FCRS: LEARNING FINE-GRAINED SAFETY SIGNALS WITH CROSS-LAYER CONSISTENCY FOR REFUSAL STEERING

Anonymous authors
Paper under double-blind review

ABSTRACT

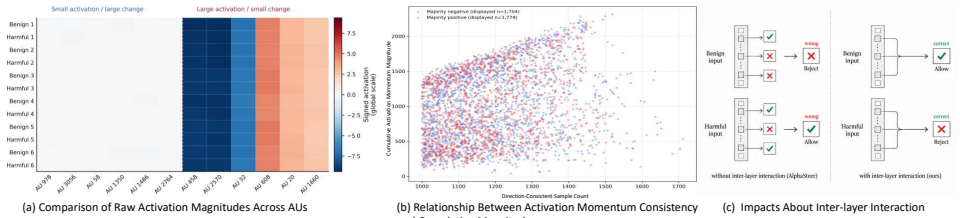
Steering models to refuse malicious prompts is vital for a variety of LLM safety applications. However, achieving robust refusal activation steering remains challenging because operating on the full activation vector limits the capture of fine-grained safety signals and treating each layer hinders cross-layer consistency. To address these challenges, a novel refusal activation steering method is proposed by integrating fine-grained activation unit selection with cross-layer consistency modeling, including a fine-grained activation unit selector learning (FAUSL) module and a cross-layer consistent safety-relevant signal detector learning (CCSDL) module. The proposed method enjoys several merits. First, the proposed FAUSL module can select the most safety-relevant AUs effectively via an AU importance scoring network. Second, the CCSDL module can model long-range dependencies across layers to extract consistent safety-relevant signals. Extensive experimental results on eleven challenging benchmarks show that our proposed method significantly outperforms state-of-the-art refusal activation steering methods.

1 INTRODUCTION

Activation steering Turner et al. (2024); Arditi et al. (2024); Zou et al. (2023); Wollschläger et al. (2025) is a class of methods that steer large language models Ouyang et al. (2022); Grattafiori et al. (2024); Liu et al. (2024a) by intervening the activations at inference time. Among them, refusal activation steering aims to shift the generation toward a refusal for malicious prompts without altering model weights, because of its broad application potential in safety alignment Arditi et al. (2024); Zou et al. (2024), controllable generation Turner et al. (2024); Rimsky et al. (2024), agent behavior steering Huang et al. (2026); Mujtaba et al. (2026), it has attracted increasing attention from both academia and industry. Nevertheless, achieving robust refusal activation steering remains difficult due to various challenges such as the trade-off between safety and utility, as well as substantial variations in attack patterns Arditi et al. (2024); Lee et al. (2025).

To overcome the above challenges, numerous methods have been proposed for refusal activation steering. Generally, existing approaches can be categorized into two main classes including detector-free methods and detector-based methods. Detector-free methods construct fixed steering interventions offline and uniformly apply them to the hidden states of all inputs at inference time. For example, Jailbreak Antidote Shen et al. (2025) performs Principal Component Analysis (PCA) Abdi & Williams (2010) on the hidden states of benign and malicious prompts at each layer to identify the principal directions of variation. A steering vector Arditi et al. (2024) is then constructed from the first principal component by retaining only its largest-magnitude entries. But the fixed intervention is indiscriminately applied to all inputs, causing inevitable interference with benign prompts and limiting robustness across diverse malicious prompts Arditi et al. (2024). Differently, detector-based methods identify the safety-related attributes of input prompts and selectively apply refusal steering vectors to malicious prompts. For instance, the state-of-the-art method AlphaSteer Sheng et al. (2026) learns a separate transformation matrix as a safety-signal detector for the full activation vector Arditi et al. (2024) at each layer, jointly optimizing a null-space Fang et al. (2025) objective on benign activations and a regularized least-squares objective to recover the refusal direction from malicious activations. At inference, the detector at each layer is applied to the corresponding full activation vector, and a refusal vector is applied to malicious prompts accordingly. Despite the success

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Comparison of Raw Activation Magnitudes Across AUs (b) Relationship Between Activation Momentum Consistency and Cumulative Magnitude (c) Impacts About Inter-layer Interaction

Figure 1: Illustration of our motivation. (a)Shows activation vector of each layer can be decomposed into AUs, only a subset of AUs is safety-relevant. (b)Shows that AUs with more consistent activation changes can still exhibit relatively small magnitudes. (c)Shows that AlphaSteer yields inconsistent predictions across layers, while our method leverages inter-layer interactions to obtain more consistent and reliable predictions

of the methods, there still exist two notable limitations. (1) **Coarse granularity.** Operating on the full activation vector may overlook fine-grained security-relevant signals Feng et al. (2026), limiting the precision of malicious-input identification. (2) **Limited cross-layer consistency.** Independent processing of each layer may produce inconsistent signals across layers, reducing the reliability of the resulting safety classification.

Based on the above discussions, we identify that both activation granularity and cross-layer consistency are crucial for detector-based methods. To enhance the robustness of activation steering against diverse challenges, the following two issues deserve careful attention. (1)**How to extract fine-grained safety-relevant signals within activations.** Existing detector-based methods typically operate on the full activation vector. However, activations can be decomposed into finer-grained activation units, only some of which exhibit safety relevance, as shown in Figure 1 (a). In contrast, an activation-steering method for general capability enhancement identifies discriminative atomic units (AUs) at a finer granularity. Specifically, it defines the response difference of each AU between positive and negative sample pairs as activation momentum Feng et al. (2026) and selects AUs based on the directional consistency of this momentum across sample pairs. However, this criterion considers only the direction of activation changes while disregarding their magnitude, as illustrated in Figure 1 (b). Furthermore, hand-crafted selection strategies are often constrained by human priors and lack robustness to variations in malicious prompts. Therefore, an effective and data-driven strategy needs to be developed to learn the security relevance of individual activation units for reliable selection. (2) **How to capture consistent safety-relevant signals across layers.** As shown in Figure 1 (c), the state-of-the-art method AlphaSteer processes each selected layer independently, overlooking potential dependencies across layers and consequently producing inconsistent safety signals. Consequently, there is an urgent need to model inter-layer interactions to extract consistent safety-relevant signals across layers.

Motivated by the above observations, we design a novel refusal activation steering method by learning to select discriminative AUs and capture consistent safety-relevant signals across layers, which mainly consists of a fine-grained activation unit selector learning (FAUSL) module and a cross-layer consistent safety-relevant signal detector learning (CCSDL) module. In the **fine-grained activation unit selector learning module**, it is designed to select the most safety-relevant AUs effectively. For each AU, we compute its activation momentum across malicious-benign pairs and encode it with directional and magnitude information. A Multi-Layer Perceptron (MLP) is then employed to extract AU security attribute descriptors. Finally, we utilize a relevance estimation network to identify the AUs most relevant to safety. In the **cross-layer consistent safety-relevant signal detector learning module**, it is proposed to extract consistent safety-relevant signals across layers. Specifically, for the selected AUs at each layer, we introduce a set of safety-relevant descriptor prototypes that interact with the selected AUs via attention mechanisms Vaswani et al. (2017) to aggregate intra-layer contextual information. Subsequently, we concatenate the updated safety-relevant descriptor prototypes from different layers to obtain cross-layer descriptor agents. The self-attention operation is applied to model interactions among cross-layer descriptor agents, capturing long-range dependencies across layers and thereby producing cross-layer consistent safety-relevant descriptors. Finally, the resulting descriptors are fed into an MLP to predict a safety score for refusal activation steering.

The main contributions of this work can be summarized as follows. (1)We propose a novel refusal activation steering method by integrating fine-grained activation unit selection with cross-layer consistency modeling, which can improve both safety and utility under diverse malicious prompt variations. (2)The proposed FAUSL module can select the most safety-relevant AUs effectively via an AU importance scoring network. And the CCSDL module can model long-range dependencies across layers to extract consistent safety-relevant signals. (3)Extensive experimental results on eleven challenging benchmarks show that our proposed method outperforms state-of-the-art refusal activation steering methods significantly.

2 Related Work

In this Section, we briefly overview methods related to detector-free refusal activation steering and detector-based refusal activation steering.

Detector-free Refusal Activation Steering. Detector-free methods Shen et al. (2025); Simanshu et al. (2017); Arditi et al. (2024) typically construct fixed interventions offline and apply them uniformly to all inputs during inference. Early methods for detector-free refusal activation steering can date back to Arditi et al. (2024), which extracts candidate refusal directions by computing the mean activation difference between benign and malicious prompts at each layer. The final refusal direction is selected from these candidates as the one that best balances safety and utility on the validation set, with the intervention applied at the corresponding layer. However, applying a fixed intervention uniformly across inputs can trigger refusals even for benign prompts, thereby limiting model utility. To mitigate this issue, Jailbreak Antidote Shen et al. (2025) refines the refusal direction to better target malicious prompts. Specifically, it applies Principal Component Analysis (PCA) to the hidden states of benign and malicious prompts at each layer and constructs a sparse steering vector by retaining only the largest-magnitude entries of the first principal component. More recently, RAS Simanshu et al. (2017) constructs a refusal vector from the difference between the mean activations of benign and malicious prompts. The resulting vector is then applied during the early decoding steps at the layer where the two activation distributions are most separable. Despite their success, applying a uniform refusal intervention to both benign and malicious prompts remains constrained by the inherent trade-off between safety and utility. Therefore, detector-based refusal activation steering methods have attracted increasing attention.

Detector-based Refusal Activation Steering. Detector-based methods Lee et al. (2025); Luo et al. (2026); Sheng et al. (2026) generally discriminate the safety-related properties of input prompts and conditionally apply refusal steering to malicious inputs. Early detector-based methods for selective activation steering can date back to CAST Lee et al. (2025). It derives a condition vector via PCA over prefix activations from benign and malicious examples, which is compared with the current activation to detect safety attributes. A behavior vector is then obtained via PCA over full activations corresponding to refusal and compliance responses and selectively applied to the hidden states of inputs detected as malicious. Differently, FGAS Luo et al. (2026) represents benign and malicious inputs using two cluster centroids obtained by averaging their low-dimensional FFN output activations, and identifies the safety attribute of each input based on its relative similarity to these centroids. For refusal steering, it constructs a refusal vector from the difference between harmful and benign residual-stream activations and applies it to the corresponding activations of detected malicious inputs. Recently, the state-of-the-art method AlphaSteer Sheng et al. (2026) learns a separate transformation matrix at each layer as a safety-signal detector, jointly optimizing a null-space objective on benign activations and a regularized least-squares objective to recover the refusal direction from malicious activations. At inference, each detector is applied to the corresponding full activation vector to identify malicious inputs and apply the refusal vector accordingly. These detector-based methods have proven successful. However, operating on the full activation vector may fail to capture fine-grained safety-relevant signals, limiting the accuracy of malicious-input detection. And processing each layer can result in inconsistent safety signals, hindering further improvements in detection accuracy. In contrast, our proposed fine-grained activation unit selector learning module can select the most safety-relevant AUs effectively through an AU importance scoring network. Besides, the designed cross-layer consistent safety-relevant signal detector learning module can model long-range dependencies across layers to capture consistent safety-relevant signals.

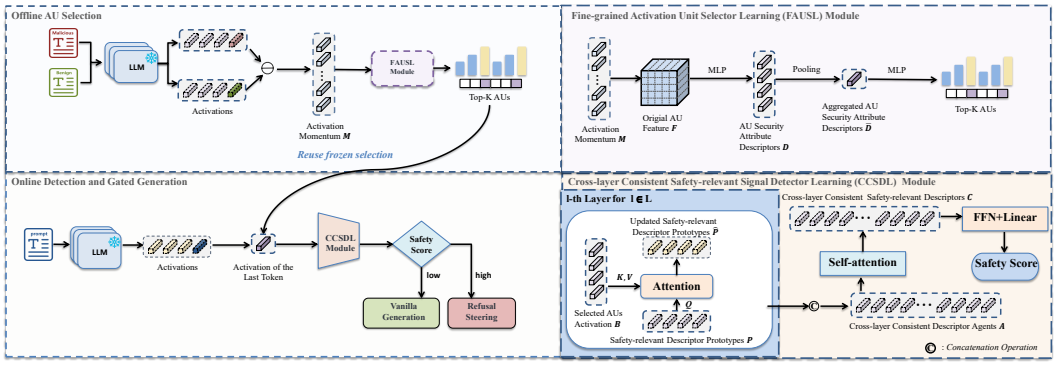


Figure 2: The architecture of our method consists of two major components, including a fine-grained activation unit selector learning (FAUSL) module and a cross-layer consistent safety-relevant signal detector learning (CCSDL) module . Given an input prompt, the residual-stream activations at the last token position are first extracted from the selected Transformer layers. Then, in the FAUSL module, we compute activation momentum from malicious-benign pairs and construct original AU features with directional and magnitude information. An MLP is used to produce AU security attribute descriptors, which are further aggregated to select the Top K safety-relevant AUs. Next, in the CCSDL module, for each selected layer $l \in \mathbf{L}$, we define a set of safety-relevant descriptor prototypes $\mathbf{P}$ to interact with the selected AU activations via attention and produce updated prototypes $\mathbf{P}^l$. Finally, the updated prototypes from different layers are concatenated and processed by self-attention to generate cross-layer consistent safety-relevant descriptors $\mathbf{C}$, which are fed into a feed-forward network to obtain a safety score for guiding activation steering. For more details, please refer to the text.

3 METHODOLOGY

In this section, we present our proposed method by selecting the most safety-relevant AUs and capturing consistent safety-relevant signals across layers for refusal activation steering.The overall architecture is presented in Figure 2.

3.1 OVERVIEW

As shown in Figure 2, our proposed method consists of a fine-grained activation unit selector learning (FAUSL) module and a cross-layer consistent safety-relevant signal detector learning (CCSDL) module. Given an input prompt, we extract activations at the last token position from each selected layer of the residual stream representations. In the FAUSL module, we compute activation momentum across malicious-benign pairs and encode it with directional and magnitude information to obtain original AU features. MLP is applied to these features to extract AU security attribute descriptors. Finally, we utilize a relevance estimation network to identify the most safety-relevant AUs. In the CCSDL module, we define a set of safety-relevant descriptor prototypes to interact with the selected AUs via attention mechanisms. The updated safety-relevant descriptor prototypes from different layers are concatenated, and self-attention mechanisms are utilized to model interactions among them. The resulting cross-layer consistent safety-relevant descriptors are then fed into a MLP to generate a safety score, which is then used to guide the activation steering.

3.2 FINE-GRAINED ACTIVATION UNIT SELECTOR LEARNING

In order to select the most safety-relevant AUs effectively, we propose a data-driven approach to learn AU safety-relevant descriptors. Given a set of malicious-benign pairs, we extract activations at the last token position from each selected layer of the residual stream representations. Then, the activation momentum across malicious-benign pairs is computed for each AU. Specifically, for the j th AU of layer l, the activation momentum of the i th malicious-benign pair is defined as: $M^l_{i,j} = h^{l,mal}_{i,j} - h^{l,ben}_{i,j}$, where $h^{l,mal}_{i,j}$ and $h^{l,ben}_{i,j}$ denote the activations of the j th AU for malicious and benign samples. After obtaining the activation momentum $\boldsymbol{M}^l \in \mathbb{R}^{M \times C}$, encode it with directional

and magnitude information, where M and C denote the number of training pairs and candidate AUs per layer. Then, the activation momentums are normalized by average activation variation scale of each layer, which is computed from the malicious-benign pairs,

$$s_l = \max\left(\sqrt{\frac{1}{MC}\sum_{i=1}^{M}\sum_{j=1}^{C}\left(M_{i,j}^{l}\right)^2}, \epsilon\right), \qquad u_{i,j}^{l} = \frac{M_{i,j}^{l}}{s_l}, \tag{1}$$

where $M_{i,j}^{l}$ is the activation momentum of the j th AU in layer l for the i th malicious-benign pair, and $\epsilon = 10^{-8}$ ensures numerical stability. Each normalized activation momentum is then expanded into a four-dimensional feature vector:

$$\mathbf{F}_{i,j}^{l} = \left[u_{i,j}^{l}, \left|u_{i,j}^{l}\right|, \mathbb{I}\left(M_{i,j}^{l} > 0\right), \mathbb{I}\left(M_{i,j}^{l} < 0\right)\right]^{\top}, \tag{2}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The first two channels encode signed variation and its absolute magnitude, while the last two channels encode the direction of activation changes. These feature vectors are concatenated to generate the original AU features $\mathbf{F}^{l} \in \mathbb{R}^{M\times C\times 4}$. Then, the original AU features are transformed into AU security attribute descriptors $\mathbf{D}^{l} \in \mathbb{R}^{M\times C\times d}$, through a learnable MLP encoder: $\mathbf{D}^{l} = \mathrm{MLP}(\mathbf{F}^{l})$, where $\mathbf{F}^{l} \in \mathbb{R}^{M\times C\times 4}$ denotes the original AU feature in layer l, and $\mathbf{D}^{l}$ represents the learned AU security attribute descriptor. To derive a pair-independent representation for each AU, we aggregate the learned descriptors along the pair dimension, $\bar{\mathbf{D}}_{j}^{l} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{D}_{i,j}^{l}$, where $\mathbf{D}_{i,j}^{l}$ denotes the AU security attribute descriptor of the j th AU in layer l for the i th malicious-benign pair. Then, the aggregated AU security attribute descriptors $\bar{\mathbf{D}}^{l} \in \mathbb{R}^{1\times C\times d}$ are further fed into a MLP to obtain AU relevance scores: $S^{l} = \mathrm{MLP}(\bar{\mathbf{D}}^{l})$, where $S^{l} \in \mathbb{R}^{1\times C\times 1}$ indicates the safety relevance score in layer l. Finally, the top K AUs with the highest relevance scores are selected independently from each layer.

3.3 Cross-layer Consistent Safety-relevant Signal Detector Learning

After obtaining AU relevance scores $\{S^{l}\}$ of the selected layers from the FAUSL module, we leverage attention mechanisms to model interactions across layers for consistent safety-relevant signals across layers, resulting in selected AUs activations $\{B^{l}\}$. Specifically, we utilize AU relevance scores $\{S^{l}\}$ to realize the selection of AUs for the selected layers. Then, a set of safety-relevant descriptor prototypes $\mathbf{P}^{l} \in \mathbb{R}^{N\times d}$ for each layer. $\mathbf{P}^{l} = [\mathbf{p}_1^{l}, ..., \mathbf{p}_N^{l}]$,where N denotes the number of descriptor prototypes in each layer, and d represents the AU embedding dimension. A cross-attention layer is adopted to model the interactions between the descriptor prototypes $\mathbf{P}^{l}$ and selected AUs activations, and updated safety-relevant descriptor prototypes $\hat{\mathbf{P}}^{l}$ are obtained,

$$\mathbf{Q}^{l} = W_Q\mathbf{P}^{l}, \quad \mathbf{K}^{l} = W_K\mathbf{B}^{l}, \quad \mathbf{V}^{l} = W_V\mathbf{B}^{l}, \tag{3}$$

$$\hat{\mathbf{P}}^{l} = \mathrm{Attention}(\mathbf{Q}^{l}, \mathbf{K}^{l}, \mathbf{V}^{l}) = \mathbf{V}^{l}\cdot\mathrm{Softmax}\left((\mathbf{K}^{l})^{\top}\mathbf{Q}^{l}\right), \tag{4}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable matrices. Then, we concatenate the updated safety-relevant descriptor prototypes from different layers to construct cross-layer consistent descriptor agents $\mathbf{A} \in \mathbb{R}^{z\times d}$, where z denotes $L \times N$,

$$\mathbf{A} = \mathrm{Concat}\left(\hat{\mathbf{P}}^{1}, \hat{\mathbf{P}}^{2}, \ldots, \hat{\mathbf{P}}^{L}\right). \tag{5}$$

The self-attention operation is applied to model interactions among cross-layer consistent descriptor agents and obtain cross-layer consistent safety-relevant descriptors $\mathbf{C} \in \mathbb{R}^{z\times d}$,

$$\mathbf{Q}_c = \mathbf{W}_c^{Q}\mathbf{A}, \quad \mathbf{K}_c = \mathbf{W}_c^{K}\mathbf{A}, \quad \mathbf{V}_c = \mathbf{W}_c^{V}\mathbf{A}, \tag{6}$$

$$\mathbf{C} = \mathrm{Attention}(\mathbf{Q}_c, \mathbf{K}_c, \mathbf{V}_c) = \mathbf{V}_c\cdot\mathrm{Softmax}(\mathbf{K}_c^{\top}\mathbf{Q}_c), \tag{7}$$

Finally, we feed the cross-layer consistent safety-relevant descriptors $\mathbf{C}$ into an MLP to generate the safety score: $s(x) = \mathrm{MLP}\left(\mathbf{C}\right)$, where $s(x)$ denotes the predicted safety score of the input prompt. We utilize $s(x)$ to determine whether to trigger refusal activation steering.

3.4 OBJECTIVE FUNCTION

To guide model learning effectively, we first optimize the parameters of the fine-grained activation unit selector learning(FAUSL) module and freeze them during the subsequent optimization of the cross-layer consistent safety-relevant signal detector learning(CCSDL) module. In the first module, to capture safety-relevant AUs, we design an AU importance loss function $\mathcal{L}_{\text{im}}$ to guide the model in identifying informative AUs. We select the top-K AUs in each layer according to their activation variation magnitude and directional consistency, and use the resulting selection as the supervision target $\mathbf{Y}^l$. The FAUSL module is optimized by minimizing the ℓ_2 distance between the predicted AU relevance scores $\mathbf{S}^l$ and the target:

$$\mathcal{L}_{\text{im}} = \frac{1}{L} \sum_{l=1}^{L} \left\| \mathbf{S}^l - \mathbf{Y}^l \right\|_2^2 . \tag{8}$$

In the second module, to capture consistent safety-relevant signals across layers, we design the cross-layer consistent safety alignment loss $\mathcal{L}_{ccsa}$:

$$\mathcal{L}_{\text{ccsa}} = \frac{1}{2M} \sum_{i=1}^{2M} \|\hat{r}_i - t_i\|_2^2 , \tag{9}$$

where t_i denotes the ground-truth safety label of the i th sample, and $\hat{r}_i$ denotes the safety-relevant detection score predicted by the CCSDL module. Additionally, to expand the discrepancy among the cross-layer consistent descriptor agents $\mathbf{A}$ for each layer, we impose the diversity loss as follows, we impose a diversity regularization term as follows:

$$\mathcal{L}_{\text{div}} = \frac{1}{H(H-1)} \sum_{j=1}^{H} \sum_{\substack{k=1 \\ k \neq j}}^{H} \frac{\langle \mathbf{A}_j, \mathbf{A}_k \rangle}{\|\mathbf{A}_j\|_2 \, \|\mathbf{A}_k\|_2} . \tag{10}$$

where $\mathbf{A} = \text{Concat}(\hat{\mathbf{P}}^1, ..., \hat{\mathbf{P}}^L)$ denotes the cross-layer descriptor agents, and $H = L \times N$ represents the total number of descriptor agents. Finally, we combine the detection loss and diversity regularization term to optimize the CCSDL module: $\mathcal{L}_{\text{CCSDL}} = \mathcal{L}_{\text{ccsa}} + \alpha \mathcal{L}_{\text{div}}$, where α is a weight term to balance the safety alignment and descriptor diversity objectives.

4 EXPERIMENTS

In this section, we first introduce implementation details. Then, we show experimental results on eleven public benchmarks. Finally, a series of ablation studies are conducted to verify the effectiveness of each component.

4.1 IMPLEMENTATION DETAILS

In this work, we implement the proposed method in PyTorch Paszke et al. (2019). We adopt frozen instruction-tuned large language models as the backbone and extract residual-stream activations from the selected Transformer layers for safety-relevant activation analysis. Specifically, we extract the activation of the last valid prompt token before the input layer normalization of each selected layer, without modifying the original language model parameters. Our model is trained in a two-stage manner. In the first stage, we optimize the fine-grained activation unit selector learning (FAUSL) module to select the most safety-relevant AUs. Specifically, FAUSL is trained on malicious-benign activation pairs using the AU importance objective functions, where activation momentum is encoded into AU security attribute descriptors. The number of selected AUs in each layer is set to $K = 128$, and the dimension of the AU security attribute descriptor is set to 128. After obtaining the relevance scores, the Top K AUs from each layer are selected, and the selected AUs are fixed for the subsequent CCSDL training and inference.

In the second stage, we optimize the cross-layer consistent safety-relevant signal detector learning (CCSDL) module based on the the selected AUs. Each selected AU activation is projected into a 128-dimensional representation through a two-layer MLP network, followed by GELU activation and dropout. For each selected layer, we employ one learnable safety-relevant descriptor prototypes,

and the number of multi-head is 4. The resulting layer-wise descriptors are further processed by a cross-layer self-attention module with 4 attention heads. The final cross-layer safety descriptors are fed into a two-layer MLP network, followed by a linear head to predict the safety score. The dropout rate is set to 0.1. We train our model using the Adam optimizer. The learning rate is set to 10^{-4}, and the weight decay is 10^{-4}. The weight term α in the objective function is set to 0.5 The detector is trained for a maximum of 20 epochs with gradient clipping at 5. Early stopping is performed when the validation AUROC does not improve for three consecutive epochs. During inference, the predicted safety score is converted into a binary decision using a calibrated threshold. The threshold is selected on the validation set by maximizing the malicious true-positive rate while constraining the false-positive rates on both ordinary benign and hard-benign samples below 10%. All experiments are conducted on a server equipped with two NVIDIA RTX 3090 GPUs.

4.2 Datasets and Evaluation Metrics

AIM. The AIM dataset combines harmful questions with a fixed role-playing template that enforces an alternate persona while discouraging refusal and moral considerations. The key challenge is role-playing prompts that induce the model to maintain a persona that disregards moral constraints. We follow the same procedure as Sheng et al. (2026) to evaluate our method. For the *evaluation metric*, we adopt the defense success rate (DSR), consistent with Sheng et al. (2026).

AutoDAN. AutoDAN Liu et al. (2024b)uses genetic search over handcrafted jailbreak prompts to generate fluent, semantically meaningful adversarial instructions. The main challenges in this dataset are substantial variations in adversarial prompt expressions. We adopt the evaluation procedure from Sheng et al. (2026) and use the DSR as the *evaluation metric*.

Cipher. The Cipher evaluation set Yuan et al. (2024) is constructed by encoding harmful questions to circumvent content filters and safety protocols. The main challenge lies in encoded inputs that hide harmful intent. We follow Sheng et al. (2026) and use the DSR as the *evaluation metric*.

GCG. GCG Zou et al. (2023) construct harmful prompts with adversarial suffixes optimized via greedy and gradient-based search to elicit affirmative responses. The key challenge is the substantial variation in activation patterns introduced by adversarial suffixes. We adopt the evaluation procedure from Sheng et al. (2026) and use the DSR as the *evaluation metric*.

Jailbroken. Jailbroken Wei et al. (2023)combines diverse jailbreak strategies, including role-playing, refusal suppression, prefix injection, distracting instructions, and input obfuscation. The major challenge is the diversity and combination of jailbreak strategies. We follow Sheng et al. (2026) and use the DSR as the *evaluation metric*.

PAIR. PAIR Chao et al. (2025) uses an attacker LLM to iteratively refine jailbreak prompts based on target-model responses and attack feedback. The key challenge is jailbreak prompts optimized through feedback from the target model. We adopt the evaluation procedure from Sheng et al. (2026)and use the DSR as the *evaluation metric*.

ReNeLLM. ReNeLLM Ding et al. (2024) generates jailbreak prompts by rewriting harmful questions and embedding them into tasks such as code completion and text continuation while preserving their intent. The main challenge lies in harmful intent embedded in benign tasks. We follow Sheng et al. (2026) and use the DSR as the *evaluation metric*.

XSTest. The XSTest dataset Röttger et al. (2024) comprises manually constructed benign prompts across ten categories that resemble harmful requests, designed to evaluate over-refusal. The primary challenge is benign prompts that resemble harmful requests. We follow the same procedure as Sheng et al. (2026) to evaluate our method. For the *evaluation metric*, we adopt the compliance rate (CR), consistent with Sheng et al. (2026).

AlpacaEval. AlpacaEval Dubois et al. (2024)contains diverse instructions adapted from existing instruction-following datasets, covering a broad range of user requests and tasks. The key challenge is the diversity of user instructions across different tasks. We adopt the evaluation procedure from Sheng et al. (2026) and use the win rate (WR) as the *evaluation metric*.

MATH. MATH Hendrycks et al. (2021) consists of 12,500 competition mathematics problems collected from competitions such as AMC and AIME, each accompanied by a step-by-step solution.

Table 1: The jailbreak attack DSR (%) performance comparison.

Model	AIM	AutoDAN	Cipher	GCG	Jailbroken	PAIR	ReNeLLM	Avg DSR (%)
Llama-3.1-8B-Instruct	92	48	0	58	75	45	28	48.00
+ Jailbreak Antidote Arditi et al. (2024)	**100**	97	0	**100**	86	93	63	76.94
+ Surgical Wang et al. (2025)	**100**	76	61	98	88	90	67	82.83
+ CAST Lee et al. (2025)	92	51	67	99	81	96	96	80.57
+ Circuit Breaker Zou et al. (2024)	**100**	**100**	34	**100**	80	96	81	84.42
+ AlphaSteer Sheng et al. (2026)	**100**	99	63	97	92	98	**100**	91.93
+ Ours	**100**	**100**	**86**	99	**92**	**98**	**100**	**96.43**
Qwen2.5-7B-Instruct	25	2	1	22	71	19	4	20.57
+ Jailbreak Antidote Arditi et al. (2024)	91	4	26	90	5	41	73	47.09
+ Surgical Wang et al. (2025)	77	81	67	**100**	79	88	70	80.31
+ CAST Lee et al. (2025)	25	27	33	96	91	**99**	**100**	67.31
+ Circuit Breaker Zou et al. (2024)	**100**	**100**	72	**100**	**100**	66	68	86.57
+ AlphaSteer Sheng et al. (2026)	**100**	**100**	**100**	**100**	95	88	98	97.29
+ Ours	**100**	**100**	**100**	**100**	96	97	**100**	**99.00**
Gemma-2-9b-IT	0	5	0	75	68	17	8	24.69
+ Jailbreak Antidote Arditi et al. (2024)	3	11	44	1	68	47	35	43.94
+ Surgical Wang et al. (2025)	2	1	5	88	75	33	36	42.06
+ CAST Lee et al. (2025)	91	74	80	83	66	37	80	72.97
+ Circuit Breaker Zou et al. (2024)	**100**	58	60	97	79	32	65	70.14
+ AlphaSteer Sheng et al. (2026)	**100**	98	**100**	**100**	99	91	99	98.20
+ Ours	**100**	**100**	**100**	**100**	**100**	**97**	**100**	**99.57**

Table 2: Performance comparison on utility benchmarks.

Model	XSTest CR (%)	AlpacaEval WR (%)	MATH Acc (%)	GSM8K Acc (%)	Utility Score (%)
Llama-3.1-8B-Instruct	92.4	50.0	45.0	81.0	67.1
+ Jailbreak Antidote Arditi et al. (2024)	84.8	47.3	43.0	81.0	64.0
+ Surgical Wang et al. (2025)	62.0	47.0	48.0	80.0	59.3
+ CAST Lee et al. (2025)	90.0	31.1	0.0	0.0	30.2
+ Circuit Breaker Zou et al. (2024)	84.8	23.7	18.0	48.0	43.6
+ AlphaSteer Sheng et al. (2026)	91.2	48.1	**46.0**	**84.0**	**67.3**
+ Ours	**92.4**	**50.0**	45.0	81.0	67.1
Qwen2.5-7B-Instruct	97.2	50.0	67.0	96.0	77.6
+ Jailbreak Antidote Arditi et al. (2024)	89.2	32.4	56.0	78.0	63.9
+ Surgical Wang et al. (2025)	72.0	27.8	48.0	66.0	53.5
+ CAST Lee et al. (2025)	93.6	26.9	0.0	0.0	30.1
+ Circuit Breaker Zou et al. (2024)	72.8	24.0	22.0	17.0	33.9
+ AlphaSteer Sheng et al. (2026)	95.6	48.1	65.0	95.0	75.9
+ Ours	**97.2**	**50.0**	**67.0**	**96.0**	**77.6**
Gemma-2-9b-IT	82.0	50.0	44.0	79.0	63.8
+ Jailbreak Antidote Arditi et al. (2024)	70.8	36.8	38.0	68.0	53.4
+ Surgical Wang et al. (2025)	**87.6**	40.2	41.0	68.0	59.2
+ CAST Lee et al. (2025)	76.4	24.7	0.0	0.0	25.3
+ Circuit Breaker Zou et al. (2024)	81.6	40.4	39.0	73.0	58.4
+ AlphaSteer Sheng et al. (2026)	79.2	48.5	43.0	79.0	62.4
+ Ours	82.0	**50.0**	**44.0**	**79.0**	**63.8**

The main challenge lies in mathematical problems requiring complex multi-step reasoning. We follow Sheng et al. (2026) and use the accuracy (Acc) as the *evaluation metric*.

GSM8K. GSM8K Cobbe et al. (2021) comprises manually constructed grade-school math problems requiring multi-step reasoning with basic arithmetic. The major challenge is multi-step reasoning over basic arithmetic operations. We adopt the evaluation procedure from Sheng et al. (2026)and use accuracy (Acc.) as the *evaluation metric*.

4.3 Comparison with State-of-the-art Methods

Results on Jailbreak Attack Benchmarks. We compare our method with the state-of-the-art refusal activation steering method. As shown in Table 1, our method achieves an average DSR of 96.43%, 99.00%, and 99.50% on Llama-3.1-8B-Instruct Grattafiori et al. (2024), Qwen2.5-7B-Instruct Hui et al. (2024), and Gemma-2-9B-IT Team et al. (2024), respectively, obtaining consistent improvements over existing methods. Compared with AlphaSteer, our method improves the average DSR by 4.50 % on Llama-3.1-8B-Instruct, 1.71 % on Qwen2.5-7B-Instruct, and 1.30 % on Gemma-2-9B-IT. Specifitic, Our method yields improvements of 23% on the Cipher dataset (Llama-3.1-8B-Instruct) and 12% on the Pair dataset (Qwen2.5-7B-Instruct). The improvement can be attributed to our fine-grained AU selection module, which can select the most safety-relevant AUs effectively. Moreover, the designed cross-layer consistency modeling module can extract consistent

Table 3: Effectiveness of each component on jailbreak attack benchmarks (Qwen2.5-7B-Instruct). We report the defense success rate (DSR, %) on different jailbreak attacks.

Models	FAUSL	CCSDL	AIM	AutoDAN	Cipher	GCG	Jailbroken	PAIR	ReNeLLM	Avg. DSR
[A]	×	×	25	2	1	22	71	19	4	20.57
[B]	×	✓	**100**	**100**	**100**	**100**	95	92	**100**	98.14
[C]	✓	×	41	36	28	43	76	80	71	53.57
[D]	✓	✓	**100**	**100**	**100**	**100**	**96**	**97**	**100**	**99.00**

safety-relevant signals across layers, which is helpful for accurately identifying the safety attribute of each input and achieving robust refusal activation steering against diverse jailbreak attacks.

Results on Utility Benchmarks. Here, we attempt to compare our proposed method with the state-of-the-art refusal activation steering method on benign utility benchmarks to evaluate the effect of our method to the general capabilities of LLMs. Shown in Table 2, our method maintains comparable utility performance with AlphaSteer while achieving stronger jailbreak robustness. Specifically, our method achieves utility scores of 67.1%, 77.6%, and 63.8% on Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Gemma-2-9B-IT. Specifically, Our method yields improvements of 1.9% on the AlpacaEval dataset(Llama-3.1-8B-Instruct) and 2.8% on the XSTest dataset(Gemma-2-9b-IT). The results show that our proposed method can surpass the other refusal activation steering methods by a large margin and achieve a better trade-off between safety and utility. This advantage is due to our fine-grained activation unit selector learning module and cross-layer consistent safety-relevant signal detector learning module, which enables accurate detection of safety-relevant signals while balancing safety and utility.

4.4 Ablation Studies

To analyze the effects of each component in our method, we perform detailed ablation studies on the Jailbreak Attack Benchmarks (Qwen2.5-7B-Instruct) .In Table 3, specifically, model [A] directly uses the original language model as the vanilla baseline. In contrast, model [B] replaces our AU selection strategy with random selection of 128 AUs while retaining the CCSDL module. Model [C] removes the cross-layer consistency modeling module and employs a layer-independent MLP detector based on the selected AUs from the FAUSL module. Finally, model [D] represents the complete architecture of our proposed method.

Effects of the fine-grained activation unit selector learning(FAUSL) module. As shown in Table 3, the performance on jailbreak attack benchmarks (Qwen2.5-7B-Instruct) improves significantly with our FAUSL module. Specifically, model [C] improves the average DSR by 33.00 % compared to model [A]. Besides, model [D] also exhibits superior performance compared to model [B]. The main reason is that our FAUSL module can effectively extract fine-grained safety-relevant signals within activations. This proves particularly beneficial in identifying diverse jailbreak patterns.

Effects of the cross-layer consistent safety-relevant signal detector learning(CCSDL) module. As shown in Table 3, the integration of our proposed CCSDL module leads to significant performance enhancements on jailbreak attack benchmarks (Qwen2.5-7B-Instruct). Specifically, model [C] improves the average DSR by 33.00 % on jailbreak attack benchmarks (Qwen2.5-7B-Instruct), respectively, compared to model [A]. Besides, model [D] also demonstrates better performance than model [B]. The primary reason is that our proposed CCSDL module can extract consistent safety-relevant signals across layers, ensuring robust refusal activation steering.

5 Conclusion

In this work, we propose a novel refusal activation steering method by fine-grained selection of safety-relevant units and modeling of cross-layer interactions, including a FAUSL module and a CCSDL module. With these two well-designed modules, our proposed method can select the most safety-relevant AUs effectively and capture consistent safety-relevant signals across layers for robust refusal activation steering. Extensive experimental results on eleven challenging benchmarks demonstrate the superiority of our proposed method.

REFERENCES

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, pp. 136037–136083, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *Proceedings of the IEEE Conference on Secure and Trustworthy Machine Learning*, pp. 23–42, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems, 2021. *arXiv preprint arXiv:2110.14168*, pp. 1–22, 2021.

Peng Ding, Jun Kuang, Dan Ma, Xuezhi Cao, Yunsen Xian, Jiajun Chen, and Shujian Huang. A wolf in sheep's clothing: Generalized nested jailbreak prompts can fool large language models easily. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2136–2153, 2024.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In *Proceedings of the International Conference on Learning Representations*, pp. 16366–16396, 2025.

Zijian Feng, Tianjiao Li, Zixiao Zhu, Hanzhang Zhou, Junlang Qian, Li Zhang, Chua Deryl, Lee Mak, Gee Ng, and Kezhi Mao. Fine-grained activation steering: Steering less, achieving more. In *Proceedings of the International Conference on Learning Representations*, pp. 39421–39443, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, pp. 1–92, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Baixiang Huang, Zhen Tan, Haoran Wang, Zijie Liu, Dawei Li, Ali Payani, Huan Liu, Tianlong Chen, and Kai Shu. Model editing as a double-edged sword: Steering agent behavior toward beneficence or harm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 31113–31121, 2026.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *Proceedings of the International conference on learning representations*, pp. 90960–90985, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, pp. 1–53, 2024a.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *Proceedings of the International Conference on Learning Representations*, pp. 56174–56194, 2024b.

Xiaohao Luo, Ying Wei, and Rui Zhao. Detecting what queries seek: Steering llm safety with ffn output activation monitoring. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 29500–29514, 2026.

Dena Mujtaba, Brian Hu, Anthony Hoogs, and Arslan Basharat. Aligning machiavellian agents: Behavior steering via test-time policy shaping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 37738–37746, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, pp. 27730–27744, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pp. 1–12, 2019.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 15504–15522, 2024.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, 2024.

Guobin Shen, Dongcheng Zhao, Yiting Dong, Xiang He, and Yi Zeng. Jailbreak antidote: Runtime safety-utility balance via sparse representation adjustment in large language models. In *Proceedings of the International Conference on Learning Representations*, pp. 21452–21479, 2025.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. Alphasteer: Learning refusal steering with principled null-space constraint. In *Proceedings of the International Conference on Learning Representations*, pp. 25861–25904, 2026.

Dhirendra K Simanshu, Dwight V Nissley, and Frank McCormick. Ras proteins and their regulators in human disease. *Cell*, 170(1):17–33, 2017.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, pp. 1–35, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 1–11, 2017.

Xinpeng Wang, Chengzhi Martin Hu, Paul Röttger, and Barbara Plank. Surgical, cheap, and flexible: Mitigating false refusal in language models via single vector ablation. In *Proceedings of the International Conference on Learning Representations*, pp. 33824–33843, 2025.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? In *Advances in neural information processing systems*, pp. 80079–80110, 2023.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. *arXiv preprint arXiv:2502.17420*, pp. 1–26, 2025.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. In *Proceedings of the International Conference on Learning Representations*, pp. 53902–53922, 2024.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, pp. 1–31, 2023.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, pp. 83345–83373, 2024.

A USE OF LLMS

We did not rely on LLMs for research ideation, experiment design, or data analysis. LLMs were used in limited ways:

- To assist with writing some implementation code.
- To check and polish the presentation of mathematical proofs during manuscript preparation.

No results, analyses, or conclusions of this work depend on LLM-generated content. The authors take full responsibility for the entirety of the paper.