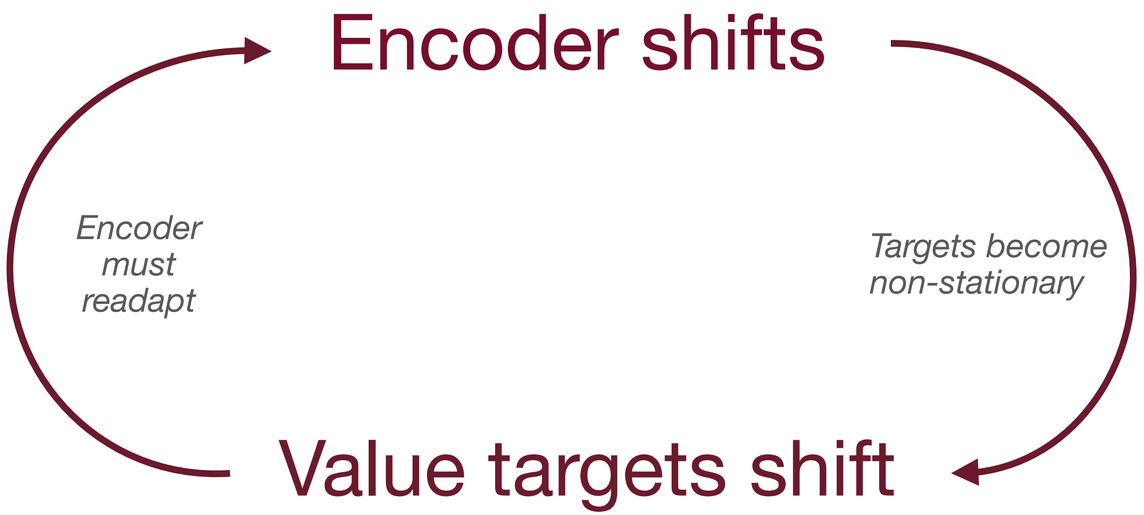
ICLR
International Conference On
Learning Representations

Stackelberg Coupling of Online Representation Learning and Reinforcement Learning

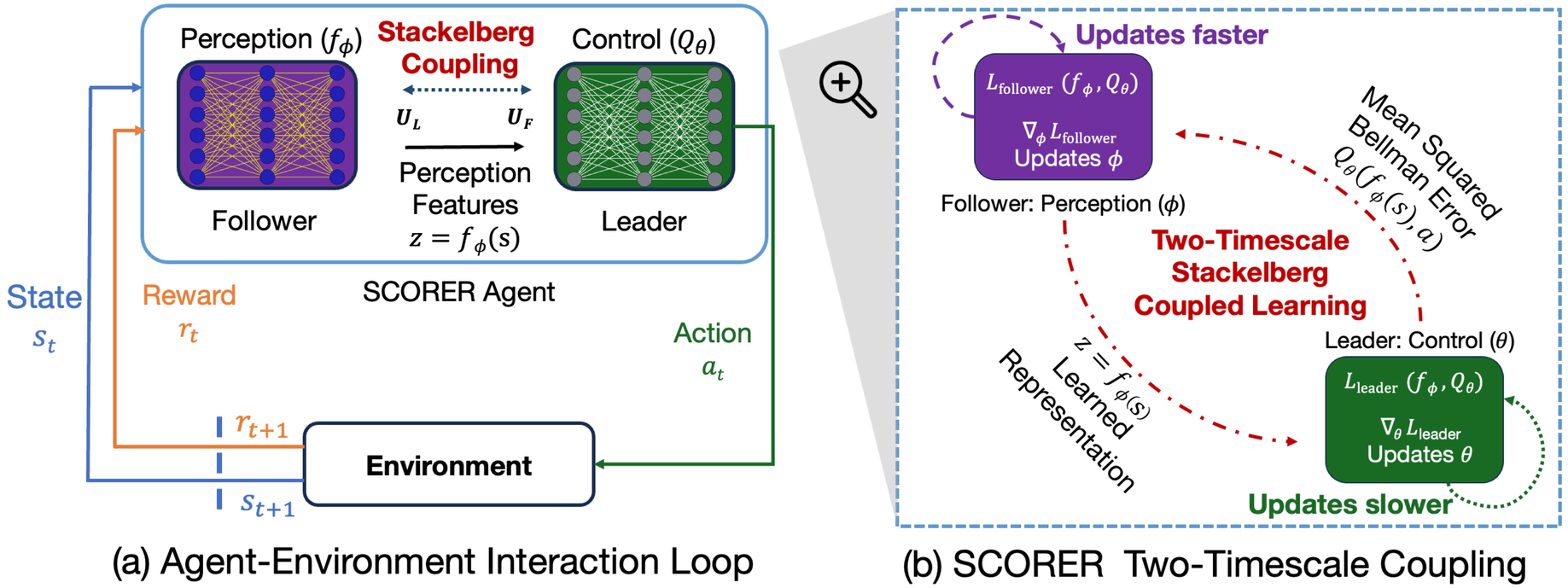
Fernando Martinez, Tao Li, Yingdong Lu, Juntao Chen

A Circular Dependency



+ Auxiliary losses attempt a fix; but conflicting gradients introduce a new obstacle to value learning.

A Hierarchical (Stackelberg) Division of Labor



(a) Agent-Environment Interaction Loop

(b) SCORER Two-Timescale Coupling

The Follower's Fast Update (Perception)

It tracks the best response to the leader's fixed strategy ($\overline{\theta_k}$) by minimizing Bellman Error Variance:

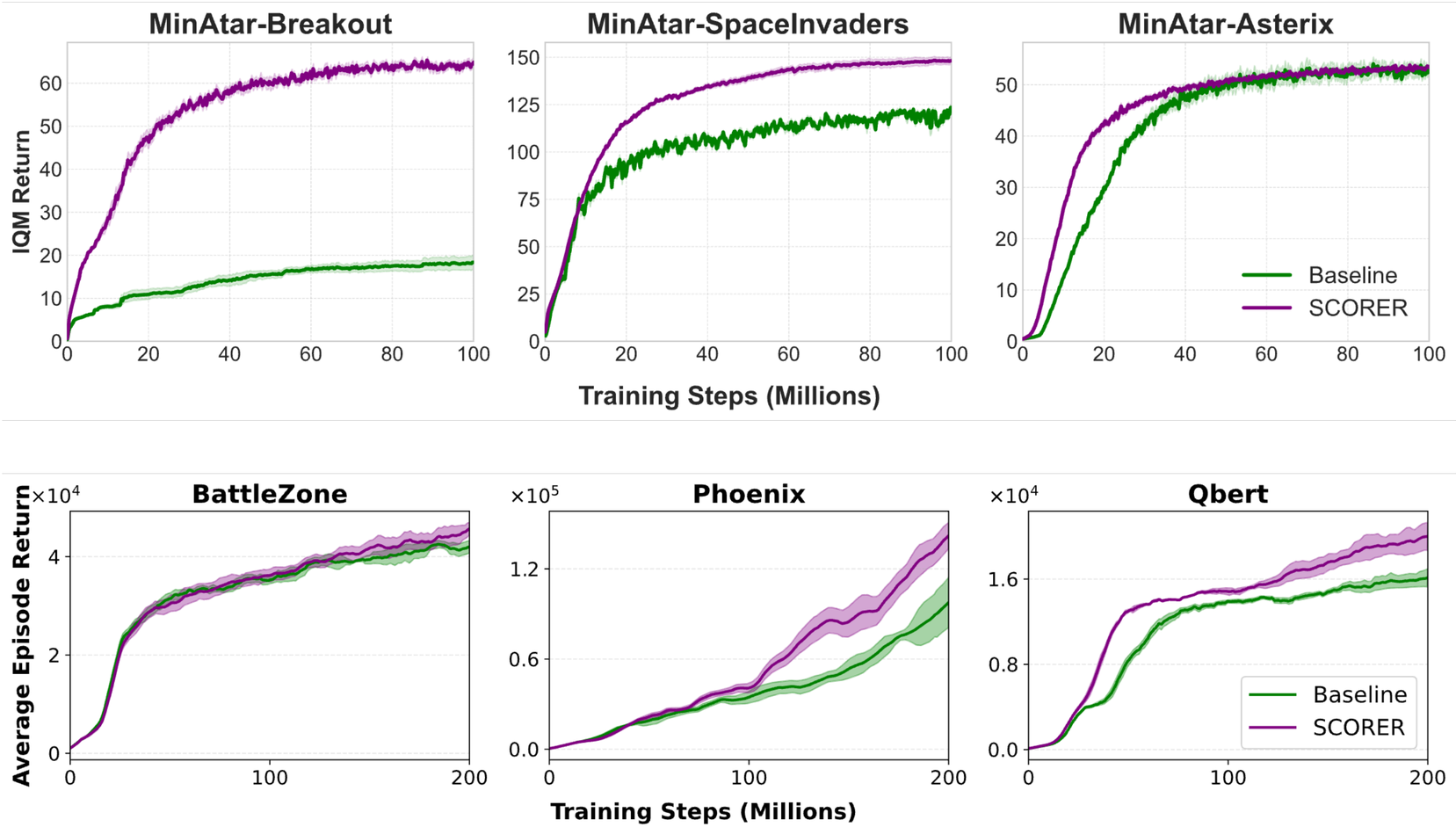
$$\phi_{k+1} \leftarrow \phi_k - \alpha_{\phi,k} \nabla_\phi \mathscr{L}_{\text{follower}}(\phi_k; B_{\text{follower}}, Y, \overline{\theta_k})$$

The Leader's Slow Update (Control)

It updates based on the follower's new, stable representation ($\overline{\phi_{k+1}}$) to minimize MSBE:

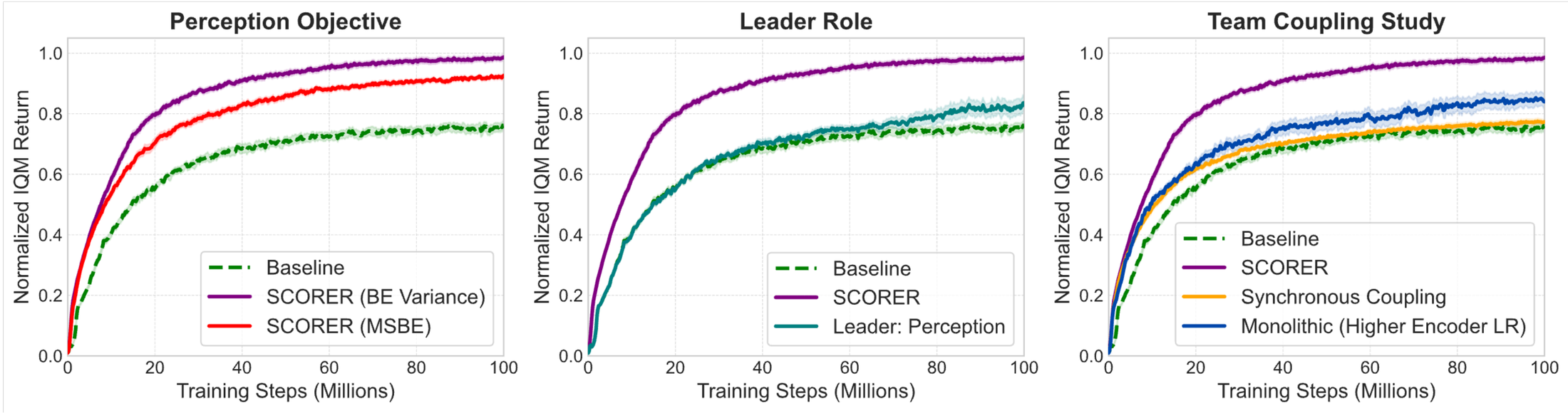
$$\theta_{k+1} \leftarrow \theta_k - \alpha_{\theta,k} \nabla_\theta \mathscr{L}_{\text{leader}}(\theta_k; B_{\text{leader}}, Y, \overline{\phi_{k+1}})$$

Gains Across Benchmarks



Every Design Choice Matters

All three design choices are necessary and complementary



- **Roles:** inversion of roles collapses performance
- **Objective**: variance beats MSBE
- **Coupling**: hierarchy beats synchronous

FORDHAM UNIVERSITY

CityU 香港城市大學 City University of Hong Kong

IBM

Thank you!

Takeaways

- Stackelberg coupling provides a principled mechanism for stable co-adaptation in value-based RL
- Minimizing Bellman error variance directly addresses the high-variance pathology of TD targets
- Consistent gains across DQN, DDQN, Dueling, R2D2, and PQN — no architectural changes required