its principal component. By rotating activations to target angles θ, it provides continuous control and generalizes both addition ($\theta < 180$) and ablation ($\theta = 90$). **Adaptive Angular Steering (AAS)** (Vu and Nguyen, 2025) adds conditional masking, applying rotation only to activations aligned with the feature direction: mask $= \max(0, \text{sign}(h \cdot d_{\text{feat}}))$. However, both methods apply steering uniformly across all layers, causing generation collapse on smaller models and poor controllability on strongly aligned models. Our analysis reveals this stems from ignoring layer-wise discriminability - early layers lack meaningful feature separation while steering them disrupts unrelated representations.

A.3 Layer-Specific Interventions

Recent work recognizes layers play heterogeneous roles. **Circuit analysis** (Wang et al., 2023; Marks et al., 2025) identifies specific attention heads and MLP neurons responsible for behaviors, enabling surgical interventions. **Mechanistic interpretability** (Elhage et al., 2021; Nanda et al., 2023) studies information flow through layer-wise transformations, revealing that features emerge progressively across depth. However, these approaches focus on understanding rather than control. Concurrent work on **layer-wise steering** (Harrasse et al., 2025) observes varying steering effectiveness across layers but lacks principled selection criteria. Our discriminative criterion $\mu_{\text{pos}}^{(k)} \cdot \mu_{\text{neg}}^{(k)} < 0$ provides a theoretically grounded, automatically computable condition for identifying steerable layers.

A.4 Comparison with Prior Methods

Table 3 contrasts Selective Steering with prior angular methods. Unlike Angular and Adaptive Angular Steering, which violate norm preservation during plane projection (Proposition 1), SS guarantees norm preservation through discriminative layer selection (Proposition 2). Our opposition-based criterion identifies layers where classes exhibit opposite-signed projections, concentrating steering effort where features naturally separate. This reduces computational overhead from $O(Ld_{\text{model}})$ to $O(|\mathcal{L}_{\text{disc}}|d_{\text{model}})$ where $|\mathcal{L}_{\text{disc}}| \ll L$, as only discriminative layers require rotation matrices.

Our method is the first to combine continuous angular control with principled layer selection, achieving robust steering without coherence degradation.

B Detailed Methodology

B.1 Proof: Norm Violation in Angular Steering

Proof of Proposition 1. We demonstrate a counterexample at the identity case $\theta = 0$, where intuitively no transformation should occur. For $\theta = 0$, the rotation matrix is:

$$\mathbf{R}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{thus} \quad \mathbf{R}_0 \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{11}$$

Substituting $\theta = 0$ into Equation 2:

$$\begin{aligned} \mathbf{h}_{\text{steered},0}^{\text{AS}} &= \mathbf{h} - \text{proj}_P(\mathbf{h}) + \|\text{proj}_P(\mathbf{h})\| \cdot [\mathbf{b}_1\ \mathbf{b}_2] \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= \mathbf{h} - \text{proj}_P(\mathbf{h}) + \|\text{proj}_P(\mathbf{h})\| \cdot \mathbf{b}_1. \end{aligned} \tag{12}$$

For $\mathbf{h}_{\text{steered},0}^{\text{AS}} = \mathbf{h}$ (identity), we require:

$$-\text{proj}_P(\mathbf{h}) + \|\text{proj}_P(\mathbf{h})\| \cdot \mathbf{b}_1 = \mathbf{0}. \tag{13}$$

Let $\text{proj}_P(\mathbf{h}) = c_1\mathbf{b}_1 + c_2\mathbf{b}_2$ where $c_1 = \mathbf{b}_1^\top \mathbf{h}$ and $c_2 = \mathbf{b}_2^\top \mathbf{h}$. Then:

$$\|\text{proj}_P(\mathbf{h})\| = \sqrt{c_1^2 + c_2^2}. \tag{14}$$

Substituting into Equation 13:

$$-(c_1\mathbf{b}_1 + c_2\mathbf{b}_2) + \sqrt{c_1^2 + c_2^2} \cdot \mathbf{b}_1 = \mathbf{0}. \tag{15}$$

Rearranging:

$$\left(\sqrt{c_1^2 + c_2^2} - c_1\right)\mathbf{b}_1 - c_2\mathbf{b}_2 = \mathbf{0}. \tag{16}$$

Since $\{\mathbf{b}_1, \mathbf{b}_2\}$ are orthonormal, both coefficients must vanish:

$$\sqrt{c_1^2 + c_2^2} - c_1 = 0 \quad \text{and} \quad c_2 = 0. \tag{17}$$

Combined with $c_2 = 0$, the first condition simplifies to $|c_1| = c_1$, requiring $c_1 \geq 0$.

Thus, $\mathbf{h}_{\text{steered},0}^{\text{AS}} = \mathbf{h}$ holds only when $\mathbf{h}$'s projection lies exactly along $\mathbf{b}_1$ with non-negative coefficient ($c_2 = 0$ and $c_1 \geq 0$). For general $\mathbf{h}$ where $c_2 \neq 0$ or $c_1 < 0$:

$$\mathbf{h}_{\text{steered},0}^{\text{AS}} \neq \mathbf{h} \quad \Rightarrow \quad \|\mathbf{h}_{\text{steered},0}^{\text{AS}}\| \neq \|\mathbf{h}\|. \tag{18}$$

This demonstrates fundamental norm violation even at the identity transformation. □

Table 3: Comparison of steering methods on key properties. ✓ indicates satisfaction, ✗ indicates violation.

Property	**ActAdd**	**DirAbl**	**SAS**	**AAS**	**SS (Ours)**
Norm preservation	✗	✗	✗	✗	✓
Layer selectivity	✗	✗	✗	✗	✓
Continuous control	✗	✗	✓	✓	✓
Fine-grained modulation	✓	✗	✓	✓	✓
Discriminability criterion	None	None	None	Alignment	Opposition
Hyperparameter sensitivity	High	Low	Low	Low	Low
Computational cost	$O(Ld_{\text{model}})$	$O(Ld_{\text{model}})$	$O(Ld_{\text{model}})$	$O(Ld_{\text{model}})$	$O(\lvert\mathcal{L}_{\text{disc}}\rvert d_{\text{model}})$

B.2 Proof: Norm Preservation in Selective Steering

Proof of Proposition 2. The rotation matrix decomposes as:

$$\mathbf{R}_\theta^P = \underbrace{[\mathbf{I} - (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top)]}_{\text{projection onto } Q} + \underbrace{[\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[\mathbf{b}_1\ \mathbf{b}_2]^\top}_{\text{rotation in plane } P}, \tag{19}$$

where Q is the orthogonal complement of $P = \text{span}\{\mathbf{b}_1, \mathbf{b}_2\}$.
Decompose $\mathbf{h} = \mathbf{h}_P + \mathbf{h}_Q$ where:

$$\mathbf{h}_P = (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top)\mathbf{h} = c_1\mathbf{b}_1 + c_2\mathbf{b}_2, \tag{20}$$

$$\mathbf{h}_Q = [\mathbf{I} - (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top)]\mathbf{h}. \tag{21}$$

Applying $\mathbf{R}_\theta^P$:

$$\begin{aligned}\mathbf{R}_\theta^P\mathbf{h} &= [\mathbf{I} - (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top)](\mathbf{h}_P + \mathbf{h}_Q) &(22)\\ &\quad + [\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[\mathbf{b}_1\ \mathbf{b}_2]^\top(\mathbf{h}_P + \mathbf{h}_Q)\\ &= \mathbf{h}_Q + [\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[c_1\ c_2]^\top, &(23)\end{aligned}$$

since projection annihilates $\mathbf{h}_P$, preserves $\mathbf{h}_Q$, and $[\mathbf{b}_1\ \mathbf{b}_2]^\top\mathbf{h}_Q = \mathbf{0}$.
The 2D rotation matrix $\mathbf{R}_\theta$ is orthogonal: $\mathbf{R}_\theta^\top\mathbf{R}_\theta = \mathbf{I}_2$. Therefore:

$$\begin{aligned}\|\mathbf{R}_\theta^P\mathbf{h}\|^2 &= \|\mathbf{h}_Q\|^2 + \|[\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[c_1\ c_2]^\top\|^2\\ &= \|\mathbf{h}_Q\|^2 + \|\mathbf{R}_\theta\,[c_1\ c_2]^\top\|^2 &(24)\\ &\quad (\{\mathbf{b}_1, \mathbf{b}_2\} \text{ orthonormal})\\ &= \|\mathbf{h}_Q\|^2 + \|[c_1\ c_2]^\top\|^2 &(25)\\ &\quad (\mathbf{R}_\theta \text{ preserves norms})\\ &= \|\mathbf{h}_Q\|^2 + c_1^2 + c_2^2 &(26)\\ &= \|\mathbf{h}_Q\|^2 + \|\mathbf{h}_P\|^2 &(27)\\ &= \|\mathbf{h}\|^2, &(28)\end{aligned}$$

where the last equality follows from orthogonality of P and Q. Thus $\|\mathbf{R}_\theta^P\mathbf{h}\| = \|\mathbf{h}\|$. □

B.3 Calibration Procedure

Step 1: Activation Extraction. Pass all prompts in $\mathcal{D}_{\text{pos}}^{(\text{train})}$ and $\mathcal{D}_{\text{neg}}^{(\text{train})}$ through the model. At each layer $k \in \{1, \ldots, L\}$ (specifically, after normalization before attention and MLP blocks), record the final token's activation vector $\mathbf{h}_p^{(k)}$ for each prompt p.

Step 2: Mean Vector Computation. For each layer k:

$$\boldsymbol{\mu}_{\text{pos}}^{(k)} = \frac{1}{|\mathcal{D}_{\text{pos}}^{(\text{train})}|}\sum_{p\in\mathcal{D}_{\text{pos}}^{(\text{train})}}\mathbf{h}_p^{(k)}, \tag{29}$$

$$\boldsymbol{\mu}_{\text{neg}}^{(k)} = \frac{1}{|\mathcal{D}_{\text{neg}}^{(\text{train})}|}\sum_{p\in\mathcal{D}_{\text{neg}}^{(\text{train})}}\mathbf{h}_p^{(k)}. \tag{30}$$

Step 3: Global Feature Direction Selection. Compute candidate directions at each layer using difference-in-means:

$$\mathbf{d}^{(k)} = \boldsymbol{\mu}_{\text{pos}}^{(k)} - \boldsymbol{\mu}_{\text{neg}}^{(k)}, \quad k = 1, \ldots, L. \tag{31}$$

Select the global feature direction as the candidate with maximum average cosine similarity to others:

$$k^* = \text{argmax}_k \frac{1}{L}\sum_{j=1}^{L}\frac{\mathbf{d}^{(k)}\cdot\mathbf{d}^{(j)}}{\|\mathbf{d}^{(k)}\|\|\mathbf{d}^{(j)}\|}, \quad \hat{\mathbf{d}}_{\text{feat}} = \frac{\mathbf{d}^{(k^*)}}{\|\mathbf{d}^{(k^*)}\|}. \tag{32}$$

This selects the direction most consistently represented across model depth.

Step 4: Discriminative Layer Identification. Project class means at each layer onto the global feature direction:

$$\tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)} = \boldsymbol{\mu}_{\text{pos}}^{(k)}\cdot\hat{\mathbf{d}}_{\text{feat}}, \quad \tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)} = \boldsymbol{\mu}_{\text{neg}}^{(k)}\cdot\hat{\mathbf{d}}_{\text{feat}}. \tag{33}$$

Identify discriminative layers as those with opposite-signed projections:

$$\mathcal{L}_{\text{disc}} = \left\{k : \tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)}\cdot\tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)} < 0\right\}. \tag{34}$$

Step 5: Steering Plane Construction. Stack candidate directions into matrix $\mathbf{D} = [\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(L)}]^\top$ and perform PCA. Extract the first principal component $\mathbf{d}_{\text{PC1}}$. Construct orthonormal basis via Gram-Schmidt:

$$\mathbf{b}_1 = \hat{\mathbf{d}}_{\text{feat}}, \tag{35}$$

$$\mathbf{b}_2 = \mathbf{d}_{\text{PC1}} - (\mathbf{d}_{\text{PC1}} \cdot \mathbf{b}_1)\mathbf{b}_1, \quad \mathbf{b}_2 \leftarrow \frac{\mathbf{b}_2}{\|\mathbf{b}_2\|}. \tag{36}$$

Store the following for inference: orthonormal basis $\{\mathbf{b}_1, \mathbf{b}_2\}$ and discriminative layer set $\mathcal{L}_{\text{disc}}$ for runtime checking.

B.4 Theoretical Analysis: Discriminability Criterion

Geometric Interpretation. The dot product criterion $\tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)} \cdot \tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)} < 0$ identifies layers where class means point in opposing directions. The squared distance between means:

$$\left\|\tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)} - \tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)}\right\|^2 = \left\|\tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)}\right\|^2 + \left\|\tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)}\right\|^2 - 2\tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)} \cdot \tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)}. \tag{37}$$

When the dot product is negative, the $-2\tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)} \cdot \tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)}$ term contributes positively, increasing separation beyond what orthogonal means would provide:

$$\left\|\tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)} - \tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)}\right\|^2 > \left\|\tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)}\right\|^2 + \left\|\tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)}\right\|^2 - 2\left\|\tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)}\right\| \cdot \left\|\tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)}\right\|. \tag{38}$$

Monotonicity of Steering Effect. Rotating activations toward angle θ monotonically increases alignment with $\mathbf{b}_1 \approx \mathbf{d}_{\text{feat}}$. For discriminative layers where $\tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)} \cdot \tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)} < 0$, this rotation consistently moves activations toward the positive class mean, providing predictable control.

C Detailed Evaluation Metrics

Coherence Metrics. We employ four complementary metrics to assess generation quality:

(1) Perplexity (PPL): Measures the model's uncertainty in generating text. For a sequence of tokens $\mathbf{x} = (x_1, \ldots, x_T)$, perplexity is computed as:

$$\text{PPL}(\mathbf{x}) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T} \log p(x_t \mid x_{<t})\right) \tag{39}$$

where $p(x_t \mid x_{<t})$ is the model's predicted probability of token x_t given previous tokens. Lower perplexity indicates more confident, fluent generation.

(2) N-gram Repetition (N-gram Rep.): Detects pathological repetition by measuring n-gram diversity. For a generated sequence with n-grams $\mathcal{N}$:

$$\text{Rep-n} = \frac{|\mathcal{N}| - |\text{unique}(\mathcal{N})|}{|\mathcal{N}|} \tag{40}$$

where $|\mathcal{N}|$ is the total count of n-grams and $|\text{unique}(\mathcal{N})|$ is the count of unique n-grams. We use $n = 4$ (4-grams). Values range from 0 (no repetition) to 1 (complete repetition). Lower is better.

(3) Language Consistency (Lang. Cons.): Detects foreign character contamination in English responses using Unicode script analysis:

$$\text{LC} = \frac{\text{\# Latin/Common characters}}{\text{\# total characters}} \tag{41}$$

We count characters from Latin, Common (punctuation, digits), and allowed scripts, excluding CJK, Arabic, Cyrillic, and other non-Latin scripts. Values range from 0 (completely foreign) to 1 (fully consistent). Higher is better.

(4) Compression Ratio (Comp. Ratio): Pattern-agnostic collapse detection using gzip compression:

$$\text{CR} = \frac{\text{compressed\_size}(\mathbf{x})}{\text{original\_size}(\mathbf{x})} \tag{42}$$

Highly repetitive or patterned text compresses well (low ratio), while diverse natural text compresses poorly (high ratio). Higher is better.

Controllability Metrics. We measure steering effectiveness using multiple attack success evaluators:

(1) Attack Success Rate (ASR): Measures the proportion of harmful prompts that successfully elicit harmful responses. For evaluation set $\mathcal{D}_{\text{eval}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ where $\mathbf{x}_i$ are harmful prompts and $\mathbf{y}_i$ are model responses:

$$\text{ASR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\text{IsHarmful}(\mathbf{y}_i)] \tag{43}$$

where $\text{IsHarmful}(\cdot)$ is a binary classifier. We use three classifiers: HarmBench (Mazeika et al., 2024), PolyGuard (Kumar et al., 2025), and LLM-as-judge with Qwen2.5-14B-Instruct (Team,