is not a realizable operation. However, the reparameterization argument shows that even if we allowed weight modifications during training, identifiability would still fail without structural constraints. Thus, non-identifiability is both a practical limitation (null-space) and a fundamental theoretical barrier (gauge symmetry).

Conclusion (Reparameterization mechanism). The inherent symmetries of neural network parameterizations induce infinitely many observationally equivalent steering vectors.

C Detailed Proof of Proposition 2

C.1 Statement and Overview

Proposition 2. Persona vectors can be identified up to scaling and permutation, thus affording reliable alignment control, under the following sufficient structural conditions:

- **Statistical Independence (ICA).** Latent components are statistically independent, allowing unique recovery via independent component analysis.
- **Sparsity constraints.** Steering directions admit sparse representations that reduce null-space ambiguity.
- **Multi-environment or interventional data.** Variation across environments or interventions breaks observational equivalences.
- **Cross-layer consistency.** Valid semantic directions propagate coherently across layers, filtering spurious components.

We provide detailed proofs for each condition.

C.2 Proof of Condition: Statistical Independence (ICA)

Setup: Assume the latent persona is a vector $z = (z_1, \dots, z_k) \in \mathbb{R}^k$ with independent components and the representation follows:

$$h_\ell = Az + \epsilon \tag{20}$$

where $A \in \mathbb{R}^{d \times k}$ is a mixing matrix and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ is Gaussian noise.

Goal: Show that A (and hence its columns, which are the steering vectors) can be recovered up to permutation and scaling.

Step 1: ICA identifiability theorem. Let $x = As$ where $s \in \mathbb{R}^k$ has independent components and $A \in \mathbb{R}^{d\times k}$ is full column rank. Under the following *sufficient conditions* (Comon, 1994; Hyvärinen and Oja, 2000):

- At most one component of s is Gaussian
- Components are statistically independent
- Sufficient samples are observed

Then A is identifiable up to column permutation and column scaling.

Important caveat: These are strong assumptions that may not hold exactly in practice for LLM personas. Statistical independence between semantic factors (e.g., formality and politeness) is an idealization; real personas may exhibit weak dependencies. The ICA framework provides a sufficient structural condition for identifiability, not a characterization of what typically holds in contemporary steering pipelines.

Step 2: Application to steering. In our setting:

- Observations: $h_\ell(x_1), \dots, h_\ell(x_N)$
- Model: $h_\ell(x_i) = Az(x_i) + \epsilon_i$
- Sources: $z(x_1), \dots, z(x_N)$ are realizations of independent persona components

Assumption verification:

- Independence: We assume $z_1, \dots, z_k$ are statistically independent
- Non-Gaussianity: At most one z_i is Gaussian (e.g., politeness and formality are typically non-Gaussian in natural language)
- Full rank: A has full column rank (each persona has a non-zero effect on representations)

Step 3: Recovery via ICA algorithm. Apply ICA algorithm (Hyvärinen and Oja, 2000) to observations $\{h_\ell(x_i)\}_{i=1}^N$:

- Whitening: Compute $\tilde{h}_\ell = \Sigma^{-1/2}(h_\ell - \mu)$ where $\mu = \mathbb{E}[h_\ell]$ and $\Sigma = \mathrm{Cov}(h_\ell)$.
- Independent component extraction: Find unmixing matrix W that maximizes non-Gaussianity of $\hat{z} = W\tilde{h}_\ell$.
- Recover mixing matrix: $\hat{A} = \Sigma^{1/2}W^{-1}$.

Step 4: Identifiability guarantee. By the ICA theorem, $\hat{A} = APD$ where:

- $P \in \mathbb{R}^{k\times k}$ is a permutation matrix
- $D \in \mathbb{R}^{k\times k}$ is a diagonal scaling matrix

Therefore, the columns of $\hat{A}$ are the columns of A up to permutation and scaling. Since steering vectors are defined as $v_i = Ae_i$ (the i-th column of A), we recover the true steering directions uniquely (up to unavoidable symmetries).

Step 5: Noise robustness. With Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, ICA remains consistent:

$$h_\ell = Az + \epsilon \tag{21}$$

Since ϵ is Gaussian and z is non-Gaussian (by assumption), the ICA algorithm separates signal from noise:

- Recovered sources: $\hat{z} = z + \tilde{\epsilon}$ where $\tilde{\epsilon}$ is small for $\sigma^2 \ll \|Az\|^2$
- Recovered mixing: $\hat{A} \approx APD$ with estimation error decreasing as $N \to \infty$

Conclusion (ICA): Under statistical independence, non-Gaussianity and full column rank—strong sufficient conditions that may not hold automatically in practice—persona vectors are identifiable up to permutation and scaling. These assumptions provide a theoretical pathway to identifiability but require careful experimental design to approximate in real steering settings.

C.3 Proof of Condition: Sparsity Constraints

Setup: Assume the true persona vector $v \in \mathbb{R}^d$ is sparse with $\|v\|_0 = s \ll d$, meaning at most s entries are non-zero. We observe the effect of steering:

$$y = J_\ell v + \eta \tag{22}$$

where $y \in \mathbb{R}^m$ are output measurements, $J_\ell \in \mathbb{R}^{m\times d}$ is the measurement matrix (Jacobian) and $\eta \sim \mathcal{N}(0, \sigma^2 I)$ is noise.

Goal: Show that v can be uniquely recovered via ℓ_1 minimization.

Step 1: Compressed sensing setup. The recovery problem is:

$$\min_{v' \in \mathbb{R}^d} \|v'\|_1 \quad \text{subject to} \quad \|J_\ell v' - y\|_2 \le \epsilon \tag{23}$$

where ϵ bounds the noise: $\epsilon \ge \|\eta\|_2$ with high probability.

Step 2: Restricted Isometry Property (RIP). A matrix J_ℓ satisfies the RIP of order s with constant δ_s if for all s-sparse vectors v:

$$(1-\delta_s)\|v\|_2^2 \le \|J_\ell v\|_2^2 \le (1+\delta_s)\|v\|_2^2 \tag{24}$$

Theorem (Candes and Tao, 2005): If J_ℓ satisfies RIP with $\delta_{2s} < \sqrt{2} - 1 \approx 0.414$, then ℓ_1 minimization recovers v exactly (in the noiseless case) or approximately (with noise) with error:

$$\|v - \hat{v}\|_2 \le C\epsilon \tag{25}$$

for some constant C depending on δ_{2s}.

Important caveat: The RIP condition is a strong assumption. For Jacobians J_ℓ arising from neural network steering, RIP typically does not hold automatically and must be verified empirically or enforced through measurement design (diverse prompt selection). This condition is sufficient for identifiability but may not be satisfied in standard steering settings without careful experimental design.

Step 3: Recovery guarantee. Solve:

$$\hat{v} = \arg\min_{v'} \|v'\|_1 \quad \text{s.t.} \quad \|J_\ell v' - y\|_2 \le \epsilon \tag{26}$$

By the compressed sensing theorem:

$$\|v - \hat{v}\|_2 \le C_1\epsilon + C_2\frac{\|v - v_s\|_1}{\sqrt{s}} \tag{27}$$

where v_s is the best s-sparse approximation to v. If v is exactly s-sparse, the second term vanishes and:

$$\|v - \hat{v}\|_2 \le C_1\epsilon \tag{28}$$

Step 4: Uniqueness. Suppose two sparse vectors v and v' both satisfy $\|J_\ell v - y\|_2 \le \epsilon$. Then:

$$\|J_\ell(v - v')\|_2 \le 2\epsilon \tag{29}$$

If $v - v'$ is $2s$-sparse and J_ℓ satisfies RIP, then by the RIP condition:

$$(1-\delta_{2s})\|v - v'\|_2^2 \le \|J_\ell(v - v')\|_2^2 \le 4\epsilon^2 \tag{30}$$

Therefore:

$$\|v - v'\|_2 \le \frac{2\epsilon}{\sqrt{1-\delta_{2s}}} \tag{31}$$

As $\epsilon \to 0$ (noiseless case), $v = v'$, establishing uniqueness.

Conclusion (Sparsity): Under sparsity assumptions and RIP conditions—strong sufficient conditions that typically require careful measurement design—persona vectors are uniquely recoverable via ℓ_1 minimization. RIP does not hold automatically for arbitrary Jacobians and must be verified or engineered through diverse prompt selection.

C.4 Proof of Condition: Multi-Environment Identification

Setup: Assume we observe the same persona z across multiple environments $e \in \{1, \dots, E\}$ where:

$$h_\ell^{(e)} = g_e(x, z) + \epsilon^{(e)} \tag{32}$$

The function g_e may vary across environments (spurious correlations change) but the causal effect of z on downstream outputs remains invariant.

Goal: Show that the invariant representation z (and its associated direction) can be identified.

Step 1: Invariant causal mechanism. Assume the causal graph:

$$\begin{tikzcd}
z \arrow[r] & h_\ell \arrow[r] & o \\
x \arrow[u] & e \arrow[u] &
\end{tikzcd}$$

where:

- z: true persona (causal factor),
- x: input prompt,
- e: environment,
- h_ℓ: internal representation,
- o: output.

The key assumption is **invariance**: the causal mechanism $h_\ell \to o$ is the same across environments, i.e., $F_{\ell \to L}$ does not depend on e.

Step 2: Invariant Risk Minimization (IRM). The objective is to find representation $\Phi : \mathbb{R}^d \to \mathbb{R}^k$ and predictor $w : \mathbb{R}^k \to \mathbb{R}$ such that:

$$\begin{aligned} \min_{\Phi, w} \quad & \sum_{e=1}^{E} R^e(\Phi, w) \\ \text{s.t.} \quad & w \in \arg\min_{w'} R^e(\Phi, w') \quad \forall e \end{aligned} \tag{33}$$

where,

$$R^e(\Phi, w) = \mathbb{E}_{(x,y)\sim\mathcal{P}^e} \left[\ell\big(w \cdot \Phi(h_\ell(x)), y\big) \right]. \tag{34}$$

Interpretation: Find a representation $\Phi(h_\ell) = z$ such that the optimal predictor w is the same across all environments. This filters out spurious correlations that vary with e.

Step 3: Identifiability under IRM. Under the following idealized conditions:

- Environments are sufficiently diverse: $\mathrm{Cov}(x|e)$ varies across e
- Causal mechanism is invariant: $p(o|z)$ is the same for all e
- Sufficient environments: $E \geq k + 1$ where $k = \dim(z)$
- No unobserved confounders between environments and outcomes

The theorem (Ahuja et al., 2022; Von Kügelgen et al., 2021) states that the invariant risk minimization objective can recover z up to an invertible transformation, though not necessarily uniquely.

Important caveat: The literature on invariant representation learning establishes conditions under which invariance constraints narrow the hypothesis class but full identifiability (unique recovery of z) requires additional assumptions beyond standard IRM formulations. The practical application of IRM to steering should be understood as identifying a stable, transferable representation rather than guaranteeing uniqueness. This is a sufficient condition in principle under strong assumptions, not a characterization of typical steering scenarios.

Step 4: Application to steering. In practice:

1. Collect multi-environment data: Extract steering vectors from diverse prompt distributions, model checkpoints or instruction formats.
2. Learn invariant representation: Train Φ via IRM objective
3. Extract steering vectors: Compute $v_i = \nabla_{h_\ell} \Phi_i(h_\ell)$ where Φ_i is the i-th component of Φ

The resulting steering vectors v_i capture invariant causal factors rather than spurious correlations.

Step 5: Theoretical guarantee. Under mild conditions (sufficient diversity, invariance, identifiable causal structure), the IRM solution satisfies:

$$\Phi(h_\ell) = T(z) \tag{35}$$

where T is an invertible transformation. Since persona steering operates on z, the directions $v_i = \nabla_{h_\ell} \Phi_i$ recover the true causal factors.