BEYOND THE POLICY TRUST REGION: ATTRIBUTING AND CONTROLLING REPRESENTATION-INDUCED MOVEMENT IN PPO

Anonymous authors
Paper under double-blind review

ABSTRACT

Trust-region methods constrain how far a policy moves, but they do not reveal which part of the policy network produces that movement. In policies with a learned representation and an action head, updates to the two components can reinforce or compensate for one another, so the complete old-to-new policy change can hide substantial component-level movement. We make these dynamics observable by recombining old and new representations with old and new action heads to form four counterfactual policies. In particular, evaluating the new representation through the frozen behavior head measures the action-distribution change attributable to the representation update. Building on this construction, we introduce the Movement-Attributed Representation Trust Region (MART), a plug-in regularizer that limits excess movement along this representation path while preserving the underlying PPO-family actor objective. Across extensive experiments in open source environments from various control scenarios, MART reduces the deviation of the attributed movement to the representation and the action ratio of the sample in all the settings tested. In traffic control, these reductions become stronger as each rollout is reused for more optimization epochs, and the action head often moves farther when the representation path is constrained, revealing compensation that is partially hidden by complete-policy diagnostics. Moreover, the results show that controlling policy movement generally benefits from considering not only the complete policy update, but also the component through which that movement arises.

1 INTRODUCTION

Trust regions are central to stable on-policy reinforcement learning (RL). TRPO constrains divergence between successive policies, while PPO uses a clipped sampled-action ratio objective as a practical first-order alternative (Schulman et al., 2015; 2017). Both regulate how far the final policy may move, but neither identifies which network component produced that movement.

Consider a policy factored into a representation network and an action head. A training step changes both components. The representation and head can each move the action distribution, can amplify one another, or can compensate so that substantial component movement appears small in the complete old-to-new comparison. Conversely, a large complete movement may arise primarily from the head while the representation changes little. Final-policy divergence is therefore insufficient for attributing the source of an update. This ambiguity matters whenever decisions depend on a learned shared representation. Feature-space regularization can stabilize latent geometry (Moalla et al., 2024), yet geometric proximity alone does not show how a feature change redistributes action probability. Heterogeneous modalities provide a useful stress test: in traffic, adding an LLM-derived text modality consistently increases fused-representation movement, but sampled-action ratios do not increase uniformly. The broader object of interest is therefore representation change measured through the policy's action geometry; multimodality supplies one demanding instance of it.

We address the ambiguity by counterfactually recombining old and new representations with old and new action heads. The resulting four policy distributions distinguish representation-attributed, head-attributed, and complete policy movement and expose compensation between components. This decomposition yields MART, the *Movement-Attributed Representation Trust Region*, which constrains

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the action-distribution movement induced by replacing the representation while holding the behavior head fixed. This target sits between feature-space matching and complete-policy trust regions: it measures a representation update by the action-distribution change it induces. MART is a plug-in penalty for PPO-family objectives. Our implementation, MART-TV, measures this path with total variation. We write MART without a suffix below. In summary, we make three contributions:

- **Counterfactual policy-movement attribution.** A four-distribution construction attributes action-space movement to representation and head updates, their complete effect, and compensation between components.
- **Movement-Attributed Representation Trust Region.** MART is a plug-in penalty for PPO, SPO, and DRPO that constrains action-space movement induced by representation updates under the frozen behavior head.
- **Mechanistic empirical study.** Experiments in various control scenarios show a mismatch between latent and action-space movement, increasingly strong MART control under optimization pressure, and objective-dependent performance consequences.

2 Related Work

Policy-level trust regions. TRPO formalizes conservative policy improvement through a constraint on the expected KL divergence between successive policies, while PPO replaces the constrained second-order update with a clipped importance-ratio surrogate that supports multiple minibatch epochs (Schulman et al., 2015; 2017). The realized behavior of both algorithms depends strongly on implementation choices, cautioning against equating an objective's nominal trust region with the update actually produced (Engstrom et al., 2020). Recent objectives continue to refine policy-level update geometry: SPO modifies PPO's surrogate to control probability ratios more directly (Xie et al., 2025), while DRPO uses a smooth divergence-weighted regularizer in LLM reinforcement learning (Yao et al., 2026). These methods regulate or characterize the complete movement of the policy . MART constrains the behavior-head-conditioned movement induced by the representation update, preventing this component path from growing arbitrarily large even when subsequent head adaptation partially hides it in the complete policy update.

Representation stability and feature regularization. A growing literature links RL failure modes to learned representations. DR3 shows that temporal-difference learning can favor degenerate features in offline value-based RL and introduces explicit regularization (Kumar et al., 2022). PPG separates policy and value optimization into phases and uses feature distillation to reduce interference (Cobbe et al., 2021). Most directly related, Moalla et al. connect PPO feature-rank deterioration and capacity loss to degraded trust-region behavior and propose Proximal Feature Optimization (PFO) to regularize representation dynamics (Moalla et al., 2024). Recent work also studies training-induced output churn and plasticity loss (Tang et al., 2025), or stabilizes large RL networks through hyperspherical feature and weight normalization (Lee et al., 2025). Together, these methods establish the importance of representation dynamics through feature geometry, interference, capacity, output stability, and architectural normalization. MART connects the feature and policy levels by passing a representation update through the frozen behavior head. It can leave a geometrically large but decision-invariant change unpenalized and detect a small change that strongly redistributes action probability. We use PFO, MSE, and cosine penalties as feature-space comparisons to this action-conditioned quantity.

Counterfactual analysis and multimodal policies. Counterfactual constructions in RL have primarily addressed temporal credit assignment. Mesnard et al. condition value functions on future information to distinguish an action's contribution to return from external factors and subsequent actions (Mesnard et al., 2021). At the representation level, interchange intervention training swaps internal representations so a network realizes the counterfactual behavior of a specified causal model (Geiger et al., 2022). MART shares the intervention intuition but applies it to component attribution within a single policy update. It recombines old and new components, i.e., old representation/old head versus new representation/old head -and measures the resulting action-space movement. The construction requires no external causal model. Multimodal policies such as VIMA and instructRL show how visual and language representations can enter action generation or policy regularization (Jiang et al., 2023; Hu & Sadigh, 2023); they motivate heterogeneous representations as a stress test.

3 Attributing and Controlling Representation-Induced Policy Movement

3.1 Why complete policy movement is ambiguous

We consider an on-policy reinforcement learning agent that interacts with an environment through observations x_t, selects actions $a_t \sim \pi(a_t \mid x_t)$, and maximizes expected discounted return (Sutton et al., 1998). Policy-gradient methods update the policy using an advantage estimate $\hat{A}_t$, which indicates whether the sampled action performed better or worse than expected (Sutton et al., 1999). In PPO (Schulman et al., 2017), the update is based on the probability ratio $r_t(\theta, \psi) = \frac{\pi_{\theta,\psi}(a_t|x_t)}{\pi_{\bar{\theta},\bar{\psi}}(a_t|x_t)}$, where $(\bar{\theta}, \bar{\psi})$ denotes the behavior policy that generated the rollout and (θ, ψ) denotes the current policy. The standard clipped PPO objective clips the surrogate contribution associated with this sampled-action ratio:

$$J_{\mathrm{PPO}} = \mathbb{E}_t\Big[\min\Big(r_t\hat{A}_t,\ \mathrm{clip}(r_t, 1-\epsilon, 1+\epsilon)\hat{A}_t\Big)\Big]. \tag{1}$$

Thus, conventional PPO clips the surrogate contribution associated with large sampled-action ratio changes, but it does not impose a hard bound on policy movement or identify which part of the network produced that movement. To expose this ambiguity, let a categorical policy factor into a representation network and an action head:

$$\pi_{\theta,\psi}(a \mid x) = \mathrm{softmax}(g_\psi(f_\theta(x)))_a\,, \tag{2}$$

where f_θ maps the observation to a learned representation and g_ψ maps that representation to action logits. A policy update changes both components, from $(\bar{\theta}, \bar{\psi})$ to (θ, ψ). Consequently, the complete old-to-new action-distribution change conflates movement induced by the representation update, movement induced by the head update, and any interaction or compensation between them. This distinction is not resolved by the PPO ratio: r_t evaluates only the probability assigned to the sampled action under the complete old and new policies. Accordingly, we distinguish *latent movement*, which denotes geometric change in representation space, from *representation-attributed movement*, which denotes the induced action-distribution change under a fixed behavior head, and *complete policy movement*, which denotes the full policy change.

3.2 Counterfactual component attribution

For each behavior observation, let $z_o = f_{\bar{\theta}}(x)$ and $z_n = f_\theta(x)$. We evaluate all four old/new representation–head combinations:

$$q_{oo} = \mathrm{softmax}(g_{\bar{\psi}}(z_o)), \qquad q_{no} = \mathrm{softmax}(g_{\bar{\psi}}(z_n)), \tag{3}$$

$$q_{on} = \mathrm{softmax}(g_{\psi}(z_o)), \qquad q_{nn} = \mathrm{softmax}(g_{\psi}(z_n)). \tag{4}$$

The first index denotes the representation and the second denotes the head; o and n mean old and new. Thus q_{no} evaluates the new representation through the old head, whereas q_{nn} is the complete current policy. Action masks, when present, are applied before normalization so every comparison has the same valid support. For distributions p and q over that support, total variation is $D_{\mathrm{TV}}(p, q) = \frac{1}{2}\sum_a |p(a) - q(a)|$. The factorized diagnostics are

$$d_{\mathrm{rep}} = D_{\mathrm{TV}}(q_{oo}, q_{no}), \qquad d^{(n)}_{\mathrm{head}} = D_{\mathrm{TV}}(q_{no}, q_{nn}), \qquad d_{\mathrm{total}} = D_{\mathrm{TV}}(q_{oo}, q_{nn}). \tag{5}$$

Here d_{rep} asks how much the action distribution moves when only the representation endpoint changes under the behavior head; d_{total} asks how much the complete policy moves. Keeping these diagnostics in TV places every path on a common scale and makes the cancellation construction below a direct consequence of the triangle inequality. Hereafter we use training distance:

$$d_{\mathrm{reg}} = D_{\mathrm{TV}}(q_{oo}, q_{no}) = d_{\mathrm{rep}}. \tag{6}$$

Because the behavior head is detached, the penalty gradient reaches the current representation but not the head; the PPO-family base objective still updates the complete current policy. We select this old-head path because it anchors the comparison to the behavior policy that generated the rollout and isolates an encoder update without allowing the regularizer to alter the action head. Thus, q_{no} is a controlled intervention rather than a policy intended for execution; it measures the action-space consequence of the representation update under the fixed behavior head. The corresponding path under the current head is retained as a diagnostic.

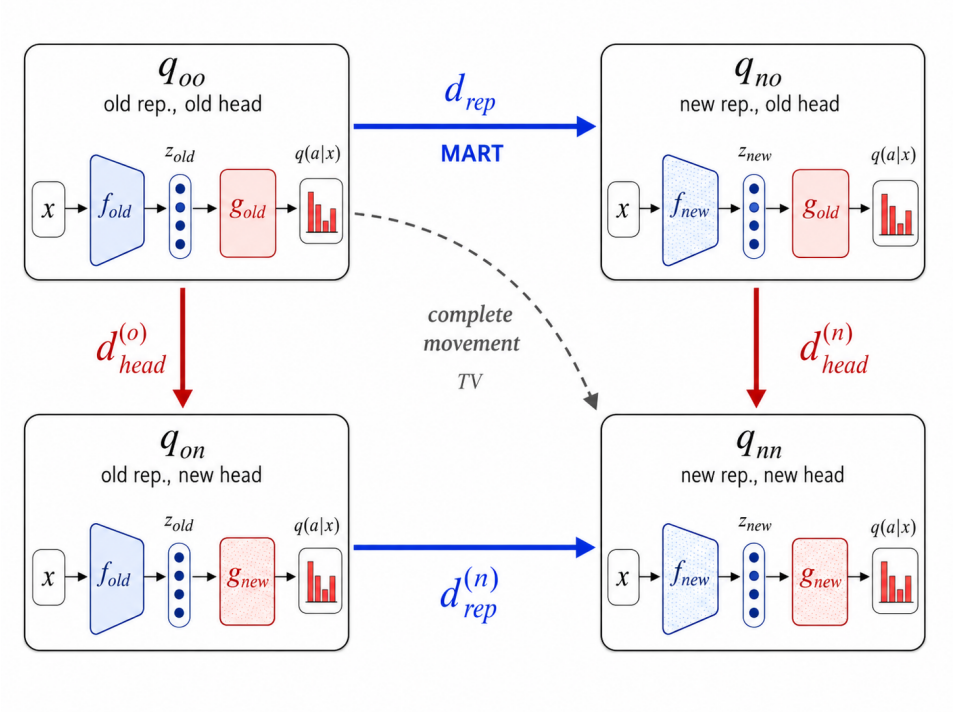

Figure 1: **Counterfactual policy recombination.** The four distributions combine old/new representations (first index) and heads (second index). Blue horizontal paths isolate representation-attributed movement and red vertical paths isolate head-attributed movement. MART constrains $q_{oo} \to q_{no}$; the dashed $q_{oo} \to q_{nn}$ path measures complete policy movement.

Supporting paths. Comparing the TV component paths with complete TV movement yields the representation-first cancellation diagnostic

$$c^{r\to h} = \left[d_{\text{rep}} + d^{(n)}_{\text{head}} - d_{\text{total}}\right]_+. \tag{7}$$

An analogous head-first diagnostic uses $d^{(o)}_{\text{head}} = D_{\text{TV}}(q_{oo}, q_{on})$ and $d^{(n)}_{\text{rep}} = D_{\text{TV}}(q_{on}, q_{nn})$. By the TV triangle inequality, both cancellation gaps are nonnegative before numerical roundoff. They are controlled component interventions, not a unique causal or additive decomposition: TV supplies a path bound. A large gap reveals component movement hidden by the complete comparison and is consistent with partial compensation.

3.3 Movement-Attributed Representation Trust Region

Following the above diagnostic, Movement-Attributed Representation Trust Region(MART) turns its central representation-attributed path into a soft control target. Notably, MART does not assume that every large d_{rep} is intrinsically harmful. Rather, it encodes the design preference that repeated optimization of a fixed rollout should not induce arbitrarily large changes in the behavior policy's action distribution through the representation component alone. For the configured distance d_{reg} from Equation 6, boundary $\delta > 0$, and numerical stabilizer $\epsilon_0 > 0$, define the normalized squared hinge

$$\ell_{\text{MART}}(d_{\text{reg}}) = \frac{1}{2}\left(\frac{[d_{\text{reg}} - \delta]_+}{\delta + \epsilon_0}\right)^2. \tag{8}$$

The hinge is zero inside the boundary and penalizes only excess movement. All reported MART configurations use uniform weighting over the complete minibatch. For minibatch $\mathcal{B}$ and penalty

coefficient $\lambda \geq 0$, the regularizer and actor objective are

$$\mathcal{R}_{\mathrm{MART}} = \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \ell_{\mathrm{MART}}((d_{\mathrm{reg}})_t), \tag{9}$$

$$J_{\mathrm{MART}} = \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} j_{\mathrm{base},t} - \lambda \mathcal{R}_{\mathrm{MART}}. \tag{10}$$

The behavior head and q_{oo} are detached, while gradients pass through the current representation in q_{no}. All main experiments use MART-TV, referred to simply as MART. The base gain j_{base} remains the original PPO, SPO, or DRPO actor gain; MART adds the excess-movement penalty to that gain.

This target differs from both adjacent forms of control: feature-space penalties such as PFO regularize latent geometry without measuring the induced action-distribution change, whereas PPO/TRPO and complete-policy KL act on the final old-to-new policy without localizing the representation component. In summary, this construction yields three testable expectations: conventional policy diagnostics need not track representation change, MART should become more consequential under stronger optimization pressure, and movement magnitude can vary independently of direction.

4 EXPERIMENTS

We first test whether latent and action-space movement agree, then measure how MART changes representation-attributed movement under increasing optimization pressure and across domains. Standard policy diagnostics and the feature-space baseline isolate the effect of the counterfactual target, while directional diagnostics connect movement control to task performance.

4.1 SETUP

Environments. The suite covers complementary sources of variation rather than repeated instances of one control problem: representation composition (numeric versus numeric+language), action structure (small fixed and dynamically masked discrete catalogs), physical dynamics (traffic, flight, and power flow), and operating condition (nominal, exogenous noise, and discrete contingencies). Traffic experiments use CityFlow (Zhang et al., 2019) under three conditions: (1) Normal traffic has no injected disruption. (2) The accident blocks one approach from simulation time 300 to 600 by reducing its capacity to 10%. (3) Over the same interval, the malfunction hides one approach with probability 0.8 and adds Gaussian observation noise with standard deviation 0.30. Besides, we assume policies to receive either numeric state or numeric state fused with LLM-derived text. Therefore, the multimodality can be considered in the analysis. LunarLander-v3 has four discrete actions and is accessed through Robust-Gymnasium (Gu et al., 2025), with clean or Gaussian state-noise conditions. Grid2Op (Marot et al., 2021) uses `l2rpn_case14_sandbox`, the LightSim2Grid backend (Donnot, 2020), and a semantically discrete catalog containing no-op and line-status actions, under clean operation or line contingencies.

Protocol and metrics. We apply MART to three popular PPO-family methods—PPO, SPO, and DRPO—and compare each against its unregularized backbone. In addition, PFO is introduced as an advanced feature-space baseline. All reported comparisons use five paired seeds. Traffic runs use 500 iterations and 128-step rollouts. Traffic effects are paired by seed and modality before averaging numeric and numeric+LLM results. LunarLander uses 1000 iterations and 1024-step rollouts; Grid2Op uses 500 iterations and 1024-step rollouts. Unless an epoch-pressure point is specified, PPO epochs are eight. We report paired bootstrap 95% confidence intervals over seed-level differences. Movement metrics include mean $|r - 1|$, clipping frequency, extreme ratios, d_{rep}, and d_{total}. Task metrics are reward and queue in traffic, return and landing success in LunarLander, and grid return and survival in Grid2Op. All detailed settings appear in Appendix B.

4.2 INPUT-MODALITY AND ACTION-SPACE MOVEMENT DIVERGE

The motivating traffic comparison asks whether adding an LLM-derived text modality changes fused-representation and sampled-action movement consistently. Figure 2 compares numeric+LLM

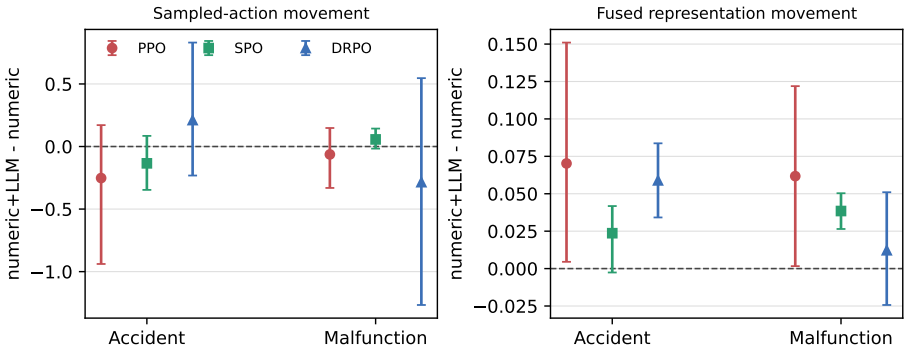

Figure 2: Effect of the LLM-derived text modality (numeric+LLM minus numeric) on unregularized traffic policies at eight PPO epochs. Points average five seeds; bars are paired bootstrap 95% intervals.

Table 1: Traffic at eight PPO epochs: MART minus base with numeric and numeric+LLM policies reported separately. Each row averages accident and malfunction after pairing by seed. Negative movement and queue changes and positive reward changes are favorable. † denotes a paired bootstrap 95% interval excluding zero; bold marks every directionally favorable MART change.

Backbone	Modality	$\Delta\lvert r-1\rvert$	Δ clipping	Δd_{rep}	Δ reward	Δ queue
PPO	Numeric	$\mathbf{-0.691}^{\dagger}$	$\mathbf{-0.168}^{\dagger}$	$\mathbf{-0.169}^{\dagger}$	$\mathbf{1.802}^{\dagger}$	$\mathbf{-20.30}^{\dagger}$
PPO	Numeric+LLM	$\mathbf{-0.384}^{\dagger}$	$\mathbf{-0.088}$	$\mathbf{-0.153}^{\dagger}$	$\mathbf{0.703}$	$\mathbf{-7.97}$
SPO	Numeric	$\mathbf{-0.316}^{\dagger}$	$\mathbf{-0.179}^{\dagger}$	$\mathbf{-0.103}^{\dagger}$	$\mathbf{1.312}^{\dagger}$	$\mathbf{-14.60}^{\dagger}$
SPO	Numeric+LLM	$\mathbf{-0.302}^{\dagger}$	$\mathbf{-0.124}^{\dagger}$	$\mathbf{-0.093}^{\dagger}$	$\mathbf{2.617}^{\dagger}$	$\mathbf{-29.28}^{\dagger}$
DRPO	Numeric	$\mathbf{-0.743}^{\dagger}$	$\mathbf{-0.117}$	$\mathbf{-0.114}^{\dagger}$	$\mathbf{0.227}$	$\mathbf{-2.50}$
DRPO	Numeric+LLM	$\mathbf{-0.735}^{\dagger}$	$\mathbf{-0.138}^{\dagger}$	$\mathbf{-0.113}^{\dagger}$	-0.105	1.11

with numeric policies at eight PPO epochs before adding MART. The text modality increases fused-representation movement in all six backbone–scenario cells, whereas sampled-action ratio deviation increases in only two and counterfactual d_{rep} is not uniformly higher. Its action-level expression therefore depends on the backbone, scenario, and head response, motivating direct measurement of component-attributed policy movement.

4.3 MART CONTROLS REPRESENTATION-ATTRIBUTED MOVEMENT

Eight PPO epochs create substantial update pressure through repeated optimization of each rollout. Table 1 reports MART minus its matched backbone in this regime. Across PPO, SPO, and DRPO, MART reduces representation-attributed and complete policy movement, sampled-action ratio deviation, and clipping in every retained traffic cell. PPO and SPO also improve reward and queue in all four cells; DRPO improves both in three, with aggregate task intervals crossing zero. The separated rows show that MART controls $|r-1|$ and d_{rep} under either input construction rather than succeeding through modality averaging. Reward and queue improve in all six numeric backbone–scenario cells and five of six numeric+LLM cells; the exception is DRPO under accidents. These results also show that reducing representation-attributed movement does not guarantee a performance improvement for every PPO-family objective. MART constrains the magnitude of movement along the representation path, while the underlying actor objective still determines whether the resulting update is beneficial. Appendix C further separates normal and abnormal effects by modality.

The component paths make this response explicit:

	Δd_{rep}	$\Delta d_{\text{head}}^{(n)}$	Δd_{total}
PPO	$\mathbf{-0.161}$	0.090	$\mathbf{-0.103}$
SPO	$\mathbf{-0.098}$	0.044	$\mathbf{-0.083}$
DRPO	$\mathbf{-0.114}$	0.064	$\mathbf{-0.089}$

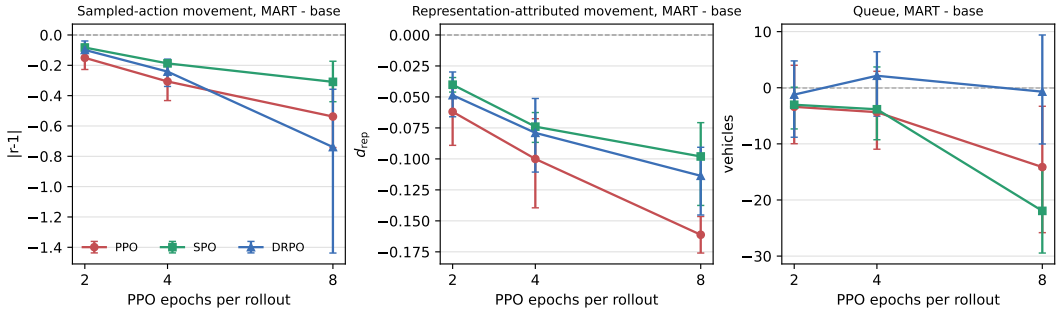

Figure 3: Traffic optimization pressure. Each point is MART minus its matched base, macro-averaged over accident/malfunction and numeric/numeric+LLM cells after pairing by seed. Bars are paired bootstrap 95% intervals; negative values favor MART for all three panels.

The representation-attributed path contracts for every backbone while the current head moves farther. Complete policy movement still decreases, but by less than d_{rep}. The difference directly exposes compensation hidden by an old-to-new policy comparison.

4.4 The effect strengthens with optimization pressure

Reusing each rollout for additional optimization epochs creates more opportunity for representation and policy movement within an update. Figure 3 reports MART-minus-base changes over accident and malfunction using numeric-only and numeric+LLM modalities. Notably, reductions in ratio deviation and representation-attributed movement grow with optimization pressure. Queue effects remain uncertain at two and four epochs but become clearly favorable for PPO and SPO at eight. In these experiments, MART becomes increasingly consequential as each rollout is reused for more optimization epochs. Normal-state traffic provides a complementary reference: MART remains active and lowers movement for all three backbones, PPO and DRPO preserve the control trade-off, and SPO has a small unresolved cost. The larger abnormal-event reductions concentrate MART's effect where rare events and repeated optimization amplify policy movement.

4.5 The mechanism transfers across domains

LunarLander and Grid2Op test the same intervention under different dynamics and discrete action structures. Table 2 reports every backbone and condition for the selected external-domain configurations and all three traffic conditions. Under normal traffic, MART reduces both movement measures for every backbone while leaving PPO and DRPO reward nearly unchanged; SPO has a small reward cost. During accidents, PPO and SPO convert the movement reduction into clear reward gains, whereas DRPO remains neutral. During malfunctions, all three backbones improve mean reward while reducing movement. Clean LunarLander MART-PPO combines reduced movement with a significant return gain. Under state noise, movement reductions persist but return effects are uncertain. In Grid2Op, all three backbones reduce ratio deviation and representation-attributed movement. PPO has favorable mean return, DRPO has favorable but uncertain mean return, and SPO incurs a return cost. Across both external domains, the selected configurations therefore preserve the movement-control mechanism while exposing backbone-dependent task effects.

4.6 Action-relevant and feature-space regularization

PFO is the closest feature-space comparison: it adds an L2 trust region on the actor's preactivation representation to the same PPO surrogate (Moalla et al., 2024). MART-PPO instead measures the consequence of representation replacement after the old action head maps it into a policy. This matched design distinguishes latent proximity from action-relevant proximity without changing the base actor objective. Notably, the traffic pressure sweep highlights the difference. At two epochs, PFO slightly reduces both sampled-action and representation-attributed movement relative to PPO. At four and eight epochs it largely follows PPO, whereas MART-PPO remains substantially lower (Figure 4). At eight epochs, MART-PPO reduces $|r-1|$ and d_{rep} relative to PFO by 33% and 59%

Table 2: Cross-domain MART-minus-base effects, averaged over five random seeds. Traffic reports the ordinary-state window for normal traffic and the active-event window for accident and malfunction; each row averages the numeric and numeric+LLM modalities after pairing. Task is reward for traffic, return for LunarLander, and grid return for Grid2Op. Bold marks every favorable MART change.

Environment	Condition	Backbone	Δ task	$\Delta\lvert r-1\rvert$	Δd_{rep}
Traffic	normal	PPO	**0.0011**	**−0.8913**	**−0.1738**
Traffic	normal	SPO	−0.0293	**−0.1753**	**−0.0929**
Traffic	normal	DRPO	**0.0010**	**−0.4224**	**−0.1103**
Traffic	accident	PPO	**1.3792**	**−0.5859**	**−0.1683**
Traffic	accident	SPO	**2.9619**	**−0.2555**	**−0.0838**
Traffic	accident	DRPO	−0.0322	**−0.4813**	**−0.1097**
Traffic	malfunction	PPO	**1.1265**	**−0.4896**	**−0.1541**
Traffic	malfunction	SPO	**0.9669**	**−0.3627**	**−0.1123**
Traffic	malfunction	DRPO	**0.1538**	**−0.9971**	**−0.1176**
LunarLander	clean	PPO	**36.3**	**−0.0312**	**−0.0229**
LunarLander	clean	SPO	−2.6	**−0.0028**	**−0.0020**
LunarLander	clean	DRPO	**8.5**	**−0.0027**	**−0.0021**
LunarLander	state noise	PPO	**7.5**	**−0.0202**	**−0.0160**
LunarLander	state noise	SPO	−12.1	**−0.0052**	**−0.0031**
LunarLander	state noise	DRPO	−5.7	**−0.0047**	**−0.0030**
Grid2Op	clean	PPO	**3**	**−0.0330**	**−0.0368**
Grid2Op	clean	SPO	−400	**−0.0057**	**−0.0053**
Grid2Op	clean	DRPO	**403**	**−0.0169**	**−0.0119**
Grid2Op	contingency	PPO	**383**	**−0.0522**	**−0.0460**
Grid2Op	contingency	SPO	−360	**−0.0081**	**−0.0069**
Grid2Op	contingency	DRPO	**645**	**−0.0211**	**−0.0113**

during accidents, and by 48% and 64% during malfunctions. The same ordering holds in every external-domain condition. Table 3 connects these movement effects to task behavior. Relative to PFO, MART-PPO improves reward and queue in both accident and malfunction scenarios, with larger gains under accidents. Under normal traffic, it incurs small reward and queue costs while retaining substantially lower action-relevant movement. MART-PPO also sharply lowers LunarLander crash frequency in both conditions, while LunarLander returns vary across seeds. Grid2Op clean outcomes are indistinguishable; under line contingency, MART-PPO improves grid return and reduces blackout frequency. Across the three domains, action-relevant representation control provides stronger policy-movement control than raw feature proximity, as reported in Figure 4.

Table 3: MART-PPO minus PFO, averaged over five random seeds. Positive task changes and negative safety/control changes favor MART-PPO; bold marks every directionally favorable change. Results are averaging numeric and numeric+LLM modalities after pairing.

Environment	Condition	Task metric	Δ task	Control metric	Δ control
Traffic	normal	Reward	−0.02	Queue	0.29
Traffic	accident	Reward	**1.32**	Queue	**−14.86**
Traffic	malfunction	Reward	**0.62**	Queue	**−6.90**
LunarLander	clean	Return	**24.4**	Crash rate	**−0.07**
LunarLander	state noise	Return	**25.6**	Crash rate	**−0.26**
Grid2Op	clean	Return	0	Blackout rate	0.00
Grid2Op	contingency	Return	**1419**	Blackout rate	**−0.14**

4.7 Movement magnitude and direction are distinct

The mixed return results expose the distinction between movement magnitude and movement direction. The wrong-way movement diagnostic weights counterfactual movement by whether it opposes the estimated advantage. MART frequently lowers d_{rep} while increasing this fraction, especially for

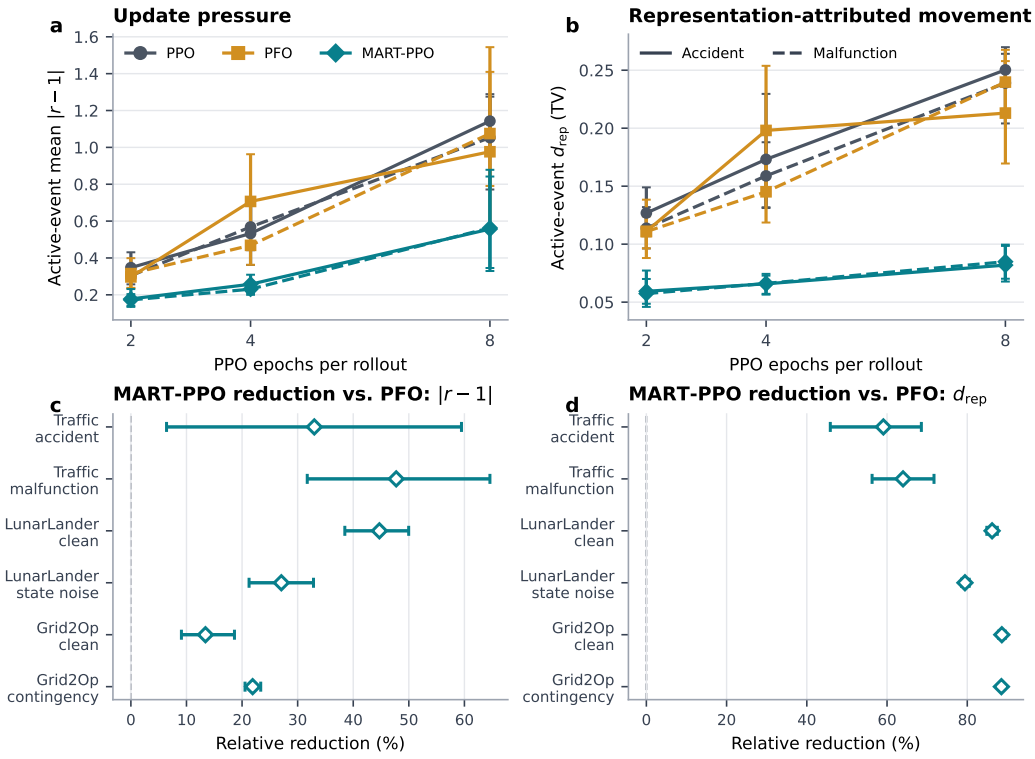

Figure 4: **Action-relevant versus feature-space regularization.** Panels (a,b) show traffic active-event movement separately for accidents (solid) and malfunctions (dashed), averaging numeric and numeric+LLM modalities after pairing. Panels (c,d) report the corresponding scenario-specific traffic reductions alongside the external-domain conditions. Values are the mean of the five seed-wise paired relative reductions of MART-PPO from PFO at eight epochs; positive values favor MART-PPO.

PPO in LunarLander and Grid2Op. In one clean Grid2Op MART-PPO run, the policy collapses late despite tightly controlled ratios and d_{rep}. MART-DRPO instead reduces wrong-way movement under line contingencies and obtains the strongest movement–performance trade-off. MART controls movement magnitude along the representation-attributed path; the base actor objective and advantage estimate still determine whether that movement has a useful direction.

5 Limitations and Conclusion

Limitations. The counterfactual attribution is defined relative to a chosen representation–head boundary and is therefore not invariant to arbitrary jointly compensated reparameterizations of these components. MART intentionally controls this path-specific component movement rather than defining an intrinsic decomposition of policy change. Extending the analysis to continuous-action policies and broader architectures is left for future work. Overall, trust-region analysis should ask not only how far a policy moves, but also where that movement originates.

Conclusion. We show that the complete-policy change can conceal substantial and compensating movement in the representation and action head. MART uses this path-specific attribution to constrain the action-distribution movement caused by representation updates. Across extensive environments, MART consistently reduces representation-attributed movement and generally improve control performance, with larger effects under stronger optimization pressure.

AI USE STATEMENT

Generative AI tools were used to provide feedback on experimental design and methodology, assist with method and software implementation, inspect and summarize experimental outputs, and support result interpretation. They were also used to edit code, draft and revise manuscript text, organize related-work notes, format tables, and prepare figures. The authors selected the research questions and methods, executed the experiments, tested AI-assisted code, checked reported numerical values against retained experiment logs and summary files, and reviewed all AI-assisted manuscript content. Generative AI was not used to generate experimental data or replace experimental execution. The authors take responsibility for the final text, claims, code, and artifacts in this work.

REFERENCES

Karl W. Cobbe, Jacob Hilton, Oleg Klimov, and John Schulman. Phasic policy gradient. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2020–2027. PMLR, 2021. URL https://proceedings.mlr.press/v139/cobbe21a.html.

Benjamin Donnot. Lightsim2grid: A c++ backend targeting the grid2op platform. https://github.com/Grid2op/lightsim2grid, 2020.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep policy gradients: A case study on PPO and TRPO. *arXiv preprint arXiv:2005.12729*, 2020. URL https://arxiv.org/abs/2005.12729.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338. PMLR, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Shangding Gu, Laixi Shi, Muning Wen, Ming Jin, Eric Mazumdar, Yuejie Chi, Adam Wierman, and Costas Spanos. Robust-gymnasium: A unified modular benchmark for robust reinforcement learning. *arXiv preprint arXiv:2502.19652*, 2025. URL https://arxiv.org/abs/2502.19652.

Hengyuan Hu and Dorsa Sadigh. Language instructed reinforcement learning for human-AI coordination. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13584–13598. PMLR, 2023. URL https://proceedings.mlr.press/v202/hu23e.html.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: Robot manipulation with multimodal prompts. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14975–15022. PMLR, 2023. URL https://proceedings.mlr.press/v202/jiang23b.html.

Aviral Kumar, Rishabh Agarwal, Tengyu Ma, Aaron Courville, George Tucker, and Sergey Levine. DR3: Value-based deep reinforcement learning requires explicit regularization. In *International Conference on Learning Representations*, 2022. URL https://iclr.cc/virtual/2022/poster/6927.

Hojoon Lee, Youngdo Lee, Takuma Seno, Donghu Kim, Peter Stone, and Jaegul Choo. Hyperspherical normalization for scalable deep reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33352–33403. PMLR, 2025. URL https://proceedings.mlr.press/v267/lee25u.html.

Antoine Marot, Benjamin Donnot, Gabriel Dulac-Arnold, Adrian Kelly, Aidan O'Sullivan, Jan Viebahn, Mariette Awad, Isabelle Guyon, Patrick Panciatici, Camilo Romero, et al. Learning to

run a power network challenge: A retrospective analysis. *arXiv preprint arXiv:2103.03104*, 2021. URL https://arxiv.org/abs/2103.03104.

Thomas Mesnard, Theophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Thomas S. Stepleton, Nicolas Heess, Arthur Guez, Eric Moulines, Marcus Hutter, Lars Buesing, and Remi Munos. Counterfactual credit assignment in model-free reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7654–7664. PMLR, 2021. URL https://proceedings.mlr.press/v139/mesnard21a.html.

Skander Moalla, Andrea Miele, Daniil Pyatko, Razvan Pascanu, and Caglar Gulcehre. No representation, no trust: Connecting representation, collapse, and trust issues in PPO. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/81166fbd9cc5adf14031cdb69d3fd6a8-Abstract-Conference.html.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897. PMLR, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Hongyao Tang, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, and Glen Berseth. Mitigating plasticity loss in continual reinforcement learning by reducing churn. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58883–58904. PMLR, 2025. URL https://proceedings.mlr.press/v267/tang25g.html.

Zhengpeng Xie, Qiang Zhang, Fan Yang, Marco Hutter, and Renjing Xu. Simple policy optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68813–68824. PMLR, 2025. URL https://proceedings.mlr.press/v267/xie25m.html.

Jiarui Yao, Xiangxin Zhou, Penghui Qi, Wee Sun Lee, Liefeng Bo, and Tianyu Pang. Rethinking the divergence regularization in LLM RL. *arXiv preprint arXiv:2606.09821*, 2026. URL https://arxiv.org/abs/2606.09821.

Huichu Zhang, Siyuan Feng, Chang Liu, Yaoyao Ding, Yichen Zhu, Zihan Zhou, Weinan Zhang, Yong Yu, Haiming Jin, and Zhenhui Li. Cityflow: A multi-agent reinforcement learning environment for large scale city traffic scenario. In *Proceedings of The Web Conference*, 2019.

A METHOD DETAILS

A.1 BACKBONE ACTOR GAINS

Let $r_t = \pi_{\theta,\psi}(a_t \mid x_t)/\pi_{\bar{\theta},\bar{\psi}}(a_t \mid x_t)$ and let $\hat{A}_t$ be a normalized advantage. The PPO gain is

$$j_{\mathrm{PPO},t} = \min\left(r_t\hat{A}_t, \mathrm{clip}(r_t, 1-\epsilon, 1+\epsilon)\hat{A}_t\right). \tag{11}$$

The implemented SPO and DRPO gains are

$$j_{\mathrm{SPO},t} = r_t\hat{A}_t - \frac{|\hat{A}_t|}{2\epsilon}(r_t - 1)^2, \tag{12}$$

$$j_{\mathrm{DRPO},t} = r_t\hat{A}_t - \frac{|\hat{A}_t|}{2\delta_D}\pi_{\bar{\theta},\bar{\psi}}(a_t \mid x_t)(r_t - 1)^2. \tag{13}$$

MART subtracts Equation 10's penalty from any of these gains without changing its backbone formula.

A.2 POLICY-INDUCED LOCAL GEOMETRY

MART measures representation movement after mapping it into action space. A forward-KL probe on the same frozen-head path gives a local geometric interpretation of this property. For distributions on the common valid-action support, let

$$D_{\mathrm{KL}}(p\|q) = \sum_a p(a)\log\frac{p(a)}{q(a)}. \tag{14}$$

Let $z_o = f_{\bar{\theta}}(x)$, $z_n = f_\theta(x)$, and $\Delta z = z_n - z_o$. Let J_g be the Jacobian of the frozen behavior-head logits $g_{\bar{\psi}}$ with respect to their representation input at z_o, and let

$$F(q_{oo}) = \mathrm{Diag}(q_{oo}) - q_{oo}q_{oo}^\top \tag{15}$$

be the categorical Fisher matrix on the common valid-action support. For a sufficiently small representation update,

$$D_{\mathrm{KL}}(q_{oo}\|q_{no}) = \frac{1}{2}\Delta z^\top J_g^\top F(q_{oo})J_g\Delta z + o(\|\Delta z\|_2^2). \tag{16}$$

For a linear-softmax behavior head with logits $W_{\mathrm{old}}z + b$, the local metric is $W_{\mathrm{old}}^\top F(q_{oo})W_{\mathrm{old}}$. Raw feature matching instead uses the isotropic metric I. Thus forward KL weights a representation direction according to its local effect on the behavior policy. Equation 16 explains the action geometry of the frozen-head path; the main MART-TV method obtains action relevance directly by measuring probability-mass movement after that head.

A.3 FORWARD-KL DIVERGENCE CHECK

For completeness, we replace the TV training distance in Equation 6 with $D_{\mathrm{KL}}(q_{oo}\|q_{no})$ while retaining the same counterfactual target and encoder-only penalty gradient. This dedicated traffic check uses $\lambda = 1$, $\delta = 0.05$, and the same values for both distances. Table 4 reports five-seed means over accident and malfunction active-event steps using only numeric and numeric+LLM inputs.

The two distances yield similar movement and reward in this matched check, with no consistent ordering across backbones. We therefore use TV throughout the main experiments for its direct probability-mass interpretation and common scale with the attribution diagnostics.

A.4 IMPLEMENTATION PSEUDOCODE

```
for each rollout:
    store actions and old log-probabilities
    store old fused representations
    snapshot the behavior action head
```

Table 4: TV and forward-KL training distances on the frozen-head representation path. Lower movement is favorable; reward is higher-is-better.

Backbone	Training distance	$\lvert r-1\rvert$	d_{rep} (TV)	Reward
PPO	TV	0.4091	0.1379	-12.090
PPO	Forward KL	0.4065	0.1377	-12.194
SPO	TV	0.2905	0.1143	-11.985
SPO	Forward KL	0.2868	0.1136	-11.901
DRPO	TV	0.3327	0.1272	-11.852
DRPO	Forward KL	0.3372	0.1278	-12.004

```
for each PPO epoch and minibatch:
    z_new = encoder_current(observation)
    q_oo = policy(head_old, z_old)
    q_no = policy(head_old, z_new)
    q_on = policy(head_current, z_old)
    q_nn = policy(head_current, z_new)

    mart_distance = total_variation(q_oo, q_no)
    mart_penalty = normalized_hinge(mart_distance, delta)
    base = mean(base_actor_gain)
    objective = base - lambda * mean(mart_penalty)
    update current encoder and head
```

The behavior head and q_{oo} are detached. The MART penalty therefore updates the encoder but not the current head.

A.5 Wrong-way drift diagnostic

Movement magnitude does not determine whether the sampled action moves in a direction favored by the estimated advantage. For each rollout sample, define the representation-attributed selected-action shift

$$\Delta \log \pi_t^{\text{rep}} = \log q_{no,t}(a_t) - \log q_{oo,t}(a_t). \tag{17}$$

With an advantage dead-zone $\tau_A \geq 0$, the logged wrong-way drift rate is

$$\text{WDR} = \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \mathbf{1}\Big[|\hat{A}_t| > \tau_A,\ \hat{A}_t \Delta \log \pi_t^{\text{rep}} < 0\Big]. \tag{18}$$

We also report the advantage-weighted statistic

$$\text{WDR}_w = \frac{\sum_{t \in \mathcal{B}} |\hat{A}_t| \mathbf{1}[|\hat{A}_t| > \tau_A,\ \hat{A}_t \Delta \log \pi_t^{\text{rep}} < 0]}{\max\Big(\epsilon_0, \sum_{t \in \mathcal{B}} |\hat{A}_t| \mathbf{1}[|\hat{A}_t| > \tau_A]\Big)}. \tag{19}$$

Advantages and counterfactual shifts are detached for these statistics. WDR is diagnostic only and is not included in the training objective.

A.6 Diagnostic identities and checks

The implementation verifies that q_{oo} reproduces the stored behavior distribution and q_{nn} reproduces the current policy to numerical precision. Each distribution is finite, normalized, and mask-consistent. We log both representation-first and head-first paths, cancellation gaps, boundary activation, penalty scale, and wrong-way movement (logged as WDR). The stored path and cancellation quantities are TV diagnostics. MART trains with the configured d_{reg}.

B EXPERIMENTAL CONFIGURATION

B.1 FEATURE-SPACE BASELINE

PFO uses the PPO clipped objective with an L2 trust region on the actor's `fused_preactivation` feature. Across all three environments we use coefficient 1.0, trust-region limit 0.1, regularization mode, and global application. PFO and MART-PPO share the same training budget, PPO epoch count, policy architecture, action masking, and paired seeds within each comparison. Thus their experimental difference is the regularization target: Euclidean preactivation displacement for PFO versus frozen-head action-distribution movement for MART.

B.2 OPTIMIZATION, ARCHITECTURE, AND REPORTING

All methods use Adam, normalized GAE advantages, discount $\gamma = 0.99$, GAE parameter 0.95, PPO clipping threshold 0.2, minibatches of 64, and gradient-norm clipping at 0.5. Table 5 lists the settings that differ by environment. The objective coefficients apply to the value loss and entropy bonus, respectively.

Table 5: Environment-specific implementation and evaluation settings.

Environment	Learning rate	Hidden width	Value coef.	Entropy coef.	Evaluation protocol
Traffic	10^{-4}	128	0.05	0.05	Stage 80; 5 diagnostic rollouts
LunarLander	3×10^{-4}	64	0.5	0.01	Every 25 iterations; 20 final episodes
Grid2Op	3×10^{-4}	128	0.5	0.01	Every 50 iterations; 20 final episodes

The categorical actor uses a Tanh numeric encoder, layer normalization, a Tanh fusion layer, and a linear action head. Traffic text policies concatenate the numeric representation with a 32-dimensional trainable projection of frozen, cached `all-MiniLM-L6-v2` sentence embeddings before fusion. The critic is a two-hidden-layer Tanh network whose input is detached from the actor encoder, preventing the value loss from updating the representation used by the actor. Grid2Op applies its valid-action mask before every categorical distribution, including the four counterfactual distributions.

The five seeds are $\{0, 21, 26, 33, 88\}$. LunarLander and Grid2Op task metrics come from the final 20-episode evaluation, while their movement statistics average the final 100 training iterations. Traffic accident and malfunction metrics use stage-80 rare rollouts restricted to active-event steps; normal traffic uses the stage-80 ordinary-state window. Each scenario is reported in the main cross-domain table after pairing by seed and modality, then averaging the numeric and numeric+LLM modalities. The modality-stratified traffic analysis retains these inputs separately. Paper-facing paired intervals use 20,000 bootstrap resamples; auxiliary environment summaries use 10,000. Exact archive paths, filters, and figure inputs are fixed in `scripts/build_paper_assets.py` and `scripts/build_pfo_comparison_figure.py`; each run directory also retains its merged `config.yaml` and per-iteration metrics.

Table 6: Common experimental budgets.

Environment	Iterations	Rollout	PPO epochs	Seeds	Conditions
Traffic	500	128	2/4/8	5	normal, accident, malfunction
LunarLander	1000	1024	8	5	clean, state noise
Grid2Op	500	1024	8	5	clean, line contingency

Traffic primary modalities are `numeric` and `numeric_llm_text`. The normal scenario injects no event. The accident reduces the capacity of approach 2 to 10% from time 300 to 600. The malfunction acts over the same interval, omitting approach 1 with probability 0.8 and adding observation noise with standard deviation 0.30. Accident and malfunction results use stage-80 rare rollouts and the active-event window; normal results use the stage-80 ordinary-state window. LunarLander state noise is Gaussian with standard deviation 0.1. Grid2Op uses a line-only catalog: one no-op plus line disconnection and reconnection actions. Invalid Grid2Op actions are masked before constructing every old/new distribution.

Table 7: MART-TV coefficients. All use uniform weighting and global regularization.

Environment	Backbone	λ	δ
Traffic epoch pressure	PPO/SPO/DRPO	0.03	0.005
LunarLander	PPO	0.01	0.005
LunarLander	SPO/DRPO	0.003	0.01
Grid2Op line-only	PPO	0.02	0.005
Grid2Op line-only	SPO	0.003	0.01
Grid2Op line-only	DRPO	0.02	0.03

Traffic DRPO-specific experiments use $\lambda = 0.01$ and $\delta = 0.005$.

C Additional Traffic Results

C.1 Full epoch-eight uncertainty

Table 8: Traffic epoch-eight MART-TV minus base. Each entry reports the estimate followed by its paired-bootstrap 95% confidence interval in parentheses. Bold marks directionally favorable MART changes.

Backbone	$\Delta\lvert r-1\rvert$	Δ clipping	Δd_{rep}	Δ reward	Δ queue
PPO	$\mathbf{-0.538}$ ($\mathbf{-0.742}$, $\mathbf{-0.329}$)	$\mathbf{-0.128}$ ($\mathbf{-0.258}$, $\mathbf{-0.043}$)	$\mathbf{-0.161}$ ($\mathbf{-0.176}$, $\mathbf{-0.146}$)	$\mathbf{1.253}$ ($\mathbf{0.293}$, $\mathbf{2.288}$)	$\mathbf{-14.13}$ ($\mathbf{-25.82}$, $\mathbf{-3.30}$)
SPO	$\mathbf{-0.309}$ ($\mathbf{-0.440}$, $\mathbf{-0.173}$)	$\mathbf{-0.152}$ ($\mathbf{-0.197}$, $\mathbf{-0.106}$)	$\mathbf{-0.098}$ ($\mathbf{-0.138}$, $\mathbf{-0.071}$)	$\mathbf{1.964}$ ($\mathbf{1.310}$, $\mathbf{2.629}$)	$\mathbf{-21.94}$ ($\mathbf{-29.46}$, $\mathbf{-14.53}$)
DRPO	$\mathbf{-0.739}$ ($\mathbf{-1.438}$, $\mathbf{-0.358}$)	$\mathbf{-0.127}$ ($\mathbf{-0.252}$, $\mathbf{-0.035}$)	$\mathbf{-0.114}$ ($\mathbf{-0.145}$, $\mathbf{-0.091}$)	$\mathbf{0.061}$ ($\mathbf{-0.835}$, $\mathbf{0.882}$)	$\mathbf{-0.70}$ ($\mathbf{-10.02}$, $\mathbf{9.40}$)

C.2 MART effects separated by traffic modality

Table 9: Traffic MART-minus-base effects separated by observation modality. Normal rows average the three backbone cells; abnormal rows average the six backbone–scenario cells from accident and malfunction. All entries are paired by seed before macro-averaging. Positive reward and negative queue and movement changes are favorable; bold marks every favorable change.

Condition	Modality	Δ reward	Δ queue	$\Delta\lvert r-1\rvert$	Δd_{rep}
Normal	Numeric	-0.004	0.04	$\mathbf{-0.507}$	$\mathbf{-0.130}$
Normal	Numeric+LLM	-0.014	0.18	$\mathbf{-0.485}$	$\mathbf{-0.122}$
Abnormal	Numeric	$\mathbf{1.114}$	$\mathbf{-12.47}$	$\mathbf{-0.584}$	$\mathbf{-0.129}$
Abnormal	Numeric+LLM	$\mathbf{1.072}$	$\mathbf{-12.05}$	$\mathbf{-0.474}$	$\mathbf{-0.120}$

The mechanism holds in every separate movement cell: MART lowers both $|r-1|$ and d_{rep} for all three backbones under normal traffic and for all backbone–scenario combinations under abnormal traffic, in both modalities. The similar aggregate effects do not establish a stronger MART benefit for numeric+LLM policies. Instead, together with Figure 2, they show why multimodal policies require an action-aware target: larger latent movement can have a nonuniform policy consequence, while MART directly controls the action-distribution movement attributable to that representation.

C.3 Cell consistency at eight epochs

At epoch eight, MART also reduces fused-representation movement by 0.141 for PPO, 0.075 for SPO, and 0.124 for DRPO. Extreme-ratio frequency decreases by 0.179, 0.143, and 0.118, respectively. These paired macro means use only the numeric and numeric+LLM modalities.

Table 10: Number of improved accident/malfunction by numeric/numeric+LLM cells out of four at eight epochs.

Backbone	$\lvert r-1\rvert$	Clip	$r>2$	d_{rep}	d_{total}	Reward	Queue
PPO	4	4	4	4	4	4	4
SPO	4	4	4	4	4	4	4
DRPO	4	4	4	4	4	3	3

Table 11: MART minus base in LunarLander. Parentheses contain paired-bootstrap 95% confidence intervals when shown. Bold marks directionally favorable MART changes.

Condition	Backbone	Δ return	Δ success	$\Delta\lvert r-1\rvert$	Δd_{rep}
Clean	PPO	$\mathbf{36.3}$ ($\mathbf{15.6}$, $\mathbf{57.1}$)	$\mathbf{0.22}$ ($\mathbf{0.12}$, $\mathbf{0.32}$)	$\mathbf{-0.0312}$	$\mathbf{-0.0229}$
Noise	PPO	$\mathbf{7.5}$ ($\mathbf{-27.1}$, $\mathbf{37.7}$)	-0.07	$\mathbf{-0.0202}$	$\mathbf{-0.0160}$
Clean	SPO	-2.6 (-18.5, 15.9)	0.00	$\mathbf{-0.0028}$	$\mathbf{-0.0020}$
Noise	SPO	-12.1 (-91.8, 53.2)	-0.07	$\mathbf{-0.0052}$	$\mathbf{-0.0031}$
Clean	DRPO	$\mathbf{8.5}$ ($\mathbf{-17.1}$, $\mathbf{34.3}$)	$\mathbf{0.05}$	$\mathbf{-0.0027}$	$\mathbf{-0.0021}$
Noise	DRPO	-5.7 (-73.7, 98.0)	$\mathbf{0.07}$	$\mathbf{-0.0047}$	$\mathbf{-0.0030}$

D Additional Cross-Domain Results

D.1 LunarLander

The noisy task remains difficult: many runs trade crashes for timeouts rather than successful landings. We therefore use clean PPO as the strongest performance result and noisy conditions as mechanism and failure-mode evidence.

D.2 Grid2Op

For MART with DRPO, clean and contingency ratio, clipping, d_{rep}, and d_{total} all improve in every paired seed. Mean grid-return effects are 403 clean and 645 under contingency; survival changes are 0.02 and 0.09. These values average five random seeds. Late contingency checkpoints remain variable, leaving these favorable task effects uncertain.

The following additional MART-SPO results use $\lambda = 0.02$ and $\delta = 0.03$; Table 2 uses $\lambda = 0.003$ and $\delta = 0.01$, as listed in Appendix B. MART with SPO has a significant task cost: approximately -402 clean and -753 under contingency. Under contingency it still lowers $|r-1|$ by 0.0030 and d_{rep} by 0.0024. The result cleanly separates smaller representation-attributed movement from better control.