# When Probes Fail to Steer: Decodability and Causal Control in Closed-Loop RL

**Anonymous authors**
Paper under double-blind review

## Abstract

In closed-loop RL, offline decodability is not evidence of steering. We test this gap in a frozen Nexto-style Rocket League policy by collecting residual activations, training a sparse autoencoder (SAE), fitting probes for simulator-derived future events, and evaluating candidate interventions through paired rollouts. Sparse-feature and raw-activation probes predict aerial touches, dribble labels, and challenges with strong offline metrics. Most probe-derived activation directions then fail as steering handles: decoder-projected directions, encoder-projected directions, exact local-gradient variants, random and shuffled controls, and action-space shaping variants are null, noisy, or high-KL rather than reliable low-KL target-specific controls. Only one held-out low-KL sparse feature boost survives: boosting feature 0 increases dribble-label frequency from 42.84 to 65.11 events per 1k steps on held-out paired seeds, with mean lift $+19.29$, 95% bootstrap interval $[+11.43, +27.13]$, positive lift on 11 of 12 seeds, and same-state KL 0.00466. Broader held-out sweeps find event-rate increases for aerial touch and challenge only through less interpretable higher-KL controls. Decoding is useful for candidate discovery, but only closed-loop intervention separates causal handles from predictive correlates.

## 1 Introduction

Activation steering has become a lightweight way to modify behavior in frozen language models. A common workflow first finds a direction or feature that predicts a behavior, then injects that representation at inference time. Closed-loop control makes this workflow suspect. In a sequential environment, an activation edit changes the current action, the next state distribution, opponent responses, and later labels. Apparent behavior control can therefore come from policy disruption, distribution shift, or label artifacts rather than targeted steering.

This paper studies that prediction-control gap. The case study is a frozen Nexto-style Rocket League policy. We collect residual-stream activations during `rlgym_sim` rollouts, learn a sparse feature basis, train future-event probes for simulator-derived labels, and test candidate interventions through closed-loop play.

The central distinction is between four claims. *Decodability* asks whether activations predict a future event. *Causal steerability* asks whether an intervention changes that event's frequency. *Specificity* asks whether the target changes without broad policy disruption. *Task preservation* asks whether gameplay diagnostics remain reasonable. Table 1 summarizes the evidence standard and findings.

Strong decodability is easy to show here. Robust low-KL steering is not: we find many predictive but non-causal or disruptive intervention surfaces, one sparse feature-level dribble-label exception, and two higher-KL dense positives for other labels.

This work makes the following contributions:

Table 1: Claim hierarchy for closed-loop activation steering.

| Claim | Evidence required | What this paper finds |
|---|---|---|
| Decodability | Probe AUROC/AUPRC | Strong for all three labels |
| Causal steerability | Paired closed-loop lift | Only robustly low-KL for dribble feature 0 |
| Specificity | Low KL/action shift | Feature 0 best; dense controls higher KL |
| Task preservation | Goals/gameplay diagnostics | Monitored as side effect; improvement not claimed |

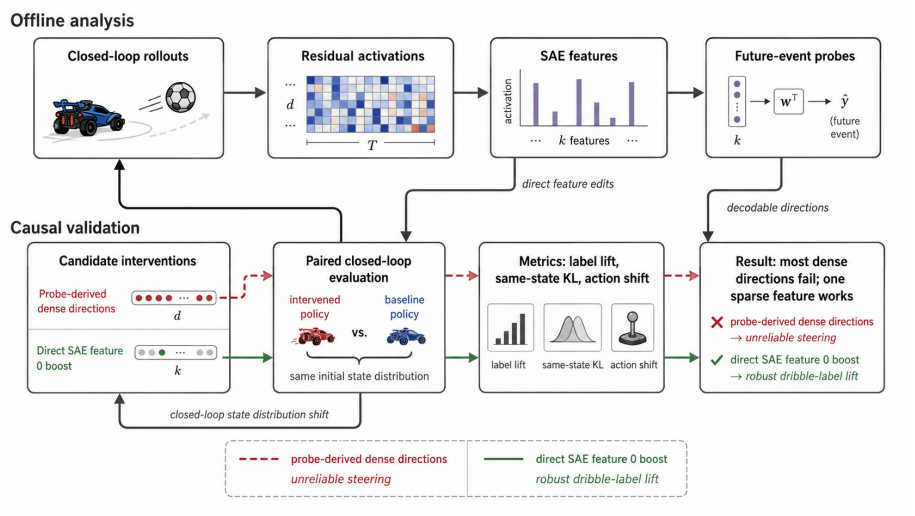


Figure 1: Evaluation pipeline. Offline probes estimate future-event decodability from activations; paired closed-loop interventions test whether candidate handles causally change event frequency while preserving low policy drift.

- a closed-loop evaluation protocol for activation interventions in a frozen Rocket League policy, measuring target-label rate, paired seed lift, same-state KL, action shift, and gameplay side effects;
- evidence that sparse-feature and raw-activation probes decode future simulator-derived events while probe-derived intervention directions mostly fail to steer those events reliably;
- a negative result for the common "probe then steer" recipe in this closed-loop setting, including decoder-projected, encoder-projected, exact local-gradient, shuffled-label, random-direction, and action-space controls;
- one narrow low-KL exception: boosting feature 0 robustly increases dribble-label frequency on held-out seeds;
- broader positive controls for aerial touch and challenge through scheduled dense/raw-activation interventions, showing event-rate steering beyond dribble but at higher KL and lower interpretability.

The main claim is intentionally narrow: closed-loop policies expose a gap between decodable features and causally useful interventions. The sparse feature result is an existence proof for one label, not evidence that sparse-feature steering generally works. The paper's center of gravity is the failure mode: offline interpretability and high probe scores are insufficient evidence for behavioral control.

## 2 Related Work

**Activation-space steering and representation engineering.** Inference-time representation interventions are the closest methodological precedent. Activation Addition and Contrastive Activation Addition derive residual-space vectors from contrastive prompt pairs or labeled examples and inject those vectors during the forward pass Turner et al. (2023); Rimsky et al. (2023). Inference-Time Intervention localizes directions associated with truthfulness in selected heads Li et al. (2023), while Representation Engineering treats population-level activations as a control surface Zou et al. (2023). Refusal directions and preference-derived steering vectors show that low-dimensional activation edits can sometimes produce targeted behavior changes in frozen language models Arditi et al. (2024); Cao et al. (2024). These results motivate activation intervention, but the evaluation setting differs: text-model interventions do not usually feed their effects back into a physical state distribution over many future control decisions.

**Sparse features, probes, and the promise of interpretable steering.** Mechanistic interpretability gives a rationale for steering through sparse latent features rather than raw hidden dimensions. Toy models of superposition explain why dense activations can entangle many latent concepts, motivating sparse decompositions Elhage et al. (2022). Sparse autoencoders have recovered more interpretable and more localizable features than raw neuron analysis in language models Huben et al. (2024). At the same time, sparse decomposition does not guarantee faithful steering: decomposing a dense vector through an SAE can distort functional effects when the vector is off the SAE training manifold or depends on signed feature contributions the SAE does not preserve Mayne et al. (2024). This paper therefore treats SAE features as candidate intervention handles that still require closed-loop causal validation.

**Interpretability and intervention in RL.** RL interpretability work has often emphasized explanations, saliency, probes, or policy summaries rather than internal activation interventions tested inside the environment Beechey et al. (2023). Closed-loop policies make intervention evidence stricter than offline prediction: a changed activation can alter the current action, future states, opponent responses, and later label frequencies. A probe that decodes a future event may therefore identify a correlate, a state-distribution marker, or a broad policy-disruption direction rather than a targeted control handle. This motivates separating decodability from causal steerability in evaluation.

**Evaluation, side effects, and the remaining gap.** Steering in sequential control must be evaluated by target-rate lift, policy drift, action distortion, seed robustness, and task side effects. A high event rate alone is insufficient because broad policy perturbations, random directions, or distribution shifts can change labels without target-specific control. We use same-state KL, action shift, paired seed deltas, bootstrap intervals, and goal diagnostics to distinguish four notions: *decodability*, whether activations predict future labels; *causal steerability*, whether intervention changes label frequency; *specificity*, whether the change avoids broad policy disruption; and *task preservation*, whether the agent still plays reasonably.

## 3 Method

### 3.1 Policy and Environment

We evaluate a frozen Nexto-style Rocket League actor in one-versus-one simulated play. Both agents use the same policy; interventions are applied to one controlled agent while the opponent remains an unmodified copy. Episodes terminate on goals, prolonged no-touch periods, or match timeout. The simulator advances the environment every eight game ticks.

We instrument intermediate policy representations at two residual blocks:

$$h^{(0)} = \mathrm{block}_0(q, x, m), \qquad h^{(1)} = \mathrm{block}_1(h^{(0)}, x, m).$$

The main experiments use residual activations from the second policy block. This activation stream gave much better sparse-autoencoder reconstruction than residual activations from the first policy block: explained variance 0.969 versus 0.091, with active fraction 0.081 versus 0.424. First-block probes were still predictive, but the second block provided a cleaner basis for steering.

### 3.2 Activations and Sparse Features

We collect observations, actions, and intermediate activations from 500 simulated episodes, with at most 4500 environment steps per episode. The sparse-autoencoder dataset contains approximately $5.4 \times 10^5$ activation samples from the selected activation stream. Each activation is represented as a 128-dimensional vector.

The SAE is a one-hidden-layer ReLU autoencoder. For $x \in \mathbb{R}^{128}$,

$$z = \mathrm{ReLU}(W_e x + b_e), \qquad \hat{x} = zW_d + b_d.$$

The decoder rows are normalized during training. The loss is

$$\mathcal{L}(x) = \mathrm{MSE}(x, \hat{x}) + \lambda \|z\|_1,$$

where the main run uses a 1024-dimensional sparse layer, $\lambda = 10.0$, batch size 2048, learning rate $3 \times 10^{-4}$, and 10 epochs.

### 3.3 Event Labels

We derive step-level labels from simulator state. Aerial touches require a ball touch while the car is off ground and above height 150. Dribbles require at least three consecutive touch steps, player and ball horizontal speed above 100, and velocity similarity at least 0.75. Challenges require simultaneous opposing touches with velocity angle at least 60 degrees and speeds above 300; the faster player is labeled. Probe and steering experiments use aerial touch, dribble, and challenge labels.

### 3.4 Future-Event Probes

For each target label, a decision at episode $e$ and step $t$ is positive if the same agent has a matching event in steps $t+1, \ldots, t+H$, with horizon $H = 32$. Sparse features are encoded for all activation samples, split by episode into an 80/20 train/validation split, standardized by training statistics, and used to train a linear binary classifier with weighted BCE loss.

Let $w_f$ be the learned feature-space probe weight after undoing feature standardization. We evaluate two dense activation-space directions:

$$v_{\mathrm{dec}} = w_f W_d, \qquad v_{\mathrm{enc}} = w_f W_e.$$

Both are normalized before evaluation. The decoder-projected direction maps the feature probe through sparse-autoencoder decoder atoms. The encoder-projected direction is a global linear approximation to the probe's input gradient. For comparison, exact local-gradient controls use the gated gradient

$$\nabla_x \ell(x) = W_e^\top D(x) w_f,$$

where $D(x) = \mathrm{diag}[\mathbf{1}\{W_e x + b_e > 0\}]$.

### 3.5 Intervention Families

Dense activation steering adds a unit vector to the selected activation stream during action selection:

$$h' = h + \alpha v,$$

where $v$ is a unit direction and $\alpha$ is a scalar from $\{-20, -10, -5, 5, 10, 20\}$. We also test exact local-gradient steering by averaging gated probe gradients over high-scoring activation states and injecting the resulting normalized direction. All reported steering runs edit only the controlled agent; the opponent remains unmodified.

Direct feature intervention modifies selected sparse-autoencoder latents. For feature set $S$, it encodes the current activation, applies boost, clamp, or ablation to $z_S$, decodes the resulting feature delta, and injects that activation delta into the policy. The feature sweep considers all 1024 sparse coordinates for the three target labels, ranks single-feature conditions on discovery seeds, then freezes selection before held-out evaluation. The confirmatory evaluation replays the top eight discovery conditions per target on untouched seeds 14000–14011, using eight episodes per seed and paired unsteered baselines.

Dense probe-derived directions aggregate many feature contributions into one activation-space vector. In a nonlinear sparse autoencoder with ReLU gates and a closed-loop policy, such directions may move activations off-manifold, activate unintended features, or perturb action logits broadly. Direct feature intervention instead modifies one latent coordinate and injects only the decoded delta associated with that coordinate, giving a more localized intervention surface.

# 4 Experimental Setup

## 4.1 Data and Hyperparameters

Table 2 summarizes the main experimental configuration.

Table 2: Experimental setup.

| Component | Value |
|---|---|
| Policy | Frozen Nexto-style actor |
| Environment | 1v1 Rocket League simulation |
| Tick skip | 8 |
| Activation stream | Residual activations from second policy block |
| Activation samples | Approximately $5.4 \times 10^5$ |
| Activation dimension | 128 |
| Sparse-feature dimension | 1024 |
| Probe horizon | 32 steps |
| Feature sweep | Top 8 discovery conditions per target |
| Confirmatory feature evaluation | 12 held-out seeds; eight episodes per seed and condition |
| Primary metrics | Target labels per 1k steps; paired seed lift; same-state KL; action shift; goal differential |

## 4.2 Metrics

Offline probes are evaluated with validation positive rate, AUROC, AUPRC, accuracy, precision at the top 1% of scores, and precision at the top 5% of scores. Closed-loop steering is evaluated by counting controlled-agent target labels and normalizing by total environment steps:

$$\mathrm{rate}(y) = 1000 \cdot \frac{\#\text{controlled target events}}{\#\text{environment steps}}.$$

The summaries report absolute lift over the unsteered baseline and relative lift. Bootstrap confidence intervals are computed by resampling episodes or seed-paired deltas, depending on the experiment. These intervals quantify rollout or seed variability within the evaluated runs.

Held-out confirmatory reruns also report controlled-agent and opponent goal rates, mean action components, action-distribution entropy, and policy KL on the same state. For each intervention

step, before the environment transition, we evaluate both steered and unsteered policies on the same current observation. The KL diagnostic compares the softmax over steered action logits against the softmax over unsteered logits, $D_{\mathrm{KL}}(\pi_{\mathrm{steered}}(\cdot|s)\|\pi_{\mathrm{base}}(\cdot|s))$, then averages over controlled-agent decisions. These diagnostics separate event-frequency control from broader gameplay effects.

# 5 Results

## 5.1 Internal Representations Decode Future Events

Sparse-autoencoder evaluation showed final MSE 0.543, L1 activation mean 0.114, explained variance 0.969, and active fraction 0.081. Because activation coordinates are not unit-normalized, the absolute MSE is mainly useful alongside explained variance: reconstruction leaves about 3.1% of activation variance unexplained. The 1024-dimensional sparse representation therefore retains most variance of the 128-dimensional residual activations while activating a small fraction of features on average, making it a useful basis for downstream probe and steering analysis.

Table 3 reports validation metrics for sparse-feature probes and baselines. Dribble has the strongest AUPRC and top-score precision. Challenge has high AUROC but much lower AUPRC, consistent with its lower validation positive rate. Direct activation probes are competitive, especially for dribble, while sparse features give their clearest advantages on aerial AUPRC and challenge AUPRC. Shuffled-label sparse-feature probes collapse toward poor or chance-level performance, showing that the probe results are not a metric artifact of class imbalance alone.

Table 3: Validation metrics for sparse-feature probes and baselines. P@1% is precision among the top 1% of scores.

| Target | Model/control | AUROC | AUPRC | P@1% |
|---|---|---|---|---|
| Aerial touch | Sparse features | 0.890 | 0.606 | 0.868 |
| Aerial touch | Raw activation | 0.873 | 0.574 | 0.891 |
| Aerial touch | Shuffled-label sparse features | 0.455 | 0.134 | 0.229 |
| Dribble | Sparse features | 0.963 | 0.842 | 0.957 |
| Dribble | Raw activation | 0.966 | 0.831 | 0.953 |
| Dribble | Shuffled-label sparse features | 0.272 | 0.067 | 0.032 |
| Challenge | Sparse features | 0.955 | 0.409 | 0.442 |
| Challenge | Raw activation | 0.920 | 0.307 | 0.365 |
| Challenge | Shuffled-label sparse features | 0.497 | 0.028 | 0.012 |

## 5.2 Probe Quality Does Not Imply Steering

High probe performance is not sufficient for closed-loop control. Table 4 summarizes the intervention-family comparison. Several alternatives change event rates or KL, but none produce a robust low-KL, target-specific positive effect comparable to direct feature intervention. Dense directions derived from the same predictive probes can be neutral, negative, or disruptive once injected into the policy loop.

Table 4: Intervention-family comparison. Target lift is mean paired target-label change in events per 1k environment steps from final control or confirmatory runs.

| Family | Target | Target lift | 95% CI | Same-state KL | Conclusion |
|---|---|---|---|---|---|
| Direct feature boost | Dribble | +19.29 | [11.43, 27.13] | 0.0047 | Low-KL positive case |
| Encoder-projected probe direction | Dribble | -30.55 | [-49.86, -12.26] | 0.1185 | Fails despite probe signal |
| Decoder-projected probe direction | Dribble | -31.38 | [-48.39, -14.58] | 0.1515 | Dense projection fails |
| Raw-activation probe direction | Dribble | -1.97 | [-17.25, 12.48] | 0.1121 | Near neutral despite decodability |
| Action-space shaping | Aerial touch | +2.60 | [1.29, 3.93] | 0.2400 | Higher-KL event-rate shift, not targeted activation steering |
| Exact local gradient | Dribble | -33.44 | [-63.70, -3.18] | 0.2990 | Local gradient fails |

The relevant comparison is not whether an intervention changes a label, but whether it changes the intended label with low policy drift and without broad action distortion.

Detailed activation-vector sweeps, random controls, action-shift diagnostics, local-gradient variants, and dense-direction controls are reported in Appendix D. The main result is unchanged by those details: probe quality alone does not identify a reliable low-KL steering direction.

### 5.3 Direct Feature Intervention Increases Dribble-Label Rate

The strongest positive result comes from directly intervening on individual sparse features rather than using dense activation directions. A discovery pass ranked single-feature boost, clamp, and ablation interventions for aerial touch, challenge, and dribble. A confirmatory pass then replayed the top eight discovery conditions per target on held-out seeds 14000–14011, using eight episodes per condition and paired unsteered baselines.

Table 5 shows the strongest confirmatory feature-level candidate for each target label. Boosting feature 0 at scale 4 is robust for dribble: mean paired lift is $+19.29$ dribble labels per 1k environment steps, the bootstrap interval is entirely positive, and 11 of 12 held-out seeds increase. A gated variant of the same feature also increases dribble rate, but clamp variants of feature 0 do not, so the evidence supports this boost operation rather than a complete semantic interpretation of the coordinate. The best aerial-touch and challenge candidates do not replicate under the same follow-up protocol. Additional dribble variants are reported in Appendix D.

Table 5: Confirmatory feature-level steering. One strongest candidate is shown for each target label. Each row uses 12 held-out seeds and eight episodes per seed. Rates and target lifts are target-label events per 1k environment steps. CI is a bootstrap interval over seed-level paired target lifts. Positive seeds counts seeds with lift $> 0$.

| Target | Intervention | Baseline rate | Steered rate | Target lift | 95% CI | Positive seeds | Same-state KL |
|---|---|---|---|---|---|---|---|
| Dribble | Boost feature 0, always on, scale 4 | 42.84 | 65.11 | 19.29 | [11.43, 27.13] | 11/12 | 0.00466 |
| Aerial touch | Clamp feature 1005, activation-gated | 4.77 | 4.71 | 0.06 | [-0.98, 1.05] | 8/12 | 0.00493 |
| Challenge | Boost feature 962, always on | 1.89 | 1.83 | -0.14 | [-0.68, 0.35] | 6/12 | 0.00052 |

Figure 2 shows the paired seed-level effect for the headline condition. The black summary markers show the across-seed mean and bootstrap interval; colored lines show each same-seed baseline and steered pair.

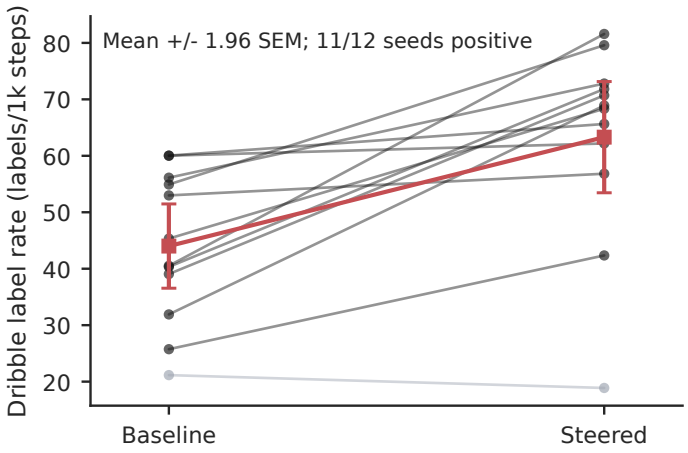


Figure 2: Held-out paired dribble rates for the feature-0 boost intervention at scale 4. Each colored line is one seed comparing unsteered baseline to intervention; black markers show the aggregate mean with 95% interval.

Figure 3 places the headline row against the other top discovery conditions tested in the same follow-up. The sparse feature-level positive effect is specific to dribble feature 0; aerial and challenge feature candidates remain near zero or negative.

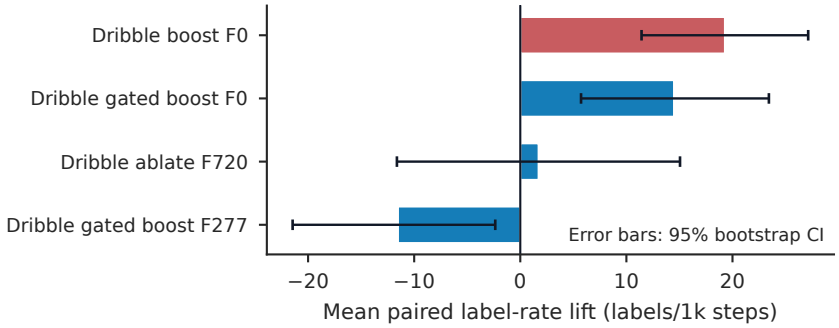


Figure 3: Confirmatory feature-level sweep summary. Bars show mean paired target lift for each top discovery condition, with bootstrap 95% intervals over seed-level target lifts. Orange marks the headline dribble feature-0 boost.

Figure 4 shows the same confirmatory sweep as a lift-versus-KL tradeoff. Feature 0 is the clearest low-KL positive outlier; other feature-level candidates are weaker, noisier, or less target-aligned.

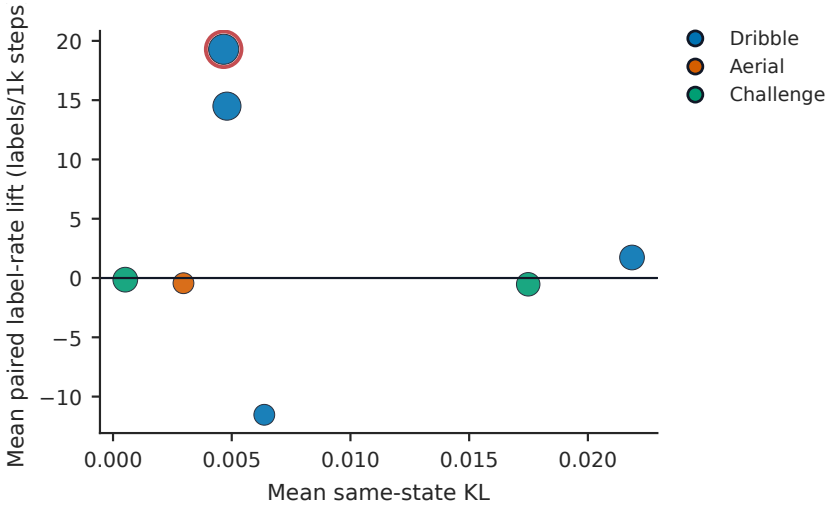


Figure 4: Feature-level lift versus same-state KL for all confirmatory feature-sweep rows. Point size is positive-seed percentage; orange ring highlights the robust dribble feature-0 intervention.

The result supports a narrow causal claim: increasing one learned feature increases dribble-label frequency in closed-loop play. It does not show that sparse-feature steering generalizes to other target labels or that probe-derived steering works in this policy. We do not yet know whether feature 0 tracks a human-legible dribbling subskill, a situational cue correlated with dribbles, or an internal action tendency; resolving that requires targeted state-conditional and action-component analysis beyond the present evidence.

## 6 Conclusion

This paper's core result is negative: decodability does not imply steerability. Sparse-feature and raw-activation probes decode future Rocket League event labels, but most dense, gradient, random,

shuffled, and action-space interventions fail to produce reliable low-KL control in this policy. Only one sparse feature boost is a real causal handle here: boosting feature 0 increases dribble-label frequency across held-out seeds with low same-state KL. Higher-KL controls can shift other event labels, but they do not provide the same evidence for interpretable low-drift steering. Feature 0 is label-specific rather than a general sparse-feature steering method. Decoding finds candidate handles; closed-loop validation decides whether they are real.

## References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024. DOI: 10.48550/arXiv.2406.11717. URL https://arxiv.org/abs/2406.11717.

Daniel Beechey, Thomas M. S. Smith, and Ozgur Simsek. Explaining reinforcement learning with shapley values. *arXiv preprint arXiv:2306.05810*, 2023. DOI: 10.48550/arXiv.2306.05810. URL https://arxiv.org/abs/2306.05810.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *arXiv preprint arXiv:2406.00045*, 2024. DOI: 10.48550/arXiv.2406.00045. URL https://arxiv.org/abs/2406.00045.

Nelson Elhage, Tristan Hume, Catherine Olsson, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. DOI: 10.48550/arXiv.2209.10652. URL https://arxiv.org/abs/2209.10652.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. DOI: 10.48550/arXiv.2309.08600. URL https://openreview.net/forum?id=F76bwRSLeK.

Kenneth Li, Oam Patel, Fernanda Viegas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023. DOI: 10.48550/arXiv.2306.03341. URL https://arxiv.org/abs/2306.03341.

Harry Mayne, Yushi Yang, and Adam Mahdi. Can sparse autoencoders be used to decompose and interpret steering vectors? *arXiv preprint arXiv:2411.08790*, 2024. DOI: 10.48550/arXiv.2411.08790. URL https://arxiv.org/abs/2411.08790.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023. DOI: 10.48550/arXiv.2312.06681. URL https://arxiv.org/abs/2312.06681.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023. DOI: 10.48550/arXiv.2308.10248. URL https://arxiv.org/abs/2308.10248.

Andy Zou, Long Phan, Sarah Chen, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023. DOI: 10.48550/arXiv.2310.01405. URL https://arxiv.org/abs/2310.01405.

# A Additional Discussion

## A.1 The Exception Is a Broad Mode, Not a Semantic Toggle

Feature 0 should not be read as a clean dribble concept. It is active on 52.6% of sampled decisions, not only immediately before dribble labels. Active states are more grounded than inactive states (83.3% versus 52.0% on ground), have more boost (0.479 versus 0.362 mean boost), and occur at higher car speed (1344 versus 1189 uu/s). Same-state intervention analysis at scale 4 gives mean KL 0.00523 and median KL $5.2 \times 10^{-5}$, so most individual action changes are small. In paired closed-loop rollouts, those small changes accumulate: mean throttle rises by 0.030, boost use by 0.024, roll falls by 0.096, ball speed rises by 66.8 uu/s, and the controlled agent is nearest to the ball more often (+0.039), while dribble labels rise from 42.84 to 65.11 per 1k steps. This pattern is more consistent with amplifying a grounded controlled-pursuit/carry mode than with toggling a semantic dribble detector.

## A.2 The Effect Is Narrow and Label-Dependent

The sparse-feature positive result is restricted to one direct feature-level dribble-label intervention. Aerial-touch and challenge feature candidates do not pass the same held-out confirmatory protocol, and dense controls do not recover a comparable dribble effect even at higher KL. This supports feature-level event-frequency control, not general sparse-feature steering.

## A.3 Other Labels Require Higher-KL Controls

The broader robustness sweep partially addresses the single-label concern, but changes the claim. Table 6 reports held-out final-test positives outside the direct sparse-feature family. A scheduled decoder-projected intervention over both residual blocks increases aerial-touch labels, and a scheduled raw-activation probe direction increases challenge labels. Both pass bootstrap and Benjamini–Hochberg criteria on the held-out final-test split, but both operate at KL over an order of magnitude larger than feature-0 dribble and are less interpretable than a single sparse coordinate. These rows should not be read as comparable successes to feature 0. They show that event-rate manipulation is possible beyond dribble, but their higher KL and denser intervention surfaces make them weaker evidence for interpretable steering. Event-rate steering is not unique to dribble, but low-KL interpretable feature steering has only one confirmed case.

Table 6: Held-out positive steering controls beyond the direct sparse-feature result. Each row uses the final-test paired seeds from the full robustness sweep. Target lift is labels per 1k environment steps.

| Target | Method | Setting | Target lift | 95% CI | BH $q$ | Same-state KL |
|---|---|---|---|---|---|---|
| Aerial touch | Decoder-projected | blocks 0+1, scale 5, proximity-gated pulse | 2.43 | [1.08, 3.79] | 0.015 | 0.0749 |
| Challenge | Raw-activation probe | block 1, scale -20, proximity-gated pulse | 2.36 | [0.66, 3.98] | 0.040 | 0.0514 |

## A.4 Post-Selection and Held-Out Validation

The single-feature discovery pass creates post-selection risk, so we separate discovery from held-out validation. Applying Benjamini–Hochberg correction to the 624 raw discovery conditions gives 11 global one-sided discoveries at $q \leq 0.05$, all for dribble; the discovery row for the feature-0 boost intervention at scale 4 has one-sided rate-test $q = 4.31 \times 10^{-4}$ globally and $q = 1.44 \times 10^{-4}$ within the dribble family. The held-out replay on seeds 14000–14011 then replicates the selected feature-0 dribble boost with paired unsteered baselines, bootstrap intervals, and positive-seed counts. A normal-approximation paired-seed power calculation using the held-out seed lifts estimates 0.996 two-sided power at $\alpha = 0.05$ for the observed mean lift 19.29 and seed-level standard deviation 14.56, with about five seeds required for 80% power at that effect size. The sign-test $p = 0.00635$

is $q = 0.0508$ within dribble-family confirmatory tests and $q = 0.152$ across all 24 confirmatory rows, so the confirmatory multiplicity-adjusted sign-test evidence is borderline within target and not global.

The claim would be weakened if similarly low-KL random, shuffled, or dense directions produced comparable paired dribble lift under the same held-out protocol; the observed controls do not show that pattern.

# B Reproducibility

The experimental pipeline consists of activation collection, sparse-autoencoder training, probe fitting, baseline evaluation, and closed-loop intervention evaluation. Reproducing the simulator experiments requires matching the Rocket League simulator, policy weights, and Torch runtime. The artifact release records commands, model hashes, intervention definitions, and run metadata for each evaluation.

# C Limitations

The selected validation and final-test protocol has been rerun for dribble, aerial, and challenge controls. Bootstrap intervals quantify variability over paired seeds/episodes, not uncertainty over policy training, alternate environment setups, or other RL policies. The feature-level dribble intervention establishes event-frequency control for one hand-defined event label, one frozen policy, one environment configuration, one activation stream, and one controlled-agent intervention setting. The aerial and challenge positives show broader event-rate steerability, but not broad sparse-feature steering: they use denser interventions, higher KL, and stronger schedule constraints. None of these results prove broad gameplay improvement.

The baselines bound method-specific claims. The broader tuning sweeps contain direct activation, shuffled-label, random-vector, decoder-projected, encoder-projected, local-gradient, action-space, and feature-level rows. Many non-feature methods are mixed or negative, and selected shuffled or random controls can change absolute target-label rates. The dense encoder-projected direction is especially limited because it drops sparse-autoencoder ReLU gates and acts as a global linear approximation. Exact local-gradient follow-ups were also negative or high-KL in our runs, so the paper does not claim that gated gradients are solved; more careful per-state or schedule-conditioned approximations remain future work. Increased event counts should therefore be attributed to the specific confirmed feature-level intervention, not to sparse-feature steering in general.

The feature-0 result should be read as a held-out replication of a selected dribble candidate with explicit FDR and power diagnostics, not as a fully familywise-controlled search result. The confirmatory sign-test evidence is borderline after within-target multiplicity correction and not significant after global correction across all confirmatory rows.

The steering evaluation changes only the controlled policy while the opponent policy remains unsteered. This is appropriate for isolating one intervention surface, but opponent and environment dynamics still affect event counts. Event labels are simulator-derived threshold labels rather than human annotations, and we have not swept the dribble threshold choices or evaluated alternative ball-control metrics such as possession duration. Because labels are simulator-thresholded, the result is best interpreted as control over the measured dribble-label rate rather than a human-validated dribbling skill. Event-rate increases can trade off against gameplay quality metrics. Goal differential is reported as a side-effect diagnostic, not as proof of task improvement; the confirmed dribble feature intervention is presented as event-frequency steering rather than improved play.

Immediate extensions are to test whether feature-level interventions can be selected reliably for additional behaviors and other closed-loop RL policies, and to measure task preservation with prespecified gameplay-quality criteria beyond same-state KL and goal differential.

# D Auxiliary Intervention Tables

These tables preserve exploratory and diagnostic runs that motivated controls but are not treated as headline evidence.

Table 7: Additional dribble feature-level variants from the confirmatory sweep. Rates and target lifts are dribble-label events per 1k environment steps. Goal-block $\Delta$ is controlled-agent minus opponent goals over each eight-episode seed block relative to paired baseline.

| Intervention | Baseline rate | Steered rate | Target lift | 95% CI | Positive seeds | Same-state KL | Goal-block $\Delta$ |
|---|---|---|---|---|---|---|---|
| Boost feature 0, always on, scale 4 | 42.84 | 65.11 | 19.29 | [11.43, 27.13] | 11/12 | 0.00466 | +11.7 |
| Boost feature 0, activation-gated, scale 4 | 42.84 | 59.02 | 14.50 | [5.73, 23.43] | 9/12 | 0.00480 | +11.8 |
| Ablate feature 720, always on | 42.84 | 48.04 | 1.73 | [-11.63, 15.05] | 6/12 | 0.02186 | +12.0 |

Table 8: Exploratory activation-vector sweep. Baseline rates and target lifts are controlled-agent target-label events per 1k environment steps.

| Direction | Target | Baseline rate | Scale | Target lift | Relative rate |
|---|---|---|---|---|---|
| Decoder | Aerial touch | 4.354 | 10 | 1.093 | 1.251 |
| Decoder | Dribble | 25.256 | -5 | -1.330 | 0.947 |
| Decoder | Challenge | 1.560 | 10 | -0.559 | 0.642 |
| Encoder-gradient | Aerial touch | 4.354 | -20 | 2.074 | 1.476 |
| Encoder-gradient | Dribble | 25.256 | 20 | 15.224 | 1.603 |
| Encoder-gradient | Challenge | 1.560 | -5 | -0.402 | 0.742 |

Table 9: Short action/goal and policy-shift diagnostics. Target rates are target-label events per 1k environment steps, aggregated over three seeds with 15 episodes per seed. KL is same-state policy KL.

| Condition | Target rate | Blue goals/ep. | Orange goals/ep. | Action shift | Same-state KL |
|---|---|---|---|---|---|
| Baseline | – | 3.42 | 3.24 | 0.000 | 0.000 |
| Dribble encoder, scale 10 | 46.46 | 2.42 | 3.87 | 0.124 | 0.280 |
| Dribble encoder, scale 20 | 35.10 | 1.62 | 5.87 | 0.238 | 0.918 |
| Dribble raw activation, scale 20 | 7.27 | 2.33 | 4.71 | 0.215 | 0.443 |
| Dribble shuffled sparse features, scale 20 | 25.50 | 1.53 | 5.31 | 0.232 | 0.747 |
| Aerial encoder-projected, scale -20 | 8.45 | 0.00 | 8.00 | 0.471 | 1.441 |
| Aerial raw activation, scale -20 | 4.95 | 2.91 | 4.11 | 0.220 | 0.450 |
| Aerial shuffled sparse features, scale -20 | 6.69 | 0.00 | 7.71 | 0.370 | 1.041 |

# E Auxiliary Result Diagrams

The main text summarizes mixed activation-vector and control experiments without treating them as headline evidence. Their diagrams are retained here for completeness.

Table 10: Exact local sparse-autoencoder gradient follow-up. Each target reports the same four schedules. Target lifts are paired against same-seed unsteered baselines over five seeds and six episodes per seed, in target-label events per 1k environment steps. Goal $\Delta$ is controlled-agent minus opponent goals per episode relative to paired baseline.

| Target | Schedule | Active frac. | Same-state KL | Target lift | 95% CI | Sign-flip $p$ | Goal $\Delta$ |
|---|---|---|---|---|---|---|---|
| Dribble | Always on | 1.000 | 0.299 | -33.44 | [-63.70, -3.18] | 0.125 | 0.07 |
| Dribble | Ball distance $< 900$ | 0.451 | 0.148 | -30.79 | [-51.60, -9.98] | 0.0625 | -0.47 |
| Dribble | Pulse 8 / cooldown 24 | 0.257 | 0.082 | -32.63 | [-51.49, -13.76] | 0.0625 | -1.43 |
| Dribble | Ball distance $< 900$ plus pulse | 0.136 | 0.051 | -19.86 | [-36.23, -3.48] | 0.0625 | -0.33 |
| Aerial touch | Always on | 1.000 | 0.242 | 1.16 | [-1.49, 3.81] | 0.375 | -2.37 |
| Aerial touch | Ball distance $< 900$ | 0.475 | 0.112 | 1.52 | [-2.56, 5.61] | 0.375 | -0.63 |
| Aerial touch | Pulse 8 / cooldown 24 | 0.259 | 0.062 | 0.04 | [-1.46, 1.54] | 0.875 | -0.27 |
| Aerial touch | Ball distance $< 900$ plus pulse | 0.121 | 0.026 | 1.41 | [-0.24, 3.06] | 0.125 | -0.43 |
| Challenge | Always on | 1.000 | 0.289 | -1.30 | [-2.27, -0.32] | 0.0625 | -0.50 |
| Challenge | Ball distance $< 900$ | 0.451 | 0.133 | 0.24 | [-2.19, 2.67] | 0.875 | -1.20 |
| Challenge | Pulse 8 / cooldown 24 | 0.256 | 0.070 | -1.86 | [-2.32, -1.41] | 0.0625 | 1.60 |
| Challenge | Ball distance $< 900$ plus pulse | 0.108 | 0.027 | 2.10 | [-0.82, 5.01] | 0.0625 | -1.03 |

Table 11: Final dense-direction and action-space controls. Each target reports the same six intervention families over 20 final-test seeds. Target lifts are target-label events per 1k environment steps, paired against same-seed unsteered baselines. Goal $\Delta$ is controlled-agent minus opponent goals per episode relative to paired baseline.

| Target | Method | Target lift | 95% CI | Same-state KL | Goal $\Delta$ |
|---|---|---|---|---|---|
| Aerial touch | Encoder-gradient | -1.12 | [-1.98, -0.19] | 0.057 | 0.28 |
| Aerial touch | Decoder-projected | 0.51 | [-0.70, 1.81] | 0.057 | -0.60 |
| Aerial touch | Raw-activation probe | 0.63 | [-0.54, 1.78] | 0.044 | 1.03 |
| Aerial touch | Shuffled-label probe | 0.67 | [-0.70, 1.86] | 0.037 | 0.68 |
| Aerial touch | Random vector | 1.17 | [0.37, 2.00] | 0.071 | 0.70 |
| Aerial touch | Action-space shaping | 2.60 | [1.29, 3.93] | 0.240 | 0.23 |
| Dribble | Encoder-gradient | -30.55 | [-49.86, -12.26] | 0.119 | -0.03 |
| Dribble | Decoder-projected | -31.38 | [-48.39, -14.58] | 0.152 | 0.83 |
| Dribble | Raw-activation probe | -1.97 | [-17.25, 12.48] | 0.112 | -0.10 |
| Dribble | Shuffled-label probe | -23.73 | [-37.17, -10.68] | 0.155 | 0.05 |
| Dribble | Random vector | -42.70 | [-62.90, -22.44] | 0.251 | -0.38 |
| Dribble | Action-space shaping | -30.86 | [-46.99, -15.89] | 0.119 | -0.40 |
| Challenge | Encoder-gradient | -0.97 | [-2.36, 0.69] | 0.135 | -0.20 |
| Challenge | Decoder-projected | -1.99 | [-3.31, -0.69] | 0.079 | 0.33 |
| Challenge | Raw-activation probe | -1.48 | [-2.37, -0.62] | 0.046 | 1.05 |
| Challenge | Shuffled-label probe | -1.54 | [-2.57, -0.62] | 0.147 | 0.15 |
| Challenge | Random vector | -2.57 | [-3.55, -1.69] | 0.083 | 0.00 |
| Challenge | Action-space shaping | -1.43 | [-2.46, -0.49] | 0.272 | 0.33 |

Table 12: Stronger dribble-only dense-control follow-up. Target lifts are dribble-label events per 1k environment steps over 20 final-test seeds.

| Method | Scale/KL target | Target lift | 95% CI | Same-state KL | Goal $\Delta$ |
|---|---|---|---|---|---|
| Encoder-gradient | -15 | -35.39 | [-54.88, -16.83] | 0.263 | 0.25 |
| Decoder-projected | -15 | -18.07 | [-38.92, 2.19] | 0.377 | -0.33 |
| Raw-activation probe | -15 | -0.84 | [-21.09, 18.94] | 0.235 | -0.55 |
| Shuffled-label probe | -15 | -6.27 | [-23.87, 11.70] | 0.325 | 0.13 |
| Random vector | -15 | -44.43 | [-59.16, -28.37] | 0.504 | -0.78 |
| Action-space shaping | KL 0.3 | -43.87 | [-63.35, -23.49] | 0.314 | 1.23 |

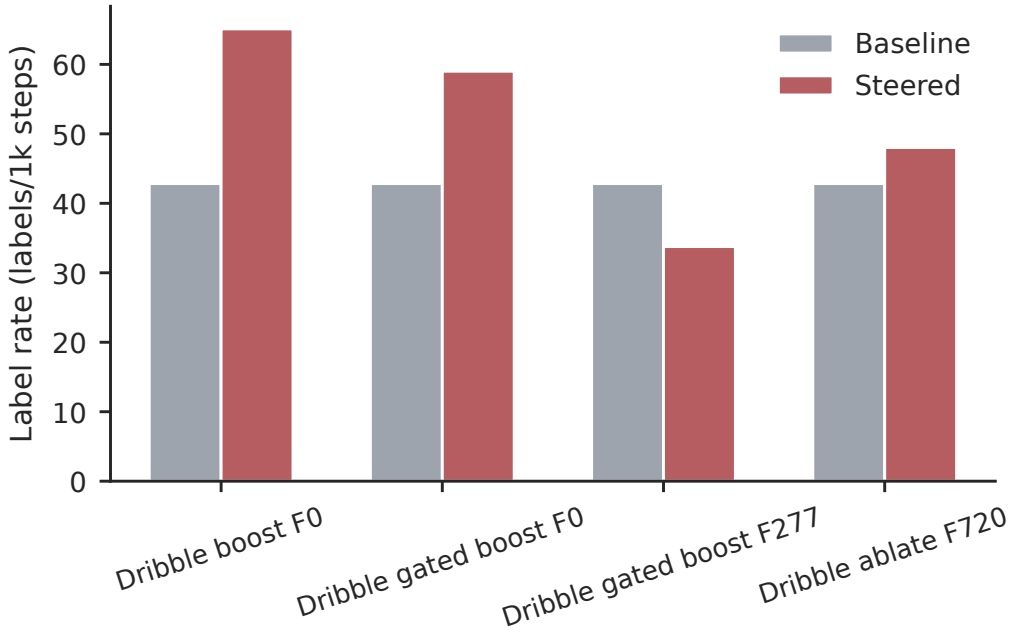


Figure 5: Pooled baseline and steered rates for the headline condition and nearest comparison rows.

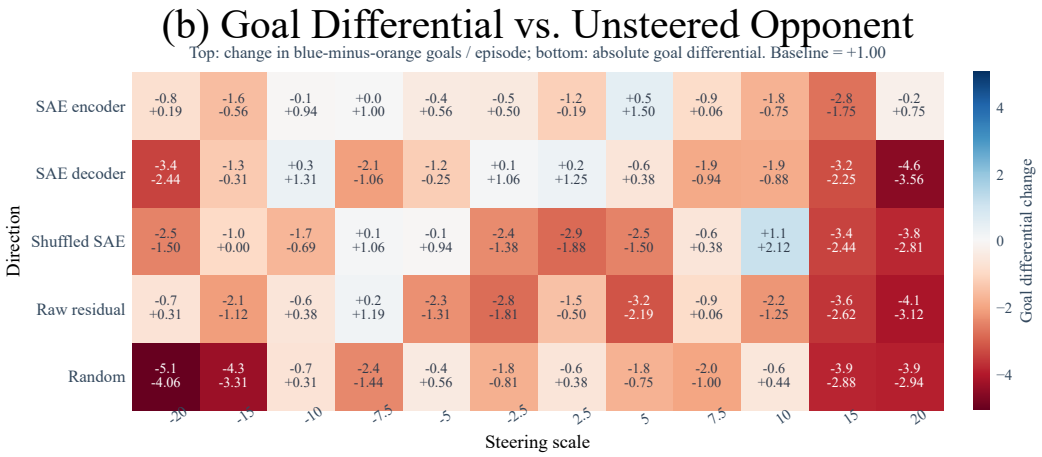


Figure 6: Goal outcome changes in the dense dribble scale scan. Each cell reports change in controlled-agent minus opponent goals per episode relative to paired unsteered baseline on top and absolute goal differential below.