\ell_S}, \\ \mathbf{u}_S &= \mathbf{u}'_S + \mathbf{L}\mathbf{Q}_S\mathbf{Q}_S^\top(\mathbf{b} - \mathbf{a}), \qquad & \beta &\in \{0.25, 0.50, 0.75\}. \end{aligned} \tag{66}$$

The columns of $\mathbf{Q}_S$ are a prefix of a fixed Gaussian-QR orthogonal basis indexed by world and support, independent of split and severity. We require every six-row block of $\mathbf{L}\mathbf{Q}_S$ to have rank six; this is checked at the narrowest prefix and at each realized severity. The baseline $\beta = 0$ uses the original conditional redraw.

For $\mathbf{P} = \mathbf{Q}_S\mathbf{Q}_S^\top$, the standardized new endpoint is $(\mathbf{I}-\mathbf{P})\mathbf{a}+\mathbf{P}\mathbf{b}$. Its covariance is $(\mathbf{I}-\mathbf{P})+\mathbf{P} = \mathbf{I}$, so both endpoints retain their conditional and unconditional Gaussian marginals. The joint change covariance is $2\mathbf{L}\mathbf{P}\mathbf{L}^\top$, of rank $\ell_S < q_S$. Each selected concept still changes with a nondegenerate rank-six marginal, so actual concept spaces and support indicators are unchanged. Generic decoder injectivity on S places the observed change in a fixed proper subspace of V_S^Δ, violating Conditional Richness on maximal supports. Local Directness and Support Intersection are preserved.

Support Intersection: always-coactive pairs. Keep the original decoder and full conditional Gaussian redraw. Bind the first $b \in \{1, 4, 8\}$ disjoint pairs $(1, 2), (3, 4), \ldots, (2b-1, 2b)$. First draw the original support size s. Draw the number J of selected bound pairs from

$$\Pr(J = j \mid s) \;\propto\; \binom{b}{j}\binom{32-2b}{s-2j}\exp(\eta_s j), \qquad \mathbb{E}[J \mid s] = \frac{bs}{32}, \tag{67}$$

on the feasible integers. Choose the J pairs uniformly and choose $s - 2J$ unbound singletons uniformly. The tilt η_s is solved by bracketed bisection to moment tolerance 10^{-12}, retaining positive probabilities on every feasible count. A bound member is active with probability $\mathbb{E}[J \mid s]/b = s/32$; an unbound member has probability $(s - 2\mathbb{E}[J \mid s])/(32 - 2b) = s/32$. Thus support sizes and all marginal activation probabilities match the baseline, although joint support patterns change. Members of a bound pair coincide in every support and cannot be separated by intersections of maximal supports. The unchanged generic decoder and full conditional redraw preserve Local Directness and Conditional Richness.

An invertible 45-degree block rotation of the two coordinate vectors, with the inverse rotation applied to the concatenated decoder, preserves that pair's combined contribution. The resulting blocks retain rank six, and both rotate to zero together on inactive pairs. Full Gaussian variation makes them nonzero almost surely on active pairs. This gives a different pair decomposition with the same observations and coactivity pattern.

Structural certificates. Generation records 60 world/condition certificates. They verify block ranks, the prescribed shared dimensions, the joint and marginal change ranks, support-size and marginal-probability identities, and explicit ambiguity witnesses where applicable. Rank checks use float64 with tolerance 10^{-10}; block orthonormality uses tolerance 2×10^{-12}. The generic-position argument in Proposition E.1 establishes the unchanged decoder's population directness. Sampled support-union singular values are numerical diagnostics, not an exhaustive population proof.

E.5 Recovery Metrics

Alignment and FVE. For a fixed predictor, the separate alignment set $\mathcal{A}$ determines

$$\sigma \in \operatorname*{arg\,min}_{\pi \in \mathfrak{S}_K} \sum_{n \in \mathcal{A}} \sum_{i=1}^{K} \|\widehat{\mathbf{G}}_{n,\pi(i)} - \mathbf{G}_{ni}\|_2^2.$$

We solve this assignment by the Hungarian algorithm and keep the same permutation on every test pair. Let $\widehat{\mathbf{d}}_n = \sum_j \widehat{\mathbf{G}}_{nj}$. The two main scores are

$$\text{Delta FVE} = 1 - \frac{\sum_n \|\widehat{\mathbf{d}}_n - \mathbf{d}_n\|_2^2}{\sum_n \|\mathbf{d}_n\|_2^2}, \tag{68}$$

$$\text{Block FVE} = 1 - \frac{\sum_{n,i} \|\widehat{\mathbf{G}}_{n,\sigma(i)} - \mathbf{G}_{ni}\|_2^2}{\sum_{n,i} \|\mathbf{G}_{ni}\|_2^2}. \tag{69}$$

These are uncentered energy-normalized scores, not centered coefficients of determination. Contributions retain their predicted amplitudes, including predictions for unchanged concepts. There is no per-sample permutation, regression alignment, or test-based grouping. FVE can be negative.

Individual blocks, affected groups, and pair sums. For a concept set A, restrict both sums in Block FVE to $i \in A$. This gives energy-weighted affected and unaffected group scores. Figure 1(b) instead uses the arithmetic mean of the two individual per-block FVE scores for $A = \{1, 2\}$, including at $h = 0$; it does not substitute the overall baseline Block FVE. The distinction avoids changing the evaluation target across levels. For a predeclared pair (i, j), define

$$\text{Pair FVE}_{ij} = 1 - \frac{\sum_n \|\widehat{\mathbf{G}}_{n,\sigma(i)} + \widehat{\mathbf{G}}_{n,\sigma(j)} - \mathbf{G}_{ni} - \mathbf{G}_{nj}\|_2^2}{\sum_n \|\mathbf{G}_{ni} + \mathbf{G}_{nj}\|_2^2}. \tag{70}$$

The main figure tracks the same pair $(1, 2)$ at every directness and binding level, including the matched baseline. The report also retains all other predeclared pair scores; the plotted pair is not selected by its test score.

Support, subspaces, and consistency. Exact predicted support is defined by literal nonzero contribution blocks. We record precision, recall, exact support accuracy, mean active count, exact-zero block fraction, and all-zero observation fraction. False-positive energy is the predicted energy on truly inactive blocks divided by total true contribution energy. In the robustness study, decoder overlap uses the alignment-only contribution assignment and averages $\|\widehat{\mathbf{Q}}_{\sigma(i)}^\top \mathbf{B}_i^\star\|_F^2/r$, with $\widehat{\mathbf{Q}}$ orthonormal. Separate optimal-subspace assignments in baseline diagnostics are distinguished from this fixed alignment. Cross-seed diagnostics align each seed separately on alignment data and compare test contributions, reporting symmetric normalized squared error, cosine similarity, and support Jaccard. Two collapsed zero predictors can agree perfectly without recovering any contributions; these cases are flagged.

Replication and uncertainty. Let q_{wt} denote a score for world $w \in \{1, 2, 3\}$ and network seed $t \in \{1, 2\}$. We report

$$\bar{q}_w = \tfrac{1}{2} \sum_{t=1}^{2} q_{wt}, \qquad \bar{q} = \tfrac{1}{3} \sum_{w=1}^{3} \bar{q}_w, \qquad s_q = \left[\tfrac{1}{2} \sum_{w=1}^{3} (\bar{q}_w - \bar{q})^2 \right]^{1/2}.$$

Error bars and tabulated uncertainties are s_q, the sample standard deviation of world means, not a standard error or a confidence interval. For the mean of the first two per-block scores, we form that mean within each world before computing s_q; we do not average the two standard deviations. All worlds and seeds are retained, including poorly recovering or collapsed models. Three-decimal rounding can display a small nonzero standard deviation as 0.000 or a near-perfect score as 1.000.

Prediction freezing and storage. All formal alignment and test predictions are completed and hashed before contribution scoring. The baseline/input study stores computed float32 contributions losslessly; the robustness study stores float32 coordinates and decoders losslessly and reconstructs contributions on CPU. Recorded GPU/CPU decoding probes use absolute tolerance 2×10^{-6} and relative tolerance 10^{-5}. The resulting arithmetic can produce small baseline differences between the two studies, so intervention comparisons use the baseline from their own study.

E.6 Real-World Experimental Protocols

The protocols below specify Figure 2. Decomposition training is unsupervised; semantic labels are used for calibration and evaluation. All results use native final-checkpoint predictions, with means

and sample standard deviations over training seeds 1, 2, and 3. These error bars are not confidence intervals.

E.6.1 StylePTB Semantic Readout

Data and models. We use atomic source–target pairs from StylePTB (Lyu et al., 2021). Auxiliary instructions are excluded from the encoded inputs, and connected sentence families remain within a single partition. Frozen Pythia-70M (Biderman et al., 2023) supplies the final-normalized EOS representation. The compared models are ConAn, SSAE, ReLU-L1 SAE, Top-K SAE, and ConCA, all with latent width $Q = 512$. ConAn uses rank-four blocks; the other models use scalar components. ConAn and SSAE encode representation differences, whereas the conventional SAEs and ConCA difference the codes of the two endpoints. Decoded contributions are expressed in the original activation units.

Readout and label hierarchy. We evaluate 17 fine factors, eight factor clusters, and four factor classes (lexical, syntax, information, and emphasis). A parent label is the logical OR of its descendants, and its score is evaluated directly against that label. Each label uses a component fixed on calibration data. Fine-factor assignments are one-to-one; cluster and class assignments allow reuse. Assignments maximize calibration normalized AP of contribution magnitudes and are held fixed for the quadratic readout.

For each selected component, a regularized logistic probe uses its contribution magnitude together with sign-invariant quadratic direction features. In orthonormal coordinates within its decoder span, these features are the magnitude and the vectorized upper triangle of the normalized direction's outer product, with off-diagonal entries scaled by $\sqrt{2}$. The direction features vanish for a zero contribution. Feature centering and a common RMS scale are fitted on calibration data; the logistic probe uses L2 regularization with $C = 1$ and balanced class weights. Probes and feature transformations remain fixed on test data.

E.6.2 Counterfactual Reconstruction and Intervention

Data and training. We use frozen Qwen3-1.7B (Yang et al., 2025) representations of counterfactual scene-description pairs, each changing object identity, color, material, or size. The evaluation covers 56 fine semantic labels. Models train on 10,000 question-format pairs, with disjoint training, calibration, and test families. Graph- and randomly initialized ConAn both use $Q = 512$, rank-eight blocks and $\lambda = 0.1$; SSAE uses $Q = 512$ and $\beta = 0.1$. All train for 10,000 updates, with both ConAn variants freezing their decoders for the first 2,500 updates.

Component selection and reconstruction. For each semantic label, we rank components by their individual calibration reconstruction benefit and freeze one ConAn block or the eight highest-ranked SSAE features. The intervention vector is their decoded sum, excluding any shared reconstruction offset. Selection is fixed across test pairs and both amplitude conditions. The reconstruction error in panel (c) uses the full model reconstruction in original units:

$$\mathcal{R}_\Delta = \frac{\sum_n \|\mathbf{d}_n - \widehat{\mathbf{d}}_n\|_2^2}{\sum_n \|\mathbf{d}_n\|_2^2}. \tag{71}$$

Thus reconstruction evaluates the aggregate prediction, whereas intervention evaluates the selected contribution.

Two intervention amplitudes. For a decoded contribution $\mathbf{v}$ and its paired representation difference $\mathbf{d}$, the two patches are

$$\mathbf{v}_{\text{native}} = \mathbf{v}, \qquad \mathbf{v}_{\text{matched}} = \begin{cases} \dfrac{\|\mathbf{d}\|_2}{\|\mathbf{v}\|_2}\mathbf{v}, & \|\mathbf{v}\|_2 > 0, \\ \mathbf{0}, & \|\mathbf{v}\|_2 = 0. \end{cases} \tag{72}$$

Rescaling is performed on the decoded vector before applying the language-model output head. Zero vectors remain in the evaluation denominator. Norm matching uses an observed paired difference.

Target success. We add the patch to the source prompt's final-normalized state for the first answer token and score complete candidate words without length normalization. Later-token conditional scores retain their unmodified source-prompt values. A trial succeeds if the target answer has the largest candidate score. Eligible pairs have correct unmodified source and target answers and distinct source and target labels. Panel (c) averages eligible rows within each semantic label, then gives each of the 56 labels equal weight. Both endpoint labels are evaluated and are not treated as independent observations.

Within-pair intervention and cross-context transfer. For panel (d), calibration donors express the same directed change as a test recipient in a different scene. Attribute changes require a different object; object changes require different attributes. Donors with exchanged endpoints are reencoded with the frozen model. Each eligible donor is scored separately, and its own paired difference sets the norm-matched amplitude; the recipient's target difference is not used to scale a transferred contribution. Within-pair intervention instead extracts the contribution and norm from the recipient pair.

Both conditions use identical donor-covered recipients with naturally correct source and target answers and at least one valid unchanged query. The reported target-success score concerns the changed query. We average donors within recipients, query views within recipient families, families within semantic labels, and then labels equally. Panel (d) therefore uses a subset and a different within-label weighting from panel (c); its scores should be compared between conditions on that matched subset. The three training seeds receive equal weight.

F Additional Experimental Results

The results below use the architecture, fixed-penalty training, and native final-checkpoint predictions specified in Appendix D. All 224 baseline/input cells and 280 robustness cells are complete, with six fits per cell. Unless indicated otherwise, values are means $\pm$ sample standard deviations over three world averages, as defined in Appendix E.5. The tables retain every coefficient and training protocol; a low score or collapsed predictor is not removed from the summary.

F.1 Reference Configurations and Input Features

Table 2 includes the four rank-six configurations and the smaller dependent/independent controls at the common coefficient $\lambda = 0.1$. With identity inputs, the graph protocol attains Block FVE 0.927 and 0.979 in the small and large dependent settings, respectively. The independent controls give similar graph scores (0.928 and 0.979). These controls show recovery in both dependent and independent Gaussian data.

Fourier features do not provide a uniform improvement. At $d = 128$, $s_{\max} = 4$, Block FVE changes from 0.983 to 0.982 for graph and from 0.337 to 0.381 for random. At $d = 64$, $s_{\max} = 4$, the graph scores are almost unchanged, whereas the random score falls from 0.685 to 0.648. Because Fourier inputs also add parameters to the input projection, these comparisons concern that complete input parameterization. Identity inputs remain the common setting for the main figure and all assumption interventions.

Table 2: Current baseline experiments at $\lambda = 0.1$. I/F denote identity/Fourier inputs; G/R denote graph/random training. A is mean nonzero block count.

Setting	Input	Protocol	Delta FVE	Block FVE	A
Small, DAG	I	G	0.991 ± 0.000	0.927 ± 0.005	3.78
Small, DAG	I	R	0.990 ± 0.000	0.469 ± 0.248	4.63
Small, DAG	F	G	0.991 ± 0.000	0.929 ± 0.001	3.61
Small, DAG	F	R	0.990 ± 0.000	0.408 ± 0.301	4.50
Small, independent	I	G	0.991 ± 0.000	0.928 ± 0.005	3.76
Small, independent	I	R	0.990 ± 0.000	0.381 ± 0.220	4.67
Small, independent	F	G	0.991 ± 0.000	0.929 ± 0.002	3.64
Small, independent	F	R	0.990 ± 0.001	0.487 ± 0.213	4.48
Large, DAG	I	G	0.986 ± 0.000	0.979 ± 0.000	3.87

Setting	Input	Protocol	Delta FVE	Block FVE	A
Large, DAG	I	R	0.975 ± 0.001	0.132 ± 0.016	9.29
Large, DAG	F	G	0.986 ± 0.000	0.979 ± 0.000	3.85
Large, DAG	F	R	0.975 ± 0.000	0.152 ± 0.016	8.95
Large, independent	I	G	0.986 ± 0.000	0.979 ± 0.000	3.88
Large, independent	I	R	0.975 ± 0.000	0.133 ± 0.023	9.36
Large, independent	F	G	0.986 ± 0.000	0.979 ± 0.000	3.84
Large, independent	F	R	0.975 ± 0.000	0.133 ± 0.014	9.06
$d = 64,\ s_{\max} = 4$	I	G	0.963 ± 0.000	0.845 ± 0.000	10.70
$d = 64,\ s_{\max} = 4$	I	R	0.966 ± 0.000	0.685 ± 0.032	15.93
$d = 64,\ s_{\max} = 4$	F	G	0.963 ± 0.000	0.845 ± 0.000	10.46
$d = 64,\ s_{\max} = 4$	F	R	0.966 ± 0.000	0.648 ± 0.052	15.77
$d = 128,\ s_{\max} = 4$	I	G	0.987 ± 0.000	0.983 ± 0.000	3.54
$d = 128,\ s_{\max} = 4$	I	R	0.975 ± 0.001	0.337 ± 0.060	8.84
$d = 128,\ s_{\max} = 4$	F	G	0.986 ± 0.000	0.982 ± 0.000	3.67
$d = 128,\ s_{\max} = 4$	F	R	0.976 ± 0.000	0.381 ± 0.052	8.46
$d = 128,\ s_{\max} = 8$	I	G	0.975 ± 0.000	0.946 ± 0.000	7.53
$d = 128,\ s_{\max} = 8$	I	R	0.966 ± 0.001	0.463 ± 0.080	15.05
$d = 128,\ s_{\max} = 8$	F	G	0.974 ± 0.000	0.945 ± 0.000	7.52
$d = 128,\ s_{\max} = 8$	F	R	0.966 ± 0.001	0.482 ± 0.054	14.56
$d = 256,\ s_{\max} = 4$	I	G	0.992 ± 0.000	0.991 ± 0.000	3.00
$d = 256,\ s_{\max} = 4$	I	R	0.977 ± 0.000	0.148 ± 0.045	8.45
$d = 256,\ s_{\max} = 4$	F	G	0.992 ± 0.000	0.991 ± 0.000	3.00
$d = 256,\ s_{\max} = 4$	F	R	0.977 ± 0.000	0.120 ± 0.021	8.70

F.2 Penalty Sensitivity

Figure 4 and Table 3 report the entire coefficient grid for both input maps and training protocols. The penalty controls a tradeoff, rather than a monotone route to better contribution recovery. For the $d = 128$, $s_{\max} = 8$ identity-input robustness baseline with graph training, $\lambda = 0.001$ gives Delta FVE 0.999984, Block FVE 0.627, and 32 active blocks on average. At $\lambda = 0.05$, these become 0.988, 0.957, and 11.43; at $\lambda = 0.1$, they are 0.975, 0.946, and 7.53. The true average support size is 5. Increasing the coefficient to 1 reduces mean activity to 2.07 but also reduces Block FVE to 0.173. Thus near-perfect reconstruction, sparse predictions, and accurate contribution recovery are distinct properties. The main figure's common coefficient is not the best test coefficient in every setting.

Table 3: Complete baseline Block FVE sweep. Entries are world-averaged means; world standard deviations are shown in the corresponding sweep figure. No coefficient is selected from test scores.

Setting / condition	Protocol	0.001	0.005	0.01	0.05	0.1	0.5	1
Small, DAG	I/G	0.761	0.669	0.660	0.919	0.927	0.785	0.546
Small, DAG	I/R	-0.091	0.144	0.214	0.396	0.469	0.353	0.414
Small, DAG	F/G	0.770	0.707	0.728	0.920	0.929	0.783	0.545
Small, DAG	F/R	-0.083	0.153	0.226	0.424	0.408	0.380	0.334
Small, independent	I/G	0.766	0.674	0.677	0.919	0.928	0.786	0.544
Small, independent	I/R	-0.097	0.136	0.215	0.359	0.381	0.358	0.411
Small, independent	F/G	0.774	0.712	0.740	0.924	0.929	0.784	0.543
Small, independent	F/R	-0.085	0.154	0.235	0.488	0.487	0.417	0.372
Large, DAG	I/G	0.622	0.662	0.953	0.986	0.979	0.790	0.416
Large, DAG	I/R	-0.318	-0.016	0.056	0.140	0.132	0.465	0.154
Large, DAG	F/G	0.636	0.662	0.954	0.986	0.979	0.789	0.416
Large, DAG	F/R	-0.324	-0.019	0.063	0.117	0.152	0.452	0.193
Large, independent	I/G	0.622	0.671	0.953	0.986	0.979	0.790	0.414
Large, independent	I/R	-0.321	-0.029	0.050	0.128	0.133	0.346	0.117
Large, independent	F/G	0.636	0.660	0.953	0.986	0.979	0.789	0.414
Large, independent	F/R	-0.328	-0.031	0.039	0.104	0.133	0.371	0.153
$d = 64,\ s_{\max} = 4$	I/G	0.316	0.311	0.629	0.785	0.845	0.713	0.353
$d = 64,\ s_{\max} = 4$	I/R	-0.203	-0.104	-0.053	0.526	0.685	0.708	0.188
$d = 64,\ s_{\max} = 4$	F/G	0.321	0.311	0.624	0.790	0.845	0.718	0.353

Setting / condition	Protocol	0.001	0.005	0.01	0.05	0.1	0.5	1
$d = 64,\ s_{\max} = 4$	F/R	-0.203	-0.101	-0.044	0.543	0.648	0.712	0.272
$d = 128,\ s_{\max} = 4$	I/G	0.629	0.889	0.980	0.990	0.983	0.786	0.358
$d = 128,\ s_{\max} = 4$	I/R	-0.317	0.187	0.521	0.278	0.337	0.746	0.009
$d = 128,\ s_{\max} = 4$	F/G	0.638	0.847	0.982	0.990	0.982	0.785	0.358
$d = 128,\ s_{\max} = 4$	F/R	-0.339	0.168	0.428	0.330	0.381	0.755	0.098
$d = 128,\ s_{\max} = 8$	I/G	0.627	0.640	0.895	0.957	0.946	0.633	0.173
$d = 128,\ s_{\max} = 8$	I/R	-0.399	0.174	0.402	0.415	0.463	0.617	0.000
$d = 128,\ s_{\max} = 8$	F/G	0.629	0.641	0.888	0.957	0.945	0.632	0.174
$d = 128,\ s_{\max} = 8$	F/R	-0.408	0.129	0.362	0.470	0.482	0.619	0.064
$d = 256,\ s_{\max} = 4$	I/G	1.000	1.000	1.000	0.998	0.991	0.803	0.357
$d = 256,\ s_{\max} = 4$	I/R	-0.290	-0.058	-0.036	0.049	0.148	0.379	0.000
$d = 256,\ s_{\max} = 4$	F/G	1.000	1.000	1.000	0.998	0.991	0.803	0.359
$d = 256,\ s_{\max} = 4$	F/R	-0.356	-0.032	-0.015	0.023	0.120	0.649	0.000

F.3 Assumption Interventions and Additional Diagnostics

Table 4 gives all ten conditions and both protocols at $\lambda = 0.1$. Figure 5 and Table 5 extend the comparison to every tested coefficient. The intervention study uses its own matched baseline; small arithmetic differences from the baseline/input study are described in Appendix E.5.

Separating affected concepts from their sum. At full overlap ($h = 6$), the energy-weighted Block FVE restricted to concepts 1 and 2 is 0.063 for $s_{\max} = 4$ and 0.083 for $s_{\max} = 8$ under graph training, whereas the corresponding pair-sum scores are 0.950 and 0.933. These subgroup values differ slightly from the arithmetic mean of per-block FVE in Figure 1(b). For eight bound pairs, the unaffected concepts retain Block FVE 0.983 and 0.947, respectively, while the affected concepts score 0.196 and 0.233. The loss in overall contribution recovery is therefore concentrated in the concepts whose support patterns were made inseparable.

Support recovery and geometry. At the unmodified robustness baseline and $\lambda = 0.1$, graph training gives mean decoder overlap 0.999931 and 0.999412 for $s_{\max} = 4$ and 8. However, support precision is only 0.846 and 0.664, with exact support accuracy 0.801 and 0.515; recall is above 0.999 in both settings. The mean predicted active counts, 3.54 and 7.53, exceed the true averages 3 and 5. Accurate spaces and high contribution FVE can therefore coexist with small spurious contributions on unchanged concepts. At higher penalties, improved support precision can instead accompany shrinkage of the true contributions.

Cross-seed consistency. For graph training at $\lambda = 0.1$, symmetric normalized cross-seed contribution error rises from approximately 0.004 to 0.468 when moving from the baseline to eight bound pairs at $s_{\max} = 4$; for $s_{\max} = 8$, it rises from 0.007 to 0.486. This variability is consistent with alternative decompositions of a stable combined contribution, and the full report retains world-level comparisons and flags agreement between collapsed predictors separately.

Table 4: Assumption interventions at $\lambda = 0.1$ (identity inputs). G/R are the two training protocols. Pair FVE uses the fixed pair $(1, 2)$; – denotes an unevaluated metric.

$s_{\max}$	Condition	Protocol	Delta FVE	Block FVE	Pair FVE
4	Baseline	G	0.987 ± 0.000	0.983 ± 0.000	0.983 ± 0.001
4	Baseline	R	0.975 ± 0.001	0.337 ± 0.060	0.438 ± 0.091
4	$h = 1$	G	0.987 ± 0.000	0.982 ± 0.000	0.982 ± 0.001
4	$h = 1$	R	0.977 ± 0.001	0.415 ± 0.036	0.345 ± 0.255
4	$h = 3$	G	0.987 ± 0.000	0.979 ± 0.000	0.981 ± 0.001
4	$h = 3$	R	0.980 ± 0.001	0.570 ± 0.055	0.593 ± 0.070
4	$h = 6$	G	0.988 ± 0.000	0.921 ± 0.008	0.950 ± 0.012
4	$h = 6$	R	0.985 ± 0.002	0.789 ± 0.086	0.858 ± 0.072
4	$\beta = .25$	G	0.987 ± 0.000	0.982 ± 0.000	–
4	$\beta = .25$	R	0.976 ± 0.001	0.354 ± 0.042	–
4	$\beta = .50$	G	0.987 ± 0.000	0.982 ± 0.000	–

$s_{\max}$	Condition	Protocol	Delta FVE	Block FVE	Pair FVE
4	$\beta = .50$	R	0.976 ± 0.000	0.375 ± 0.037	–
4	$\beta = .75$	G	0.987 ± 0.000	0.983 ± 0.000	–
4	$\beta = .75$	R	0.976 ± 0.000	0.355 ± 0.066	–
4	$b = 1$	G	0.987 ± 0.000	0.938 ± 0.003	0.986 ± 0.001
4	$b = 1$	R	0.976 ± 0.001	0.361 ± 0.042	0.761 ± 0.116
4	$b = 4$	G	0.988 ± 0.000	0.788 ± 0.005	0.986 ± 0.001
4	$b = 4$	R	0.978 ± 0.001	0.298 ± 0.018	0.735 ± 0.092
4	$b = 8$	G	0.988 ± 0.000	0.581 ± 0.017	0.986 ± 0.001
4	$b = 8$	R	0.979 ± 0.000	0.227 ± 0.016	0.734 ± 0.172
8	Baseline	G	0.975 ± 0.000	0.946 ± 0.000	0.948 ± 0.004
8	Baseline	R	0.966 ± 0.001	0.463 ± 0.080	0.524 ± 0.089
8	$h = 1$	G	0.975 ± 0.000	0.945 ± 0.000	0.947 ± 0.004
8	$h = 1$	R	0.966 ± 0.001	0.478 ± 0.071	0.486 ± 0.098
8	$h = 3$	G	0.975 ± 0.000	0.941 ± 0.000	0.949 ± 0.003
8	$h = 3$	R	0.969 ± 0.001	0.613 ± 0.033	0.696 ± 0.165
8	$h = 6$	G	0.976 ± 0.000	0.892 ± 0.005	0.933 ± 0.034
8	$h = 6$	R	0.973 ± 0.002	0.745 ± 0.116	0.860 ± 0.117
8	$\beta = .25$	G	0.975 ± 0.000	0.946 ± 0.000	–
8	$\beta = .25$	R	0.966 ± 0.000	0.507 ± 0.054	–
8	$\beta = .50$	G	0.974 ± 0.000	0.946 ± 0.000	–
8	$\beta = .50$	R	0.966 ± 0.001	0.460 ± 0.047	–
8	$\beta = .75$	G	0.975 ± 0.000	0.947 ± 0.000	–
8	$\beta = .75$	R	0.966 ± 0.000	0.473 ± 0.027	–
8	$b = 1$	G	0.975 ± 0.000	0.900 ± 0.004	0.953 ± 0.004
8	$b = 1$	R	0.967 ± 0.001	0.484 ± 0.022	0.717 ± 0.167
8	$b = 4$	G	0.975 ± 0.000	0.767 ± 0.004	0.954 ± 0.003
8	$b = 4$	R	0.968 ± 0.000	0.427 ± 0.009	0.741 ± 0.221
8	$b = 8$	G	0.975 ± 0.000	0.583 ± 0.028	0.954 ± 0.003
8	$b = 8$	R	0.970 ± 0.001	0.349 ± 0.070	0.883 ± 0.058

Table 5: Complete robustness Block FVE sweep. Entries are world-averaged means; world standard deviations are shown in the corresponding sweep figure. No coefficient is selected from test scores.

Setting / condition	Protocol	0.001	0.005	0.01	0.05	0.1	0.5	1
$s_{\max} = 4$, Baseline	G	0.629	0.891	0.980	0.990	0.983	0.786	0.358
$s_{\max} = 4$, Baseline	R	-0.317	0.187	0.521	0.278	0.337	0.746	0.009
$s_{\max} = 4$, $h = 1$	G	0.629	0.897	0.977	0.989	0.982	0.785	0.357
$s_{\max} = 4$, $h = 1$	R	-0.314	0.196	0.488	0.307	0.415	0.765	0.007
$s_{\max} = 4$, $h = 3$	G	0.628	0.894	0.967	0.985	0.979	0.783	0.356
$s_{\max} = 4$, $h = 3$	R	-0.310	0.192	0.492	0.424	0.570	0.756	0.002
$s_{\max} = 4$, $h = 6$	G	0.597	0.859	0.916	0.926	0.921	0.745	0.341
$s_{\max} = 4$, $h = 6$	R	-0.302	0.213	0.553	0.566	0.789	0.736	0.007
$s_{\max} = 4$, $\beta = .25$	G	0.629	0.891	0.979	0.990	0.982	0.786	0.365
$s_{\max} = 4$, $\beta = .25$	R	-0.316	0.196	0.480	0.278	0.354	0.760	0.011
$s_{\max} = 4$, $\beta = .50$	G	0.629	0.883	0.977	0.990	0.982	0.787	0.380
$s_{\max} = 4$, $\beta = .50$	R	-0.318	0.192	0.482	0.309	0.375	0.782	0.096
$s_{\max} = 4$, $\beta = .75$	G	0.629	0.914	0.982	0.991	0.983	0.794	0.414
$s_{\max} = 4$, $\beta = .75$	R	-0.309	0.179	0.472	0.298	0.355	0.781	0.166
$s_{\max} = 4$, $b = 1$	G	0.599	0.856	0.932	0.943	0.938	0.757	0.346
$s_{\max} = 4$, $b = 1$	R	-0.317	0.168	0.453	0.241	0.361	0.721	0.011
$s_{\max} = 4$, $b = 4$	G	0.493	0.710	0.770	0.779	0.788	0.669	0.309
$s_{\max} = 4$, $b = 4$	R	-0.312	0.121	0.399	0.263	0.298	0.613	0.005
$s_{\max} = 4$, $b = 8$	G	0.335	0.493	0.547	0.563	0.581	0.527	0.254
$s_{\max} = 4$, $b = 8$	R	-0.307	0.016	0.256	0.139	0.227	0.479	0.006
$s_{\max} = 8$, Baseline	G	0.627	0.639	0.896	0.957	0.946	0.633	0.173
$s_{\max} = 8$, Baseline	R	-0.399	0.174	0.402	0.415	0.463	0.617	0.000
$s_{\max} = 8$, $h = 1$	G	0.626	0.639	0.893	0.955	0.945	0.632	0.173
$s_{\max} = 8$, $h = 1$	R	-0.397	0.171	0.404	0.483	0.478	0.619	0.000
$s_{\max} = 8$, $h = 3$	G	0.626	0.638	0.886	0.950	0.941	0.630	0.173
$s_{\max} = 8$, $h = 3$	R	-0.394	0.166	0.389	0.538	0.613	0.612	0.000

Setting / condition	Protocol	0.001	0.005	0.01	0.05	0.1	0.5	1
$s_{\max} = 8, h = 6$	G	0.596	0.610	0.849	0.901	0.892	0.606	0.167
$s_{\max} = 8, h = 6$	R	-0.390	0.184	0.423	0.692	0.745	0.600	0.000
$s_{\max} = 8, \beta = .25$	G	0.627	0.639	0.895	0.957	0.946	0.634	0.179
$s_{\max} = 8, \beta = .25$	R	-0.398	0.159	0.402	0.438	0.507	0.626	0.001
$s_{\max} = 8, \beta = .50$	G	0.626	0.638	0.893	0.956	0.946	0.636	0.190
$s_{\max} = 8, \beta = .50$	R	-0.398	0.165	0.401	0.435	0.460	0.628	0.006
$s_{\max} = 8, \beta = .75$	G	0.626	0.640	0.903	0.958	0.947	0.649	0.224
$s_{\max} = 8, \beta = .75$	R	-0.394	0.164	0.387	0.429	0.473	0.648	0.084
$s_{\max} = 8, b = 1$	G	0.592	0.605	0.848	0.907	0.900	0.609	0.167
$s_{\max} = 8, b = 1$	R	-0.397	0.153	0.379	0.391	0.484	0.587	-0.000
$s_{\max} = 8, b = 4$	G	0.500	0.513	0.713	0.768	0.767	0.543	0.153
$s_{\max} = 8, b = 4$	R	-0.390	0.102	0.313	0.416	0.427	0.499	0.001
$s_{\max} = 8, b = 8$	G	0.349	0.365	0.521	0.571	0.583	0.440	0.128
$s_{\max} = 8, b = 8$	R	-0.379	0.033	0.205	0.277	0.349	0.402	0.001

F.4 Numerical Diagnostics and Reproducibility

The robustness experiment completed all 1,680 fits and their prediction and scoring stages, with no missing or failed stages. Of its 120 selected graph initializers, 73 recorded self-expression convergence warnings at the 10^{-7} target. These finite-budget solutions were retained under the eligibility rule in Appendix D.1. In addition, 79 of the 1,680 predictors output zero on at least 99% of test observations.

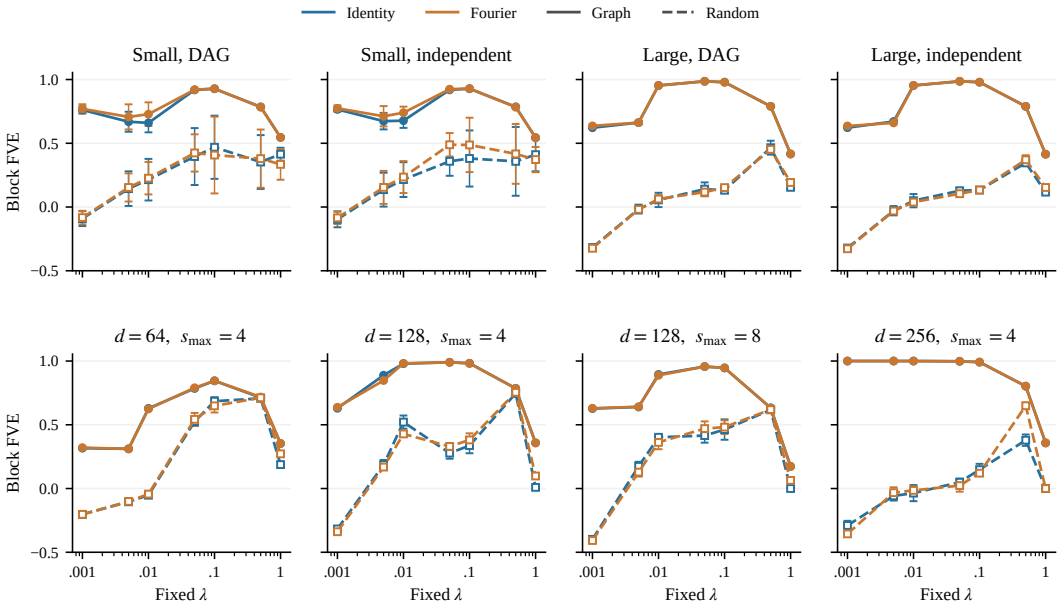

Figure 4: **Input features and penalty sensitivity.** Block FVE over all seven fixed coefficients for eight configurations. Color denotes identity or Fourier input; solid and dashed lines denote graph and random training. Reference settings are specified in Appendix E.2. Error bars show world-level standard deviations; negative FVE values are retained.

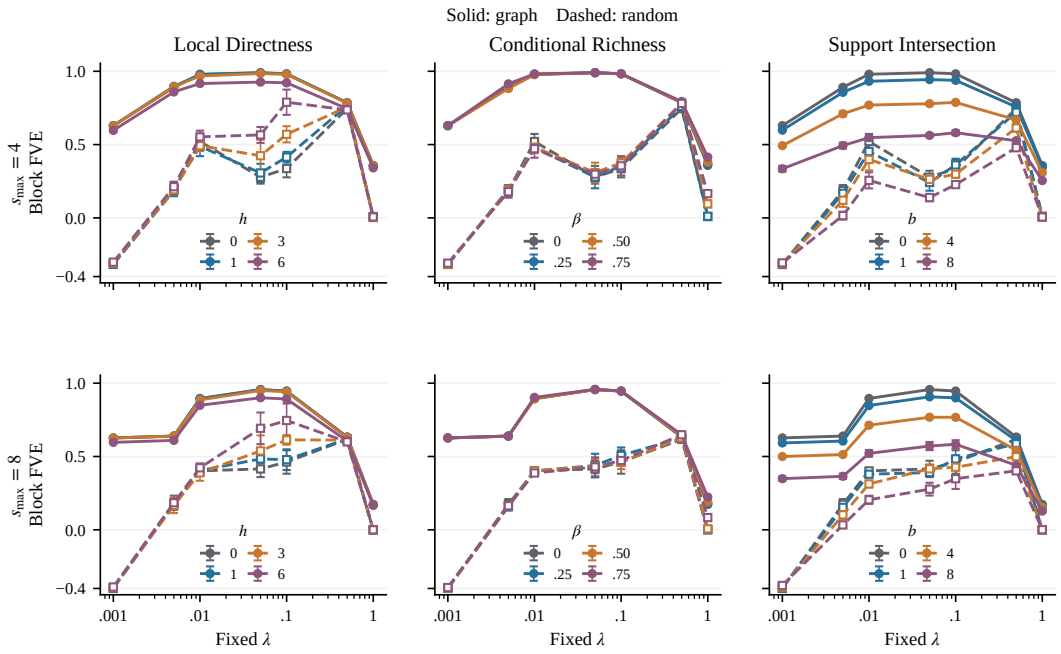

Figure 5: **Penalty sweeps under the three interventions.** All curves use identity inputs and $d = 128$; rows show $s_{\max} = 4$ and 8. Colors indicate intervention severity, with zero denoting the matched baseline. Solid and dashed lines denote graph and random, respectively. Error bars are sample standard deviations across world means.