FIRE: LEARNING IDENTIFIABLE OPERATORS FOR MULTI-ENVIRONMENT GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-environment data mix shared subject structure with environment-specific distortion. Domain generalization targets robust prediction and harmonization removes environment effects; neither is designed to identify the shared latent directions. Factored Identifiable Representation via Environments (FIRE) models each environment's observations as unknown full-rank linear transforms of a common subject representation. Stage 1 decomposes a reference environment's third cumulant; Stage 2 freezes the corresponding recovered rank-one operators during supervised prediction. Independent loadings with nonzero third cumulants and noise with zero third cumulant identify directions up to sign, permutation, and reference coordinates from one environment, without pairing. Under sub-Gaussian tails, a whitening-free estimator attains $\sqrt{d/N}$ error up to a conditioning factor once $N \gtrsim d^2$; without structural loading constraints, identification is impossible. Paired cross-environment second moments cancel independent noise. Synthetic recovery follows $N^{-1/2}$ scaling and remains accurate under poor conditioning; perturbation sweeps measure deviations from latent directions. Updating operators raises input-direction recovery error from 0.036 to 1.075; freezing preserves geometry at equal training loss. Against ERM and six domain-generalization methods using matched classification losses, features, and splits, FIRE has the highest mean out-of-distribution accuracy on four of five benchmarks, with gains modest relative to seed variability. Random frozen directions predict equally well on Camelyon17. Spectral recovery supplies identifiable structure: its TCGA-BRCA directions separate PAM50 subtypes more strongly than random or PCA directions.

1 INTRODUCTION

Biomedical datasets routinely span environments (Yan et al., 2022): scanners and acquisition protocols in neuroimaging, hospitals in histopathology, and processing sites in gene expression. Observed features mix shared subject structure with environment-specific distortion. Post hoc analysis requires both robust prediction and identifiable latent directions determined by the data, rather than an arbitrary basis.

Domain-generalization methods are not designed to recover these directions (Wang et al., 2022). Adversarial alignment (Ganin et al., 2016), invariant risk minimization (Arjovsky et al., 2019), and group-robust objectives (Sagawa et al., 2020) can improve out-of-distribution (OOD) prediction, although well-tuned empirical risk minimization (ERM) remains difficult to beat (Gulrajani & Lopez-Paz, 2021). Invariance can isolate a predictive shared subspace, but any invertible change of basis preserves invariance and can be absorbed by the predictor. Harmonization and mixed-effects methods (Johnson et al., 2007; Bates et al., 2015; Dinsdale et al., 2021; Abbasi et al., 2024) explicitly model and remove site effects as nuisance, leaving the underlying shared structure unrecovered. Neither family imposes the structural assumptions on latent loadings needed to identify individual directions.

Factored Identifiable Representation via Environments (**FIRE**) models observations through unknown environment-specific linear transforms of a low-dimensional subject representation comprising a small set of shared program directions. Unknown transforms restrict identification to $A_{e_{\mathrm{ref}}} b_k$ in reference coordinates, rather than the environment-free direction b_k. Stage 1 estimates these directions by tensor decomposition; Stage 2 fixes their rank-one operators inside a supervised predictor (Figure 1).

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

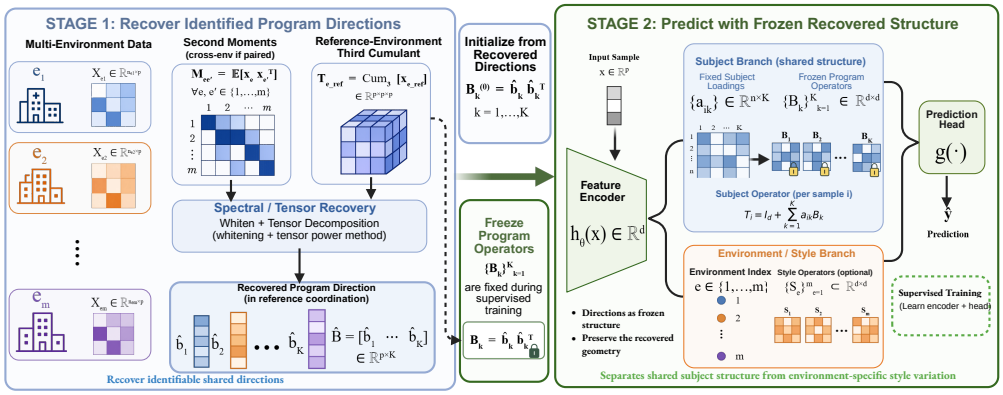


Figure 1: FIRE overview. Stage 1 decomposes the reference third cumulant into individual components, yielding normalized directions $\hat{b}_k$ that estimate $A_{e_{\text{ref}}} b_k$ up to sign and permutation. Appropriately aligned paired cross-environment second moments provide whitening free of independent noise; unpaired real-data benchmarks use the reference environment's own covariance. Recovered directions can be back-projected to the original feature space for inspection. Stage 2 freezes rank-one operators $B_k = \hat{b}_k \hat{b}_k^\top$ and combines them with subject loadings, encoder $h_\theta(x)$, an optional environment-specific style branch, and a prediction head.

One environment suffices for identification. Independent loadings with nonzero third cumulants and noise with zero third cumulant yield K rank-one terms along $A_{e_{\text{ref}}} b_k$. Kruskal uniqueness (Kruskal, 1977) fixes their directions up to sign and permutation, without paired subjects. Under sub-Gaussian tails, whitening-free simultaneous diagonalization (Leurgans et al., 1993) gives angular error $\sqrt{d/N}$ up to a conditioning factor once $N \gtrsim d^2$. Symmetric loadings leave no third-cumulant trace; without structural loading constraints, no estimator achieves vanishing error (Proposition 3). Additional environments aid robustness: their second cross-moments are noise-free in expectation, lowering synthetic recovery error from 0.307 to 0.186 at noise level $\sigma = 2$.

Freezing the recovered operators preserves their geometry while training the encoder and prediction head: updating them raises synthetic recovery error from 0.036 to 1.075 at unchanged training loss. Benchmark subjects appear in one environment, allowing unpaired cumulant estimation but using reference auto-covariance whitening, which mixes signal and noise. A vanishing-error guarantee requires consistent signal-covariance estimation, including observation-noise correction. The linear model describes a regime after feature extraction, not physical scanner, staining, or sequencing effects. Reference-coordinate recovery tolerates condition numbers up to 1000. Nonlinear and spatial perturbations increase latent-direction deviation, which need not measure error against the identifiable reference target.

The experiments distinguish the stages' roles. With matched classification loss, features, and splits, FIRE has the highest mean OOD accuracy against ERM and six domain-generalization methods on four of five benchmarks. Margins are modest; IRM leads on OASIS-1. Random frozen directions leave Camelyon17 OOD accuracy unchanged. Spectral recovery instead supplies structure: synthetic program error is 0.029 versus 1.315 for random directions, and TCGA-BRCA directions separate PAM50 subtypes more strongly than random or raw PCA directions.

Our contributions are: (1) applying classical cumulant uniqueness to establish single-environment reference-coordinate identification without pairing, a whitening-free finite-sample rate, and failure conditions; (2) measuring sensitivity to conditioning and observation-map perturbations; and (3) showing that freezing preserves recovered input directions at unchanged training loss, while distinguishing predictive performance from spectral recovery's identifiable structure.

2 Related Work

Identifiability from moments. Low-order tensor moments identify mixtures, topic models, independent component analysis (ICA), and multi-view models (Anandkumar et al., 2014); Kruskal's

condition ensures uniqueness (Kruskal, 1977), and simultaneous diagonalization constructs the factors (Leurgans et al., 1993). FIRE's single-environment model is undercomplete linear ICA with mixing matrix $A_{e_{\text{ref}}} D$ and zero-third-cumulant noise; nonzero skewness supplies the identifying information associated with non-Gaussianity in ICA (Comon, 1994). This classical argument removes pairing, without claiming a new identification principle. Cross-environment second moments similarly cancel view-specific noise. FIRE contributes the multi-environment operator setting, reference-coordinate target, frozen-operator predictor, and empirical sensitivity analysis.

Identifiability from environments. Nonlinear ICA and identifiable variational autoencoders use auxiliary variables, often environment or segment indices, across which latent distributions change (Hyvarinen et al., 2019; Khemakhem et al., 2020); extensions address causal and object-centric representations (Varici et al., 2025; Kori et al., 2024). FIRE shares loading distributions across environments while changing observation maps. The closest setting, unpaired multi-domain causal representation learning (Sturma et al., 2023), studies domain-specific linear mixing of non-Gaussian latents and identifies their joint distribution and shared causal graph from per-domain marginals under suitable conditions. FIRE targets the narrower object of shared directions in one reference environment's coordinates, used inside a predictor rather than for causal discovery. Invariant causal prediction also uses environment heterogeneity, to select causal outcome predictors rather than recover latent directions (Peters et al., 2016).

Domain generalization and harmonization. DANN (Ganin et al., 2016), IRM (Arjovsky et al., 2019), group DRO (Sagawa et al., 2020), CORAL (Sun & Saenko, 2016), Mixup (Zhang et al., 2018), and Spectral Decoupling (Pezeshki et al., 2021) target cross-environment robustness; DomainBed finds well-tuned ERM hard to beat consistently (Gulrajani & Lopez-Paz, 2021). Shared-private models, including domain separation networks (Bousmalis et al., 2016), separate shared and environment-specific representations. Neither they nor invariance methods identify a shared-subspace basis. ComBat (Johnson et al., 2007), mixed-effects models (Bates et al., 2015), and site-invariant deep MRI (Dinsdale et al., 2021; Abbasi et al., 2024) remove nuisance site effects before analysis. FIRE models environment effects through recovery transforms and predictive style operators. Unlike neural operators mapping function spaces (Lu et al., 2021; Li et al., 2021), its operators act on subject representations and are tied to a recovered dictionary.

3 METHOD

3.1 PROBLEM SETUP AND NOTATION

For subjects $i \in \{1, \dots, N\}$, environments $e \in \{1, \dots, E\}$ and programs $k \in \{1, \dots, K\}$, observations $u_{ie} \in \mathbb{R}^d$ follow

$$u_{ie} = A_e m_i + \varepsilon_{ie}, \qquad m_i = \sum_{k=1}^{K} a_{ik} b_k = D a_i. \tag{1}$$

Here $D = [b_1, \dots, b_K] \in \mathbb{R}^{d \times K}$ is a dictionary of shared program directions (latent factors), $a_i \in \mathbb{R}^K$ contains subject-specific loadings, $A_e \in \mathbb{R}^{d \times d}$ is an unknown full-rank environment transform, and $\varepsilon_{ie} \sim \mathcal{N}(0, \sigma^2 I)$ is independent noise. We seek the directions without knowing the transforms. Centered, mutually independent loadings with finite second and third moments make the moments of m_i decompose along D.

The model covers paired subjects measured in several environments and unpaired subjects measured once. Identification uses one environment's marginal distribution (Section 3.4). Paired cross-moments cancel independent noise, but whitening in reference coordinates also requires alignment. Other environments supply the Stage 2 environment branch and the populations to which recovered directions are applied, under a shared-program assumption:

Assumption 1 (Cross-environment exchangeability)**.** The loading distribution is shared across environments: $a_i \perp e$. Environments change the observation map A_e, not the population of latent subject types.

This is exact in Colored MNIST, most defensible in the Alzheimer's Disease Neuroimaging Initiative (ADNI), Camelyon17 and The Cancer Genome Atlas breast cancer cohort (TCGA-BRCA), and weaker in the Open Access Series of Imaging Studies (OASIS-1), whose age-quartile environments

may change latent composition. Proposition 1 concerns only the reference marginal and does not need this assumption or constrain other environments. Exchangeability describes a shared latent population; it does not align $A_e b_k$ across environments. Reusing reference operators in a common feature space is an empirical modeling choice, not a transport guarantee.

3.2 Tensor Spectral Recovery

Principal component analysis (PCA) and factor analysis recover a subspace; Stage 1 seeks its individual directions. Third-order structure resolves the rotational ambiguity through weights $\mu_{3,k} = \mathbb{E}[a_{ik}^3]$. The identifiable target is the direction of $A_{e_{\text{ref}}} b_k$ in reference coordinates, rather than the environment-free b_k. We reserve D for the Stage 1 dictionary and B_k for Stage 2 rank-one operators.

Population object. Centered independent loadings and independent noise with zero third cumulant give the reference-environment tensor

$$\mathcal{T} = \mathrm{Cum}_3(u_{ie_{\text{ref}}}) = \sum_{k=1}^{K} \mu_{3,k} (A_{e_{\text{ref}}} b_k)^{\otimes 3}. \tag{2}$$

Its K rank-one terms contain no noise contribution or other environment.

Estimator. We decompose the sample cumulant by whitening, tensor power iteration and unwhitening. Whitening requires the reference signal covariance $C_{\text{ref}} = A_{e_{\text{ref}}} \Sigma A_{e_{\text{ref}}}^\top$, with $\Sigma = D\,\mathrm{diag}(\mathbb{E}[a_{ik}^2]) D^\top$. Paired subjects provide

$$M_{ee'} = \mathbb{E}[u_{ie} u_{ie'}^\top] = A_e \Sigma A_{e'}^\top, \qquad e \neq e', \tag{3}$$

which is noise-free but generally differs from the symmetric C_{ref}. Known alignment, or identity environment maps, supplies the required covariance. Without pairing, the auto-covariance is $C_{\text{ref}} + \mathrm{Cov}(\varepsilon)$: the whitening guarantee needs absent noise or consistent noise-covariance correction. Ignoring even small fixed noise can leave asymptotic bias. Identification itself requires neither pairing nor signal covariance (Section 3.4).

The paired implementation uses cross-moments to select a subspace, then whitens the projected auto-covariance; observation noise therefore remains. Its synthetic reference maximizes the whitened tensor norm (Appendix A). For signal-covariance whitening, keep the K leading eigenpairs (U_K, Λ_K) of its estimate and set $W = U_K \Lambda_K^{-1/2}$. If this matrix consistently estimates the signal covariance, applying W along each mode gives a whitened sample tensor with orthogonally decomposable population form

$$\widetilde{\mathcal{T}} = \sum_{k=1}^{K} \tilde{\lambda}_k q_k^{\otimes 3}, \qquad \tilde{\lambda}_k = \mu_{3,k} \|W^\top A_{e_{\text{ref}}} b_k\|^3, \tag{4}$$

with orthonormal q_k. We symmetrize, apply robust tensor power iteration with deflation and reorthogonalization (Anandkumar et al., 2014), and unwhiten through the pseudoinverse: $\hat{b}_k = (W^\top)^+ q_k / \|(W^\top)^+ q_k\|$. These estimate normalized reference-coordinate directions up to sign and permutation. Real-data recovery uses a training environment as reference and its own covariance (Appendix A), without target labels or statistics. PCA loadings map the estimated directions back to the original feature space.

3.3 Predictive Operator Model and Frozen Programs

Initialize $B_k^{(0)} = \hat{b}_k \hat{b}_k^\top$ and set $T_i = I + \sum_k a_{ik} B_k$. The implementation stores $B_k = s_k u_k v_k^\top$ (rank one). Trainable factors are unconstrained after initialization; the recovery score uses $v_k / \|v_k\|$, not the left factor or the full matrix (Appendix E). For encoder output $h_\theta(x) \in \mathbb{R}^d$, the implemented predictor is

$$f(x; i) = g(T_i h_\theta(x)), \qquad g(z) = W_{\text{head}} z + b_{\text{head}}, \quad W_{\text{head}} \in \mathbb{R}^{C \times d}. \tag{5}$$

Here $C = 2$ for binary classification, 5 for TCGA-BRCA and 1 for synthetic regression. The benchmarks minimize task loss with B_k frozen, training the encoder and head. The optional style branch is disabled; its proposed formulation needs additional constraints (Appendix A). Freezing preserves stored geometry, not necessarily its alignment with the learned encoder.

Synthetic runs use generating loadings. Benchmark runs draw each training subject's loadings once from $\mathcal{N}(0, I_K)$; source-only ridge regression maps features to these loadings for new subjects. No target labels or target statistics enter fitting, training or model selection. Fixed loadings make $\sum_{i,k} |a_{ik}|$ constant. No unseen style operator is fitted; the optional branch uses $S_e = I$ for a new environment.

The benchmark branch therefore uses random training loadings and source-ridge predictions, rather than identified subject codes. Spectral and random frozen controls have equal Camelyon17 mean accuracy (Appendix E). Only the synthetic runs use generating subject codes.

3.4 Theory Summary

One environment identifies directions; additional environments can help estimate whitening moments. Write $c_k = A_{e_{\mathrm{ref}}} b_k$ and $\bar{c}_k = c_k / \|c_k\|$.

Proposition 1 (Single-environment identifiability)**.** Under model (1) in the reference environment, suppose $A_{e_{\mathrm{ref}}}$ is full rank and $b_1, \ldots, b_K$ are linearly independent ($K \le d$). Let the loadings be centered and mutually independent, with finite nonzero third cumulants $\mu_{3,k}$, and the noise independent of the loadings with zero third cumulant, as for Gaussian or symmetric noise. Then the third cumulant of $u_{ie_{\mathrm{ref}}}$ determines $\{\bar{c}_k\}_{k=1}^K$ up to sign and permutation.

Noise drops out of (2), which depends only on the reference distribution. The factor matrix $[c_1, \ldots, c_K]$ has Kruskal rank K, satisfying $3K \ge 2K + 2$ for $K \ge 2$ and giving essential uniqueness (Kruskal, 1977); for $K = 1$, rank-one structure fixes the direction. Neither additional environments, pairing, signal-covariance knowledge nor Assumption 1 is needed. Symmetric loadings leave their directions unidentified at third order. Appendix B gives the proof.

Proposition 2 (Finite-sample rate, whitening-free estimator)**.** Assume Proposition 1 and i.i.d. reference observations with $R = \sup_{\|v\|=1} \|\langle v, u - \mathbb{E}u\rangle\|_{\psi_2} < \infty$. For $0 < \delta < 1$, set $h = 3d \log 13 + \log(6/\delta)$, assume $N \ge h$, and define $\widehat{\mathcal{T}} = N^{-1} \sum_i (u_i - \bar{u})^{\otimes 3}$ and $\eta_N = R^3[\sqrt{h/N} + h^{3/2}/N]$. With probability at least $1 - \delta$, $\|\widehat{\mathcal{T}} - \mathcal{T}\|_{\mathrm{op}} \le C\eta_N$, with universal C. For known K, use the rank-K restricted matrix pencil from two independent random unit contractions drawn independently of the data (Appendix B). Conditional on nonzero contraction coefficients and distinct pencil eigenvalues, if $C\eta_N$ is below the appendix's perturbation threshold, then

$$\min_{\pi \in S_K} \max_k \sin \angle(\hat{b}_k, \bar{c}_{\pi(k)}) \le C \kappa_{\mathrm{CP}} \eta_N. \tag{6}$$

The appendix specifies κ_{CP} and the threshold through factor conditioning, contraction magnitudes and the pencil eigengap. Probability is over observations, conditional on the contractions.

At fixed confidence, the terms balance at $N \asymp d^2$; fixed dimension and conditioning give $O_{\mathbb{P}}(N^{-1/2})$, with a constant depending on realized contraction margins and eigengap. The implemented whitening requires consistent signal-covariance estimation. Theorem 4 in Appendix C gives its conditional bound, including whitening and unwhitening errors. Known aligned paired observations with finite sixth moments provide $N^{-1/2}$ moment rates. Pairing alone does not supply alignment, and uncorrected noisy auto-covariance is not consistent for the signal covariance.

Proposition 3 (Non-identifiability under latent reparameterization)**.** Fix $2 \le K \le d$ and nonsingular environment transforms. Let $\mathcal{M}$ contain model (1) with unknown full-column-rank dictionaries and unknown centered loading distributions, without independence or third-cumulant constraints. For unit-normalized estimates and $\bar{C} = [\bar{c}_1, \ldots, \bar{c}_K]$, the direction loss satisfies

$$\begin{aligned} d_{\pm}(\hat{B}, \bar{C}) &= \min_{\pi,\, s_k \in \{-1,1\}} \max_k \|\hat{b}_k - s_k \bar{c}_{\pi(k)}\|, \\ \inf_{\hat{B}} \sup_{M \in \mathcal{M}} \mathbb{E}_M[d_{\pm}(\hat{B}, \bar{C}(M))] &\ge c_0 > 0 \qquad \text{for every } N. \end{aligned} \tag{7}$$

The constant may depend on the fixed transforms, but not on N.

Environments alone cannot identify unconstrained loading models: $D \mapsto DQ$, $a \mapsto Q^{-1}a$ leaves observations unchanged in every environment. The appendix applies Le Cam's two-point method to obtain the minimax lower bound. Identification therefore depends on loading structure, not environment count. Independence and nonzero skewness fix normalized directions up to signs and permutations, as in Proposition 1.

4 EXPERIMENTAL SETUP

We evaluate synthetic recovery and five benchmarks spanning digits, histopathology, neuroimaging and gene expression. Additional recovery analyses, benchmark summaries and implementation details appear in Appendix F–I.

Protocol. Methods share features, splits, optimizer settings and seeds. PCA gives $d = 50$ inputs. FIRE uses a ReLU encoder $50 \to H_1 \to H_2 \to 50$, without layer normalization or dropout: $(H_1, H_2) = (128, 64)$ on Colored MNIST, Camelyon17 and ADNI sMRI, and $(64, 32)$ on OASIS-1 and TCGA-BRCA. Every baseline instead uses $50 \to H_1 \to 50$, with one hidden layer. We use Adam and early stopping. Following DomainBed (Gulrajani & Lopez-Paz, 2021), comparisons share the benchmark-specific settings across methods, without per-method retuning or target-informed model selection. We compare FIRE with ERM, DANN, IRM, Group-DRO, CORAL, Mixup and Spectral Decoupling (Pezeshki et al., 2021), using published default penalties: $\lambda_{\mathrm{IRM}} = 1.0$, $\eta_q = 0.01$, $\lambda_{\mathrm{CORAL}} = 1.0$, $\lambda_{\mathrm{adv}} = 1.0$, $\alpha = 0.2$. Binary tasks use single-logit cross-entropy for baselines and equivalent two-logit softmax cross-entropy for FIRE; all TCGA-BRCA heads use five-way softmax cross-entropy. Earlier baselines used squared error on a single logit, understating their performance; we report corrected runs. Accuracies are mean $\pm$ s.d. over 10 seeds under leave-one-environment-out evaluation. Comparisons are descriptive because aggregate statistics lack the paired seed-level differences needed for significance tests.

FIRE configuration. Main prediction runs fix $K = 3$ a priori to retain several interpretable programs, with $\lambda_1 = \lambda_2 = 0.01$. The post hoc $K \in \{3, 5, 7\}$ scan changes accuracy by less than one percentage point and was not used for selection (Appendix I.6). Loadings are fixed and the style branch is disabled (Section 3.3). Each real-data subject appears once, so Stage 1 uses the reference environment's own subjects and second-order whitening matrix (Section 3.2). Reuse across environments is evaluated empirically; Assumption 1 supplies no alignment and is likely weakest on OASIS-1, where age quartiles may change latent composition as well as the observation map.

4.1 BENCHMARKS

Colored MNIST (Arjovsky et al., 2019) uses three color-label correlations $[0.9, 0.8, 0.1]$ and leave-one-environment-out evaluation. **Camelyon17** (Bandi et al., 2018; Koh et al., 2021) contains $455{,}954$ histopathology patches from five hospitals. We reduce 512-dimensional ResNet-18 features to 50 principal components (PCs), subsample $5{,}000$ patches per hospital and evaluate leave-one-hospital-out. **OASIS-1** (Marcus et al., 2007) uses 226 subjects: voxel-based morphometry gray-matter measurements are aggregated to 100 Schaefer regions of interest (ROIs) and reduced to 50 PCs for dementia classification across four age quartiles. **ADNI structural MRI (sMRI)** (Jack Jr et al., 2008) uses 616 subjects and eight scanner protocols (ADNI-2, ADNI-3 and ADNI-GO crossed with GE, Philips and Siemens), with gray-matter features reduced to 50 PCs for cognitively normal versus impaired classification. **TCGA-BRCA** (Cancer Genome Atlas Network et al., 2012) uses 896 subjects from 13 tissue source sites: log-transformed RNA-seq, the $5{,}000$ most variable genes and 50 PCs predict five PAM50 subtypes. Appendix I gives training settings, baseline descriptions and evaluation metrics.

5 RESULTS

5.1 RECOVERY SCALING AND CONDITIONING

We test recovery scaling and conditioning, and sensitivity to observation-map perturbations. The score is the largest sign-aligned chord distance after greedy cosine matching, in $[0, \sqrt{2}]$; random directions give 1.30 ± 0.05. The scaling experiment uses $K = 3$, $d = 20$, $E = 5$, $\sigma = 0.1$, identity transforms, and centered exponential loadings; Appendix I.1 discusses alternative loading distributions.

Error falls from 0.216 ± 0.053 at $N = 100$ to 0.023 ± 0.009 at $N = 10{,}000$, with log-log slope -0.448 versus the -0.500 reference (Figure 2(b)). Perturbing every environment, including the reference, raises reference-coordinate error from 0.028 ± 0.002 at $\kappa = 1$ to 0.098 ± 0.078 at $\kappa = 1000$, still an order of magnitude below random directions (Figure 2(c)).

Table 1: Leave-one-environment-out accuracy (mean ± s.d., 10 seeds). Features, splits, optimizer, cross-entropy and seeds are shared, without per-method retuning or target information. Bold marks each row's highest mean.

Benchmark	FIRE	ERM	DANN	IRM	G-DRO	CORAL	Mixup	SD
Camelyon17	**.851±.031**	.848±.030	.837±.027	.821±.020	.633±.038	.823±.030	.816±.027	.833±.029
Colored MNIST	**.863±.085**	.844±.084	.830±.085	.799±.111	.607±.045	.826±.082	.814±.097	.852±.099
TCGA-BRCA	**.870±.052**	.857±.050	.815±.054	.858±.053	.858±.052	.843±.053	.588±.084	.855±.046
OASIS-1	.509±.070	.492±.075	.511±.086	**.531±.074**	.437±.150	.482±.077	.509±.073	.494±.061
ADNI sMRI	**.770±.128**	.764±.120	.768±.121	.629±.067	.728±.185	.757±.117	.748±.126	.749±.118

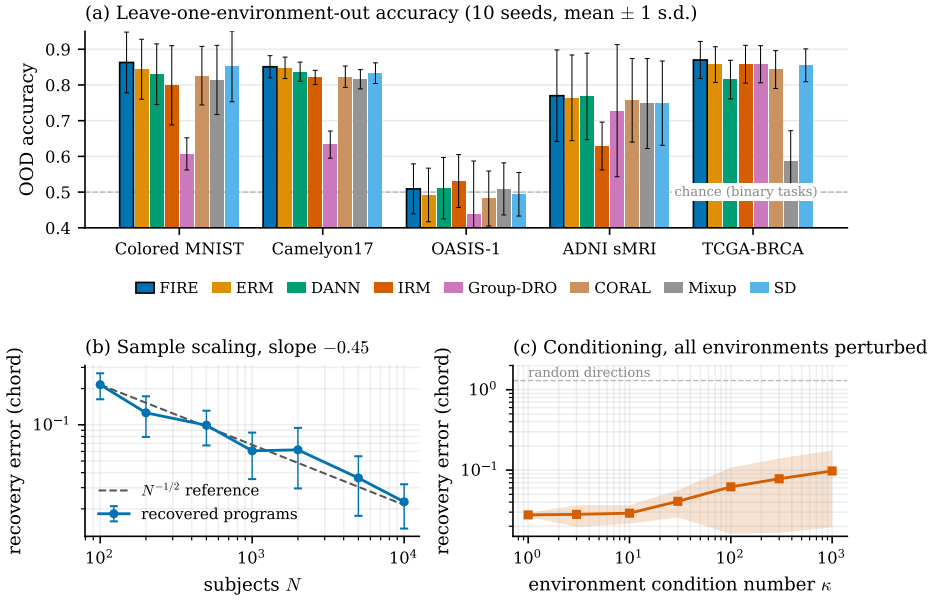


Figure 2: Accuracy and synthetic recovery. (a) FIRE, ERM and six DG baselines: leave-one-environment-out means over 10 matched seeds; one-s.d. bars exceed every margin over the strongest competitor (shared classification loss, features and splits). The accuracy axis starts at 0.4, clipping the lower OASIS-1 Group-DRO error bar, whose endpoint is 0.287. (b) Recovery versus sample size with identity transforms: log-log slope -0.45, reference $-1/2$. (c) Conditioning sweep over every environment, including the reference, scored in reference coordinates ($N = 6000$, $E = 5$, $K = 3$, $d = 20$, 5 seeds); shading: one s.d.; dashed: random-direction level.

The perturbation sweeps compare with latent b_k, without transforming the target. At strength $\varepsilon = 1$ ($N = 6000$, $E = 5$, $K = 3$, $d = 20$, 5 seeds), deviation is 0.096 ± 0.007 for subject-dependent maps and 0.488 ± 0.050 for spatial perturbations. For nonlinearity it is 0.046 ± 0.006 at $\varepsilon = 0.2$ and 1.338 ± 0.079 at $\varepsilon = 1$; counts below 0.15 fall from $5/5$ to $0/5$. Three of five runs score below 0.15 at $\varepsilon_n = 0.5$ alone and at 0.35 with $\varepsilon_s = 0.2$. Joint deviation at $(0.5, 0.2)$ is 1.102 ± 0.544, versus 0.587 and 0.113 separately. These are latent-target sensitivities: reference selection was not logged, and the spatial construction remains globally linear while changing reference directions. They do not establish failure of reference-coordinate identification (Appendix I.1).

5.2 Benchmark Comparison

Spectral trainable and frozen programs attain equal final training loss, but only freezing preserves the recovered input directions (Table 2). Contrastive training raises direction error further. This supports a geometry benefit at equal training fit, not a held-out prediction claim. Freeze–thaw leaves training loss at 0.0101 while raising direction error from 0.031 to 0.376 (Appendix E).

Table 2: Stage 2 strategies.

K=3, d=20, N=5,000, 5 seeds.

Strategy	Train. loss	Rec. error
Random, frozen	.0116±.0004	1.335±.035
Spectral, trainable	.0102±.0002	1.075±.292
Spectral, frozen	**.0102±.0003**	**.036±.019**
Spectral + contrastive	.0495±.0036	1.293±.080

All rows report final training loss. The random frozen row comes from the separate freeze–thaw run (Appendix E).

Table 3: TCGA-BRCA direction quality.

Fisher scores over 8 subject-level bootstraps.

Directions	Subtype	Site	Bio./site
Spectral	.993±.125	.102±.024	10.52±3.85
ICA	.859±.108	.105±.010	8.25±1.44
Raw PCA	.824±.070	.108±.010	7.67±.71
Random	.32±.12	–	–

Ratios average the per-bootstrap subtype/site ratios from the same run (Appendix H.1).

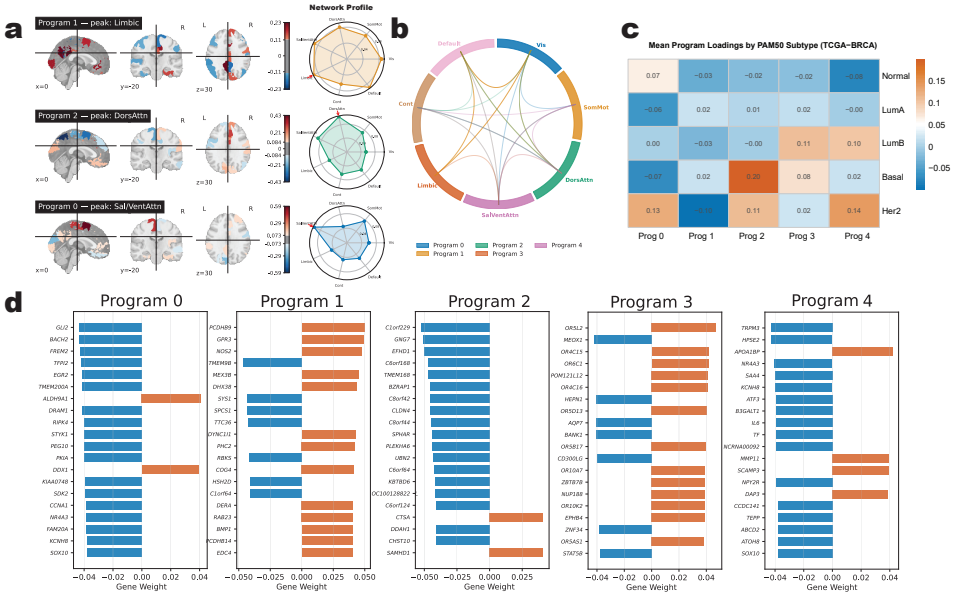


Figure 3: Application-case illustration from an earlier implementation. (a) OASIS-1 atlas projections and network profiles; (b) network-weight summaries; (c) TCGA-BRCA loading means by PAM50 subtype; (d) gene-weight profiles. These panels used cross-environment whitening on unpaired data, a degenerate configuration. They illustrate visualization outputs, not validation of the corrected recovery method. Table 3 uses a separate single-environment recovery.

FIRE leads four of five benchmarks in the reported comparison (Table 1; Figure 2(a)), by 0.011 on Colored MNIST, 0.003 on Camelyon17, 0.002 on ADNI sMRI and 0.012 on TCGA-BRCA; IRM leads OASIS-1 by 0.022. Every margin lies within one seed standard deviation. Without paired seed-level differences, these are descriptive comparisons, not significance claims; the non-exhaustive suite does not establish state of the art. Well-tuned ERM stays close throughout, consistent with DomainBed (Gulrajani & Lopez-Paz, 2021). Failures are structured: Group-DRO is weak on Colored MNIST and Camelyon17, and Mixup collapses on five-class TCGA-BRCA. On Camelyon17, spectral and random frozen operators both score 0.854 ± 0.032, random trainable operators 0.852 ± 0.030, and a shallower dense MLP 0.847 ± 0.029. The operator models share an encoder; the dense model has one fewer hidden layer, so it does not isolate operatorization. These controls share the dimensionality-scan seeds; Table 1 reports the separate main FIRE/ERM runs. The bases tie in mean accuracy; freezing shows no predictive penalty. Recovery differs: spectral directions recover synthetic generating programs (0.029 ± 0.002), random directions fail (1.315 ± 0.035), and raw PCA trails (0.379 ± 0.183). TCGA-BRCA spectral directions also separate PAM50 subtypes better than ICA or raw PCA (Section 5.3).

5.3 Direction Quality and Application Illustration

The application illustration uses $K = 5$; prediction uses prespecified $K = 3$. In the post hoc $K \in \{3, 5, 7\}$ scan, accuracy varies by less than one percentage point; held-out environments were not used to choose K (Appendix I.6).

On TCGA-BRCA, Fisher ratios of subject projections measure PAM50 subtype separation and tissue-source-site alignment (Table 3). Spectral directions improve biological separation without stronger site alignment, yielding the best biology-to-batch ratio. All three columns use eight subject-level bootstraps; the last averages per-bootstrap ratios (Appendix H.1). These results describe directions, not accuracy. Figure 3 illustrates how candidate directions can be displayed in anatomical and molecular coordinates. Its earlier implementation differs from the single-environment recovery used for Table 3. We therefore do not use the illustration to infer functional organization, subtype-specific programs, enrichment, or phenotype associations.

ADNI fMRI marks a boundary: FIRE trails every baseline by roughly 10 points on leave-one-protocol-out connectivity (Appendix G.2). Global rather than per-site z-normalization preserves the ordering, making normalization artifacts unlikely; model mismatch and poor covariance conditioning remain plausible.

6 Discussion

Accuracy alone gives no reason to prefer FIRE to ERM: its four benchmark leads are smaller than seed variability, and the Camelyon17 basis controls have equal mean accuracy. Its value is inspecting identifiable axes whose geometry freezing preserves: synthetic programs and TCGA-BRCA subtype separation without stronger site alignment. Shared PCA may remove environment-specific variation, leaving little benefit for additional DG constraints and their optimization trade-offs. ADNI fMRI is the only benchmark where FIRE trails every baseline. Model mismatch and whitening instability remain hypotheses; the predictive comparison does not distinguish them. OASIS-1 favors $K = 1$ in the post hoc scan; this does not establish a rank-one latent model. The synthetic perturbation scores measure latent-target deviation, without establishing an identification boundary. Biological interpretation requires corrected recovery, stability assessment and external validation; the application illustration supplies no such validation. The directions remain reference-dependent.

7 Limitations

Equation (1) is a feature-space identification regime, not scanner, staining or sequencing physics; synthetic sensitivity tests do not locate the benchmarks relative to that regime. Symmetric loadings are outside its third-order guarantee. Interpretation depends on reference distortion: the directions are $A_{e_{\mathrm{ref}}} b_k$, not b_k. Exchangeability alone aligns neither environment coordinates nor the learned encoder with these directions. Both implemented whitening procedures retain noise, so their outputs do not inherit Proposition 2; whitening-free estimation is a next step. The real-data predictive benchmarks use PCA and small MLPs rather than end-to-end training. The $O(Nd^3)$ moment cost motivates $d = 50$; stability through $d = 200$ does not replace scalable tensor approximation for high-dimensional or sparse recovery. PCA can discard sparse or low-variance signals, limiting recovery and biological interpretation to its retained subspace. Benchmark loadings are fixed random training inputs; generating subject codes are used only in the synthetic experiments. Jointly learning loadings and the encoder is an extension.

8 Conclusion

FIRE combines reference-direction identification with frozen-operator prediction. Independent, nonzero-skew loadings identify directions from one environment's third cumulant without pairing; the impossibility result establishes the need for loading constraints. Freezing preserves recovered operators at unchanged synthetic training loss. Benchmark accuracy is competitive; Camelyon17 spectral and random bases tie, while the shallower dense MLP prevents isolating an operatorization effect. Future work includes whitening-free estimation, nonlinear or regionally varying maps, and recovery beyond PCA.

AI USE STATEMENT

We used generative AI tools to review theoretical arguments, experimental methodology, and interpretations of reported results, and to assist with proof corrections. We also used these tools for language polishing, typo checking, and manuscript formatting. The authors take responsibility for the final content of the paper.

REFERENCES

Soolmaz Abbasi et al. Deep learning for the harmonization of structural MRI scans: a survey. *BioMedical Engineering OnLine*, 23(1):90, 2024.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1): 2773–2832, 2014.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Peter Bandi et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2): 550–560, 2018.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumors. *Nature*, 490(7418):61, 2012.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

Nicola K Dinsdale, Mark Jenkinson, and Ana IL Namburete. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *NeuroImage*, 228:117689, 2021.

Yaroslav Ganin et al. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

Clifford R Jack Jr et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Pang Wei Koh et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Avinash Kori, Francesco Locatello, Ainkaran Santhirasekaram, Francesca Toni, Ben Glocker, and Fabio De Sousa Ribeiro. Identifiable object-centric representation learning via probabilistic slot attention. *Advances in Neural Information Processing Systems*, 37:93300–93335, 2024.

Joseph B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2): 95–138, 1977. doi: 10.1016/0024-3795(77)90069-6.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, 2022. doi: 10.1093/imaiai/iaac012. URL https://doi.org/10.1093/imaiai/iaac012.

Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pp. 38–53, 1973.

S. E. Leurgans, R. T. Ross, and R. B. Abel. A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, 1993. doi: 10.1137/0614071.

Zongyi Li et al. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

Daniel S Marcus et al. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, 2007.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016. doi: 10.1111/rssb.12167.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1256–1272, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34465–34492, 2023.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision Workshops*, pp. 443–450, 2016.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112):1–90, 2025.

Jindong Wang et al. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8052–8072, 2022.

Junqi Wang et al. Functional network estimation using multigraph learning with application to brain maturation study. *Human Brain Mapping*, 42(9):2880–2892, 2021.

Weizheng Yan et al. Deep learning in neuroimaging: Promises and challenges. *IEEE Signal Processing Magazine*, 39(2):87–98, 2022.

Bin Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, pp. 423–435, 1997.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Yipu Zhang et al. Multi-paradigm fMRI fusion via sparse tensor decomposition in brain functional connectivity study. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1712–1723, 2020.

A ALGORITHM PSEUDOCODE

This pseudocode describes the unpaired benchmark pipeline. The paired synthetic implementation is described below. Lemma 5 gives a separate known-alignment construction for the conditional recovery analysis; it is not the implemented estimator.

Step 1: Estimate sample moments. For each training environment, center observations by their sample mean and compute $\hat{\mathbf{M}}_{ee} = N_e^{-1} \sum_i u_{ie} u_{ie}^\top$ and $\hat{\mathcal{T}}_e = N_e^{-1} \sum_i u_{ie}^{\otimes 3}$. Each subject is measured once, so cross-environment moments are unavailable; equal loading marginals do not permit pairing independent subjects.

Step 2: Choose a reference environment and whiten. Choose a reference from the training environments. Eigendecompose its second-order matrix $\hat{\mathbf{M}}_{e_{\mathrm{ref}} e_{\mathrm{ref}}} = U \Lambda U^\top$, keep the K leading eigenpairs after discarding eigenvalues below $\tau_{\mathrm{cut}} \Lambda_{11}$, and form $\mathbf{W} = U_K \Lambda_K^{-1/2}$. This K-component procedure requires at least K positive eigenvalues to survive truncation. No target labels or target statistics enter this choice.

Step 3: Construct the whitened tensor. Apply $\mathbf{W}$ along each mode to obtain $\widetilde{\mathcal{T}} = \hat{\mathcal{T}}_{e_{\mathrm{ref}}}(\mathbf{W}, \mathbf{W}, \mathbf{W})$.

Step 4: Recover whitened components. Run the tensor power method with L random restarts. At each restart, iterate $\theta \leftarrow \widetilde{\mathcal{T}}(\mathbf{I}, \theta, \theta) / \|\widetilde{\mathcal{T}}(\mathbf{I}, \theta, \theta)\|$ to convergence, keep the best component, subtract it by deflation, and re-orthogonalize the recovered directions.

Step 5: Unwhiten the program directions. Map each recovered q_k back to the observation space via $\hat{\mathbf{b}}_k = (\mathbf{W}^\top)^+ q_k / \|(\mathbf{W}^\top)^+ q_k\|$, where $+$ denotes the Moore–Penrose pseudoinverse.

Step 6: Initialize and freeze the program operators. Set $\mathbf{B}_k = \hat{\mathbf{b}}_k \hat{\mathbf{b}}_k^\top$ and keep these operators fixed during the predictive stage.

Step 7: Train the predictive model. In the reported benchmarks, optimize the encoder h_θ and prediction head under $\mathcal{L}_{\mathrm{pred}}$ until convergence or early stopping. The per-subject loadings are held fixed, making $\mathcal{L}_{\mathrm{sparse}}$ constant, and the style branch is disabled. The optional style formulation has the limitation below.

Step 8: Infer loadings for unseen subjects. Fit a ridge regression from training features to training loadings on source data only and apply it to the test features. No target labels or target statistics enter training, fitting, or model selection; target labels are used only for evaluation. For an unseen environment, use the identity style operator when the style branch is enabled, without fitting to target data.

Paired synthetic implementation. Let $X_e \in \mathbb{R}^{N \times d}$ contain observations centered within environment e. For each environment, compute

$$M_e = \frac{1}{E-1} \sum_{e' \neq e} \frac{X_e^\top X_{e'}}{N}.$$

Let U_e contain its top-K left singular vectors and compute $G_e = (X_e U_e)^\top (X_e U_e)/N$. With $G_e = Q_e \Lambda_e Q_e^\top$, retain eigenvalues at least $10^{-6} \lambda_{\max}(G_e)$ and set $W_e = U_e Q_{e,r_e} \Lambda_{e,r_e}^{-1/2}$. Use the common retained dimension $r = \min(K, \min_e r_e)$. Form same-environment third tensors $\widetilde{\mathcal{T}}_e = N^{-1} \sum_i (W_e^\top x_{ie})^{\otimes 3}$ in this dimension and choose $e_{\mathrm{ref}} = \arg\max_e \|\widetilde{\mathcal{T}}_e\|_F$. Symmetrize that tensor over all six permutations, use power iteration with 10 restarts, at most 1000 iterations and the directional stopping tolerance below, then deflate with re-orthogonalization. Unwhiten via $W_{\mathrm{ref}} (W_{\mathrm{ref}}^\top W_{\mathrm{ref}})^{-1} q_k$ and normalize. Pairing enters only the subspace estimate; no cross-environment third tensor or alignment is used. The projected auto-covariance G_e retains observation noise. This is not the signal-covariance estimator in Lemma 5, and the conditional whitening theorem does not give an unconditional consistency guarantee for these runs. In the unpaired benchmarks, cross-environment whitening is disabled.

Scope of the optional style objective. The optional scalar formulation is $g_s(\langle w, T_i h_\theta(x) \rangle + \langle v, S_e h_\theta(x) \rangle)$, with $w, v \in \mathbb{R}^d$ and $S_e \in \mathbb{R}^{d \times d}$. It is distinct from the implemented vector-input head.

Its proposed separation term is $\mathcal{L}_{\text{style}} = -E^{-1}\sum_e \|S_e - \bar{S}\|_F^2$, where $\bar{S} = E^{-1}\sum_e S_e$, within $\mathcal{L}_{\text{pred}} + \lambda_1 \sum_{i,k} |a_{ik}| + \lambda_2 \mathcal{L}_{\text{style}}$. For $E \geq 2$ and $\lambda_2 > 0$, this objective is unbounded without an additional constraint: setting $v = 0$, $S_1 = tH$, $S_2 = -tH$ and the other operators to zero leaves predictions unchanged while the style term tends to $-\infty$ for nonzero H. We therefore do not present this unconstrained expression as a valid training prescription or claim an evaluated benefit for the optional branch. The reported benchmarks disable it; their objective is the task loss, since fixed loadings make the sparsity term constant. A constrained style model would require its own specification and evaluation.

B Identifiability from a Single Environment

This appendix proves Proposition 1 and gives the argument behind Proposition 2. Both concern the reference environment alone. Write $c_k = A_{e_{\text{ref}}} b_k$ and $C = [c_1, \ldots, c_K] \in \mathbb{R}^{d \times K}$, and let $u = u_{ie_{\text{ref}}} - \mathbb{E}[u_{ie_{\text{ref}}}]$ denote a centered observation from that environment.

Proof of Proposition 1. The centered observation is $u = \sum_{k=1}^K a_{ik} c_k + \tilde{\varepsilon}$ with $\tilde{\varepsilon}$ the centered noise. Third cumulants are additive over independent summands, and for a centered scalar a_{ik} the third cumulant of $a_{ik} c_k$ is $\mathbb{E}[a_{ik}^3]\, c_k^{\otimes 3}$. Mutual independence of the loadings makes all cross-cumulants vanish, and the noise has zero third cumulant by assumption, so

$$\mathcal{T} \;=\; \text{Cum}_3(u) \;=\; \sum_{k=1}^K \mu_{3,k}\, c_k^{\otimes 3},$$

which is (2). The noise variance does not appear, so $\mathcal{T}$ is a functional of the reference environment's distribution.

Since $A_{e_{\text{ref}}}$ is nonsingular and $b_1, \ldots, b_K$ are linearly independent, the columns of C are linearly independent and the Kruskal rank of C is K. The decomposition above uses C in all three modes, so Kruskal's condition reads $3K \geq 2K + 2$, which holds for every $K \geq 2$; the decomposition into K rank-one terms is therefore essentially unique (Kruskal, 1977), that is, unique up to permutation of the terms and scalings of the three factors of each term whose product is one. Writing $v_k = c_k / \|c_k\|$ and $\rho_k = \mu_{3,k} \|c_k\|^3$, each term is $\rho_k v_k^{\otimes 3}$. Unit normalization leaves only $v_k \mapsto s_k v_k$, $\rho_k \mapsto s_k \rho_k$ and permutation ambiguity. For $K = 1$ any flattening of the rank-one tensor fixes its direction. Thus the normalized directions $\{v_k\}$, rather than their unconstrained magnitudes, are determined by the cumulant. The assumption $\mu_{3,k} \neq 0$ is what places all K terms in the decomposition: if $\mu_{3,k} = 0$ then c_k is absent from $\mathcal{T}$ and no procedure based on third-order information can recover it. □

Proposition 1 is an instance of the classical identifiability of linear, non-Gaussian latent variable models (Comon, 1994; Anandkumar et al., 2014): in the reference environment, (1) is undercomplete linear independent component analysis with mixing matrix $A_{e_{\text{ref}}} D$ and noise of vanishing third cumulant. We state it because it removes the pairing requirement from the identifiability claim, not because the argument is new.

Finite-sample rate. We use the injective norm $\|\mathcal{H}\|_{\text{op}} = \sup_{\|x\|=\|y\|=\|z\|=1} |\mathcal{H}(x, y, z)|$. Let u_i denote population-centered observations and assume they are i.i.d. with $\sup_{\|v\|=1} \|\langle v, u_i \rangle\|_{\psi_2} \leq R$, where $\|X\|_{\psi_\alpha} = \inf\{s > 0 : \mathbb{E}\exp(|X/s|^\alpha) \leq 2\}$. This is an explicit joint sub-Gaussian assumption; its scale includes the loadings, mixing matrix, and noise.

Concentration. For fixed unit x, y, z, the centered product $Z_i = \langle x, u_i \rangle \langle y, u_i \rangle \langle z, u_i \rangle - \mathbb{E}[\langle x, u_i \rangle \langle y, u_i \rangle \langle z, u_i \rangle]$ has $\|Z_i\|_{\psi_{2/3}} \leq CR^3$ by Hölder's inequality. The sub-Weibull sum bound of Kuchibhotla & Chakrabortty (2022, Theorem 3.1) therefore gives

$$\Pr\left\{ \left| \frac{1}{N} \sum_i Z_i \right| > CR^3 \left[\sqrt{\frac{t}{N}} + \frac{t^{3/2}}{N} \right] \right\} \leq 2e^{-t}.$$

A $1/6$-net of the unit sphere has at most 13^d points. Applying the bound to all triples of net points and using multilinearity gives the same bound, up to a factor of two, for $\|N^{-1} \sum_i u_i^{\otimes 3} - \mathcal{T}\|_{\text{op}}$ with $t = h = 3d \log 13 + \log(6/\delta)$ and failure probability at most $\delta/3$. Indeed, replacing the three unit

arguments by their nearest net points changes a trilinear form by at most $3(1/6)$ times its injective norm.

Sample centering preserves this rate, but is not generally of lower order. With $S_2 = N^{-1}\sum_i u_i u_i^\top$ and $\bar{u} = N^{-1}\sum_i u_i$, the exact identity is

$$\widehat{\mathcal{T}} = \frac{1}{N}\sum_i u_i^{\otimes 3} - \mathrm{Sym}(\bar{u} \otimes S_2) + 2\bar{u}^{\otimes 3},$$

where Sym sums the three placements of the vector argument. Sphere-net bounds for the sample mean and second moment give, on an event of probability at least $1 - 2\delta/3$, $\|\bar{u}\| \le CR\sqrt{h/N}$ and $\|S_2\|_{\mathrm{op}} \le CR^2(1 + h/N)$. For $N \ge h$, the centering correction is thus at most $3\|\bar{u}\|\|S_2\|_{\mathrm{op}} + 2\|\bar{u}\|^3 \le CR^3\sqrt{h/N}$. Combining these events proves the concentration claim in Proposition 2. This uses concentration for unbounded sub-Weibull products, rather than applying bounded Bernstein directly to cubic sub-Gaussian observations. The exponential loadings used in the recovery experiments are not covered by this sub-Gaussian assumption; their measured scaling remains an empirical result.

Contractions and estimator. Set $v_k = c_k/\|c_k\|$, $F = [v_1, \dots, v_K]$, and $\rho_k = \mu_{3,k}\|c_k\|^3$. For two unit vectors w_1, w_2, drawn independently of the data, define

$$M_j = \mathcal{T}(I, I, w_j) = FD_jF^\top, \qquad D_j = \mathrm{diag}(\rho_k\langle v_k, w_j\rangle), \quad j = 1, 2.$$

Assume $\min_k |\rho_k\langle v_k, w_2\rangle| > 0$ and, for $K > 1$, $g = \min_{k\neq\ell}|r_k - r_\ell| > 0$, where $r_k = \langle v_k, w_1\rangle/\langle v_k, w_2\rangle$. Independent uniform sphere draws satisfy these conditions almost surely for linearly independent factors, but they do not give a uniform positive gap. Let $s = \sigma_K(M_2) > 0$, $m = \|M_1\|_{\mathrm{op}}$, and $\chi = \|F\|_{\mathrm{op}}/\sigma_K(F)$. The contraction margin enters through $s \ge \sigma_K(F)^2 \min_k |\rho_k\langle v_k, w_2\rangle|$. For $K = 1$, set $g = \infty$.

Form $\widehat{M}_j = \widehat{\mathcal{T}}(I, I, w_j)$, let $\widehat{U} \in \mathbb{R}^{d\times K}$ contain the K leading left singular vectors of $\widehat{M}_2$, and set $\widehat{B}_j = \widehat{U}^\top\widehat{M}_j\widehat{U}$. Recover unit directions by lifting the eigenvectors of $\widehat{B}_1\widehat{B}_2^{-1}$ through $\widehat{U}$ and normalizing (Leurgans et al., 1993). Restricting to this estimated rank-K subspace prevents the null space of the ambient $d \times d$ pencil from introducing spurious components.

Local perturbation bound. Write $\epsilon = \|\widehat{\mathcal{T}} - \mathcal{T}\|_{\mathrm{op}}$. Then $\|\widehat{M}_j - M_j\|_{\mathrm{op}} \le \epsilon$. If $\epsilon \le cs$, the singular subspace error is at most $C\epsilon/s$. Consequently $H = \widehat{U}^\top F$ is invertible, has condition number at most 2χ, and $B_j^0 = \widehat{U}^\top M_j\widehat{U} = HD_jH^\top$ satisfies $\|(B_2^0)^{-1}\|_{\mathrm{op}} \le 2/s$. Since $\|\widehat{B}_j - B_j^0\|_{\mathrm{op}} \le \epsilon$, the inverse identity gives

$$\left\|\widehat{B}_1\widehat{B}_2^{-1} - B_1^0(B_2^0)^{-1}\right\|_{\mathrm{op}} \le C\frac{1 + m/s}{s}\epsilon.$$

The population compressed pencil is $B_1^0(B_2^0)^{-1} = H\,\mathrm{diag}(r_k)H^{-1}$. For

$$\epsilon \le cs\min\left\{1, \frac{g}{\chi(1 + m/s)}\right\},$$

its eigenvalues remain separated. Expressing an estimated eigenvector in the columns of H bounds its off-component coordinates by the pencil perturbation divided by g; multiplying by H costs at most its condition number. After lifting and normalizing, the eigenvector error and the subspace error give

$$\min_{\pi\in S_K}\max_k \sin\angle(\hat{b}_k, v_{\pi(k)}) \le C\left[\frac{1}{s} + \frac{\chi(1 + m/s)}{sg}\right]\epsilon = C\kappa_{\mathrm{CP}}\epsilon.$$

Here $c, C > 0$ are universal and the second term is zero for $K = 1$. This proves the local deterministic stability claim and, with the concentration bound, equation (6). It does not assert a dimension-independent conditioning constant for arbitrary random contractions.

Metric. Angles are taken between unoriented lines. After sign and permutation alignment, the chord distance between unit directions is $2\sin(\angle/2) \le \sqrt{2}\sin\angle$ for $0 \le \angle \le \pi/2$. Thus, when greedy matching recovers the correct correspondence, the maximum-angle bound also controls the worst matched chord error reported in the experiments; that error lies in $[0, \sqrt{2}]$.

What the implemented pipeline does instead. Proposition 2 describes a whitening-free estimator. The pipeline in Section 3.2 whitens first; its conditional recovery guarantee requires the reference-coordinate signal covariance C_{ref}. The paired implementation uses cross-moments only for subspace selection and then whitens a projected auto-covariance, retaining observation noise (Appendix A). Paired cross-environment second moments are noise-free in expectation, but require appropriate known alignment to supply this covariance; with identity maps they supply it directly. Without pairing the pipeline uses $\mathrm{Cov}(u_{ie_{\mathrm{ref}}}) = C_{\mathrm{ref}} + \mathrm{Cov}(\varepsilon)$, so orthogonal decomposability is not generally preserved. Absent noise or a consistent noise-covariance correction supplies the signal-covariance whitening required by the recovery guarantee. Under signal-covariance whitening, the component weights are $\mu_{3,k}/\sigma_k^3$, where $\sigma_k^2 = \mathrm{Var}(a_{ik})$, and whitening becomes less stable as the smallest signal-covariance eigenvalue decreases. We make no convergence-rate exponent claim for the implemented real-data whitening procedure. The directions reported on the real benchmarks therefore do not inherit Proposition 2; implementing the whitening-free solver would close this gap and is left for future work. The next section proves the conditional recovery bound for estimated signal-covariance whitening and states what aligned paired data supply.

C Proof of Theorem 4 (Program recovery with signal-covariance whitening)

This section gives a recovery guarantee conditional on estimating the signal covariance in the same coordinates as the third cumulant. Paired observations provide such a covariance when their environment maps are known and aligned; pairing alone is insufficient. Identifiability itself is established by Proposition 1 without pairing. The result below concerns a whitened estimator with the robust tensor-power algorithm and sufficient restarts, rather than an unconditional guarantee for either implemented whitening procedure.

Fix the reference environment and write $c_k = A_{e_{\mathrm{ref}}} b_k$, $\sigma_k^2 = \mathrm{Var}(a_{ik}) > 0$, and

$$S = C_{\mathrm{ref}} = \sum_{k=1}^{K} \sigma_k^2 c_k c_k^\top, \qquad \mathcal{T} = \sum_{k=1}^{K} \mu_{3,k} c_k^{\otimes 3}.$$

Assume that the c_k are linearly independent, the loading coordinates are independent and centered, and the observation noise is independent of the loadings and has zero third cumulant. Let $m = \lambda_K(S) > 0$, $M = \lambda_1(S)$, and $\lambda_* = \min_k |\mu_{3,k}|/\sigma_k^3 > 0$. Matrix norms below are operator norms; for tensors, $\|\mathcal{T}\|_{\mathrm{op}} = \sup_{\|x\|=\|y\|=\|z\|=1} |\mathcal{T}(x,y,z)|$.

Theorem 4 (Program recovery with estimated signal covariance)**.** Let $\hat{S}$ be a symmetric estimate of S and $\hat{\mathcal{T}}$ a symmetric estimate of $\mathcal{T}$, with

$$\|\hat{S} - S\|_{\mathrm{op}} \le \epsilon_S \le m/4, \qquad \|\hat{\mathcal{T}} - \mathcal{T}\|_{\mathrm{op}} \le \epsilon_T.$$

Using the top K eigenpairs of $\hat{S}$, define $\hat{W} = \hat{U}_K \hat{\Lambda}_K^{-1/2}$ and $\hat{B} = \hat{U}_K \hat{\Lambda}_K^{1/2} = \hat{S}\hat{W}$. Set

$$\tau = \frac{2\sqrt{2}\,\epsilon_T}{m^{3/2}} + \frac{12\|\mathcal{T}\|_{\mathrm{op}}\epsilon_S}{m^{5/2}},$$
$$\beta = \left(\sqrt{\frac{2}{m}} + \frac{2M}{m^{3/2}}\right)\epsilon_S.$$

For $K \ge 2$ and $0 < \tau \le c\lambda_*/K$, with a sufficiently small universal constant c, apply the robust tensor-power algorithm of Anandkumar et al. (2014, Algorithm 1 and Theorem 5.1) to $\hat{\mathcal{T}}(\hat{W}, \hat{W}, \hat{W})$ with its prescribed iteration and restart budgets. For any algorithmic failure probability $\zeta \in (0,1)$, the resulting normalized estimates $\hat{b}_k = \hat{B}\hat{q}_k/\|\hat{B}\hat{q}_k\|$ satisfy, with probability at least $1-\zeta$ over the algorithm's random starts,

$$\max_k \left\| \hat{b}_k - s_k \frac{c_{\pi(k)}}{\|c_{\pi(k)}\|} \right\| \le \frac{2}{\sqrt{m}} \left(\beta + \frac{8\sqrt{M+\epsilon_S}}{\lambda_*}\tau \right) \tag{8}$$

for a permutation π and signs s_k. For fixed dimension and fixed nondegenerate population moments, if $\epsilon_S, \epsilon_T = O_{\mathbb{P}}(N^{-1/2})$, then the direction error is $O_{\mathbb{P}}(N^{-1/2})$ when the algorithmic failure probability tends to zero. No separation between different third-cumulant weights is required. For $K = 1$ use direct scalar decomposition; when $\tau = 0$, exact orthogonal decomposition gives zero recovery error.

For the thresholded procedure in Appendix A, this K-component guarantee also requires that the top K eigenpairs of $\hat{S}$ survive truncation: $\hat{\lambda}_K \geq \tau_{\mathrm{cut}} \hat{\lambda}_1 > 0$. A fixed relative cutoff can exclude a nonzero signal eigenvalue; consistency of the covariance estimate alone does not ensure this condition.

Lemma 5 (Aligned paired moments and concentration)**.** Suppose subject pairs are independent and identically distributed across subjects. For two environments, let the known alignment be $R = A_{e_{\mathrm{ref}}} A_{e'}^{-1}$, and put $X_i = u_{ie_{\mathrm{ref}}}$ and $Y_i = R u_{ie'}$. Assume centered observations, independent noise across environments and from the loadings, and finite sixth moments of X_i and Y_i. Then

$$\mathbb{E}[X_i Y_i^\top] = S, \qquad \mathbb{E}[X_i^{\otimes 3}] = \mathcal{T}.$$

Define $Z_i = (X_i Y_i^\top + Y_i X_i^\top)/2$, $V_2 = \mathbb{E}\|Z_i - S\|_F^2$, and $V_3 = \mathbb{E}\|X_i^{\otimes 3} - \mathcal{T}\|_F^2$. For $\hat{S}_0 = N^{-1} \sum_i Z_i$ and $\hat{\mathcal{T}}_0 = N^{-1} \sum_i X_i^{\otimes 3}$, with probability at least $1 - \delta$ both bounds hold:

$$\|\hat{S}_0 - S\|_{\mathrm{op}} \leq \sqrt{\frac{2V_2}{N\delta}}, \tag{9}$$

$$\|\hat{\mathcal{T}}_0 - \mathcal{T}\|_{\mathrm{op}} \leq \sqrt{\frac{2V_3}{N\delta}}. \tag{10}$$

Replacing population centering by sample centering preserves the $O_{\mathbb{P}}(N^{-1/2})$ rate at fixed dimension.

Proof. The aligned observations have the same signal $\sum_k a_{ik} c_k$. Independence and zero means eliminate the signal–noise and cross-environment noise terms in the second cross-moment. For the third moment, independence makes cumulants additive; the noise contributes zero and mixed loading cumulants vanish. Finite sixth moments imply $V_2, V_3 < \infty$. Independence across subjects gives $\mathbb{E}\|\hat{S}_0 - S\|_F^2 = V_2/N$ and $\mathbb{E}\|\hat{\mathcal{T}}_0 - \mathcal{T}\|_F^2 = V_3/N$. Markov's inequality, the operator-to-Frobenius norm bound, and a union bound prove the displays. Sample means are $O_{\mathbb{P}}(N^{-1/2})$; expanding the centered products yields an $O_{\mathbb{P}}(N^{-1})$ correction to the second moment and an $O_{\mathbb{P}}(N^{-1/2})$ correction to the third. This argument does not assert a dimension-free constant or a sub-Gaussian concentration bound. □

Remark 1 (What supplies the signal covariance). The cross-moment before alignment is $A_e \Sigma A_{e'}^\top$, which is generally neither symmetric nor the covariance in the reference coordinates. The alignment in Lemma 5 is additional information; it is not supplied by pairing or by a rank condition on the environment maps. Identity maps require no alignment. Alternatively, a consistent estimate of the noise covariance can be subtracted from the reference auto-covariance. Theorem 4 applies to either construction if its stated estimation errors hold, including any error from estimating the alignment or noise covariance.

Lemma 6 (Whitening perturbation)**.** Under Theorem 4, let $G = \hat{W}^\top S \hat{W}$ and $W = \hat{W} G^{-1/2}$. Then $W^\top S W = I_K$, and

$$\|\hat{W}\|_{\mathrm{op}}, \|W\|_{\mathrm{op}} \leq \sqrt{2/m}, \qquad \|\hat{W} - W\|_{\mathrm{op}} \leq \frac{2\epsilon_S}{m^{3/2}}. \tag{11}$$

Proof. Weyl's inequality gives $\hat{\lambda}_K \geq m - \epsilon_S > 0$ and $\|\hat{W}\|_{\mathrm{op}} \leq (m - \epsilon_S)^{-1/2}$. Since $\hat{W}^\top \hat{S} \hat{W} = I_K$,

$$\|G - I_K\|_{\mathrm{op}} \leq r := \frac{\epsilon_S}{m - \epsilon_S} \leq \frac{1}{3}.$$

Thus G is positive definite, W whitens S, and $\|W\|_{\mathrm{op}} \leq (m - 2\epsilon_S)^{-1/2} \leq \sqrt{2/m}$. Spectral calculus gives $\|G^{-1/2} - I_K\|_{\mathrm{op}} \leq (1 - r)^{-1/2} - 1$. Multiplication by $(m - \epsilon_S)^{-1/2}$ and $\epsilon_S \leq m/4$ yield the second bound in (11). This sample-dependent population whitening avoids assuming that empirical and population eigenvectors agree within the signal subspace. □

Lemma 7 (Whitened tensor perturbation)**.** With W as above, $\widetilde{\mathcal{T}} = \mathcal{T}(W, W, W)$ is orthogonally decomposable and

$$\|\hat{\mathcal{T}}(\hat{W}, \hat{W}, \hat{W}) - \widetilde{\mathcal{T}}\|_{\mathrm{op}} \leq \tau. \tag{12}$$

Proof. Let $L = [\sigma_1 c_1, \ldots, \sigma_K c_K]$ and $Q = W^\top L$. Since $QQ^\top = W^\top S W = I_K$, the columns q_k of the square matrix Q form an orthonormal basis. Consequently

$$\widetilde{\mathcal{T}} = \sum_k \frac{\mu_{3,k}}{\sigma_k^3} q_k^{\otimes 3}.$$

Absorb each coefficient's sign into q_k when applying the tensor-power theorem. To control perturbations, first change the tensor at fixed $\hat{W}$, then telescope the three whitening arguments:

$$\begin{aligned}\|\hat{\mathcal{T}}(\hat{W},\hat{W},\hat{W}) - \mathcal{T}(W,W,W)\|_{\mathrm{op}} \\ \le \epsilon_T \|\hat{W}\|_{\mathrm{op}}^3 + \|\mathcal{T}\|_{\mathrm{op}} \|\hat{W} - W\|_{\mathrm{op}} \\ \cdot \big(\|\hat{W}\|_{\mathrm{op}}^2 + \|\hat{W}\|_{\mathrm{op}}\|W\|_{\mathrm{op}} + \|W\|_{\mathrm{op}}^2\big).\end{aligned}$$

Lemma 6 bounds this by τ. The inequality includes the error from using estimated whitening with the estimated tensor, without an omitted mixed remainder. □

Lemma 8 (Robust tensor-power recovery)**.** For a symmetric orthogonal tensor $\sum_k \lambda_k q_k^{\otimes 3}$ with positive coefficients, let $\lambda_* = \min_k \lambda_k > 0$. For $K \ge 2$, if the symmetric perturbation is at most $0 < \tau \le c\lambda_*/K$, the robust algorithm of Anandkumar et al. (2014, Theorem 5.1), with the iteration and restart counts specified there, satisfies

$$\|\hat{q}_k - q_{\pi(k)}\| \le \frac{8\tau}{\lambda_{\pi(k)}}, \qquad |\hat{\lambda}_k - \lambda_{\pi(k)}| \le 5\tau \tag{13}$$

simultaneously with probability at least $1 - \zeta$. Distinct coefficients are unnecessary. For $K = 1$, direct scalar decomposition gives the same bounds when $\tau < \lambda_*/2$; zero perturbation admits exact decomposition. The cited guarantee applies to that algorithm, including its restart selection and deflation; an arbitrary fixed-budget implementation does not automatically satisfy it.

Lemma 9 (Estimated unwhitening)**.** Let $B = SW$ and $\hat{B} = \hat{S}\hat{W}$. Then $BB^\top = S$, $Bq_k = \sigma_k c_k$ before the sign absorption above, $\|\hat{B} - B\|_{\mathrm{op}} \le \beta$, and $\|\hat{B}\|_{\mathrm{op}} \le \sqrt{M + \epsilon_S}$. If aligned whitened directions have error at most ϵ_q, then

$$\max_k \left\| \frac{\hat{B}\hat{q}_k}{\|\hat{B}\hat{q}_k\|} - s_k \frac{c_{\pi(k)}}{\|c_{\pi(k)}\|} \right\| \le \frac{2}{\sqrt{m}} \big(\beta + \sqrt{M + \epsilon_S}\, \epsilon_q\big). \tag{14}$$

Proof. The notation in Lemma 7 gives $B = LQ^\top$, proving the first two identities and $\|Bq_k\| \ge \sqrt{m}$. Also

$$\|\hat{B} - B\|_{\mathrm{op}} \le \|\hat{S} - S\|_{\mathrm{op}} \|\hat{W}\|_{\mathrm{op}} + \|S\|_{\mathrm{op}} \|\hat{W} - W\|_{\mathrm{op}} \le \beta.$$

The sample eigenvalues give the stated norm bound for $\hat{B}$. Thus $\|\hat{B}\hat{q}_k - s_k B q_{\pi(k)}\| \le \sqrt{M + \epsilon_S}\epsilon_q + \beta$. The normalization inequality $\|x/\|x\| - y/\|y\|\| \le 2\|x - y\|/\|y\|$ proves (14). The matrix $\hat{B} = (\hat{W}^\dagger)^\top$ is the appropriate rectangular unwhitening map; no ordinary inverse of the $d \times K$ whitening matrix is used. □

Proof of Theorem 4. Lemma 6 constructs an exact population whitening in the sample coordinates. Lemma 7 bounds the resulting tensor perturbation by τ. Lemma 8 then gives $\epsilon_q \le 8\tau/\lambda_*$, and substitution into Lemma 9 proves (8). These arguments are deterministic conditional on the moment estimates, apart from the power algorithm's random starts. If the moment bounds hold with probability $1 - \delta$, the combined success probability is at least $1 - \delta - \zeta$. With fixed m, M, λ_* and $\|\mathcal{T}\|_{\mathrm{op}}$, both τ and β are $O_{\mathbb{P}}(N^{-1/2})$. The small-perturbation condition therefore holds with probability tending to one. Lemma 5 supplies such rates for known aligned pairs under finite sixth moments. □

Remark 2 (Conditioning dependence). The bound depends on the smallest and largest eigenvalues of the signal covariance, the third-moment norm, and the smallest standardized third cumulant. These quantities depend on the program matrix and loading distribution as well as the reference environment map. A condition number for the environment map alone does not control them, so we do not assign a universal exponent of $\kappa(\mathbf{A})$ or a sample threshold depending only on d and K. The conditioning sweep in Table 6 measures empirical sensitivity under its specified generator.

Remark 3 (Unpaired practical estimator). On the real benchmarks, each subject is observed in one environment and aligned paired cross-moments are unavailable. The implemented whitening uses the reference auto-covariance $S + \mathrm{Cov}(\varepsilon)$, not an estimate converging to S. With nonzero uncorrected noise covariance, its error relative to S need not vanish as N grows, so Theorem 4 does not establish consistency for that procedure. A separately consistent noise-covariance correction would meet the theorem's covariance requirement. Otherwise, the whitening-free approach in Appendix B is a distinct estimator; population identifiability remains valid under Proposition 1.

D Proof of Proposition 3 (Non-identifiability under latent reparameterization)

Proof. We construct two models with identical observation laws and separated sets of normalized reference-coordinate directions. This is the zero-total-variation case of Le Cam's two-point argument (LeCam, 1973; Yu, 1997).

Step 1: Setup. Choose a dictionary $D \in \mathbb{R}^{d\times K}$ with orthonormal columns, $K \geq 2$, and any centered loading distribution with finite third moments. Keep the nonsingular transforms $\{A_e\}$ and the independent Gaussian observation noise fixed in both models. Let Q rotate the first two coordinate axes through $\pi/4$ and leave the remaining coordinates unchanged. Both D and DQ have unit, linearly independent columns.

Step 2: Constructing indistinguishable models. Set

$$D^{(1)} = D, \qquad D^{(2)} = DQ, \qquad a^{(2)} \overset{d}{=} Q^\top a^{(1)}. \tag{15}$$

The loading distribution is unknown and unrestricted in $\mathcal{M}$, so both models are admissible. Under this coupling,

$$A_e D^{(2)} a^{(2)} = A_e DQQ^\top a^{(1)} = A_e D^{(1)} a^{(1)} \tag{16}$$

for every environment. Thus the entire observation law is the same, including paired observations when available. Extra environments cannot distinguish these two models.

Let $\bar{C}^{(j)}$ contain the normalized columns of $A_{e_{\mathrm{ref}}} D^{(j)}$. The two sets differ even after signs and permutations: each rotated column is a nontrivial linear combination of the first two original columns, and nonsingularity of the reference transform preserves their linear independence. Consequently,

$$\rho = d_{\pm}(\bar{C}^{(1)}, \bar{C}^{(2)}) > 0.$$

For identity reference transform, $\rho = \sqrt{2-\sqrt{2}}$; for any fixed nonsingular reference transform it remains a positive constant independent of N.

Step 3: Le Cam bound. The loss in (7) is a metric on direction sets modulo signs and permutations. Its triangle inequality gives, for every estimate,

$$d_{\pm}(\hat{B}, \bar{C}^{(1)}) + d_{\pm}(\hat{B}, \bar{C}^{(2)}) \geq \rho.$$

Taking expectations under the common observation law shows that at least one risk is $\rho/2$ or larger. Taking the infimum over estimators proves the proposition with $c_0 = \rho/2$. This bound concerns directional ambiguity, not arbitrary scaling or labels. □

Remark 4 (How the theorem's conditions restore identifiability). The reparameterization preserves observations but generally changes the dependence between loading coordinates. Requiring independent loadings with nonzero third cumulants rules out the rotated alternative: Proposition 1 fixes the normalized directions from the reference tensor alone. The loading law need not be known, and its third cumulants need not be distinct. The remaining scaling ambiguity disappears on normalization, leaving signs and permutations. The paired recovery theorem additionally requires a consistent estimate of the reference signal covariance for whitening.

Remark 5 (Number of environments). One environment suffices for third-cumulant identifiability. Two paired views with independent noise can also provide the signal covariance when their relative alignment is known; more views can supply additional moment estimates. A count of environments or a rank condition on their transforms alone does not provide that alignment. Theorem 4 therefore states the required signal-covariance estimation condition explicitly.

Remark 6 (Non-degeneracy boundary). If a loading has zero third cumulant, its direction is absent from the third-order tensor; this does not by itself rule out recovery using higher moments. Gaussian loadings give a stronger impossibility example. With $a \sim \mathcal{N}(0, I_K)$, both a and $Q^\top a$ have independent standard Gaussian coordinates, and rotating D leaves every observed signal covariance $A_e D D^\top A_e^\top$ unchanged. The preceding ambiguity then survives even under loading independence, as in classical Gaussian-source ICA (Comon, 1994).

E Training Dynamics

The Stage 2 strategy comparison appears in Table 2 in the main text.

Rank-one operators and recovery scoring. Each operator is stored as $B_k = U_k \operatorname{diag}(s_k) V_k^\top$ with fixed rank $r = 1$. The factors are orthonormalized only at initialization. Random initialization uses QR factors and $s_k = 0.01$; spectral initialization factors $\hat{b}_k \hat{b}_k^\top$, with $U_k = V_k = \pm\hat{b}_k$ and $s_k = 1$. Updates do not preserve equal factors or unit norms. The scored direction is the normalized right factor $V_k[:, 1]$, so the reported error measures input-side alignment, not the left factor or complete operator. For a nonzero rank-one operator this is its right singular direction up to sign. Frozen spectral factors retain the Stage-1 direction. Matching and aggregation are defined in Appendix I.5.

Synthetic Stage-2 protocol. The three spectral rows of Table 2 use $N = 5000$, $E = 5$, $K = 3$, $d = 20$, $\sigma = 0.1$, identity transforms and five seeds. The encoder is $20 \to 64 \to 20$ with a scalar linear head. Each subject is replicated across the five environments with generating loadings and the common regression target $y_i = (m_i)_1$. Training uses Adam at 10^{-2}, 50 epochs, batch size 256 and patience 20; the style branch is off and the loading penalty is constant. The contrastive row uses weight 0.1. The reported loss is the last training-epoch loss, not a held-out loss. The random frozen row is drawn from the separate freeze–thaw experiment below; it is not the random-trainable arm of this script.

Table 4 uses the same synthetic data, model, inputs, target and Adam settings as above. Operators are frozen for 50 epochs, then trained for another 50; early stopping is disabled. Loss is the final training-epoch loss from each phase, and summaries use population standard deviations. Thawing preserves the spectral arm's training loss while increasing its direction error.

Table 4: Freeze and thaw on synthetic data: 50 epochs per phase, without early stopping. Loss is the final training-epoch loss; summaries are mean ± population s.d. In the spectral arm, thawing leaves training loss essentially unchanged and raises recovery error by an order of magnitude. Random directions remain near the random-direction level. Five seeds were run; the random-thawed recovery entry uses the three finite seeds after two diverged.

	Training loss		Recovery error	
Initialization	frozen	thawed	frozen	thawed
Spectral	$.0101 \pm .0001$	$.0101 \pm .0001$	$.031 \pm .005$	$.376 \pm .166$
Random	$.0116 \pm .0004$	$.0115 \pm .0004$	$1.335 \pm .035$	$1.340 \pm .044$

F Additional Synthetic Experiments

F.1 Robustness to Environment Conditioning

This section reports the grids behind Section 5.1. The conditioning and perturbation grids use $N = 6000$, $E = 5$, $K = 3$, $d = 20$, noise $\sigma = 0.1$, and 5 seeds. All environments are perturbed. Conditioning is scored in the selected reference coordinates; the other grids compare with latent directions without transforming the target (Appendix I.1). The unperturbed control is 0.029 ± 0.002 and a random-direction baseline is 1.30 ± 0.05.

The observed gap grows with noise. The implemented cross-moments improve subspace estimation but do not remove noise from the projected whitening covariance. Identification does not require a

Table 5: Operator and encoder controls on Camelyon17 (leave-one-hospital-out, shared data and splits). The dense MLP has one hidden layer; the operator models have two, so this comparison does not isolate operatorization. Spectral recovery is run on the training hospitals alone. The four configurations differ by less than the seed variability. These controls resolve neither an accuracy effect of the basis nor a predictive gain large relative to variability over the shallower dense MLP. What separates the bases is recovery of the generating programs on synthetic data and biological alignment on TCGA-BRCA (Section 5.3), not accuracy. These controls use the dimensionality-scan seeds; the corresponding main-table means are 0.851 ± 0.031 for FIRE and 0.848 ± 0.030 for ERM.

Configuration	OOD accuracy
Dense MLP (no operators)	0.847 ± 0.029
Random frozen operators	0.854 ± 0.032
Random trainable operators	0.852 ± 0.030
Spectral frozen operators	0.854 ± 0.032

Table 6: Recovery error as the environment transforms become ill-conditioned. Error grows with κ but stays an order of magnitude below the random-direction level throughout, so the generator tolerates substantial ill-conditioning. Theorem 4 depends on signal-covariance eigenvalues and cumulant strengths as well as the environment map; the sweep does not establish a universal condition-number exponent.

κ	1	3	10	30	100	300	1000
Recovery error	$.028 \pm .002$	$.028 \pm .009$	$.029 \pm .008$	$.041 \pm .015$	$.062 \pm .046$	$.078 \pm .061$	$.098 \pm .078$

second environment; these results do not isolate variance from whitening bias. At low noise FastICA on one environment is as good or better; this baseline is distinct from the whitening-free estimator analyzed in Proposition 2.

A fourth sweep adds $\Delta s_{p(i)}$ to each subject's loadings, where $p(i)$ is a balanced random group assignment and $s_e \sim \mathcal{N}(0, I_K)$ (Table 11). Paired observations retain every subject in every environment, so all environments share the same mixture loading law; this arm does not violate Assumption 1. The mixture can instead induce dependence between loading coordinates. In the pseudobulk arm, subject i is observed only in $p(i)$; missing cells are filled by each environment's observed mean and become zero after centering. Centering removes this group's constant mean shift, not the paired mixture's between-group variation. These arms do not isolate an exchangeability effect or establish robustness to changes in variance, skewness or higher moments.

G Per-Benchmark Results

G.1 Benchmark Summary and Auxiliary fMRI Results

Table 1 reports the numbers behind Figure 2(a), and Table 12 the ADNI fMRI boundary case discussed below.

The complete comparison appears in the main text as Table 1. For Colored MNIST, Camelyon17, OASIS-1, ADNI sMRI and TCGA-BRCA, respectively, the strongest competing baselines are SD, ERM, IRM, DANN and the IRM/Group-DRO tie. The corresponding FIRE-minus-baseline mean gaps are $+.011$, $+.003$, $-.022$, $+.002$ and $+.012$. Their 95% confidence intervals are not reported: aggregate means and standard deviations do not supply the paired seed-level differences needed to estimate them. The comparisons are descriptive.

FIRE has the highest mean accuracy on four of the five benchmarks, with margins over the strongest competing baseline of 0.011 on Colored MNIST, 0.003 on Camelyon17, 0.002 on ADNI sMRI and 0.012 on TCGA-BRCA; on OASIS-1 IRM is ahead by 0.022. The margins are small relative to the reported seed variability, and paired seed-level results are not available for significance tests. FIRE's mean gaps over Group-DRO on Camelyon17 ($+0.218$) and Mixup on TCGA-BRCA ($+0.282$) are larger, but we do not attach unsupported confidence intervals to them. Group-DRO is the one method

Table 7: Latent-direction deviation under three observation-map perturbations of strength ε. Spatial variation remains a global linear transform; the untransformed target and unlogged reference selection prevent interpreting these values as reference-coordinate recovery errors. Counts denote scores below 0.15.

ε	Nonlinear	Subject-dependent	Spatially heterogeneous
0.0	$.029 \pm .002$ (5/5)	$.026 \pm .002$ (5/5)	$.027 \pm .002$ (5/5)
0.1	$.034 \pm .005$ (5/5)	$.026 \pm .002$ (5/5)	$.062 \pm .008$ (5/5)
0.2	$.046 \pm .006$ (5/5)	$.027 \pm .003$ (5/5)	$.113 \pm .011$ (5/5)
0.35	$.300 \pm .505$ (4/5)	$.033 \pm .006$ (5/5)	$.191 \pm .019$ (0/5)
0.5	$.587 \pm .654$ (3/5)	$.039 \pm .007$ (5/5)	$.266 \pm .026$ (0/5)
0.75	$.863 \pm .642$ (1/5)	$.061 \pm .007$ (5/5)	$.383 \pm .039$ (0/5)
1.0	$1.338 \pm .079$ (0/5)	$.096 \pm .007$ (5/5)	$.488 \pm .050$ (0/5)

Table 8: Joint perturbation: nonlinearity ε_n against spatial heterogeneity ε_s, scored against latent directions as in Table 7. At $(\varepsilon_n, \varepsilon_s) = (0.5, 0.2)$ the error is 1.102, against 0.587 and 0.113 for each alone and about 0.67 after adding their increments above the control. This descriptive comparison is not an identification-failure test. The $\varepsilon_s = 0$ column reproduces the nonlinearity sweep of Table 7, a consistency check on the two runs.

ε_n	$\varepsilon_s = 0$	$\varepsilon_s = 0.2$	$\varepsilon_s = 0.5$
0.0	$.029 \pm .002$ (5/5)	$.113 \pm .011$ (5/5)	$.266 \pm .026$ (0/5)
0.1	$.034 \pm .005$ (5/5)	$.117 \pm .014$ (5/5)	$.268 \pm .028$ (0/5)
0.2	$.046 \pm .006$ (5/5)	$.123 \pm .016$ (5/5)	$.488 \pm .495$ (0/5)
0.35	$.300 \pm .505$ (4/5)	$.345 \pm .455$ (3/5)	$1.094 \pm .469$ (0/5)
0.5	$.587 \pm .654$ (3/5)	$1.102 \pm .544$ (1/5)	$1.092 \pm .471$ (0/5)
0.75	$.863 \pm .642$ (1/5)	$1.111 \pm .532$ (0/5)	$1.095 \pm .461$ (0/5)
1.0	$1.338 \pm .079$ (0/5)	$1.120 \pm .513$ (0/5)	$1.291 \pm .113$ (0/5)

that fails outright, on Colored MNIST and Camelyon17, and Mixup collapses on the five-class TCGA-BRCA task. The suite is non-exhaustive and we make no state-of-the-art claim.

G.2 ADNI fMRI Functional Connectivity

We also evaluate FIRE on ADNI functional MRI (fMRI) data to probe a more difficult multi-site neuroimaging setting, where multigraph and tensor approaches are common alternatives for functional-connectivity analysis (Wang et al., 2021; Zhang et al., 2020). The dataset contains $1{,}070$ scans from 923 unique subjects across seven scanner protocols (ADNI-2/3/GO $\times$ GE/Philips/Siemens). Features are 116×116 Fisher-z transformed ROI correlation matrices, vectorized to $6{,}670$ dimensions and reduced to 50 by principal component analysis (PCA). The task is binary classification of cognitively normal versus impaired subjects (AD+MCI+EMCI+LMCI), evaluated with leave-one-protocol-out splits over three seeds.

FIRE trails all baselines by roughly 10 percentage points. One possible explanation was that per-site z-normalization removed the site effects on which FIRE relies. Repeating the experiment with global z-normalization, which preserves site-specific scale shifts, produces nearly identical results: FIRE 0.583 ± 0.089, IRM 0.680 ± 0.149, ERM 0.674 ± 0.141, and DANN 0.667 ± 0.143. The preserved ordering does not identify the cause of the deficit. Mismatch between the data and the operator model $T_i = I + \sum_k a_{ik} B_k$, or conditioning of the covariance used for whitening, remain possible explanations.

H Direction Quality

H.1 What the Spectral Basis Adds (TCGA-BRCA)

The Camelyon17 controls resolve no predictive effect of the basis (Table 5); the direction-quality comparison below instead concerns TCGA-BRCA. We compare four bases in the same 50-dimensional

Table 9: What several environments buy. With paired subjects, cross-environment second moments are free of independent noise in expectation and select the recovery subspace; projected auto-covariance whitening still includes noise. The observed multi-environment scores degrade more slowly. FastICA uses environment 0. Synthetic, 5 seeds.

Noise σ	Multi-environment	Single-environment	FastICA
0.1	$.029 \pm .002$	$.029 \pm .003$	$.023 \pm .008$
0.5	$.042 \pm .005$	$.047 \pm .008$	$.047 \pm .006$
1.0	$.080 \pm .010$	$.104 \pm .014$	$.110 \pm .020$
2.0	$.186 \pm .022$	$.307 \pm .019$	$.457 \pm .156$
4.0	$.699 \pm .173$	$1.124 \pm .157$	$1.184 \pm .165$

Table 10: Cost and benefit of a larger recovery dimension on Camelyon17. Stage 1 costs $O(Nd^3)$ in time and $O(d^3)$ in memory, so d is a computational choice rather than a modeling limit at this scale. Accuracy improves slightly for both methods, so $d = 50$ is not a bottleneck on this benchmark, though a larger PCA dimension is not automatically better.

d	Stage-1 recovery time	FIRE	ERM
50	1.0 s	0.854 ± 0.032	0.847 ± 0.029
100	4.4 s	0.862 ± 0.028	0.859 ± 0.028
200	12.7 s	0.867 ± 0.025	0.865 ± 0.026

PCA space: the spectral directions, FastICA components, the raw PCA axes, and random unit directions. Table 13 scores them on TCGA-BRCA by the Fisher ratio of subject projections across PAM50 subtypes, which measures biological separation, and across tissue source sites, which measures batch alignment, over 8 subject-level bootstraps. The subtype, site, and biology-to-batch columns come from the same bootstrap run. The ratio is the mean of the per-bootstrap ratios, not the quotient of the two reported means. A separate bootstrap run, reported in the response to reviewers, gave a raw-PCA subtype Fisher ratio of 0.85 ± 0.04. We retain the single-run values here so that each ratio is paired with its corresponding subtype and site estimates.

The comparisons expose a trade-off. Spectral directions have the highest subtype-to-site Fisher ratio, and on synthetic data they recover the generating programs more accurately than random or raw PCA directions: error 0.029 ± 0.002, against 1.315 ± 0.035 for random directions and 0.379 ± 0.183 for raw PCA, with FastICA comparable to spectral at low noise (0.023 ± 0.008) and worse as noise grows (Table 9). Against that, PCA and ICA directions are markedly more stable across recovery seeds on this cohort than the spectral ones, limiting per-direction interpretation. The Fisher comparison uses a separate single-environment recovery; the application illustration in Figure 3 is not its source.

I Experimental Details

I.1 Synthetic Data Generation

For recovery rate and robustness experiments (Section 5.1 and Appendix F.1):

- **Dimension:** $d = 20$.
- **Programs:** $K = 3$ program directions $\mathbf{b}_k \in \mathbb{R}^d$, drawn i.i.d. from $\mathcal{N}(0, \mathbf{I}_d)$ and orthonormalized.
- **Environments:** $E = 5$ environment transforms $\mathbf{A}_e \in \mathbb{R}^{d \times d}$. For the identity setting, $\mathbf{A}_e = \mathbf{I}$. For the conditioning sweep, $\mathbf{A}_e = U_e \mathrm{diag}(\sigma_e) V_e^\top$ where U_e, V_e are random orthogonal matrices and σ_e is log-spaced from 1 to $1/\kappa$. The generic generator's optional nonidentity setting instead uses linearly spaced singular values from 1 to κ; it is not used for the reported conditioning sweep.
- **Loadings:** Proposition 1 requires nonzero third cumulants, not distinct values across programs. The generator draws $a_{ik} \sim \mathrm{Exponential}(1)$ and centers observations within each environment before moment estimation. For fixed linear maps this is equivalent to centered exponential loadings, with equal nonzero third cumulants; centering before a nonlinear or

Table 11: Group-dependent loading-mean shifts of strength Δ. Pseudobulk centering removes the within-environment constant shift; paired environments contain all groups and retain their mixture variation. The arms differ in observation design.

Δ	Paired multi-environment	Per-environment-centered pseudobulk
0.0	$.032 \pm .013$	$.074 \pm .041$
0.3	$.050 \pm .017$	$.074 \pm .041$
0.5	$.093 \pm .037$	$.074 \pm .041$
1.0	$.254 \pm .092$	$.074 \pm .041$
2.0	$.629 \pm .251$	$.074 \pm .041$

Table 12: ADNI fMRI, the boundary case. These numbers predate the corrected training setup of Table 1: the baselines were trained with squared error on a single logit, so these results do not establish the gap under the corrected matched-loss protocol.

Method	OOD accuracy
IRM	0.680 ± 0.149
ERM	0.674 ± 0.141
DANN	0.664 ± 0.142
FIRE	0.578 ± 0.084

subject-dependent perturbation is not equivalent. For comparison with these assumptions, consider (a) Gaussian loadings $a_{ik} \sim \mathcal{N}(0,1)$, whose third cumulants vanish, and (b) per-program skewed loadings $a_{ik} \sim \mathrm{Gamma}(\alpha_k, 1) - \alpha_k$ with $\alpha_k \in \{1, 2, 4\}$, which have distinct nonzero third cumulants. The Gaussian case lies outside the third-order identifiability condition. Neither recovery result requires distinct third cumulants. Exponential and gamma loadings are not sub-Gaussian, so Proposition 2 does not directly cover these generators. They do have finite sixth moments, as required by the fixed-dimensional moment bound in Lemma 5; applying the whitened recovery theorem additionally requires its signal-covariance and algorithmic conditions.

- **Noise:** $\varepsilon_{ie} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ with $\sigma = 0.1$.
- **Sample sizes:** $N \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$.
- **Seeds:** 5 independent random seeds per configuration.

Perturbation definitions and scoring targets. Let $z_i = Da_i$ use the uncentered exponential draws. Scaling, Stage-2 strategies, noise and loading-shift sweeps use identity transforms and score against b_k. Conditioning uses the transforms above in all environments and scores against $A_{e_{\mathrm{ref}}} b_k / \|A_{e_{\mathrm{ref}}} b_k\|$, with the selected reference read from the recovery diagnostics. For the remaining perturbation sweeps, $A_0 = I$ and other base maps have condition number 3. With scalar strengths ε, ε_n and ε_s, the noiseless perturbations are:

- **Nonlinear:** $t_{ie} = A_e z_i$ and $x_{ie} = t_{ie} + \varepsilon s_e \tanh(t_{ie}/s_e)$ elementwise, where s_e is the scalar standard deviation across the environment's entries.
- **Subject-dependent:** $x_{ie} = A_e(I + \varepsilon D_i) z_i$, with $D_i = M_i / \|M_i\|_{\mathrm{op}}$, independent Gaussian $d \times d$ matrices M_i, and one fixed D_i per subject across environments.
- **Spatial:** divide coordinates into four contiguous blocks with diagonal selectors P_b. Draw one normalized Gaussian matrix $G_{eb} = M_{eb} / \|M_{eb}\|_{\mathrm{op}}$ per environment and block, and set $x_{ie} = \sum_b P_b (A_e + \varepsilon G_{eb}) z_i$.
- **Combined:** apply the spatial map at strength ε_s, then the nonlinear map at strength ε_n, with the scale computed from the spatially transformed values.

All environments, including environment 0, receive the perturbation. These runs compare the estimated directions with b_k without a reference transform. Their reference is selected by tensor norm, not fixed to environment 0, and was not logged. An identity base map does not guarantee its selection. Moreover, the spatial map is $\widetilde{A}_e = A_e + \varepsilon \sum_b P_b G_{eb}$, a single fixed global linear map; when full rank, it remains within model (1). Even selecting environment 0 would give spatial reference directions

Table 13: Quality of the recovered directions on TCGA-BRCA. Higher subtype separation is better, lower site alignment is better, and the ratio summarizes the two. The spectral basis separates subtypes best while being no more site-aligned than ICA or PCA, so its advantage is not bought by absorbing batch structure. Cross-seed stability is the mean absolute cosine similarity between directions recovered from different random seeds.

Directions	Subtype Fisher ↑	Site Fisher ↓	Biology / batch ↑	Cross-seed stability
Spectral (ours)	$\mathbf{0.993 \pm 0.125}$	0.102 ± 0.024	$\mathbf{10.52 \pm 3.85}$	0.52 ± 0.10
ICA	0.859 ± 0.108	0.105 ± 0.010	8.25 ± 1.44	0.94 ± 0.04
Raw PCA	0.824 ± 0.070	0.108 ± 0.010	7.67 ± 0.71	0.94 ± 0.04
Random	0.32 ± 0.12	–	–	0.27 ± 0.04

$\widetilde{A}_0 b_k$, generally not b_k. We therefore report these grids as deviation from the latent directions, not evidence that spatial variation violates the linear model or defeats reference-coordinate recovery.

Pairing and noise sweep. The base generator measures every subject in all E environments with independent Gaussian observation noise. The noise sweep uses $\sigma \in \{0.1, 0.5, 1, 2, 4\}$: (i) the paired implementation in Appendix A; (ii) environment 0 with auto-covariance whitening; and (iii) FastICA on environment 0, with K components, 1000 maximum iterations, tolerance 10^{-4} and normalized components.

I.2 Training Hyperparameters

All Stage 2 training uses the following defaults unless otherwise noted:

- **Optimizer:** Adam, $\beta_1 = 0.9$, $\beta_2 = 0.999$; learning rate 10^{-3} on Colored MNIST and Camelyon17 and 5×10^{-4} on OASIS-1, ADNI sMRI and TCGA-BRCA.
- **Batch size:** 256 on Colored MNIST and Camelyon17, 32 on the three smaller cohorts.
- **Maximum epochs:** 30 on Colored MNIST and Camelyon17, 100 elsewhere.
- **Early stopping:** patience of 10 epochs on the larger benchmarks and 20 on the smaller ones, on validation loss.
- **Encoder architecture:** FIRE uses $d \to H_1 \to H_2 \to d$, ReLU after each hidden layer, no layer normalization and zero dropout. Hidden widths are $128{-}64$ on Colored MNIST, Camelyon17 and ADNI sMRI, and $64{-}32$ on OASIS-1 and TCGA-BRCA. The final affine layer returns to $d = 50$; rank-one 50×50 operators act there. A linear head maps to two class logits, or five for TCGA-BRCA. All baselines, including SD, use $d \to H_1 \to d$, with one hidden layer. Their binary heads output one logit; TCGA-BRCA heads output five. The Camelyon17 dense control uses this shallower encoder.
- **Programs:** $K = 3$ for all main predictive experiments, fixed a priori.
- **Seeds:** 10 per fold under leave-one-environment-out evaluation.
- **Loss weights:** $\lambda_1 = 0.01$ (sparsity, ℓ_1 on loadings), $\lambda_2 = 0.01$ (style separation).
- **Sparsity:** $\mathcal{L}_{\text{sparse}} = \sum_{i,k} |a_{ik}|$; inactive in the reported runs, since the loadings are inputs rather than parameters.
- **Style separation:** the proposed negative-dispersion term is inactive in the reported benchmarks and is not a well-posed unconstrained objective; see Appendix A.
- **Prediction loss:** binary cross-entropy with logits for single-logit baseline heads; softmax cross-entropy for FIRE's two-logit binary heads and all five-logit TCGA-BRCA heads. Regression uses MSE.
- **Test-time loadings:** ridge regression from training features onto the training loadings, fitted on source data only.

I.3 Spectral Recovery Hyperparameters

- **Whitening truncation:** eigenvalues below $\tau_{\text{cut}} \lambda_{\max}$ are discarded, with $\tau_{\text{cut}} = 10^{-6}$. The K-component description requires K retained positive eigenvalues.
- **Tensor power method restarts:** $L = 10$ (default), increased to $L = 20$ for $K > 5$.
- **Convergence tolerance:** the directional stopping criterion is $\|\theta_{t+1} - \theta_t\|_2 < 10^{-8}$ after normalization.

- **Maximum iterations per restart:** 1000.
- **Moment accumulation:** computed in `float64` to prevent catastrophic cancellation, even when the model operates in `float32`.
- **Condition number monitoring:** a warning is logged if $\kappa(\hat{\mathbf{M}}_{ee}) > 10^6$.

I.4 Baseline Details

For the five main benchmark comparisons, all baselines, including SD, use one hidden layer of width H_1 (Appendix I.2), with shared benchmark-specific optimizer settings, schedules, early stopping and seeds. Binary baseline heads use single-logit binary cross-entropy; FIRE uses two-logit softmax cross-entropy, equivalent at the logit difference. All TCGA-BRCA heads use five-way softmax cross-entropy. The correction replaced squared error with cross-entropy without enlarging the binary baseline heads.

ERM (Empirical Risk Minimization). A feature encoder and task head without operator structure, $f(x) = g_{\mathrm{ERM}}(h_{\mathrm{ERM}}(x))$.

DANN (Domain-Adversarial Neural Network) (Ganin et al., 2016). Encoder shared with a domain classifier attached via a gradient reversal layer. Domain classifier: two-layer MLP matching the encoder width. Gradient reversal weight $\lambda_{\mathrm{adv}} = 1.0$, annealed from 0 to 1 over the first 10 epochs following Ganin et al. (2016). The main task and the domain classifier share the benchmark's learning rate.

IRM (Invariant Risk Minimization) (Arjovsky et al., 2019). Same encoder and prediction head as ERM. The IRM penalty $\lambda_{\mathrm{IRM}}\|\nabla_{w|w=1.0}\mathcal{L}_e(w \cdot \Phi)\|^2$ is applied per environment, with $\lambda_{\mathrm{IRM}} = 1.0$. Following standard practice, the penalty is annealed in from 0 over the first 10 epochs to allow initial feature learning.

DRO (Group Distributionally Robust Optimization) (Sagawa et al., 2020). Same encoder and prediction head as ERM. Optimizes worst-case loss across environments: $\min_\theta \max_e \mathcal{L}_e(\theta)$. Environment weights are updated via exponentiated gradient ascent with step size $\eta_q = 0.01$, following Sagawa et al. (2020).

CORAL (Deep Correlation Alignment) (Sun & Saenko, 2016). Same encoder and prediction head as ERM. Adds a penalty $\lambda_{\mathrm{CORAL}}\|C_s - C_t\|_F^2$ on the difference between feature covariance matrices of two training environments C_s, C_t, with $\lambda_{\mathrm{CORAL}} = 1.0$. Computed over all environment pairs during training.

Mixup (Inter-Domain Mixup) (Zhang et al., 2018). Same encoder and prediction head as ERM. During training, inputs and labels from different environments are convexly combined: $\tilde{x} = \lambda x_i + (1-\lambda)x_j$, $\tilde{y} = \lambda y_i + (1-\lambda)y_j$, where $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$ with $\alpha = 0.2$. Pairs (i, j) are sampled from different environments.

SD (Spectral Decoupling) (Pezeshki et al., 2021). Same encoder and prediction head as ERM, with the weight-decay term replaced by an ℓ_2 penalty on the network's logits, which decouples feature-learning dynamics and counteracts gradient starvation. Applied with the penalty weight recommended in the original work (Pezeshki et al., 2021).

I.5 Evaluation Metrics

- **Recovery error:** Normalize estimated and target directions. Repeatedly select the unused pair with the largest absolute cosine, then remove its row and column. This gives a greedy assignment π_g, not an optimal assignment. The per-seed score is

$$\mathrm{err} = \max_k \min_{s \in \{-1,+1\}} \|\hat{\mathbf{b}}_k - s\bar{\mathbf{b}}_{\pi_g(k)}\|_2.$$

  Use $\bar{b}_k = A_{e_{\mathrm{ref}}} b_k / \|A_{e_{\mathrm{ref}}} b_k\|$ for conditioning and normalized latent b_k for identity-map runs. Perturbation grids also use the latent target, with the limitations in Appendix I.1.

Stage-2 directions are normalized right factors (Appendix E). Report means and standard deviations across seeds. The score is in $[0, \sqrt{2}]$; random directions give 1.30 ± 0.05 in $d = 20$. For a correctly matched pair, chord distance is $2\sin(\angle/2) \leq \sqrt{2}\sin\angle$ for $\angle \leq \pi/2$. The maximum-angle theorem bounds the worst chord score locally when greedy matching selects the correct correspondence; it does not guarantee an arbitrary greedy assignment.

- **Loss:** Stage-2 strategy and freeze–thaw rows report final training-epoch loss with generating loadings. The regression task loss is MSE; classification uses the binary or multiclass cross-entropy described in Appendix I.4.
- **OOD accuracy:** classification accuracy on subjects from a held-out environment not seen during training.
- **Log-log slope:** slope of $\log(\text{error})$ against $\log(N)$ by least squares. A slope of -0.5 corresponds to the $O(N^{-1/2})$ rate.
- **Reporting:** the five main benchmark accuracies in Table 1 are means $\pm$ one standard deviation over 10 seeds under leave-one-environment-out evaluation. Auxiliary experiments state their seed counts separately. We report the comparison descriptively; paired seed-level results are needed for significance tests.

I.6 Sensitivity to Number of Programs

We evaluate FIRE's accuracy for $K \in \{1, 3, 5, 7, 10\}$ on three representative benchmarks (Table 14), using leave-one-site-out cross-validation with a single seed per configuration and frozen programs. Accuracy is stable across $K \in \{3, 5, 7\}$ on all three (< 1 percentage point of variation: 0.6 on Colored MNIST, 0.9 on Camelyon17, 0.9 on OASIS-1), so the prespecified $K = 3$ used in all main predictive experiments sits on a plateau rather than at a tuned optimum. This scan is post hoc and was not used for model selection; no held-out-environment information entered the choice of K, and the leave-one-environment-out protocol is intact. On OASIS-1 accuracy is highest at $K = 1$. The cohort has $N = 226$ subjects. The predictive scan favors a parsimonious model, but accuracy alone does not identify the latent rank or establish that additional programs contain only estimation noise.

Table 14: OOD accuracy versus the number of programs.

K	1	3	5	7	10
CMNIST	.899	.908	.902	.903	.901
Camelyon17	.856	.848	.853	.857	.859
OASIS-1	.548	.515	.506	.506	.530

I.7 Computational Resources

All experiments were run on a workstation with an NVIDIA RTX 6000 Ada Generation GPU (48 GB VRAM), an AMD Ryzen Threadripper PRO 5975WX 32-core CPU, and 504 GB system RAM. No cloud compute or multi-node training was used.

Peak GPU memory. All experiments fit within 8 GB of GPU memory; the largest models, Camelyon17 with ResNet-18 features, use approximately 4 GB. Feature extraction for Camelyon17 was run once and cached, requiring roughly 2 hours and about 6 GB VRAM.

Total project compute. Including the reported experiments (among them the ADNI fMRI negative result), preliminary runs, hyperparameter exploration, baseline tuning and unreported ABIDE-I pilots, we estimate total project GPU time at 40–60 hours on a single GPU. This total includes exploratory work as well as the experiments reported here.