SPHERE: Mitigating the Loss of Spectral Plasticity in Mixture-of-Experts for Deep Reinforcement Learning

Lirui Luo, Guoxi Zhang, Hongming Xu, Cong Fang, Qing Li

A practical Parseval penalty that keeps MoE policy updates spectrally diverse during continual RL.

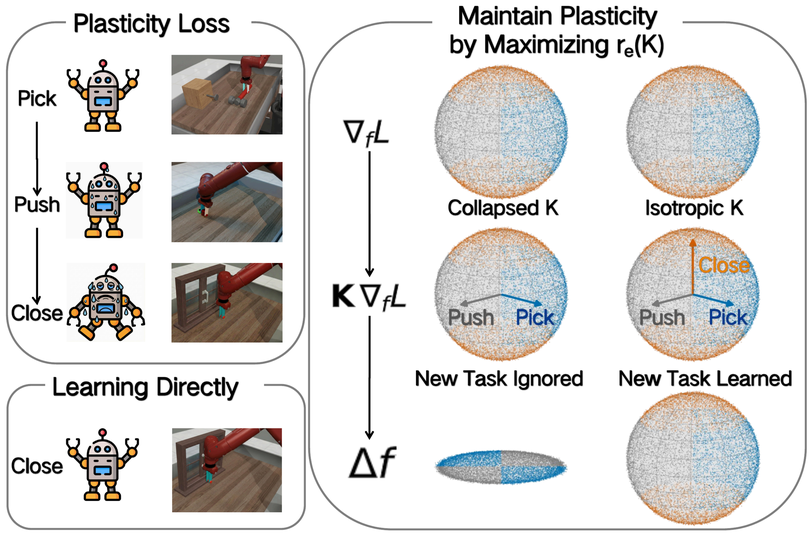
ICML 2026 Accepted

Motivation and Gap

MoE policies add capacity, yet continual RL still loses the ability to update in new directions.

- **Plasticity loss:** later tasks fail even when each task is learnable in isolation.
- **MoE gap:** expert routing does not by itself prevent update collapse.
- **Goal:** preserve functional update diversity in MoE policies during RL and CRL.

Plasticity Loss as Spectral Collapse



Paper Fig. 1. Low-rank spectra filter new-task gradients; isotropic spectra keep multiple update directions available.

Experimental Setup

Protocols. RL trains each task independently; CRL resumes one policy through the task sequence.
Benchmarks. MetaWorld CW10 and HumanoidBench H1.
Metrics. Final success and eNTK effective rank r_e(K); higher r_e(K) means broader functional updates.

Contributions

- **Lens:** formalize plasticity loss as loss of spectral plasticity using eNTK effective rank.
- **Method:** derive a tractable feature-Gram proxy and introduce SPHERE.
- **Evidence:** show improved success and sustained spectral plasticity on MetaWorld and HumanoidBench.

Spectral Plasticity

The empirical NTK K maps gradients to functional change:

$$\Delta f = -\eta K \nabla_f L$$

We quantify update breadth by spectral-entropy effective rank:

$$r_e(K) = \exp(-\sum_i p_i \log p_i),\ p_i = \lambda_i / \sum_j \lambda_j$$

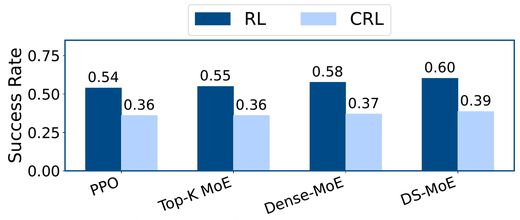
Collapsed spectrum

low r_e(K)

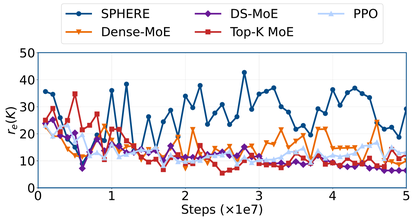
Isotropic spectrum

high r_e(K)

Evidence of Spectral Plasticity Loss



Paper Fig. 2. CRL degrades success relative to independent RL across dense PPO and MoE policies.



Paper Fig. 3. Baselines lose eNTK effective rank, while SPHERE maintains higher spectral plasticity.

SPHERE Method

Fix: regularize the last hidden expert features of the MoE actor before functional update directions collapse.

① Actor MoE **last expert features** → ② weighted feature Gram **cross-expert correlations** → ③ Parseval target **flatten spectrum**

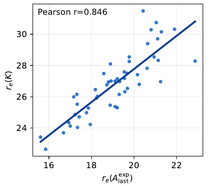
SPHERE keeps routing-weighted features isotropic, so continual updates keep multiple directions alive.

$$\mathcal{L}_{SPHERE}(A) = \|A - \tfrac{\mathrm{Tr}(A)}{m} I\|_F^2 \qquad \mathcal{L} = \mathcal{L}_{PPO} + \lambda^e \mathcal{L}_{SPHERE}$$

Theory. The penalty contracts the anisotropic component of the feature Gram.

Practicality. It avoids explicitly forming the eNTK during PPO updates.

Feature Isotropy Tracks Spectral Plasticity

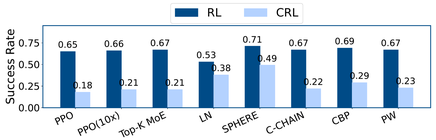


Feature Gram proxy. Higher r_e(A_last) predicts higher eNTK effective rank r_e(K).

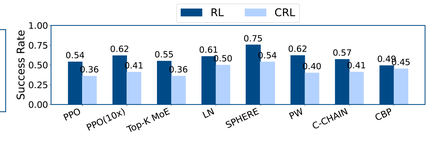
The last-layer feature spectrum is therefore a practical surrogate for tracking spectral plasticity during training.

Paper Fig. 7. Positive co-variation between r_e(A_last) and r_e(K).

Main Results

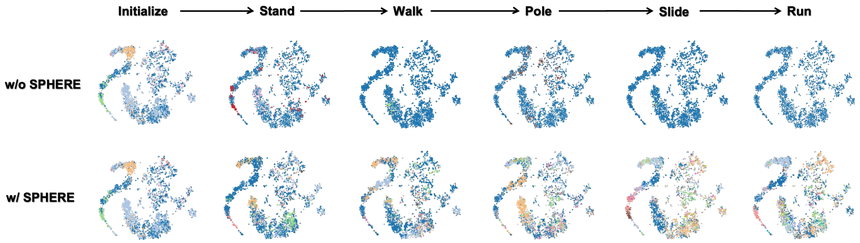


Paper Fig. 4. MetaWorld CRL success improves from 0.21 to 0.49 over Top-K MoE.



Paper Fig. 5. HumanoidBench success improves by 36% under RL and 50% under CRL.

Qualitative Evolution



Paper Fig. 6. Without SPHERE, states concentrate on one direction. With SPHERE, multiple components stay active.

Design Choices

Layer placement. Regularizing the last hidden expert layer outperforms applying the penalty to all hidden expert layers.
Cross-expert structure. Concatenated expert features preserve cross-expert correlation terms; per-expert penalties are weaker.
Feature Gram. Shaping A_last is more practical and stronger than gradient-Gram regularization.

Conclusion

- **Bottleneck:** MoE-CRL can collapse into few functional update directions.
- **Principle:** maintain spectral plasticity by flattening the routing-weighted expert-feature spectrum.
- **Result:** SPHERE improves success while sustaining eNTK effective rank on both benchmarks.