Real-DRL: Teach and Learn in Reality

Yanbing Mao[1,*], Yihao Cai[1,*], Lui Sha[2]

[1] Wayne State University, [2] University of Illinois Urbana-Champaign

[*] Indicates Equal Contribution

Scan To Know More

The Thirty-Ninth Annual Conference on Neural Information Processing Systems

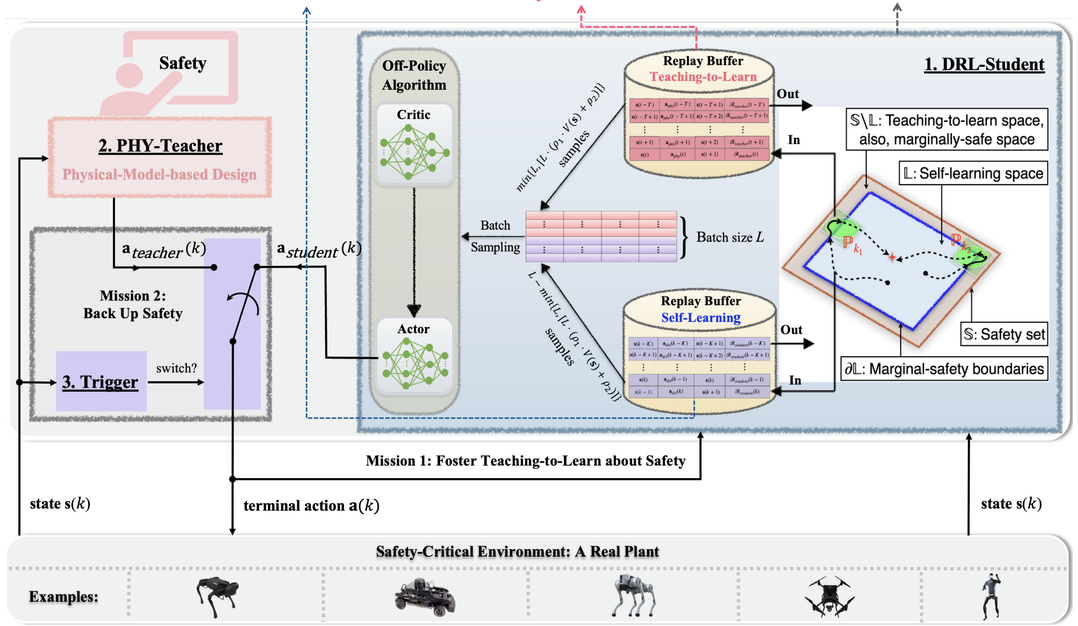
Motivation and Challenge

Applications: AI-enabled Autonomous Systems

- **Runtime Learning Safety**
  - The risky nature of trial-and-error exploration in Deep Reinforcement Learning
  - Learning in real-world (e.g., unknown environments) requires timely and adaptive responses
- **Safety-related Data Imbalance Issue**
  - Underrepresentation of rare but crucial data → poor safety performance at critical moments
  - Leading to training bias and limited generalization capability
- **Sampling Efficiency**
  - High-quality data fosters efficient and safe learning
  - Inefficient sampling prolongs training, and increases runtime safety risks

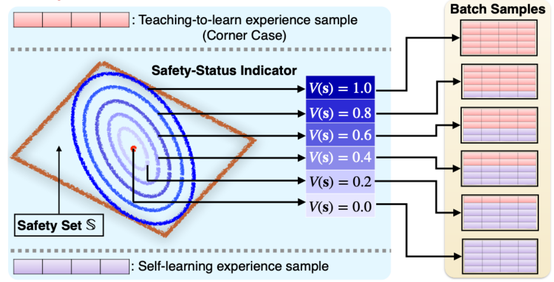
Framework Overview

Real-world Data + Safety-critical Data + Novel RL Architecture



Three Core Components

Component-I: DRL-Student



Safety-status indicator

$V(s) \triangleq s^T \cdot P \cdot s$

s – *real time state s(t)*

Total Sampled Batch Size

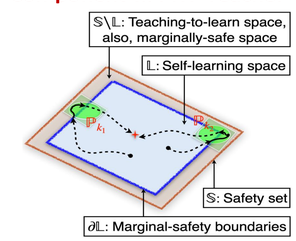
$L = L_t + L_s$

$L_t = \min\{L, [L\cdot(\rho_1\cdot V(s(t))+\rho_2]\}$

$L_s = L - \min\{L, [L\cdot(\rho_1\cdot V(s(t))+\rho_2]\}$

ρ_1, ρ_2 – *hyperparameters*

Component-II: PHY-Teacher



System Dynamics: $s(t+1) = f(s(t), a(t)), t \in \mathbb{N}$

Safety Set: $\mathbb{S} \triangleq \{s \in \mathbb{R}^n \mid -c < C \cdot s < c\}$

Self-Learning Space: $\mathbb{L} \triangleq \{s \in \mathbb{R}^n \mid -\eta \cdot c < C \cdot s < \eta \cdot c\}, 0 < \eta < 1$

Real-time Patch Design $\mathbb{P}_k$

① **Control Goal** : $s_k^* \triangleq \chi \cdot s(k)$, $s(k\text{-}1) \in \mathbb{L}$ and $s(k) \in \partial\mathbb{L}$

② **LMI Feasibility** : Construct LMIs and optimize for F_k

③ **Action Policy** : $a_{teacher}(t) = F_k \cdot (s(t) - s_k^*)$, $t \in \mathbb{T}_k$

Component-III: Trigger

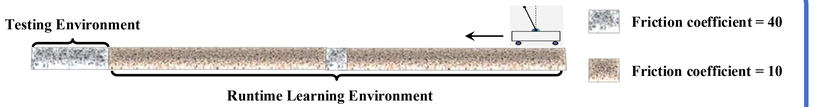
Triggering Condition $\mathcal{T}$: $s(k\text{-}1) \in \mathbb{L}$ and $s(k) \in \partial\mathbb{L}$

PHY-Teacher Action Step: $\mathbb{T}_k \triangleq \{k+1, k+2, \ldots, k+\tau_k\}$

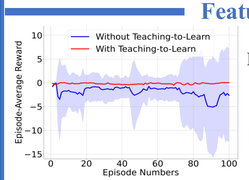
Switching Law

$$a(t) = \begin{cases} a_{teacher}(t), & \text{if } \mathcal{T} \text{ holds at } k, \text{ and } t \in \mathbb{T}_k \\ a_{student}(t), & \text{if } s(k) \in \mathbb{L} \end{cases}$$

Three Notable Features



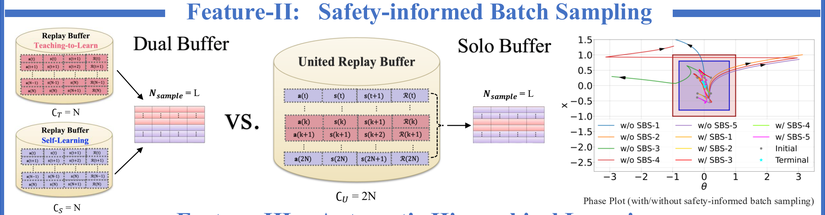
Feature-I: Teaching-to-Learn Mechanism



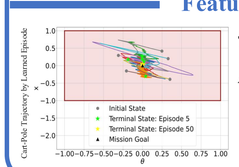
Episode-Average Reward = $\frac{\text{Return (i.e., cumulative reward) in one episode}}{\text{DRL-Student's total activation time in one episode}}$

Adopting Teaching-to-Learn paradigm leads to overall improved episode-average reward and stable learning

Feature-II: Safety-informed Batch Sampling



Feature-III: Automatic Hierarchical Learning



Task Goal: $(\bar{x}^*, \bar{\theta}^*) = (0, 0)$

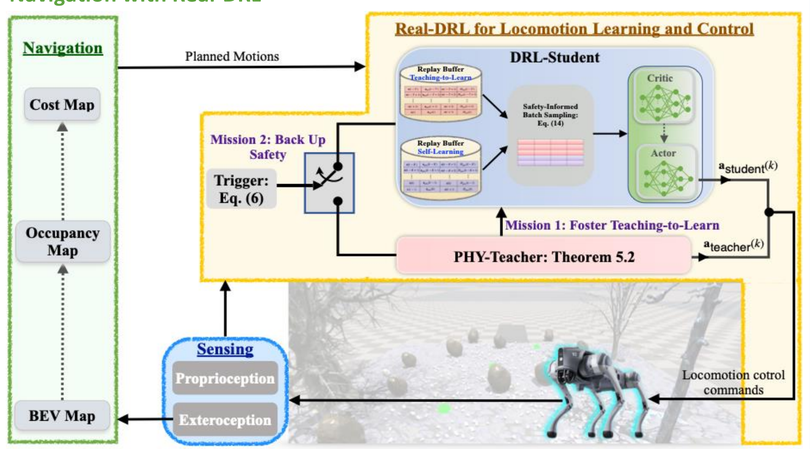
Automatic Hierarchical Learning:

Safety-first (Episode 5) ⇢ **High-Performance** (Episode 50)

DRL-Student become independent of **PHY-Teacher**

Experiment (Wild Environment)

Navigation with Real-DRL

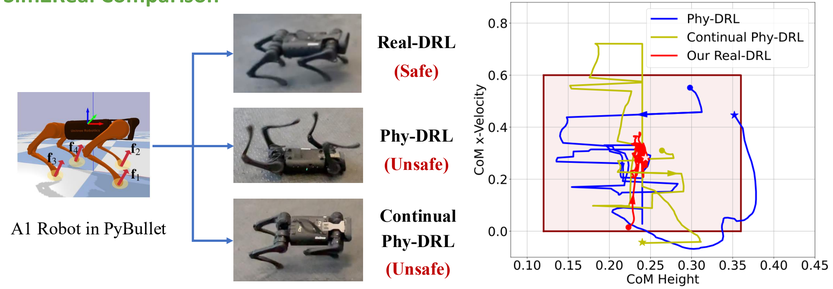


Comparison with SOTA

Model ID	Navigation Performance					Energy Efficiency	
	Success	Is Fall	Collision	Num (wp)	Travel Time (s)	Avg Power (W)	Total Energy (J)
CLF-DRL	No	Yes	No	0	N/A	N/A	N/A
Phy-DRL	No	No	Yes	1	∞	507.9441	∞
Runtime Assurance	No	Yes	No	2	N/A	N/A	N/A
Neural Simplex	No	No	Yes	2	∞	487.9316	∞
PHY-Teacher	**Yes**	No	No	**4**	55.5327	482.8468	26817.68
Our Real-DRL	**Yes**	No	No	**4**	**45.3383**	**479.4638**	**21742.42**

Experiment (Real Environment)

Sim2Real Comparison



Conclusion

- ***Core Contribution***
  - Real-world data collection from the unknown and hard-to-predict environment
  - Safety-critical data generation enabled by a verifiable and robust PHY-Teacher
  - An innovative RL architecture that supports modular and extensible design
- ***Three Notable Features***
  - Teaching-to-learn Mechanism (foster safe learning and fast convergence)
  - Automatic Hierarchy Learning (learn safety first and high-performance policy)
  - Safety-informed batch sampling (resolve data imbalance caused by corner cases)