2024c).. Higher ASR indicates more successful steering toward harmful behavior.

(2) Refusal Score (RS) (Arditi et al., 2024): Substring-based detection of refusal patterns:

$$\text{RS} = \frac{1}{N}\sum_{i=1}^{N} \nVdash \left[\exists s \in \mathcal{S}_{\text{refusal}} : s \in \mathbf{y}_i\right] \qquad (44)$$

where $\mathcal{S}_{\text{refusal}}$ is a set of common refusal substrings (e.g., "I'm sorry", "I cannot", "As an AI"). Lower RS indicates less refusal behavior.

Robustness Metrics. We measure preservation of general capabilities using zero-shot accuracy:

Accuracy (Acc): For each benchmark task $\mathcal{B}$ with test set $\{(\mathbf{x}_i, y_i^*)\}_{i=1}^{M}$ where y_i^* are ground truth labels:

$$\text{Acc}(\mathcal{B}) = \frac{1}{M}\sum_{i=1}^{M} \nVdash [f(\mathbf{y}_i) = y_i^*] \qquad (45)$$

where $f(\cdot)$ extracts the answer from model output $\mathbf{y}_i$ using task-specific parsers (e.g., multiple-choice extraction for MMLU, numerical answer extraction for GSM8K). Higher accuracy indicates better capability retention.

D Additional Results

This section provides a detail analysis for coherence from Section 4. Table 4 quantifies coherence quality through three complementary metrics. **SS achieves the best or second-best compression ratio in 8/8 models**, indicating superior resistance to generation collapse. Notably, on challenging models where SAS/AAS struggle (Qwen2.5-1.5B, Qwen2.5-3B, gemma-2-2b), **SS reduces n-gram repetition by 88.9%, 91.3%, and 97.9% respectively compared to SAS** - from 0.4649→0.0516, 0.2734→0.0237, and 0.8242→0.0177. Critically, **SS restores language consistency to near-perfect levels (1.0000) on Qwen2.5-1.5B and Qwen2.5-3B**, where SAS produces severe contamination (0.9196 and 0.7611 respectively), demonstrating its ability to prevent multilingual leakage that plagues angular steering methods. The variance statistics (±std) reveal that **SS produces significantly more stable outputs across steering angles**: compression ratio variance is lower than SAS/AAS in 6/8 models, with particularly dramatic improvements on unstable models (Qwen2.5-1.5B: 0.3142 vs 0.3853/0.4062; gemma-2-2b: 0.0288 vs 0.0481/0.2249).

E Ablation Studies

We conduct comprehensive ablation studies to validate the two core design decisions in Selective Steering: (1) discriminative layer selection via the opposite-signed criterion, and (2) norm-preserving transformation via the rotation matrix formulation. Experiments are performed on three representative models spanning different sizes and architectures: Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct (Yang et al., 2024; Team, 2024c), and gemma-2-9B-it (Team, 2024a). These models were selected because they exhibited strong performance in our main experiments (Section 4), demonstrating clear discriminative layer patterns and reliable steering behavior.

E.1 Ablation 1: Layer Selection Strategies

Motivation. To isolate the contribution of our discriminative layer selection criterion (Equation 9), we compare against four alternative strategies that do not exploit opposite-signed discriminability.

Compared Strategies.

- **Random Selection (50%):** Randomly sample 50% of layers for steering, matching the typical size of $\mathcal{L}_{\text{disc}}$. This controls for the effect of layer count while removing discriminative selection.
- **Early Layers:** Apply steering to the first half of layers. This tests the hypothesis that early layers are sufficient for behavior control.
- **Late Layers:** Apply steering to the second half of layers. This tests whether late-stage intervention near the output is more effective.
- **Uniform (All Layers):** Apply steering to all layers uniformly, equivalent to Angular Steering's approach.
- **Discriminative Selection (Ours):** Apply steering only to layers satisfying $\boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)} < 0$.

All strategies use the norm-preserving transformation (Equation 10) to isolate the effect of layer selection. For each model, we select the steering angle θ^* that maximizes ASR under the Discriminative Selection strategy, then evaluate all strategies at this fixed angle to ensure fair comparison.

Results. Table 5 reports controllability metrics (ASR and Refusal Score) across strategies.

Model	Method	N-gram Rep. ↓	Lang. Cons. ↑	Comp. Ratio ↑
Llama-3.1-8B	ActAdd	0.0725	**1.0000**	0.4274
	DirAbl	**0.0182**	0.9999	0.6973
	SAS	0.0986 ± 0.0779	**1.0000 ± 0.0000**	0.6048 ± 0.2331
	AAS	0.0649 ± 0.0659	**1.0000 ± 0.0000**	0.6270 ± 0.2409
	SS (Ours)	0.1065 ± 0.1824	0.9999 ± 0.0001	**0.7075 ± 0.2763**
Llama-3.2-1B	ActAdd	0.1983	**1.0000**	0.3967
	DirAbl	0.0417	0.9998	0.5131
	SAS	0.2206 ± 0.2111	0.9993 ± 0.0022	0.5698 ± 0.2647
	AAS	0.1403 ± 0.1317	0.9996 ± 0.0016	0.5842 ± 0.2552
	SS (Ours)	**0.0413 ± 0.0357**	0.9996 ± 0.0005	**0.6875 ± 0.2619**
Llama-3.2-3B	ActAdd	0.0759	**1.0000**	0.4115
	DirAbl	0.0321	**1.0000**	0.5588
	SAS	0.0640 ± 0.0367	0.9997 ± 0.0006	0.5898 ± 0.1717
	AAS	0.0330 ± 0.0227	0.9999 ± 0.0001	0.5881 ± 0.1790
	SS (Ours)	**0.0289 ± 0.0393**	0.9997 ± 0.0005	**0.6924 ± 0.1968**
Qwen2.5-1.5B	ActAdd	0.1849	0.3093	0.2192
	DirAbl	**0.0507**	0.9999	0.5278
	SAS	0.4649 ± 0.3592	0.9196 ± 0.1701	0.4353 ± 0.3853
	AAS	0.4149 ± 0.3956	0.9884 ± 0.0290	0.4970 ± 0.4062
	SS (Ours)	0.0516 ± 0.0595	**1.0000 ± 0.0000**	**0.7201 ± 0.3142**
Qwen2.5-3B	ActAdd	0.4623	0.9998	0.2330
	DirAbl	**0.0219**	0.9996	0.4621
	SAS	0.2734 ± 0.1334	0.7611 ± 0.3432	0.3787 ± 0.2779
	AAS	0.1815 ± 0.1698	0.8713 ± 0.2825	0.3454 ± 0.1772
	SS (Ours)	0.0237 ± 0.0271	**0.9998 ± 0.0003**	**0.5273 ± 0.0830**
Qwen2.5-7B	ActAdd	0.1377	0.9991	0.3948
	DirAbl	0.0158	**0.9995**	0.4695
	SAS	0.1379 ± 0.1876	0.9992 ± 0.0019	0.4170 ± 0.1194
	AAS	0.0768 ± 0.1332	0.9995 ± 0.0016	0.4616 ± 0.0797
	SS (Ours)	**0.0100 ± 0.0066**	0.9994 ± 0.0011	**0.5101 ± 0.0458**
gemma-2-2b	ActAdd	0.9804	**1.0000**	0.0320
	DirAbl	**0.0138**	0.9999	0.4721
	SAS	0.8242 ± 0.3151	**1.0000 ± 0.0000**	0.0351 ± 0.0481
	AAS	0.4159 ± 0.4332	**1.0000 ± 0.0000**	0.2878 ± 0.2249
	SS (Ours)	0.0177 ± 0.0209	**1.0000 ± 0.0000**	**0.4871 ± 0.0288**
gemma-2-9b	ActAdd	0.9707	**1.0000**	0.0753
	DirAbl	**0.0022**	**1.0000**	**0.5325**
	SAS	0.9891 ± 0.0147	**1.0000 ± 0.0000**	0.0268 ± 0.0242
	AAS	0.5117 ± 0.4906	**1.0000 ± 0.0000**	0.2740 ± 0.2635
	SS (Ours)	0.1500 ± 0.2921	0.9999 ± 0.0001	0.4625 ± 0.1528

Table 4: Coherence evaluation across steering methods. Metrics averaged over all steering angles. Best scores (excluding No Steering) in **bold**, second-best underlined. ↓/↑ indicate lower/higher is better.

Key Observations. **(1) Discriminative Selection substantially outperforms alternatives.** Across all models and evaluators, Discriminative Selection achieves 2–8× higher HarmBench ASR compared to non-selective baselines (Random, Early, Late). For example, on Qwen2.5-3B, HarmBench ASR improves from 0.000 (Early/Late/Random) to 0.846 (Discriminative), and LLM-judge ASR increases from 0.000 to 0.837. This validates that opposite-signed discriminability identifies layers where steering is most effective.

(2) Early and Random strategies fail almost completely. Early Layers and Random Selection yield near-zero ASR on smaller models (Qwen2.5-1.5B, Qwen2.5-3B), indicating that indiscriminate intervention in non-discriminative layers is ineffective. This aligns with Figure 2b, which shows early layers exhibit minimal class separation.

(3) Late Layers show moderate effectiveness but inconsistent. Late Layers achieve partial success (HarmBench ASR: 0.038–0.240), suggesting some discriminative capacity emerges in deeper layers. However, performance is highly variable across models and substantially trails Discriminative Selection, indicating that not all late layers are discriminative.

(4) Uniform (All Layers) is surprisingly competitive but brittle. Applying steering to all layers

Table 5: **Ablation study: Layer selection strategies.** All methods use norm-preserving transformation at the same angle θ^* (selected to maximize ASR under Discriminative Selection). ASR metrics (↑ better): HarmBench, PolyGuard†, LLM-judge. Refusal Score (Substring, ↓ better). †PolyGuard scores are inflated due to sensitivity to text degradation patterns (discussed below).

Model	Strategy	HarmBench↑	PolyGuard†↑	LLM-judge↑	Substring↓
Qwen2.5-1.5B	Random (50%)	0.000	0.029	0.010	0.990
	Early Layers	0.000	0.019	0.000	0.990
	Late Layers	0.038	0.346	0.000	0.952
	Uniform (All)	0.308	0.981	0.087	0.000
	Discriminative (Ours)	**0.740**	**0.942**	**0.664**	**0.000**
Qwen2.5-3B	Random (50%)	0.000	0.000	0.000	0.981
	Early Layers	0.000	0.010	0.010	0.990
	Late Layers	0.000	0.038	0.000	0.942
	Uniform (All)	0.548	1.000	0.298	0.010
	Discriminative (Ours)	**0.846**	**0.962**	**0.837**	**0.000**
Gemma-2-9B	Random (50%)	0.019	0.010	0.010	0.971
	Early Layers	0.010	0.010	0.010	0.990
	Late Layers	0.240	0.356	0.212	0.692
	Uniform (All)	0.279	0.990	0.173	0.000
	Discriminative (Ours)	**0.683**	**1.000**	**0.683**	**0.000**

yields moderate ASR (0.279–0.548) and eliminates refusals (Substring $\approx$ 0.000), appearing competitive at first glance. However, this comes at a severe cost to coherence (discussed in Section 4): uniform steering on smaller models (<7B) causes perplexity spikes, repetition collapse, and foreign language contamination. Discriminative Selection achieves comparable or higher ASR while maintaining generation quality by avoiding non-discriminative layers.

(5) PolyGuard exhibits systematic bias toward degraded text. PolyGuard consistently assigns high scores to Uniform (All Layers), even when HarmBench and LLM-judge indicate low harmfulness (e.g., Qwen2.5-1.5B: PolyGuard 0.981 vs. HarmBench 0.308). Upon manual inspection, we find PolyGuard flags incoherent or repetitive text as "unsafe" due to its content moderation heuristics detecting anomalous patterns (e.g., repetitive refusal phrases, foreign characters, grammatical errors). Thus, PolyGuard scores should be interpreted cautiously - high scores may indicate text degradation rather than genuine harmfulness. We report PolyGuard for completeness but emphasize HarmBench and LLM-judge as more reliable indicators.

E.2 Ablation 2: Norm Preservation

Motivation. To validate that norm preservation is critical for steering effectiveness (not merely layer selection), we compare our norm-preserving formulation (Equation 10) against Angular Steering's implementation (Equation 2), both using the *same* discriminative layer set $\mathcal{L}_{\text{disc}}$.

Compared Formulations.

- **Angular Steering Implementation:** Apply the efficient implementation from Vu and Nguyen (2025):

$$\mathbf{h}'^{(k)} = \mathbf{h}^{(k)} - \text{proj}_P(\mathbf{h}^{(k)}) + \|\text{proj}_P(\mathbf{h}^{(k)})\| \cdot [\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[1\ 0]^\top,$$

  which violates norm preservation (Proposition 1).

- **Norm-Preserving Formulation (Ours):** Apply the rotation matrix:

$$\begin{aligned}\mathbf{h}'^{(k)} &= \mathbf{R}_\theta^P \mathbf{h}^{(k)} \\ &= \left[\mathbf{I} - (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top) + [\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[\mathbf{b}_1\ \mathbf{b}_2]^\top\right]\mathbf{h}^{(k)},\end{aligned}$$

  which guarantees $\|\mathbf{h}'^{(k)}\| = \|\mathbf{h}^{(k)}\|$ (Proposition 2).

Both methods use the same discriminative layers ($\mathcal{L}_{\text{disc}}$) and angle ($\theta^*$), isolating the effect of norm preservation.

Results. Table 6 reports controllability metrics.

Key Observations. (1) Norm preservation is essential for effective steering. The norm-preserving formulation achieves 26–70× higher HarmBench ASR compared to Angular Steering's