Computational Linguistics, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model, 2024. URL https://arxiv.org/abs/2306.03341.

Li, Y., Cao, Y., He, H., Cheng, Q., Fu, X., Xiao, X., Wang, T., and Tang, R. M²IV: Towards efficient and fine-grained multimodal in-context learning via representation engineering. In *Second Conference on Language Modeling*, 2025a. URL https://openreview.net/forum?id=9ffYcEiNw9.

Li, Z., Xu, Z., Han, L., Gao, Y., Wen, S., Liu, D., Wang, H., and Metaxas, D. N. Implicit in-context learning, 2025b. URL https://arxiv.org/abs/2405.14660.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL https://arxiv.org/abs/2109.07958.

Lindsey, J., Templeton, A., Marcus, J., Conerly, T., Batson, J., and Olah, C. Sparse crosscoders for cross-layer features and model diffing. https://transformer-circuits.pub/2024/crosscoders/index.html, October 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html. Transformer Circuits Thread (online article), published October 2024.

Liu, S., Ye, H., Xing, L., and Zou, J. In-context vectors: Making in context learning more effective and controllable through latent space steering, 2024a. URL https://arxiv.org/abs/2311.06668.

Liu, W., Wang, X., Wu, M., Li, T., Lv, C., Ling, Z., Zhu, J., Zhang, C., Zheng, X., and Huang, X. Aligning large language models with human preferences through representation engineering, 2024b. URL https://arxiv.org/abs/2312.15997.

Mayne, H., Yang, Y., and Mahdi, A. Can sparse autoencoders be used to decompose and interpret steering vectors?, 2024. URL https://arxiv.org/abs/2411.08790.

Nguyen, D., Prasad, A., Stengel-Eskin, E., and Bansal, M. Multi-attribute steering of language models via targeted intervention, 2025. URL https://arxiv.org/abs/2502.12446.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering llama 2 via contrastive activation addition, 2024. URL https://arxiv.org/abs/2312.06681.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. Film: Visual reasoning with a general conditioning layer, 2017. URL https://arxiv.org/abs/1709.07871.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Khundadze, G., Kernion, J., Landis, J., Kerr, J., Mueller, J., Hyun, J., Landau, J., Ndousse, K., Goldberg, L., Lovitt, L., Lucas, M., Sellitto, M., Zhang, M., Kingsland, N., Elhage, N., Joseph, N., Mercado, N., DasSarma, N., Rausch, O., Larson, R., McCandlish, S., Johnston, S., Kravec, S., El Showk, S., Lanham, T., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Clark, J., Bowman, S. R., Askell, A., Grosse, R., Hernandez, D., Ganguli, D., Hubinger, E., Schiefer, N., and Kaplan, J. Discovering language model behaviors with model-written evaluations, 2022. URL https://arxiv.org/abs/2212.09251.

Pres, I., Ruis, L., Lubana, E. S., and Krueger, D. Towards reliable evaluation of behavior steering interventions in llms, 2024. URL https://arxiv.org/abs/2410.17245.

Scalena, D., Sarti, G., and Nissim, M. Multi-property steering of large language models with dynamic activation composition. In Belinkov, Y., Kim, N., Jumelet, J., Mohebbi, H., Mueller, A., and Chen, H. (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 577–603, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.34. URL https://aclanthology.org/2024.blackboxnlp-1.34/.

Tan, D., Chanin, D., Lynch, A., Kanoulas, D., Paige, B., Garriga-Alonso, A., and Kirk, R. Analyzing the generalization and reliability of steering vectors, 2025. URL https://arxiv.org/abs/2407.12404.

Tanneru, S. H., Ley, D., Agarwal, C., and Lakkaraju, H. On the hardness of faithful chain-of-thought reasoning in large language models, 2024. URL https://arxiv.org/abs/2406.10625.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

van der Weij, T., Poesio, M., and Schoots, N. Extending activation steering to broad skills and multiple behaviours, 2024. URL https://arxiv.org/abs/2403.05767.

Wehner, J., Abdelnabi, S., Tan, D., Krueger, D., and Fritz, M. Taxonomy, opportunities, and challenges of representation engineering for large language models, 2025. URL https://arxiv.org/abs/2502.19649.

Weng, Y., He, S., Liu, K., Liu, S., and Zhao, J. Controllm: Crafting diverse personalities for language models, 2024. URL https://arxiv.org/abs/2402.10151.

Wu, M., Liu, W., Wang, X., Li, T., Lv, C., Ling, Z., JianHao, Z., Zhang, C., Zheng, X., and Huang, X. Advancing parameter efficiency in fine-tuning via representation editing. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13445–13464, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.726. URL https://aclanthology.org/2024.acl-long.726/.

Wu, Z., Arora, A., Wang, Z., Geiger, A., Jurafsky, D., Manning, C. D., and Potts, C. Reft: Representation finetuning for language models, 2024b. URL https://arxiv.org/abs/2404.03592.

Wu, Z., Arora, A., Geiger, A., Wang, Z., Huang, J., Jurafsky, D., Manning, C. D., and Potts, C. Axbench: Steering llms? even simple baselines outperform sparse autoencoders, 2025. URL https://arxiv.org/abs/2501.17148.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yin, F., Ye, X., and Durrett, G. Lofit: Localized fine-tuning on llm representations, 2024. URL https://arxiv.org/abs/2406.01563.

Zhang, B. and Sennrich, R. Root mean square layer normalization, 2019. URL https://arxiv.org/abs/1910.07467.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A. Softmin Composition for Multi-Attribute Steering

Given concept scorers $\{f_i\}_{i=1}^m$, we define the target region for each attribute as $\mathcal{R}_i = \{h : f_i(h) > 0\}$. The conjunction corresponds to the intersection $\mathcal{R}_\wedge = \bigcap_{i=1}^m \mathcal{R}_i = \{h : \forall i,\ f_i(h) > 0\}$. The hard minimum $g(h) = \min_i f_i(h)$ satisfies $g(h) > 0$ iff $\forall i,\ f_i(h) > 0$, but is non-differentiable at ties. We therefore use the smooth softmin,

$$g_\tau(h) \;=\; -\tau \log \sum_{i=1}^{m} \exp\left(-\frac{f_i(h)}{\tau}\right), \tag{5}$$

which converges to $\min_i f_i(h)$ as $\tau \to 0$.

Let $Z(h) = \sum_{i=1}^m \exp(-f_i(h)/\tau)$. Differentiating (5) gives

$$\nabla g_\tau(h) \;=\; \sum_{i=1}^{m} w_i(h)\, \nabla f_i(h), \qquad w_i(h) = \frac{\exp(-f_i(h)/\tau)}{Z(h)}. \tag{6}$$

Hence, attributes with smaller $f_i(h)$ receive larger weights, so the composed direction emphasizes the currently limiting constraint, which stabilizes multi-attribute steering without manual reweighting.

In SVF, steering directions are computed from the local normal of the chosen score function. For multi-attribute steering, we replace the single-concept score $f(\cdot)$ with $g_\tau(\cdot)$ from (5), and compute the steering direction based on it.

B. Additional Experiment Settings

B.1. Configurations

Following Tan et al. (2025), we use a 40/10/50 train/validation/test split for the datasets. To collect representations for training, we flatten each multiple-choice instance by pairing the question with each option independently. The target response and the opposite response are appended to the same question as separate samples. We then run a forward pass and extract the last-token hidden representation as the training feature for each sample. For normalization, we apply RMS normalization (Zhang & Sennrich, 2019) as the first step in our pipeline.

For the projection module in $u^{(\ell)} = R\hat{h}^{(\ell)}$, we initialize the trainable projection matrix with PCA computed on the pooled hidden representations from the selected training layers, and project representations into a 64-dimensional subspace. For the FiLM-style modulation in Eq. 3, the layer embeddings are 8-dimensional vectors initialized randomly. The associated linear projectors are also randomly initialized. The MLP used to train the boundary has a single hidden layer with 64 hidden units.

We use hidden representations from layers 15–24 for Llama-2-7b-Chat-hf and layers 20–29 for Qwen3-14b during training, selected via validation over contiguous 10-layer windows. We train for 5 epochs using AdamW with a learning rate of 3e-4 and a weight decay of 1e-2 for regularization.

At inference time, we extract the last-token representation of the prompt and compute the steering direction from the learned boundary. We inject the resulting steering vector into the last-token representation with a scaling factor that is selected based on the validation split. Across our tasks, the best-performing values typically fall in the range 30–50. To keep interventions lightweight, we steer only a contiguous 4-layer window instead of all layers in the 10-layer stage. The window is chosen by a small validation sweep. The intervention is applied to layers 15-18 for Llama-2-7b-Chat-hf and layers 20–23 for Qwen3-14b. For open-ended generation tasks, the refresh window K of SVF is set to 1. We use greedy decoding with a maximum of 128 new tokens.

B.2. Datasets

Model-Written Evaluations Datasets Anthropic's Model-Written Evaluations (MWE) (Perez et al., 2022) is a large benchmark suite of over 100 categories designed to probe model personas and behavioral tendencies. Each dataset contains 1000 examples, where each example includes a prompt and two candidate continuations: a target response and an opposite response in an A/B format. The answer options vary in length across categories, ranging from long continuations (e.g., wealth-seeking and myopic) to short responses (e.g., interest-in-science and narcissism). We include categories from both regimes to test SVF under different conditions.