the unified direction. Formally, the set of vectors is stacked into a matrix:

$$V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix} \in \mathbb{R}^{k \times d} \tag{8}$$

We then perform PCA on V to identify the dominant direction of maximum variance across all domains, and normalize the first principal component as:

$$v_{\text{pca}} = \text{PCA}(V), v_{\text{shared}} \leftarrow \frac{v_{\text{pca}}}{\|v_{\text{pca}}\|_2} \tag{9}$$

Orthogonal Projection (OP). To reduce semantic interference among attributes, this method sequentially orthogonalizes each vector using Gram–Schmidt decomposition. For $i \geq 2$, the vector v_i is projected onto the orthogonal complement of the subspace spanned by the previous vectors:

$$\tilde{v}_i = v_i - \sum_{j=1}^{i-1} \frac{\langle v_i, \tilde{v}_j \rangle}{\|\tilde{v}_j\|^2} \tilde{v}_j, \qquad v_{\text{shared}} = \sum_{i=1}^{k} \tilde{v}_i. \tag{10}$$

Shared Feature Selection (SFS). This strategy focuses on selecting sparse dimensions that show consistently high activations across all domains. Specifically, it enhances dimensions where all v_i exceed a positive threshold τ_+ and suppresses those where all v_i fall below a negative threshold τ_-. Let $[v_i]_j$ denotes the j-th coordinate of v_i. The composed vector is defined as

$$[v_{\text{shared}}]_j = \begin{cases} \frac{1}{k} \sum_{i=1}^{k} [v_i]_j & \text{if } [v_i]_j \geq \tau_+ \text{ for all } i = \{1, \ldots, k\} \\ \frac{1}{k} \sum_{i=1}^{k} [v_i]_j & \text{if } [v_i]_j \leq \tau_- \text{ for all } i = \{1, \ldots, k\} \\ 0 & \text{otherwise} \end{cases}$$

Linear Weighting (LW). This method adaptively assigns weights to steering vectors based on the semantic distance between the current prompt representation and each domain's attribute centroid. Let $h \in \mathbb{R}_d$ denote the prompt's hidden state of the intervention layer, and $c_i \in \mathbb{R}_d$ the centroid of domain i. The unified vector is computed as:

$$d_i = \|h - c_i\|_2, \alpha_i = \frac{d_i}{\sum_{j=1}^{k} d_j}, v_{\text{shared}} = \sum_{i=1}^{k} \alpha_i v_i. \tag{11}$$

In the following sections, unless explicitly stated otherwise, we apply Principal Component Analysis (PCA) as the default method for composing multi-attribute steering vectors.

4.2. Main Experimental Results

We evaluate the performance of SRS on both single-attribute and multi-attribute steering tasks. In single-attribute setting, the goal is to steer LLM's output along a single attribute direction, such as fairness, safety or truthfulness, with a steering vector tailored to that specific attribute. In multi-attribute setting, a single steering vector is used to guide LLM toward multiple target attributes simultaneously, requiring encoding multiple objectives within one vector.

4.2.1. Performance on Single-attribute Tasks. Tab. 1 presents the experimental results of different methods on three domain tasks, respectively. SRS consistently achieves the highest refusal rate (1.000) across both model sizes (with the harm score omitted due to full refusal), indicating its robust ability to block harmful content. Compared to strong baselines like CAA (0.990 and 0.985 on the two models, respectively) and SAE-TS (0.985 and 0.990), SRS achieves stricter refusal. For the fairness domain, SRS achieves the best fairness scores across both model sizes, reaching 0.965 on 2B and 0.974 on 9B, outperforming all prior baselines. The improvements over CAA (0.958 and 0.967) and SAE-TS (0.955 and 0.965) are notable. In the truthfulness domain, SRS again surpasses competing methods on both models, achieving 0.982 and 0.991 of Truth metric, and 0.993 and 0.994 of Info. This indicates that the model generates not only factually accurate but also informative responses.

In addition, detailed results on impact score of each sparse representation dimension calculated by SRS are shown in Appendix. A.

To intuitively illustrate behavioral differences, Fig. 2 compares responses from `Gemma-2-2B-it` under three configurations, i.e., no control, steering with CAA, and steering with SRS, using a harmful prompt that solicits illegal script generation. Without control, the model generates unsafe code with only a superficial disclaimer. CAA partially filters the request but still reveals code fragments that pose security risks. In contrast, SRS performs a complete behavioral override, rejecting the request outright and returning a structured response that highlights the associated legal, ethical, and security concerns.

Finally, to enhance transparency, we employ Neuronpedia [2] to interpret the top-ranked sparse features identified by SRS. As shown in Tab. 3, the positively associated features (i.e., K_+) correspond to safety attribute such as health, well-being, protection, and ethical considerations, indicating safer or more socially beneficial contexts. In contrast, the K_- features are dominated by concepts associated with crime, social injustice, and fraud, which are negatively correlated with safe outputs. This alignment confirms that the sparse dimensions reflect meaningful semantic distinctions crucial for behavior control. Interpretability results for other domains are included in Appendix. C.

4.2.2. Performance on multi-attribute Tasks. Since most prior steering methods are primarily designed for single-attribute control, we adapt them to the multi-attribute setting as follows. For LAT [44], we independently run the LAT process for each target attribute to obtain its individual steering vector. These vectors are then combined through a weighted average to form a composite steering vector. For

Model	Method	Safety		Fairness	Truthfulness		Side-effect		
		Refusal ↑	Harm ↓	Fairness ↑	Truth ↑	Info ↑	Grammar ↓	Diversity ↑	Utility ↑
Gemma-2-2B	**Base**	0.940	0.489	0.825	0.926	0.808	0.872	8.658	0.579
	LAT	0.960	0.413	0.932	0.943	0.954	1.549	6.752	0.508
	ActAdd	0.975	0.357	0.954	0.947	0.957	0.947	7.941	0.523
	CAA	0.990	0.326	0.958	**0.982**	0.977	1.256	8.181	0.517
	SAE-FS	0.970	0.352	0.951	0.956	0.963	1.042	8.352	0.552
	SAE-TS	0.985	0.327	0.955	0.973	0.971	0.937	8.9425	0.559
	SRS	**1.000**	-	**0.965**	**0.982**	**0.993**	**0.951**	**9.132**	**0.573**
Gemma-2-9B	**Base**	0.970	0.437	0.862	0.945	0.911	0.613	10.132	0.739
	LAT	0.975	0.384	0.919	0.965	0.967	1.113	8.611	0.621
	ActAdd	0.985	0.366	0.942	0.975	0.982	0.962	9.573	0.660
	CAA	0.985	0.312	0.967	0.989	0.992	1.012	10.269	0.656
	SAE-FS	0.985	0.380	0.934	0.972	0.986	0.837	9.878	0.705
	SAE-TS	0.990	0.341	0.965	0.987	0.994	**0.751**	10.314	0.712
	SRS	**1.000**	-	**0.974**	**0.991**	0.994	0.772	**10.476**	**0.735**

TABLE 1: Steering performance comparison of different methods on single-attribute tasks, (e.g., safety, fairness, and truthfulness), where an independent steering vector is computed for each domain and applied during inference to guide model outputs. ↑ means higher is better and ↓ means lower is better. (Note: *Harm* score in safety domain score is computed only on responses that do not refuse the malicious prompt. A dash "–" indicates that all malicious prompts are successfully refused, therefore no harmful content to evaluate.)

both ActAdd [39] and CAA [30], we independently compute the steering direction for each attribute using their respective methods—mean activation difference for ActAdd, and contrastive mean difference for CAA. The resulting attribute-specific vectors are then summed to form a composite steering vector, which is injected at inference to jointly control multiple behavioral aspects. For both SAE-FS [28] and SAE-TS [11], we apply the same steering procedure used in the single-attribute setting to each attribute, and combine these vectors into one. Our proposed method SRS supports multi-attribute steering through the composition methods introduced in Sec. 4.1.4.

The results of multi-attribute cases are shown in Tab. 2. The experimental results clearly demonstrate that our proposed method SRS outperforms existing baselines across all alignment objectives, while maintaining or improving overall generation quality. Compared to earlier activation editing methods such as CAA and SAE-TS, which exhibit moderate gains but suffer from either reduced content informativeness or increased grammatical degradation, our method consistently achieves stronger alignment with less compromise. For instance, on the `Gemma-2-9B-it`, our approach with PCA composition yields the highest refusal rate (0.990), the lowest harmfulness (0.312), and among the best fairness and truthfulness scores (0.972 and 0.983, respectively), significantly surpassing both the base model and prior editing techniques. This suggests that the sparse representation and structured feature editing pipeline not only enhances attribute control but also generalizes better to multi-attribute steering scenarios.

Impact of Composition Strategy. Since SRS first disentangles the features and applies steering in the sparse latent space, we go beyond the conventional linear addition composition (such as linear weighting) and explore several alternative strategies for vector composition in multi-attribute tasks. Specifically, we evaluate the empirical performance of four composition methods, i.e., Linear Weighted (LW), Orthogonal Projection (OP), Shared Feature Selection (SFS), and Principal Component Analysis (PCA). The results are shown in Tab. 2.

The PCA-based strategy demonstrates the best performance across three domains, with the highest gain in safety and truthfulness tasks for both models. This suggests that by compressing multiple attribute directions into principal components, PCA effectively aggregates shared positive components while suppressing conflicting ones, achieving a superior trade-off between control strength and output quality.

Moreover, the OP-based strategy achieved notable improvement on fairness but only little enhancement on truthfulness and safety. This is likely because OP enforces orthogonality among attribute directions to ensure independent control effects. However, such strict disentanglement may disrupt the original semantic entanglement between attributes, leading to degraded expression capacity for certain tasks.

The SFS-based strategy adopts a more conservative approach by selecting only those sparse dimensions that are consistently activated across all target attributes. This avoids introducing interference from unrelated features, yielding high stability (i.e., the lowest variance in content quality across experiments). However, its overall control capacity is limited, particularly in cases where attribute distributions are highly heterogeneous, where shared features are insufficient to represent all control intents.

In summary, the experimental results reveal that the effectiveness of multi-attribute composition strategies funda-

Model	Method	Safety		Fairness	Truthfulness		Side-effect		
		Refusal ↑	Harm ↓	Fairness ↑	Truth ↑	Info ↑	Grammar ↓	Diversity ↑	Utility ↑
Gemma-2-2B	**Base**	0.940	0.489	0.825	0.926	0.808	0.872	8.658	0.579
	LAT	0.945	0.476	0.843	0.933	0.950	1.256	7.133	0.512
	ActAdd	0.955	0.447	0.859	0.921	0.886	0.947	8.235	0.524
	CAA	0.955	0.433	0.877	0.929	0.895	1.015	8.571	0.520
	SAE-FS	0.960	0.412	0.895	0.936	0.923	**0.913**	8.793	0.549
	SAE-TS	0.970	0.378	0.912	0.943	0.931	0.935	8.726	0.567
	SRS_{LW}	0.975	0.387	0.922	0.933	0.936	0.926	8.512	0.570
	SRS_{OP}	0.975	0.419	**0.944**	0.928	0.931	0.927	**8.976**	0.575
	SRS_{SFS}	0.970	0.436	0.915	0.924	0.927	1.043	8.586	0.569
	SRS_{PCA}	**0.980**	**0.365**	0.934	**0.946**	**0.954**	0.951	8.943	**0.577**
Gemma-2-9B	**Base**	0.970	0.437	0.862	0.945	0.911	0.613	10.132	0.739
	LAT	0.970	0.403	0.868	0.947	0.919	1.089	8.659	0.625
	ActAdd	0.975	0.392	0.874	0.958	0.923	0.751	9.437	0.656
	CAA	0.975	0.384	0.897	0.954	0.834	0.964	9.754	0.658
	SAE-FS	0.975	0.380	0.934	0.956	0.962	0.723	10.362	0.702
	SAE-TS	0.985	0.337	0.965	0.977	0.975	0.774	10.829	0.707
	SRS_{LW}	0.985	0.334	**0.974**	0.981	0.994	0.701	**10.766**	0.713
	SRS_{OP}	0.980	0.357	0.969	0.977	0.987	0.686	10.427	0.720
	SRS_{SFS}	0.980	0.346	0.965	0.978	**0.992**	0.712	10.046	0.708
	SRS_{PCA}	**0.990**	**0.312**	0.972	**0.983**	0.987	**0.631**	10.536	**0.731**

TABLE 2: Steering performance comparison of different methods on multi-attribute tasks, where a shared steering vector is computed from the three domains (e.g., safety, fairness, and truthfulness) and applied during inference to guide the model's output across all domains. Our proposed framework (SRS) is evaluated under four distinct composition strategies, i.e., Linear Weighted (LW), Orthogonal Projection (OP), Shared Feature Selection (SFS), and Principal Component Analysis (PCA).

mentally depends on how well they model the structural relationships among attributes. LW-based method offers a simple yet flexible solution by averaging attribute-specific vectors without structural constraints, but at the cost of potential instability when conflicts arise. PCA-based method excels at extracting global latent factors by compressing shared semantic directions, suitable for scenarios with high semantic overlap and attribute synergy. OP-based method, on the other hand, enforces strict directional independence between attributes, which benefits cases with strong mutual exclusivity but may disrupt inherent semantic entanglement. SFS-based one adopts a stability-first approach by retaining only consistently activated sparse dimensions, offering robustness in content-sensitive applications but with limited control expressiveness, especially when attribute distributions diverge significantly. Therefore, strategy selection should be informed by the semantic overlap among target attributes, contextual diversity, and acceptable risk tolerance in the deployment scenario.

4.3. Safeguard Robustness under Jailbreaks

This experiment aims to investigate whether the learned steering vectors truly capture task-level semantics (e.g., "safety enhancement") or merely overfit to specific prompt phrasings. Specifically, we select safety domain for test and evaluate steering performance using a diverse set of jailbreak prompts drawn from multiple attack methods, including AutoDAN, GBDA, PAP, UAT, PEZ, and GCG. These prompts all contain harmful instructions, but differ significantly in expression style, often using indirect or obfuscated language to bypass refusal mechanisms. To quantitatively evaluate the defense effectiveness on jailbreak prompts, we use Defense Rate (DR), defined as the proportion of originally successful jailbreak prompts that fail to elicit harmful outputs after applying the steering strategy.

As shown in Fig. 3, our proposed method SRS achieves the highest DSR across all attack methods, with an average rate exceeding 81%. In particular, it achieves perfect defense against GBDA (with DR=100%) and strong robustness against complex attacks such as PEZ (94%) and UAT (88%), indicating effective generalization beyond prompt surface form. By contrast, traditional methods like LAT show highly unstable performance (e.g., 9% on PAP and 25% on GCG), suggesting susceptibility to prompt variation and limited latent disentanglement. Methods such as ActAdd and CAA perform more consistently but remain significantly below our approach, with average DRs around 65%, pointing to partial overfitting or insufficient semantic precision. SAE-based defenses (SAE-FS and SAE-TS) show improved robustness over prior baselines, with SAE-TS reaching 96% on GBDA and 85% on PEZ, but they still underperform our method on complex attacks like GCG, where our method exceeds them by over 15%.

These results show that our method remains effective even when facing with adversarial prompts with different styles and phrasings, indicating that the steering vector learned by SRS captures the underlying harmful intent rather than