STEERING LLMS FOR MULTI-AGENT DECISION-MAKING USING REPRESENTATION LEARNING

Dom Huh
University of California, Davis
dhuh@ucdavis.edu

Prasant Mohapatra
University of South Florida
pmohapatra@usf.edu

ABSTRACT

Activation steering offers a lightweight mechanism for controlling large language models (LLMs), but existing approaches have yet been integrated within strategic multi-agent decision-making settings. In this work, we propose a representation learning framework for activation steering tailored to multi-agent decision-making, optimizing steering representations directly from interaction trajectories by grounding latent variables in multi-agent dynamics and enforcing latent self-consistency over time. Our approach disentangles latent factors underlying strategic interaction, enabling fine-grained behavioral control without modifying model parameters or relying on task-specific supervision but on the nature of the multi-agent dynamics. We evaluate our method on γ-Bench, a diverse suite of cooperative, competitive, and mixed-motive games, and demonstrate consistent improvements in social and strategic performance across multiple open-source LLM families. These results suggest that representation learning provides a scalable and interpretable foundation for activation steering in multi-agent systems.

1 INTRODUCTION

Large language models (LLMs) have achieved impressive performance across reasoning, planning, and interaction-heavy tasks (Xi et al., 2025). As a result, they are increasingly deployed as autonomous agents operating in multi-agent environments, where success depends not only on individual reasoning ability but also on strategic adaptation, coordination, and robustness to non-stationary opponents (Guo et al., 2024; Huang et al., 2024). However, effectively steering LLM behavior in such settings remains a fundamental challenge.

Most existing approaches rely on prompt engineering or fine-tuning to induce desired behaviors. While effective in isolated or static tasks, these methods scale poorly to multi-agent settings, where several agents must be modeled to adapt to their own unique actions, beliefs, and incentives of others. Prompt-based control is brittle to distributional shifts and interaction effects, while fine-tuning is costly, inflexible, and difficult to apply when multiple agents or objectives are involved. These limitations motivate alternative mechanisms for controlling LLM behavior that are both lightweight and sensitive to interaction dynamics.

Activation steering, which modifies internal model activations at inference time, has recently emerged as a promising alternative to parameter updates. Prior work has shown that steering vectors extracted from contrastive data or in-context examples can reliably influence model behavior without retraining (Hendel et al., 2023; Liu et al., 2024; Panickssery et al., 2024; Zou et al., 2025; Lee et al., 2024). However, existing activation steering methods have little regard for the underlying multi-agent-specific dynamics. As a result, these approaches struggle to generalize to strategic multi-agent environments, where behavior is shaped by latent beliefs, temporal dependencies, and inter-agent reasoning.

In this paper, we argue that effective activation steering for multi-agent decision-making requires learning structured, task-grounded representations to guide and construct a meaningful search space for residual vectors. Thereby, we propose a representation learning framework that learns steering vectors directly from multi-agent interaction trajectories. Our approach grounds latent steering representations in the environment dynamics by reconstructing observations, actions, and rewards from each agent's perspective, while simultaneously enforcing latent self-consistency across time

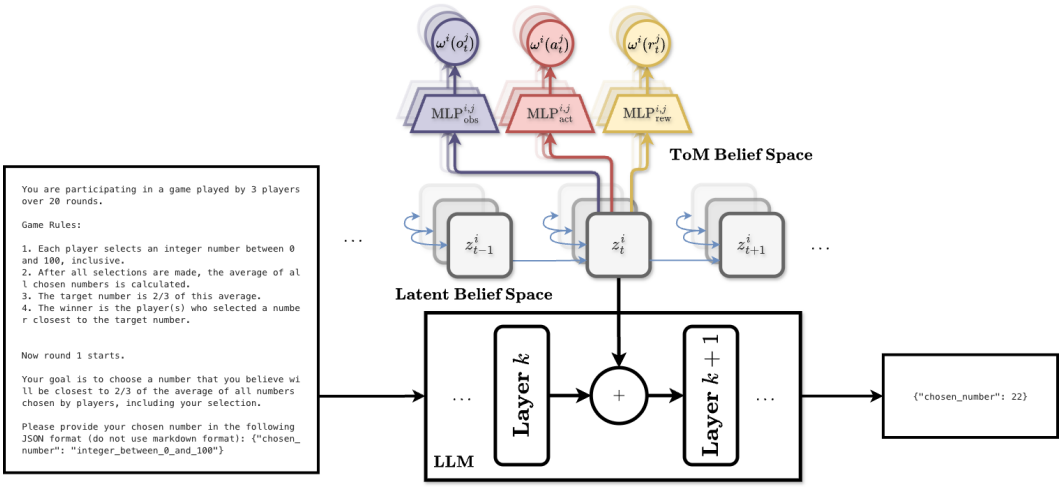


Figure 1: An illustration of our representation learning process applied to the steering vector z_t^i for agent i at timestep t. The latent belief space depicts the self-consistency that is enforced within the latent space dynamics and the theory-of-mind (ToM) belief space shows the grounding of steering vectors to the multi-agent environment, including the observation o, action a and rewards r.

and between agents. This induces a latent belief space that captures strategic structure, including theory-of-mind–like representations, and enables fine-grained behavioral modulation through activation steering alone.

We evaluate our method on γ-Bench (Huang et al., 2024), a challenging benchmark of cooperative, competitive, and mixed-motive games designed to probe social reasoning and strategic adaptation. Across multiple open-source LLM families and game types, our representation learning–based steering consistently outperforms both unsteered models and initialization-only steering baselines. These gains are especially pronounced in non-stationary and adversarial settings, where heuristic steering methods often fail. We provide empirical evidence that representation-driven steering scales across models and strategic settings. More broadly, our results suggest that representation learning offers a path toward unifying activation steering with multi-agent learning and decision theory, moving LLM controllability beyond prompt-level heuristics toward structured, task-aware control.

2 METHOD

We extend the representation learning paradigm introduced in Huh and Mohapatra (2024) to enrich the steering vector that stimulates the activations of the LLM agent in a self-supervised learning manner. As shown in Figure 1, this process consists of two central themes: enforcing latent self-consistency and grounding latent variables to the task dynamics. To realize these ideas, we implement the proposed objectives to optimize a set of steering vectors directly.

Initialization We first initialize a set of steering vectors $\{z_0 \dots z_{k-1}\}_{i \in N}$ over k selected middle layers of the LLMs for each agent i using contrastive activations addition (CAA) (Panickssery et al., 2024) on a curated dataset D of interaction trajectories τ collected using pre-trained LLMs using dataset aggregation and statistical rejection sampling (Liu et al., 2023).

$$\tau = \{o_0, a_0, r_0, \dots o_T, a_T, r_T\}_{i \in N} \tag{1}$$

where N is the set of all agents and T is the terminal time-step.

Instead of directly labeling samples as positive and negative, we normalize the expected returns over the trajectories on the sampled batch and introduced a dynamic threshold value to label the upper p-percentile as positive, where p is a hyperparameter tuned during experimentation, to compute the mean difference vector. For each LLM model, we ran layer-wise analysis to target activations that were most positively affected to the performance using grid-search (Lee et al., 2024). For MoE

Table 1: Performance of different LLMs comparing steering vector methods on γ-Bench, where Base refers to no steering, Init. refers to initialization only, and Repr. refers to our Representation Learning approach.

γ-Bench Scores	LLaMA-4			Gemma-3			Qwen-3		
	Base	Init.	Repr.	Base	Init.	Repr.	Base	Init.	Repr.
Guess 2/3 of the Average	83.6	84.5	87.0	87.0	91.6	93.2	41.8	56.4	59.3
El Farol Bar	70.0	68.5	70.0	59.7	64.3	68.5	23.0	33.5	33.5
Divide the Dollar	79.0	88.4	90.3	87.6	87.6	94.9	56.4	64.3	65.0
Public Goods Game	52.5	52.5	68.5	48.8	51.2	52.5	11.6	20.0	27.6
Diner's Dilemma	43.5	43.5	44.8	37.8	39.6	40.0	1.1	3.2	3.2
Sealed-Bid Auction	13.2	26.9	31.6	7.6	24.2	27.1	2.5	5.6	7.9
Battle Royale	65.5	86.8	92.7	35.8	77.7	81.0	12.8	18.9	19.5
Pirate Game	78.1	85.4	94.3	78.8	86.1	88.2	57.4	60.3	61.2

architectures, our empirical experimentation supports restricting steering layers within the gating networks, proving most stable and effective.

Representation Learning Given the initialized steering vectors, we apply two learning objectives: belief space construction $\mathcal{L}_{bc}$ and latent self-consistency $\mathcal{L}_{sc}$. For belief space reconstruction, we define a belief space ω^i for each agent i as a function of its steering vectors, where $\omega^i(o^j), \omega^i(a^j), \omega^i(r^j)$ represents agent i's internal models of other agent j's observation, action and reward respectively. Hence, for this objective, we reconstruct the multi-agent environment from the local perspective of each agent using their steering vectors, effectively instilling a ToM belief space at the latent representation level.

$$\mathcal{L}_{bc} = \mathbb{E}_{\tau \sim D}[-\ln P(\omega^i(\tau) = \tau)] \tag{2}$$

On the other hand, latent self-consistency enforces a consistency within the joint steering vectors space, achieved through applying forward and inverse dynamic coherency as well as an inter-predictive property within the joint space. In practice, this objective is realized using three auxiliary tasks using contrastive learning techniques.

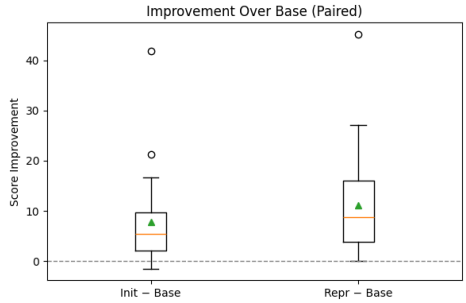
$$\mathcal{L}_{sc} = \mathbb{E}_{\tau \sim D}[-\ln P(z_{t+1}|z_t) - \ln P(a_t|z_t, z_{t+1}) - \ln P(z_t|m(z_t)] \tag{3}$$

where $m(z_t)$ applies a masking over steering vectors along the agent dimension. We follow prescribed design choices from MA-LSO (Huh and Mohapatra, 2024), such as the Bayesian network representation and the scaling parameters. We optimize this objective using gradient learning (Oozeer et al., 2025) using the same RSO dataset aggregation framework as in the initialization step.

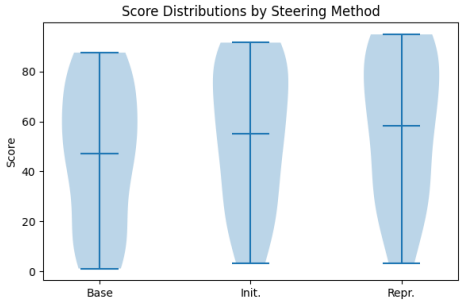
3 Results

Evaluation Benchmark. We evaluate our approach on γ-Bench (Huang et al., 2024), a suite of multi-agent strategic games designed to probe coordination, adaptation, and robustness under non-stationary opponents. γ-Bench spans both cooperative and competitive settings, including matrix games, sequential social dilemmas, and partial-information environments. We test using the default settings provided, and adopt the scoring scheme with score standardization proposed in the original work. Our selection of OSS LLM models include: Llama-4 Maverick 17B-128E Instruct FP8 (AI, 2025), Gemma-3 12B-IT (Kamath et al., 2025) and Qwen-3 4B-IT (Team, 2025).

Performance Analysis We compare our representation learning activation steering against the base LLM with no steering and with initialization steering only. From Table 1, our results suggest notable and consistent improvements from representation learning across all three model families and nearly all games. In cooperative and mixed-motive settings such as Public Goods, Divide the Dollar, and Pirate Game, representation learning yields substantial gains over both the base and initialization baselines, indicating improved coordination and social reasoning. In competitive and non-stationary environments, including Battle Royale and Sealed-Bid Auction, representation learning achieves the largest relative improvements, particularly for smaller models such as Qwen-3, where initialization alone provides limited benefits. While initialization steering occasionally improves



(a) Box plot of paired score differences.



(b) Violin plot over score distribution.

performance, its effects are inconsistent and in some cases neutral or negative (e.g., El Farol Bar for LLaMA-4), whereas representation learning reliably matches or exceeds the best baseline in every task. Overall, these results demonstrate that learning task-relevant steering representations provides a more robust and scalable mechanism for adapting LLM behavior in multi-agent strategic settings than initialization-based approaches alone.

From Figure 2a, the observed trends directly compares the improvement from the baseline to both the initialization and representation learning methods, emphasizes the consistency of performance gains across tasks. Figure 2b illustrates the distribution of scores for each steering method, revealing the underlying shape and density of the data, highlighting similar variability but slightly higher central tendency with our representation learning approach.

4 Conclusion

We introduced a representation learning framework for activation steering that enables principled, fine-grained control of LLM behavior in multi-agent decision-making settings. Rather than relying on heuristic steering directions, prompt engineering, or task-specific supervision, our approach learns latent steering vectors directly from interaction trajectories by grounding them in multi-agent dynamics and enforcing latent self-consistency. This allows behavioral control to be applied entirely at inference time, without modifying model parameters or reward functions. More broadly, our findings suggest that representation learning offers a scalable path toward steering LLM-based agents in complex multi-agent systems. By aligning internal activations with task-grounded latent structures, activation steering can move beyond ad hoc control toward a theoretically grounded tool for coordination, robustness, and interpretability. We view this work as a step toward unifying representation learning, multi-agent learning, and LLM controllability, opening the door to future work on compositional steering, formal guarantees, and deployment in real-world multi-agent settings.

References

Meta AI. Llama-4: Multimodal intelligence, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges, 2024. URL https://arxiv.org/abs/2402.01680.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors, 2023. URL https://arxiv.org/abs/2310.15916.

Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael R Lyu. How far are we on the decision-making of llms? evaluating llms' gaming ability in multi-agent environments. *arXiv preprint arXiv:2403.11807*, 2024.

Dom Huh and Prasant Mohapatra. Representation learning for efficient deep multi-agent reinforcement learning, 2024. URL https://arxiv.org/abs/2406.02890.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, et al. Gemma 3 technical report. *CoRR*, 2025.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. *arXiv preprint arXiv:2409.05907*, 2024.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering, 2024. URL https://arxiv.org/abs/2311.06668.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*, 2023.

Narmeen Oozeer, Luke Marks, Fazl Barez, and Amirali Abdullah. Beyond linear steering: Unified multi-attribute control for language models, 2025. URL https://arxiv.org/abs/2505.24535.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024. URL https://arxiv.org/abs/2312.06681.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.