中国科学院大学
University of Chinese Academy of Sciences

NEURAL INFORMATION PROCESSING SYSTEMS

Rainbow Delay Compensation: A Multi-Agent Reinforcement Learning Framework for Mitigating Delayed Observation

Songchen Fu, Siang Chen, et al.

Laboratory of Speech and Intelligent Information Processing, Institute of Acoustics, University of Chinese Academy of Sciences

Department of Electronic Engineering, Tsinghua University

NEURAL INFORMATION PROCESSING SYSTEMS

Background

- Multi-agent RL widely applied
- Delays ubiquitous in real-world
- Delays in MAS more complex
  - Different observation components have different delay characteristics
  - Stochastic delays more impactful than fixed delays

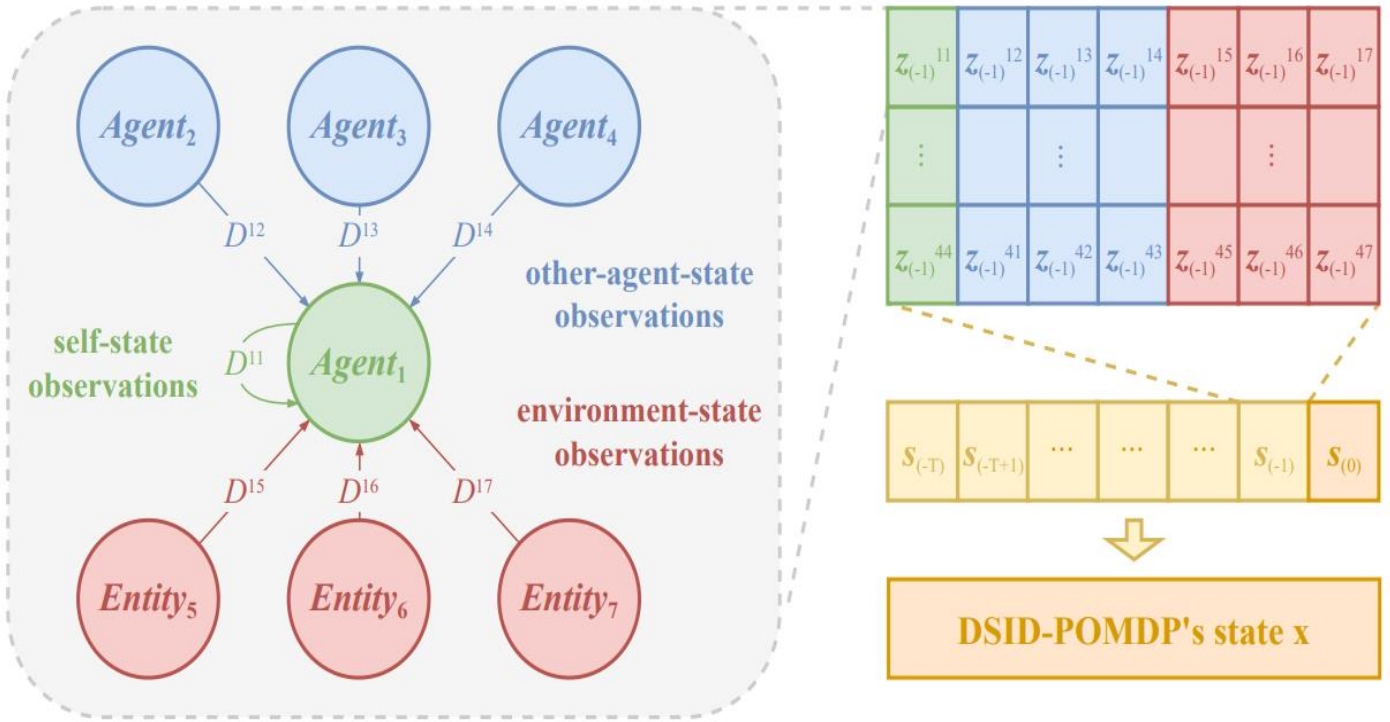
NEURAL INFORMATION PROCESSING SYSTEMS

Main Contributions

- Define DSID-POMDP
- Propose RDC training framework with four core components
- Introduce two compensator modes (*Echo* and *Flash*)
- Validate on MPE and SMAC, achieving near delay-free performance

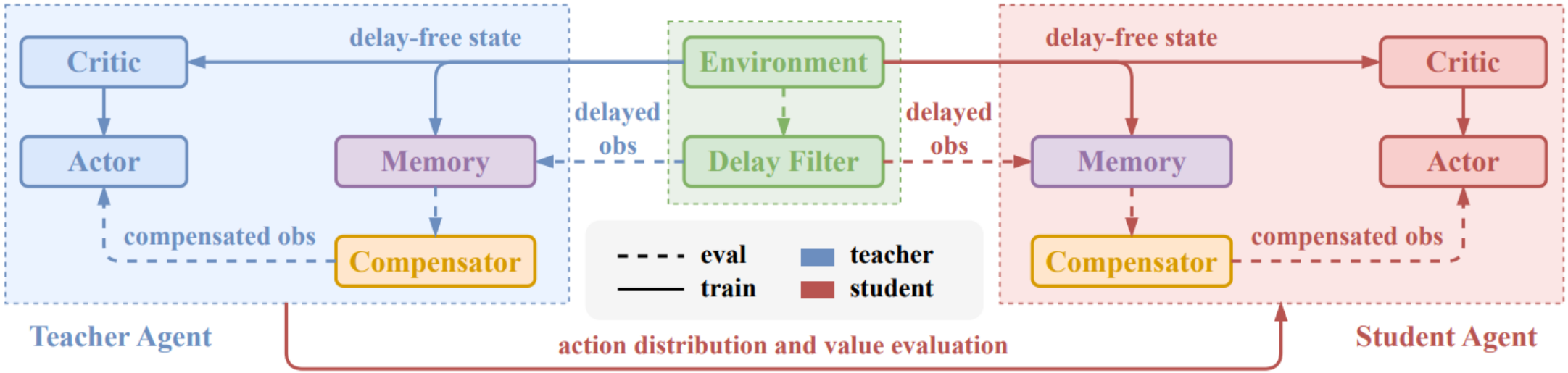
Problem Formulation

- **DSID-POMDP:** Decentralized Stochastic Individual Delay POMDP
- **Key features:**
  - Each agent's observation contains multiple components
  - Different components have independent delay distributions
  - Extended state space includes historical states



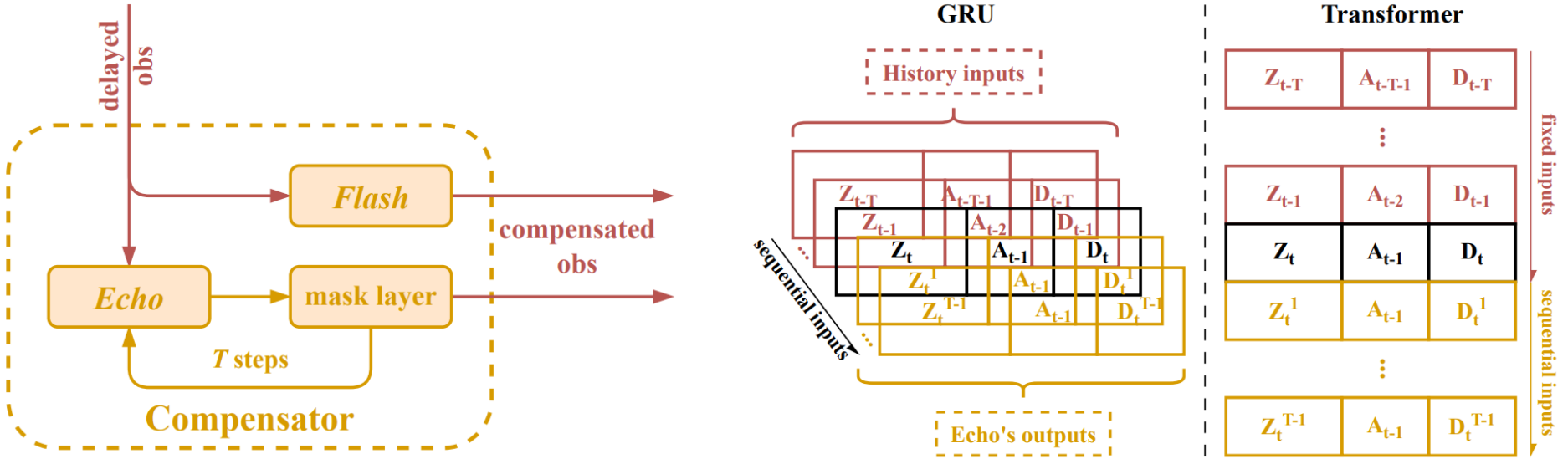
RDC Framework Overview

- **Delay Compensator** (reconstructs delay-free observation)
- **Delay-reconciled Critic** (uses delay-free states)
- **Curriculum Learning Actor** (gradual transition)
- **Knowledge Distillation** (teacher-student learning)



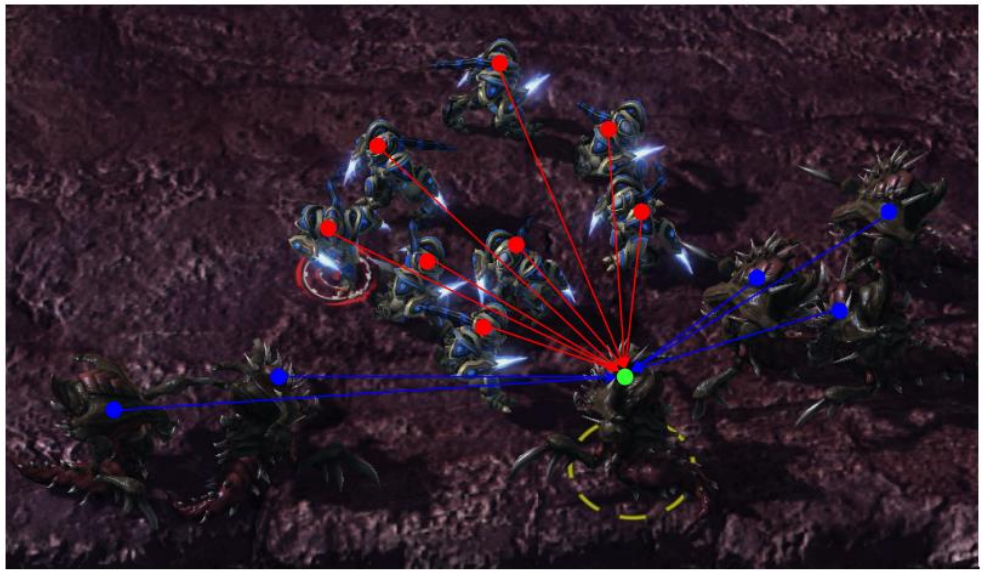
Delay Compensator

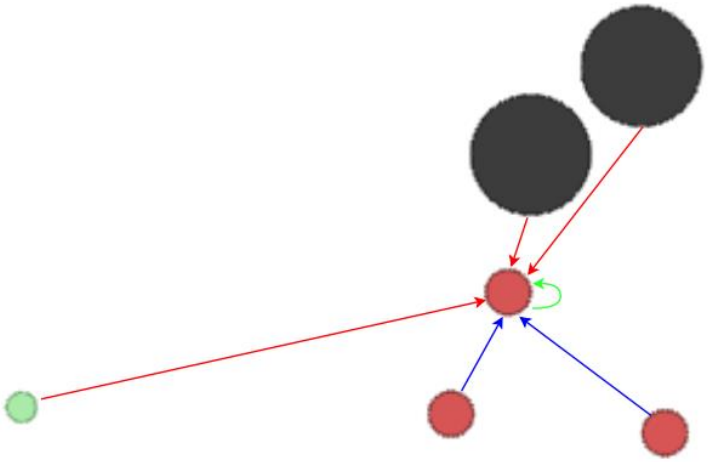
- **Two modes:**
  - *Flash*: Direct output, fast, suitable for small delay variations
  - *Echo*: Autoregressive model, step-by-step output, adapts to variable delays
- **Implementation:** GRU-based and Transformer-based
- **Input:** Historical observation sequence + action sequence + delay value vector



Experimental Setup

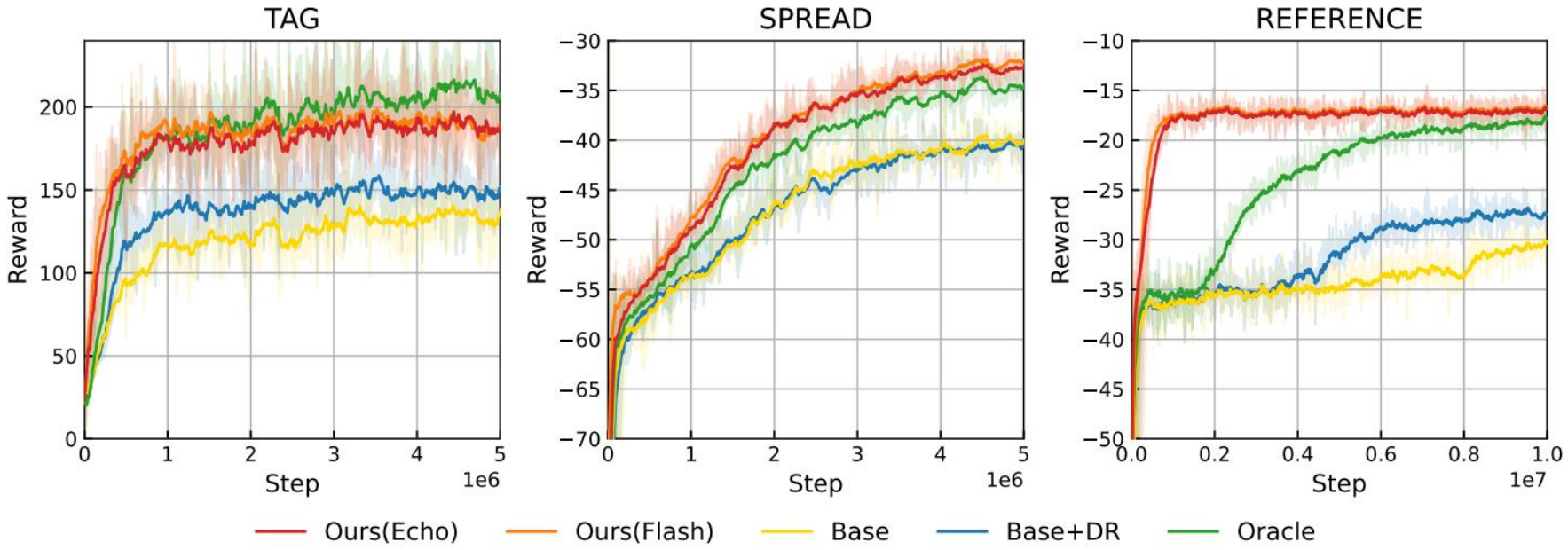
- **Environments:**
  - Multiagent Particle Environment (simple-tag, simple-spread, simple-reference)
  - StarCraft Multi-Agent Challenge (3s_vs_5z, 5m_vs_6m, 6h_vs_8z)
- **Baselines:** FT-QMIX, FT-VDN
- **Delay settings:** Fixed delays (0-12) and unfixed delays (uniform distribution)





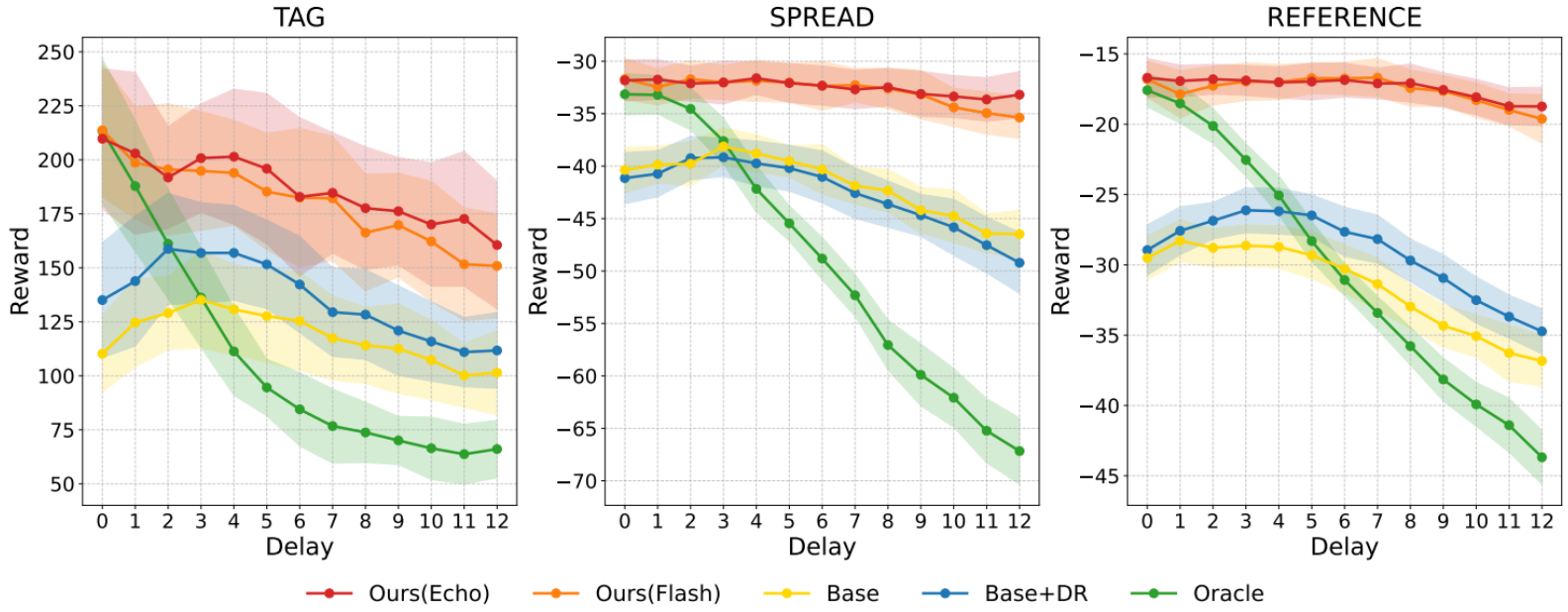
Main Results - Training Performance

- Baseline methods suffer severe performance degradation under delayed observation
- RDC-enhanced models converge faster
- Achieve or exceed delay-free Oracle performance
- Knowledge distillation accelerates training



Main Results - Fixed Delay Testing

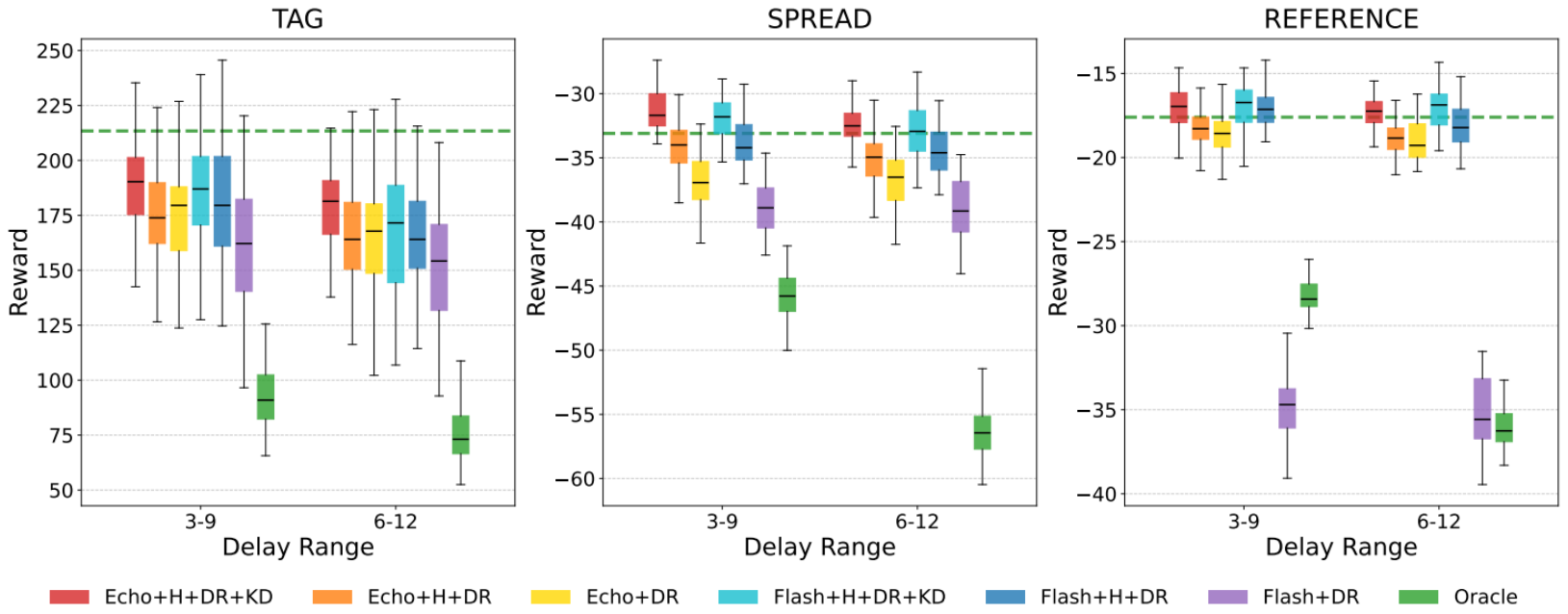
- Oracle performance degrades significantly as delay increases
- Baseline methods cannot generalize to unseen delays
- RDC-enhanced models maintain excellent performance across all delay settings
- Particularly robust on SPREAD and REFERENCE tasks



NEURAL INFORMATION PROCESSING SYSTEMS

Main Results - Unfixed Delay Testing

- **Testing:** In-distribution and half-out-of-distribution delays
- RDC-enhanced models show only marginal performance drop in out-of-distribution tests
- **Ablation study:** Each module contributes



Conclusion

- RDC framework effectively addresses delayed observation in MAS
- Compensator is the core component, directly reconstructing delay-free observations
- Curriculum learning and knowledge distillation provide additional support
- Future work: Improve compensator architecture and theoretical analysis