FROM INFLUENCE TO UTILITY: REFERENCE-ALIGNED MESSAGE SCHEDULING FOR FROZEN MULTI-AGENT POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

A learned communication policy can generate more messages than a deployment link can carry. Selecting among them requires distinguishing a useful receiver response from a merely large one. We study this problem with the actor and communication stack frozen. CA-RAPU evaluates each candidate-induced policy shift against an action reference from other senders addressing the same receiver, then combines positive progress with influence, delivery reliability, confidence, and cost. The analysis characterizes the ambiguity of unsigned influence and gives sufficient conditions under which reference error preserves the selected set. Across 20 independently trained stacks in an assignment-controlled Spread task, CA-RAPU improves return over unsigned KL by 0.0277 [0.0153, 0.0401] at identical traffic. When each method calibrates its own transmission threshold, the net-return gain is 0.132 [0.112, 0.153]. Brackets give paired 95% confidence intervals. Redistributing peer support reverses the comparison with KL even when queue size stays fixed. Reference alignment thus helps allocate a frozen policy's messages when peers provide informative evidence about receiver actions.

1 INTRODUCTION

A learned communication policy may produce more messages than its deployment link can carry. Retraining the actor and communication models for each bandwidth limit, message price, or channel condition is costly. A shared inference service or outbound relay can instead allocate capacity among the messages already waiting in its queue. This leaves the learned policy fixed while adapting its communication to the available link.

The difficulty is recognizing a useful message before transmitting it. Comparing a receiver's action distributions with and without a candidate measures how strongly that candidate changes behavior (Lowe et al., 2019; Jaques et al., 2019). TEM and VIL2C use receiver KL in communication learning and wireless allocation (Guo et al., 2023; Zhang et al., 2026). A strong response, however, can move probability toward a helpful action or away from it. A fitted value function can distinguish these directions when suitable task supervision is available. We ask whether evidence already supplied by other senders can guide message selection without fitting a message-value predictor.

Receiver influence becomes a more useful scheduling signal when its direction is evaluated against task-relevant evidence from peers. CA-RAPU forms an action reference from messages addressed to the same receiver, excluding the focal sender. In cooperative navigation, several senders may provide noisy clues about that receiver's destination. Their aggregate supplies a direction against which to evaluate the policy shift caused by a candidate. Positive reference progress and receiver influence define its utility, while reliability, confidence, and cost determine transmission priority. The reference adapter and directional ranking rule have no trainable parameters. Our evaluation uses a shared confidence provider fitted offline and transmission thresholds calibrated on development data.

The mathematical argument connects this information to allocation. Unsigned responses can leave the preferred candidate unidentified even when a selector may abstain. Under a common continuation, an ideal reference gives the sign and ordering of local value shifts. We bound how reference error affects positive-progress normalization and the transmission score, then give sufficient condi-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tions for preserving the selected set. These results explain both the role of direction and the accuracy required at a capacity boundary.

The experiments separate message choice, transmission rate, and reference quality. At exactly 128 attempts per trajectory, CA-RAPU improves return over unsigned KL and a critic-supervised DeepSets ranker on independently trained Spread policies. Calibrating each method's rate also gives CA-RAPU higher net return. Reference interventions test the proposed explanation. Redistributing peer support can reverse the comparison with KL while leaving the queue size unchanged. A separate CommNet study examines the dependence on the communication interface.

2 Related Work

Influence and message value. Receiver interventions measure whether agents listen to one another (Lowe et al., 2019) and can provide intrinsic rewards for communication (Jaques et al., 2019). TEM uses absent/present-message KL to train communication, while VIL2C uses receiver response in latency-aware wireless allocation (Guo et al., 2023; Zhang et al., 2026). Task-value methods obtain a directional signal through supervision. I2C uses changes in joint action value, SMV assigns Shapley message credit, and CMVC learns counterfactual message values (Ding et al., 2020; Xue et al., 2022; Gao et al., 2025). SCoUT derives communication advantages for send and recipient decisions during training, then discards its grouping module and critics at execution (Vora et al., 2026). CA-RAPU constructs its ranking direction from decoded nonfocal peer evidence at scheduling time. The confidence model and development calibration are shared with the ranking controls, and scheduling uses frozen models without a reward or critic query.

Learning communication and allocating access. End-to-end approaches co-adapt messages and control (Foerster et al., 2016; Sukhbaatar et al., 2016; Jiang & Lu, 2018; Das et al., 2019). SchedNet learns medium access with sender encoders. IMAC and CAIC learn compression or sending frequency with the communication policy (Kim et al., 2019; Wang et al., 2020; Li et al., 2026). Other methods learn partners, ordering, topology, or temporal message integration (Sun et al., 2024; Ding et al., 2024; Zhang et al., 2020; Song et al., 2025; Li et al., 2025; Yang et al., 2025). Execution-time allocation can use beliefs, dynamics, or task-value estimates (Goldman & Zilberstein, 2003; Roth et al., 2005; Macke et al., 2021). DALA learns auction-based access for language agents (Fan et al., 2026). Our setting begins after the directed messages have been generated. This boundary permits a scheduler to be attached to an existing communication stack.

Allocation in collaborative perception. RDcomm jointly optimizes task-aware coding and selection, and COOPERTRIM adjusts feature sharing using temporal uncertainty (Liu et al., 2026; Mukhopadhyay et al., 2026). These methods allocate perceptual representations. CA-RAPU evaluates a generated message through its effect on a receiver's policy. Appendix C compares the information access and allocation units of the nearest methods.

3 Problem Setting

At time t, a joint policy controls n agents and has generated directed messages $e = (i, j, m_e) \in E_t$, with sender i, receiver j, and payload m_e. The scheduler selects at most B candidates from this queue. It cannot change the generator or task policy.

We consider a centralized egress queue co-located with the candidate generator, as in a shared inference service or outbound relay. Payloads and pre-transmission metadata are already available within this boundary. A read-only copy of the frozen receiver permits queries with and without a candidate, and a task adapter maps peer evidence to the receiver's finite action set $\mathcal{A}_j$. Queries leave the live communication state unchanged. Rewards, critics, future states, and realized channel outcomes are unavailable to CA-RAPU and the four ranking controls at scheduling time. A separate value-based comparator, Joint-DirectQ, receives critic access as described in the results.

Candidate e has attempt cost c_e, predicted dropout probability p_e^{drop}, and expected delay d_e. The constrained link begins at transmission from the queue, so cost is charged for each attempted delivery, including messages that are dropped. Collecting payloads into the queue is outside this cost model; a system that uses its scarce link for that collection would also need upstream allocation.

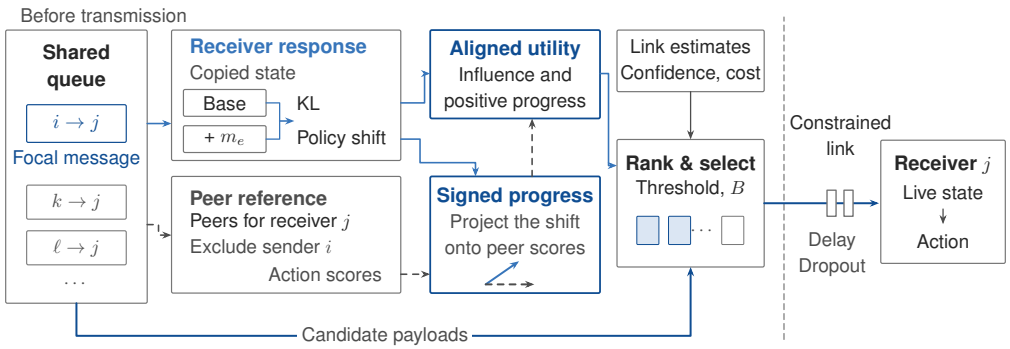

Figure 1: A candidate from sender i to receiver j is evaluated using copied receiver queries and evidence from other senders addressing j. The policy shift and peer action scores give signed progress, which combines with receiver KL to guide selection under capacity B. Solid arrows carry payloads or receiver-response computations; dashed arrows carry peer-derived computations. Only selected messages enter the constrained link; successful deliveries update the live receiver.

Let $\mathcal{I}_t$ contain the information available before channel realization. A scheduling policy σ selects $A_t = \sigma(\mathcal{I}_t) \subseteq E_t$, $|A_t| \le B$, to maximize

$$\max_{\sigma} \mathbb{E}_{\sigma}\left[R_{\text{task}} - \lambda_{\text{comm}} \sum_{t} \sum_{e \in A_t} c_e\right]. \tag{1}$$

CA-RAPU uses a local score to approximate a candidate's usefulness, then applies a deterministic capacity rule. Closed-loop rollouts evaluate Equation (1), including message interactions and the future states induced by the selected messages.

4 Reference-Aligned Receiver Influence

CA-RAPU combines two properties of a message: how strongly it changes the receiver and whether that change agrees with evidence from peers. Figure 1 shows the resulting scheduling path.

Measuring the receiver response. Let $o_{j,t}$ be receiver j's observation, $x_{j,t}$ its current communication state, and π_θ the frozen receiver policy. Candidate $e = (i, j, m_e)$ defines two queries from the same copied state,

$$\pi_j^0 = \pi_\theta(\cdot \mid o_{j,t}, x_{j,t}), \qquad \pi_e^+ = \pi_\theta(\cdot \mid o_{j,t}, x_{j,t} \oplus (m_e, 0)). \tag{2}$$

Here $\oplus(m_e, 0)$ appends the focal message at age zero, leaving other messages unchanged. Influence is $u_e^{\text{KL}} = D_{\text{KL}}(\pi_e^+ \| \pi_j^0)$. For numerical evaluation we use $\widetilde{u}_e = \sum_a \pi_e^+(a) \log[(\pi_e^+(a) + \epsilon_{\text{KL}})/(\pi_j^0(a) + \epsilon_{\text{KL}})]$, with $\epsilon_{\text{KL}} = 10^{-8}$. For an inventory with at least two candidates, let μ_t and v_t be the mean and population standard deviation of $\widetilde{u}_e$. When $v_t \ge \epsilon_{\text{KL}}$, set $\widehat{u}_e = \max\{0, (\widetilde{u}_e - \mu_t)/(v_t + \epsilon_{\text{KL}})\}$. Set $\widehat{u}_e = 0$ when the inventory has fewer than two candidates or $v_t < \epsilon_{\text{KL}}$. This normalization measures positive departures from the queue's mean response.

Building a reference from peers. Let $\mathcal{P}_{j,t}^{-e}$ denote current evidence addressed to receiver j after excluding the focal sender i. A fixed task adapter returns the bounded action score $h_{e,t} = H_{\mathcal{T}}(x_{j,t}, \mathcal{P}_{j,t}^{-e}; \pi_\theta) \in [-1, 1]^{|\mathcal{A}_j|}$. The adapter uses current receiver state, decoded peer messages, and fixed action semantics. With receiver state and peer evidence held fixed, changing m_e leaves the reference unchanged.

In Spread, $\overline{m}_{j,t}^{(-i)}$ aggregates clues addressed to j after removing every contribution from sender i. The adapter projects each action's movement vector onto the direction decoded from this aggregate. Candidate progress is then

$$g_e = \sum_a [\pi_e^+(a) - \pi_j^0(a)] h_{e,t}(a) = \mathbb{E}_{\pi_e^+}[h_{e,t}] - \mathbb{E}_{\pi_j^0}[h_{e,t}]. \tag{3}$$

Thus g_e is positive when the candidate moves the receiver's expected command toward the peer reference. Other tasks need an adapter that expresses their peer evidence as action scores. The reported adapters have no trainable parameters and are frozen before scheduler evaluation.

Ranking and deciding when to send. Let $P_t = \{e \in E_t : g_e > 0\}$. For nonempty P_t, define $\rho_t = (|P_t|^{-1} \sum_{e \in P_t} g_e^2)^{1/2}$ and set $\widehat{g}_e = \max(g_e, 0)/\rho_t$ when $\rho_t > \kappa$, where $\kappa = 10^{-10}$. Set $\widehat{g}_e = 0$ when P_t is empty or the scale is at most κ. An unavailable peer reference gives $g_e = 0$. The directional utility is

$$z_e = \sqrt{\widehat{u}_e \widehat{g}_e}. \tag{4}$$

The geometric mean gives high priority to messages that both change the receiver and move it toward the peer reference. Zero progress or zero influence gives zero utility. Appendix J compares this geometric score with an arithmetic alternative.

Let $\eta \geq 0$ and $\alpha \geq 0$ control delay and confidence sensitivity. Channel terms convert utility into a transmission score. Predicted delivery and delay give $r_e = (1 - p_e^{\mathrm{drop}}) \exp(-\eta d_e)$. An optional confidence head estimates bounded payload–sender confidence $\widehat{q}_e$ from the generated payload and available metadata. In Spread, it is fitted offline to latent clue precision and frozen with the communication stack. Comparisons share the same provider. Pooling the sender estimates through step t gives $\overline{q}_{i,t}$ and weight $w_e = r_e \overline{q}_{i,t}^{\alpha}$. CA-RAPU then selects

$$\begin{aligned} s_e &= w_e z_e - \lambda_{\mathrm{score}} c_e, \\ A_t &= \text{Top-}B\{e \in E_t : z_e > 0,\ s_e > \tau\}. \end{aligned} \tag{5}$$

The threshold τ lets the scheduler leave capacity unused. λ_{score} shapes candidate ordering, while λ_{comm} prices attempted communication in the evaluation objective. Without confidence or delay estimates, the neutral settings are $\overline{q} = 1$ and $d = 0$. Algorithm 1 gives the full decision path and deterministic tie rule.

5 From Reference Information to Allocation

The analysis distinguishes the information needed to value a response from the accuracy needed to preserve a scheduling decision. Fix one decision context and suppress t in the reference and its normalization.

Why unsigned influence can be insufficient. Consider actions $0, 1$, a base probability $\pi^0(1) = 1/2$, and candidates e_+, e_- inducing probabilities $1/2 + \varepsilon$ and $1/2 - \varepsilon$, for $\varepsilon \in (0, 1/2)$. Both have the same receiver KL and response norms. Let their metadata be identical, attempt cost zero, and capacity one. Two possible continuations exchange the action values, $Q_+(1) = Q_-(0) = 1$ and $Q_+(0) = Q_-(1) = 0$.

Proposition 1 (Ambiguity of unsigned response)**.** *Any randomized selector whose observable inputs have the same joint law in these two continuations has worst-continuation regret at least ε relative to the better candidate. This includes selectors using only fixed candidate labels, common metadata, and sign-invariant response magnitudes, with or without abstention.*

Proof. Let r_+, r_-, r_0 be its common selection probabilities for the two candidates and abstention. The regrets are $2\varepsilon r_- + \varepsilon r_0$ and $2\varepsilon r_+ + \varepsilon r_0$. Their average is $\varepsilon(r_+ + r_- + r_0) = \varepsilon$. □

The missing information is which action is valuable. Peer evidence can supply this information when it distinguishes the continuations. Its relevance to the task, rather than the presence of a reference alone, is what matters.

Relating reference progress to local value. For candidate e and receiver j, consider interventions differing only in the receiver's current action distribution, π_j^0 or π_e^+. Conditional on each action a, assume that other agents' actions and the subsequent trajectory have the same joint law under both interventions. Let this common continuation have integrable return and conditional action value $Q_e(a)$. Define $\Delta_e = \mathbb{E}_{\pi_e^+}[Q_e] - \mathbb{E}_{\pi_j^0}[Q_e]$. Choose a common scale $M_t > 0$ and offsets b_e so that $h_e^\star(a) = (Q_e(a) - b_e)/M_t \in [-1, 1]$. This ideal reference is an analytical quantity. CA-RAPU uses the peer reference h_e.

Proposition 2 (Value identity and reference error)**.** *Let* $g_e^\star = \langle \pi_e^+ - \pi_j^0, h_e^\star \rangle$. *If* $\|h_e - h_e^\star\|_\infty \le \delta_e$, *then*

$$\Delta_e = M_t g_e^\star, \qquad |\Delta_e - M_t g_e| \le M_t \delta_e \|\pi_e^+ - \pi_j^0\|_1. \tag{6}$$

Proof. Substitute $Q_e = M_t h_e^\star + b_e$. The offset cancels because the two action distributions each sum to one. Hölder's inequality then bounds the inner product of $\pi_e^+ - \pi_j^0$ with the reference error. □

The common positive scale makes ideal progress preserve the sign and ordering of local value shifts. For an approximate reference, the error also depends on how strongly the receiver responds. Under absolute continuity, Pinsker's inequality bounds the response norm by $\sqrt{2D_{\mathrm{KL}}(\pi_e^+ \| \pi_j^0)}$. The common continuation isolates an action intervention. Actual message persistence and subsequent communication are measured in closed-loop evaluation.

How informative peers improve the reference. Reference error reflects both uncertainty in peer observations and any shared bias. A bounded-peer model separates these effects. For candidate e, suppose $K_e \ge 1$ peer action-score vectors $Y_{e,k} \in [-1,1]^{|\mathcal{A}_j|}$ are conditionally independent given the decision context. Let $h_e = K_e^{-1} \sum_k Y_{e,k}$, and assume $\|\mathbb{E}[h_e \mid \text{decision context}] - h_e^\star\|_\infty \le \beta_e$.

Proposition 3 (Concentration of peer evidence)**.** *For* $\zeta \in (0,1)$, *with conditional probability at least* $1-\zeta$,

$$\|h_e - h_e^\star\|_\infty \le \beta_e + \sqrt{\frac{2\log(2|\mathcal{A}_j|/\zeta)}{K_e}}. \tag{7}$$

Replacing ζ *by* $\zeta/|E_t|$ *makes these bounds simultaneous over candidates, without requiring independence between them.*

Proof. Hoeffding's inequality bounds each coordinate's tail probability by $2\exp(-K_e r^2/2)$. A union bound over actions and the bias bound give the result. A second union bound covers the inventory. □

Additional independent peers reduce sampling error as $K_e^{-1/2}$, while shared bias persists. These are sufficient conditions on peer scores. The Spread adapter aggregates decoded clues before mapping them to actions, so its reference quality is assessed empirically rather than assumed to follow this averaging model. The fixed-inventory topology experiment isolates how the distribution of actual peer support changes selection.

Preserving the selected set. Hold receiver queries, weights, costs, threshold, and tie rule fixed, with $1 \le B < |E_t|$. Assume $\|h_e - h_e^\star\|_\infty \le \delta_e$ for every $e \in E_t$. Starred quantities apply the same computations with $h_e^\star$. Write $\varepsilon_e = \delta_e \|\pi_e^+ - \pi_j^0\|_1$ and $P^\star = \{e : g_e^\star > 0\}$. Suppose

$$g_e^\star > \varepsilon_e \quad (e \in P^\star), \qquad g_e^\star + \varepsilon_e \le 0 \quad (e \notin P^\star). \tag{8}$$

These margins preserve positive support. For nonempty $P^\star$, let $\rho^\star$ be its ideal RMS scale and set $\bar{\varepsilon} = (|P^\star|^{-1} \sum_{e \in P^\star} \varepsilon_e^2)^{1/2}$. Assume $\rho^\star - \bar{\varepsilon} > \kappa$. For $e \in P^\star$ define

$$\Gamma_e = \frac{\varepsilon_e + \widehat{g}_e^\star \bar{\varepsilon}}{\rho^\star - \bar{\varepsilon}}, \qquad \xi_e = |w_e| \sqrt{\widehat{u}_e \Gamma_e}, \qquad \xi = \max_e \xi_e, \tag{9}$$

with $\Gamma_e = \xi_e = 0$ outside $P^\star$.

Proposition 4 (Allocation stability)**.** *Under these conditions,* $|s_e - s_e^\star| \le \xi_e$. *If* $|s_e^\star - \tau| > \xi$ *for every* e *with* $z_e^\star > 0$, *the ideal and deployed adaptive eligible sets coincide. The selected sets coincide when this set has at most* B *candidates, or when its ordered ideal scores satisfy*

$$s_{(B)}^\star - s_{(B+1)}^\star > 2\xi. \tag{10}$$

For fixed traffic, the same gap over the full inventory is sufficient.

Proof. The reverse triangle inequality gives $|\rho - \rho^\star| \le \bar{\varepsilon}$. Subtracting the normalized ratios yields $|\widehat{g}_e - \widehat{g}_e^\star| \le \Gamma_e$. The inequality $|\sqrt{a} - \sqrt{b}| \le \sqrt{|a-b|}$ gives the score bound. Common positive support and fixed $\widehat{u}_e$ preserve positive utility, and the threshold margin preserves eligibility. Each selected–unselected score difference changes by at most 2ξ, so the capacity gap preserves membership. □

Under the same support margins, an empty $P^\star$ makes both adaptive rules abstain. For nonempty $P^\star$, the same holds when $\rho^\star + \bar{\varepsilon} \le \kappa$. In either case, the fixed-traffic rules use identical costs and tie keys. Appendix D gives tie-sensitive conditions. Together the results connect peer information to allocation through the response size, the positive-progress scale, and the distance from the selection boundary.

6 Experimental Setup

We study an assignment-controlled version of MPE2 Simple Spread (Lowe et al., 2017; Terry et al., 2021). It retains the cooperative movement and collision dynamics, while fixing an agent–landmark assignment at reset and generating noisy receiver-directed landmark clues. Each communication stack contains a movement actor, sender encoder, receiver decoder, candidate generator, and confidence head. We train twenty stacks for three-agent closed-loop evaluation and a separate twenty-stack four-agent cohort for reference and short-horizon diagnostics. The corresponding queues contain six and twelve directed messages per step. Architecture, training, and clue generation are specified in Appendix E.

Message choice and communication rate. To isolate ranking, the fixed-traffic experiment disables the positive-utility filter and threshold in Equation (5). It sends the two highest-scoring candidates at every one of 64 steps, using the common tie rule and continuing the prescribed attempts in an absorbing state after early success. Each method therefore makes exactly 128 unit-cost attempts. Comparators are progress magnitude, unsigned receiver KL, action-distribution L2, and a critic-supervised DeepSets ranker (Zaheer et al., 2017). They share the frozen stack, pre-transmission information, resets, and channel potentials; their trajectories evolve separately after their selections differ. In the adaptive studies, episodes terminate naturally and each method selects its own threshold on five disjoint development stacks. Evaluation averages the nine dropout–delay conditions equally within each trained stack.

Outcomes and statistical units. Let $\bar{d}_T$ be the final mean distance under the initial minimum-distance assignment. Success requires every assigned distance to be at most 0.12. We report

$$R_{\text{task}} = \mathbf{1}_{\text{success}} + 3[1 - \min\{\bar{d}_T/\sqrt{2}, 1\}], \qquad R_{\text{net}} = R_{\text{task}} - \lambda_{\text{comm}} C_{\text{attempted}}.$$

Here $C_{\text{attempted}} = \sum_t \sum_{e \in A_t} c_e$. This assignment-based task return lies in $[0, 4]$ and differs from native MPE reward. Dropped messages still incur attempt cost. A trained stack is one inferential unit. We average episodes within channel condition and then the nine conditions within stack. Between-method differences use paired 95% Student-t intervals. The p-values are Holm-adjusted within each specified comparison family. Fixed-state diagnostics separately compare peer progress with four-step conditional action values on states sampled independently of the selectors. Appendix A maps the studies and their outcomes.

7 Results

7.1 Better Message Choice at the Same Traffic

At 128 attempts per trajectory, CA-RAPU obtains mean task return 2.7029. Table 1 reports all four paired comparisons. The mean return gain over unsigned receiver KL is 0.0277 [0.0153, 0.0401]. Seventeen of twenty stacks favor CA-RAPU, and the mean gain is positive in every channel condition. On those trajectories, success increases from 13.16% to 15.29%, an absolute gain of 2.13 percentage points [1.18, 3.08]. Success measures completed assignments, while task return also accounts for final distance.

Table 1: **Fixed-traffic return gains.** CA-RAPU minus each control over twenty independently trained Spread-3 stacks. Every arm makes 128 attempts; p-values are Holm-adjusted across the four comparisons.

Control	Return difference [95% CI]	Positive stacks	Holm p
Progress magnitude	.0186 [.0029, .0344]	16/20	.0454
Unsigned receiver KL	.0277 [.0153, .0401]	17/20	6.59×10^{-4}
Action-distribution L2	.0137 [−.0045, .0319]	14/20	.132
Critic-supervised DeepSets	.0287 [.0116, .0458]	16/20	.0070

The progress-magnitude control removes the sign from the directional signal; CA-RAPU improves on it after multiple-comparison correction. The DeepSets ranker sees the complete candidate set, including payloads, link metadata, and the same base and counterfactual receiver distributions. It is trained on critic-derived targets and uses no critic during scheduling. An independent episode sample from the same twenty trained stacks gives a CA-RAPU margin of 0.0339 [0.0201, 0.0476]. The L2 comparison is inconclusive because its interval includes zero, and three channel cells have negative mean margins. Appendix H reports the full protocol, channel breakdown, and all secondary outcomes.

A fitted pair-value comparator, Joint-DirectQ, can additionally query a centralized critic and enumerate all message pairs. CA-RAPU's return difference from this comparator is 0.0162 [-0.0037, 0.0361]. The one-sided test meets the prespecified non-inferiority margin of -0.03, while the two-sided comparison remains inconclusive (Appendix H.5).

7.2 Higher Net Return When Communication Is Priced

At communication price $\lambda_{\text{comm}} = 0.005$, a secondary study lets CA-RAPU, unsigned KL, and progress magnitude each select a threshold on development stacks. CA-RAPU obtains net return 2.434, exceeding KL by 0.132 [0.112, 0.153] and magnitude by 0.139 [0.118, 0.160], with positive paired differences on all twenty evaluation stacks. It sends 0.813 messages per decision, compared with 0.631 for KL and 0.651 for magnitude. CA-RAPU earns enough additional task return to pay for its higher transmission rate.

A separate abstention study shows how much communication the same scheduler can avoid relative to a fixed rate. Adaptive CA-RAPU obtains task return 2.6160 and net return 2.4587 [2.4311, 2.4863], using 31.45 attempts per episode versus 116.73 for fixed-$B = 2$ CA-RAPU. Relative to silence, this preserves 80% of the fixed-rate method's task-return gain with 26.9% of its attempts. The paired net-return gains are 0.3228 [0.2960, 0.3496] over fixed-rate CA-RAPU, 0.2722 [0.2402, 0.3042] over a calibrated metadata-only scheduler, and 0.2569 [0.2250, 0.2888] over silence. All twenty stacks favor CA-RAPU in each comparison, with Holm-adjusted $p \leq 6.90 \times 10^{-13}$.

The benefit over KL also appears across communication prices. In a separate five-price study, CA-RAPU has the highest mean net return at every tested price. The KL contrasts at the four positive prices survive correction across all ten comparisons; the zero-price contrast is inconclusive. The L2 contrasts survive correction at prices 0.0025 and 0.01. Appendix I gives the thresholds, rates, and paired results for each study. The fixed-traffic and calibrated-rate experiments answer complementary questions. Direction improves which messages are sent, and thresholding determines how much of that improvement is worth its cost.

7.3 Where the Directional Advantage Comes From

Peer progress carries information about value. On the four-agent fixed-state diagnostic, the sign of peer progress agrees with the sign of four-step value shifts with balanced accuracy 0.742 [0.708, 0.777] and Matthews correlation 0.501 [0.438, 0.564]. Within-state pairwise rank agreement is 0.751 [0.735, 0.767]. Positive shifts account for 86.0% of candidates; recall is 93.7% for positive shifts and 54.8% for negative ones. The peer signal is therefore informative but less reliable at recognizing harm. The pair selected by CA-RAPU has higher mean candidate-local value than the KL-selected pair on 19 of 20 stacks (Appendix F).

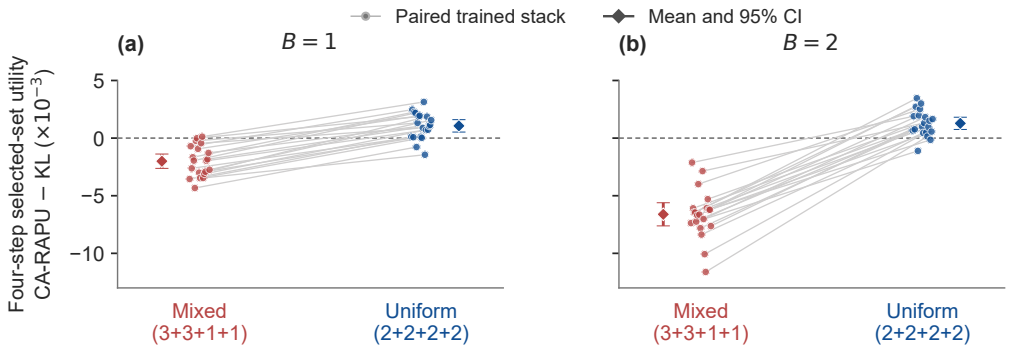

Figure 2: **Redistributing peer support reverses the comparison with KL.** Capacities are (a) $B = 1$ and (b) $B = 2$. Both support patterns contain eight messages; receiver indegrees are $3+3+1+1$ for mixed support and $2+2+2+2$ for uniform support. Small points show CA-RAPU-minus-KL four-step selected-set utility for each of twenty independently trained stacks; lines pair the same stack across patterns. Diamonds and bars show means and 95% Student-t intervals. Positive values favor CA-RAPU. Evaluations branch from fixed states.

The sign and the source both matter. A magnitude-matched intervention changes only the treatment of negative progress. Giving negative progress zero utility improves selected-pair one-step task-aligned gain by 0.01096 [0.01002, 0.01191]. In a separate comparison, replacing the receiver's own preferred-action reference with peer evidence raises fixed-traffic return by 0.0995 [0.0773, 0.1217], with positive differences on all twenty stacks. Peers thus supply useful information beyond the receiver's existing preference. The CA-RAPU-minus-self-inclusive return difference is -0.0062 [-0.0173, 0.0048], so the additional effect of excluding the focal sender remains inconclusive (Appendix G).

How peer support is distributed can change the outcome. In a four-step branch evaluation, we hold the four-agent queue at eight messages and redistribute receiver-matched support. A mixed pattern leaves each candidate with either zero or two peers after excluding its sender; a uniform pattern supplies one peer to every candidate. As Figure 2 shows, CA-RAPU trails KL under mixed support and leads it under uniform support at both tested capacities. Averaged over $B = 1, 2$, the uniform-minus-mixed change in the CA-RAPU–KL margin is 0.00548 [0.00490, 0.00606], with positive differences on all twenty stacks (Appendix K). The same queue size can therefore support different allocation outcomes, depending on which receivers have informative peers.

Reference perturbations give a related result. Moderate noise, missing peers, and peer conflict preserve a positive one-step selection advantage. Reversing one of two peers by $180°$ leaves no clear advantage (Appendix G.1). Both experiments connect the observed gains to the quality of the peer direction, consistent with the role of reference accuracy in the analysis.

8 Discussion

The results identify a way to adapt learned communication to a constrained deployment link. A frozen receiver exposes how messages change its policy, and informative peers can supply the direction needed to rank those changes. In Spread, this improves selection over unsigned KL at equal traffic and supports a higher net-return operating point after rate calibration. The reference avoids fitting a message-value predictor for each stack, but needs a meaningful map from peer evidence to receiver actions.

Transfer depends on what the peers know. Spread supplies redundant clues about assigned movement targets. A different communication interface may offer weaker evidence. In a six-policy CommNet Predator–Prey study, the CA-RAPU-minus-KL native-return difference at capacity $B = 5$ is -0.00855 [-0.01831, 0.00122], with one positive policy. The adaptive net-return difference is 0.02001 [0.00280, 0.03721], with all six differences positive. These are exploratory paired

95% Student-t intervals. The prespecified fixed-capacity qualification criterion fails, so the planned twenty-policy study was not run (Appendix L). An available consensus must therefore be assessed for its relevance to the receiver's task. Rate adaptation can also change the comparison by changing both the attempted traffic and the trajectory.

Message interactions and computation. The value analysis isolates one action intervention under a common continuation. Actual messages also update receiver state, interact with other deliveries, and influence later communication. Closed-loop Spread evaluation includes these effects. Larger tasks and stronger learned set selectors would test how far the local ranking signal continues to help.

Reference construction adds computation. With six candidates, query-inclusive median CPU latency is 3.872 ms for cached CA-RAPU, 2.749 ms for KL, and 3.432 ms for DeepSets. Reusing additive peer summaries reduces reference work while preserving the tested selections. Appendix M reports the timing protocol and scaling measurements.

9 Conclusion

Peer evidence gives a directional interpretation to the response of a frozen receiver. CA-RAPU uses this reference to rank generated messages, improving over unsigned KL at equal traffic and under calibrated communication prices in controlled Spread. The analysis shows when reference accuracy preserves a local allocation decision, and the interventions demonstrate why the distribution of peer support matters. The usefulness of reference alignment depends on what the queue reveals about receiver actions.

Reproducibility Statement

Appendix B specifies the scheduler and numerical conventions, Appendix D gives the full assumptions and proofs, and Appendix E describes training and data generation. The evaluation sections specify development splits, paired measurements, independent statistical units, and comparison families.

AI Use Statement

Generative AI assisted with hypotheses and experimental design, research code, mathematical reasoning, analysis and interpretation, literature search, figure preparation, writing, and translation. AI-assisted code and analyses were checked through targeted tests and recomputation from recorded experimental outputs. The authors are responsible for the final text, claims, and artifacts.

References

Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. TarMAC: Targeted Multi-Agent Communication. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1538–1546. PMLR, 2019.

Ziluo Ding, Tiejun Huang, and Zongqing Lu. Learning Individually Inferred Communication for Multi-Agent Cooperation. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22069–22079, 2020.

Ziluo Ding, Zeyuan Liu, Zhirui Fang, Kefan Su, Liwen Zhu, and Zongqing Lu. Multi-agent coordination via multi-level communication. In *Advances in Neural Information Processing Systems*, volume 37, pp. 118513–118539, 2024. doi: 10.52202/079017-3763.

Yijia Fan, Jusheng Zhang, Kaitong Cai, Jing Yang, Chengpei Tang, Jian Wang, and Keze Wang. Cost-effective communication: An auction-based method for language agent interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29412–29420, 2026. doi: 10.1609/aaai.v40i35.40182.

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*, volume 29, pp. 2137–2145, 2016.

Zihong Gao, Changsheng Qu, Yuanhui Hao, and Liangjun Ke. Communication in multiagent reinforcement learning via counterfactual message value. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 55(11):8152–8165, 2025. doi: 10.1109/TSMC.2025.3604230.

Claudia V. Goldman and Shlomo Zilberstein. Optimizing information exchange in cooperative multi-agent systems. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 137–144. ACM, 2003. doi: 10.1145/860575.860598.

Xudong Guo, Daming Shi, and Wenhui Fan. Scalable Communication for Multi-Agent Reinforcement Learning via Transformer-Based Email Mechanism. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 126–134, 2023. doi: 10.24963/ijcai.2023/15.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z. Leibo, and Nando de Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3040–3049. PMLR, 2019.

Jiechuan Jiang and Zongqing Lu. Learning attentional communication for multi-agent cooperation. In *Advances in Neural Information Processing Systems*, volume 31, pp. 7254–7264, 2018.

Daewoo Kim, Sangwoo Moon, David Earl 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*, 2019.

Pei Li, Yongkang Zhang, Zhonglin Lv, Jinmin Zhu, and Jiangjin Yin. CAIC: Congestion-Aware Intent Communication for Multi-Agent Reinforcement Learning. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 3705–3718, 2026.

Xinran Li, Xiaolu Wang, Chenjia Bai, and Jun Zhang. Exponential Topology-Enabled Scalable Communication in Multi-Agent Reinforcement Learning. In *International Conference on Learning Representations*, 2025.

Genjia Liu, Anning Hu, Yue Hu, Wenjun Zhang, and Siheng Chen. Rate-Distortion Optimized Pragmatic Communication for Collaborative Perception. In *International Conference on Learning Representations*, volume 2026, pp. 119998–120019, 2026.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6379–6390, 2017.

Ryan Lowe, Jakob N. Foerster, Y-Lan Boureau, Joelle Pineau, and Yann N. Dauphin. On the pitfalls of measuring emergent communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 693–701. IFAAMAS, 2019.

William Macke, Reuth Mirsky, and Peter Stone. Expected value of communication for planning in ad hoc teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11290–11298, 2021. doi: 10.1609/aaai.v35i13.17346.

Shilpa Mukhopadhyay, Amit Roy-Chowdhury, and Hang Qiu. COOPERTRIM: Adaptive Data Selection for Uncertainty-Aware Cooperative Perception. In *International Conference on Learning Representations*, volume 2026, pp. 124109–124126, 2026.

Maayan Roth, Reid G. Simmons, and Manuela M. Veloso. Reasoning about joint beliefs for execution-time communication decisions. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 786–793. ACM, 2005. doi: 10.1145/1082473.1082593.

Shoucheng Song, Youfang Lin, Sheng Han, Chang Yao, Hao Wu, Shuo Wang, and Kai Lv. CoDe: Communication delay-tolerant multi-agent collaboration via dual alignment of intent and timeliness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23304–23312, 2025. doi: 10.1609/aaai.v39i22.34497.

Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2244–2252, 2016.

Chuxiong Sun, Zehua Zang, Jiabao Li, Jiangmeng Li, Xiao Xu, Rui Wang, and Changwen Zheng. T2MAC: Targeted and Trusted Multi-Agent Communication through Selective Engagement and Evidence-Driven Integration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15154–15163, 2024. doi: 10.1609/aaai.v38i13.29438.

J. K. Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S. Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, Niall L. Williams, Yashas Lokesh, and Praveen Ravi. PettingZoo: Gym for multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 34, pp. 15032–15043, 2021.

Manav Vora, Gokul Puthumanaillam, Hiroyasu Tsukamoto, and Melkior Ornik. SCoUT: Scalable Communication via Utility-Guided Temporal Grouping in Multi-Agent Reinforcement Learning. *Reinforcement Learning Journal*, 7, 2026.

Rundong Wang, Xu He, Runsheng Yu, Wei Qiu, Bo An, and Zinovi Rabinovich. Learning efficient multi-agent communication: An information bottleneck approach. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9908–9918. PMLR, 2020.

Di Xue, Lei Yuan, Zongzhang Zhang, and Yang Yu. Efficient Multi-Agent Communication via Shapley Message Value. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 578–584, 2022. doi: 10.24963/ijcai.2022/82.

Guang Yang, Tianpei Yang, Jingwen Qiao, Yanqing Wu, Jing Huo, Xingguo Chen, and Yang Gao. Multi-Agent Reinforcement Learning with Communication-Constrained Priors. In *Advances in Neural Information Processing Systems*, volume 38, pp. 154040–154060, 2025. doi: 10.52202/085713-5150.

Lebin Yu, Yunbo Qiu, Quanming Yao, Yuan Shen, Xudong Zhang, and Jian Wang. Robust communicative multi-agent reinforcement learning with active defense. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17575–17582, 2024. doi: 10.1609/aaai.v38i16.29708.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3391–3401, 2017.

Qian Zhang, Zhuo Sun, Yao Zhang, Zhiwen Yu, Bin Guo, and Jun Zhang. VIL2C: Value-of-information aware low-latency communication for multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29874–29882, 2026. doi: 10.1609/aaai.v40i35.40234.

Sai Qian Zhang, Qi Zhang, and Jieyu Lin. Succinct and robust multi-agent communication with temporal message control. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17271–17282, 2020.

A Evaluation Overview

Table 2 links each main claim to its cohort, communication constraint, endpoint, and inferential unit. The detailed protocols and comparison families appear in the cited sections.

Table 2: Studies organized by the questions asked in the main text.

Question	Evidence	Communication and endpoint	Inference	Main conclusion
Task-relevant direction	Four-agent reference–value diagnostics (Section F) and Spread-3 sign/source interventions (Section G)	Candidate-local four-step utility; source comparison uses exact-$B = 2$ episodic return	Twenty independently trained stacks per cohort; paired stack-level estimates within each declared family	Peer direction aligns with short-horizon value, and replacing a receiver-self reference improves matched-traffic return
Matched-traffic selection	Unified response and learned rankers (Section H), Joint-DirectQ (Section H.5), and fixed-$M = 8$ topology intervention (Section K)	Exact $B = 2$ and equal unit-cost traffic; task return in Spread-3 and four-step selected-set utility in Spread-4	Twenty stacks; paired Student-t intervals and declared multiplicity families	Signed direction improves message choice over unsigned KL and DeepSets; peer-support topology changes the effect
Priced deployment	Calibrated abstention and price studies (Section I) plus component interventions (Section J)	Net return and attempt rate at $\lambda_{\mathrm{comm}} = .005$, five-price frontier, and heterogeneous costs	Thresholds selected on five development stacks and evaluated on twenty held-out stacks	Directional ranking supports a high-net-return operating point; each deployment factor contributes under the tested costs

B Receiver-Policy Scheduling Interface

This section specifies the receiver queries, peer reference, normalization, selection rule, and information available to the scheduler.

B.1 Candidates and receiver-policy queries

At step t, the candidate inventory E_t contains the available directed transmissions $e = (i \to j, m_e)$, with sender i, receiver j, and payload m_e, in a fixed canonical order. Scheduling takes place after payload generation, with the movement policy and communication model frozen. The scheduler allocates the limited transmission attempts among these candidates.

For receiver j with finite action set $\mathcal{A}_j$, local observation $o_{j,t}$, and current policy and message state $x_{j,t}$, the frozen policy returns an action distribution. We denote by

$$\pi_j^0 = \pi_\theta(\cdot \mid o_{j,t}, x_{j,t})$$

the distribution without candidate e, and by

$$\pi_e^+ = \pi_\theta(\cdot \mid o_{j,t}, x_{j,t} \oplus (m_e, 0))$$

the distribution after the task interface's focal age-zero update. Both queries start from the same receiver state and retain every non-focal buffer entry. In Spread, the focal update appends m_e at age zero to a copied buffer. Previously delivered messages, including those from sender i, remain in both queries. These counterfactual queries leave the live buffer, policy state, and subsequent action and channel randomness unchanged.

B.2 Task adapter and shared scheduling core

CA-RAPU isolates the only task-specific operation in a reference adapter. Let $x_{j,t}$ be the retained receiver state and let $\mathcal{P}_{j,t}^{-e}$ contain current evidence addressed to receiver j from senders other than i. The adapter returns

$$h_{e,t} = H_{\mathcal{T}}(x_{j,t}, \mathcal{P}_{j,t}^{-e}; \pi_\theta) \in [-1, 1]^{|\mathcal{A}_j|}.$$

Its inputs are the retained receiver state, decoded messages from other senders, fixed action semantics, and queries to the frozen receiver. Rewards, critics, planners, privileged or future environment

states, and realized channel outcomes are unavailable. Holding $x_{j,t}$ and this peer evidence fixed makes $h_{e,t}$ invariant to the current focal payload m_e. This condition concerns the evidence used to construct the reference; it leaves previously received messages in $x_{j,t}$. The adapter is fixed before scheduler development and has no trainable parameters in the reported tasks. The same normalization, signed projection, geometric combination, deployment score, and capacity rule then apply across task interfaces.

The adapter supplies a bounded action score $h_{e,t}(a) \in [-1, 1]$ from candidate-external evidence available before transmission. In Spread, this score is $h_{e,t}(a) = \langle v_a, d_{e,t}^{(-i)} \rangle$, where v_a is the unit movement vector for action a (zero for staying still) and $d_{e,t}^{(-i)}$ is the unit direction from receiver j toward the peer reference. The direction is zero when the reference coincides with the receiver. This reference is formed from current candidates addressed to j after excluding sender i. For each remaining sender, the adapter averages its decoded vectors and pooled confidence. Confidence values are clipped to $[10^{-3}, 1]$ before averaging. Each resulting sender weight is capped at 1.5 times the median sender weight plus 10^{-9}, and the reference is their weighted mean. Invalid weights use a coordinatewise median, and the resulting coordinates are clipped to $[0, 1]^2$. When no peer remains, $g_e = 0$. Adaptive selection then omits the candidate, while fixed traffic applies the shared force-fill rule. The same two policy queries define the exact divergence used in the analysis, its numerically stabilized scoring counterpart, and signed progress,

$$\begin{aligned} u_e^{\mathrm{KL}} &= D_{\mathrm{KL}}(\pi_e^+ \| \pi_j^0), \\ \widetilde{u}_e &= \sum_{a \in \mathcal{A}_j} \pi_e^+(a) \left[\log\big(\pi_e^+(a) + \epsilon_{\mathrm{KL}}\big) - \log\big(\pi_j^0(a) + \epsilon_{\mathrm{KL}}\big) \right], \\ g_e &= \sum_{a \in \mathcal{A}_j} [\pi_e^+(a) - \pi_j^0(a)] h_{e,t}(a). \end{aligned} \tag{11}$$

Here $\epsilon_{\mathrm{KL}} = 10^{-8}$. The analytical inequalities use the exact divergence u_e^{KL}, while the numerical score uses $\widetilde{u}_e$. The sign of g_e records whether the receiver response follows the reference.

The score standardizes $\widetilde{u}_e$ within the current inventory: $\widehat{u}_e = \max\{0, (\widetilde{u}_e - \mu_t)/(v_t + 10^{-8})\}$ when $|E_t| > 1$ and $v_t \geq 10^{-8}$, and zero otherwise. Let μ_t and v_t denote the population mean and standard deviation of $\{\widetilde{u}_e : e \in E_t\}$. Let $P_t = \{e : g_e > 0\}$ and $\rho_t = (|P_t|^{-1} \sum_{e \in P_t} g_e^2)^{1/2}$ when P_t is nonempty, and $\rho_t = 0$ otherwise. It sets $\widehat{g}_e = \max(g_e, 0)/\rho_t$ when $\rho_t > 10^{-10}$ and sets all $\widehat{g}_e$ to zero otherwise. Directional utility, deployment weight, and the final score are

$$\begin{aligned} z_e &= \sqrt{\widehat{u}_e \widehat{g}_e}, \qquad r_e = (1 - p_e^{\mathrm{drop}}) \exp(-\eta d_e), \\ w_e &= r_e \overline{q}_{i,t}^{\alpha}, \qquad s_e = w_e z_e - \lambda_{\mathrm{score}} c_e, \\ A_t &= \text{Top-}B\{e \in E_t : z_e > 0,\ s_e > \tau\}. \end{aligned} \tag{12}$$

Here $\overline{q}_{i,t}$ pools frozen confidence predictions from sender i up to step t. No realized dropout or delay is visible to the selector. The shared core uses $\eta = 0.25$, $\alpha = 1.5$, and $\lambda_{\mathrm{score}} = 0.03$, fixed before the evaluation cohorts. Section J reports local sensitivity. Adaptive thresholds are selected on five development stacks. The score coefficient λ_{score} controls ranking; the reporting price $\lambda_{\mathrm{comm}} = 0.005$ charges attempts in the primary Spread studies. Fixed-traffic comparisons disable eligibility and thresholding and select exactly B candidates by the specified ordering. In the fixed-$M = 8$ study, the mask defines the scheduler's eight-edge input: peer references, confidence history, normalizers, and scores are all recomputed within that queue.

Table 3 summarizes the information available to each component and when that component is fixed.

B.3 Stable selection and exact tie handling

Every study uses an exact, deterministic tie rule fixed before evaluation; none defines an epsilon-wide near-tie region. The current selector used by the unified, receiver-source, component, and fixed-M topology studies orders candidates lexicographically by

$$(s_e, w_e, -k_e), \qquad w_e = r_e \overline{q}_{i,t}^{\alpha},$$

in descending order, where k_e is the canonical candidate index. The deployment weight is consulted only when two floating-point scores are exactly equal, and the index only when both preceding keys tie.

Algorithm 1 CA-RAPU at one decision step. Receiver queries are side-effect-free and do not alter the live rollout.

Input. Frozen actor and codec; inventory E_t; frozen confidence head and task adapter $H_{\mathcal{T}}$; pre-channel $(p_e^{\mathrm{drop}}, d_e, c_e)$; capacity B; threshold τ.

1. Query the receiver. From the same retained-buffer snapshot, compute π_j^0 once per receiver and π_e^+ once per candidate, holding every non-focal entry fixed. Apply only the task-defined focal age-zero update; in Spread, this appends the candidate to the copied buffer.

2. Form a candidate-external reference. For each e, evaluate $h_{e,t} = H_{\mathcal{T}}(x_{j,t}, \mathcal{P}_{j,t}^{-e}; \pi_\theta)$ without exposing the focal payload to the reference. The Spread adapter realizes this call by removing sender i and mapping the remaining receiver-matched vectors to action scores.

3. Build directional utility. Compute $\widetilde{u}_e$ and $g_e = \langle \pi_e^+ - \pi_j^0, h_{e,t} \rangle$; normalize over E_t to obtain $\widehat{u}_e$ as defined above. Set $\widehat{g}_e = \max(g_e, 0)/\rho_t$ if $\rho_t > 10^{-10}$ and $\widehat{g}_e = 0$ otherwise, then set $z_e = \sqrt{\widehat{u}_e \widehat{g}_e}$.

4. Apply deployment conditions. Pool the frozen confidence estimates to $\overline{q}_{i,t}$ and compute $w_e = (1 - p_e^{\mathrm{drop}}) e^{-\eta d_e} \overline{q}_{i,t}^{\alpha}$ and $s_e = w_e z_e - \lambda_{\mathrm{score}} c_e$.

5. Allocate. Among candidates with $z_e > 0$ and $s_e > \tau$, return up to B in descending lexicographic order $(s_e, w_e, -k_e)$, where k_e is the canonical edge index.

Table 3: **Information and freeze boundary.** Task adapters may differ; every operation after $h_{e,t}$ is shared. "Held-out use" records whether a held-out evaluation outcome can affect the object.

Object	Source	Fixed before	Selection-time role	Held-out use
Actor, codec, candidate generator	Upstream task training	Scheduler development	Frozen inference only	None
Task adapter $H_{\mathcal{T}}$	Retained receiver state, decoded nonfocal evidence, fixed action semantics, and frozen-receiver queries	Scheduler development	Forms $h_{e,t}$ from pre-transmission evidence	None
Confidence provider	Upstream auxiliary training; neutral $\overline{q} = 1$ if absent	Scheduler development	Frozen inference only	None
Normalization, η, α, λ_{score}, capacity and tie rule	Shared method specification	Scheduler development	Fixed computation	None
Rate threshold τ	Disjoint development trajectories	Held-out evaluation	Fixed operating point	None
Reported effects	Held-out trajectories	Not applicable	Never visible to selector	Evaluation only

Auxiliary diagnostics use the deterministic keys specified for each study. Replaying the adaptive and five-price studies with the earlier two-key rule and the three-key rule above gives identical selected edges.

The selector can be maintained in $O(|E_t| \log B)$ time and $O(B)$ auxiliary memory. Because B is fixed in the reported deployments, its selection stage is linear in the candidate count. Receiver-policy queries and score construction account for essentially all measured latency; Top-B maintenance is negligible.

B.4 What the interface assumes

The interface requires a frozen receiver policy with candidate-absent and candidate-present queries, and payloads that exist before attempts are allocated. It fits communication stacks with separable encoders, decoders, and actors. Systems that jointly generate messages and choose task actions online would require a different interface.

The score evaluates one candidate at a time. The identity below therefore concerns a conditional intervention at a fixed state. Interactions between simultaneous messages require a separate analysis; Section K examines selected sets with a controlled four-step branch intervention.

C Nearest-Work Interface Comparison

Table 4 distinguishes information used during training from inputs available when communication is selected. Here "post-generation" means that a selector can be attached after the actor, codec, and candidate generator have been fixed; it is stronger than merely freezing parameters at ordinary test time.

Table 4: **Nearest methods by training and deployment boundary.** Entries describe the original published mechanisms rather than queue-level controls adapted in this paper. Selection-time inputs exclude signals used only during training.

Method	Training boundary	Selection-time inputs	Allocation unit	Post-generation	Abstention and capacity semantics
VIL2C (Zhang et al., 2026)	MAPPO task policy and CTDE-trained resource allocator	Local observation and channel state; receiver action entropy during reception	Link bandwidth and power	No	Continuous resource budgets; entropy-triggered stopping with a maximum wait
TEM (Guo et al., 2023)	Action and Transformer message/recipient networks trained jointly; receiver KL enters the communication loss	Local observation and current message-chain buffer	No-send or one-recipient forwarding action	No	Explicit no-send; forwarding also stops at a repeated recipient or fixed chain limit
SchedNet (Kim et al., 2019)	Encoder, action selector, and scheduling-weight generator trained jointly with centralized critics	Per-agent scheduling weights computed from local observations	Broadcasting sender on a shared medium	No	Top-k or softmax access for a fixed number of simultaneous senders
SCoUT (Vora et al., 2026)	Action, send, and recipient heads trained jointly under CTDE; grouping and critics are training-only	Decentralized policy state and mailbox inputs	Send/no-send plus recipient choice	No	Learned silence and recipient decisions; no priced global-capacity term
SMS/SMV (Xue et al., 2022)	A Shapley-value message selector is trained with task-derived coalition utilities	Locally predicted Shapley message values	Peer request before receiving the selected message	No	Requests peers with positive predicted value; no priced global Top-B allocation
CMVC (Gao et al., 2025)	Actor, communication module, and message-value predictor trained from centralized-critic labels	Locally predicted counterfactual message values	Peer request followed by value-weighted message aggregation	No	Requests peers with positive predicted value; no priced global Top-B allocation
CAIC (Li et al., 2026)	Temporal intent encoder, asynchronous fusion, and frequency controller learned as one communication stack	Agent history, predicted intent change, and queue-delay estimate	Whether a new intent message enters the shared queue	No	Learned sending frequency under state-dependent queue service; no post-generation Top-B choice
CA-RAPU	Actor, codec, and generator fixed; adapter and optional confidence provider frozen before evaluation	Absent/present receiver queries, candidate-external peer evidence, and pre-channel link metadata	Queued directed message	Yes	Strict threshold permits silence; deterministic Top-B capacity and attempt-cost term

D Direction Identification and Single-Decision Perturbation Analysis

We first show what signed information adds to receiver influence, then relate an ideal action reference to a single-decision value shift. The remaining results bound how reference error affects the implemented score and selected set. The analysis assumes a common continuation after the current action; message interactions and changing future communication are evaluated through rollouts.

D.1 IDENTIFICATION LOWER BOUND FOR SIGN-INVARIANT RESPONSE

Proposition 5 (Two-environment lower bound for unsigned selectors)**.** *Let $\mathcal{A}_j = \{0, 1\}$, $B = 1$, and $\varepsilon \in (0, 1/2)$. Set $p^0(1) = 1/2$ and let candidates e_+ and e_- induce $p^+(1) = 1/2 + \varepsilon$ and $p^-(1) = 1/2 - \varepsilon$. Their confidence, reliability, delay, and other deployment metadata are identical, and communication cost is fixed to zero. Two task continuations exchange values:*

$$Q_+(1) = Q_-(0) = 1, \qquad Q_+(0) = Q_-(1) = 0. \tag{13}$$

Consider any randomized selector over $\{e_+, e_-, \varnothing\}$, where abstention $\varnothing$ retains p^0. All inputs observable to the selector have the same joint law in the two continuations. Its decision may use fixed labels, the common metadata, and sign-invariant response magnitudes equal for the two candidates (including equal KL, an equal ℓ_p response norm, or equal $|g|$), but not continuation identity, reward, value, or a signed reference. Its worst-continuation expected task-value regret relative to the better candidate is at least ε. An oracle signed observation, which is not assumed deployable, selects the better candidate in both continuations and has zero regret.

Proof. Symmetry gives $D_{\mathrm{KL}}(p^+\|p^0) = D_{\mathrm{KL}}(p^-\|p^0)$, so the selector uses the same probabilities r_+, r_-, and r_0 in both continuations. Its two regrets are $2\varepsilon r_- + \varepsilon r_0$ and $2\varepsilon r_+ + \varepsilon r_0$. Their average is $\varepsilon(r_+ + r_- + r_0) = \varepsilon$, so the larger is at least ε. The bound is attained whenever $r_+ = r_-$, including pure abstention. Ideal signed progress is $+\varepsilon$ for the preferred candidate and $-\varepsilon$ for the other, yielding zero regret. □

A reward-aware selector or directional reference can distinguish the two continuations. The experiments test whether peer evidence provides useful direction under the stated scheduling interface.

D.2 IDEAL-REFERENCE IDENTITY UNDER COMMON CONTINUATION

Fix context x_t, receiver j with finite action set $\mathcal{A}_j$, and candidate e, and impose:

- (A1) the interventions differ only in whether receiver j's current action is drawn from π_j^0 or π_e^+;
- (A2) conditional on every receiver action $a \in \mathcal{A}_j$, the joint law of the other agents' current actions and the entire subsequent return trajectory is identical under the two interventions, including all later transitions, observations, actions, messages, internal states, and rewards;
- (A3) return R is integrable under the common continuation, with conditional value $Q_e(a) = \mathbb{E}[R \mid x_t, A_j = a]$; and
- (A4) for the finite inventory E_t, there is a common scale $M_t > 0$ and candidate-specific offsets $b_e \in \mathbb{R}$ such that $h_e^\star(a) = (Q_e(a) - b_e)/M_t \in [-1, 1]$ for every candidate and action.

Proposition 6 (Current-decision identity under common continuation)**.** *Under (A1)–(A4), define $g_e^\star = \mathbb{E}_{\pi_e^+}[h_e^\star] - \mathbb{E}_{\pi_j^0}[h_e^\star]$ and $\Delta_e = \mathbb{E}_{\pi_e^+}[Q_e] - \mathbb{E}_{\pi_j^0}[Q_e]$. Then $\Delta_e = M_t g_e^\star$.*

Proof. Substitute $Q_e(a) = M_t h_e^\star(a) + b_e$. The b_e term cancels because both action distributions sum to one, leaving $\Delta_e = M_t \sum_a (\pi_e^+(a) - \pi_j^0(a)) h_e^\star(a) = M_t g_e^\star$. □

The common positive scale makes ideal progress preserve both the sign and the within-decision ordering of the conditional current-decision value shift. Reliability, confidence, cost, and reference approximation can still change the ordering of the deployed score.

The common continuation isolates a transient action intervention. Actual delivery also updates the receiver buffer and can alter later communication; the closed-loop experiments measure these additional effects.

D.3 SURROGATE-REFERENCE PERTURBATION

Let h_e be the deployed reference, $g_e = \mathbb{E}_{\pi_e^+}[h_e] - \mathbb{E}_{\pi_j^0}[h_e]$, and assume $\|h_e - h_e^\star\|_\infty \le \delta_e$.

Proposition 7 (Surrogate-reference error)**.** *Under (A1)–(A4),*

$$|\Delta_e - M_t g_e| \le M_t \delta_e \|\pi_e^+ - \pi_j^0\|_1. \tag{14}$$

If $\pi_e^+ \ll \pi_j^0$, Pinsker's inequality further gives

$$|\Delta_e - M_t g_e| \le M_t \delta_e \sqrt{2 D_{\mathrm{KL}}(\pi_e^+ \| \pi_j^0)}. \tag{15}$$

Without absolute continuity, the KL statement is valid but uninformative.

Proof. Proposition 6 and Hölder's inequality give

$$|\Delta_e - M_t g_e| = M_t \left| \sum_a (\pi_e^+(a) - \pi_j^0(a))(h_e^\star(a) - h_e(a)) \right| \tag{16}$$

$$\le M_t \delta_e \|\pi_e^+ - \pi_j^0\|_1. \tag{17}$$

Pinsker's inequality, $\|p - q\|_1 \le \sqrt{2 D_{\mathrm{KL}}(p\|q)}$, gives the second bound. □

The reference error δ_e is an assumption in this bound. Our diagnostic measures sign, rank, selected-set, and cosine alignment on sampled states.

D.4 Reference concentration with multiple peer signals

The preceding result treats reference error as an assumption. A simple model of noisy peers makes its dependence on effective support explicit. For a fixed candidate e with $K_e \ge 1$ peers, let $Y_{e,1}, \dots, Y_{e,K_e} \in [-1,1]^{|\mathcal{A}_j|}$ be candidate-external peer score vectors, conditionally independent given the decision context. Define

$$h_e = \frac{1}{K_e} \sum_{k=1}^{K_e} Y_{e,k}, \qquad \bar{\mu}_e = \frac{1}{K_e} \sum_{k=1}^{K_e} \mathbb{E}[Y_{e,k} \mid x_t], \tag{18}$$

and allow a deterministic approximation bias $\|\bar{\mu}_e - h_e^\star\|_\infty \le \beta_e$. We use a distinct symbol here because this bias is unrelated to the additive offset in assumption (A4).

Proposition 8 (Concentration of a bounded peer reference)**.** *For any $\zeta \in (0,1)$, with conditional probability at least $1 - \zeta$,*

$$\|h_e - h_e^\star\|_\infty \le \beta_e + \sqrt{\frac{2 \log(2|\mathcal{A}_j|/\zeta)}{K_e}}. \tag{19}$$

Consequently, on the same event,

$$|\Delta_e - M_t g_e| \le M_t \|\pi_e^+ - \pi_j^0\|_1 \left[\beta_e + \sqrt{\frac{2 \log(2|\mathcal{A}_j|/\zeta)}{K_e}} \right]. \tag{20}$$

For a finite inventory, applying the statement with failure probability $\zeta/|E_t|$ and taking a union bound makes both inequalities simultaneous for all candidates; no independence across candidates is required.

Proof. For each action coordinate, Hoeffding's inequality for conditionally independent variables in $[-1,1]$ gives $\Pr(|h_e(a) - \bar{\mu}_e(a)| > r \mid x_t) \le 2\exp(-K_e r^2/2)$. A union bound over actions with $r = \sqrt{2\log(2|\mathcal{A}_j|/\zeta)/K_e}$, followed by the triangle inequality and Proposition 7, proves the claim. The final statement is another union bound. □

The stochastic term decreases as $K_e^{-1/2}$, whereas systematic bias β_e remains. Conditional independence and centering are sufficient modeling conditions, not properties guaranteed by the interface. Correlated, duplicated, or jointly biased peers may provide little additional information; the fixed-inventory topology intervention probes this distinction empirically. Equation (19) may be substituted for δ_e in the support and score-boundary conditions below.

D.5 SINGLE-DECISION SCORE AND SELECTION PERTURBATION

We now propagate reference error through the score that CA-RAPU actually deploys. Fix a finite inventory E_t, $1 \le B < |E_t|$, the receiver distributions and hence u_e^{KL}, $\widetilde{u}_e$, and $\widehat{u}_e$, the deployment weights w_e, costs c_e, threshold τ, and the stable tie rule. Let starred quantities use $h_e^\star$ and suppose $\|h_e - h_e^\star\|_\infty \le \delta_e$. Proposition 7 gives

$$|g_e - g_e^\star| \le \varepsilon_e, \qquad \varepsilon_e := \delta_e \|\pi_e^+ - \pi_j^0\|_1 \le \delta_e \sqrt{2u_e^{\mathrm{KL}}}. \tag{21}$$

Write $P^\star = \{e : g_e^\star > 0\}$. Assume the reference error cannot change positive support:

$$g_e^\star > \varepsilon_e \quad (e \in P^\star), \qquad g_e^\star + \varepsilon_e \le 0 \quad (e \notin P^\star). \tag{22}$$

Then $P = \{e : g_e > 0\} = P^\star$. If $P^\star$ is nonempty, set

$$\rho_t^\star = \left(\frac{1}{|P^\star|} \sum_{e \in P^\star} (g_e^\star)^2 \right)^{1/2}, \qquad \overline{\varepsilon}_P = \left(\frac{1}{|P^\star|} \sum_{e \in P^\star} \varepsilon_e^2 \right)^{1/2}. \tag{23}$$

For the implementation floor $\kappa = 10^{-10}$, if $\rho_t^\star + \overline{\varepsilon}_P \le \kappa$, both references take the zero-RMS branch. If instead

$$\rho_t^\star - \overline{\varepsilon}_P > \kappa, \tag{24}$$

both take the nonzero branch and

$$|\widehat{g}_e - \widehat{g}_e^\star| \le \Gamma_e, \qquad \Gamma_e := \begin{cases} \dfrac{\varepsilon_e + \widehat{g}_e^\star \overline{\varepsilon}_P}{\rho_t^\star - \overline{\varepsilon}_P}, & e \in P^\star, \\ 0, & e \notin P^\star. \end{cases} \tag{25}$$

No stability guarantee is made in the boundary band $\rho_t^\star - \overline{\varepsilon}_P \le \kappa < \rho_t^\star + \overline{\varepsilon}_P$. If $P^\star$ is empty, or if both references take the zero-RMS branch, every z_e is zero and both adaptive selectors abstain. Fixed-traffic selection still force-fills B slots and is handled separately below. The remainder of the adaptive argument considers the nonzero branch.

For $z_e = \sqrt{\widehat{u}_e \widehat{g}_e}$ and $s_e = w_e z_e - \lambda_{\mathrm{score}} c_e$, define

$$|z_e - z_e^\star| \le \sqrt{\widehat{u}_e \Gamma_e}, \qquad |s_e - s_e^\star| \le \xi_e := |w_e| \sqrt{\widehat{u}_e \Gamma_e}, \quad \xi = \max_e \xi_e. \tag{26}$$

Proposition 9 (One-decision adaptive selection invariance under bounded reference perturbation)**.** *Under Equations* (22) *and* (24)*, let $\mathcal{I}^\star = \{e : z_e^\star > 0\}$. If this set is empty, both adaptive selectors abstain. Otherwise, suppose it satisfies*

$$\min_{e \in \mathcal{I}^\star} |s_e^\star - \tau| > \xi. \tag{27}$$

Then the deployed and ideal eligible sets coincide. Their selected sets also coincide if the eligible set has at most B candidates or, otherwise, if the ideal eligible scores obey

$$s_{(B)}^\star - s_{(B+1)}^\star > 2\xi. \tag{28}$$

Proof. Equation (22) preserves positive support, while the reverse triangle inequality gives $|\rho_t - \rho_t^\star| \le \overline{\varepsilon}_P$. Equation (24) keeps both normalizations on the nonzero branch and yields Equation (25). The square root is $1/2$-Hölder on nonnegative inputs, so $|\sqrt{a} - \sqrt{b}| \le \sqrt{|a-b|}$ gives Equation (26); the fixed cost term cancels. Positive support and fixed $\widehat{u}_e$ make the informative sets identical. Equation (27) then preserves strict eligibility, and Equation (28) prevents any selected–unselected pair from reversing order. □

Corollary 1 (One-decision fixed-traffic selection invariance)**.** *Let*

$$T^\star = \mathrm{LexTopB}_{e \in E_t}(s_e^\star, w_e, -k_e), \qquad T = \mathrm{LexTopB}_{e \in E_t}(s_e, w_e, -k_e),$$

where both sets use descending lexicographic order. If both references take the zero-RMS branch, then $T = T^\star$. On the nonzero branch, the same conclusion holds if, for every $e \in T^\star$ and $f \notin T^\star$, either

$$s_e^\star - s_f^\star > \xi_e + \xi_f,$$

or $\xi_e = \xi_f = 0$ and the ideal tuple $(s_e^\star, w_e, -k_e)$ precedes $(s_f^\star, w_f, -k_f)$. In particular, a strict full-inventory boundary gap $s_{(B)}^\star - s_{(B+1)}^\star > 2\xi$ is sufficient.

Proof. In the zero-RMS branch, $z_e = z_e^\star = 0$ and both scores equal $-\lambda_{\text{score}} c_e$; the remaining lexicographic keys are fixed. On the nonzero branch, the first condition and Equation (26) preserve every ideal selected–unselected strict score order. When both error radii are zero, the two tuples are unchanged, so the fixed tie rule preserves their order. No candidate can cross the top-B boundary. □

The proposition applies to adaptive selection; the corollary covers the fixed-traffic rule. Both are sufficient conditions at a fixed decision. Redundancy, cross-receiver effects, and selected-set-dependent continuations are studied separately in the four-agent branch experiments.

D.6 Comparator signals

The progress-magnitude control removes the sign and uses the magnitude of reference-projected receiver-policy change. The unsigned receiver-KL control uses the same positive-standardized receiver divergence as CA-RAPU, without the directional progress term. Both controls use the same frozen actor, payloads, candidate inventory, link information, traffic rule, and exact tie convention. They differ only in the ranking signal.

Unsigned KL measures the extent of the receiver's response. Progress magnitude measures its size along the reference direction, while signed progress distinguishes movement toward and away from that direction. All three signals share the same deployment weights, so the comparisons isolate the information contributed by the ranking signal.

E Frozen Learned-Payload Stacks and Training

E.1 Stack construction

The Spread studies use the discrete-action MPE2 Spread transition system, with the local reward mixing coefficient set to 0.5, a simulator limit of 65 cycles, and the agent count determined by the cohort. The upstream reset samples agent and landmark positions on $[-1, 1]^2$ and retains native collision dynamics; we map positions to $[0, 1]^2$. At reset, the lexicographically first permutation with minimum total Euclidean agent–landmark distance defines a one-to-one assignment that remains fixed for the episode. The actor chooses among stay, up, down, left, and right. The reported return is the assignment-based success–distance metric below, not the native MPE reward.

Each learned-payload stack combines a movement actor trained under uniformly sampled communication with a two-dimensional sender encoder, a receiver decoder, a metadata-only native scheduler, and a separate CA-RAPU confidence head. The communication interface is fitted for its actor after actor training. Every component is frozen before scheduler evaluation.

At each active step, after buffered messages age and in-flight deliveries arrive, the environment generates one candidate for every ordered pair $i \to j$ with $i \neq j$. Each candidate contains a two-coordinate noisy clue about receiver j's assigned landmark, a noisy reported-quality field, and pre-transmission link descriptors. Self-edges are absent, so the complete three- and four-agent inventories contain six and twelve candidates, respectively. Candidate generation precedes scheduling: the selector can rank or reject available candidates but cannot synthesize a new payload.

The actor input for agent j has exactly ten entries: normalized position (two), its private noisy assigned-landmark clue (two), the decoded aggregate of delivered messages (two), a message-present bit, mean reported quality, normalized mean message age, and normalized time. Delivered encoded payloads are aggregated with age decay and reported-quality weighting. The decoder maps that two-vector together with mean reported quality and normalized mean age to a two-dimensional assigned-landmark estimate, which occupies the actor's two message coordinates; with no delivery, the original no-message feature row is preserved exactly. The reconstruction target is receiver j's assigned landmark, and reconstruction MSE is applied only when that receiver has a delivered training message.

Delivered messages remain eligible through age eight. At the beginning of a decision, retained messages age and expired entries are removed before due in-flight messages arrive. Their encoded payloads are aggregated with weights $\exp(-0.35\,\text{age})\max(0.05, \widetilde{q})$ before receiver decoding; a

successful zero-delay transmission enters the buffer before the actors choose that decision's actions. A counterfactual receiver query transiently adds the focal age-zero payload to a copy of the current buffer. It neither writes the live buffer nor removes other retained messages.

Behavior cloning initializes the actor by cross-entropy to a deterministic five-action teacher: move along the larger-magnitude coordinate of the vector from the current position to the assigned landmark, or stay at the landmark. The cloned data randomize position and time and, independently for each receiver, include a noisy receiver-directed training message with probability 0.9. PPO then optimizes a shared task reward. If $\bar{d}_t$ is the mean agent-to-assigned-landmark distance, write $D_t = 1 - \min\{\bar{d}_t/\sqrt{2}, 1\}$ for normalized assignment proximity. The per-step training reward is

$$r_t^{\mathrm{PPO}} = 1.25(D_{t+1} - D_t) + 0.01 D_{t+1} - 0.002 + \mathbf{1}_{\mathrm{success}} + 0.5\mathbf{1}_{\mathrm{terminal}} D_{t+1}, \tag{29}$$

where terminal means success or the final horizon step. For the frozen stacks used here, actor PPO sets the communication-cost weight to zero and retains a 0.05 cross-entropy anchor to the same teacher labels.

The encoder is a $2 \to 64 \to 64 \to 2$ MLP, and the decoder is a $4 \to 64 \to 64 \to 2$ MLP; both use tanh, tanh, and sigmoid activations. The native scheduler is a $9 \to 64 \to 64 \to 1$ ReLU MLP with a linear output. Its deployment-visible inputs are current noisy reported quality, normalized sender and receiver identities, effective dropout, normalized expected delay, edge cost, normalized time, base dropout, and normalized maximum delay. It is trained jointly with the encoder and decoder on 32,768 state samples, with 4,096 validation states, for 20 epochs using batches of 512 and AdamW with learning rate 10^{-3} and weight decay 10^{-5}. Its objective gives equal weight to the frozen actor's action cross-entropy and receiver-target reconstruction MSE. This task-supervised native scheduler supplies a comparison method. CA-RAPU instead uses the separate confidence estimate defined below.

At each reset, sender i receives an episode-stable latent quality

$$q_i^\star = \mathrm{clip}\left(q_{\mathrm{mid}} + 2.2 \left[1 - \frac{\sigma_i}{2.1\sigma_{\mathrm{clue}}} - q_{\mathrm{mid}} \right], 0.15, 0.95 \right), \tag{30}$$

where $\sigma_i \sim \mathrm{Uniform}(0.55\sigma_{\mathrm{clue}}, 1.85\sigma_{\mathrm{clue}})$, $\sigma_{\mathrm{clue}} = 0.45$, and $q_{\mathrm{mid}} = 0.55$. A candidate exposes only the noisy report $\widetilde{q}_{e,t} = \mathrm{clip}(q_i^\star + \epsilon_{e,t}, 0.15, 0.95)$, $\epsilon_{e,t} \sim \mathcal{N}(0, 0.15^2)$; it never exposes $q_i^\star$.

The sender's private assigned-landmark clue adds $\mathcal{N}(0, \sigma_i^2 I)$. Its reset-time receiver-directed clue adds noise with standard deviation $0.03[0.5 + 1.5(1 - q_i^\star)]$; each current candidate then adds independent noise with standard deviation $0.03[1 + 1.5(1 - q_i^\star)]$. All clue coordinates are clipped to $[0, 1]$.

CA-RAPU fits a separate $11 \to 64 \to 64 \to 1$ ReLU MLP with a sigmoid output to estimate $q_i^\star$ from the two-dimensional encoded payload and the same nine visible metadata fields. The latent target is used only for offline MSE training and is unavailable to every rollout selector. The disjoint training, validation, and development splits contain 12,288, 3,072, and 3,072 candidate rows, corresponding to 2,048, 512, and 512 independently seeded states with six directed candidates per state. Training uses 12 epochs, batches of 512, AdamW with learning rate 10^{-3}, weight decay 10^{-5}, and gradient norm at most 1, retaining the checkpoint with the lowest validation MSE. For the four-agent confidence head, the same row budgets correspond to 1,536, 384, and 384 balanced eight-edge cycle states; the three Hamiltonian cycles are equally represented. The downstream fixed-$M = 8$ intervention evaluates eight-edge queues drawn from the balanced mask families defined in Section K.

At step t, the frozen confidence head predicts every candidate before ranking. Its clipped predictions are pooled by sender as

$$\overline{q}_{i,t} = \frac{\sum_{\tau=0}^{t} \sum_{e \in E_\tau : s(e) = i} \mathrm{clip}(\widehat{q}_{e,\tau}, 0, 1)}{\sum_{\tau=0}^{t} |\{e \in E_\tau : s(e) = i\}|}. \tag{31}$$

Here $s(e)$ denotes the sender of candidate e. The pool resets between episodes, includes the current step exactly once, and assigns the same $\overline{q}_{i,t}$ to every current outgoing candidate from sender i.

Actor, codec, confidence, and threshold fitting use data disjoint from evaluation trajectories. Diagnostic oracle labels are used only to evaluate selection.

Table 5 summarizes movement-actor training for the two evaluation cohorts.

Table 5: Frozen movement-actor training. Training-time traffic does not set the post-freezing evaluation budget.

Setting	Three-agent cohort	Four-agent cohort
Cohort	20 independent actors	20 independent actors
Actor	$10 \to 128 \to 128 \to 5$ tanh policy	Same
Behavior cloning	12,000 resets (36,000 rows); 6 epochs; batch 2,048; AdamW 3×10^{-4}	Same resets (48,000 rows); otherwise identical
PPO	4 updates; 64 episodes/update; 2 epochs; minibatch 1,024; AdamW 5×10^{-5}	Same
PPO constants	$\gamma = .98$, GAE $= .95$, clip $= .2$, value $= .5$, entropy $= 0$, BC anchor $= .05$, target KL $= .03$	Same
Regularization	Weight decay 10^{-4}; gradient norm $\leq .5$; no centralized critic	Same
Training traffic	Uniform $B = 2$ of 6	Uniform $B = 3$ of 12
Checkpoint	Final update; every seed retained	Same

E.2 Evaluation cohort and statistical unit

Unless a section states otherwise, learned-interface studies use twenty independently trained frozen stacks. Within a paired key, methods share the reset, canonical directed-edge identities and order, cost rule, and complete edge–time channel-potential table. Each method nevertheless owns its physical state, message buffers, pending-delivery queue, payload history, confidence pool, and receiver-policy queries. Candidate values and trajectories may therefore diverge after different communication decisions. Channel cells are averaged within a stack. The frozen stack, not an episode or candidate row, is the inferential unit.

The paired design removes variation that is irrelevant to ranking. A channel potential contains all counterfactual dropout and delay draws needed by every method. Realized outcomes are revealed only after selection. Methods therefore share exogenous randomness without receiving privileged channel information.

Initial task states and potential channel outcomes use independent random streams. This separates task variation from channel variation while preserving paired comparisons across methods.

E.3 Task return and channel grid

Spread episodes have horizon 64 and success radius 0.12. At reset, the agent–landmark assignment is the permutation minimizing total Euclidean distance. If $\bar{d}_T$ is the final mean assigned distance in the normalized two-dimensional world, define

$$D_T = 1 - \min\{\bar{d}_T/\sqrt{2}, 1\}, \qquad R_{\text{task}} = \mathbf{1}_{\text{success}} + 3D_T \in [0, 4]. \tag{32}$$

Success requires every assigned distance to be at most 0.12. Net return is $R_{\text{net}} = R_{\text{task}} - \lambda_{\text{comm}} C_{\text{attempted}}$, where attempted cost includes messages later dropped by the channel. The exact-traffic designs use unit edge costs, hence $C_{\text{attempted}} = N_{\text{attempted}}$.

PPO rollouts and evaluation cross the nine Cartesian-product conditions $p_{\text{base}} \in \{0, 0.5, 0.75\}$ and $L_{\max} \in \{0, 4, 8\}$. Evaluation episode seeds are disjoint from actor and communication-interface training. Behavior cloning is channel-agnostic; PPO uses uniformly sampled candidate identities under the training-time budget.

Each directed edge has one of five channel types, determined by $k = (37i + 19j) \bmod 5$. Its dropout bias b_{ij} and delay scale s_{ij} are the zero-indexed kth entries of

$$\mathbf{b} = (-0.15, -0.05, 0, 0.10, 0.20), \qquad \mathbf{s} = (0.55, 0.75, 1, 1.25, 1.55).$$

In a channel cell,

$$p_{ij}^{\text{drop}} = \text{clip}(p_{\text{base}} + b_{ij}, 0, 0.95), \qquad L_{ij} = \max\{0, \text{round}(L_{\max} s_{ij})\}. \tag{33}$$

Realized delay is discrete uniform on $\{0, \ldots, L_{ij}\}$, while the selector sees only $L_{ij}/2$. Before any method runs, counter-based draws keyed by episode, directed edge, decision, and random stream

materialize the complete dropout–delay potential table. Methods share this exogenous table but never observe a realized outcome before selection. Every attempted edge is charged its declared cost even if it is subsequently dropped.

E.4 Fixed traffic and adaptive traffic

Exact-traffic studies disable signal eligibility and thresholding, then force-fill the fixed number of attempted messages from the declared eligible set. This is the complete inventory in the standard comparisons. In Section K, each eight-edge mask instead defines the complete scheduler input for that decision, and all inventory-dependent quantities are recomputed after masking. These studies use an absorbing-state suffix after early task success: the physical state is held fixed while the remaining prescribed attempts continue and are charged. With unit edge costs, every ranker therefore incurs the same attempt count and attempted cost. The separate rate-adaptation study instead follows natural episode termination for every arm, including its fixed-$B = 2$ reference, so its attempt counts can vary with the time of success.

Adaptive studies instead allow a selector to abstain. Thresholds and any operating-point coefficients are chosen on five disjoint development stacks and frozen before held-out evaluation. Adaptive comparisons therefore measure the joint effect of the ranking rule and its calibrated deployment rate, whereas fixed-traffic comparisons measure ordering under an exact attempt budget.

E.5 Stability of development-set calibration

We test whether a single development stack determines the selected operating point by repeating calibration five times, each time omitting one of the five frozen development stacks. Each fold reuses the same 15-point ordered threshold grid, rebuilds the episode–channel–stack aggregation from the raw development table, and applies the same objective and tie rule. The full-development threshold and every held-out outcome remain unchanged.

At the primary reporting price, the signed CA-RAPU threshold is recovered exactly in all five folds. Unsigned receiver KL is recovered in four folds and moves by one adjacent grid point in the fifth. Progress magnitude is less stable as a grid label, but its largest regret when the reselected threshold is reevaluated on the complete development stream is 0.0127 return units, or 0.55% of the full-development objective. Across the 75 method–price–fold decisions in the multi-price sensitivity analysis, 62 (82.7%) recover the full-development point and 72 (96.0%) remain within one grid step; the largest complete-development regret is 0.0149 return units (0.61%). Independent recomputation agrees on all 90 reselection decisions. These results assess calibration sensitivity under the original primary endpoints, thresholds, and multiplicity families.

F Oracle Alignment

F.1 Truncated action-value diagnostic

The oracle-alignment study asks whether the peer-derived direction agrees with a task-grounded short-horizon value on frozen states. For each receiver action a, exact replay forces that action at the current decision while the other agents and the subsequent no-message continuation follow a matched frozen rule, producing $Q_e^{(4)}(a)$. Candidate-absent and candidate-present values are then the action-probability expectations of this same forced-action table under π_j^0 and π_e^+. This mirrors the expectation in the conditional identity; it is distinct from Joint-DirectQ, which evaluates the frozen actor's greedy joint action at each queried observation. Neither diagnostic changes the deployed scheduler.

The diagnostic equally averages three representative diagonal channel profiles, $(p_{\mathrm{base}}, L_{\max}) \in \{(0, 0), (0.5, 4), (0.75, 8)\}$. Closed-loop studies otherwise average all nine cells of the fixed dropout–delay grid.

The diagnostic reports sign agreement after excluding oracle shifts with absolute value at most 10^{-10}, within-state pairwise rank agreement, centered score–value association, and the difference in mean candidate-local oracle value between the pair selected by CA-RAPU and the pair selected

by unsigned receiver KL. Aggregation occurs within state and episode before stack-level inference. Sign agreement uses three-way signs with the same tolerance.

The diagnostic measures local direction and selection quality. Joint delivery is evaluated separately because candidate-local values can interact.

F.2 Class-balance decomposition of sign agreement

Positive oracle shifts are common, so raw sign agreement is accompanied by a descriptive class-balance decomposition in Table 6. Every quantity is reconstructed from the saved candidate rows after excluding oracle shifts with absolute value at most 10^{-10}. A positive prediction requires strictly positive peer progress; near-zero predictions are counted separately using the same tolerance. Point estimates average the twenty stacks equally; intervals are descriptive 95% Student-t intervals over stacks and are shown for balanced accuracy and Matthews correlation coefficient (MCC).

Table 6: Class-balance decomposition of peer-direction sign agreement with the four-step value diagnostic. Majority accuracy always predicts a positive shift. Intervals are descriptive 95% Student-t intervals across twenty independently trained stacks per cohort.

Metric	Spread-3	Spread-4
Positive prevalence	.866	.860
Majority accuracy	.866	.860
Raw sign agreement	.898	.885
Positive-shift recall	.953	.937
Negative-shift recall	.540	.548
Balanced accuracy	.747 [.716,.777]	.742 [.708,.777]
MCC	.525 [.467,.584]	.501 [.438,.564]

In both cohorts, raw agreement exceeds the majority baseline. The reference recovers about 54–55% of negative shifts and over 93% of positive shifts. Balanced accuracy and MCC are descriptive supplements to the original raw-agreement endpoint.

F.3 Magnitude-matched direction attribution

A separate fixed-state experiment isolates the sign of peer progress from its magnitude on the raw-clue Spread-3 interface. It loads the frozen movement policy, queries the generated raw clues, and uses the checkpoint-native confidence estimator rather than the learned two-dimensional codec and its auxiliary confidence head. This provides a mechanism-transfer diagnostic across communication interfaces. Both forced-$B = 2$ rankings use the same candidates, receiver-policy queries, normalized receiver KL, checkpoint-native confidence, link reliability, unit cost, positive-progress RMS scale, and tie rule. For candidates with positive progress, their signals and final scores are identical across arms. The signed arm clips negative progress, whereas the magnitude-matched control reflects it to its absolute value. When no positive candidate defines a nonzero scale, the fixed fallback assigns zero to both arms; this occurs in 166 of 286,115 reference states. The primary endpoint is the stack-level difference in selected-pair mean one-step true task-aligned expected-action gain.

The study contains 20 frozen Spread-3 stacks, nine channel cells, 25 shared no-communication reference rollouts per cell, and 286,115 shared reference states. The true task-aligned expected-action outcome uses the receiver's assigned landmark, but is read only after both score vectors and selected sets have been fixed; it never enters either ranking. Signed ranking improves the endpoint by 0.010964 [0.010020, 0.011908] ($p = 8.99 \times 10^{-16}$), with positive differences on all 20 stacks. After averaging states within rollouts, rollouts within cells, and cells equally within stacks, the mean set-difference rate is 10.9%, with 94.0% mean identity overlap. An independent recomputation checks the shared queries, positive-candidate equality, budget, and aggregation from the saved state and episode rows. Because the trajectories do not evolve separately under the two rankings, this result attributes local selection quality to direction; it is not a closed-loop return comparison.

F.4 Agreement with Value-Based Selection

A secondary analysis of the complete four-agent cohort replaces only the deployed raw peer gain by the exact $H = 4$ forced-action gain. It retains normalized receiver KL, pooled confidence, link reliability, unit cost, positive-gain RMS normalization, $\lambda_{\text{score}} = 0.03$, exact $B = 2$, and canonical ties. Thus it compares the deployed scheduler with its ideal-direction counterpart, not with unweighted Top-2 oracle gains. Table 7 reports selection agreement for the four-agent analysis and an independently trained three-agent cohort. Figure 3 shows agreement by ideal-margin quartile.

Table 7: Descriptive agreement between deployed and ideal-direction Top-2 selection. Intervals are pointwise 95% Student-t intervals across twenty independently trained stacks; margin strata are equal-count bins formed only from the normalized ideal Top-2/Top-3 score boundary.

Setting	Exact Top-2	Mean Top-2 overlap
Four-agent complete cohort	.7222 [.6867,.7577]	.8451 [.8243,.8660]
Ideal-margin stratum 1	.5530	.7566
Ideal-margin stratum 2	.7583	.8660
Ideal-margin stratum 3	.7572	.8688
Ideal-margin stratum 4	.8126	.8859
Three-agent replication	.8375 [.8103,.8647]	.9167 [.9019,.9314]

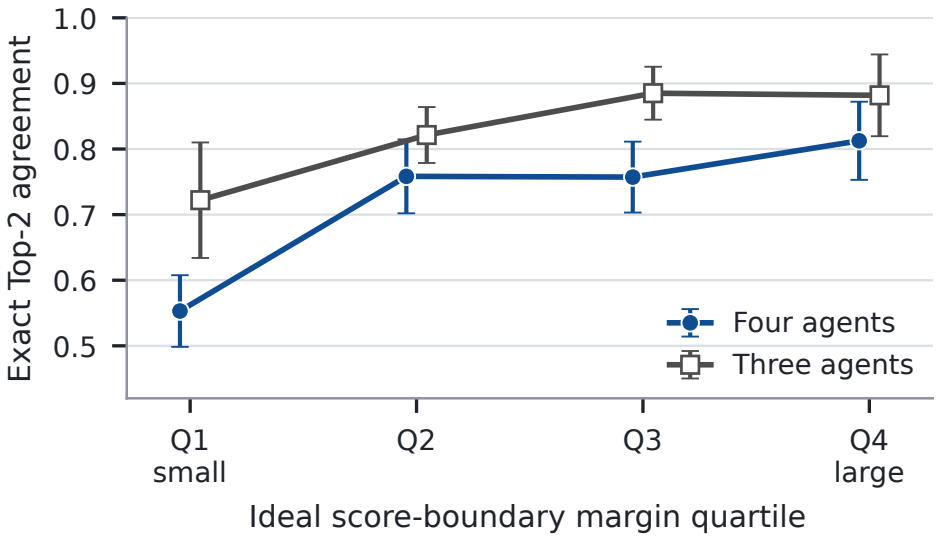

Figure 3: **Selection agreement across ideal score-margin quartiles.** Points show exact Top-2 agreement between the peer-referenced scheduler and the otherwise identical scheduler using four-step action-value direction, stratified by quartiles of the ideal normalized boundary margin. Error bars are descriptive, pointwise 95% Student-t intervals over twenty independently trained stacks; an independently trained three-agent cohort provides a separate replication.

We also test the selection-margin statement on these decisions. The observed margin condition requires every ideal selected–unselected score gap to exceed the sum of the two absolute deployed–ideal score residuals. It holds for 772 of 1,440 four-agent blocks (53.61% [49.88%, 57.34%]) and 1,027 of 1,440 three-agent blocks (71.32% [67.59%, 75.05%]). Every block satisfying it preserves the exact Top-2 set. Among the remaining blocks, stack-level exact agreement is 0.405 [0.354, 0.456] and 0.428 [0.360, 0.495], respectively. This retrospective condition explains fixed-state

agreement using observed score residuals. The stronger uniform sufficient condition derived through $\delta_e \to \xi_e$ holds for 2 and 48 blocks on the same records.

Every four-agent state has positive ideal gain; only 1 of 480 states has no positive ideal gain–KL product, and the deployed product is zero in 0.0083 of state–profile blocks. Disagreement is therefore concentrated near the ideal selection boundary rather than being driven by empty ideal support. This analysis addresses selection agreement; closed-loop effects are evaluated separately.

G Reference-Source Controls

We isolate the source of the directional reference while keeping the remaining scheduler components fixed. For candidate $e = (i \to j, m_e)$, let π_j^0 and π_e^+ be the frozen receiver-policy distributions without and with the candidate, and write $\Delta\pi_e(a) = \pi_e^+(a) - \pi_j^0(a)$. CA-RAPU uses a receiver-matched decoded-peer direction formed from the current candidates after excluding sender i:

$$g_e^{\mathrm{peer}} = \sum_a \Delta\pi_e(a) h_{e,t}^{(-i)}(a). \tag{34}$$

The receiver-self control instead defines $a_j^{\mathrm{self}} = \arg\max_a \pi_j^0(a)$ and

$$g_e^{\mathrm{self}} = \Delta\pi_e(a_j^{\mathrm{self}}), \tag{35}$$

so a candidate is rewarded when it increases the probability of the receiver's own candidate-absent preference. A third, self-inclusive control constructs the same receiver-matched decoded-peer aggregate as CA-RAPU but retains the focal sender, giving $g_e^{\mathrm{incl}} = \sum_a \Delta\pi_e(a) h_{e,t}^{(\mathrm{incl})}(a)$. The CA-RAPU–receiver-self contrast tests the source of the reference within a common scheduling interface. The CA-RAPU–self-inclusive contrast then isolates the effect of allowing the focal sender to contribute to that reference.

For each arm, positive progress is normalized by its within-decision positive RMS and combined geometrically with the same positive normalized receiver KL. The resulting signal z_e enters the common score

$$s_e = w_e z_e - 0.03 c_e, \qquad w_e = r_e \, \bar{q}_{i,t}^{1.5}, \tag{36}$$

where r_e is the pre-transmission delivery–delay factor and $\bar{q}_{i,t}$ is the pooled sender-confidence estimate. All arms use the same frozen actor, encoder, decoder, confidence head, counterfactual query, KL definition, channel model, cost rule, candidate-generation rule, traffic constraint, and tie rule. They share paired resets and edge-time channel potentials, while each trajectory evolves under its own selected messages. Each arm force-fills $B = 2$ candidates at each of 64 steps, without abstention, using descending score, then descending deployment weight, then canonical candidate index. Every trajectory therefore makes 128 attempts.

We evaluate the three arms on twenty frozen Spread-3 communication stacks, nine dropout–delay conditions, and 25 held-out episodes per condition. Episode means are first formed within condition and then averaged equally over the nine conditions, leaving one paired task-return difference per stack. The two primary contrasts, CA-RAPU minus receiver-self and CA-RAPU minus self-inclusive, use two-sided paired Student-t intervals and form one two-test Holm family. Self-inclusive minus receiver-self is descriptive. Exact traffic makes every paired task-return margin identical to its net-return margin. The evaluation uses independently sampled episodes and channel potentials on the same twenty trained stacks.

CA-RAPU exceeds the receiver-self control by 0.099512 [0.077290, 0.121735] task-return units (raw $p = 1.48 \times 10^{-8}$; Holm-adjusted $p = 2.96 \times 10^{-8}$), with positive differences on all twenty stacks. The CA-RAPU–self-inclusive margin is -0.006225 [-0.017259, 0.004810] ($p_{\mathrm{Holm}} = 0.2523$; 9 of 20 positive). Thus the experiment identifies peer-derived direction, rather than the receiver's own candidate-absent preference, as the material reference-source improvement. The self-inclusive comparison is inconclusive about the return benefit of sender exclusion. Exclusion makes the current reference invariant to the focal payload, conditional on the retained receiver state and the other senders' evidence.

The receiver-self arm uses the receiver-best-action reference of ADMAC (Yu et al., 2024) inside CA-RAPU's otherwise unchanged scheduling interface; we therefore refer to it as an *ADMAC-inspired receiver-self control*.

G.1 Sensitivity to Reference Quality

We next ask how accurately the peer reference must represent the useful action direction. This fixed-state study uses twenty Spread-4 stacks. With the complete twelve-edge inventory, each focal edge has two receiver-matched peer senders after excluding the focal sender. We reuse eight no-message episodes per stack, snapshots at steps 0, 4, and 8, and all nine channel profiles, with capacity fixed at $B = 2$. Candidate payloads, confidence histories, receiver-policy queries, normalized receiver KL, deployment weights, costs, and tie handling are shared across conditions; only the peer reference changes. Each selected pair is fixed before either the true landmark assignment or the saved four-step branch utility is consulted.

The five pre-specified stress conditions rotate the aggregate reference by $-30°$ or $+30°$, apply independent clipped Gaussian angular noise with standard deviation $30°$ to the two peer vectors, retain only one of the two peers, or rotate one peer by $\pm 90°$. A one-peer $180°$ reversal provides a severe descriptive condition. The primary endpoint is the selected pair's mean one-step expected-action gain toward the receivers' true assignments. Profiles are averaged within state, the three snapshots within episode, and the eight episodes within stack. The joint criterion requires every moderate condition, relative to direction-removed ranking, to have a two-sided 95% stack-level interval above zero and at least 16 of 20 positive stack differences. The clean reference, severe reversal, and four-step branch endpoint are secondary. Figure 4 and Table 8 report the paired effects.

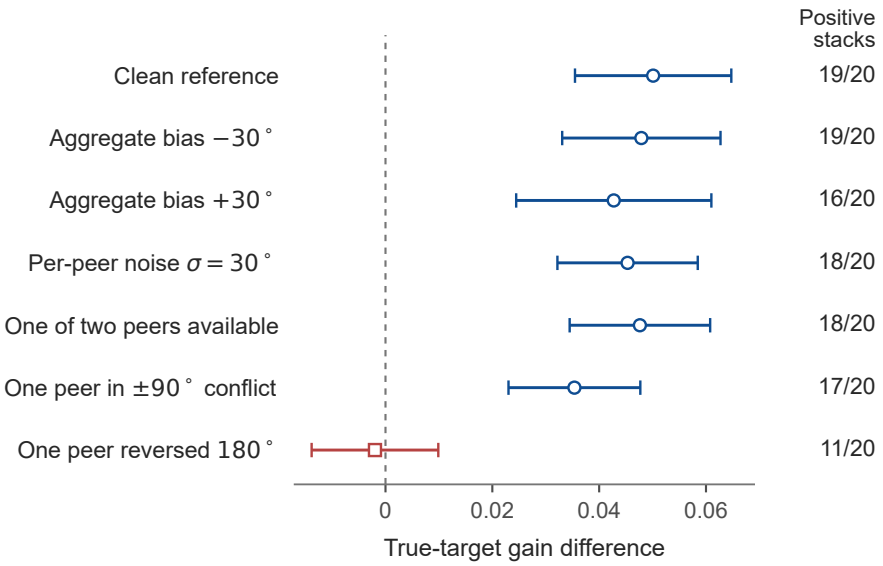

Figure 4: **Fixed-state sensitivity to reference perturbations.** Points are paired mean improvements over the direction-removed control in selected-pair one-step true-target gain; bars are two-sided 95% Student-t intervals over twenty Spread-4 stacks. All five moderate stress conditions pass the pre-specified joint criterion. The $180°$ reversal is a severe descriptive condition whose interval crosses zero. This is a fixed-state mechanism test rather than an episode-return experiment.

Table 8: Spread-4 reference-quality sensitivity at exact $B = 2$. Entries are paired improvements over direction-removed ranking with two-sided 95% Student-t intervals across twenty stacks. "Pos." counts positive stack differences. The first five stressed rows comprise the joint criterion; the 180° reversal and all four-step branch contrasts are descriptive.

Reference condition	One-step true-target gain	Pos.	$H = 4$ branch utility
Clean reference	0.05011 [0.03548, 0.06474]	19/20	0.001429 [0.000854, 0.002003]
Aggregate bias -30°	0.04790 [0.03308, 0.06272]	19/20	0.000831 [-0.000079, 0.001742]
Aggregate bias $+30^\circ$	0.04274 [0.02447, 0.06101]	16/20	0.000334 [-0.000449, 0.001116]
Per-peer noise, $\sigma = 30^\circ$	0.04532 [0.03219, 0.05846]	18/20	0.001425 [0.000862, 0.001989]
One of two peers available	0.04763 [0.03448, 0.06078]	18/20	0.001591 [0.000962, 0.002219]
One peer in $\pm 90^\circ$ conflict	0.03537 [0.02303, 0.04771]	17/20	0.001491 [0.000648, 0.002333]
One peer reversed 180°	-0.00197 [-0.01383, 0.00989]	11/20	0.000214 [-0.000829, 0.001256]

All five moderate conditions pass the joint criterion, retaining 70.6–95.6% of the clean-reference margin. Reversing one peer by 180° leaves no clear one-step advantage: the interval includes zero. For the secondary four-step endpoint, the clean, continuous-noise, single-peer, and $\pm 90^\circ$ conditions remain positive, whereas the two aggregate-rotation intervals include zero. We therefore use this experiment to characterize the direction mechanism, without extrapolating it to closed-loop return under corrupted references.

H Unified Exact-Traffic Evidence

The unified study places four ranking controls on one common exact-traffic stream. CA-RAPU, progress magnitude, unsigned receiver KL, action-distribution L2, and set-aware DeepSets share twenty frozen stacks, nine channel conditions, 25 paired episodes per condition, reset seeds, and complete edge–time channel-potential tables. Every arm force-fills two attempts at each of 64 steps, including the absorbing suffix after success, for 128 unit-cost attempts per trajectory. The four CA-RAPU-minus-control task-return contrasts form one Holm family. Equal attempt cost makes each paired task-return margin identical to its net-return margin; this design isolates ordering, not rate efficiency.

All five arms compute $w_e = r_e \,\mathrm{clip}(\bar{q}_{i,t}, 10^{-6}, 1)^{1.5}$ before applying the common score. Table 9 compares their information access and supervision.

Table 9: **Selector information and supervision.** Receiver queries are side-effect-free calls to the frozen actor. "Offline value" marks critic-derived selector targets; "critic at selection" marks critic access during scheduling. The first five selectors share frozen actor and head checkpoints, query and scoring implementations, paired resets, edge–time channel potentials, and exact $B = 2$ traffic. Once selections diverge, query outputs, confidence histories, deployment weights, buffers, and physical states remain method-specific.

Selector	Receiver query	Peer reference	Offline value	Critic at selection	Ranking statistic
CA-RAPU	yes	yes	no	no	signed reference-aligned response
Progress magnitude	yes	yes	no	no	sign-removed reference response
Unsigned receiver KL	yes	no	no	no	receiver action KL
Action-distribution L2	yes	no	no	no	$\lVert \pi_e^+ - \pi_j^0 \rVert_2$
DeepSets ranker	yes	no	yes	no	fitted set-aware $\widehat{\Delta Q}_e$
Joint-DirectQ	joint	no	yes	yes	best pair under fitted critic

H.1 Custom critic-supervised set ranker

DeepSets is fitted after the actor and communication interface have been frozen. Each candidate feature $x_e \in \mathbb{R}^{27}$ contains encoded payload and visible link metadata (11 dimensions), sender and receiver one-hot identities (six), the receiver's base action probabilities (five), and the candidate-

induced probability shift (five). The permutation-equivariant head is

$$h_e = \phi(x_e), \qquad \bar{h} = \frac{1}{6}\sum_{f=1}^{6}\phi(x_f), \qquad \widehat{\Delta Q}_e = \rho([h_e, \bar{h}]), \tag{37}$$

where ϕ is a 27–64–64 MLP and ρ is a 128–64–64–1 MLP, for 18,433 parameters. Targets are the mean of three frozen greedy-action critic differences between candidate-present and candidate-absent observations.

The label trajectories for this head and for the DirectQ critic below use the same fixed random-communication state generator: at each step, K is uniform on $\{0, 1, 2\}$ and a K-subset is then drawn uniformly from the candidate set. Training and validation use disjoint seed sets and are balanced over the nine channel profiles.

Each stack uses 2,048 training and 512 validation episodes disjoint from outcome evaluation, AdamW with learning rate 10^{-3} and weight decay 10^{-5}, and a validation-plateau rule. On the reserved labels, mean RMSE is 0.03443, same-decision pairwise ranking accuracy is 0.7383, exact top-two recovery is 0.463, and summed target regret is 0.01904 [0.01696, 0.02112]. These are offline prediction diagnostics, not environment-return guarantees. At outcome evaluation, the frozen head sees only its stated pre-channel features and is combined with the same reliability, pooled confidence, edge-cost term, traffic, and tie rule as the other arms.

Table 10: Unified exact-traffic paired contrasts. Intervals and tests use the twenty independently trained frozen stacks as inferential units. Three contrasts pass; the prespecified joint criterion of improvement over all four controls is not met.

Contrast	Paired task-return result
CA-RAPU − progress magnitude	0.0186 [0.0029, 0.0344], Holm-adjusted $p = 0.0454$, 16/20 positive
CA-RAPU − unsigned receiver KL	0.0277 [0.0153, 0.0401], Holm-adjusted $p = 6.59 \times 10^{-4}$, 17/20 positive
CA-RAPU − action-distribution L2	0.0137 [-0.0045, 0.0319], Holm-adjusted $p = 0.132$, 14/20 positive
CA-RAPU − set-aware DeepSets	0.0287 [0.0116, 0.0458], Holm-adjusted $p = 0.0070$, 16/20 positive

Table 10 reports the four paired contrasts. The comparison holds attempted traffic fixed. Appendix I reports the separate communication-rate experiments.

H.2 Effect size and channel variation

Table 11 summarizes the distribution of paired stack effects rather than relying on significance alone. The unsigned-KL contrast is the most stable response-based comparison: its corrected paired d_z is 1.00, its median and interquartile range are positive, and its mean is positive in every channel cell. The L2 contrast is smaller and more variable; three of its nine cell means are negative, consistent with its overall interval crossing zero. Figure 5 shows these cell-level differences and their uncertainty.

Table 11: Distribution of CA-RAPU-minus-control task-return effects across twenty stacks. d_z is the small-sample-corrected paired standardized effect; the final column counts channel cells with a negative mean out of nine.

Control	Mean [95% CI]	Median [IQR]	d_z	Positive stacks	Negative cells
Progress magnitude	.0186 [.0029,.0344]	.0136 [.0047,.0436]	.53	16/20	2/9
Unsigned receiver KL	.0277 [.0153,.0401]	.0298 [.0134,.0352]	1.00	17/20	0/9
Action-distribution L2	.0137 [−.0045,.0319]	.0197 [−.0037,.0325]	.34	14/20	3/9
Set-aware DeepSets	.0287 [.0116,.0458]	.0234 [.0114,.0425]	.75	16/20	0/9

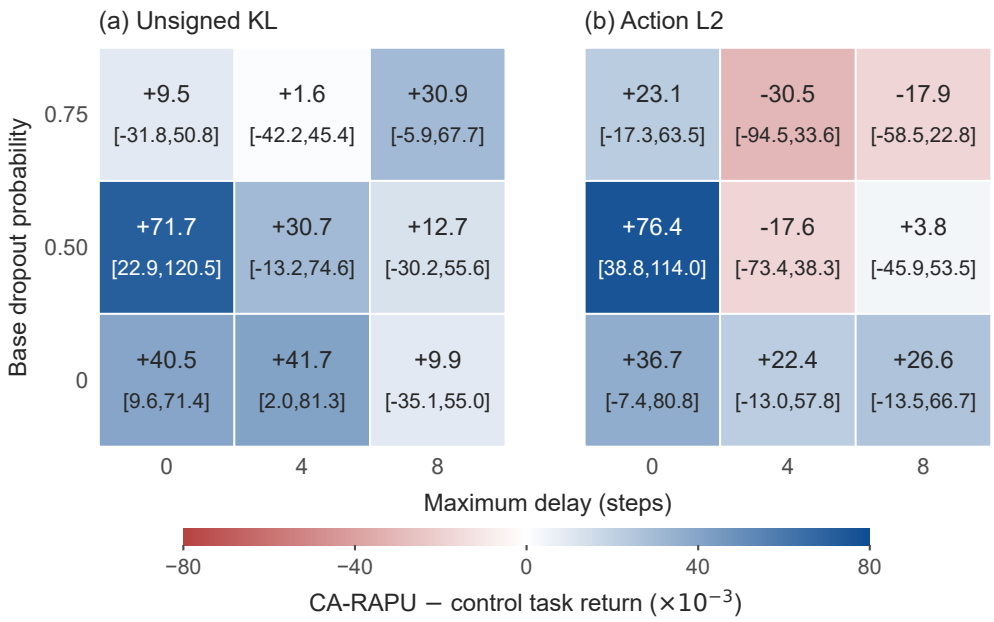

Figure 5: **Fixed-traffic margins across channel conditions.** Each cell gives the mean paired task-return difference and a pointwise 95% Student-t interval across twenty stacks. Positive values favor CA-RAPU. The CA-RAPU-minus-unsigned-KL mean is positive throughout the dropout–delay grid. Three action-L2 cell means are negative, but all three intervals include zero. All entries are in 10^{-3} return units. Cell intervals are descriptive and are not adjusted for multiple comparisons.

H.3 Secondary outcomes and episode-stream replication

The learned comparison also has a dedicated replication on a disjoint episode stream. Under the same $B = 2$, 64-step traffic protocol, CA-RAPU exceeds the set-aware DeepSets ranker by 0.03388 [0.02013, 0.04763], with positive differences on 17 of 20 stacks. This complements the Holm-4 estimate of 0.02868 [0.01160, 0.04577] without reusing its episode or channel-potential stream.

To interpret the composite task-return endpoint, we also analyze binary success and terminal mean distance on the same paired trajectories. This secondary analysis leaves the task-return comparison unchanged. CA-RAPU's mean success rate is 0.1529, compared with 0.1393 for progress magnitude, 0.1316 for unsigned KL, 0.1424 for action L2, and 0.1422 for DeepSets. Its success advantage over unsigned KL is 0.0213 [0.0118, 0.0308] ($p_{\text{Holm}} = 6.14 \times 10^{-4}$). The corresponding terminal-distance improvement is 0.00300 [0.00099, 0.00500] ($p_{\text{Holm}} = 0.0167$). Table 12 reports all four contrasts and shows that the secondary endpoints follow the same ordering most clearly against unsigned KL; intervals for several other controls include zero.

Table 12: Secondary outcomes on the paired unified exact-traffic trajectories. Success is CA-RAPU minus control; distance is control minus CA-RAPU, so positive values favor CA-RAPU in both columns. Intervals are pointwise 95% Student-t intervals; the reported adjusted p-values use separate four-comparison families for the two outcomes.

Control	Success difference	Terminal-distance improvement
Progress magnitude	.0136 [.0023,.0248]	.00239 [−.00048,.00527]
Unsigned receiver KL	.0213 [.0118,.0308]	.00300 [.00099,.00500]
Action-distribution L2	.0104 [−.0034,.0243]	.00152 [−.00173,.00478]
Set-aware DeepSets	.0107 [−.0016,.0230]	.00849 [.00512,.01187]

H.4 External learned and execution-time controls

Two additional studies compare CA-RAPU with recognizable allocation principles adapted to the frozen directed-message queue. Each study reuses the twenty frozen Spread-3 communication stacks but evaluates an independently sampled episode stream, separate from the unified study and from the other external control. Both comparisons use exact traffic and a single standalone primary contrast; neither is added to the four-control Holm family.

The *SchedNet-style frozen-queue learned scheduler* instantiates the learned-weight/top-k abstraction of SchedNet's deterministic Top(k) WSA variant (Kim et al., 2019). Full SchedNet jointly trains communication encodings, task actions, and communication weights, then uses the selected agents as broadcast senders. Our control instead operates on an already-trained directed-message stack. For each frozen stack, one edge-shared $13 \to 64 \to 64 \to 1$ ReLU network with 5,121 parameters scores all six directed candidates. Its input contains the two-dimensional frozen sender payload, reported quality, effective dropout probability, expected delay divided by eight, unit edge cost, normalized decision time, and sender and receiver one-hot identities. A separate scorer is fitted for each stack because independently trained codecs do not share an identified latent coordinate system. The scorer receives no peer reference, progress value, confidence-head output, receiver counterfactual action probabilities, critic value, Q-value, or oracle label. At evaluation, candidates are ordered by decreasing learned priority and then by canonical edge index.

Training is on-policy and model-free. At each step, a Plackett–Luce policy samples an ordered pair without replacement from the six priorities, and REINFORCE optimizes potential-shaped return-to-go. The shaping term telescopes to terminal task return up to a reset-state constant. Adam runs for 48 fixed updates, each with two episodes from every one of the nine channel profiles, for 864 training episodes per stack. The final update is retained without model selection; a disjoint ten-episode-per-profile validation stream is diagnostic only. Training, validation, and formal evaluation use disjoint episode streams, and the actor, codec, confidence model, and candidate generator remain frozen. Across the twenty retained training traces, the mean change from the first to the final update is 0.0265 [-0.0486, 0.1017], and sampling entropy remains near 1.700. We therefore treat this experiment as a narrow test of learned priority plus top-k access, not as the paper's principal learned comparator; the set-aware DeepSets study above fills that role.

TMC trains temporally smooth messages and action-confidence estimates; a sender transmits when message change crosses a threshold or a timeout expires, and the receiver reuses an unexpired cached message (Zhang et al., 2020). Our *TMC-inspired age/payload-change control* uses only two observable clues—age since the last attempt and payload change—to form an exact-budget directed-edge ranking. It does not implement TMC's training losses, threshold gate, or full sender–receiver cache semantics. For each directed edge, the control stores the payload and step of its most recent attempt. An unseen edge enters the priority group with zero payload change; otherwise, age is the number of steps since the last attempt and payload change is the Euclidean distance between the current and most recently attempted two-dimensional payloads. Edges whose age exceeds four also enter the priority group. Priority edges are ordered by decreasing age and payload change, while the remaining edges are ordered by decreasing payload change and age; canonical edge index resolves exact ties. The first two edges are attempted, and their histories are updated before channel outcomes are drawn, so dropped and delivered attempts follow the same temporal rule. The control uses no abstention threshold, channel or cost modifier, reward, critic, privileged task state, or task-quality input.

Formal evaluation for each external control uses all twenty stacks, nine dropout–delay profiles, 25 episodes per stack and profile, and 64 communication steps. Both methods force-fill $B = 2$ messages at every step, including after task success, for 128 unit-cost attempts per trajectory. Within a comparison, the two arms share reset conditions and edge–time channel potentials but run in separate environments once their selections diverge. Episode differences are averaged within channel profile and then equally across profiles, leaving one task-return difference per stack. We report the paired mean, two-sided 95% Student-t interval, two-sided paired p-value, and number of positive stack differences in Table 13.

Table 13: Standalone exact-traffic comparisons with external controls. The two rows use separate held-out episode streams; p-values are unadjusted and do not enter the unified Holm-4 family.

Comparator	CA-RAPU	Comparator return	Difference [95% CI]	Positive	p
TMC-inspired	2.71288	2.64207	0.07081 [0.05317, 0.08844]	19/20	7.99×10^{-8}
SchedNet-style frozen-queue learned scheduler	2.68749	2.44623	0.24126 [0.21060, 0.27193]	20/20	1.05×10^{-12}

These comparisons evaluate queue-level adaptations of temporal and learned priority rules. The SchedNet-style learner shows little training improvement or entropy reduction, which limits the interpretation of its large return gap. DeepSets provides the principal learned-ranking comparison. Figure 6 summarizes the paired margins.

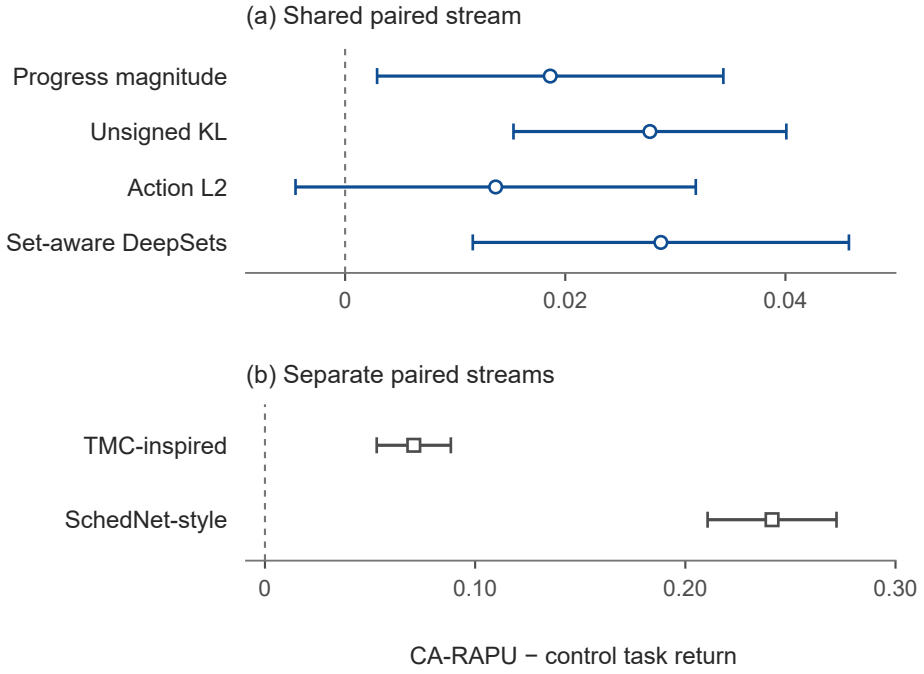

Figure 6: **Signed receiver response improves exact-traffic message selection.** Points are mean paired CA-RAPU-minus-control task-return margins; bars are 95% Student-t confidence intervals across twenty frozen stacks. Positive values favor CA-RAPU. (a) Four controls share one paired stream and a Holm-4 family; only the action-L2 interval crosses zero. (b) The TMC-inspired age/payload-change control and SchedNet-style learned-priority control use independent paired streams and standalone tests. They adapt specific allocation ideas to the frozen directed queue rather than reproduce the complete TMC or SchedNet systems. Every arm attempts exactly two messages per step, so task-return and net-return margins coincide within each comparison. Panels use different horizontal-axis ranges.

H.5 Centralized task-aware pair comparator

Joint-DirectQ is a task-aware value control on the same message queue. Each member of its frozen three-critic ensemble is a 128–128–1 MLP over the joint observation and a one-hot joint action. The target is $\mathbf{1}_{\text{success}} + 3(D_{\text{final}} - D_t)$ under the frozen actor's greedy joint action. Per stack, critic fitting uses 4,096 training and 1,024 validation episodes, AdamW with learning rate 10^{-3}, weight decay 10^{-5}, batch size 1,024, and gradient norm at most 1. Training uses a 20-epoch minimum and a patience of 20 epochs for a 0.25% validation-RMSE improvement. The nominal cap is 200 epochs; runs that have not met this stopping condition continue to a maximum of 400 epochs. Across the

twenty evaluation stacks, mean held-out critic RMSE is 0.3681 (stack SD 0.0290; range 0.3059–0.4211), and mean critic-member disagreement on reserved candidate queries is 0.0225. These diagnostics characterize the quality of its fitted value signal.

The ensemble observes the joint observation and evaluates the empty communication state, all six singleton insertions, and all 15 simultaneous message pairs. Let o_A be the actor-visible joint observation after inserting subset A, let $a_g(o_A)$ be the frozen actor's greedy joint action, and define

$$\Delta Q(A) = \frac{1}{3}\sum_{m=1}^{3}\left[Q_m(o_A, a_g(o_A)) - Q_m(o_\varnothing, a_g(o_\varnothing))\right]. \tag{38}$$

For pair $\{e, f\}$, its second-order critic residual is

$$I_{ef} = \Delta Q(\{e, f\}) - \Delta Q(\{e\}) - \Delta Q(\{f\}), \tag{39}$$

and Joint-DirectQ scores the pair as

$$S_{ef} = w_e \Delta Q(\{e\}) + w_f \Delta Q(\{f\}) + w_e w_f I_{ef} - 0.03(c_e + c_f). \tag{40}$$

This is the bi-affine interpolation induced by treating the deployment weights as endpoint delivery factors. Joint-DirectQ enumerates every pair and resolves exact score ties by the larger endpoint weight, the smaller endpoint weight, and then canonical pair order. It introduces no pair head, pair labels, fitted interaction model, or pair-specific coefficient. Singleton DirectQ uses the same $\Delta Q(\{e\})$, deployment weights, and cost term but ranks candidates independently and selects the top two.

The comparison uses twenty frozen stacks, nine channel conditions, and 25 paired episodes per condition. CA-RAPU, singleton DirectQ, and Joint-DirectQ make exactly two attempts at each of 64 steps and share reset seeds and edge-time channel potentials. The primary contrast is CA-RAPU minus Joint-DirectQ under a non-inferiority margin of -0.03, assessed by a one-sided 95% Student-t lower bound. This margin was chosen before outcome evaluation. It is 0.75% of the $[0, 4]$ return range and, on the continuous distance component of the metric, corresponds to a mean assigned-distance change of $0.03\sqrt{2}/3 = 0.0141$, less than one eighth of the 0.12 success radius. We additionally report the same lower-bound decision at thresholds 0, -0.01, and -0.02 as sensitivity analyses. CA-RAPU minus singleton DirectQ and Joint-DirectQ minus singleton DirectQ are secondary two-sided contrasts. Because both DirectQ variants use centralized task supervision unavailable to CA-RAPU, they serve as value-informed controls on the same message queue.

CA-RAPU attains mean task return 2.6935, compared with 2.6773 for Joint-DirectQ and 2.6773 for singleton DirectQ. All three arms make 128 attempts, so their paired task- and net-return margins coincide. The paired CA-RAPU–Joint-DirectQ difference is 0.01621 [-0.00365, 0.03607], with 14 of 20 stack-level differences positive. Its one-sided 95% lower bound is -0.00020, which exceeds the -0.03 non-inferiority margin ($p = 5.32 \times 10^{-5}$). The lower bound also exceeds -0.02 and -0.01, but not zero. The declared margin is met, while the corresponding two-sided contrast remains inconclusive ($p = 0.104$). CA-RAPU minus singleton DirectQ is 0.01623 [-0.00237, 0.03484] ($p = 0.0836$), while Joint-DirectQ minus singleton DirectQ is 0.00002 [-0.00736, 0.00740] ($p = 0.995$). These secondary contrasts are reported for context.

I Rate Adaptation across Communication Prices

The single-price calibrated rate-adaptation study is the primary deployment analysis. It uses its own development episode stream to freeze the CA-RAPU and native-scheduler thresholds, then evaluates those operating points on an independently seeded held-out stream under natural episode termination. The signed-versus-unsigned calibrated-abstention comparison and the five-price frontier are secondary operating-point and robustness analyses. Each uses a separate development stream, freezes its method-specific thresholds before formal evaluation, and runs on its own independently seeded held-out stream; the frontier freezes a distinct threshold for every method–price pair. Every contrast is paired only within its study stream. Absolute outcome levels are neither pooled nor compared across streams.

All adaptive calibrations search the same fixed set: all positive-signal candidates eligible; score thresholds $\{-0.02, 0, 0.005, 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64, 1.28, 2.56, 5.12\}$; and the si-

lence endpoint. Eligibility still requires the method's informative signal to be strictly positive. Development selection maximizes equal-cell, equal-stack net return; exact objective ties prefer fewer attempts and then the more abstaining threshold.

I.1 Single-price rate adaptation

The adaptive selector chooses up to two eligible candidates at each active step. Eligibility is determined by a frozen threshold selected on development stacks. The held-out study uses the same twenty frozen stacks on an independent paired stream. Threshold selection does not access held-out task outcomes.

At reporting price $\lambda_{\mathrm{comm}} = 0.005$, rate-adaptive CA-RAPU achieves mean task return 2.6160 with 95% interval $[2.5890, 2.6429]$, mean net return 2.4587 $[2.4311, 2.4863]$, mean success 0.1500 $[0.1337, 0.1663]$, and mean attempt count (and, under unit edge cost, attempted cost) 31.45 $[30.66, 32.24]$. The corresponding mean communication penalty is $0.1573 = 0.005 \times 31.45$. Table 14 reports the absolute operating points. The paired tests compare the adaptive selector with fixed-$B = 2$ CA-RAPU, the calibrated native metadata scheduler, and silence, with Holm adjustment across the three contrasts.

Table 14: Held-out single-price rate results. Means and intervals summarize twenty independently trained stacks. Intervals are two-sided 95% Student-t intervals.

Mode	Task return	Net return	Success	Attempts
CA-RAPU Adaptive	2.6160 [2.5890,2.6429]	2.4587 [2.4311,2.4863]	.1500 [.1337,.1663]	31.45 [30.66,32.24]
CA-RAPU Fixed $B = 2$	2.7195 [2.6903,2.7487]	2.1359 [2.1009,2.1708]	.1713 [.1532,.1895]	116.73 [115.49,117.97]
Native metadata Calibrated	2.2088 [2.1971,2.2205]	2.1865 [2.1704,2.2026]	.0122 [.0081,.0163]	4.47 [.20,8.74]
Native metadata Fixed $B = 2$	2.4540 [2.4324,2.4756]	1.8302 [1.8057,1.8546]	.0473 [.0375,.0572]	124.77 [124.04,125.49]
Silence	2.2018 [2.1932,2.2104]	2.2018 [2.1932,2.2104]	.0102 [.0075,.0129]	0

The adaptive-minus-fixed CA-RAPU net-return difference is 0.322840 $[0.296044, 0.349635]$, with 20/20 positive stacks and $p_{\mathrm{Holm}} = 1.36 \times 10^{-15}$. Against the calibrated native metadata scheduler, the difference is 0.272200 $[0.240153, 0.304248]$, with 20/20 positive and $p_{\mathrm{Holm}} = 5.36 \times 10^{-13}$. Against silence it is 0.256869 $[0.224988, 0.288750]$, again with 20/20 positive and $p_{\mathrm{Holm}} = 6.90 \times 10^{-13}$.

These results describe rate selection at one price under natural episode termination.

I.2 Direction under calibrated abstention

A secondary study gives each direction signal its own development-selected abstention threshold. Methods share the actor, codec, confidence model, candidate-generation and channel rules, and cost accounting.

The outcome is net return at the selected rate. At $\lambda_{\mathrm{comm}} = 0.005$, CA-RAPU obtains mean net return 2.434. Its paired margin is 0.139 [0.118, 0.160] over progress magnitude and 0.132 [0.112, 0.153] over unsigned receiver KL, with all twenty stack differences positive in both comparisons. CA-RAPU attempts 0.813 messages per decision, compared with 0.651 and 0.631 for the controls. Its additional task benefit more than offsets the cost of the higher rate.

I.3 Multi-price frontier

The multi-price study evaluates five fixed reporting prices. For each method and price, one threshold is selected on development stacks from the same fixed grid and then frozen. Held-out task outcomes are generated once and are repriced exactly with the common per-attempt cost.

CA-RAPU has the highest mean net return at all five reporting prices. The family contains two control contrasts at five prices and is adjusted jointly with Holm's method. The clearest familywise-supported separation occurs in the priced operating regimes; the evidence does not imply uniform significance at every individual price. Tables 15 and 16 report the absolute frontier and its paired contrasts.

The score construction and candidate inventory remain fixed across prices; each method–price pair uses its own development-selected threshold.

Table 15: Held-out multi-price net return; parentheses give mean attempts.

λ_{comm}	CA-RAPU	Action L2	Unsigned KL
0	2.6989 (60.43)	2.6864 (118.18)	2.6909 (110.14)
.0025	2.5478 (60.43)	2.4610 (35.30)	2.4432 (89.23)
.005	2.3968 (60.43)	2.3728 (35.30)	2.3060 (36.74)
.01	2.3165 (23.25)	2.2540 (8.74)	2.2283 (4.14)
.02	2.2213 (5.84)	2.2088 (.96)	2.2043 (.70)

Table 16: CA-RAPU minus control on the multi-price frontier. Intervals are pointwise paired 95% Student-t intervals across twenty stacks. The ten contrasts form one Holm family for the adjusted p-values.

λ_{comm}	Control	Difference [95% CI]	Positive	Holm p
0	Action L2	.01250 [−.00288,.02787]	14/20	.2104
0	Unsigned KL	.00798 [−.01315,.02911]	12/20	.4389
.0025	Action L2	.08684 [.06492,.10876]	20/20	6.85×10^{-7}
.0025	Unsigned KL	.10467 [.08462,.12473]	20/20	1.24×10^{-8}
.005	Action L2	.02402 [.00058,.04745]	12/20	.1805
.005	Unsigned KL	.09081 [.06751,.11412]	19/20	7.55×10^{-7}
.01	Action L2	.06248 [.04780,.07716]	20/20	2.78×10^{-7}
.01	Unsigned KL	.08815 [.06751,.10878]	20/20	2.78×10^{-7}
.02	Action L2	.01252 [−.00014,.02518]	13/20	.1805
.02	Unsigned KL	.01702 [.00665,.02739]	15/20	.01386

J Score Ablations and Local Sensitivity

This study reuses the twenty frozen Spread-3 stacks on an episode stream disjoint from every preceding evaluation. It provides closed-loop component evidence within the existing cohort; model training is unchanged. Each stack is evaluated on 25 episodes in each of the nine channel cells and under two heterogeneous edge-cost assignments. Both assignments use the multiset $\{0.5, 0.5, 1, 1, 1.5, 1.5\}$ and therefore have the same mean cost:

	0→1	0→2	1→0	1→2	2→0	2→1
receiver-tiered	1	1.5	0.5	1.5	0.5	1
sender-tiered	0.5	0.5	1	1	1.5	1.5

(41)

Episode seeds exclude the arm and cost-profile identities, so all arms and both cost assignments share the reset and canonical edge–time channel potentials.

Communication is enabled only at decisions $0, \ldots, 15$, with two nominal selections per decision; the physical task may continue through decision 63. Before success, selected edges are real attempts and incur their declared cost whether delivered or dropped. After success, the physical state is absorbing and each remaining communication slot records the arm-independent pair $(0, 1)$ only for accounting. These entries are neither transmitted nor billed. Thus every trajectory has exactly 32 nominal selections, while its billable attempt count and cost can be smaller.

The full score is compared with five one-component interventions: arithmetic rather than geometric conjunction; magnitude-matched removal of the progress sign; removal of the explicit confidence

multiplier ($\alpha = 0$); unit explicit reliability; and removal of the score-side cost term ($\lambda_{\text{score}} = 0$). Each intervention changes exactly one formula field. Under the confidence intervention, pooled confidence still contributes to the sender-excluded peer reference; only its explicit multiplicative score factor is removed. Under the reliability intervention, channel metadata remains available to the frozen confidence head.

Every arm owns an independent environment, buffer, pending-delivery queue, payload history, confidence pool, and physical state. Only reset identity, canonical edges, costs, and channel potentials are shared. The study therefore preserves closed-loop consequences instead of replaying candidate or receiver-query values across arms.

The primary endpoint is $R_{\text{task}} - 0.005C_{\text{billable}}$. Episodes are first averaged within stack–cost-profile–channel–arm cells, the nine channel cells are weighted equally, and the two cost profiles are then weighted equally. The five full-minus-control contrasts use the twenty stack-level differences, two-sided 95% Student-t intervals, and one Holm correction over exactly five p-values.

Table 17: One-component interventions under heterogeneous edge costs. The estimand is complete CA-RAPU minus the indicated intervention in stack-averaged net return; positive values favor the complete score. Intervals are paired 95% Student-t intervals, and p-values are Holm-adjusted over all five rows.

Intervention	Only change from the complete score	Net-return difference [95% CI]	Positive	Holm p
Arithmetic conjunction	$\sqrt{\widehat{g}_e \widehat{u}_e} \to (\widehat{g}_e + \widehat{u}_e)/2$	0.01705 [0.00555, 0.02856]	15/20	0.01173
Sign-removed, magnitude-matched progress	Negative raw progress is reflected under the same positive-progress RMS; positive-progress candidates are unchanged	0.02244 [0.01511, 0.02976]	19/20	1.51×10^{-5}
No explicit confidence multiplier	$\overline{q}^{\alpha}_{i,t} \to 1$; confidence remains in the peer reference	0.01075 [0.00110, 0.02041]	15/20	0.03083
No explicit reliability multiplier	$r_e \to 1$; channel metadata remains available to the frozen confidence head	0.01341 [0.00579, 0.02102]	15/20	0.00470
No score-side cost	$\lambda_{\text{score}} \to 0$; heterogeneous attempted cost remains in net return	0.04835 [0.03745, 0.05926]	20/20	8.63×10^{-8}

All five interventions reduce mean net return in this Spread-3 study (Table 17). The magnitude-matched comparison isolates direction by changing only the treatment of negative progress. Removing cost from ranking produces the largest loss under the two heterogeneous cost assignments. The other rows measure the contributions of geometric combination and the explicit confidence and reliability multipliers.

A separate descriptive grid varies one factor at a time: $\alpha \in \{1, 1.5, 2\}$, $\eta \in \{0.125, 0.25, 0.5\}$, and $\lambda_{\text{score}} \in \{0.015, 0.03, 0.06\}$. The default configuration is shared rather than rerun. The six nondefault variant-minus-default differences use Bonferroni intervals with 95% family coverage. For each factor, local regret is the distribution across stacks of the larger $\max\{0, R_{\text{default}} - R_{\text{variant}}\}$ over its two nondefault levels. Table 18 reports these descriptive comparisons, which do not enter the confirmatory Holm family.

Table 18: Descriptive one-factor sensitivity around the reported default. The last three columns summarize the stack-level worst regret over the two nondefault levels and therefore appear once per factor. Intervals provide 95% family coverage over the six variant-minus-default estimates.

Factor	Level	Net-return difference [simultaneous 95% CI]	Worst regret		
			Mean	Median	Maximum
α	1.0	-0.003698 [-0.011585, 0.004188]	0.011349	0.011816	0.028469
	2.0	-0.004594 [-0.012532, 0.003345]			
η	0.125	-0.005798 [-0.011257, -0.000339]	0.011523	0.009991	0.033421
	0.5	-0.004454 [-0.013762, 0.004853]			
λ_{score}	0.015	-0.005731 [-0.013025, 0.001562]	0.012993	0.015624	0.028771
	0.06	-0.005441 [-0.013671, 0.002789]			

All six local-sensitivity estimates lie between -0.0058 and -0.0037 net-return units; only the $\eta = 0.125$ simultaneous interval excludes zero. The factor-wise mean worst regret ranges from 0.01135 to 0.01299.

K Peer-Support Topology at Fixed Inventory Size

We vary peer-support topology without changing the number of eligible messages. The study uses twenty independently trained four-agent stacks and 480 states drawn from no-message rollouts. At every state, two exposure-balanced families of 24 masks retain exactly eight of the twelve directed candidates. The families have identical marginal edge exposure, sender-outdegree distributions, and reciprocal-dyad counts, but different receiver indegrees. The uniform family has topology $2+2+2+2$, so every candidate has one sender-excluded peer for its receiver. The mixed family has topology $3 + 3 + 1 + 1$: candidates addressed to the two degree-one receivers have no such peer, whereas those addressed to the degree-three receivers have two. Both selectors recompute their references, normalizers, confidence histories, and scores from the masked queue before force-filling $B \in \{1, 2\}$ messages.

Selection is evaluated with a four-step branch return. For each state, the no-message, singleton, and pair branches share their continuation randomness; the branch outcomes are unavailable to either selector. Results are averaged over masks, channel profiles, and states within each trained stack, which is the inferential unit.

Figure 2 in the main text plots the paired CA-RAPU-minus-KL four-step set utilities for the two support patterns.

Under uniform support, CA-RAPU exceeds unsigned KL by 0.00106 [0.00052, 0.00160] for $B = 1$ and 0.00128 [0.00075, 0.00181] for $B = 2$. Under mixed support, the corresponding margins are -0.00200 [-0.00263, -0.00138] and -0.00661 [-0.00762, -0.00561]. The mean-over-capacity topology interaction is 0.00548 [0.00490, 0.00606] (Holm-adjusted $p = 1.11 \times 10^{-13}$), with the same sign on all twenty stacks. The interactions at $B = 1$ and $B = 2$ are 0.00307 [0.00263, 0.00350] and 0.00789 [0.00695, 0.00884], respectively. These three interaction tests form one Holm-adjusted family. Holding $M = 8$ rules out candidate count alone as an explanation for the topology contrast: reorganizing receiver-matched support changes whether the directional reference improves selection. The intervention shows that the comparison can reverse as support topology changes; it does not attribute the effect to zero-peer edges alone because the mixed topology also redistributes one-peer support into two-peer support.

L TRANSFER TO COMMNET PREDATOR–PREY

We evaluated six independently trained CommNet Predator–Prey policies. Each was trained from scratch with full communication for 20,000 updates. All six met the initial task-competence and full-versus-silence criteria. We then evaluated seven schedulers on each final checkpoint, using 256 episodes per scheduler and policy (10,752 episodes total).

The primary outcome was CA-RAPU minus unsigned receiver KL in native task return at per-decision capacity $B = 5$. A secondary comparison used threshold 0.64 for both methods and attempt price 0.005. Before evaluation, progression to a twenty-policy study required a positive mean difference and at least five positive policies out of six for each outcome. Fixed-capacity return averaged -0.008548, with one positive policy, so the study stopped after this stage. Adaptive net return averaged $+0.020007$, with all six differences positive. Table 19 reports all six policy differences. Exploratory paired 95% Student-t intervals across policies are -0.00855 [-0.01831, 0.00122] for fixed-capacity return and 0.02001 [0.00280, 0.03721] for adaptive net return. These intervals summarize the six reported differences; the prespecified qualification rule still determines progression to the larger study.

Table 19: CommNet results for all six policies, CA-RAPU minus unsigned receiver KL. Adaptive native return and saved attempt cost sum to the adaptive net margin, up to rounding. Fixed capacity is five messages per decision.

Policy	Fixed native	Adaptive native	Saved cost	Adaptive net
1	-0.010938	+0.005313	+0.007676	+0.012988
2	-0.010039	+0.019805	+0.011074	+0.030879
3	-0.008594	+0.021797	+0.005781	+0.027578
4	-0.020469	+0.000000	+0.001074	+0.001074
5	+0.008242	+0.003437	+0.001191	+0.004629
6	-0.009492	+0.021641	+0.021250	+0.042891
Mean	-0.008548	$+0.011999$	$+0.008008$	$+0.020007$

The adaptive net-return margin comprises a native-return gain of $+0.011999$ and saved attempt cost of $+0.008008$. Native return is higher on five policies and tied on one; mean attempts are lower on all six. The five negative fixed-capacity contrasts also have longer CA-RAPU episodes. Because episodes terminate naturally, equal per-decision capacity can produce different total traffic. This adapter therefore shows different behavior under fixed and adaptive communication; the current data leave its mechanism unresolved.

M COMPUTATIONAL COST AND EVALUATION DETAILS

M.1 INFERENCE PATH AND RUNTIME

In the simulated cooperative tasks, the scheduler reads the current queue and pre-transmission link descriptors, queries the frozen receiver, constructs the task reference, and applies stable Top-B selection. Its use requires access to queued payloads and a representative receiver query. Reference quality, confidence calibration, and channel estimates can change after deployment. Model parameters remain fixed, and channel outcomes are observed only after selection.

Let C_π denote the cost of one receiver-policy evaluation and E the current candidate inventory. Within one decision, the candidate-absent distribution depends only on the receiver and can be cached after n rows; candidate-present inputs are edge-specific but batch over $|E|$ rows. For the Spread coordinate adapter, the reference is algebraic and policy-query work is $O((n+|E|)C_\pi)$. For a complete directed graph, $|E| = n(n-1)$. Given the reference vectors, score algebra is $O(|E|)$ and stable top-B maintenance is $O(|E| \log B)$, or linear for fixed B.

The formal evaluations used a direct reference routine that scans the queue for each focal candidate. We also implement a cached reference for the additive adapter used here. It first aggregates evidence by receiver and sender, then reuses leave-one-sender-out statistics; its reference stage costs $O(|E| + \sum_j k_j \log k_j)$ for k_j represented senders at receiver j, or $O(|E| \log n)$ on a complete directed graph.

Before timing, the cached and formal routines produced exactly matching Top-2 selections and zero observed score difference on deterministic inventories from 6 to 380 candidates.

Table 20 reports query-inclusive latency in the six-edge evaluation setting. Table 21 isolates scorer scaling by holding receiver-query outputs fixed and varying a deterministic synthetic complete directed inventory.

Table 20: Query-inclusive CPU latency in milliseconds per scheduling decision for the six-edge, $B = 2$ evaluation setting.

Method	Median	p95
CA-RAPU, cached	3.872	3.887
CA-RAPU, direct scan	4.379	4.398
Unsigned receiver KL	2.749	2.757
Set-aware DeepSets	3.432	3.439

Table 21: CPU scorer-only scaling after a precomputed receiver query. Inventories are synthetic complete directed graphs; values are median milliseconds per decision. The final column is direct divided by cached latency.

Candidates	Direct scan	Cached reference	Speedup
6	2.087	1.571	1.33×
12	3.837	2.156	1.78×
30	9.279	3.128	2.97×
90	32.083	7.483	4.29×
210	98.631	14.741	6.69×
380	232.297	26.139	8.89×

Measurements use Intel Xeon Platinum 8260 CPUs on Linux x86-64, 48 PyTorch CPU threads, four warm-up calls, and 15 samples of four calls. Medians and p95 values summarize per-call sample means. Query-inclusive timing covers feature construction, decoding, actor queries, scoring, and selection. It excludes model loading, candidate generation and encoding, environment steps, transport, and movement selection. The scaling experiment fixes query outputs and varies a synthetic inventory.

M.2 Paired evaluation

Each comparison pairs methods on the same trained stack, initial states, and potential channel outcomes. Inference aggregates episodes and channel conditions within stack. Episode-stream replications reuse the trained stacks with distinct evaluation episodes; evaluation seeds are disjoint from training.

Fixed-action replays across the two Spread evaluation environments matched reset observations, world states, transitions, rewards, and termination flags over 32 steps for each of four seeds, supporting comparable dynamics on those trajectories.

M.3 Statistical reporting

The inferential unit is an independently trained frozen stack or policy; episodes and channel cells are repeated measurements within that unit. Unless specified otherwise, paired contrasts use two-sided Student-t intervals after within-unit aggregation and Holm adjustment within each comparison family. Joint-DirectQ instead uses a one-sided lower bound for non-inferiority. The choice-set–capacity analysis also reports an exact sign-flip test for its primary interaction.