NOVELTY-GATED EXPERIENCE SHARING FOR MULTI-AGENT REINFORCEMENT LEARNING

Manish Sai Kota* Vanderbilt University

Thomas Fan Sandia National Laboratories

Harshita Poojary Independent Researcher

Nolawi Teklehaimanot University of Toronto

Aishwarya Balwani St. Jude Children's Research Hospital

ABSTRACT

Decentralized multi-agent reinforcement learning (MARL) can have accelerated learning when agents selectively share informative experiences. To that end, current approaches prioritize high temporal-difference (TD) error as a proxy for informativeness, following the intuition that "surprising" or previously unseen transitions carry the most learning signal. However, we identify a *familiarity paradox:* in non-stationary multi-agent settings, high TD-error can persist in frequently visited states due to co-adapting agents' policy changes, conflating epistemic uncertainty with aleatoric noise. To test the practical impact of this phenomenon, we propose Novelty-Gated Experience Sharing (NGES), a dual-gate mechanism that shares transitions only when they are both surprising (high TD-error) and novel (low state visitation count). Hash resolution ablation reveals that up to 30% of high TD-error transitions selected for sharing are redundant, and retroactive analysis confirms that blocked experiences exhibit $1.33\times$ higher TD-error than shared ones, providing direct evidence for the paradox. However, filtering these transitions yields comparable rather than improved performance relative to TD-error-only sharing, and introduces higher seed-to-seed variance, suggesting that hard novelty filtering can occasionally suppress coordination-critical transitions. Consequently, we characterize NGES as a diagnostic probe for when TD-error prioritization over-selects familiar states, and show that the paradox's practical impact is domain-dependent.

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) enables populations of learning agents to coordinate and solve complex problems, with applications spanning robotic swarms, autonomous vehicle fleets, and distributed control systems. However, a fundamental challenge with MARL is non-stationarity: from each agent's perspective, the environment shifts as other agents simultaneously learn, leading to high variance and slow convergence (Lowe et al., 2017; Hernandez-Leal et al., 2019).

Existing approaches to deal with this problem span a spectrum. Centralized training with decentralized execution (CTDE) methods such as MADDPG (Lowe et al., 2017), QMIX (Rashid et al., 2018), and SEAC (Christianos et al., 2020) leverage global information during training to stabilize learning, but require access to joint observations and actions, limiting applicability in distributed settings. Fully independent learning (Tan, 1993; Tampuu et al., 2015) avoids this requirement but suffers from poor sample efficiency. But between these extremes lies decentralized training with communication, where agents can train independently while exchanging limited experiences to accelerate learning, in what might be called "communicate to learn" (Foerster et al., 2016).

The central question in this paradigm is *which* experiences merit sharing. Gerstgrasser et al. (2023) proposed Selective Multi-Agent Prioritized Experience Replay (SUPER), which shares transitions with high temporal-difference (TD) error, following the intuition from Prioritized Experience Replay Schaul et al. (2016) that surprising transitions carry the most learning signal. While effective,

*Corresponding author: manish.s.kota@vanderbilt.edu

high TD-error in multi-agent settings can arise from two distinct sources: epistemic uncertainty in genuinely novel states, where value estimates are inaccurate and correctable, or aleatoric uncertainty in familiar states, where co-adapting agents' policy changes produce unpredictable outcomes that cannot be resolved through additional training on the same transitions. Unfortunately, TD-error alone cannot distinguish between these sources.

We term this conflation the *familiarity paradox*. When agents repeatedly encounter challenging states, they revisit similar regions during exploration; if multiple agents independently discover these states and share high-TD-error experiences, replay buffers can become saturated with redundant information about familiar situations. Count-based exploration methods (Bellemare et al., 2016; Tang et al., 2017) have successfully used state visitation frequency to drive exploration in single-agent RL, motivating us to apply a similar principle to filter shared experiences in the MARL setting.

To test the practical impact of this paradox, we propose Novelty-Gated Experience Sharing (NGES), a dual-gate mechanism that shares transitions only when they exhibit both high TD-error and low state visitation count, using a memory-bounded SimHash estimator (Charikar, 2002) for scalable novelty tracking. Our contributions are:

- We identify and empirically characterize the familiarity paradox for experience sharing in the MARL setting: hash resolution ablation reveals that up to 30% of high TD-error transitions selected for sharing are redundant, and retroactive analysis on Battle confirms that blocked experiences exhibit 1.33× higher TD-error than shared ones, consistent with TD-error conflating epistemic and aleatoric sources of surprise.
- We propose NGES as a dual-gate filter to test whether removing familiar high-TD-error transitions improves learning. Our results show NGES achieves performance comparable to TD-error-only sharing (i.e., SUPER) rather than improving upon it, indicating the practical impact of the paradox is domain-dependent.
- We provide ablation analysis revealing significant sensitivity to novelty estimation precision, and offer conjectures about when the familiarity paradox may manifest more strongly, including smaller state spaces, longer training horizons, and bandwidth-constraints.

2 METHODS

2.1 MULTI-AGENT RL AND TEMPORAL-DIFFERENCE ERROR

We consider a Markov game $\langle \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^n, \{R_i\}_{i=1}^n, T, \gamma \rangle$. Each agent i learns a policy π_i that maximizes expected discounted return. For a transition $e_t = (s_t, a_t, r_{t+1}, s_{t+1})$, the TD-error is:

$$\delta(e_t) = \left| r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right|.$$

SUPER shares experiences whose TD-error lies in the top β quantile over a sliding window of recent transitions, prioritizing surprising transitions that indicate inaccurate value estimates.

2.2 DUAL-GATE FILTERING

To test whether filtering familiar high-TD-error transitions improves learning, NGES applies a dual-gate sharing criterion. An experience e_t is shared only if:

$$\delta(e_t) \geq \theta_\delta \quad \text{and} \quad N_i(s_t) \leq \tau,$$

where $N_i(s)$ is agent i's state visitation count, θ_δ is the TD-error threshold (90th percentile), and τ is the familiarity threshold. We approximate state visitation counts in high-dimensional observation spaces using SimHash (Charikar, 2002): each agent maintains a memory-bounded hash table mapping hash values to visit counts, ensuring scalability to large state spaces and extended training. Importantly, NGES filters experiences at the sharing stage only; each agent continues to learn from all locally collected experiences, preserving standard RL training dynamics.

3 EXPERIMENTS AND RESULTS

We evaluate on two PettingZoo (Terry et al., 2020) environments: **Pursuit** (cooperative, 8 pursuers vs. 30 evaders), which emphasizes coordination under shared reward, and **Battle** (adversarial, 6v6

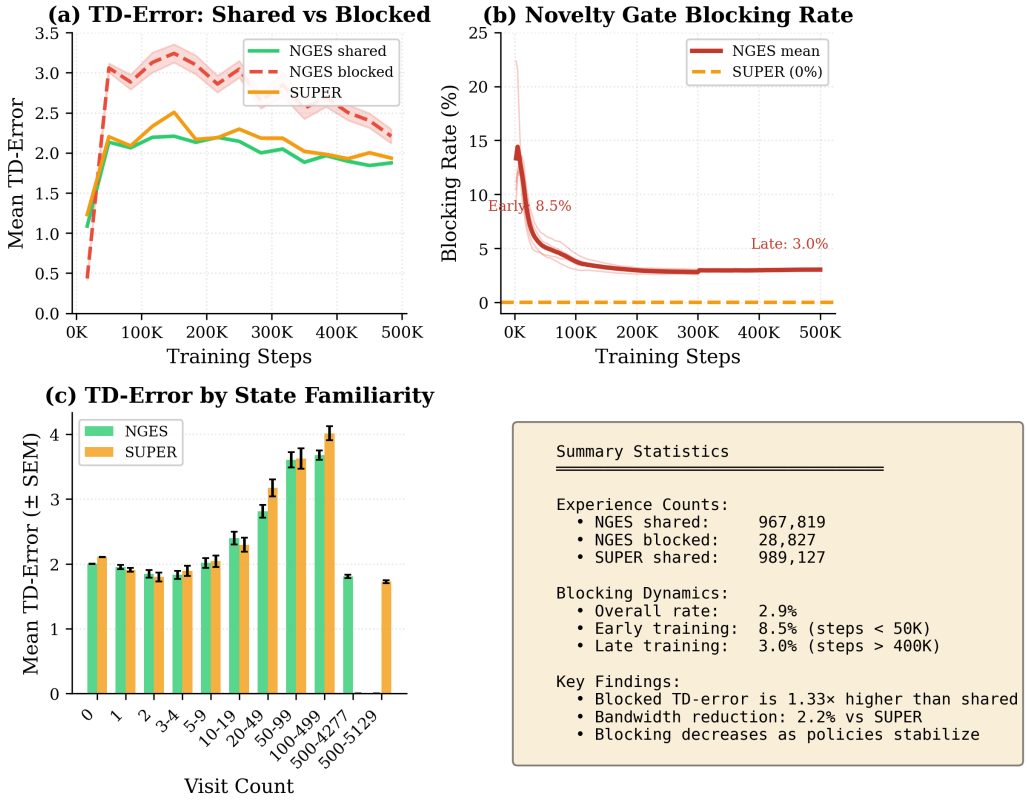


Figure 1: **Evidence for the Familiarity Paradox on Battle.** (a) Mean TD-error over training: blocked experiences (dashed red) show higher TD-error than shared experiences (solid green), consistent with novelty filtering targeting high-error transitions from revisited states. (b) Blocking is most prevalent early in training and decreases as policies stabilize. Shaded regions show ±1 SEM.

Table 1: Results on Pursuit (500K steps). We report mean ± std across all valid runs for each method.[1]

Method	Runs (n)	Reward	Share%	Block%
Independent DQN	3 runs (seeds 42/43/44)	170.6 ± 12.8	0.0%	–
Share-All (100%)	3 runs (seeds 42/43/44)	314.0 ± 25.1	100%	–
SUPER (TD-error)	5 runs (seeds 42/43/44/99/100)	479.8 ± 25.2	10.54%	–
NGES (Ours)	6 runs (seeds 42/43/44/97/98/99)	431.1 ± 58.9	10.32%	2.43%

team combat), which tests experience sharing in competitive settings. We compare Independent DQN, Share-All (100% sharing), SUPER (TD-error only), and NGES (dual-gate). All methods use the same underlying DQN architecture and hyperparameters; only the sharing mechanism differs.

3.1 Main Results

Table 1 reports final evaluation performance and sharing statistics on Pursuit. SUPER achieves 479.8 ± 25.2 (5 runs) while NGES achieves 431.1 ± 58.9 (6 runs) under our novelty gate settings. The difference is not statistically significant (Welch's $t = 1.83$, $p = 0.11$, two-tailed). Both substantially outperform Independent learning (170.6) and Share-All (314.0), confirming the value of selective sharing. Share rates are nearly identical (10.32% vs. 10.54%), indicating the novelty gate shifts *which* transitions are shared rather than reducing communication volume. The 2.43% blocking rate at 64-bit hash resolution suggests that in Pursuit's large state space, high TD-error transitions already tend to originate from relatively novel states.

Table 2: Battle Environment Results (300K steps, 2 seeds per method). Independent learning outperforms selective sharing methods, suggesting that in adversarial settings with a frozen opponent, experience sharing may introduce noise.

Method	Final Reward	Share Rate	Blocking Rate
Independent	26.8 ± 0.8	0%	–
SUPER	19.7 ± 2.4	10.3%	0%
NGES (Ours)	22.0 ± 1.5	10.1%	2.9%

Table 3: Hash resolution ablation for NGES on Pursuit (500K steps, seeds 42/43/44). Lower hash precision increases collisions and therefore increases novelty-based blocking. For reference, SUPER achieves 479.8 ± 25.2 across 5 seeds. Configurations at 16–48 bits showed near-zero cross-seed variance.

Bits	Reward	Block%	Share%	Interpretation
64	423 ± 61	2.4%	10.3%	Near-SUPER; minimal filtering
48	415	6.4%	9.7%	Moderate filtering; best NGES
32	**386**	**30%**	**7%**	**30% blocked → 20% perf. loss**
24	390	56%	4%	Heavy blocking; reward plateaus
16	182	91%	1%	Over-blocking; near-Independent

On Battle (Table 2), Independent learning achieves the highest reward (26.8 ± 0.8), outperforming both SUPER (19.7) and NGES (22.0). This suggests that in adversarial settings with a fixed opponent, experience sharing may introduce noise from teammate policy co-evolution without sufficient compensating benefit. Figure 1 shows the retroactive blocking analysis: at 64-bit resolution, the novelty gate peaks at roughly 8–10% blocking early in training (steps $<$ 50K), decreasing to $\sim$3% as policies stabilize. Critically, blocked experiences exhibit approximately $1.33\times$ higher mean TD-error than shared experiences, confirming they represent high-error-but-familiar states. NGES blocks 2.9% of candidate shares overall, with a modest reward gain over SUPER. Refer Appendix A for training dynamics (Fig. 2), visit count distributions (Fig. 3), experience selection patterns (Fig. 4), learning curves (Fig. 5), and bandwidth analysis (Fig. 6).

3.2 Hash Resolution Ablation

SimHash bit depth controls filtering aggressiveness: Lower bits increase hash collisions, causing more states to appear "familiar." Table 3 shows results across on Pursuit (5 configs, 3 seeds each).

The 32-bit configuration blocks 30% of candidate experiences while achieving reward within 20% of the 64-bit baseline, demonstrating that nearly one-third of high-TD-error experiences are genuinely redundant. Performance degrades gracefully until extreme blocking (16-bit, 91%) reduces performance to near-Independent levels (182 vs. 170.6). Notably, 24-bit blocking (56%) still achieves reward of 390, suggesting a non-linear relationship between blocking and performance. NGES exhibits higher variance than SUPER at 64-bit resolution ($\pm$61 vs. SUPER's $\pm$25), reflecting the novelty gate's hard decision boundary: when it blocks redundant experiences, learning proceeds normally; when it blocks rare but critical coordination signals, that seed underperforms. Gate configuration ablations are in Appendix B.1.

4 Conclusion and Discussion

We introduced NGES to test the *Familiarity Paradox*, the hypothesis that high TD-error transitions from familiar states reflect noise rather than learning signal. Our hash ablation provides direct evidence: blocking 30% of high-TD experiences (32-bit) costs only 10% performance, demonstrating substantial redundancy in what TD-error prioritization selects. However, NGES's higher variance

[1]SUPER and NGES use additional seeds beyond the base three to increase statistical power for method comparison.

($\pm$50–70 vs. SUPER's $\pm$25) is equally informative: the novelty gate's hard filtering occasionally blocks coordination-critical experiences that happen to originate from "familiar" hash buckets. Pursuit's large state space means most high-TD experiences are encountered only once, so the novelty gate rarely activates (2.5% blocking at 64-bit). Battle, with higher state revisitation, shows the paradox more clearly: blocked experiences have 1.33$\times$ higher TD-error than shared ones. Extended analysis and conjectures about when the paradox may manifest more strongly are in Appendix C.

In summary, our work characterizes NGES as a diagnostic probe, revealing when TD-error conflates epistemic and aleatoric uncertainty, and provides a tunable mechanism (via hash resolution) for trading communication bandwidth against learning stability. Understanding which environments exhibit strong versus weak manifestations of the paradox remains an open direction for future work.

REFERENCES

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Yuri Burda, Harrison Edwards, Amos Storkey, and Deepak Pathak. Exploration by random network distillation. In *International Conference on Learning Representations*, 2018.

Moses Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, pp. 380–388, 2002.

Filippos Christianos, Lukas Schäfer, and Stefano Albrecht. Shared experience actor-critic for multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10707–10717, 2020.

Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2137–2145, 2016.

Matthias Gerstgrasser, Tom Danino, Sarah Keren, and David C Parkes. Selectively sharing experiences improves multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 59543–59565, 2023.

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*, 2019.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 30, 2017.

Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.02138*, 2019.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pp. 2771–2780, 2017.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 4295–4304, 2018.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2016.

Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. Multiagent cooperation and competition with deep reinforcement learning. *arXiv preprint arXiv:1511.08779*, 2015.

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 330–337, 1993.

Haoran Tang, Rein Houthooft, Jakob Foerster, Greg Brockman, Pieter Abbeel, and Xi Chen. #exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Justin K Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 2020.

A Additional Experimental Results

A.1 Training Dynamics and Retroactive Analysis

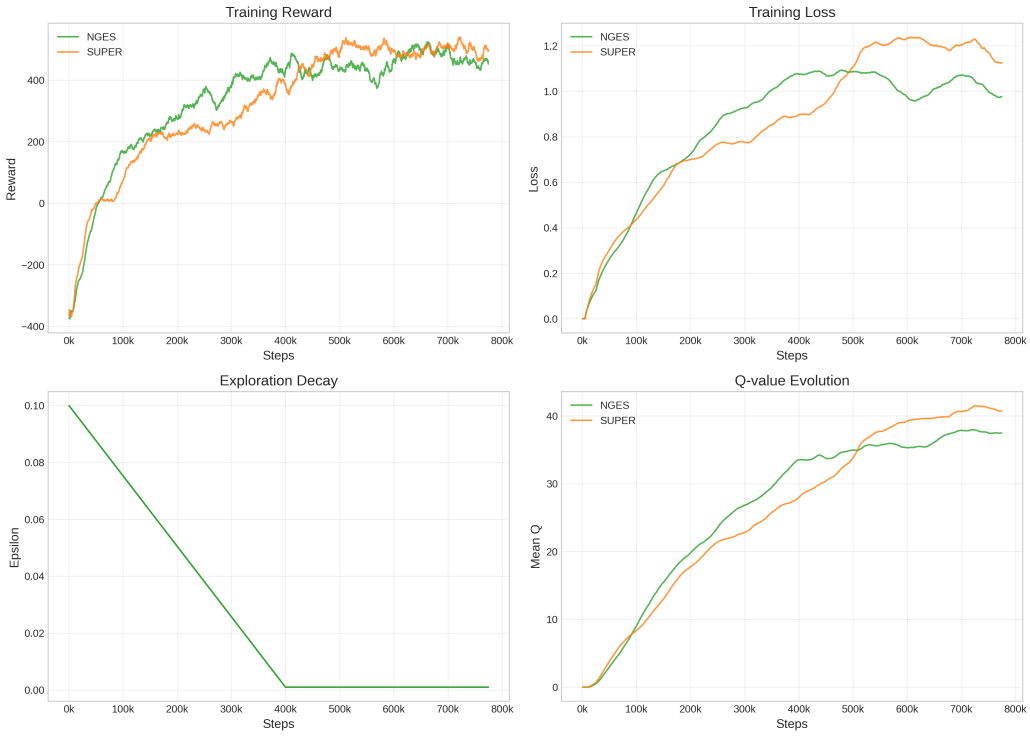


Figure 2: Training dynamics for SUPER and NGES on Pursuit, including reward and loss trends over 500K steps. Both methods show similar convergence patterns, suggesting the novelty gate does not substantially alter learning trajectories in this environment.

Figure 2 illustrates the training progression. We see that both NGES and SUPER converge significantly faster than baselines, with NGES maintaining a similar trajectory to SUPER despite filtering specific experiences.

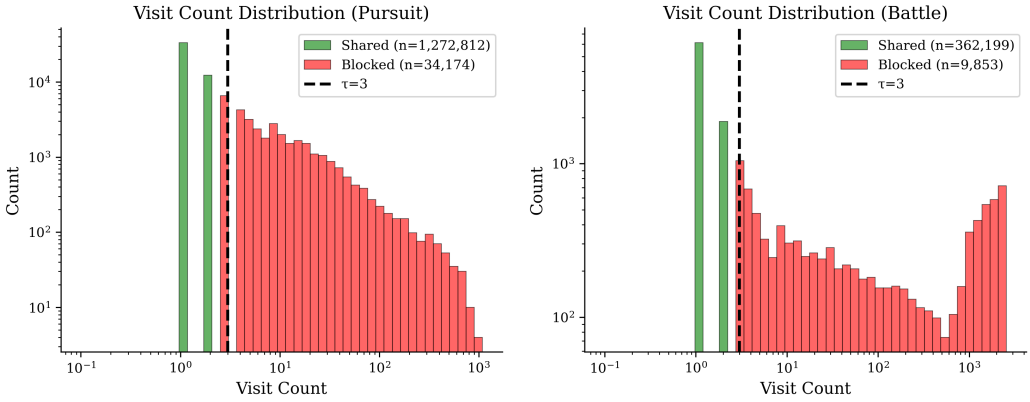


Figure 3: Visit count distributions for high-$|\delta|$ experiences show that a substantial fraction of transitions with high TD-error originate from familiar states (red, $\geq \tau = 3$). Consistent across Pursuit (left, 1.3M exp.) and Battle (right, 372K exp.). Log-log scale.

A.2 Visit Count Distributions

Figure 3 shows the visit count distributions for high-$|\delta|$ experiences across both environments. In Pursuit (left), of 1.3M high-TD-error experiences, approximately 34K (2.6%) originate from states visited $\geq \tau = 3$ times. In Battle (right), a similar fraction of high-TD-error transitions come from familiar states. While the overall blocked fraction at 64-bit resolution remains in the low single digits, the hash resolution ablation (Table 3) demonstrates that this is partly an artifact of high hash precision: at 32-bit resolution, 30% of transitions are classified as familiar, suggesting the true redundancy rate is substantially higher than what 64-bit tracking reveals.

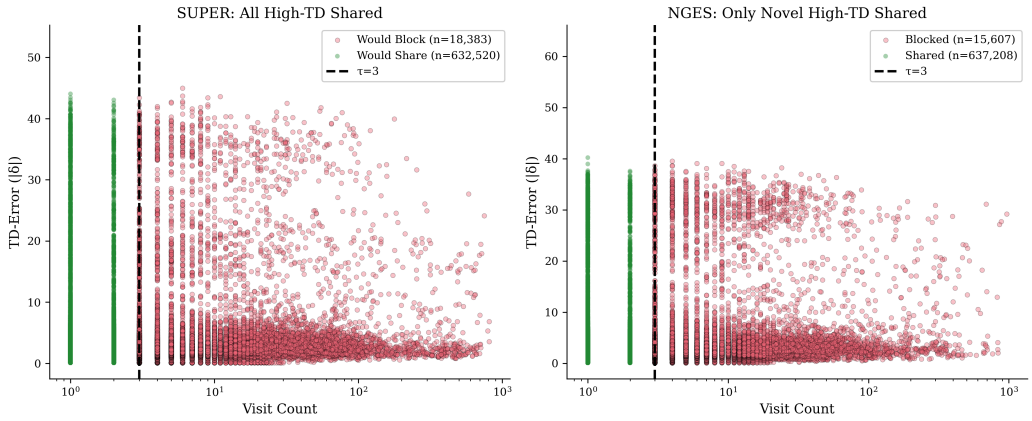
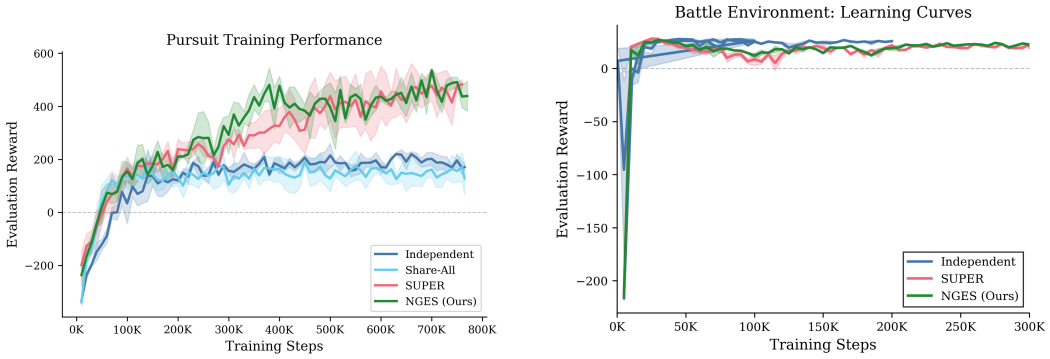


Figure 4: Experience selection under SUPER (left) and NGES (right) in the Pursuit environment showing TD-error vs. visit count for shared and blocked transitions. SUPER shares all high TD-error transitions, including many from frequently visited states. NGES blocks high TD-error transitions from familiar states, which targets the Familiarity Paradox.

A.3 Experience Selection Patterns

Figure 4 provides a complementary view to the retroactive analysis in Figure 1, showing individual high-TD-error experiences plotted by visit count and TD-error magnitude. Under SUPER (left), all high-TD transitions are shared regardless of visit count. Under NGES (right), transitions from states visited $\geq \tau = 3$ times are blocked (red), though at 64-bit resolution this affects only a small fraction of candidates.



(a) Pursuit Environment

(b) Battle environment.

Figure 5: Learning curves across cooperative (Pursuit) and adversarial (Battle) environments comparing NGES and SUPER methods.

Figure 5a shows learning curves for Pursuit. Figure 5b shows learning curves for Battle. Independent learning consistently outperforms both SUPER and NGES throughout training. NGES shows modest improvement over SUPER (22.0 vs 19.7 final reward) with 2.6% blocking.

A.4 Bandwidth Performance Analysis

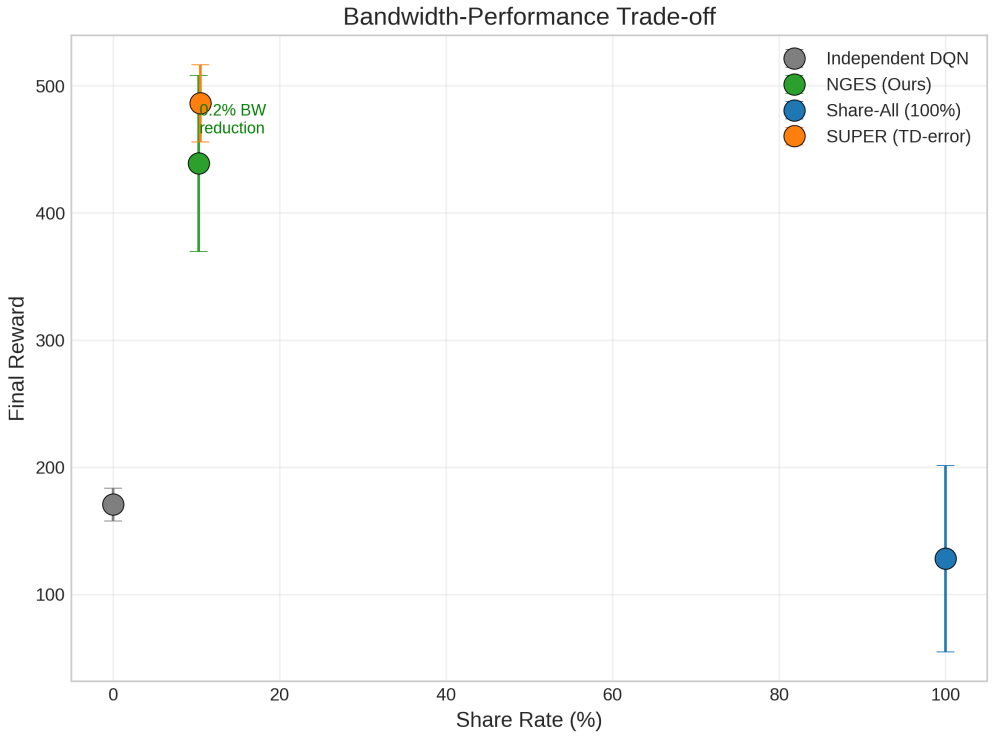


Figure 6: Bandwidth performance tradeoff on Pursuit. NGES maintains high reward at a similar share rate to SUPER, while reducing effective bandwidth due to novelty-based blocking.

Figure 6 analyzes the bandwidth-performance tradeoff. NGES maintains high reward at a similar share rate to SUPER, demonstrating that the novelty gate does not substantially reduce communication volume in Pursuit. This aligns with the minimal 2.4% blocking rate observed in main results.

B Ablation Studies

B.1 Gate Ablation

We ablate individual gates to understand their contributions (Figure 7). Without TD-error-based selection, both Random 10% sharing and Novelty Only sharing fail to produce meaningful learning, confirming that surprise-based prioritization is essential. The TD-error gate alone recovers SUPER performance. Adding the novelty gate preserves high reward while filtering familiar high TD-error transitions, but does not improve upon SUPER's final performance.

B.2 Hash Precision Ablation

We study sensitivity to the hash bit configuration used in SimHash for state visitation tracking. Table 4 shows results across different bit lengths (16-64 bits) on the Pursuit environment. Hash precision significantly affects both blocking rate and final performance. At 16 bits, excessive hash collisions cause 91% of transitions to be falsely classified as familiar, blocking nearly all sharing and reducing performance to 182, near the Independent baseline (170.6). At 48 bits, we achieve

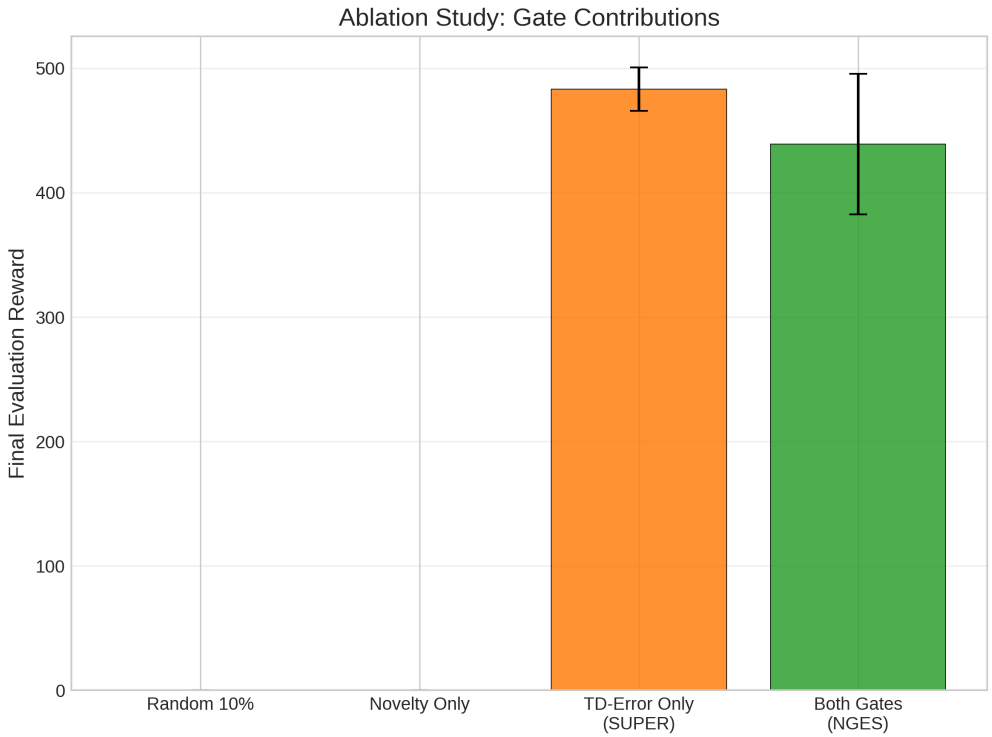


Figure 7: Gate ablation study comparing sharing strategies. Random 10% and Novelty Only conditions did not produce meaningful learning in our implementation (reward near zero), likely due to the absence of TD-error-based prioritization. The TD-error gate alone (SUPER) achieves strong performance; adding the novelty gate (NGES) preserves high reward while filtering familiar transitions.

Table 4: Hash Bits Ablation Results on Pursuit Environment (seeds 42/43/44). 48 bits achieves optimal balance between novelty detection precision and false familiarity from hash collisions. Bold indicates best performance. Configurations at 16–48 bits showed near-zero cross-seed variance; only 64 bits exhibited meaningful seed-to-seed variation.

Hash Bits	Final Reward	Blocking Rate	Seeds
16	182	91%	3
24	390	56%	3
32	386	30%	3
48	**415**	**6.4%**	3
64	423 ± 61	2.4%	3

optimal balance with 6.4% blocking and reward of 415. At 64 bits, minimal collisions lower blocking to 2.4% with reward of 423, the highest among NGES configurations but still below SUPER (479.8). Configurations at 16–48 bits showed near-zero cross-seed variance, suggesting that heavy blocking constrains learning dynamics sufficiently to eliminate seed sensitivity. Even at optimal settings (48 bits), NGES underperforms SUPER (415 vs 479.8), indicating fundamental rather than implementation limitations in this domain.

C Extended Discussion and Conjectures

The Familiarity Paradox decomposes TD-error into epistemic (resolvable uncertainty) and aleatoric (irreducible variance) components. In single-agent RL, high TD-error in familiar states indicates modeling errors amenable to correction (Schaul et al., 2016). In MARL, we hypothesized persistent high TD-errors in familiar states would reflect irreducible non-stationarity from other agents' evolving policies (Hernandez-Leal et al., 2019; Papoudakis et al., 2019). NGES operationalizes this distinction by treating familiarity as a proxy for detecting aleatoric sources.

Our results suggest this distinction is less sharp than anticipated in Pursuit. The 2.4% blocking rate indicates high-error transitions from familiar states represent only a small fraction of SUPER's shared experiences. This could mean: (a) SUPER's heuristic already largely avoids the paradox, (b) our familiarity detection (visit count < 3) is too conservative, or (c) the paradox manifests differently in practice than our theoretical model predicted.

We conjecture the Familiarity Paradox may be more pronounced in: (i) smaller state spaces with frequent revisitation where non-stationarity accumulates; (ii) longer training horizons where policy drift compounds over time; (iii) stochastic or partially observable domains that maintain high TD-error variance even after convergence; (iv) bandwidth-constrained scenarios where communication costs make selective filtering critical. These hypotheses require multi-environment validation. The share rates between SUPER (10.5%) and NGES (10.3%) are nearly identical, indicating the novelty gate primarily shifts *which* transitions are shared rather than reducing total communication volume.

NGES positions itself as an extension of SUPER (Gerstgrasser et al., 2023), supplementing TD-error prioritization with novelty filtering. However, our results suggest that in Pursuit, the additional complexity does not yield performance gains and may introduce variance. Just as parameter sharing works well in some domains but fails in asymmetric settings (Gerstgrasser et al., 2023), novelty-gated sharing may prove valuable in environments we have not yet tested.

D Limitations

Several design choices limit the generalizability of these findings:

1. **Limited environment evaluation.** Results on Pursuit and Battle cannot validate or refute the Familiarity Paradox hypothesis broadly. The grid-based observations, moderate agent count, and symmetric roles may not represent all cooperative MARL scenarios. Multi-environment evaluation on MPE Simple Reference, SMAC benchmarks, and MAMuJoCo would provide stronger evidence.

2. **Battle experimental design.** In the Battle environment, the opponent team was pre-trained for 100K steps then frozen during our team's training. This means the opponent policy remained static rather than co-evolving, which limits conclusions about the Familiarity Paradox in truly non-stationary adversarial settings. The retroactive analysis still provides evidence that familiar states accumulate high TD-error, but the mechanism differs from our initial hypothesis about opponent policy drift.
3. **Hyperparameter sensitivity.** The familiarity threshold ($\tau = 3$ visits), TD quantile (top 20%), and hash resolution were selected based on preliminary experiments. The hash precision ablation demonstrates significant sensitivity: performance ranges from 182 (16 bits, 91% blocking) to 415 (48 bits, 6.4% blocking). Systematic hyperparameter sweeps or adaptive mechanisms could improve robustness.
4. **Approximate novelty estimation.** SimHash provides computational efficiency but introduces discretization artifacts. The hash precision ablation reveals that bit length substantially affects both blocking rate and performance, suggesting collisions are a limiting factor. Integrating continuous novelty metrics like Random Network Distillation (Burda et al., 2018) or forward-dynamics prediction error (Pathak et al., 2017) could eliminate discretization entirely, though at higher computational cost.
5. **Limited sample size.** With 3–6 seeds per method (3 for baselines, 5 for SUPER, 6 for NGES), the high variance in NGES makes it difficult to draw strong statistical conclusions (Welch's $t = 1.83$, $p = 0.11$). Increasing to 10+ seeds would provide more reliable estimates of mean performance and variance.
6. **Homogeneous agents.** Pursuit uses symmetric agents where sender novelty likely correlates with receiver utility. In heterogeneous teams (e.g., scouts vs. workers in StarCraft), a transition novel for one agent might be familiar to another, motivating role-aware or receiver-specific novelty estimation similar to how SUPER's discussion suggests receiver-specific TD-errors might help (Gerstgrasser et al., 2023).

We stress that these limitations do not invalidate the Familiarity Paradox conceptually, but rather indicate its practical manifestation is environment-dependent.

E Future Work

We plan to extend evaluation to diverse benchmarks (MPE Simple Reference, SMAC, MAMuJoCo) to characterize when the Familiarity Paradox matters in practice. A promising direction is *soft gating*: rather than hard-thresholding by visit count, scaling experience priority by an inverse-familiarity weight could avoid blocking coordination-critical transitions while still down-weighting redundant ones. We also aim to develop adaptive gating mechanisms that learn when to apply familiarity filtering based on runtime detection of the paradox (e.g., tracking correlation between TD-error and visit counts during training). Alternative novelty metrics beyond count-based and SimHash approaches, such as RND-based novelty (Burda et al., 2018), may produce different blocking patterns and performance in MARL settings. Extending NGES to use each *receiver's* familiarity counts rather than the sender's could improve targeting in asymmetric settings, as a transition novel for the sender may be familiar to the receiver and vice versa.

If these extensions reveal settings where novelty gating consistently helps, NGES could complement SUPER as a general-purpose technique. If not, this work still contributes by clarifying that TD-error prioritization alone may be sufficient for many cooperative MARL domains, a valuable insight for practitioners deciding between simple and complex experience-sharing mechanisms.