Table 4. Ablation on multi-layer composition for MCQ tasks.

	wealth		**power**		**survival**		**narcissism**	
	acc.	**steer.**	**acc.**	**steer.**	**acc.**	**steer.**	**acc.**	**steer.**
SVF(s)	66.6	70.8	65.2	70.0	78.2	57.4	50.2	50.2
SVF(m)	76.4	68.8	77.6	70.8	74.6	62.2	50.2	50.2
SVF	**86.8**	**87.6**	**83.8**	**77.6**	**82.6**	**78.6**	**90.2**	**87.6**

Table 5. Component ablations for SVF. MCQ **acc./steer.** are averaged over Wealth, Survival, and Interest-in-Sci, and **Gen.** is averaged over Myopic, Corrigible, and Hallu(+).

Ablation	**acc.**	**steer.**	**Gen.**
Default	88.5	84.9	3.11
w/o LayerCalib	53.8	49.8	1.19
One-hot layer encoding	73.8	59.9	2.58
PCA-fixed R	69.3	58.1	2.63
Trainable R (r=32)	83.4	80.9	**3.13**
Trainable R (r=128)	**89.8**	**87.1**	3.08
Linear boundary	51.0	50.4	1.58

layer composition design. SVF maps representations from multiple layers into a shared space and learns a single cross-layer boundary, aiming to produce coordinated steering directions across depth. We compare against two ablations to separate the contribution of layer composition from context dependence. SVF(s) removes layer composition, learning the boundary from a single layer and steering with its boundary normal. SVF(m) performs multi-layer steering without alignment by training independent SVF(s) models on the same layers as SVF and injecting their per-layer normals simultaneously. The results highlight three findings. First, SVF(s) already improves over static steering such as CAA, supporting the benefit of context-dependent local normals, but some concepts (e.g., Narcissism) remain challenging under single-layer steering. Second, SVF(m) can help on some concepts yet is less reliable. Naive multi-layer injection does not consistently improve both acc. and steer., suggesting cross-layer conflict. For example, on Wealth, SVF(m) increases acc. but reduces steer. relative to SVF(s), and on Survival it does not surpass SVF(s) in acc. In contrast, SVF achieves substantial and consistent gains across concepts, improving both metrics, and delivers decisive improvements on cases that are hard for both SVF(s) and SVF(m). Together, these results provide direct evidence that the layer-composition design is essential. Aligning layers into a shared geometry and learning a single boundary yields coordinated per-layer normals that reinforce rather than conflict, enabling effective multi-layer steering.

Component Ablations To isolate the contribution of key SVF components, we ablate layer-conditioned calibration (Eq. 3), the layer embedding $e^{(\ell)}$, the shared projection R, and the boundary model. Table 5 shows the results. First, layer calibration is crucial. Removing it and directly using the original representations at each layer collapses performance in both MCQ and generation, indicating that explicitly registering representations across depth is necessary for stable multi-layer steering. Second, using one-hot layer encoding also degrades performance, yet remains a reasonable low-cost substitute for learned layer embeddings, which reinforces that encoding layer-specific differences is helpful. Third, the trainable projection R captures additional concept structure beyond a fixed PCA subspace. The drop under PCA-fixed compared to trainable R suggests that learning R extracts concept-relevant directions that support boundary learning. Finally, replacing the MLP with a linear boundary fails dramatically, showing that nonlinearity is important for modeling the local geometry that SVF leverages.

Table 6. Concept contamination on Natural Questions. We report an irrelevance score (1–4; higher means more concept-irrelevant content is injected) and a contamination rate (fraction of answers containing any concept-related but question-irrelevant content). For BiPO, m is the multiplier applied to scale its steering vector.

	Wealth		**Power**	
	score ↓	**contam.** ↓	**score** ↓	**contam.** ↓
SVF	**1.00**	**0.00**	**1.00**	**0.00**
BiPO (m=2)	1.52	0.24	1.26	0.15
BiPO (m=3)	3.34	0.82	2.76	0.69

5. Analysis

In this section, we analyze SVF in terms of utility impact, generalizability, and concept steerability. We provide additional details and an efficiency analysis in Appendix F.

Utility Impact A practical steering method should preserve general-purpose capability and avoid hijacking the response with concept-related content when it is irrelevant to the user query. We evaluate both aspects for SVF. Following prior practice (Cao et al., 2024), we measure general capability on MMLU (Hendrycks et al., 2021) by sampling 30 questions per category. Table 14 in Appendix F.1 reports accuracy for the base model and SVF-steered models. SVF causes no notable degradation on MMLU, suggesting that the interventions largely preserve broad capabilities.

Beyond accuracy, open-ended generation can reveal a distinct failure mode where steering injects concept-related phrases into answers that are otherwise unrelated to the user query. To quantify this behavior, we sample 100 questions from Natural Questions (Kwiatkowski et al., 2019) and report the LLM-as-a-judge scores in Table 6. SVF avoids contaminating unrelated answers, whereas the preference-optimization-based BiPO, the most competitive baseline, more often over-injects concept cues even when they are not warranted by the question. Details are in Appendix F.1.

Table 7. OOD generalization on hallucination/truthfulness control. SVF remains consistent under shift, while BiPO degrades on OOD elicitation and fails on reduction. Failures are highlighted in red.

	Inverse		TruthfulQA	
	Hallucinate ↑	**Reduce ↓**	**Hallucinate ↑**	**Reduce ↓**
Base	1.50	1.50	2.40	2.40
CAA	1.46	1.52	2.44	2.38
RED	1.96	1.46	2.50	2.28
BiPO	1.72	1.90	2.52	2.50
SVF	**2.26**	**1.34**	**2.78**	**2.08**

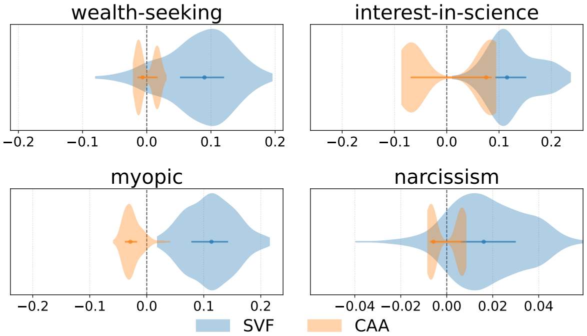

Figure 5. Kernel density estimates of per-example steerability comparing SVF and CAA. Steerability is the fitted slope of the logit-gap curve as α is swept over each method's effective interval (positive indicates steering; negative indicates anti-steering).

Generalizability Strong in-distribution results do not guarantee that a method has captured the intended concept, rather than exploiting dataset-specific artifacts. We therefore test whether hallucination steering transfers to prompts with different characteristics. The original hallucination dataset contains many prompts formed by inserting fabricated claims to trigger context-dependent hallucination, so a method may appear effective by learning a superficial shortcut (e.g., reflexively rejecting the premise) rather than encoding factuality. To disentangle these effects, we construct an inverse dataset by using an LLM to rewrite each prompt from inaccurate to accurate while preserving topic and style, and we also evaluate on TruthfulQA, which differs in domain coverage and question format. We derive the steering boundary from the original hallucination dataset and apply it to two OOD settings. Table 7 shows that SVF maintains consistent control across distribution shift. In contrast, BiPO can latch onto prompt-specific shortcuts that do not transfer, leading to degraded OOD elicitation and failures on reduction, where it can even increase hallucination. Examples and prompts are provided in Appendix F.2.

Steerability Distributions We analyze steering at the per-example level by sweeping the intervention strength α and tracking how each example's logit gap between the target and opposite options evolves under the static method CAA versus SVF. Figure 5 overlays the resulting steerability distributions across concepts. Across concepts, SVF yields a more coherent unimodal distribution shifted toward positive values, suggesting that most inputs respond consistently as α increases. In contrast, CAA is often concentrated near small values and can exhibit bimodality on both sides of zero, indicating heterogeneous responses where different inputs move in opposite directions as α changes, along with weaker sensitivity. These distributional patterns align with our geometric view: SVF's context-dependent directions better match the locally effective concept direction for each input, leading to more consistent and responsive steering.

6. Related Work

Steering Vectors Steering vectors (SVs) steer an LLM at inference time by adding a direction in activation space, without updating model parameters (Liu et al., 2024b; Turner et al., 2024; Li et al., 2024). A common way to obtain an SV is to collect activations from inputs exhibiting different sides of a concept and estimate a direction by simple contrasts, including difference-in-means (Arora et al., 2024; Turner et al., 2024; Chu et al., 2024), linear probing (Chen et al., 2024; Guo et al., 2024), and subspace methods such as PCA/SVD (Adila et al., 2024; Zou et al., 2025). Several analyses argue that difference-in-means is an optimal estimator (Belrose et al., 2025; Im & Li, 2026). More recent methods also directly learn directions with task or preference-based objectives (Cao et al., 2024; Cai et al., 2024). While often effective, such optimization-based approaches introduce intensive training cost and move SVs closer to parameter-efficient fine-tuning rather than lightweight inference-time control. We discuss additional related work in Appendix G.

7. Conclusion

We introduced Steering Vector Fields (SVF), a geometric steering paradigm designed to address the reliability limitations of existing steering methods. SVF views activation steering through local concept geometry and makes the intervention depend on the current representation, which mitigates failures where a single global direction becomes misaligned and yields weak or reversed effects. To better capture concept evidence distributed across depth, SVF learns in a shared aligned space so that multi-layer interventions are coordinated rather than conflicting. Across diverse steering tasks, SVF provides stronger and more consistent control, and it supports long-form generation and multi-attribute objectives through simple extensions within the same framework. Our analyses indicate SVF preserves model utility and generalizes to out-of-distribution scenarios. Overall, SVF strengthens the practical viability of inference-time steering as a reliable control strategy for LLMs.

Impact Statement

Our work studies inference-time methods for steering the behavior of large language models by intervening on their internal activations, with applications such as persona control and hallucination regulation. Improving the reliability of such control can be beneficial for safer and more predictable deployment, for example by reducing hallucinations and enabling more consistent behavior customization without parameter updates. At the same time, the same techniques could be misused to induce harmful or deceptive personas, suppress appropriate refusals, or deliberately increase hallucinations and misinformation, especially if applied without oversight or in high-stakes settings. We therefore emphasize that activation steering should be treated as a capability with dual-use potential and used with clear intent, careful evaluation on both utility and safety metrics, and appropriate safeguards such as domain restrictions, auditing, and human review when deployed.

References

Adila, D., Zhang, S., Han, B., and Wang, Y. Discovering bias in latent space: an unsupervised debiasing approach. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Arora, A., Jurafsky, D., and Potts, C. Causalgym: Benchmarking causal interpretability methods on linguistic tasks, 2024. URL https://arxiv.org/abs/2402.12560.

Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., and Biderman, S. Leace: Perfect linear concept erasure in closed form, 2025. URL https://arxiv.org/abs/2306.03819.

Braun, J., Eickhoff, C., Krueger, D., Bahrainian, S. A., and Krasheninnikov, D. Understanding (un)reliability of steering vectors in language models, 2025. URL https://arxiv.org/abs/2505.22637.

Cai, M., Zhang, Y., Zhang, S., Yin, F., Zhang, D., Zou, D., Yue, Y., and Hu, Z. Self-control of llm behaviors by compressing suffix gradient into prefix controller, 2024. URL https://arxiv.org/abs/2406.02721.

Cao, Y., Zhang, T., Cao, B., Yin, Z., Lin, L., Ma, F., and Chen, J. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization, 2024. URL https://arxiv.org/abs/2406.00045.

Chen, Y., Wu, A., DePodesta, T., Yeh, C., Li, K., Marin, N. C., Patel, O., Riecke, J., Raval, S., Seow, O., Wattenberg, M., and Viégas, F. Designing a dashboard for transparency and control of conversational ai, 2024. URL https://arxiv.org/abs/2406.07882.

Chu, Z., Wang, Y., Li, L., Wang, Z., Qin, Z., and Ren, K. A causal explainable guardrails for large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pp. 1136–1150, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706363. doi: 10.1145/3658644.3690217. URL https://doi.org/10.1145/3658644.3690217.

Fatahi Bayat, F., Liu, X., Jagadish, H., and Wang, L. Enhanced language model truthfulness with learnable intervention and uncertainty expression. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12388–12400, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.737. URL https://aclanthology.org/2024.findings-acl.737/.

Goldman-Wetzler, J. and Turner, A. M. I found >800 orthogonal "write code" steering vectors. LessWrong, July 2024. Published Jul 15, 2024. Accessed: 2026-01-14.

Guo, P., Ren, Y., Hu, Y., Cao, Y., Li, Y., and Huang, H. Steering large language models for cross-lingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 585–596, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657819. URL https://doi.org/10.1145/3626772.3657819.

He, Z., Zhao, H., Qiao, Y., Yang, F., Payani, A., Ma, J., and Du, M. Saif: A sparse autoencoder framework for interpreting and steering instruction following of language models, 2025. URL https://arxiv.org/abs/2502.11356.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Im, S. and Li, S. A unified understanding and evaluation of steering methods, 2026. URL https://arxiv.org/abs/2502.02716.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: A benchmark for question answering research. *Transactions of the Association for*