Conclusion (Multi-environment): With diverse environments and invariant causal mechanisms—idealized sufficient conditions that narrow the hypothesis class—persona vectors corresponding to stable causal factors can be recovered up to invertible transformations. Full uniqueness requires additional assumptions beyond standard IRM formulations. This provides a principled approach to filtering spurious correlations but should be understood as identifying transferable representations rather than guaranteeing unique recovery.

C.5 Proof of Condition: Cross-Layer Consistency

Setup: Assume persona vectors exhibit consistent geometric relationships across layers:

$$v_{\ell+1} = W_\ell v_\ell + \delta_\ell \tag{36}$$

where $W_\ell \in \mathbb{R}^{d\times d}$ is the weight matrix connecting layers and $\|\delta_\ell\|$ is small.

Goal: Show that cross-layer constraints reduce the solution space and improve identifiability.

Step 1: Single-layer null space. From proposition 1, at layer ℓ:

$$\mathcal{N}_\ell = \{v_0 : J_\ell v_0 = 0\} \tag{37}$$

with $\dim(\mathcal{N}_\ell) = d - r_\ell$ where $r_\ell = \text{rank}(J_\ell)$.

Step 2: Cross-layer propagation. If $v_\ell \in \mathcal{N}_\ell$, does $v_{\ell+1} = W_\ell v_\ell \in \mathcal{N}_{\ell+1}$?

Generally, no. The null space changes across layers:

$$v_\ell \in \mathcal{N}_\ell \not\Rightarrow W_\ell v_\ell \in \mathcal{N}_{\ell+1} \tag{38}$$

Step 3: Intersection of constraints. Consider steering vectors observed at multiple layers $\ell_1, \ldots, \ell_L$. Each layer provides a constraint:

$$J_{\ell_i} v_{\ell_i} = y_i \tag{39}$$

If we additionally require consistency:

$$v_{\ell_{i+1}} = W_{\ell_i} v_{\ell_i} + \delta_{\ell_i} \tag{40}$$

Then the solution must satisfy:

$$v_{\ell_i} \in \{v : J_{\ell_i} v = y_i\} \cap \{v : W_{\ell_i} v \approx v_{\ell_{i+1}}\} \tag{41}$$

Step 4: Reduced null space (qualitative characterization). The cross-layer constraints create an overdetermined system. The effective null space is:

$$\mathcal{N}_{\text{eff}} = \bigcap_{i=1}^{L-1} \{v : W_{\ell_i} v \in \mathcal{N}_{\ell_{i+1}}\} \tag{42}$$

Since each W_{ℓ_i} generically has full rank and maps null space vectors to non-null-space vectors, we expect:

$$\dim(\mathcal{N}_{\text{eff}}) \ll \dim(\mathcal{N}_\ell)$$

Intuition: Each additional layer imposes new independent constraints. If the propagation matrices W_{ℓ_i} are sufficiently "generic" (full rank with uncorrelated null space mappings), then each layer reduces the effective null-space dimension. For sufficiently many informative layers with uncorrelated constraint structures, the effective null space can shrink dramatically or even vanish.

Conclusion (Cross-layer): Cross-layer consistency constraints can substantially reduce null-space dimensionality by creating overdetermined systems. The extent of reduction depends on the specific geometric structure of propagation matrices and layer-wise Jacobians. While this approach does not guarantee complete identifiability in general, it provides a practical method for filtering spurious null-space components that lack geometric stability across layers.

D Intuitive Explanations

D.1 Geometric Intuition for Null-Space Ambiguity

Visual analogy: Consider a simplified example where a 3D steering vector $v \in \mathbb{R}^3$ affects 2D outputs $o \in \mathbb{R}^2$ through a projection matrix $J \in \mathbb{R}^{2\times 3}$. By the rank-nullity theorem, the null space $\ker(J)$ is a 1D subspace since $\dim(\ker(J)) = 3 - \text{rank}(J) = 3 - 2 = 1$.

For concreteness, suppose $J = [I_2 \mid 0]$ projects onto the first two coordinates. Then $\ker(J) = \text{span}\{(0,0,1)\}$ is the v_3 axis. The key insight generalizes: for any projection matrix J, directions in $\ker(J)$ are invisible to outputs. Figure 5 illustrates this geometric intuition.

Any vector of the form $v' = v + \alpha v_3$ for $\alpha \in \mathbb{R}$ produces identical outputs:

$$\begin{aligned} J(v + \alpha v_3) &= Jv + \alpha J(0,0,1)^\top \\ &= Jv + \alpha \cdot 0 \\ &= Jv \end{aligned} \tag{43}$$

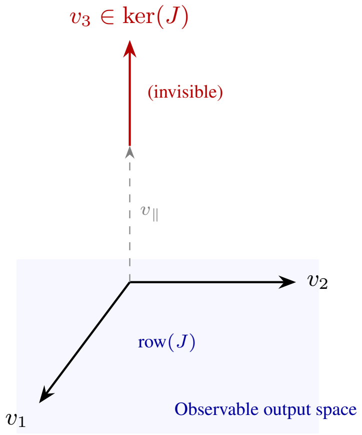


Figure 5: Geometric intuition for null-space ambiguity. The output observes only the (v_1, v_2) components (blue shaded region). The v_3 component lies in $\ker(J)$ and is invisible to the output. Adding any αv_3 to v leaves the output unchanged: $J(v + \alpha v_3) = Jv$ for all $\alpha \in \mathbb{R}$.

Since α can take infinitely many values and $v' \not\propto v$ for generic choices of α, there exist infinitely many geometrically distinct steering vectors that are observationally equivalent.

D.2 Why More Data Doesn't Help

Common misconception: "If we collect more steering examples, we can pin down the unique vector."

Why this fails: Consider observing steering effects on N prompts $\{x_1, \ldots, x_N\}$. Each observation provides:

$$o_i = J_\ell(x_i)v + \eta_i \tag{44}$$

Stacking these:

$$\begin{bmatrix} o_1 \\ o_2 \\ \vdots \\ o_N \end{bmatrix} = \begin{bmatrix} J_\ell(x_1) \\ J_\ell(x_2) \\ \vdots \\ J_\ell(x_N) \end{bmatrix} v + \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_N \end{bmatrix} \tag{45}$$

The stacked Jacobian $J_{\text{stack}} \in \mathbb{R}^{(N \cdot V) \times d}$ has null space:

$$\ker(J_{\text{stack}}) = \bigcap_{i=1}^{N} \ker(J_\ell(x_i)) \tag{46}$$

Critical observation: If all $J_\ell(x_i)$ share a common null space (e.g., all prompts probe similar aspects of the model), then:

$$\ker(J_{\text{stack}}) = \ker(J_\ell(x_1)) \neq \{0\} \tag{47}$$

This means adding more prompts does not reduce the null space *when all prompts probe the same effective subspace*—the ambiguity persists. Figure 6 illustrates this phenomenon.

D.3 Why Orthogonal Perturbations Preserve Semantics

Setup: We test steering with v versus $v + v_\perp$, where $v_\perp \perp v$ and $\|v_\perp\| = 1$. Empirically, adding a random orthogonal direction produces nearly equivalent semantic effects. Why?

Decomposing the perturbation. Any perturbation $v_\perp$ can be decomposed into observable and invisible components:

$$v_\perp = v_{\perp,\text{row}} + v_{\perp,\text{null}} \tag{48}$$

where $v_{\perp,\text{row}} \in \text{row}(J)$ (observable through outputs) and $v_{\perp,\text{null}} \in \ker(J)$ (invisible to outputs).

For a random $v_\perp \perp v$, the expected fraction in the null space is:

$$\mathbb{E}[\|v_{\perp,\text{null}}\|^2] \approx \frac{\dim(\ker(J))}{d} \approx 0.20\text{–}0.25 \tag{49}$$

This means ∼20–25% of the perturbation is automatically invisible to model outputs.

Observable effect is diffuse and unstructured. The actual output change from adding $v_\perp$ is:

$$J(v + v_\perp) = Jv + Jv_{\perp,\text{row}} \tag{50}$$

While $Jv_{\perp,\text{row}}$ is not necessarily small in norm, it represents a *random direction* in the ∼3200-dimensional row space of J. In contrast, Jv is a structured, semantically coherent steering effect (e.g., consistently shifting outputs toward "honesty" or "sycophancy").

The key distinction is not magnitude but *semantic coherence*: Jv produces aligned, directional changes across the vocabulary, while $Jv_{\perp,\text{row}}$ produces diffuse, incoherent perturbations that average out across tokens and do not systematically shift semantic meaning in any consistent direction.

Analogy: If Jv is a strong wind blowing consistently north, then $Jv_{\perp,\text{row}}$ is turbulence—it may have comparable energy but lacks directional coherence, so the overall trajectory remains northward.

D.4 ICA Intuition: Why Independence Helps

Setup: Suppose representations encode two independent factors:

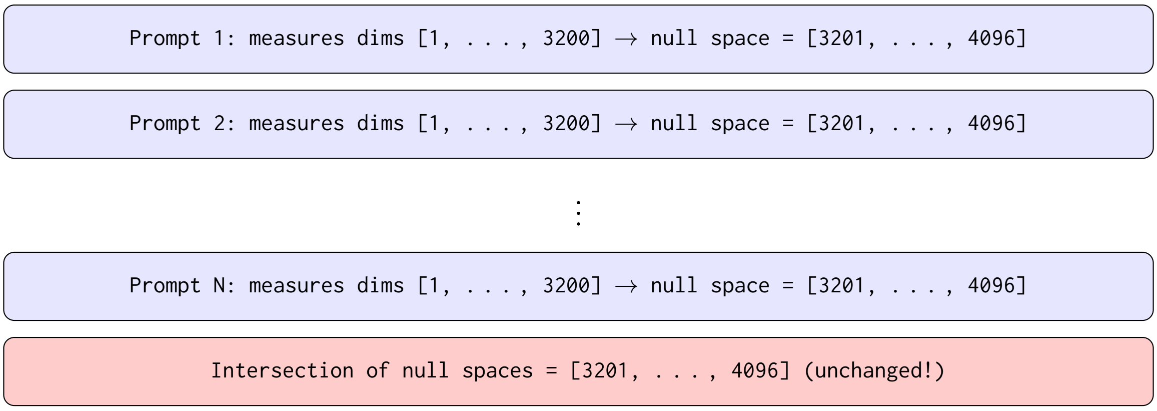

Prompt 1: measures dims [1, . . . , 3200] → null space = [3201, . . . , 4096]
Prompt 2: measures dims [1, . . . , 3200] → null space = [3201, . . . , 4096]
Prompt N: measures dims [1, . . . , 3200] → null space = [3201, . . . , 4096]
Intersection of null spaces = [3201, . . . , 4096] (unchanged!)

Figure 6: Illustration of why prompt diversity does not resolve null-space ambiguity. Each prompt induces a Jacobian that measures the same effective subspace, leaving the null-space intersection unchanged.

- z_1: formality (independent)
- z_2: politeness (independent)

And mix them:

$$h = v_1 z_1 + v_2 z_2 \tag{51}$$

Problem without independence. If z_1 and z_2 are correlated (e.g., formal text tends to be polite), we can reparameterize:

$$\tilde{z}_1 = z_1 + \alpha z_2, \quad \tilde{z}_2 = z_2 \tag{52}$$

$$h = (v_1 - \alpha v_2)\tilde{z}_1 + (v_2 + \alpha v_1)\tilde{z}_2 \tag{53}$$

giving infinitely many equivalent decompositions (different α values).

Solution with independence. If $z_1 \perp z_2$ (statistically independent), then $\tilde{z}_1 = z_1 + \alpha z_2$ is NOT independent of $\tilde{z}_2 = z_2$ for $\alpha \neq 0$:

$$I(\tilde{z}_1; \tilde{z}_2) = I(z_1 + \alpha z_2; z_2) > 0 \tag{54}$$

ICA finds the unique decomposition where sources are independent (refer Figure 7), breaking the symmetry.

D.5 Sparsity Intuition: Occam's Razor

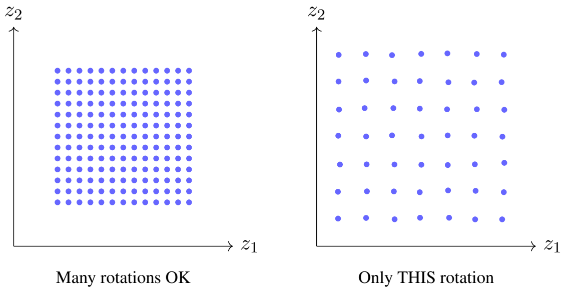
Setup: Suppose we observe:

$$y = Jv + \eta \tag{55}$$

and both v_1 and v_2 fit the data:

$$\|Jv_1 - y\| \approx \|Jv_2 - y\| \approx \epsilon \tag{56}$$

Question: Which is the "true" vector?



(a) Correlated sources (b) Independent sources

Figure 7: Visual comparison of correlated versus independent sources.

Sparsity principle. Prefer the simplest explanation. If:

- v_1 has 100 non-zero entries
- v_2 has 10 non-zero entries

then v_2 is more likely to be the true sparse signal.

Why this works. If the true vector is sparse, then dense solutions must include spurious noise components:

$$v_{\text{dense}} = v_{\text{true}} + v_{\text{noise}}$$

The ℓ_1 penalty penalizes density, filtering out v_{noise}.

D.6 Multi-Environment Intuition: Finding What's Stable

Setup: Observe "formality" steering in 3 environments:

- **E1 (academic):** Formality correlates with technical jargon
- **E2 (business):** Formality correlates with professional tone
- **E3 (legal):** Formality correlates with precise language