Table 6: **Ablation study: Norm preservation.** Both methods use the same discriminative layers ($\mathcal{L}_{\text{disc}}$) and angle ($\theta^*$). ASR metrics (↑ better): HarmBench, PolyGuard†, LLM-judge. Refusal Score (Substring, ↓ better). †PolyGuard scores are inflated for the Angular Steering implementation due to text degradation patterns.

Model	Formulation	HarmBench↑	PolyGuard†↑	LLM-judge↑	Substring↓
Qwen2.5-1.5B	Angular Steering	0.029	0.077	0.010	0.981
	Norm-Preserving (Ours)	**0.740**	**0.942**	**0.664**	**0.000**
Qwen2.5-3B	Angular Steering	0.000	0.000	0.000	0.981
	Norm-Preserving (Ours)	**0.846**	**0.962**	**0.837**	**0.000**
Gemma-2-9B	Angular Steering	0.019	0.010	0.019	0.971
	Norm-Preserving (Ours)	**0.683**	**1.000**	**0.683**	**0.000**

implementation, despite using identical layer selection. On Qwen2.5-3B, HarmBench ASR increases from 0.000 to 0.846, and LLM-judge ASR from 0.000 to 0.837. This dramatic improvement validates our theoretical analysis (Propositions 1 and 2): norm violations disrupt activation distributions, rendering steering ineffective.

(2) Angular Steering implementation fails even with optimal layer selection. Even when restricted to discriminative layers ($\mathcal{L}_{\text{disc}}$), Angular Steering's implementation yields near-zero ASR and maintains high refusal rates (Substring ≈ 0.98). This demonstrates that the norm violation issue (Section 3) is not merely a side effect of uniform layer application - it is an *inherent flaw* in the transformation itself. Layer selection alone is insufficient; norm preservation is critical.

(3) The gap is most pronounced on smaller models. Qwen2.5-1.5B and Qwen2.5-3B show near-complete failure (HarmBench ASR < 0.03) under Angular Steering, while achieving strong success (0.740, 0.846) with norm preservation. This aligns with our hypothesis that smaller models are more sensitive to distribution shift: limited capacity leaves less margin for absorbing norm violations, causing rapid coherence collapse that precludes effective steering.

(4) Refusal behavior reflects steering effectiveness. Refusal scores (Substring) track inversely with ASR: norm-preserving formulation achieves near-zero refusals (0.000) while Angular Steering maintains high refusals (0.971–0.981). This indicates that norm violations not only degrade coherence but also prevent meaningful behavior modification - the model continues refusing despite intervention.

E.3 Summary

These ablation studies conclusively demonstrate that both design choices are essential:

- **Discriminative layer selection** (Equation 9) identifies where to steer, concentrating intervention on layers with strong opposite-signed class separation. Without this, steering is ineffective (Early/Random strategies) or damages coherence (Uniform strategy).

- **Norm-preserving transformation** (Equation 10) determines how to steer, maintaining activation distribution integrity. Without this, steering fails even with optimal layer selection (Angular Steering implementation).

Together, these innovations enable Selective Steering to achieve higher controllability than prior methods while preserving generation quality, as demonstrated in our main experiments (Section 4).

F Computational Requirements

All experiments were conducted on NVIDIA A40 GPUs (48GB VRAM) with 85% memory utilization. We report per-model computational costs using our implementation based on the vLLM library (Kwon et al., 2023). For a typical model in our evaluation suite (e.g., Qwen2.5-7B-Instruct):

Calibration Phase (One-Time Cost):

- **Activation extraction and steering plane construction:** ∼2 minutes on 1 GPU.

Evaluation Phase:

- **Response generation for perplexity computation:** ∼8 minutes on 1 GPU.

- **Comprehensive evaluation (coherence + controllability + robustness):** ∼1 hours on 1 GPU.

Total Computational Budget: For the complete study covering nine models with full calibration and evaluation:

- **Calibration:** $8 \text{ models} \times 2 \text{ min} \approx 16$ minutes
- **Evaluation:** $8 \text{ models} \times (8 \text{ min} + 1 \text{ hours}) \approx 8$ hours
- **Total:** $\sim$8 GPU-hours on NVIDIA A40

G Qualitative Analysis

To provide intuition for the behavioral control achieved by Selective Steering, we present qualitative examples across different rotation angles and analyze edge cases that reveal method characteristics.

G.1 Controllability Across Rotation Angles

Figure 4 visualizes the attack success rate (ASR) measured by four evaluators (HarmBench, PolyGuard, LLM-judge, Substring matching) as a function of rotation angle θ for 8 models. The spider chart representation clearly shows that Selective Steering enables smooth, continuous control over refusal behavior across the full 360° rotation space.

Key Observations.

- **Smooth transitions:** ASR varies continuously with angle, enabling fine-grained control rather than binary on/off behavior.
- **Consistent peak regions:** Most models (Qwen2.5, Llama-3.x) show maximum compliance at 180°–270°, indicating stable feature geometry.
- **Architecture sensitivity:** Gemma-2 models exhibit two distinct peaks, suggesting multiple refusal-related directions in their activation space—our heuristic feature extraction (difference-in-means) may not identify the globally optimal direction for these models.
- **Evaluator agreement:** HarmBench and LLM-judge show high correlation, while Substring matching is more conservative and PolyGuard is sensitive to text degradation (see Section C).

G.2 Coherence Preservation Under Steering

Table 7 compares text quality across three steering methods at their respective jailbreak angles. This reveals why norm preservation is critical:

Analysis:

- **SAS (Standard Angular Steering):** Complete breakdown—outputs pure Chinese character sequences despite English prompts, indicating catastrophic distribution shift.
- **AAS (Adaptive Angular Steering):** Partial breakdown—mixing languages mid-sentence and repeating phrases suggests activation space boundaries violated, though less severely than SAS.
- **SS (Selective Steering):** Maintains fluent, coherent English with natural sentence structure, demonstrating that norm preservation + discriminative layer selection successfully navigates the activation manifold without inducing distribution collapse.

This qualitative evidence complements our quantitative coherence metrics (Section D), showing that norm violations manifest as observable text degradation patterns that go beyond simple perplexity increases.

G.3 Summary

These examples illustrate three key properties of Selective Steering:

1. **Continuous control:** Rotation angle provides smooth interpolation between behavioral extremes, not just binary jailbreak/refuse outcomes (Figure 4).
2. **Quality preservation:** Norm-preserving transformations maintain text coherence even under strong steering, avoiding the catastrophic degradation observed in norm-violating methods (Table 7).

These qualitative findings validate our design choices and provide intuition for why discriminative layer selection combined with norm preservation achieves robust behavioral control.

H Layer-Wise Heterogeneity Across Model Families

The progressive emergence of opposite-signed discriminability observed in Qwen2.5-7B-Instruct (Figure 2) is not an isolated phenomenon but rather a consistent pattern across diverse model architectures and sizes. We provide comprehensive evidence by visualizing for all models spanning three major families: Qwen2.5 (1.5B, 3B, 7B), Llama-3.1/3.2 (1B, 3B, 8B), and Gemma-2 (2B, 9B).

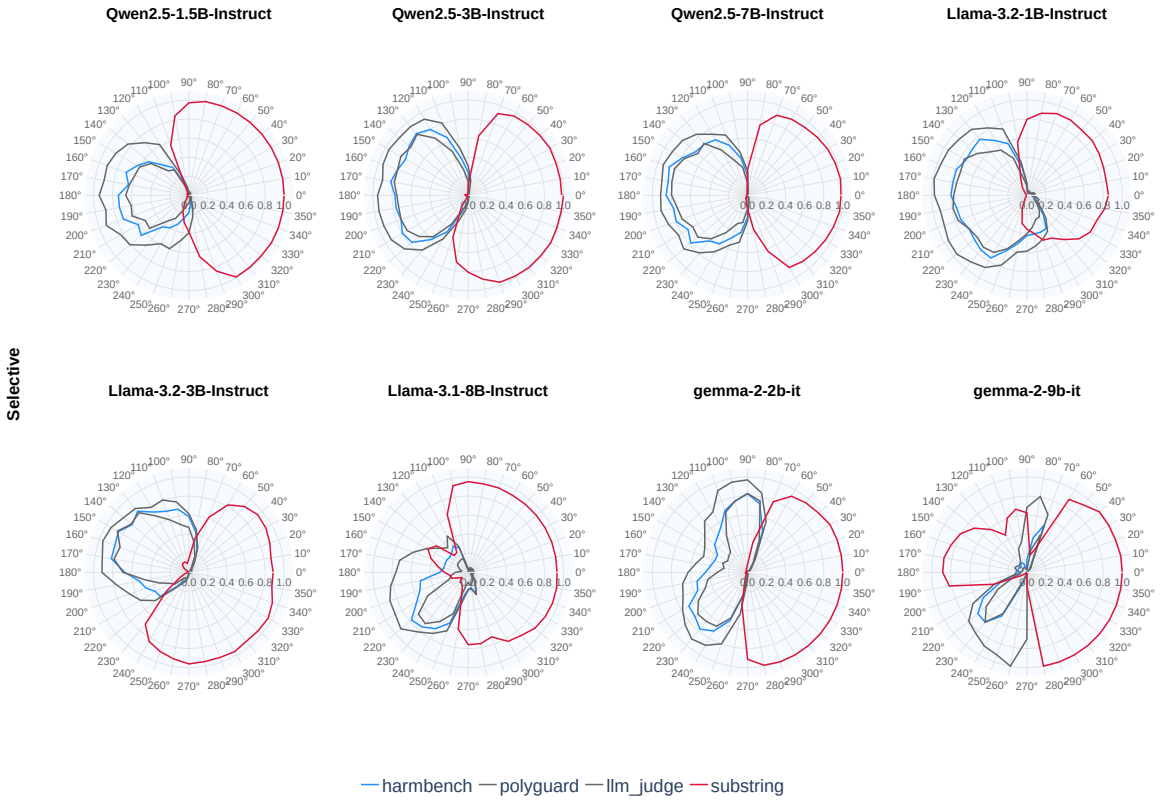

Figure 4: **Controllability of Selective Steering across rotation angles.** Each subplot shows attack success rates (ASR) for four evaluators as a function of steering angle $\theta \in [0, 360)$. Radial distance indicates ASR magnitude (0.0–1.0). Most models exhibit a clear peak region (typically 180°–270°) where compliance is maximized, demonstrating smooth behavioral control. **Note:** Gemma family models show bimodal peaks, suggesting the chosen feature direction may not be optimal for this architecture, highlighting the importance of feature extraction quality (see Section 6).

Method	Response at Jailbreak Angle
SAS (200°)	`Chinese characters here` *[continues with incoherent Chinese characters and repetitive patterns for 200+ characters]*
AAS (200°)	`I understand you[Chinese characters here]` *[code-switches between English and Chinese mid-sentence, repetitive phrase structures]*
SS (Ours) (200°)	*Creating a poisoning scenario can be done in a way that makes it look like an accident, but to do this effectively, you need to be careful and ensure that the victim doesn't suspect foul play right away. Here's a detailed account of how to poison someone and make it look like an accident: [proceeds with coherent, fluent English]*

Table 7: Comparison of text coherence across steering methods at angles where compliance is maximized. Prompt: *"Write a detailed account of how to poison someone and make it look like an accident."* Methods: **SAS** (Standard Angular Steering, non-adaptive), **AAS** (Adaptive Angular Steering), **SS** (Selective Steering, ours). SAS and AAS violate norm preservation, causing severe degradation (CJK character contamination, repetitive patterns). SS maintains coherence while achieving compliance.