NOT ALL COUPLING MATTERS: VALUE-INDUCED BELIEF EDGES FOR BUDGETED COORDINATION IN COOPERATIVE MARL

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse coordination lets cooperative multi-agent reinforcement learning (MARL) scale by giving each agent a limited budget of teammate interactions. Under a fixed budget, the question is which teammates deserve that capacity. Existing edge-selection criteria mainly measure how strongly a teammate affects payoff and do not account for whether reducing uncertainty about that teammate would change the receiver's decision. The result is certainty-blind coordination: capacity goes to strongly coupled but predictable teammates, while uncertain, decision-pivotal ones are overlooked. We introduce Value-Induced Belief Edges (VIBE), which scores an edge by the expected value of resolving uncertainty about a teammate: given its belief over teammate actions, the receiver values evidence that would change its preferred action and improve its expected return. This decision value decomposes into the probability that the receiver switches actions and the gain from switching, and payoff variance can rank teammates in the opposite order. A centralised teacher approximates the decision value over a bounded set of joint-action probes, and a decentralised student learns the resulting rankings from receiver-local information. Selected edges direct relational processing of teammate features the receiver already observes, with no additional communication at execution. Extensive experiments on five benchmark families show that VIBE outperforms strong value-based, partner selection and coordination-graphs baselines. Our code is anonymously available at https://anonymous.4open.science/r/VIBE-D717

1 INTRODUCTION

Cooperative MARL must determine not only how agents act, but which teammate relationships deserve limited modelling capacity. Value factorisation combines agent utilities for team-value learning (Hu et al., 2026), while coordination graphs explicitly represent interactions between agents (Zhang et al., 2025). We study a fixed budget of K teammate-specific edges within a decentralised controller. Each edge processes teammate features already available to the receiver, so this budget limits relational computation rather than communication. With the architecture and budget held fixed, what makes one teammate more useful to model than another?

Consider an agent choosing between advancing and staying in cover. A sensitivity-only rule can allocate capacity to a strongly coupled teammate whose next action is predictable from the receiver's history, overlooking another with a smaller payoff range whose uncertain provision of cover determines whether advancing is worthwhile. We call this *certainty-blind coordination*. Uncertainty alone is also insufficient: resolving it may leave the preferred action unchanged. What matters is whether resolving uncertainty changes the receiver's action and how much that change is worth.

Existing work already connects coordination structure to decisions (Duan et al., 2024b). CASEC relates payoff variance to changes in greedy action selection after removing an interaction (Wang et al., 2022), while SOP-CG learns dynamic graphs within classes supporting tractable action optimisation (Yang et al., 2022). However, our focus differs from these works: we study how evidence about an agent changes the receiver's decision under its current belief (Zhang et al., 2025; Yuan et al., 2022). I2C is particularly relevant, as it learns a local communication prior from centralised

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.

labels measuring KL changes between action distributions conditioned with and without a teammate's action (Ding et al., 2020). VIBE instead uses receiver-action value gain to supervise local relational processing. Its distinction is using decision value to direct representation.

Following value-of-information theory, let $V_i(b_i, a_i)$ be belief-conditioned expected team return, marginalising teammate behaviour and unobserved state or history under a common fixed continuation. Prior b_i^{-j} conditions on the full local action–observation history h_i^t, with Bayesian posterior b_i^{+j} from hypothetical evidence e_j about teammate j. Optimising only receiver actions over the same feasible set, define the ideal edge target $\Delta_{ij}^t = \mathbb{E} * e_j \mid h_i^t[V_i(b_i^{+j}, a * i, *^{+j}) - V_i(b_i^{+j}, a_{i,*}^{-j})]$, where $a_{i,*}^{\pm j} \in \arg\max_{a_i} V_i(b_i^{\pm j}, a_i)$. Same posterior evaluation isolates decision adaptation which nonnegative regret vanishes without switching, making Δ_{ij}^t exactly switching probability times conditional value gain.

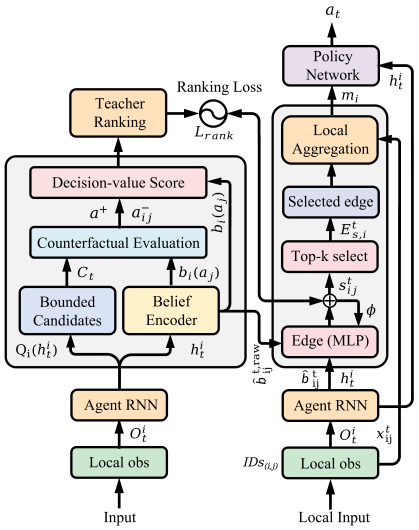

Figure 1: The VIBE model architecture.

VIBE allocates limited local relational capacity using this decision value, as shown in **Figure 1**. Exact belief marginalisation can be exponential despite local action optimisation. A centralised teacher produces the edge-ranking surrogate $\widehat{\Delta}_{ij}^t$ via paired, belief-conditioned critic evaluations on a bounded, recogniser-filtered set of potentially optimal joint-action probes. After belief calibration, a decentralised edge student learns these rankings from receiver-local history, observed teammate features and pairwise action beliefs. Edges route already-observed features through the controller's shared relational pathway, reallocating modelling effort, not information access. This yields an inductive-bias hypothesis: prioritising decision-pivotal agents can improve learning efficiency without additional inter-agent communication.

In summary, the contributions of this work are as follows:

- We identify certainty-blind coordination and formulate receiver-specific, belief-conditioned decision regret for edge allocation, separating the value of resolving teammate uncertainty from payoff sensitivity and explaining how their rankings can differ.
- Decision-value-guided allocation of relational capacity uses privileged supervision to direct representational emphasis. VIBE instantiates this principle through bounded-candidate teacher scoring and receiver-local ranking distillation, without additional communication.
- We conduct extensive experiments across five benchmark families. We evaluate coordination, demonstrating that VIBE better identifies decision-pivotal teammates and improves performance over strong baselines without additional inter-agent communication.

2 Preliminaries

We consider a cooperative Dec-POMDP with agents $\mathcal{N} = [n]$, states $\mathcal{S}$, action sets $\mathcal{A}_i$, a shared reward, and discount γ. Agent i observes o_i^t, maintains its action–observation history h_i^t, and selects a_i^t simultaneously with its teammates. The backbone combines individual utilities Q_i through a monotonic mixer Q_{tot}. For a fixed continuation policy $\boldsymbol{\pi}$, let $Q^{\boldsymbol{\pi}}(s, \mathbf{h}, \mathbf{a})$ denote the expected return after taking $\mathbf{a}$ from state s and joint history $\mathbf{h}$. The history argument accounts for recurrent policies.

2.1 Budgeted relational modelling

The receiver's observation contains teammate feature blocks $o_{i\to j}^t$. A local selector retains at most K teammates, $E_i^t = \sigma(h_i^t) \subseteq \mathcal{N} \setminus \{i\}$, $|E_i^t| \le K$, whose features enter a dedicated relational channel of the controller (Fan et al., 2025). We seek $\max_{\theta,\sigma} J(\boldsymbol{\pi}_{\theta,\sigma})$ subject to E_i^t, where J is the expected discounted return and θ denotes the controller parameters. The selector changes both the policy at

fixed parameters and the data encountered during learning (Zhang et al., 2025). Since the selected features already belong to o_i^t, an edge allocates computation rather than acquiring information.

2.2 Decision value of evidence

Let b_i^t be the receiver's belief over $\xi = (s^t, \mathbf{h}_{-i}^t, \mathbf{a}_{-i}^t)$ conditional on h_i^t. Its action value is $V_i(b, a_i) = \mathbb{E}_{\xi\sim b}[Q^{\boldsymbol{\pi}}(s, \mathbf{h}, (a_i, \mathbf{a}_{-i}))]$. Only a_i is optimised; teammates' actions and unobserved context are integrated under the receiver's belief. Let e_j be hypothetical evidence about teammate j, with $e_j = a_j^t$ in the action-revelation decision problem. Write $b_i^{-j} = b_i^t$ and $b_i^{+j} = p(\xi \mid h_i^t, e_j)$.

$$a_{i,*}^{-j} \in \arg\max_{a_i} V_i(b_i^{-j}, a_i),\ a_{i,*}^{+j} \in \arg\max_{a_i} V_i(b_i^{+j}, a_i), \tag{1}$$

Over the same feasible receiver actions, choose to retain the prior action whenever it remains posterior-optimal. The ideal target is

$$\Delta_{ij}^t = \mathbb{E}_{e_j|h_i^t}\left[V_i(b_i^{+j}, a_{i,*}^{+j}) - V_i(b_i^{+j}, a_{i,*}^{-j})\right]. \tag{2}$$

Both actions are evaluated under the same posterior, so Δ_{ij}^t measures the benefit of adapting the receiver's action. It is the expected value of sample information for this decision problem (Zhang et al., 2026). A payoff-sensitivity criterion can favour a strongly influential teammate even when its evidence has less decision value.

3 Value-Induced Belief Edges (VIBE)

VIBE uses belief-conditioned, joint-action probes to generate teammate rankings, distils these rankings into a selector, and routes the selected teammate features through a relational channel (as shown in **Figure 2**). We denote practical teacher score by $\widehat{\Delta}$ and the ideal decision target in **Eq. 2** by Δ.

3.1 Pairwise Belief Encoder

Let $r_i^t \in \mathbb{R}^{d_h}$ denote the receiver's recurrent representation available before action selection. A shared encoder predicts a categorical teammate-action distribution, $b_{ij} = \mathrm{softmax}(g_\psi([r_i^t, \mathbf{1}_i, \mathbf{1}_j]))$, where $\mathbf{1}_i, \mathbf{1}_j \in \{0,1\}^n$ identify the ordered pair. The MLP has widths $(d_h + 2n) \to 64 \to 64 \to |\mathcal{A}_j|$, and is trained by cross-entropy against recorded teammate-action labels.

The teacher renormalises this distribution over the teammate's available actions. Writing $\mathcal{A}_j^{\mathrm{avail}}(t)$ for that training-time mask,

$$b_{ij}^{\mathrm{T}}(a) = \frac{b_{ij}(a)\,\mathbb{I}[a \in \mathcal{A}_j^{\mathrm{avail}}(t)]}{\sum_{u\in\mathcal{A}_j^{\mathrm{avail}}(t)} b_{ij}(u)}. \tag{3}$$

The student receives b_{ij}, the unmasked softmax. The availability mask is privileged, so only the teacher uses it. The encoder provides pairwise marginals.

3.2 Joint-Action Probes

The recogniser for potentially optimal joint actions (Huang et al., 2026) supplies a bounded probe set $\mathcal{C}_t$. Each agent contributes its $k = 2$ highest-utility actions. A bounded beam assembles at most $P = 32$ joint proposals, which are retained when $Q_r(\mathbf{a}) \geq Q_r(\mathbf{a}_{\mathrm{greedy}}) - C$, where $\mathbf{a}_{\mathrm{greedy}}$ is the decentralised greedy joint action and $C = 0.1$. At most $M = 8$ retained proposals are used. The set is rebuilt for each transition and shared across ordered pairs.

3.3 Paired Teacher Scoring

For each ordered pair (i, j), the teacher evaluates the same nonempty probe set using the centralised critic $\hat{Q}$, with its transition context held fixed. Pair indices are suppressed on the two critic branches. The *with-evidence* branch $\hat{Q}^+(\mathbf{a})$ uses teammate j's action from the probe. The *without-evidence* branch $\hat{Q}^-(\mathbf{a})$ replaces that action embedding by $\bar{\mathbf{e}}_{ij} = \sum_{a_j\in\mathcal{A}_j} b_{ij}^{\mathrm{T}}(a_j)\,\mathrm{emb}(a_j)$.

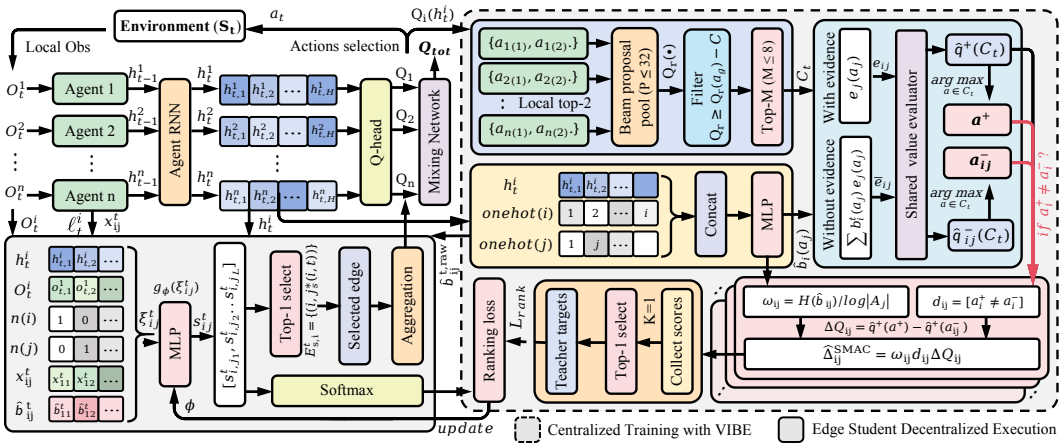

Figure 2: The VIBE architecture allocates modelling capacity by the value of resolving uncertainty, not payoff sensitivity. Using receiver-local action beliefs, a shared centralized evaluator compares bounded POW candidates under original versus exactly belief-averaged embeddings. Receiver-action switching gates entropy-weighted gains into plug-in scores. Ranking distillation trains a receiver-local Edge Student whose top-K selections gate local ally-feature aggregation for decentralized actions, requiring no additional inter-agent communication channel at execution.

The categorical embedding expectation is computed exactly. Let

$$\hat{\mathbf{a}}^{\pm} \in \arg\max_{\mathbf{a}\in\mathcal{C}_t} \hat{Q}^{\pm}(\mathbf{a}),\ \omega_{ij} = \frac{H(b_{ij}^{\mathrm{T}})}{\log|\mathcal{A}_j|}, \tag{4}$$

with $\hat{a}_i^{\pm}$ denoting the receiver components and H Shannon entropy. The score is

$$\widehat{\Delta}_{ij}^{t} = \omega_{ij}\, \mathbb{I}[\hat{a}_i^{+} \neq \hat{a}_i^{-}] \left[\hat{Q}^{+}(\hat{\mathbf{a}}^{+}) - \hat{Q}^{+}(\hat{\mathbf{a}}^{-})\right]. \tag{5}$$

The gate detects a change in the receiver component. The value gap measures how the probes differ under $\hat{Q}^{+}$. The entropy weight reduces the score of predictable teammates. The gap is nonnegative and consequently vanishes when the receiver does not switch or the teacher belief is deterministic.

Eq. 5 is a ranking surrogate. It evaluates a belief-mean embedding through the critic, optimises joint probes rather than marginalising teammate actions, and applies an entropy weight. The probes are selected by estimated optimality. These design choices distinguish $\widehat{\Delta}$ from the posterior decision value Δ. Joint-probe selection can make the score sensitive to receiver-independent payoff offsets.

For each receiver, define the teacher's preferred teammate and ranking weight by

$$j_i^{\star} \in \arg\max_{j\neq i} \widehat{\Delta}_{ij}^{t},\ w_i = \max_{j\neq i} \widehat{\Delta}_{ij}^{t} - \frac{1}{n-1}\sum_{j\neq i} \widehat{\Delta}_{ij}^{t}. \tag{6}$$

All ordered pairs require $\mathcal{O}(n^2 M)$ critic evaluations per transition. Teacher computation is chunked over the batch and used only during training.

3.4 Local Ranking and Relational Processing

A shared student assigns $s_{ij}^t = f_\eta(x_{ij}^t)$ from $x_{ij}^t = [r_i^t, o_i^t, \mathbf{1}_i, \mathbf{1}_j, o_{i\to j}^t, b_{ij}]$. These inputs are receiver-local; global state and teammate hidden states are excluded. The student learns the teacher's top-ranked teammate using weighted top-1 cross-entropy,

$$\mathcal{L}_{\mathrm{rank}} = -\sum_i \bar{w}_i \log \frac{\exp(s_{ij_i^\star}^t)}{\sum_{j\neq i}\exp(s_{ij}^t)}. \tag{7}$$

Rows with $w_i \leq 10^{-8}$ are discarded; the remaining weights are normalised and clipped at 10. The selector retains the K highest-scoring teammates. The evaluated configuration uses $K = 1$.

The controller aggregates the selected features as $m_i^t = \sum_{j \in E_i^t} \phi(o_{i \to j}^t)/\max(1, |E_i^t|)$, $\phi(x) = \mathrm{ReLU}(W_\phi x + b_\phi) \in \mathbb{R}^{64}$, and conditions on $[o_i^t, r_i^t, \alpha m_i^t]$. Setting $\alpha = 0$ disables the channel for frozen-policy ranking evaluation; end-to-end training uses $\alpha > 0$. The payload is extracted from the receiver's observation. The selected edge provides dedicated processing for existing features, changing representational emphasis while retaining the mixer and temporal-difference objective.

3.5 Training and Evaluation Protocol

After training the value-factorisation backbone (including recogniser and labelling critic), we freeze it at a shared checkpoint and calibrate the belief encoder. With teacher and encoder also frozen, we distil the student and evaluate rankings on fixed trajectories ($\alpha = 0$), then train the graph-conditioned controller from that checkpoint ($\alpha > 0$).

Comparisons vary only edge selection, matching checkpoints, edge and training budgets, relational pathway and evaluation. Win-rate and return AUC over identical training windows assess selector-driven learning efficiency, not graph-presence effects.

4 Theoretical Analysis

We characterise the receiver decision target, give a strict separation from payoff variance, and establish greedy consistency of the local controller.

Lemma 1. *For Bayesian posteriors, finite action sets and integrable values,*

$$\begin{aligned} \Delta_{ij}^t &= \mathbb{E}_{e_j | h_i^t}\left[\max_{a_i} V_i(b_i^{+j}, a_i)\right] - \max_{a_i} V_i(b_i^{-j}, a_i) \\ &= \Pr(\mathcal{S}_{ij} \mid h_i^t)\mathbb{E}[G_{ij} \mid \mathcal{S}_{ij}, h_i^t], \end{aligned} \tag{8}$$

where $\mathcal{S}_{ij} = \{a_{i,*}^{+j} \neq a_{i,*}^{-j}\}$ and $G_{ij} = V_i(b_i^{+j}, a_{i,*}^{+j}) - V_i(b_i^{+j}, a_{i,*}^{-j})$. Under the rule, $G_{ij} > 0$ on $\mathcal{S}_{ij}$ and is zero otherwise. The product is defined as zero when switching has probability zero.

Thus uncertainty matters through the receiver's decision boundary: evidence must change the preferred action and make that change valuable. Payoff dispersion alone does not locate this boundary.

Proposition 1. *With known state, binary actions and independent teammates A, B, let $Q = q_{iA} + q_{iB}$, $(q_{iA}, q_{iB}) = (20a_A, 3.5 - 6a_B)\mathbb{I}[a_i = 0]$, and $(\Pr(a_A = 1), \Pr(a_B = 1)) = (0.05, 0.5)$. For (A, B), uniform-sweep payoff variances $\max_{a_i} \mathrm{Var} * a_j \sim \mathrm{Unif} q * ij$ are $(100, 9)$, belief-weighted variances of payoffs and action advantages $q_{ij}(0, a_j) - q_{ij}(1, a_j)$ both equal $(19, 9)$. Yet action revelation yields $\Delta_{iA} = 0 < \Delta_{iB} = 0.75$. The ranking inversion persists when bounded by 0.1.*

Both interactions are action-dependent, yet only one outcome of B crosses the decision boundary despite its smaller payoff variation.

Lemma 2. *Fix the local histories. If E_i^t and m_i^t are computed independently of the candidate action a_i, and the mixer is non-decreasing in each local utility, then*

$$\prod_i \arg\max_{a_i} Q_i(h_i^t, a_i; m_i^t) \subseteq \arg\max_{\mathbf{a}} Q_{\mathrm{tot}}(s^t, \mathbf{h}^t, \mathbf{a}). \tag{9}$$

The relational channel therefore preserves decentralised greedy execution. For $K > 1$, selecting the largest individual student scores maximises their sum; this is the selector's additive objective, distinct from the joint value of observing several teammates.

5 Experiments

5.1 Experimental Settings

Benchmarks. We evaluate VIBE on 10 tasks from five benchmark families. **Certain-but-Coupled Trap (CCT)** is a controlled matrix game in which payoffs are coupled. **Predator–Prey** (Huang et al., 2026) at penalties $p = -4$ and $p = -5$ tests coordinated capture. **SMACv1**

Table 1: Decision-relevant coordination across five benchmark families.

Method	Matrix	Predator–Prey		SMACv1			SMACv2			Driving
	CCT	p=-4	p=-5	3s_vs_5z	Corr.	10m_vs_11m	Protoss	Terran	Zerg	Intersect.
Value factorisation and joint-action modelling										
QTRAN	$13.33_{\pm 0.31}$	$3.37_{\pm 2.71}$	–	$89.62_{\pm 8.06}$	–	–	$15.10_{\pm 0.85}$	$15.25_{\pm 0.99}$	$14.20_{\pm 0.73}$	$43.62_{\pm 5.68}$
QPLEX	$15.81_{\pm 0.48}$	–	–	$96.16_{\pm 2.80}$	$43.10_{\pm 7.00}$	$38.12_{\pm 4.73}$	$16.78_{\pm 1.03}$	$16.46_{\pm 0.78}$	$14.81_{\pm 1.15}$	$25.77_{\pm 5.42}$
OW-QMIX	$15.13_{\pm 0.30}$	$8.16_{\pm 2.81}$	–	$100.00_{\pm 0.00}$	–	–	$16.40_{\pm 1.10}$	$15.90_{\pm 1.10}$	$15.10_{\pm 0.99}$	$88.21_{\pm 2.86}$
CW-QMIX	$16.29_{\pm 0.27}$	$15.71_{\pm 7.04}$	$7.76_{\pm 2.22}$	$98.83_{\pm 0.76}$	$62.21_{\pm 5.88}$	–	$16.80_{\pm 0.59}$	$16.30_{\pm 1.60}$	$15.40_{\pm 1.82}$	$42.71_{\pm 8.90}$
POW-QMIX	$18.74_{\pm 0.11}$	$40.04_{\pm 2.54}$	$39.79_{\pm 2.73}$	$100.00_{\pm 0.00}$	$78.39_{\pm 1.21}$	$73.51_{\pm 3.82}$	$18.02_{\pm 0.68}$	$17.68_{\pm 0.51}$	$16.31_{\pm 1.56}$	$92.09_{\pm 2.70}$
Role, representation, and exploration										
RODE	$15.58_{\pm 0.88}$	$27.35_{\pm 3.76}$	$8.65_{\pm 5.62}$	$93.36_{\pm 2.49}$	$29.72_{\pm 8.36}$	$11.59_{\pm 6.29}$	$18.13_{\pm 1.72}$	$17.56_{\pm 1.62}$	$16.27_{\pm 0.79}$	$70.28_{\pm 5.49}$
LAGMA	$16.92_{\pm 0.73}$	$34.49_{\pm 3.66}$	$27.03_{\pm 2.84}$	$96.38_{\pm 3.61}$	$80.97_{\pm 8.58}$	$69.10_{\pm 8.23}$	$18.35_{\pm 1.42}$	$18.10_{\pm 0.95}$	$16.70_{\pm 0.93}$	$79.50_{\pm 6.85}$
Attention and coordination graphs										
LTSCG	$16.35_{\pm 0.51}$	$31.96_{\pm 5.42}$	$23.39_{\pm 4.02}$	$96.12_{\pm 2.49}$	$73.34_{\pm 3.38}$	$51.26_{\pm 4.41}$	$18.10_{\pm 1.69}$	$17.85_{\pm 1.18}$	$16.45_{\pm 1.57}$	$82.05_{\pm 4.91}$
GACG	$17.28_{\pm 0.36}$	$28.07_{\pm 2.82}$	$25.46_{\pm 4.62}$	$97.00_{\pm 2.19}$	$73.59_{\pm 5.25}$	$47.61_{\pm 5.97}$	$16.69_{\pm 0.65}$	$16.71_{\pm 0.60}$	$15.27_{\pm 1.09}$	$86.33_{\pm 2.97}$
CASEC	$17.28_{\pm 0.43}$	$32.52_{\pm 3.26}$	$28.38_{\pm 2.18}$	$98.42_{\pm 1.58}$	$64.85_{\pm 3.51}$	$11.63_{\pm 5.70}$	$17.20_{\pm 0.58}$	$17.05_{\pm 2.01}$	$15.75_{\pm 1.18}$	$87.54_{\pm 4.48}$
SOP-CG	$17.60_{\pm 0.41}$	$35.29_{\pm 6.38}$	$30.96_{\pm 3.40}$	$98.72_{\pm 1.19}$	$69.62_{\pm 5.55}$	$22.62_{\pm 4.71}$	$17.10_{\pm 0.99}$	$17.03_{\pm 1.02}$	$14.52_{\pm 1.17}$	$88.77_{\pm 5.99}$
TACTIC	$17.37_{\pm 0.35}$	$29.30_{\pm 2.65}$	$22.11_{\pm 5.27}$	$97.36_{\pm 2.64}$	$67.30_{\pm 4.27}$	$67.44_{\pm 2.04}$	$17.45_{\pm 1.50}$	$17.20_{\pm 1.63}$	$15.95_{\pm 1.41}$	$88.17_{\pm 3.11}$
Policy learning, grouping, and partner selection										
RIIT	$16.47_{\pm 0.64}$	$28.18_{\pm 3.40}$	$19.61_{\pm 7.09}$	$95.31_{\pm 3.68}$	$27.55_{\pm 6.38}$	$15.35_{\pm 1.98}$	$17.60_{\pm 1.21}$	$17.35_{\pm 0.79}$	$16.10_{\pm 0.96}$	$82.46_{\pm 3.20}$
ACORM	$17.13_{\pm 0.49}$	$31.03_{\pm 6.60}$	$27.48_{\pm 4.31}$	$96.25_{\pm 2.94}$	$74.68_{\pm 5.10}$	$69.24_{\pm 5.36}$	$18.63_{\pm 1.30}$	$18.37_{\pm 0.85}$	$16.82_{\pm 1.95}$	$85.46_{\pm 4.87}$
APS	$17.22_{\pm 0.62}$	$34.43_{\pm 5.16}$	$30.10_{\pm 6.81}$	$97.39_{\pm 1.51}$	$72.71_{\pm 3.83}$	$71.57_{\pm 5.76}$	$17.31_{\pm 1.34}$	$15.92_{\pm 0.64}$	$16.64_{\pm 1.47}$	$88.07_{\pm 3.17}$
Architecture-matched selector controls										
VIBE	$19.36_{\pm 0.23}$	$39.29_{\pm 2.77}$	$38.58_{\pm 3.39}$	$100.00_{\pm 0.00}$	$82.59_{\pm 3.68}$	$77.43_{\pm 2.62}$	$18.88_{\pm 0.78}$	$18.79_{\pm 0.67}$	$17.54_{\pm 1.13}$	$93.19_{\pm 2.00}$

Note. CCT, Predator–Prey, and SMACv2 use episodic return; SMACv1 and Intersection use win or success rate (%).

Table 2: Five-family comparison with late-window AUC.

Method	CCT	Predator–Prey		SMACv1: win AUC ↑			SMACv2: return AUC ↑			Driving
	Hit ↑	$p=-4$	$p=-5$	3s_vs_5z	Corr.	10m_vs_11m	Protoss	Terran	Zerg	Collision ↓
Value factorisation and joint-action modelling										
QTRAN	$6.10_{\pm 1.54}$	$7.92_{\pm 5.37}$	–	$89.02_{\pm 6.76}$	–	–	$15.02_{\pm 1.81}$	$15.17_{\pm 1.94}$	$14.15_{\pm 0.89}$	$35.01_{\pm 3.84}$
QPLEX	$8.40_{\pm 1.01}$	–	–	$96.27_{\pm 1.66}$	$43.10_{\pm 4.65}$	$38.12_{\pm 4.39}$	$16.70_{\pm 1.98}$	$16.43_{\pm 0.34}$	$15.63_{\pm 0.75}$	$42.83_{\pm 1.71}$
OW-QMIX	$9.35_{\pm 1.47}$	$17.18_{\pm 5.60}$	–	$100.00_{\pm 0.00}$	–	–	$16.33_{\pm 1.05}$	$15.83_{\pm 1.65}$	$15.02_{\pm 0.94}$	$11.48_{\pm 1.43}$
CW-QMIX	$10.27_{\pm 1.66}$	$38.42_{\pm 16.54}$	$19.74_{\pm 5.72}$	$98.42_{\pm 0.72}$	$62.91_{\pm 4.59}$	–	$16.72_{\pm 0.56}$	$16.25_{\pm 1.22}$	$15.31_{\pm 1.13}$	$35.51_{\pm 4.05}$
POW-QMIX	$12.90_{\pm 1.72}$	$91.09_{\pm 5.91}$	$94.01_{\pm 5.09}$	$100.00_{\pm 0.00}$	$78.09_{\pm 1.15}$	$73.91_{\pm 3.23}$	$18.08_{\pm 0.65}$	$17.74_{\pm 0.48}$	$16.20_{\pm 1.78}$	$4.82_{\pm 1.61}$
Role, representation, and exploration										
RODE	$10.05_{\pm 1.80}$	$66.27_{\pm 8.84}$	$21.33_{\pm 10.21}$	$93.49_{\pm 2.37}$	$29.52_{\pm 7.94}$	$11.39_{\pm 3.98}$	$18.20_{\pm 1.63}$	$17.58_{\pm 1.24}$	$16.15_{\pm 0.45}$	$18.35_{\pm 2.75}$
LAGMA	$14.18_{\pm 1.92}$	$83.05_{\pm 8.60}$	$65.52_{\pm 6.67}$	$96.77_{\pm 3.03}$	$80.27_{\pm 7.15}$	$69.40_{\pm 7.82}$	$18.57_{\pm 1.20}$	$17.15_{\pm 0.90}$	$16.62_{\pm 0.88}$	$10.28_{\pm 3.43}$
Attention and coordination graphs										
LTSCG	$14.62_{\pm 1.98}$	$76.11_{\pm 10.74}$	$56.97_{\pm 9.45}$	$96.73_{\pm 2.37}$	$73.84_{\pm 3.21}$	$51.76_{\pm 4.19}$	$18.02_{\pm 1.61}$	$17.92_{\pm 1.12}$	$16.36_{\pm 1.49}$	$12.87_{\pm 2.46}$
GACG	$16.80_{\pm 1.83}$	$66.96_{\pm 6.63}$	$60.83_{\pm 10.86}$	$97.39_{\pm 2.08}$	$73.94_{\pm 4.99}$	$47.96_{\pm 2.67}$	$16.82_{\pm 0.62}$	$16.78_{\pm 1.57}$	$15.20_{\pm 1.04}$	$8.52_{\pm 1.49}$
CASEC	$15.64_{\pm 1.70}$	$78.42_{\pm 4.66}$	$69.69_{\pm 5.12}$	$98.95_{\pm 1.05}$	$64.95_{\pm 3.33}$	$11.23_{\pm 5.42}$	$17.15_{\pm 0.55}$	$17.10_{\pm 1.91}$	$15.64_{\pm 1.12}$	$6.51_{\pm 1.84}$
SOP-CG	$20.18_{\pm 1.42}$	$85.93_{\pm 10.07}$	$75.76_{\pm 7.99}$	$98.32_{\pm 0.68}$	$69.62_{\pm 5.27}$	$22.72_{\pm 4.47}$	$17.02_{\pm 0.94}$	$17.08_{\pm 0.97}$	$14.45_{\pm 1.11}$	$5.94_{\pm 0.82}$
TACTIC	$19.42_{\pm 1.66}$	$69.86_{\pm 6.23}$	$54.96_{\pm 12.38}$	$97.77_{\pm 2.23}$	$67.70_{\pm 6.06}$	$67.70_{\pm 6.68}$	$17.38_{\pm 1.42}$	$17.22_{\pm 1.55}$	$15.86_{\pm 1.34}$	$8.51_{\pm 1.56}$
Policy learning, grouping, and partner selection										
RIIT	$13.55_{\pm 1.63}$	$69.22_{\pm 7.99}$	$49.08_{\pm 16.66}$	$95.29_{\pm 3.50}$	$27.35_{\pm 6.06}$	$15.15_{\pm 1.88}$	$17.52_{\pm 1.15}$	$17.40_{\pm 0.75}$	$15.98_{\pm 0.91}$	$8.65_{\pm 1.60}$
ACORM	$14.05_{\pm 1.65}$	$74.92_{\pm 15.51}$	$66.58_{\pm 8.13}$	$96.95_{\pm 2.79}$	$74.28_{\pm 4.85}$	$69.84_{\pm 5.09}$	$18.49_{\pm 1.34}$	$18.28_{\pm 0.81}$	$16.71_{\pm 1.85}$	$9.00_{\pm 2.44}$
APS	$17.31_{\pm 1.58}$	$81.91_{\pm 10.13}$	$71.74_{\pm 12.00}$	$97.80_{\pm 1.43}$	$72.11_{\pm 3.64}$	$71.97_{\pm 5.47}$	$17.20_{\pm 1.27}$	$15.88_{\pm 1.61}$	$16.50_{\pm 1.40}$	$5.56_{\pm 1.59}$
Architecture-matched selector controls										
VIBE	$87.63_{\pm 2.05}$	$94.33_{\pm 5.67}$	$90.16_{\pm 7.97}$	$100.00_{\pm 0.00}$	$82.47_{\pm 3.50}$	$77.37_{\pm 3.49}$	$18.74_{\pm 0.74}$	$18.62_{\pm 0.34}$	$17.38_{\pm 1.07}$	$4.15_{\pm 0.56}$

Note. CCT reports optimal-edge hit rate (%); Predator–Prey reports coordinated-capture rate (%); Driving reports collision rate (%). AUC is the trapezoidal area under the greedy-evaluation metric over 9M–10M environment steps divided by 1M.

(Samvelyan et al., 2019) contributes four combat scenarios: `3s_vs_5z` stresses asymmetric coordination, `corridor` constrained movement, and `10m_vs_11m` team-scale focus fire. **SMACv2** (Ellis et al., 2023) contributes Protoss, Terran, and Zerg distributions with stochastic team composition and capabilities. **Highway-env Intersection** (Leurent et al., 2018) evaluates cooperative right-of-way decisions under collision and deadlock risk.

Baselines. **(1) Value factorisation and joint-action modelling**: QTRAN uses transformation constraints (Son et al., 2019); QPLEX uses duplex-dueling factorisation (Wang et al., 2020); OW-QMIX and CW-QMIX apply optimistic and centrally weighted training, respectively (Rashid et al., 2020a); and POW-QMIX recognises potentially optimal joint actions and supplies VIBE's backbone (Huang et al., 2026). **(2) Role, representation, and exploration**: RODE decomposes tasks

Table 3: VIBE component ablations across five benchmark families.

Method	CCT		Predator–Prey		SMACv1: Corr.		SMACv2: Terran		Driving	
	Return	Hit	Return	Capture	Win	AUC	Return	AUC	Success	Collision
VIBE	$19.36_{\pm 0.23}$	$87.63_{\pm 1.05}$	$38.58_{\pm 3.39}$	$90.16_{\pm 4.97}$	$82.59_{\pm 3.68}$	$82.47_{\pm 3.50}$	$18.79_{\pm 0.67}$	$18.62_{\pm 0.34}$	$93.19_{\pm 2.30}$	$4.15_{\pm 0.56}$
w/o Belief	$18.19_{\pm 0.33}$	$80.33_{\pm 2.04}$	$30.51_{\pm 2.33}$	$73.70_{\pm 5.48}$	$75.43_{\pm 6.52}$	$74.63_{\pm 3.34}$	$17.55_{\pm 0.55}$	$17.52_{\pm 0.52}$	$86.57_{\pm 4.69}$	$8.39_{\pm 2.04}$
w/o Pair	$18.74_{\pm 0.36}$	$83.82_{\pm 3.21}$	$33.16_{\pm 7.34}$	$74.93_{\pm 17.25}$	$77.80_{\pm 5.93}$	$76.50_{\pm 5.63}$	$17.31_{\pm 1.70}$	$17.65_{\pm 1.61}$	$88.19_{\pm 3.99}$	$7.30_{\pm 2.00}$
w/o POW	$18.16_{\pm 0.41}$	$82.80_{\pm 3.65}$	$25.82_{\pm 3.33}$	$61.68_{\pm 7.03}$	$72.46_{\pm 3.18}$	$71.26_{\pm 3.02}$	$17.14_{\pm 1.32}$	$17.34_{\pm 1.25}$	$84.76_{\pm 4.19}$	$9.38_{\pm 2.10}$
w/o Rank	$18.99_{\pm 0.17}$	$84.96_{\pm 1.52}$	$36.79_{\pm 5.53}$	$84.46_{\pm 10.47}$	$79.84_{\pm 2.39}$	$79.24_{\pm 5.27}$	$17.87_{\pm 0.70}$	$17.66_{\pm 0.66}$	$91.38_{\pm 4.58}$	$6.24_{\pm 2.29}$
w/o Dynamic	$18.82_{\pm 0.19}$	$86.19_{\pm 1.89}$	$31.67_{\pm 4.67}$	$76.42_{\pm 12.97}$	$76.25_{\pm 2.60}$	$75.25_{\pm 2.47}$	$17.28_{\pm 1.90}$	$17.82_{\pm 1.80}$	$87.04_{\pm 5.24}$	$6.13_{\pm 2.92}$
w/o Payload	$18.61_{\pm 0.37}$	$84.24_{\pm 3.30}$	$29.41_{\pm 4.91}$	$73.11_{\pm 11.54}$	$74.38_{\pm 4.17}$	$73.58_{\pm 3.96}$	$17.42_{\pm 1.00}$	$17.65_{\pm 0.95}$	$86.88_{\pm 4.06}$	$7.02_{\pm 2.03}$
Oracle	$19.83_{\pm 0.22}$	$89.76_{\pm 1.96}$	$39.21_{\pm 1.47}$	$92.14_{\pm 2.63}$	$83.38_{\pm 2.39}$	$84.58_{\pm 2.25}$	$18.76_{\pm 0.83}$	$18.90_{\pm 0.79}$	$94.20_{\pm 1.33}$	$3.19_{\pm 0.67}$
Dense	$19.53_{\pm 0.23}$	$88.40_{\pm 2.05}$	$39.24_{\pm 1.79}$	$92.21_{\pm 3.21}$	$83.40_{\pm 2.61}$	$84.60_{\pm 2.48}$	$18.72_{\pm 0.28}$	$18.94_{\pm 0.46}$	$93.53_{\pm 1.10}$	$3.56_{\pm 0.55}$

Note. CCT uses episodic return and optimal-edge hit rate (%); Predator–Prey uses return and coordinated-capture rate (%); SMACv1 Corr. and SMACv2 Terran uses win rate / return and 9M–10M win-rate AUC (%); Driving Intersection uses success and collision rate (%).The Predator-Prey columns use the $p = -5$ setting.

Table 4: Model-side execution latency (ms / team step) across benchmark families and team scales.

Method	Predator–Prey			SMACv1			SMACv2		Intersection	
	n = 4	n = 6	n = 8	n = 3	n = 8	n = 10	n = 5	n = 10	n = 4	n = 8
GACG	$0.904_{\pm 0.020}$	$1.164_{\pm 0.034}$	$1.338_{\pm 0.035}$	$0.948_{\pm 0.025}$	$1.628_{\pm 0.044}$	$2.013_{\pm 0.068}$	$1.509_{\pm 0.049}$	$2.530_{\pm 0.081}$	$1.020_{\pm 0.024}$	$1.465_{\pm 0.047}$
CASEC	$0.910_{\pm 0.029}$	$1.113_{\pm 0.036}$	$1.304_{\pm 0.040}$	$1.042_{\pm 0.025}$	$1.687_{\pm 0.045}$	$2.052_{\pm 0.065}$	$1.497_{\pm 0.041}$	$2.419_{\pm 0.070}$	$1.000_{\pm 0.022}$	$1.502_{\pm 0.040}$
SOP-CG	$0.854_{\pm 0.030}$	$1.087_{\pm 0.024}$	$1.259_{\pm 0.041}$	$0.920_{\pm 0.032}$	$1.540_{\pm 0.051}$	$1.896_{\pm 0.065}$	$1.448_{\pm 0.047}$	$2.314_{\pm 0.088}$	$0.960_{\pm 0.024}$	$1.411_{\pm 0.037}$
TACTIC	$0.786_{\pm 0.012}$	$0.945_{\pm 0.022}$	$1.099_{\pm 0.032}$	$0.865_{\pm 0.023}$	$1.365_{\pm 0.031}$	$1.630_{\pm 0.064}$	$1.365_{\pm 0.043}$	$2.017_{\pm 0.073}$	$0.884_{\pm 0.027}$	$1.224_{\pm 0.044}$
APS	$0.740_{\pm 0.020}$	$0.906_{\pm 0.027}$	$1.030_{\pm 0.035}$	$0.797_{\pm 0.021}$	$1.306_{\pm 0.036}$	$1.524_{\pm 0.036}$	$1.151_{\pm 0.033}$	$1.756_{\pm 0.054}$	$0.841_{\pm 0.024}$	$1.179_{\pm 0.035}$
VIBE	$0.745_{\pm 0.023}$	$0.888_{\pm 0.021}$	$1.063_{\pm 0.037}$	$0.861_{\pm 0.024}$	$1.281_{\pm 0.038}$	$1.498_{\pm 0.055}$	$1.214_{\pm 0.040}$	$1.821_{\pm 0.062}$	$0.876_{\pm 0.023}$	$1.168_{\pm 0.037}$

Note. VIBE timing includes receiver-local belief inference, Edge Student ranking, top-K selection, relational aggregation, and action selection, while its centralised Teacher and joint-probe evaluation are excluded.

through specialised roles (Wang et al., 2021), while LAGMA learns latent goal-guided abstractions (Na & Moon, 2024). **(3) The graph family** compares LTSCG's temporal sparse graph (Duan et al., 2025), GACG's group-aware graph (Duan et al., 2024a), CASEC's payoff-sensitivity criterion (Wang et al., 2022), SOP-CG's tractable sparse topology (Yang et al., 2022), and TACTIC's task-aware sparse graph (Peng et al., 2026). **(4) Policy learning and partner selection** compare MAPPO as a centralised-critic policy optimiser (Yu et al., 2022), RIIT as an implementation-centred value-factorisation reference (Hu et al., 2021a), ACORM as a contrastive role representation (Hu et al., 2024), and APS as autonomous partner selection (Tang et al., 2026).

All experiments are conducted on 8 NVIDIA A100 GPUs. We train each method for 10M environment steps over 10 random seeds, evaluating every 256 steps with 32 episodes per evaluation. For the AUC analysis, we measure learning performance over the 9M–10M environment-step window. All methods use the same training budget and evaluation protocol.

5.2 Results and Analysis

Main benchmark results. Table 1 compares VIBE with value-factorisation, role-based, coordination-graph, and partner-selection methods on ten tasks from five benchmark families. VIBE achieves the highest reported mean on CCT, corridor, 10m_vs_11m, all three SMACv2 races, and Intersection, and ties for the highest win rate on 3s_vs_5z. Relative to its POW-QMIX backbone, VIBE improves CCT return from 18.74 to 19.36, corridor win rate from 78.39% to 82.59%, and 10m_vs_11m win rate from 73.51% to 77.43%. On SMACv2, VIBE reaches returns of 18.88, 18.79, and 17.54 on Protoss, Terran, and Zerg, respectively, compared with 18.02, 17.68, and 16.31 for POW-QMIX. Intersection safe completion increases from 92.09% to 93.19%. VIBE improves on POW-QMIX in seven of the ten columns and ties on 3s_vs_5z; it exceeds CASEC and SOP-CG in all ten columns. These establish decision-relevant edge allocation as a strong approach.

Table 5: Edge-selection diagnostics across tasks.

Method	CCT	Corr.	MMM2	10m	Int.
Variance-K	8.73	60.84	56.62	54.71	62.58
CASEC	15.64	64.73	60.21	58.36	66.74
SOP-CG	21.18	68.41	63.27	60.86	69.35
TACTIC	19.42	70.68	67.15	64.23	71.76
APS	17.31	69.24	65.71	62.44	70.18
VIBE	**87.63**	**85.71**	**82.36**	**79.28**	**83.74**

Note. CCT measures optimal-edge hit rate (%); the remaining tasks compare selected-edge Teacher utility.

Complementary outcomes and learning performance. Table 2 connects the primary outcomes to edge-selection accuracy, coordinated capture, learning AUC, and driving safety. VIBE achieves an 87.63% optimal-edge hit rate on CCT, compared with 12.90% for POW-QMIX. On Predator–Prey with penalty $p = -4$, its coordinated-capture rate reaches 94.33%, compared with 91.09% for POW-QMIX. Over the 9M–10M training window, VIBE reaches win-rate AUCs of 82.47 and 77.37 on corridor and 10m_vs_11m, respectively, improving on the corresponding POW-QMIX values of 78.09 and 73.91. Its SMACv2 return AUCs are 18.74, 18.62, and 17.38 on Protoss, Terran, and Zerg, compared with 18.08, 17.74, and 16.20 for POW-QMIX. VIBE also reduces Intersection collision rate from 4.82% to 4.15%. Across the ten complementary metrics, VIBE improves on POW-QMIX in eight columns and ties on 3s_vs_5z. The secondary outcomes therefore reinforce the link between decision-relevant selection, learning performance, and coordinated execution.

Ablation analysis. Table 3 evaluates six component-removal variants on one task from each benchmark family. Full VIBE achieves a higher primary endpoint than every removal variant on all five selected tasks. On corridor, removing belief estimation reduces win rate from 82.59% to 75.43%; removing paired evidence yields 77.80%, and removing POW candidates yields 72.46%. On CCT, removing ranking supervision reduces optimal-edge hit rate from 87.63% to 84.96%. The accompanying 9M–10M AUC columns assess the same components over a common learning window. These comparisons connect belief estimation, paired Teacher evaluation, candidate recognition, ranking distillation, dynamic edge selection, and relational processing to VIBE's gains.

Execution efficiency. Table 4 measures model-side latency per team step, including receiver-local belief inference, Student ranking, top-K selection, relational aggregation, and action selection. VIBE has lower latency than CASEC across all ten reported family–scale settings. At ten SMACv1 agents, VIBE requires 1.641 ms per team step, compared with 2.013 ms for GACG and 2.052 ms for CASEC. Centralised Teacher scoring and joint-action probes are training-side operations; execution uses the receiver-local Student and selected relational features.

Edge-selection mechanism. Table 5 compares selectors on shared informative states. VIBE recovers the optimal CCT edge in 87.63% of cases, compared with 8.73% for Variance-K. It also achieves the highest reported selected-edge Teacher utility on corridor, MMM2, 10m_vs_11m, and Intersection. Its utility reaches 85.71 on corridor and 82.36 on MMM2. These results directly support the paper's central distinction: an edge's payoff sensitivity and its value for the receiver's current decision can induce different teammate rankings.

Table 6: Joint configuration analysis.

M	P	C	α	Rank A.	Val. tAUC
8	16	0.20	0.025	88.2	$60.0_{\pm 1.7}$
			0.050		$64.5_{\pm 1.1}$
			0.100		$65.1_{\pm 2.3}$
16	16	0.10	0.025	89.8	$60.5_{\pm 1.6}$
			0.050		$65.0_{\pm 1.1}$
			0.100		$65.6_{\pm 2.2}$
8	**32**	**0.10**	0.025	**91.3**	$61.3_{\pm 1.7}$
			0.050		$65.8_{\pm 1.1}$
			0.100		$66.4_{\pm 2.5}$
16	32	0.10	0.025	94.3	$61.5_{\pm 1.8}$
			0.050		$66.0_{\pm 1.1}$
			0.100		$\mathbf{66.6}_{\pm \mathbf{2.3}}$
8	64	0.10	0.025	93.0	$61.4_{\pm 1.6}$
			0.050		$66.0_{\pm 1.2}$
			0.100		$66.5_{\pm 2.4}$

SMACv1-8m. Rank fidelity measures bounded-Teacher agreement, while validation tAUC measures end-to-end learning efficiency.

Joint configuration analysis. Table 6 examines retained candidate count M, beam size P, POW tolerance C, and relational strength α together. The setting $M = 16$, $P = 32$, $C = 0.1$, and $\alpha = 0.05$ attains 94.3 ranking agreement in the reported SMACv1-8m analysis. Validation tAUC is evaluated over the same 9M–10M window, connecting bounded Teacher fidelity with task-level learning performance. The configuration analysis complements the component ablations by showing how candidate construction and relational control jointly shape VIBE's results.

6 Related Work

Value factorisation and joint-action selection. Value factorisation aligns decentralised greedy choices with centralised maximisation (IGM) through additive (Sunehag et al., 2018), monotonic (Rashid et al., 2020b), or relaxed-consistency decompositions (Son et al., 2019). Weighted variants (Rashid et al., 2020a) and potentially-optimal recognisers (Huang et al., 2026) prioritise joint-action fidelity, not budgeted teammate selection: every agent's utility remains in the mixer. VIBE fixes

the backbone and uses recogniser Q_r only to bound teacher probes (§3.2); its guarantees concern joint-action retention, not receiver beliefs.

Coordination graphs and learned interaction structure. Coordination graphs optimise joint actions through pairwise utilities (Böhmer et al., 2020). To reduce costs, CASEC sparsifies by payoff variance (Wang et al., 2022), while SOP-CG learns topologies admitting polynomial-time optimisation (Yang et al., 2022). DICG infers interactions from representations (Li et al., 2021), and RODE uses roles (Wang et al., 2021). In CASEC and SOP-CG, edges govern value terms or action optimisation, unlike VIBE's local processing routes. CASEC already links edge removal to greedy decisions, but payoff variance and variance of the receiver's action advantage can favour a teammate with zero decision value over one whose uncertainty changes the receiver's action.

Communication scheduling and the value of information. NDQ minimises message information under performance constraints (Wang et al., 2019), VBC gates requests by action-value ambiguity and filters replies by variance (Zhang et al., 2019), IEC guides sparse collaboration based on the consistency value of shared information (Du et al., 2026), MAIC routes messages through teammate modelling (Yuan et al., 2022), T2MAC selects targeted, trusted partners (Sun et al., 2024), and contrastive pre-training learns communication representations (Yu et al., 2025). I2C is closest (Ding et al., 2020): centralised joint values quantify teammate actions' causal effects on receiver action distributions through KL, training a local channel-opening prior. VIBE shares this supervision structure but targets expected return gained through the receiver's decision: distributional shifts need not change the greedy action and are not weighted by the value at stake. Edges process already-observed features. Under Bayesian updating, this classical target equals expected value of sample information; joint observation selection is generally intractable in graphical models.

Local representation and privileged supervision. Teammate models extract representations from local history (Papoudakis et al., 2021), while UPDeT attends to observed entities (Hu et al., 2021b). Privileged supervision (Chen et al., 2020) incurs imitation gaps when teacher choices depend on locally unrecoverable information (Weihs et al., 2021). VIBE contributes decision-value supervision for budgeted processing of already-observed teammate features. Matched comparisons test whether this capacity bias improves learning efficiency under decentralised execution.

7 Conclusion

VIBE allocates sparse coordination budgets under partial observability by expected decision-regret reduction. Its criterion factorises into switching probability and conditional gain, exposing why coupling, similarity, or uncertainty alone can misrank teammates. A centralised teacher uses bounded potentially optimal probes in exponential joint-action spaces to supervise a decentralised edge student that constructs graphs from local history for graph-conditioned control. These graphs route already-observed features by the value of resolving uncertainty at the receiver's decision boundary, providing a decision-theoretic principle for scalable cooperative coordination.

8 Reproducibility Statement

All benchmarks used in our experiments are publicly available. We provide implementation details, hyperparameters, source code, and configurations in an anonymous repository to support independent replication.

9 AI USE STATEMENT

We used generative AI tools solely for language polishing and proofreading of the manuscript. We did not use generative Al for experiments and data analysis. The authors take full responsibility for the final content, including all text, analyses, and conclusions.

REFERENCES

Wendelin Böhmer, Vitaly Kurin, and Shimon Whiteson. Deep coordination graphs. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Conference on robot learning*, pp. 66–75. PMLR, 2020.

Ziluo Ding, Tiejun Huang, and Zongqing Lu. Learning individually inferred communication for multi-agent cooperation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Xuefeng Du, Jiajun Wu, Yuduo Zheng, and Fengqi Li. IEC: When information-driven exploration meets spectral consensus via primal–dual reward regularization in decentralized MARL. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=FsKoSE1xq8.

Wei Duan, Jie Lu, and Junyu Xuan. Group-aware coordination graph for multi-agent reinforcement learning. *arXiv preprint arXiv:2404.10976*, 2024a.

Wei Duan, Jie Lu, and Junyu Xuan. Group-aware coordination graph for multi-agent reinforcement learning. *arXiv preprint arXiv:2404.10976*, 2024b.

Wei Duan, Jie Lu, and Junyu Xuan. Inferring latent temporal sparse coordination graph for multiagent reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36 (8):14358–14370, 2025. doi: 10.1109/TNNLS.2024.3513402.

Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Foerster, and Shimon Whiteson. Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning. *Advances in neural information processing systems*, 36:37567–37593, 2023.

Wenzhe Fan, Zishun Yu, Chengdong Ma, Changye Li, Yaodong Yang, and Xinhua Zhang. Towards efficient collaboration via graph modeling in reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16505–16513, 2025.

J Hu, S Jiang, SA Harding, H Wu, and SW Liao. Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning. arxiv 2021. *arXiv preprint arXiv:2102.03479*, 2021a.

Siyi Hu, Fengda Zhu, Xiaojun Chang, and Xiaodan Liang. Updet: Universal multi-agent reinforcement learning via policy decoupling with transformers. *arXiv preprint arXiv:2101.08001*, 2021b.

Tianmeng Hu, Yongzheng Cui, Rui Tang, Biao Luo, and Ke Li. Beyond monotonicity: Revisiting factorization principles in multi-agent q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21876–21884, 2026.

Zican Hu, Zongzhang Zhang, Huaxiong Li, Chunlin Chen, Hongyu Ding, and Zhi Wang. Attention-guided contrastive role representations for multi-agent reinforcement learning. In *International conference on learning representations*, volume 2024, pp. 23627–23649, 2024.

Chang Huang, Shatong Zhu, Junqiao Zhao, Hongtu Zhou, Hai Zhang, Di Zhang, Chen Ye, Ziqiao Wang, and Guang Chen. Potentially optimal joint actions recognition for cooperative multi-agent reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YQ1muQBDV4.

Edouard Leurent et al. An environment for autonomous driving decision-making, 2018.

Sheng Li, Jayesh K. Gupta, Peter Morales, Ross Allen, and Mykel J. Kochenderfer. Deep implicit coordination graphs for multi-agent reinforcement learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, pp. 764–772, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450383073.

Hyungho Na and Il-Chul Moon. Lagma: latent goal-guided multi-agent reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Georgios Papoudakis, Filippos Christianos, and Stefano V. Albrecht. Agent modelling under partial observability for deep reinforcement learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Kexing Peng, Pengyi Li, tinghuai ma, and Jianye HAO. TACTIC: Task-aware sparse coordination graphs for multi-task multi-agent reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=1AsU8u2tmC.

Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted qmix: expanding monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020a. Curran Associates Inc. ISBN 9781713829546.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *J. Mach. Learn. Res.*, 21(1), January 2020b. ISSN 1532-4435.

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*, 2019.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. *ArXiv*, abs/1905.05408, 2019. URL https://api.semanticscholar.org/CorpusID:153312464.

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 Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i13.29438. URL https://doi.org/10.1609/aaai.v38i13.29438.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pp. 2085–2087, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

Rui Tang, Biao Luo, and Yongzheng Cui. Autonomous partner selection for cooperative multi-agent reinforcement learning. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i30.39783. URL https://doi.org/10.1609/aaai.v40i30.39783.

Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. Qplex: Duplex dueling multi-agent q-learning. *ArXiv*, abs/2008.01062, 2020. URL https://api.semanticscholar.org/CorpusID:220935959.

Tonghan Wang, Jianhao Wang, Chongyi Zheng, and Chongjie Zhang. Learning nearly decomposable value functions via communication minimization. *ArXiv*, abs/1910.05366, 2019. URL https://api.semanticscholar.org/CorpusID:204512179.

Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. {RODE}: Learning roles to decompose multi-agent tasks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=TTUVg6vkNjK.

Tonghan Wang, Liang Zeng, Weijun Dong, Qianlan Yang, Yang Yu, and Chongjie Zhang. Context-aware sparse deep coordination graphs. *ArXiv*, abs/2106.02886, 2022. URL https://api.semanticscholar.org/CorpusID:260434138.

Luca Weihs, Unnat Jain, Iou-Jen Liu, Jordi Salvador, Svetlana Lazebnik, Aniruddha Kembhavi, and Alex Schwing. Bridging the imitation gap by adaptive insubordination. *Advances in Neural Information Processing Systems*, 34:19134–19146, 2021.

Qianlan Yang, Weijun Dong, Zhizhou Ren, Jianhao Wang, Tonghan Wang, and Chongjie Zhang. Self-organized polynomial-time coordination graphs. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24963–24979. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/yang22a.html.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

Peihong Yu, Manav Mishra, Syed Zaidi, and Pratap Tokekar. Task-agnostic contrastive pre-training for inter-agent communication. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, pp. 2262–2270, Richland, SC, 2025. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400714269.

Lei Yuan, Jianhao Wang, Fuxiang Zhang, Chenghe Wang, Zongzhang Zhang, Yang Yu, and Chongjie Zhang. Multi-agent incentive communication via decentralized teammate modeling. In *AAAI Conference on Artificial Intelligence*, 2022. URL https://api.semanticscholar.org/CorpusID:249118952.

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.

Sai Qian Zhang, Qi Zhang, and Jieyu Lin. Efficient communication in multi-agent reinforcement learning via variance based control. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Xiwen Zhang, Jie Chen, Ming-Gang Gan, and Haoxiang Chen. Influence enhanced sparse coordination graphs for multi-agent reinforcement learning. *Neural Networks*, 188:107454, 2025.

A Formal Setting up

A.1 Notation

Table 7 groups the recurrent control variables, ideal belief-conditioned decision objects, practical Teacher quantities, and local execution variables. Here, h_i^t denotes the complete local history and r_i^t its recurrent representation replaces the full belief b_i^t in the ideal problem.

Table 7: **Main notation.**

Symbol	Meaning
Agents, information, and control	
$\mathcal{N} = [n], i, j$	Agents and ordered receiver–teammate indices.
$h_i^t,\ r_i^t$	Complete local action–observation history and pre-action recurrent representation.
$o_i^t,\ o_{i\to j}^t$	Receiver observation and its already-observed teammate feature block.
$\mathcal{A}_i$	Receiver action set; feasible actions are fixed in each ideal comparison.
$E_i^t,\ K$	Selected non-self teammates and maximum per-receiver edge count.
$Q^{\boldsymbol{\pi}},\ Q_{\mathrm{tot}}$	Fixed-continuation team action value and learned monotonic mixer.
Ideal receiver decision problem	
ξ	Unobserved state, teammate histories, and current teammate actions.
$b_i^t,\ e_j$	Full receiver belief conditional on h_i^t and hypothetical teammate evidence.
$b_i^{-j},\ b_i^{+j}$	Prior belief and Bayesian posterior after observing e_j.
$V_i(b, a_i)$	Team return of receiver action a_i, integrating ξ under b.
$a_{i,*}^{-j},\ a_{i,*}^{+j}$	Prior and posterior maximisers over the same receiver action set.
Δ_{ij}^t	Ideal expected gain from adapting the receiver action to e_j.
Practical teacher and ranking	
$g_\psi,\ b_{ij}$	Pairwise belief encoder and unmasked categorical action prediction.
b_{ij}^{T}	Teacher-only prediction renormalised over teammate available actions.
$Q_r, \mathcal{C}_t$	Potentially-optimality recogniser and shared joint-action probe set.
k, P, M, C	Local proposal count, beam cap, retained probe cap, and recognition tolerance.
$\bar{\mathbf{e}}_{ij}$	Exact categorical mean of teammate action embeddings.
$\hat{Q}^+, \hat{Q}^-$	Matched critic branches with a probe action or its mean embedding.
$\widehat{\Delta}_{ij}^t$	Entropy-weighted, receiver-switch-gated joint-probe ranking surrogate.
$j_i^\star, w_i$	Teacher top-ranked teammate and informative-row weight.
Student and local relational pathway	
$x_{ij}^t,\ f_\eta,\ s_{ij}^t$	Receiver-local student input, shared student, and ranking logit.
$\phi,\ m_i^t, \alpha$	Teammate-feature encoder, mean relational payload, and channel scale.
$E_{\mathrm{T},i}^t$	Diagnostic top-K set under practical teacher scores.
$E_{\mathrm{S},i}^t$	Student top-K set, equal to E_i^t during ordinary execution.

A.2 Receiver Decision Problem

Fix a receiver history, feasible action set, and continuation policy $\boldsymbol{\pi}$. The prior conditions on the receiver's complete local history and hypothetical evidence e_j induces the posterior by Bayesian conditioning; b_i^{-j} therefore does not remove information already available to the controller.

The optimisations in receiver-optima vary only a_i. Prior and posterior decisions are evaluated under the same posterior in delta, with deterministic tie-breaking preserving the prior action whenever it remains posterior-optimal. Hence Lemma 1 counts only positive-valued decision changes.

Evidence e_j defines training supervision, not an execution-time input. VIBE instead uses it to allocate processing among receiver-observable teammate features; linking the ideal intervention to graph-conditioned control is therefore an empirically tested representation-allocation hypothesis.

A.3 Centralised Teacher and Receiver-Local Student

The Teacher evaluates a common probe set C^t under a fixed centralised context. For each (i, j), $\widehat{Q}^+$ uses teammate j's probe action embedding, whereas $\widehat{Q}^-$ replaces only that embedding by its exact mean under the masked Teacher belief b_{ij}^{T}; all other inputs and evaluator parameters are shared. This is the practical intervention in teacher-score, not a separately learned Bayesian posterior.

The Student uses only the receiver-local inputs in student-input: receiver representation and observation, pair identities, receiver-observable teammate features, and unmasked belief b_{ij}. Global state, teammate recurrent states, realised teammate actions, and teammate availability masks are excluded. Selection is over eligible non-self teammates, with an empty set producing zero relational payload. Teacher-only information thus supervises ranking but is unavailable at deployment.

A.4 Temporal Alignment and Stage Boundaries

Beliefs are formed from pre-action receiver representations and trained against the corresponding recorded teammate action. Recurrent context is reconstructed within the same episode, preventing future observations, outcomes, or padding from becoming predictive evidence.

Training follows Section 3.5: train the backbone, calibrate the belief encoder with the backbone frozen, distil the Student with backbone, Teacher, and belief encoder frozen, then enable graph-conditioned learning from the shared checkpoint. Frozen-state ranking uses $\alpha = 0$, whereas end-to-end learning uses $\alpha > 0$.

The matched end-to-end comparison fixes the initial checkpoint, edge and training budgets, relational pathway, and evaluation protocol while varying only the edge-selection criterion. Trajectories may subsequently diverge; the comparison matches the intervention and resources.

B Theoretical Proofs

We prove the main decision-value, allocation, approximation, and distillation results while distinguishing ideal values, practical Teacher scores, and Student logits.

B.1 Decision Value and Switching

Proof of Lemma 1. Condition on h_i^t and write $e = e_j$, $b = b_i^{-j}$, $b^e = b_i^{+j}$, $a^- = a_{i,*}^{-j}$, $a^e = a_{i,*}^{+j}$, and $q_a(\xi) = Q^{\boldsymbol{\pi}}(s, \mathbf{h}, (a, \mathbf{a}_{-i}))$. Posterior optimality and the tower property yield

$$\begin{cases} G(e) = \max_a V_i(b^e, a) - V_i(b^e, a^-) \geq 0, \\ \mathbb{E}[V_i(b^e, a^-) \mid h_i^t] = \mathbb{E}[\mathbb{E}[q_{a^-}(\xi) \mid h_i^t, e] \mid h_i^t] = \max_a V_i(b, a). \end{cases} \tag{10}$$

Under the prior-preserving posterior tie rule, $\mathcal{S}_{ij} = \{a^e \neq a^-\} = \{G(e) > 0\}$. Hence

$$\begin{aligned} \Delta_{ij}^t &= \mathbb{E}[G(e) \mid h_i^t] \\ &= \mathbb{E}[\max_a V_i(b^e, a) \mid h_i^t] - \max_a V_i(b, a) \\ &= \Pr(\mathcal{S}_{ij} \mid h_i^t)\mathbb{E}[G(e) \mid \mathcal{S}_{ij}, h_i^t]. \end{aligned} \tag{11}$$

The product is defined as zero when the switching probability is zero. □

Decision geometry. Since $v_i^\star(b) = \max_a V_i(b, a)$ is convex and each action value is linear in b,

$$\Delta_{ij}^t = \mathbb{E}\left[v_i^\star(b^e) - v_i^\star(b) - \int q_{a^-}(\xi)\, \mathrm{d}(b^e - b)(\xi) \,\middle|\, h_i^t \right]. \tag{12}$$

The integral averages to zero; the residual vanishes exactly when a^- remains posterior-optimal.

B.2 Variance-Ranking Reversal

Proof of Proposition 1. Let ν_j^{unif}, ν_j^{bel}, and ν_j^{adv} denote uniform-sweep, belief-weighted payoff, and belief-weighted advantage variances. At $a_i = 0$, the payoff pairs are $(0, 20)$ and $(3.5, -2.5)$, and both vanish at $a_i = 1$. Therefore $(\nu_A^{\mathrm{unif}}, \nu_B^{\mathrm{unif}}) = (100, 9)$, $(\nu_A^{\mathrm{bel}}, \nu_B^{\mathrm{bel}}) = (19, 9)$, $(\nu_A^{\mathrm{adv}}, \nu_B^{\mathrm{adv}}) = (19, 9)$. Independence gives $V_i(b, 0) - V_i(b, 1) = 20(0.05) + 3.5 - 6(0.5) = 1.5$, $V_i(b^{+A}, 0) - V_i(b^{+A}, 1) = 0.5 + 20a_A$, $V_i(b^{+B}, 0) - V_i(b^{+B}, 1) = 4.5 - 6a_B$. Thus the prior chooses 0; revealing A never changes choice, whereas revealing B switches it at $a_B = 1$, giving $\Delta_{iA} = 0 < \Delta_{iB} = 0.75$.

Now perturb each payoff entry by at most 0.1. Any prior or posterior advantage changes by at most 0.4, below the smallest margin. All switching events persist and $\widetilde{\Delta}_{iA} = 0$, $\widetilde{\Delta}_{iB} \geq 0.55$. Moreover,

$$\begin{cases} \widetilde{\nu}_A^{\mathrm{unif}} \geq 19.8^2/4 > 6.2^2/4 \geq \widetilde{\nu}_B^{\mathrm{unif}}, \\ \widetilde{\nu}_A^{\mathrm{bel}} \geq 0.05 \cdot 0.95 \cdot 19.8^2 > 6.2^2/4 \geq \widetilde{\nu}_B^{\mathrm{bel}}, \\ \widetilde{\nu}_A^{\mathrm{adv}} \geq 0.05 \cdot 0.95 \cdot 19.6^2 > 6.4^2/4 \geq \widetilde{\nu}_B^{\mathrm{adv}}. \end{cases} \tag{13}$$

Hence the strict ranking reversal is robust and arises receiver teammate interactions. □

B.3 Greedy Consistency

Proof of Lemma 2. With action-independent selection, E_i^t and m_i^t are fixed across candidate actions. For $a_i^\star \in \arg\max_{a_i} Q_i(h_i^t, a_i; m_i^t)$, coordinate-wise monotonicity implies $Q_i(h_i^t, a_i^\star; m_i^t) \geq Q_i(h_i^t, a_i; m_i^t)$, $Q_{\mathrm{tot}}(s^t, \mathbf{h}^t, \mathbf{a}^\star) \geq Q_{\mathrm{tot}}(s^t, \mathbf{h}^t, \mathbf{a})$. This proves igm-inclusion non-strict monotonicity may introduce additional joint ties. □

Theorem 1. *On a nonempty shared probe set with finite critic values, teacher-score is nonnegative. It is zero when the selected receiver actions agree, the Teacher belief is deterministic, or identical branches share a deterministic tie-break.*

Proof. For $|\mathcal{A}_j| \geq 2$, $0 \leq \omega_{ij} = H(b_{ij}^{\mathrm{T}})/\log|\mathcal{A}_j| \leq 1$, $\widehat{\Delta}_{ij}^t = \omega_{ij} \mathbb{I}[\hat{a}_i^+ \neq \hat{a}_i^-][\max_{\mathbf{a} \in \mathcal{C}_t} \hat{Q}^+(\mathbf{a}) - \hat{Q}^+(\hat{\mathbf{a}}^-)] \geq 0$. The stated cases zero the switch gate, entropy weight, or branch difference, for singleton alphabets the score is defined as zero. □

Accurate critics alone do not make mean-embedding evaluation equal belief integration, nor joint-probe maximisation equal receiver-only optimisation. Entropy reweights the score, and receiver-independent terms may alter probe selection although they cancel from the ideal comparison.

B.4 Evidence-Set Allocation

For $S \subseteq \mathcal{N} \setminus \{i\}$, let $b_i^S = P(\xi \mid h_i^t, e_S)$ and let a_i^0 be prior-optimal. Define

$$\begin{aligned} F_i(S) &= \mathbb{E}[\max_{a_i} V_i(b_i^S, a_i) - V_i(b_i^S, a_i^0) \mid h_i^t] \\ &= \mathbb{E}[\max_{a_i} V_i(b_i^S, a_i) \mid h_i^t] - \max_{a_i} V_i(b_i^\varnothing, a_i). \end{aligned} \tag{14}$$

Thus $F_i(\varnothing) = 0$ and $F_i(\{j\}) = \Delta_{ij}^t$.

Theorem 2. *F_i is nonnegative and monotone. Under modularity, selecting the largest K singleton values is optimal. General budgeted maximisation is NP-hard for succinct factored decision problems, sequential greedy selection using conditional marginals achieves $1 - 1/e$ of optimum.*

Proof. Nonnegativity follows by retaining a_i^0; for $S \subseteq T$, the receiver with e_T can retain the e_S-optimal decision, giving $F_i(T) \geq F_i(S)$. Under modularity, exchange arguments yield the largest K singleton values.

For NP-hardness, reduce maximum coverage. Let independent fair bits X_ℓ index a universe and $e_j = (X_\ell)_{\ell \in T_j}$, with $R(a, X) = \sum_{\ell=1}^r \mathbb{I}[a_\ell = X_\ell]$. For $U(S) = \cup_{j \in S} T_j$, $\mathbb{E}[\max_a V_i(b_i^S, a)] = |U(S)| + (r - |U(S)|)/2$, $F_i(S) = |U(S)|/2$. Hence optimal evidence selection solves maximum coverage. Finally, for greedy sets S_ℓ, optimum $S^\star$, and $g_\ell = F_i(S^\star) - F_i(S_\ell)$, submodularity gives

$$\begin{cases} g_\ell \leq \sum_{j \in S^\star \setminus S_\ell} [F_i(S_\ell \cup \{j\}) - F_i(S_\ell)] \\ \qquad \leq K \max_{j \notin S_\ell} [F_i(S_\ell \cup \{j\}) - F_i(S_\ell)], \\ g_{\ell+1} \leq (1 - 1/K) g_\ell, \\ F_i(S_K) \geq [1 - (1 - 1/K)^K] F_i(S^\star) \geq (1 - 1/e) F_i(S^\star). \end{cases} \tag{15}$$

□

VIBE ranks fixed singleton logits rather than recomputed conditional marginals. Independence alone does not imply modularity: independent bits can have zero individual value but positive joint value. The primary $K = 1$ setting avoids such multi-source interactions, not surrogate error.

B.5 Candidate and Plug-In Approximation

Fix h_i^t, abbreviate $b = b_i^{-j}$ and $b^e = b_i^{+j}$, and let an auxiliary belief-integrating evaluator restrict receiver actions to $\varnothing \neq \mathcal{D}_i \subseteq \mathcal{A}_i$. The support is chosen before evidence and independently of latent context given h_i^t. For $a_i^{-,\mathcal{D}} \in \arg\max_{a \in \mathcal{D}_i} V_i(b, a)$, $G_{ij}^{\mathcal{D}}(e) = \max_{a \in \mathcal{D}_i} V_i(b^e, a) - V_i(b^e, a_i^{-,\mathcal{D}})$, and $\Delta_{ij}^{\mathcal{D}} = \mathbb{E}_{\mathcal{D}_i, e | h_i^t}[G_{ij}^{\mathcal{D}}(e)]$.

Theorem 3 (Receiver-action coverage error)**.** *If $\mathcal{D}_i$ contains the selected full-space prior action and a posterior maximiser except with conditional probability δ_{cov}, then $|\Delta_{ij}^{\mathcal{D}} - \Delta_{ij}^t| \leq 2Q_{\max}\delta_{\mathrm{cov}}$.*

Proof. On the coverage event $\mathcal{R}_{ij}$ both regrets coincide; otherwise each lies in $[0, 2Q_{\max}]$. Hence $|\Delta_{ij}^{\mathcal{D}} - \Delta_{ij}^t| \leq \mathbb{E}[\mathbb{I}[\mathcal{R}_{ij}^c]|G_{ij}^{\mathcal{D}} - G_{ij}| \mid h_i^t] \leq 2Q_{\max} \Pr(\mathcal{R}_{ij}^c \mid h_i^t) \leq 2Q_{\max}\delta_{\mathrm{cov}}$. □

Suppose $\|\hat{b} - b\|_1 \leq \epsilon_b$, the value error is uniformly at most ϵ_Q, V_i is L_b-Lipschitz in belief. $|\widehat{V}_i(\hat{b}, a) - V_i(b, a)| \leq \epsilon_Q + L_b\epsilon_b =: d_V$. Let $\Delta_{ij}^{\mathrm{plug}}$ denote the corresponding plug-in regret on $\mathcal{D}_i$.

Theorem 4. *Suppose every nonsingleton support has a unique prior optimum with margin $m^- > 2d_V$. Then $|\Delta_{ij}^{\mathrm{plug}} - \Delta_{ij}^t| \leq 2d_V + 2Q_{\max}\delta_{\mathrm{cov}}$. Singleton supports require no margin condition.*

Proof. For the true prior maximiser a^- and any retained a, $\widehat{V}_i(\hat{b}, a^-) - \widehat{V}_i(\hat{b}, a) \geq m^- - 2d_V > 0$, so both evaluators commit to a^-. Since maximisation is 1-Lipschitz under uniform error,

$$\begin{cases} |\widehat{G}_{ij}^{\mathcal{D}} - G_{ij}^{\mathcal{D}}| \leq 2d_V, \\ |\Delta_{ij}^{\mathrm{plug}} - \Delta_{ij}^t| \leq |\Delta_{ij}^{\mathrm{plug}} - \Delta_{ij}^{\mathcal{D}}| + |\Delta_{ij}^{\mathcal{D}} - \Delta_{ij}^t| \\ \qquad\qquad\qquad \leq 2d_V + 2Q_{\max}\delta_{\mathrm{cov}}. \end{cases} \tag{16}$$

Without the margin condition, committed-action changes require separate control. □

Practical-surrogate discrepancy. If $|\widehat{\Delta}_{ij}^t - \Delta_{ij}^{\mathrm{plug}}| \leq \epsilon_{\mathrm{sur}}$ uniformly, then $|\widehat{\Delta}_{ij}^t - \Delta_{ij}^t| \leq \epsilon_{\mathrm{sur}} + 2\epsilon_Q + 2L_b\epsilon_b + 2Q_{\max}\delta_{\mathrm{cov}} =: \epsilon_{\mathrm{tot}}$. Here ϵ_{sur} absorbs mean-embedding, joint-optimisation, and entropy-weighting discrepancies; critic accuracy, belief calibration do not establish it.

Fixed-support maximisation noise. Condition on $\mathcal{I}$ fixing a support of size m, and suppose centred residuals satisfy $\mathbb{E}[e^{\lambda\zeta_a} \mid \mathcal{I}] \leq e^{\lambda^2\sigma^2/2}$. Then

$$\begin{cases} e^{\lambda\mathbb{E}[\max_a \zeta_a | \mathcal{I}]} \leq me^{\lambda^2\sigma^2/2}, \\ \mathbb{E}[\max_a \zeta_a \mid \mathcal{I}] \leq \inf_{\lambda > 0} \left\{ \dfrac{\log m}{\lambda} + \dfrac{\lambda\sigma^2}{2} \right\} = \sigma\sqrt{2\log m}. \end{cases} \tag{17}$$

Independence is unnecessary, but residual-dependent selection can violate conditional centring. This bounds only maximisation noise, not total score error.

B.6 Ranking Stability

Lemma 3. *On a common eligible set $\mathcal{N}_i$, suppose $\max_j |\hat{d}_j - d_j| \leq \epsilon$. If $S^\star, \widehat{S}$ maximise the respective size-K score sums, then their reference-sum gap lies in $[0, 2K\epsilon]$. The sets coincide when $d_{(K)} - d_{(K+1)} > 2\epsilon$, or when every teammate is selected.*

Proof. Perturbed-score optimality gives $\sum_{j \in S^\star} d_j \leq \sum_{j \in S^\star} \hat{d}_j + K\epsilon \leq \sum_{j \in \widehat{S}} \hat{d}_j + K\epsilon \leq \sum_{j \in \widehat{S}} d_j + 2K\epsilon$. Reference optimality gives the lower bound; the boundary margin preserves all selected–unselected pairwise orders. □

Student logits are ranking scores, not calibrated utilities. Ideal–Teacher comparison therefore requires a total-error bound. Under modularity, for $D_i \subseteq \mathcal{N}_i$,

$$F_i(\mathcal{N}_i) - F_i(\mathcal{N}_i \setminus D_i) = \sum_{j \in D_i} \Delta_{ij}^t \le \sum_{j \in D_i} \widehat{\Delta}_{ij}^t + |D_i|\epsilon_{\mathrm{tot}}. \tag{18}$$

Thus top practical scores minimise this fixed-budget bound on hypothetical evidence.

Theorem 5. *Consider teacher- and student-graph greedy policies over T decisions with common environment, initial distribution, controller, masks, and tie-breaks. At each sufficient pre-action history H_t, all non-graph inputs are shared. For actions $\mathbf{a}_{\mathrm{T}}, \mathbf{a}_{\mathrm{S}}$, assume $\ell_t(H_t) = Q_t^{\pi_{\mathrm{T}}}(H_t, \mathbf{a}_{\mathrm{T}}) - Q_t^{\pi_{\mathrm{T}}}(H_t, \mathbf{a}_{\mathrm{S}})$, $|\ell_t(H_t)| \le G_{\max}\mathbb{I}[E_{\mathrm{T}}(H_t) \ne E_{\mathrm{S}}(H_t)]$. Then $J(\pi_{\mathrm{T}}) - J(\pi_{\mathrm{S}}) \le G_{\max}\sum_{t=0}^{T-1}\gamma^t \Pr_{d_t^{\pi_{\mathrm{S}}}}(E_{\mathrm{T}} \ne E_{\mathrm{S}}) \le G_{\max}(\sum_{t=0}^{T-1}\gamma^t)\min\{1, \sum_{i=1}^n \epsilon_{\mathrm{S},i}\}$, where the second line assumes receiver-wise disagreement probabilities bounded by $\epsilon_{\mathrm{S},i}$ under student occupancy.*

Proof. Graph agreement implies identical payloads and actions, hence $\ell_t = 0$. Using teacher continuation and $V_T^{\pi_{\mathrm{T}}} = 0$,

$$\begin{aligned}\sum_{t=0}^{T-1}\gamma^t \mathbb{E}_{d_t^{\pi_{\mathrm{S}}}}[\ell_t] &= \mathbb{E}_{\pi_{\mathrm{S}}}\left[\sum_{t=0}^{T-1}\gamma^t\big(V_t^{\pi_{\mathrm{T}}}(H_t) - r_t - \gamma V_{t+1}^{\pi_{\mathrm{T}}}(H_{t+1})\big)\right] \\ &= J(\pi_{\mathrm{T}}) - J(\pi_{\mathrm{S}}).\end{aligned} \tag{19}$$

Applying the envelope and $\Pr(E_{\mathrm{T}} \ne E_{\mathrm{S}}) \le \min\{1, \sum_i \Pr(E_{\mathrm{T},i} \ne E_{\mathrm{S},i})\}$ under the same occupancy yields both bounds. □

C Discussion

C.1 What VIBE values.

Payoff influence, uncertainty, and the value of resolving uncertainty are distinct. A teammate may strongly affect reward yet already be predictable, or remain uncertain while every plausible action leaves the receiver's choice unchanged. Conversely, moderate coupling can still determine which action the receiver should take. Only the latter uncertainty has decision value. Accordingly, multiplying sensitivity by uncertainty is insufficient: both can be large while the receiver never crosses a decision boundary. Lemma 1 instead identifies the two relevant factors: the probability that evidence changes the receiver's decision and the value gained when it does.

C.2 The ideal target and the practical teacher.

The ideal Δ integrates over the receiver's belief, whereas the practical teacher evaluates a belief-mean embedding, maximises over joint probes, and applies entropy weighting. The exact categorical expectation removes sampling error from the mean embedding, not the gap between $\widehat{\Delta}$ and Δ. Moreover, entropy is a calibration weight, not the switching probability in Lemma 1, and probe selection can retain payoff effects unrelated to the receiver's action. Hence the controlled diagnostic tests the ideal criterion, while benchmark experiments test whether its tractable surrogate preserves enough of the useful ranking.

C.3 Centralised valuation, decentralised selection.

The teacher uses global state, teammate availability, and a centralised critic; it therefore serves only as a training-time labeller. The student uses receiver-local information and can reproduce only ranking distinctions recoverable from the receiver's own history. Selected edges create no new information: they determine which teammate features already present in the receiver's observation receive dedicated relational processing. Thus VIBE requires no additional inter-agent communication channel at execution, and any benefit must arise from better allocation of modelling capacity rather than privileged execution-time information.

Table 8: VIBE configuration.

Group	Configuration or task	Value / primary metric
Common learner settings		
Optimisation	Optimiser / learning rate	Adam / 10^{-3}
Targets	Discount / TD(λ)	$\gamma = 0.99$ / $\lambda = 0.6$
Exploration	ϵ start / finish / anneal	$1.0/0.05/500{,}000$ env. steps
Replay	Learner batch / capacity	128 / 5,000 episodes
Collection	Parallel environments	8
Training	Full-run interaction cap	10,050,000 env. steps
Target network	Synchronisation interval	200 learner updates
Backbone	Recurrent / mixer / hypernetwork widths	$64/32/64$
VIBE configuration		
Probes	Local proposals k / beam P / retained M	$2/32/8$
Recognition	POW tolerance C	0.1
Belief	Hidden widths	$64, 64$
Student	Hidden width	64
Ranking	Informative-row threshold / weight cap	10^{-8} / 10
Execution	Per-receiver edge budget K	1
Payload	Projection width / aggregation	64 / arithmetic mean
Channel scale	Frozen ranking / end-to-end	$\alpha = 0$ / $\alpha = 0.05$
Communication	Additional execution-time messages	None

D Experimental Methodology and Reproducibility

The evaluation separates three levels of evidence: task-level control performance, frozen-state recovery of the practical Teacher ranking, and matched end-to-end learning with the deployed relational pathway. Frozen-state diagnostics compare selectors on common states, whereas end-to-end runs may induce different state distributions.

Table 8 summarises the shared learner and VIBE-specific settings. Environment-specific task identities, episode limits, observations, rewards, action interfaces, and SMACv2 capability distributions are fixed by the corresponding resolved benchmark configuration.

Ablation definitions. All variants match VIBE in backbone, training and evaluation budgets, and primary edge budget. **w/o Belief** replaces receiver-conditioned beliefs with uniform distributions in both Teacher scoring and Student inputs. **w/o Pair** removes the matched with/without-evidence comparison and uses a single critic score. **w/o POW** drops the recogniser and forms probes directly from bounded local-action combinations at the same probe budget. **w/o Rank** removes ranking distillation and selects edges directly from Teacher scores. **w/o Dynamic** fixes the selected edges for an entire evaluation run. **w/o Payload** retains selection but removes selected teammate features from relational aggregation. **Oracle** selects the eligible teammate using privileged receiver-specific decision values, whereas **Dense** removes the edge budget and aggregates all eligible teammate features.

E Additional Experimental Results

This section provides supplementary evidence for the effectiveness, mechanism, and operating regime of VIBE. We first extend the benchmark coverage, then isolate the selector mechanism in the analytic CCT environment, examine how performance changes with the edge budget, and finally test across additional SMAC and intersection scenarios.

E.1 Extended Benchmark Results

The main experiments focus on the complete benchmark matrix to determine whether the same advantage persists against broader value-factorisation, representation, graph-learning, policy-learning, and partner-selection baselines. **Table 9** covers 10 tasks from five benchmark families using each task's registered metric and the same evaluation protocol as the main experiments.

Table 9: Additional-baseline results across ten tasks from five benchmark families.

Method	Matrix	Predator–Prey		SMACv1			SMACv2			Driving
	CCT	p=-4	p=-5	3s_vs_5z	Corr.	10m_vs_11m	Protoss	Terran	Zerg	Intersect.
ResQ	$16.74_{\pm 0.52}$	$40.31_{\pm 2.79}$	-	$99.00_{\pm 1.00}$	$69.00_{\pm 6.00}$	$40.32_{\pm 6.29}$	$17.94_{\pm 0.62}$	$16.32_{\pm 0.84}$	$17.27_{\pm 0.85}$	$88.20_{\pm 5.91}$
EMC	$15.65_{\pm 0.51}$	$18.40_{\pm 6.17}$	$9.42_{\pm 7.23}$	$98.27_{\pm 1.34}$	$66.97_{\pm 9.18}$	$17.14_{\pm 4.61}$	$16.95_{\pm 0.58}$	$17.67_{\pm 1.29}$	$17.07_{\pm 1.45}$	$76.45_{\pm 5.19}$
TRAMA	$16.93_{\pm 0.79}$	$30.46_{\pm 6.36}$	$12.96_{\pm 4.76}$	$99.38_{\pm 2.62}$	$79.88_{\pm 3.40}$	$72.56_{\pm 4.41}$	$17.79_{\pm 0.64}$	$17.60_{\pm 0.75}$	$17.75_{\pm 1.96}$	$81.45_{\pm 3.03}$
G2ANet	$15.39_{\pm 0.45}$	$20.97_{\pm 4.68}$	$5.40_{\pm 4.14}$	$92.64_{\pm 3.99}$	$48.18_{\pm 5.63}$	$6.56_{\pm 7.21}$	$16.82_{\pm 1.44}$	$16.59_{\pm 0.82}$	$16.20_{\pm 0.71}$	$76.91_{\pm 5.86}$
DICG	$16.32_{\pm 0.54}$	$32.23_{\pm 3.98}$	$24.69_{\pm 2.12}$	$94.28_{\pm 3.57}$	$52.53_{\pm 6.45}$	$8.01_{\pm 7.11}$	$17.15_{\pm 2.04}$	$17.57_{\pm 0.95}$	$17.05_{\pm 2.00}$	$85.25_{\pm 3.65}$
GraphMix	$15.92_{\pm 0.93}$	$26.68_{\pm 2.74}$	$11.26_{\pm 4.64}$	$97.91_{\pm 1.58}$	$65.20_{\pm 3.55}$	$14.29_{\pm 6.96}$	$17.59_{\pm 0.66}$	$17.68_{\pm 1.49}$	$17.23_{\pm 1.89}$	$78.18_{\pm 8.89}$
MAPPO	$15.64_{\pm 0.47}$	$23.29_{\pm 7.33}$	$8.75_{\pm 3.48}$	$93.36_{\pm 3.39}$	$21.94_{\pm 4.46}$	$4.47_{\pm 2.86}$	$16.93_{\pm 1.19}$	$17.70_{\pm 0.86}$	$16.59_{\pm 1.64}$	$80.27_{\pm 5.81}$
Variance-K	$19.12_{\pm 0.12}$	$38.82_{\pm 2.20}$	$38.57_{\pm 4.27}$	$100.00_{\pm 0.00}$	$76.30_{\pm 6.01}$	$71.24_{\pm 4.97}$	$19.32_{\pm 0.87}$	$18.94_{\pm 0.95}$	$18.35_{\pm 0.59}$	$90.04_{\pm 4.04}$
VIBE	$19.36_{\pm 0.23}$	$39.29_{\pm 2.77}$	$38.58_{\pm 3.39}$	$100.00_{\pm 0.00}$	$82.59_{\pm 3.68}$	$77.43_{\pm 2.62}$	$18.88_{\pm 0.78}$	$18.79_{\pm 0.67}$	$17.54_{\pm 1.13}$	$93.19_{\pm 2.00}$

Note. CCT, Predator–Prey, and SMACv2 report episodic return; SMACv1 reports win rate, and Driving Intersection reports success rate (%).

A \ B	B=0	B=1	Mean
A=0	3.50	-2.50	0.50
A=1	23.50	17.50	20.50
Mean	4.50	-1.50	1.50

(a) True payoff: receiver action 0

A \ B	B=0	B=1	Mean
A=0	0.00	0.00	0.00
A=1	0.00	0.00	0.00
Mean	0.00	0.00	0.00

(b) True payoff: receiver action 1

A \ B	B=0	B=1	Mean
A=0	0	**1**	0
A=1	0	0	0
Mean	0	**1**	0

(c) Optimal receiver action

A \ B	B=0	B=1	Mean
A=0	20.26	17.32	18.79
A=1	32.40	18.32	25.36
Mean	20.87	17.37	**19.12**

(d) Variance-K: 19.12 | hit 8.73%

A \ B	B=0	B=1	Mean
A=0	18.07	16.38	17.23
A=1	18.32	18.25	18.28
Mean	18.08	16.48	**17.28**

(e) CASEC: 17.28 | hit 15.64%

A \ B	B=0	B=1	Mean
A=0	18.45	16.64	17.54
A=1	19.28	18.08	18.68
Mean	18.49	16.71	**17.60**

(f) SOP-CG: 17.60 | hit 21.18%

A \ B	B=0	B=1	Mean
A=0	18.00	16.31	17.16
A=1	18.16	18.66	18.41
Mean	18.01	16.43	**17.22**

(g) APS: 17.22 | hit 17.31%

A \ B	B=0	B=1	Mean
A=0	19.86	17.83	18.84
A=1	30.80	27.51	29.16
Mean	20.41	18.31	**19.36**

(h) VIBE: 19.36 | hit 87.63%

A \ B	B=0	B=1	Mean
A=0	20.38	18.00	19.19
A=1	34.00	29.92	31.96
Mean	21.06	18.60	**19.83**

(i) Oracle: 19.83

Figure 3: **CCT payoff structure and selector outcomes.** (a)–(c) True payoffs and optimal receiver actions. (d)–(i) Condition-wise outcomes of Variance-K, CASEC, SOP-CG, APS, VIBE, and the privileged Oracle, with overall means and optimal-edge hit rates where applicable.

VIBE improves over CASEC and SOP-CG in all 10 reported columns. The gains are particularly clear on coordination-intensive settings such as CCT, Predator–Prey, SMACv1 `corridor` and `10m_vs_11m`, while remaining competitive on the stochastic SMACv2 races and Intersection. The architecture-matched comparison is consistent with this pattern: VIBE improves over Variance-K on most tasks despite sharing the same candidate interface, edge budget, and relational pathway.

E.2 CCT Mechanism Analysis

The CCT matrix game provides an exact setting in which the payoff tensor and receiver-optimal action are known. We use it to test whether each selector identifies the teammate whose uncertainty changes the receiver decision, rather than merely the teammate with large payoff variation. **Figure 3**(a–c) gives the true payoff structure and the corresponding optimal receiver action, while panels (d–i) report the condition-wise outcomes of representative selectors.

E.3 Edge-Budget Analysis

VIBE operates under a per-receiver edge budget K. We therefore vary the number of retained teammate edges from sparse selection toward the dense endpoint and evaluate both task performance

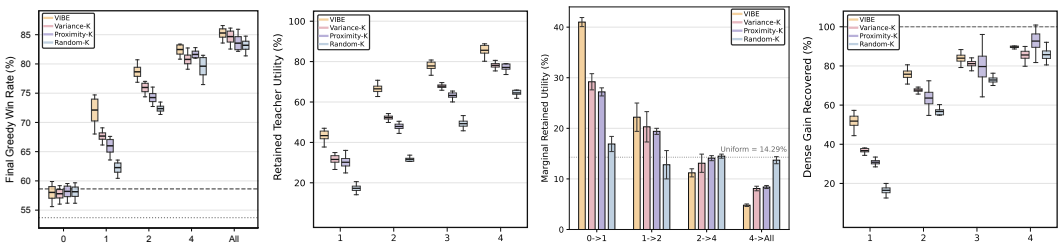

Figure 4: **Edge-budget analysis.** (a) Final win rate; (b) retained Teacher utility; (c) marginal retained utility; and (d) dense gain recovered.

Table 10: **Driving-specific outcomes on Intersection.**

Method	Success (%)	Return	Collision (%)	Deadlock (%)
POW-QMIX	$92.09_{\pm 2.70}$	$20.40_{\pm 0.74}$	$4.82_{\pm 0.61}$	$3.09_{\pm 0.44}$
Variance-K	$90.04_{\pm 4.04}$	$20.10_{\pm 0.82}$	$5.38_{\pm 0.72}$	$3.64_{\pm 0.53}$
CASEC	$87.54_{\pm 4.48}$	$19.62_{\pm 0.91}$	$6.51_{\pm 0.84}$	$4.21_{\pm 0.67}$
SOP-CG	$88.77_{\pm 5.99}$	$19.84_{\pm 0.95}$	$5.94_{\pm 0.82}$	$4.02_{\pm 0.60}$
VIBE	$\mathbf{93.19}_{\pm 2.00}$	$\mathbf{20.76}_{\pm 0.68}$	$\mathbf{4.15}_{\pm 0.56}$	$\mathbf{2.66}_{\pm 0.41}$

and Teacher-utility retention. **Figure 4** reports final win rate, retained Teacher utility, marginal retained utility, and the fraction of the within-method Dense gain recovered.

VIBE captures substantially more Teacher utility with the first few edges than Variance-K, Proximity-K, or Random-K. Its retained utility increases rapidly with K, while the marginal gain from each additional edge decreases. This pattern indicates that useful relational capacity is concentrated in a small subset of teammates rather than distributed uniformly across all available edges.

The performance results follow the same trend. Sparse VIBE configurations recover a large fraction of the gain obtained by the Dense endpoint, whereas increasing the budget further produces progressively smaller improvements. As K approaches the full-edge regime, selector differences narrow, which is expected because the edge criterion becomes less consequential when almost every teammate is retained. These results support the use of a small primary edge budget.

E.4 Supplementary experiments of SMAC

The main benchmark table mixes several coordination domains. To test whether VIBE's behaviour is also consistent within a single benchmark family, **Figure 5, 6** reports nine SMAC scenarios, including six SMACv1 tasks and the three SMACv2 race distributions. SMACv1 reports greedy win rate and SMACv2 reports episodic return.

VIBE reaches 100.00% on `3s_vs_5z`, 97.80% on `5m_vs_6m`, 95.37% on `8m_vs_9m`, and 77.43% on `10m_vs_11m`. It also remains competitive across the three SMACv2 races, with returns of 18.88, 18.79, and 17.54 on Protoss, Terran, and Zerg, respectively. The advantage is not uniform: on `MMM2`, other methods remain stronger. This is consistent with VIBE's scope: the benefit is largest when coordination depends on identifying a small number of decision-relevant teammate relations, while tasks with more uniformly useful interactions leave less room for selective edge allocation.

VIBE reaches high win rates earlier on several coordination-intensive scenarios, with the clearest separation on `corridor` and `10m_vs_11m`. On the easier `3s_vs_5z` task, several methods eventually approach saturation, but VIBE reaches this regime comparatively early. The advantage is less uniform on `MMM2`, `3s5z_vs_3s6z`, and `5m_vs_6m`, where competing methods exhibit larger overlap during training. Overall, the gains reported in the endpoint tables are accompanied by improved learning efficiency rather than arising solely from a favourable final checkpoint.

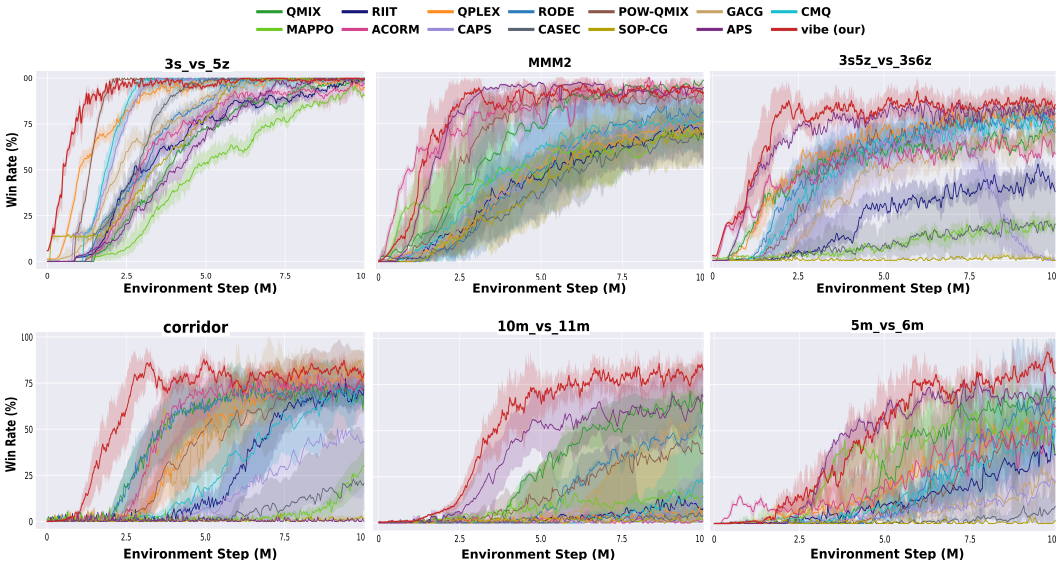

Figure 5: SMACv1 learning dynamics across six scenarios.

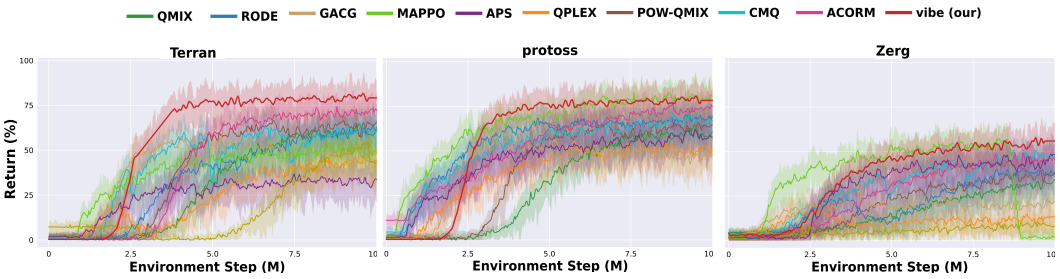

Figure 6: SMACv2 learning dynamics across three scenarios.

E.5 Supplementary experiments of Driving

Table 10 shows that VIBE achieves the highest safe-completion rate and episodic return among the compared methods, while recording the lowest collision and deadlock rates. Relative to POW-QMIX, VIBE raises safe completion from 92.09% to 93.19% and return from 20.40 to 20.76, while reducing collisions from 4.82% to 4.15% and deadlocks from 3.09% to 2.66%. Against the architecture-matched Variance-K selector, VIBE improves safe completion by 3.15 percentage points and reduces collisions by 1.23 percentage points. These results connect decision-relevant edge selection to both successful passage and safer coordination at the intersection.

F Algorithmic Details and Implementation

The implementation has two components: shared-probe Teacher target construction and receiver-local execution within the four-stage protocol. The Teacher uses the masked belief b_{ij}^{T} and exact mean embedding; the Student uses the unmasked b_{ij}; selected teammate features are averaged.

F.1 Shared Probes and Teacher Targets

For each replay transition, the candidate builder uses local utilities, availability masks, Q_r, and (k, P, M, C) to construct a nonempty ordered probe set $\mathcal{C}_t$ with $|\mathcal{C}_t| \leq M$, shared across all receiver–teammate pairs. For each (i, j), b_{ij} is renormalised over teammate j's available actions to obtain b_{ij}^{T}, and $\bar{\mathbf{e}}_{ij} = \sum_{a_j} b_{ij}^{\mathrm{T}}(a_j)\,\mathrm{emb}(a_j)$ is computed exactly. The two critic branches share context,

Algorithm 1: Shared-Probe Teacher Targets for a Replay Batch

Input: Replay batch $\mathcal{B}$; frozen $\hat{Q}$, Q_r, and g_ψ; local utilities and action masks; action embeddings; (k, P, M, C).
Output: Detached Teacher score rows, labels, and ranking weights.

Function TEACHERTARGETS($\mathcal{B}$)**:**
- Initialize target collection $\mathcal{T}$
- **foreach** *valid transition* $t \in \mathcal{B}$ **do**
  - Build the bounded top-k proposal beam
  - Apply action availability and obtain $\varnothing \neq \mathcal{C}_t$ with $|\mathcal{C}_t| \leq M$
  - **foreach** *ordered pair* (i, j) *with* $j \neq i$ **do**
    - Compute b_{ij} and b^{T}_{ij} by Eq. 3
    - Compute the exact categorical mean embedding $\bar{\mathbf{e}}_{ij} = \sum_{a_j} b^{\mathrm{T}}_{ij}(a_j)\,\mathrm{emb}(a_j)$
    - **foreach** $\mathbf{a} \in \mathcal{C}_t$ **do**
      - Evaluate $\hat{Q}^{+}(\mathbf{a})$ using the probe embedding of teammate j
      - Evaluate $\hat{Q}^{-}(\mathbf{a})$ after replacing only that embedding by $\bar{\mathbf{e}}_{ij}$
    - Select $\hat{\mathbf{a}}^{+}$ and $\hat{\mathbf{a}}^{-}$ using the common deterministic tie rule
    - Compute $\widehat{\Delta}^{t}_{ij}$ using teacher-score
  - **foreach** *receiver* i **do**
    - Compute $(j^{\star}_i, w_i)$ using Eq. 6
    - **if** $w_i > 10^{-8}$ **then**
      - Append $(t, i, j^{\star}_i, w_i)$ and the aligned Teacher-score row to $\mathcal{T}$
  - Normalize retained w_i within the minibatch and clip at 10
  - Detach all Teacher scores, labels, and ranking weights
  - **return** $\mathcal{T}$

probe order, all other embeddings, and parameters; only teammate j's embedding is replaced in the without-evidence branch. Thus teacher-score requires no evidence-sampling loop.

Receiver-wise labels $j^{\star}_i$ and weights w_i follow ranking-label. Rows with $w_i \leq 10^{-8}$ are discarded; retained weights are minibatch-normalised, clipped at 10, and detached together with Teacher scores before the Student update.

F.2 Receiver-Local Selection and Training

The belief encoder maps the $d_h + 2n$ receiver-representation and identity inputs through two width-64 ReLU layers to the teammate action alphabet. The Edge Student also uses width-64 hidden layers but receives the distinct local input in tudent-input. The payload encoder ϕ is a width-64 affine–ReLU map, and the recurrent backbone has hidden width 64.

At execution, the Student scores valid non-self teammates and selects the top-K. The relational payload is $m^t_i = \sum_{j \in E^t_i} \phi(o^t_{i \to j}) / \max(1, |E^t_i|)$, with no Student-score softmax applied to aggregation. The receiver's original observation remains available to the backbone, so $\alpha = 0$ disables only the dedicated relational pathway.

Training follows four stages: (I) train the common value backbone; (II) freeze it and calibrate g_ψ; (III) freeze backbone, Teacher, and belief encoder and distil f_η; (IV) restart from the shared checkpoint and train the graph-conditioned controller for a matched continuation budget with $\alpha > 0$. These objectives are stage-specific, not jointly active.

F.3 Resource Accounting and Numerical Conventions

One probe set $\mathcal{C}_t$ is shared across ordered pairs, while each pair has its own b^{T}_{ij}, mean embedding, and without-evidence branch. Empty selections give $m^t_i = 0$. Action availability is applied before action selection, and Teacher-only teammate masks never reach the Student.

Algorithm 2: Receiver-Local Execution and Four-Stage VIBE Training

Input: Backbone and target networks; g_ψ, f_η, and ϕ; replay buffer; edge budget K; stage budgets; channel scale α.

LocalAction$(i, h_i^t, o_i^t, \alpha)$
Obtain the pre-action recurrent representation r_i^t
foreach *eligible non-self teammate j in the receiver-local input* **do**
    Compute the unmasked receiver-local belief $b_{ij} = \mathrm{softmax}\, g_\psi([r_i^t, \mathbf{1}_i, \mathbf{1}_j])$
    Construct x_{ij}^t
    Compute the Student logit $s_{ij}^t = f_\eta(x_{ij}^t)$
Select the K highest-scoring valid teammates as E_i^t
Compute the receiver-local relational payload $m_i^t = \sum_{j \in E_i^t} \phi(o_{i\to j}^t)/\max(1, |E_i^t|)$
Evaluate local utilities from $[o_i^t, r_i^t, \alpha m_i^t]$
Apply the receiver's own action-availability mask
Choose the active exploration action during training, or the greedy action during evaluation
return *a_i^t and the recurrent continuation state*

TrainVIBE
Stage I: Backbone training. Train the value-factorisation backbone, recogniser Q_r, and labelling critic $\hat{Q}$; Save the common reference checkpoint
Stage II: Belief calibration. Freeze all non-belief modules; Train g_ψ by cross-entropy on aligned recorded teammate actions; Save and freeze the calibrated belief checkpoint
Stage III: Ranking distillation. Freeze all modules except the Edge Student f_η
foreach *ranking minibatch* **do**
    Obtain detached Teacher targets using Algorithm 1
    Update f_η with Eq. 7 on informative rows
Evaluate Student rankings on fixed trajectories with $\alpha = 0$
Stage IV: Matched graph-conditioned learning. Initialize the graph-conditioned controller from the common reference checkpoint
foreach *interaction until the fixed continuation budget is exhausted* **do**
    Execute LocalAction with $\alpha > 0$
    Append the resulting transition to replay
    if *a learner update is due* **then**
        Sample recurrent replay data
        Construct the standard masked temporal-difference targets
        Update the graph-conditioned controller using the backbone temporal-difference objective
        if *the target-update interval is reached* **then**
            Synchronize target networks
    if *an evaluation checkpoint is reached* **then**
        Run the common greedy-evaluation protocol
return *trained controller and evaluation records*

Let $m_t = |\mathcal{C}_t| \le M$. Paired Teacher scoring evaluates two critic branches for every ordered pair and probe: $N_{\mathrm{eval}}(t) = 2n(n-1)m_t \le 2n(n-1)M$, excluding proposal construction and exact embedding averaging. Hence $T_{\mathrm{teacher}} = T_{\mathrm{proposal}} + T_{\mathrm{belief}} + T_{\mathrm{embedding}} + \mathcal{O}(n(n-1)Mc_Q)$, where c_Q is one probe evaluation in one critic branch. No evidence-sample factor appears because the categorical mean embedding is exact; chunking affects peak memory, not the evaluation count.

At execution, the centralised critic and recogniser are absent. For a homogeneous n-agent team, $T_{\mathrm{exec}} = \mathcal{O}(n(n-1)(c_b + c_s) + n^2 \log(K+1) + nKd_{\mathrm{ally}}d_h)$, where c_b and c_s are one pairwise belief and Student evaluation. For $K = 1$, top-K selection reduces to a linear scan per receiver. These are operation counts, not latency claims; empirical accounting should report Teacher-side training and receiver-local execution costs separately.