Broadcast-Gain: A 2-Byte, Stop-Gradient Control Plane to Trim Long-Tail Latency in Cooperative MARL

Anonymous Authors

Abstract

Cooperative Multi-Agent Reinforcement Learning (MARL) over bursty, lossy links faces delayed/sparse rewards, high-variance gradients, and learned communication that assumes smooth channels. We introduce *Broadcast–Gain (BG)*, a fixed-rate, **2-byte**, stop-gradient neighbor broadcast that overlays a standard PPO+GAE policy with no changes to training or rewards. Each cycle, an agent sends one byte encoding a residual of local pressure-progress and one byte of coarse context (axis bit, distance bin). Receivers keep the freshest packets and form a confidence-weighted consensus that gates a simple phase scheduler; the overlay only nudges the MOVE logit via a tiny multiplier and a narrow, distance-decayed push near the gate. Bandwidth is $\sim 0.24\,\mathrm{kbit/s}$ per agent; compute is a few scalar ops per step.

We evaluate a single-junction grid with a c-step clearance lock across $N \in \{100, 120, 140\}$, per-tick packet drop probabilities $\{0.60, 0.65, 0.70\}$, and `cycle_len` $\in \{3, 5, 6\}$. For each cell we compare a *frozen* baseline to the same frozen policy with BG (constants fixed). BG trims tails where it matters: on the hardest cell (N=120, drop 0.70, 6-step cycle) p95 wait (95th-percentile steps-to-clear) drops by **4.97** steps and near-gate flow rises by **+392**/1k, with idle-red ≈ 0. Across 108 cells BG wins 78 (72%), with gains concentrated at larger N and longer cycles and graceful degradation as drops increase. Mechanism checks show reallocation into green (+16–20 pp) and higher near-gate flow, consistent with a consensus gate that stretches minimum green under weak information and flips knife-edge outcomes without thrash.

BG is neighbor-only, event-based, robust to drops, and drops in without touching the learner.

Keywords: multi-agent reinforcement learning; long-horizon control; bandwidth-efficient communication; stop-gradient coordination; neuromodulatory gain

1 Introduction

Cooperative MARL over long horizons breaks when bandwidth is scarce and delivery is bursty. Delayed, sparse rewards raise gradient variance; learned communication often assumes rich, differentiable channels [1–4]; and centralized critics or value factorization stabilize training only with wide access and heavier models—poor fits when agents get a few bits per tick and links drop packets [5–8]. Bandwidth-aware schedulers and information-efficiency methods adapt what/when to talk but add complexity and training burden [9, 10]. The gap is a tiny, robust control-plane signal that works under packet loss and stays compatible with standard policy learning.

We propose Broadcast–Gain (BG): a fixed-rate, two-byte, stop-gradient broadcast that supplies a small, confidence-weighted global cue and a targeted push near the junction. It is neighbor-only, requires no learned protocol or back-propagation through the channel, and overlays a standard PPO+GAE policy [11, 12]. In short, BG trades rich messages and attention for a minimal cue that gates phase by consensus and lengthens minimum green when information is weak.

Despite its size, BG moves the needle. On a hard evaluation cell (N=120, dropout 0.70, cycle_len=6), it reduces the 95^{th}-percentile wait by **4.97** steps and adds **+392** near-gate crossings per 1k steps, with idle-red ≈ 0, at ∼0.24 kbit/s per agent. Across settings, gains concentrate where tails are largest and degrade gracefully as loss increases. Our contributions are a two-byte stop-gradient broadcast primitive, a confidence-aware gate that tolerates loss, and evidence that such a minimal overlay reliably trims long-tail latency without changing the base learner.

2 Method: Broadcast–Gain

Broadcast–Gain is a stop-gradient overlay on a standard policy. It adds a fixed neighbor broadcast each cycle, fuses received hints into a single confidence-weighted cue, drives a phase scheduler for the junction, and applies a near-gate push that adjusts the move logit.

Setting. Two perpendicular corridors share one junction with a c-step clearance lock (Fig. 1). Agents act every step with local observations. Communication is neighbor-only and fixed-rate (once per cycle). The junction exposes a served axis $S \in \{+1, -1\}$ that may switch at cycle boundaries.

Once per cycle, each agent i broadcasts two bytes: (1) a one-byte residual z_i summarizing local progress/pressure (int8; optional μ-law), and (2) a one-byte meta tag (axis bit, distance bin). Messages are sent within a small Manhattan radius; receivers keep the freshest packet per sender under a short TTL. Unique senders are aggregated into per-axis estimates. A moving average of coverage/freshness yields an information weight $w_{\text{info}} \in [0, 1]$, and a simple consensus score rises when most senders favor the same axis. These combine into a gate $w_{\text{cons}} = \mathrm{gate}(\text{consensus},\ w_{\text{info}}) \in [0, 1]$, which increases with agreement and coverage and decays smoothly as packets are lost.

The scheduler maintains S. Each cycle it enforces a minimum green that stretches when w_{info} is low, then switches when an advantage built from the fused signals clears a confidence-scaled threshold or when a max-green limit hits. If a cycle was wasted-clear (lock held, no crossing), the next minimum green is shortened to damp oscillations. This procedure is local and carries no gradients. The overlay touches only the move logit:

$$y^{(i)}_{\text{move}} = g^{(i)}_{\text{mul}}\,\ell^{(i)}_{\text{move}} + g^{(i)}_{\text{add}}, \qquad p^{(i)}_{\text{move}} = \sigma\big(y^{(i)}_{\text{move}}\big),$$

with a tiny multiplicative term $g^{(i)}_{\text{mul}} \approx 1$ and a near-gate additive push

$$g^{(i)}_{\text{add}} = \mathrm{clip}\Big(\Lambda\,\mathrm{sgn}(s_i S)\,e^{-d_i/\tau}\,w_{\text{cons}} + \gamma_{\text{fair}}\,w_{\text{cons}}\,\phi_i,\ -A,\ A\Big).$$

Here $s_i \in \{+1, -1\}$ is the agent's axis, d_i its grid distance to the gate, and ϕ_i a green-only fairness term that grows with near-gate wait. Green receives a small positive push; red a soft brake, with a hard-stop band for $d_i \le d_{\text{stop}}$. A short open window just past the gate enables platooning. Bandwidth is fixed at two bytes per agent per cycle (e.g., $2\,\text{B} \times 15\,\text{Hz} \times 8 = 240\,\text{bps} \approx 0.24\,\text{kbit/s}$ with 60 Hz and C=4).

Small-perturbation guarantee. Let $|\delta| \le A$ denote the *total* clipped shift BG applies to the MOVE logit at a state (we fold the tiny multiplicative term into δ via its effect on the logit and clip). Then the overlay changes the policy only a little:

Theorem 1 (Tight drift for a single-logit push). *For any observation o, if π_{BG} is obtained from π by shifting only the* MOVE *logit by δ (others unchanged), then*

$$D_{\text{TV}}\big(\pi_{\text{BG}}(\cdot|o), \pi(\cdot|o)\big) = \big|\pi_{\text{BG}}(\text{MOVE}|o) - \pi(\text{MOVE}|o)\big| \ \le\ \tanh\Big(\frac{|\delta|}{4}\Big),$$

and

$$D_{\text{KL}}\big(\pi_{\text{BG}}(\cdot|o)\,\|\,\pi(\cdot|o)\big) \ \le\ \frac{\delta^2}{8}, \qquad D_{\text{KL}}\big(\pi(\cdot|o)\,\|\,\pi_{\text{BG}}(\cdot|o)\big) \ \le\ \frac{\delta^2}{8}.$$

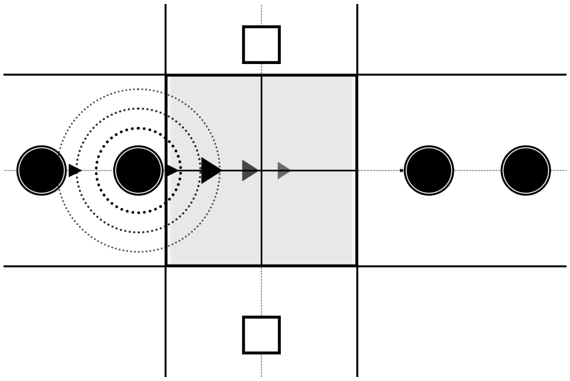

Figure 1: Experimental setup visualization.

Eval: N=120, dropout =0.70, cycle_len =6 (*frozen* → BG), 3 seeds			
Variant (eval)	SII	Tail p95 Δ↓ (steps)	Near-gate Δ↑ (/1k)
Frozen baseline (ref.)	0.000	0.00	0.0
BG (TD)	**0.450**	**-4.97**	**+391.9**
BG (RawEnt)	0.240	-2.35	+328.5

Mechanism (train): N=140, dropout =0.70, cycle_len =6 - BG (TD)			
Variant	Δshare_att_green (pp)	Δshare_real_green (pp)	Δnear-gate
BG (TD)	+16.17	+19.89	+160.6

Table 1: **Broadcast–Gain (BG) results.** *Right, top:* strongest eval cell (N=120, dropout =0.70, cycle_len =6), comparing a frozen baseline to the same policy with BG. *Right, bottom:* mechanism check on the train run cell (N=140, dropout =0.70, cycle_len =6).

3 Experiments

Setup and metrics. Single-junction grid with a c-step clearance lock. Factors: $N \in \{100, 120, 140\}$, per-tick packet dropout $\{0.60, 0.65, 0.70\}$, and `cycle_len` $\in \{3, 5, 6\}$. For each cell we run matched seeds and compare a frozen PPO+GAE policy to the same frozen policy with the BG overlay; BG constants are fixed across cells (no per-cell tuning). The primary metric is tail_wait_p95 (↓). Secondaries are near-gate realized crossings (per 1k steps, ↑), idle-red (↓), and (train-only) gate efficiency (↑) used for mechanism checks. For ranking only, we report a Signed Improvement Index (SII): a signed z-score combining $(-\text{p95}, +\text{near-gate})$ relative to the frozen baseline (SII>0 favors BG).

Results. *Reference stress cell* (N=120, dropout 0.70, `cycle_len`=6): BG(TD) reduces p95 by 4.97 steps and increases near-gate crossings by 391.9 per 1k steps, with idle-red ≈ 0 (SII $= 0.450$). The RawEnt variant yields 2.35 and +328.5, respectively. These shifts trim the tail without inducing red-time idling, consistent with a targeted near-gate push (Table. 1).

Across cells, BG wins 78/108 (72%). Gains concentrate at longer cycles and larger N; very short cycles (3) can be neutral or negative. Typical near-gate improvements are +249–+354 per 1k steps. We observe a small dip in direction-normalized gate efficiency (mean ≈ -0.02). Better tails correlate with this dip (Spearman $r_s \approx 0.53$; scatter in the appendix), consistent with reallocating green time where it matters.

Mechanism checks and ablations (train reference case: N=140, 0.70, 6): BG(TD) shifts attention-green by +16.17 pp, realized-green by +19.89 pp, and near-gate by +160.6 per 1k. Removing the near-gate push, removing confidence, or compressing to one byte each weakens or eliminates these gains (appendix).

4 Conclusion

Broadcast–Gain is a two-byte, stop-gradient control-plane overlay that reduces long-tail latency under bursty delivery at negligible cost (0.24 kbit/s per agent and a few scalar ops per step). In the most demanding evaluation case (N=120, dropout =0.70, `cycle_len`=6), BG lowers tail p95 by 4.97 steps and increases near-gate crossings by 392 per 1k steps while keeping idle-red near zero. Gains are strongest at longer cycles and larger N, and the method degrades gracefully as packet loss increases; very short cycles can be neutral or slightly negative.

Mechanistically, BG supplies a small, reliable global cue without learning through the channel: neighbors form a confidence-weighted consensus that gates phase decisions; a narrow near-gate push resolves knife-edge conflicts; and a light damping term reduces wasted clear. This recovers much of the effect of max-pressure with microscopic bandwidth and without altering the underlying PPO policy or rewards.

The approach aligns with event-driven, local-to-global coordination trends (e.g., robot-centric and graph-floor models with asynchronous updates). A pragmatic integration is to pair a learned short-horizon predictor with a 2-byte BG gate at execution time, keeping learning off the link while remaining robust to bursty loss [13].

Looking ahead, the most impactful extensions are: adaptive rate/quantization and TTL driven by uncertainty; forecast-to-gate fusion that modulates the consensus weight and near-gate strength; generalization to merges, splits, and multi-phase controllers; multi-hop sparse consensus for larger floors; and sim-to-real studies on MAPF-like layouts with congested Wi-Fi. Our view is that tiny, stop-gradient broadcasts are an underused lever in long-horizon MARL - practical to deploy, robust under adversity, and complementary to richer learned predictors rather than competing with them.

References

[1] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. URL `https://arxiv.org/abs/1605.07736`.

[2] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. URL `https://arxiv.org/abs/1605.06676`.

[3] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Michael Rabbat, and Joelle Pineau. Tarmac: Targeted multi-agent communication. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, 2019. URL `https://proceedings.mlr.press/v97/das19a/das19a.pdf`.

[4] Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. Learning when to communicate at scale in multiagent cooperative and competitive tasks. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://arxiv.org/abs/1812.09755`. IC3Net.

[5] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275*, 2017. URL `https://arxiv.org/abs/1706.02275`.

[6] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *AAAI Conference on Artificial Intelligence*, 2018. URL `https://arxiv.org/abs/1705.08926`.

[7] Peter Sunehag, Guy Lever, Audrūnas Gruslys, Wojciech M. Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017. URL `https://arxiv.org/abs/1706.05296`.

[8] Tabish Rashid, Mikayel Samvelyan, Christian Schröder de 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 (ICML)*, 2018. URL `https://arxiv.org/abs/1803.11485`.

[9] Daewoo Kim, Sangwoo Moon, David Hostallero, Wan Ju Kang, Taeyoung Lee, Kyunghwan Son, and Yung Yi. Learning to schedule communication in multi-agent reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://arxiv.org/abs/1902.01554`. SchedNet.

[10] Rui Wang, Xiaolong Guo, Chao Yu, Zhen Xiao, Changjie Fan, and Jun Wang. Learning efficient multi-agent communication: An information-theoretic perspective. *arXiv preprint arXiv:1911.06992*, 2019. URL `https://arxiv.org/abs/1911.06992`.

[11] John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015. URL `https://arxiv.org/abs/1506.02438`.

[12] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL `https://arxiv.org/abs/1707.06347`.

[13] Ameya Agaskar, Sriram Siva, William Pickering, Kyle O'Brien, Charles Kekeh, Ang Li, Brianna Gallo Sarker, Alicia Chua, Mayur Nemade, Charun Thattai, Jiaming Di, Isaac Iyengar, Ramya Dharoor, Dino Kirouani, Jimmy Erskine, Tamir Hegazy, Scott Niekum, Usman A. Khan, Federico Pecora, and Joseph W. Durham. Deepfleet: Multi-agent foundation models for mobile robots. *arXiv preprint arXiv:2508.08574*, 2025. URL `https://arxiv.org/abs/2508.08574`.

A Proofs for §2: Small-perturbation guarantees

Proof of Theorem 1. Let the action set have size $K \geq 2$ and let $\ell \in \mathbb{R}^K$ be baseline logits with $\pi = \mathrm{softmax}(\ell)$. Denote the MOVE action by m and set $p = \pi(m|o)$. BG shifts only the MOVE logit: $\ell'_m = \ell_m + \delta$, $\ell'_j = \ell_j$ for $j \neq m$, with $|\delta| \leq A$ (after absorbing the multiplicative term into δ and clipping as stated in the main text). Then

$$p' \triangleq \pi_{\mathrm{BG}}(m|o) = \frac{e^{\ell_m+\delta}}{e^{\ell_m+\delta} + \sum_{j\neq m} e^{\ell_j}} = \sigma\big(\underbrace{\ell_m - \log\sum_{j\neq m} e^{\ell_j}}_{\mathrm{logit}(p)} + \delta\big) = \sigma(\theta + \delta),$$

where $\theta = \mathrm{logit}(p)$ and $p = \sigma(\theta)$. For $j \neq m$, probabilities rescale by a common factor $\alpha = \frac{1-p'}{1-p}$, i.e., $\pi'(\,j\,|o) = \alpha\,\pi(j|o)$.

Total variation. Because all non-MOVE coordinates scale identically,

$$\|\pi' - \pi\|_1 = |p' - p| + \sum_{j\neq m} |\alpha\,\pi(j|o) - \pi(j|o)| = |p' - p| + |\alpha - 1|\,(1-p) = 2|p' - p|.$$

Hence $D_{\mathrm{TV}}(\pi', \pi) = \frac{1}{2}\|\pi' - \pi\|_1 = |p' - p|$. To bound $|p' - p|$, define $g(x) = \sigma(x+\delta) - \sigma(x)$. Then $g'(x) = \sigma'(x+\delta) - \sigma'(x)$ with $\sigma'(u) = \sigma(u)\big(1 - \sigma(u)\big)$; by symmetry of σ' about 0, $g'(x) = 0$ iff $x = -\delta/2$. Evaluating,

$$\max_x |g(x)| = \big|\sigma(\delta/2) - \sigma(-\delta/2)\big| = 2\sigma(\delta/2) - 1 = \tanh(\delta/4),$$

so $|p' - p| \leq \tanh(|\delta|/4)$. (Also $|p' - p| \leq \|\sigma'\|_\infty |\delta| = |\delta|/4$ for a linear small-shift bound.)

KL bounds. Because only one logit changes and the rest redistribute proportionally, both divergences reduce to the Bernoulli KL between $(p', 1-p')$ and $(p, 1-p)$. Let $A(\theta) = \log(1 + e^\theta)$ be the Bernoulli log-partition with $A''(\theta) = \sigma(\theta)(1 - \sigma(\theta)) \leq \frac{1}{4}$. By L-smoothness ($L = \frac{1}{4}$) and standard exponential-family identities,

$$D_{\mathrm{KL}}\big(\mathrm{Bern}(\sigma(\theta+\delta))\,\|\,\mathrm{Bern}(\sigma(\theta))\big) \leq \tfrac{L}{2}\delta^2 = \tfrac{\delta^2}{8},$$

and symmetrically $D_{\mathrm{KL}}\big(\mathrm{Bern}(\sigma(\theta))\,\|\,\mathrm{Bern}(\sigma(\theta+\delta))\big) \leq \frac{\delta^2}{8}$. □

Hard-stop safety. *(a.k.a. Lemma A)* If $s_i \neq S$ and $d_i \leq d_{\mathrm{stop}}$, the overlay clamps $g_{\mathrm{add}} = -A \leq 0$. Since softmax is monotone in each coordinate, decreasing the MOVE logit cannot increase its probability, i.e., $\pi_{\mathrm{BG}}(\mathrm{MOVE}|o) \leq \pi(\mathrm{MOVE}|o)$.

Conservative performance bound. *(Corollary A)* Let $J(\pi)$ be the γ-discounted return and $\epsilon = \max_s \big|\mathbb{E}_{a\sim\pi_{\mathrm{BG}}(\cdot|s)}[A_\pi(s,a)]\big|$. A standard TV-based performance difference bound yields

$$J(\pi_{\mathrm{BG}}) \;\geq\; J(\pi) + \mathbb{E}_{s\sim d_\pi}\Big[\sum_a \pi_{\mathrm{BG}}(a|s)\,A_\pi(s,a)\Big] - \frac{2\gamma}{(1-\gamma)^2}\,\epsilon\,D_{\mathrm{TV}}^{\max}(\pi_{\mathrm{BG}}, \pi),$$

and Theorem 1 gives worst-case regret $O\big(\tanh(A/4)\big)$.

Absorbing the multiplicative term. *(Lemma A)* If the effective perturbation on the MOVE logit is $y = g_{\mathrm{mul}}\ell + g_{\mathrm{add}}$ with $g_{\mathrm{mul}} \in [1-\varepsilon, 1+\varepsilon]$, $|g_{\mathrm{add}}| \leq A$, and logits clipped $|\ell| \leq L$, then $y = \ell + \delta$ with $\delta = (g_{\mathrm{mul}} - 1)\ell + g_{\mathrm{add}}$ and $|\delta| \leq A + \varepsilon L$. Thus Theorem 1 holds with $A \mapsto A + \varepsilon L$.

No-Zeno switching. *(Lemma A)* If `min_green_eff` $\geq m > 0$ at every cycle boundary, then over T steps with cycle length C, the number of flips is at most $\lceil T/(C\,m)\rceil$ (each flip forces at least m full cycles of hold).

Remark. BG is a small, stop-gradient perturbation: per state D_{TV} is at most $\tanh(A/4)$, KL drift is $O(A^2)$, hard-stop cannot increase red encroachment, and a positive minimum green rules out pathological flip rates.

B Protocol, Experiments, and Robustness

Cycle and neighborhood. A cycle groups C environment steps. Each agent transmits at most once per cycle to neighbors within Manhattan radius R; per sender, only the freshest packet is kept for up to T_{TTL} cycles.

Two bytes. Each agent i broadcasts $\texttt{pkt}_i = [\, z_i \mid m_i \,] \in \{-128, \ldots, 127\} \times \{0, \ldots, 255\}$. *Byte 0* ($z_i$): signed int8 residual via μ-law companding with μ=255. Default (**TD**): with $\delta_{\mathrm{TD}}^{(i)} = r + \gamma V(o') - V(o)$, normalize $x = \mathrm{clip}(\delta_{\mathrm{TD}}^{(i)}/s_\delta, -1, 1)$ and compand

$$q = \mathrm{sign}(x)\, \frac{\ln(1 + \mu|x|)}{\ln(1+\mu)}, \qquad z_i = \mathrm{clip}\big(\lfloor 127\, q \rceil, -127, 127\big).$$

Alternative (**RawEnt**): $x = 1 - H(\pi(\cdot|o_i))/H_{\max}$, then compand/quantize as above. *Byte 1* (m_i): packs axis and distance,

$$m_i = (\texttt{axis\_bit} \ll 7) \mid (\texttt{dist\_bin} \,\&\, 0x7F), \quad \texttt{axis\_bit} \in \{0,1\} \Leftrightarrow s_i \in \{-1,+1\},\ \texttt{dist\_bin} = \min(\lfloor d_i/\Delta_d \rfloor, 127).$$

Let $\mathcal{N}_a$ be the set of unique fresh senders supporting axis $a \in \{-1, +1\}$. Decompand z_j via $\hat{z}_j = \mathrm{sign}(z_j)\big((1+\mu)^{|z_j|/127} - 1\big)/\mu$. Weight freshness by $\eta_j = \exp(-\Delta t_j/\tau_{\mathrm{fresh}})$ and form axis scores

$$Z_a = \sum_{j \in \mathcal{N}_a} \eta_j\, \hat{z}_j, \qquad w_{\mathrm{info}} = c \cdot \mathrm{mean}_j(\eta_j), \quad c = \min\Big(\frac{\sum_a |\mathcal{N}_a|}{N_{\mathrm{ref}}}, 1\Big) \in [0,1].$$

Consensus:

$$\rho = \tanh\Big(\frac{Z_{+1} - Z_{-1}}{\kappa}\Big), \qquad w_{\mathrm{cons}} = \sigma(\alpha\,\rho)\, w_{\mathrm{info}}, \ \ \sigma(x) = \tfrac{1}{1+e^{-x}}.$$

Near-gate logit adjustment (per step). BG touches only the MOVE logit $\ell_{\mathrm{move}}^{(i)}$:

$$y_{\mathrm{move}}^{(i)} = \ell_{\mathrm{move}}^{(i)} + g_{\mathrm{add}}^{(i)}, \qquad p_{\mathrm{move}}^{(i)}{}' = \sigma\big(\mathrm{logit}(p_{\mathrm{move}}^{(i)}) + g_{\mathrm{add}}^{(i)}\big),$$

so the change is a shift by $g_{\mathrm{add}}^{(i)}$ in the *log-odds* of moving. With served axis $S \in \{-1, +1\}$, agent axis s_i, grid distance d_i, and fairness accumulator $\phi_i \in [0,1]$,

$$g_{\mathrm{add}}^{(i)} = \mathrm{clip}\Big(\Lambda\, \mathrm{sgn}(s_i S)\, e^{-d_i/\tau}\, w_{\mathrm{cons}} + \gamma_{\mathrm{fair}}\, w_{\mathrm{cons}}\, \phi_i,\ -A,\ A\Big),$$

$$\phi_i \leftarrow \begin{cases} \min(1, \phi_i + 1/K_{\mathrm{fair}}), & s_i{=}S,\ d_i \le d_{\mathrm{fair}}, \\ \max(0, \phi_i - 1/K_{\mathrm{fair}}), & \text{otherwise,} \end{cases} \qquad \text{reset } \phi_i \text{ on crossing.}$$

Safety: if $s_i \neq S$ and $d_i \le d_{\mathrm{stop}}$, force $g_{\mathrm{add}}^{(i)} = -A$ (hard-stop). A short "open window" (W_{open} steps) with $\Lambda_{\mathrm{open}} \le \Lambda$ can enable small platoons.

Scheduler (cycle boundary).

Algorithm 1: BG scheduler (compact)

state: $S \in \{\pm 1\}$, `green_age`, `min_green_0`
inputs: $Z_{\pm 1}$, w_{info}, last-cycle crossings x
`green_age` $\leftarrow$ `green_age` $+1$; `min_green_eff` $\leftarrow$ `min_green_0` $+\lambda_{\mathrm{stretch}}(1 - w_{\mathrm{info}})$
if `green_age` $<$ `min_green_eff` **then**
    return HOLD
end if
$\Delta Z \leftarrow Z_{-S} - Z_S$; thresh $\leftarrow \theta_0 + \theta_1(1 - w_{\mathrm{info}})$
if $\Delta Z >$ thresh **or** `green_age` $\ge$ `max_green` **then**
    $S \leftarrow -S$; `green_age` $\leftarrow 0$
    if $x{=}0$ **then**
        `min_green_0` $\leftarrow$ max(`min_green_min`, `min_green_0` $- \Delta_{\mathrm{wc}}$)
    else
        `min_green_0` $\leftarrow (1 - \beta_{\mathrm{mg}}) \cdot$ `min_green_0` $+ \beta_{\mathrm{mg}} \cdot$ `min_green_tgt`
    end if
else
    return HOLD
end if

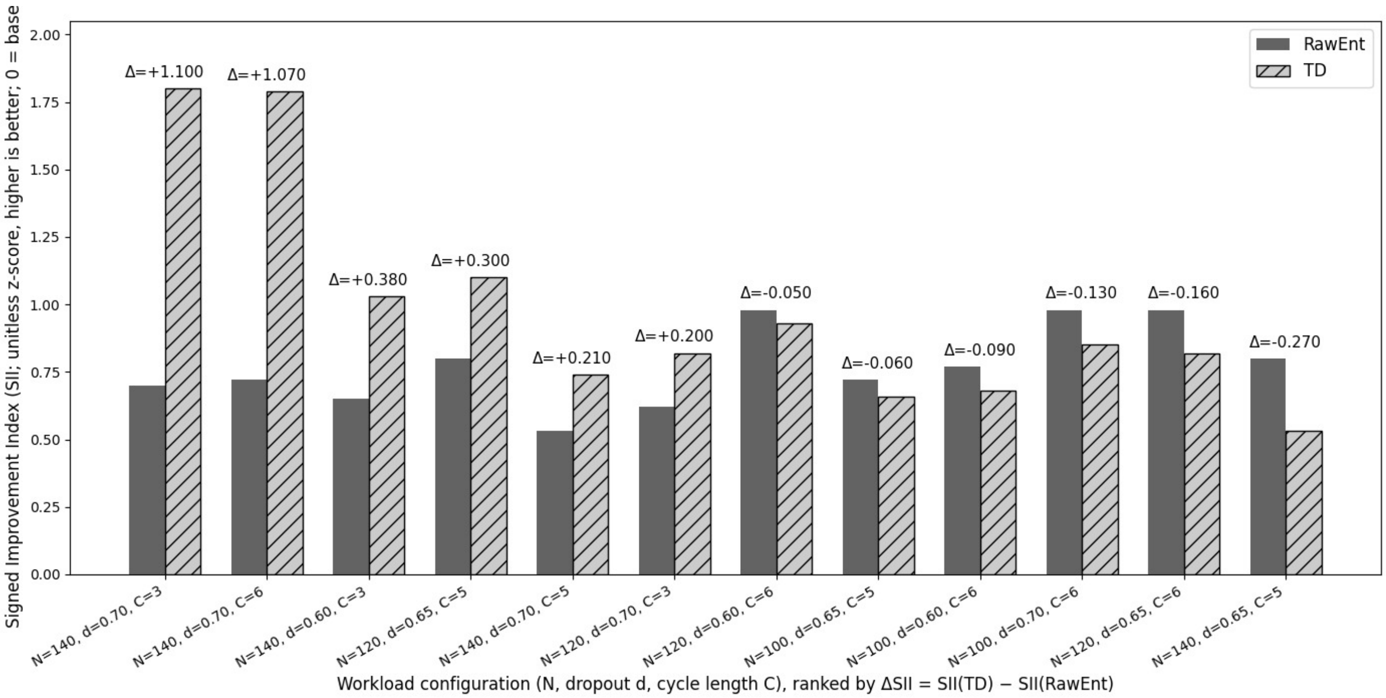

Figure 2: **TD vs. RawEnt.** Signed Improvement Index (higher is better) for BG with TD (hatched) vs. RawEnt (solid) on selected cells, ordered by $\Delta\mathrm{SII} = \mathrm{SII(TD)} - \mathrm{SII(RawEnt)}$. TD dominates on heavier loads/longer cycles; RawEnt is competitive on lighter cells.

Bitrate and compute. Per agent: $16\,\text{bits} \times f_{\text{step}}/C$ bps (e.g., $60\,\text{Hz}$ and $C{=}4 \Rightarrow \sim 240\,\text{bps}$). Per step: $O(|\mathcal{N}|)$ scalar ops (decompand, freshness, two sums) and one logit add; no extra networks.

Environment, training, evaluation. Single four-way junction with clearance lock $c{=}C$. Local observations; actions $\{\textsc{move},\textsc{wait}\}$. Directed links drop with Bernoulli rate $p \in \{0.60, 0.65, 0.70\}$. Train PPO+GAE *without* BG, then freeze. For each cell $(N, \text{dropout}, C)$, evaluate baseline vs. BG from identical RNG snapshots and seeds $\{13, 17, 23\}$. Training per seed: 2000 PPO updates (rollout 2048, 32 minibatches, 4 epochs), Adam lr $3{\times}10^{-4}$, $\gamma{=}0.99$, $\lambda{=}0.95$, clip 0.2, entropy/value coefs $0.01/0.5$, grad-norm clip 0.5. Evaluation: $4{\times}10^5$ env steps/seed/cell at $60\,\text{Hz}$. Observations and returns use running mean/var normalization.

Metrics and ranking. Primary: near-gate wait p95—for each crossing, count steps from first entry to $d \le d_{\text{fair}}$ until crossing; take pooled empirical 95^{th} percentile across matched seeds. Secondaries: near-gate crossings per 1k steps and idle-red (fraction of steps with near-gate demand held red). For compact comparison: Signed Improvement Index (SII) from standardized deltas: $\mathrm{SII} = \frac{1}{2}\big(-z(\Delta\mathrm{p95}) + z(\Delta\mathrm{NG})\big)$.

Sensitivity and robustness (brief). Gains concentrate at larger N and longer cycles; very short cycles ($C{=}3$) can be neutral/negative. Neighborhood and staleness matter: $R \in [2, 3]$ and $T_{\mathrm{TTL}} \in \{1, 2\}$ give stable wins—larger values add coverage but inject stale/conflicting packets that raise idle-red. Gate/push should be tuned jointly: prefer raising Λ (push) before clip A; large A can induce oscillations. For $C{=}3$, increase min-green stretch and slightly reduce switch-threshold slope to avoid premature flips under weak information. Under bursty loss (Gilbert–Elliott with mean burst length L_B and average loss $\bar{\epsilon}$; transitions $r{=}1/L_B$, $p = r(\bar{\epsilon} - \epsilon_G)/(\epsilon_B - \bar{\epsilon})$), freshness weighting and TTL degrade gracefully with L_B.

Sanity check vs. ε-max-pressure. A non-learning controller that flips when $\Delta Q{=}Q_{-S}{-}Q_S$ exceeds a threshold Θ (with tie-region randomization ε) recovers part of BG's benefit but idles red more; BG's consensus $\times$ confidence gate plus near-gate push reallocates green without idling.