SEGUE: BRIDGING SYMBOLIC COORDINATION AND EMBODIED EXECUTION FOR SEQUENTIAL SOCIAL DILEMMAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon coordination in embodied shared-resource environments requires agents to follow collective strategies while adapting to local observations, execution constraints, and failures. Symbolic solvers such as Value-Set Iteration (VSI) can compute effective high-level coordination policies, but their abstract models do not specify how recommendations should be grounded under changing embodied conditions. We propose **SEGUE** (**S**ymbolic-to-**E**mbodied **G**uidance for **U**ncertain **E**xecution), which treats private VSI recommendations as revisable guidance and grounds action type and quantity in each agent's current context. We further introduce **ESD-Bench**, a closed-loop benchmark that evaluates the full symbolic-to-embodied process from recommendation and program generation to execution and resource transition. In free-action evaluation, SEGUE achieves welfare of 8.80 versus 8.20 for the strongest baseline in the Balanced setting, achieves the best observed welfare of 5.80 under Scarcity, and executes 100% of requested actions and quantities in both settings. It also matches symbolic VSI welfare under aligned dynamics and maintains 100% joint feasibility under action-set perturbations. Ablations show that the gains stem from current coordination guidance rather than additional context or a particular backbone.

1 INTRODUCTION

Real-world multi-agent problems often involve shared resources that benefit multiple agents but deteriorate under individual over-exploitation. As illustrated in Figure 1, extracting more resources provides immediate private gains, yet repeated exploitation can reduce regeneration, accelerate environmental degradation, or even trigger irreversible resource collapse. Sustaining the resource requires restraint or maintenance investment, both of which typically yield lower immediate rewards. This creates a fundamental tension between short-term individual incentives and long-term collective welfare, underlying a broad class of social dilemmas such as common-pool resource management (Dawes, 1980; Hardin, 1968; Ostrom, 1990).

The challenge becomes substantially harder when such coordination must be sustained through embodied agents. These problems are inherently *sequential*: every joint action changes the future resource state, rewards, and available actions. They are also naturally *infinite-horizon*: agents interact continuously without a predefined terminal state. Most existing benchmarks and methods study only parts of this setting. Early work often imposes finite horizons or artificial termination (Leibo et al., 2017; Pérolat et al., 2017; Leibo et al., 2021), while recent methods such as Value-Set Iteration (VSI) Gan & Majumdar (2025) address infinite-horizon coordination through discounted correlated equilibria. However, these formulations largely abstract away the execution layer: a symbolic recommendation is assumed to translate reliably into the intended action. In embodied settings, this assumption can break down because the symbolic state is an abstraction of the environment, the available action set may differ from that assumed by the coordinator, and the agent may fail to execute a recommended action exactly. Consequently, a coordination strategy that is optimal at the symbolic level may produce poor outcomes after deployment.

To study this symbolic-to-embodied deployment gap, we introduce **ESD-Bench**, a closed-loop Minecraft benchmark for multi-agent social dilemmas with dynamic shared resources. In the *Cleanup* task, agents obtain immediate private rewards by harvesting resources, while cleaning incurs an immediate cost but reduces pollution and supports future regeneration. ESD-Bench explicitly

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.

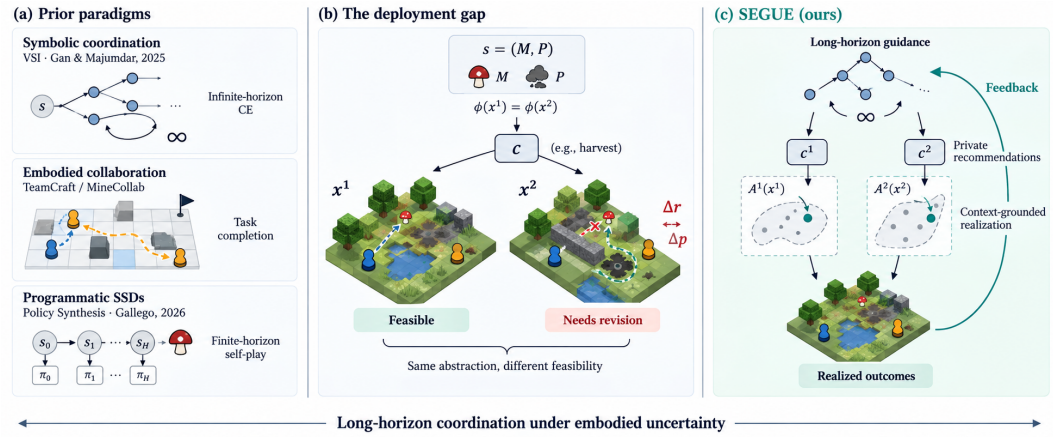


Figure 1: Long-horizon coordination under embodied uncertainty. (a) Prior paradigms emphasize symbolic coordination, embodied collaboration, or finite-horizon social dilemmas. (b) Physically distinct states can share symbolic representation yet differ in action feasibility. (c) SEGUE treats private symbolic recommendations as revisable guidance, grounds actions in local context, and feeds realized outcomes back into coordination.

traces the complete decision-to-outcome process, from a symbolic recommendation to the intended action, generated program, environmental execution, and resulting resource transition. This enables us to evaluate not only whether agents coordinate effectively, but also whether coordination decisions remain executable and effective after grounding in the environment.

We then propose **SEGUE** (**S**ymbolic-to-**E**mbodied **G**uidance for **U**ncertain **E**xecution), a centralized-guidance, decentralized-execution framework that separates long-horizon coordination from local action grounding. At each turn, the environment is mapped to a compact symbolic state, which VSI uses together with interaction history to generate a joint high-level recommendation. Each agent receives only its private component of this recommendation and independently grounds it using local observations, task rules, interaction history, and execution feedback. Rather than treating the recommendation as a hard command, SEGUE interprets it as *guidance*: an agent may revise the recommended action or quantity when the recommendation is incompatible with its current execution context. The resulting programs are executed in parallel, and only the realized outcomes are used to update the shared resource state and symbolic history. An independent LLM judge converts execution outcomes into structured feedback for subsequent decisions. In this way, VSI provides the global coordination signal, while decentralized LLM agents bridge the gap between symbolic guidance and executable behavior.

This separation allows SEGUE to preserve the long-horizon coordination capability of symbolic solvers while remaining robust to uncertainty at execution time. Across our evaluations, SEGUE achieves the highest mean welfare in the Balanced setting (8.80 versus 8.20 for the strongest baseline), matches the best-performing method in the Scarce setting (5.80), and successfully executes all requested actions and quantities in both settings. Under action-set perturbations, SEGUE maintains 100% joint feasibility and consistently achieves higher welfare than an LLM-based mediator across the tested perturbation rates. Ablation studies further show that the current, state-aligned VSI recommendation is the key coordination signal, while SEGUE consistently outperforms a context-only LLM baseline across DeepSeek, GPT-4o, and Qwen. These results suggest that effective multi-agent coordination requires not only a good long-horizon policy, but also an explicit mechanism for grounding symbolic coordination in uncertain execution environments.

Our contributions are summarized as follows:

- We propose **SEGUE**, a symbolic-to-embodied coordination framework that combines long-horizon VSI guidance with decentralized, context-grounded execution under local execution uncertainty.
- We formalize the **symbolic-to-embodied deployment gap** and derive a discounted value bound that characterizes how reward and transition errors affect long-horizon performance.
- We introduce **ESD-Bench**, a closed-loop Minecraft benchmark that evaluates multi-agent coordination jointly from symbolic recommendations to executable programs and realized resource dynamics.

2 Related Work

Sequential social dilemmas and symbolic coordination. Sequential social dilemmas model how current actions affect future resources and incentives. Prior work studies common-pool resource use, inequity aversion, social influence, and evaluation suites such as Melting Pot (Leibo et al., 2017; Pérolat et al., 2017; Hughes et al., 2018; Jaques et al., 2019; Leibo et al., 2021). These settings capture delayed resource effects through predefined action interfaces. VSI computes approximately optimal history-dependent correlated equilibria in discounted infinite-horizon stochastic games (Gan & Majumdar, 2025), but assumes predefined symbolic states and actions. We use VSI as the strategic coordinator and study how its recommendations are grounded into executable behavior under physical constraints.

Coordination frameworks for multi-agent LLMs. Multi-agent LLM methods differ in where coordination occurs and what information is shared. Existing systems use centralized mediators, structured interaction topologies, or negotiation and plan sharing (Zhu et al., 2025; Tewolde et al., 2026; White et al., 2025). SEGUE combines centralized VSI guidance with decentralized action selection and execution. VSI samples a joint recommendation, while each agent receives only its private component and uses local context to select an action and quantity and generate a program.

Evaluation environments for social and embodied agents. SOTOPIA, MultiAgentBench, and CoopEval emphasize social reasoning and coordination (Zhou et al., 2024; Zhu et al., 2025; Tewolde et al., 2026). AgentBench and embodied benchmarks emphasize task completion and collaboration (Liu et al., 2024; Fan et al., 2022; Long et al., 2024; White et al., 2025; Liu et al., 2026), while Policy Synthesis evaluates LLM-generated programs in Cleanup-like social dilemmas (Gallego, 2026). ESD-Bench combines dynamic shared resources, program execution, and a discounted infinite-horizon objective. It traces the recommendation-to-outcome pipeline to distinguish coordination failures from execution failures. Table 1 compares evaluation settings.

Table 1: Comparison of evaluation settings.

Setting	MA	LLM	Emb.	SR	PA	IH
SSD/CPR	✓	–	✓	✓	–	–
Melting Pot	✓	–	✓	△	–	–
SOTOPIA	✓	✓	–	–	–	–
MultiAgentBench	✓	✓	–	–	–	–
CoopEval	✓	✓	–	–	–	–
MineDojo	–	–	✓	–	–	–
AgentBench	–	✓	△	–	△	–
TeamCraft	✓	✓	✓	–	–	–
MineCollab	✓	✓	✓	–	–	–
MECoBench	✓	✓	✓	–	–	–
Policy Synthesis	✓	✓	✓	✓	✓	–
ESD-Bench	✓	✓	✓	✓	✓	✓

MA: multi-agent; Emb.: embodied interaction; SR: dynamic shared resource; PA: program-generated actions; IH: discounted infinite-horizon objective. ✓: explicit support; △: partial or task-dependent support; –: not a primary focus.

3 Methods

3.1 Preliminaries

Problem Formulation. We consider a multi-agent correlated-equilibrium (CE) problem in an infinite-horizon stochastic game (von Stengel & Forges, 2008). Here, n agents interact with a central coordinator. At each time step, the game continues with probability γ and terminates with probability $1-\gamma$. The coordinator maximizes the expected long-term task reward, but agents may deviate from its recommendations to improve their own payoffs. Recommendations must therefore be both valuable for the task and incentive-compatible. The game is $\mathcal{G} = \langle S, \mathbf{A}, p, \mathbf{r}, \gamma \rangle$, where S is the state space used by the coordinator, $\mathbf{A} = A_1 \times \cdots \times A_n$ is the joint action space of the agents, $p(\cdot \mid s, \mathbf{a})$ is the transition kernel, and $\mathbf{r} = (r_0, r_1, \ldots, r_n)$ is the reward vector. Here r_0 denotes the coordinator's task reward, and r_i denotes the reward of agent i.

At each time step t, the coordinator privately recommends an action $a_i^t \in A_i$ to each agent i. The joint recommendation is $\mathbf{a}^t = (a_1^t, \ldots, a_n^t) \in \mathbf{A}$. If all agents follow the recommendation, the realized joint action is $\mathbf{a}^t$, and the environment transitions according to $s^{t+1} \sim p(\cdot \mid s^t, \mathbf{a}^t)$, while player i receives reward $r_i(s^t, \mathbf{a}^t)$. A history-dependent coordinator policy π maps an interaction history to a distribution over joint recommendations. Let Σ denote the set of possible histories, and let $\pi(\cdot \mid \sigma, s) \in \Delta(\mathbf{A})$ be the recommendation distribution at history $\sigma \in \Sigma$ and current state s. Since the probability of reaching round t is γ^t, the expected discounted value of player i under

policy π is

$$V_i^\pi(s) = \mathbb{E}_{\pi,p}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s^t, \mathbf{a}^t) \mid s^0 = s\right]. \tag{1}$$

For compactness, Eq. (1) suppresses the history argument when no deviation is considered; the continuation values in the Q-functions below keep the realized history explicit.

Let $\rho_i : \Sigma \times S \times A_i \to A_i$ be a deviation plan for agent i. Given a recommended joint action $\mathbf{a}$, define $\mathbf{a}^{\rho_i} = \mathbf{a} \oplus_i \rho_i(\sigma, s, a_i)$, where $\oplus_i$ replaces only the i-th component of $\mathbf{a}$. If agent i deviates, the realized joint action becomes $\mathbf{a}^{\rho_i}$. The discounted Q-values for following the recommendation and deviating are, respectively,

$$Q_i^\pi(\sigma, s, \mathbf{a}) = r_i(s, \mathbf{a}) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,\mathbf{a})} V_i^\pi(\sigma, s, \mathbf{a}, \mathbf{a}; s'), \tag{2}$$

$$Q_i^{\pi,\rho_i}(\sigma, s, \mathbf{a}) = r_i(s, \mathbf{a}^{\rho_i}) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,\mathbf{a}^{\rho_i})} V_i^{\pi,\rho_i}(\sigma, s, \mathbf{a}, \mathbf{a}^{\rho_i}; s'), \tag{3}$$

where s' is the next state. A policy π is an exact CE if, for every agent i, every history-state pair (σ, s), every recommendation $\alpha \in A_i$, and every deviation plan ρ_i, following the recommendation yields at least as much expected discounted value as deviating:

$$\sum_{\mathbf{a}: a_i = \alpha} \pi(\mathbf{a} \mid \sigma, s) Q_i^\pi(\sigma, s, \mathbf{a}) \geq \sum_{\mathbf{a}: a_i = \alpha} \pi(\mathbf{a} \mid \sigma, s) Q_i^{\pi,\rho_i}(\sigma, s, \mathbf{a}). \tag{4}$$

Thus, conditioned on receiving recommendation α, agent i cannot improve its discounted value by deviating. Let $\mathcal{C}$ denote the set of exact CE policies. The coordinator's optimal CE problem is

$$\pi^* \in \arg\max_{\pi \in \mathcal{C}} V_0^\pi(s_{\text{init}}), \tag{5}$$

where s_{init} is the initial state.

The Optimal CE Problem. Solving Eq. (5) exactly is intractable for general stochastic games. We therefore use Value-Set Iteration (VSI) (Gan & Majumdar, 2025) as the theoretical symbolic benchmark. VSI solves the same discounted stochastic-game setting on the finite symbolic model $\mathcal{G}_{\text{VSI}} = \langle S_{\text{VSI}}, \mathbf{A}, p_{\text{VSI}}, \mathbf{r}_{\text{VSI}}, \gamma \rangle$, where S_{VSI} is a finite symbolic state space, p_{VSI} is the symbolic transition model, $\mathbf{r}_{\text{VSI}}$ is the symbolic reward matrix, and γ is the same continuation probability as in the embodied game.

VSI does not enumerate a separate policy for every history. Instead, it characterizes the set of CE-inducible continuation values at each symbolic state. Let $V^\star(s) \subseteq \mathbb{R}^{n+1}$ be the set of value vectors inducible at $s \in S_{\text{VSI}}$. For a candidate value-set function V, VSI defines the operator

$$\Phi(V)(s) = \Big\{ \mathbb{E}_{\mathbf{a} \sim \bar{\pi}} \left[\mathbf{r}_{\text{VSI}}(s, \mathbf{a}) + \gamma \mathbb{E}_{s' \sim p_{\text{VSI}}(\cdot|s,\mathbf{a})} w(\mathbf{a}, \mathbf{a}; s') \right] \\ \mid (\bar{\pi}, w) \in \mathcal{F}_s(V) \Big\}. \tag{6}$$

Here $\bar{\pi} \in \Delta(\mathbf{A})$ is the current recommendation distribution, and $w(\mathbf{a}, \mathbf{a}^{\rho_i}; s') \in \mathbb{R}^{n+1}$ is the continuation value after recommendation $\mathbf{a}$, realized action $\mathbf{a}^{\rho_i}$, and next state s'. The notation $\mathbf{a}^{\rho_i}$ follows Eq. (3): it denotes the joint action obtained when one agent deviates from the recommendation. The feasible set $\mathcal{F}_s(V)$ contains pairs $(\bar{\pi}, w)$ that satisfy the symbolic counterpart of the IC constraint in Eq. (4) and the onward-value condition

$$w(\mathbf{a}, \mathbf{a}^{\rho_i}; s') \in V(s') \quad \forall i,\ \mathbf{a} \in \mathbf{A},\ \rho_i,\ s' \in S_{\text{VSI}}. \tag{7}$$

Gan & Majumdar (2025) show that the inducible value-set function is the greatest fixed point of this monotone operator, $V^\star = \Phi(V^\star)$. Thus, VSI provides a principled optimality benchmark under exact symbolic state, reward, and transition models.

3.2 Limitations of VSI in Embodied Environments

VSI characterizes optimality in abstract stochastic games, but transferring its policies to embodied multi-agent systems is complicated by its finite abstraction of the physical state space.

In embodied settings, computing the symbolic reward matrix becomes increasingly costly as the numbers of modalities and agents grow. Information such as images, depth, and terrain states cannot

be effectively encoded in reward signals, creating a gap between embodied environments and VSI's assumptions. Let

$$\phi : S_{\text{real}} \to S_{\text{VSI}} \tag{8}$$

be the encoder from real embodied states to symbolic VSI states. For readability, write a real state as $x = (s, \eta)$ when $s = \phi(x)$, where η collects unmodeled residual information. Since S_{VSI} is finite while S_{real} may be continuous or extremely large, many physically different states can map to the same symbolic state.

This abstraction induces reward residuals:

$$r_i^{\text{real}}(x, \mathbf{a}) = r_i^{\text{VSI}}(\phi(x), \mathbf{a}) + \delta_i(x, \mathbf{a}), \tag{9}$$

where δ_i captures the part of the real reward that is not represented by the symbolic reward matrix. It also induces transition mismatch. The real environment defines a transition over physical states, while VSI uses a transition over symbolic states. The real transition pushed through the encoder is

$$\widetilde{p}(s' \mid x, \mathbf{a}) = \Pr_{x' \sim p_{\text{real}}(\cdot \mid x, \mathbf{a})}[\phi(x') = s']. \tag{10}$$

The transition error is the discrepancy between $\widetilde{p}$ and p_{VSI}.

Theorem 1 (Discounted Symbolic-to-Real Gap). Let π_{VSI} be a policy computed in the symbolic VSI model and deployed in the real environment through the encoder ϕ. For any real state x, let $V_{\text{real}}^{\pi_{\text{VSI}}}(x)$ denote its discounted coordinator value in the real embodied environment, and let $V_{\text{VSI}}^{\pi_{\text{VSI}}}(\phi(x))$ denote its discounted value in the symbolic VSI model from the matched symbolic state. Assume that $|r_0^{\text{real}}(x, \mathbf{a})|$ and $|r_0^{\text{VSI}}(s, \mathbf{a})|$ are bounded by $R_{\max}$. Define

$$\Delta_r = \sup_{x \in S_{\text{real}},\, \mathbf{a} \in \mathbf{A}} \left| r_0^{\text{real}}(x, \mathbf{a}) - r_0^{\text{VSI}}(\phi(x), \mathbf{a}) \right|, \tag{11}$$

$$\Delta_p = \sup_{x \in S_{\text{real}},\, \mathbf{a} \in \mathbf{A}} \text{TV}\left(\widetilde{p}(\cdot \mid x, \mathbf{a}), p_{\text{VSI}}(\cdot \mid \phi(x), \mathbf{a}) \right), \tag{12}$$

where TV denotes total-variation distance. Then

$$|V_{\text{real}}^{\pi_{\text{VSI}}}(x) - V_{\text{VSI}}^{\pi_{\text{VSI}}}(\phi(x))| \le \frac{\Delta_r}{1-\gamma} + \frac{2\gamma R_{\max}}{(1-\gamma)^2} \Delta_p. \tag{13}$$

Therefore, A larger γ extends the expected interaction horizon but amplifies the worst-case effects of reward and transition mismatch through the factors $(1-\gamma)^{-1}$ and $(1-\gamma)^{-2}$, respectively, as characterized by Theorem 1.

3.3 Context-Grounded VSI Coordination

Private symbolic guidance. At turn t, SEGUE encodes the physical state x_t as $s_t = \phi(x_t)$ and maintains the symbolic interaction history σ_t. VSI then samples a joint high-level recommendation

$$c_t \sim \pi_{\text{VSI}}(\cdot \mid \sigma_t, s_t), \qquad c_t = (c_t^1, \ldots, c_t^n). \tag{14}$$

Each agent observes only its private component c_t^i. The recommendation provides a long-horizon strategic role, but it is not an executable command because the symbolic model omits local geometry, reachability, program validity, and physical interference.

Context-grounded realization. Agent i grounds its private recommendation using

$$z_t^i = (D, o_t^i, j_{t-1}^i, h_t^i, m_{t-1}^i), \tag{15}$$

where D is the task specification, o_t^i is the current local observation, j_{t-1}^i is the previous judge feedback, h_t^i is the available interaction history, and m_{t-1}^i is the reflection memory. The agent planner produces

$$u_t^i = P_{\theta_i}(z_t^i, s_t, c_t^i) = (a_t^i, q_t^i, \psi_t^i), \tag{16}$$

where a_t^i, q_t^i, and ψ_t^i denote the intended action, requested quantity, and executable program. The agent may revise the action type or quantity when the recommendation is incompatible with current execution constraints.

Under the free-action protocol, each agent independently scores candidate action–quantity pairs using its private recommendation c_t^i and available context, selects a pair, and generates an executable program. The selected action may differ from c_t^i.

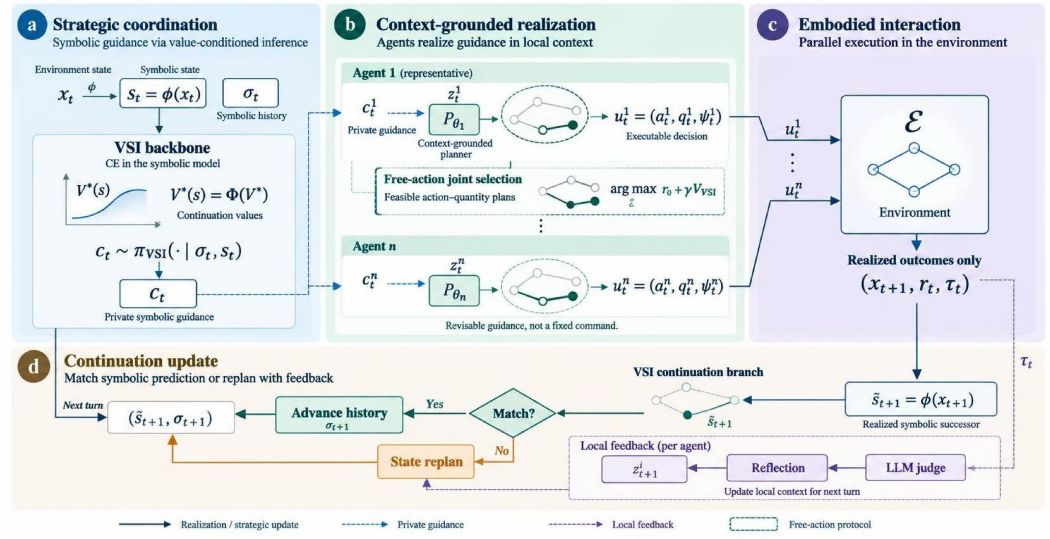


Figure 2: SEGUE combines centralized VSI guidance with decentralized action selection and execution. VSI samples a joint recommendation and sends each agent only its private component. Each agent uses this recommendation and local context to select an action and quantity and generate a program. Programs execute in parallel; realized outcomes update symbolic history and agent memory.

Execution and feedback. The generated programs execute in parallel in the shared environment:

$$(x_{t+1}, r_t, \tau_t) \sim \mathcal{E}(\cdot \mid x_t, u_t^1, \ldots, u_t^n), \tag{17}$$

where τ_t records program validity, completed action types and quantities, and execution duration. Rewards and successor states depend only on completed actions and realized quantities. An independent LLM judge structures execution feedback but does not determine rewards or transitions.

Each agent reflects on the feedback, and the resulting memory becomes part of z_{t+1}^i. SEGUE advances the symbolic history along the VSI continuation branch if its predicted successor matches $\phi(x_{t+1})$; otherwise, it replans from the realized symbolic state.

Scope of the coordination guarantee. The correlated-equilibrium guarantees of VSI apply to the symbolic game G_{VSI}, not to the complete embodied policy after grounding and execution. SEGUE therefore uses VSI for long-horizon strategic coordination and relies on the context-grounded LLM loop to handle information outside the symbolic abstraction, as shown in Figure 2.

4 ESD-BENCH

Task and interface. ESD-Bench evaluates embodied coordination in Minecraft *Cleanup*, where harvesting mushrooms yields immediate reward, whereas cleaning pollution supports regeneration. The symbolic state $s_t = (M_t, P_t)$ records available resources and pollution but omits execution-relevant details such as geometry and reachability. Methods provide an intended action, requested quantity, and executable program for each agent at each turn.

Execution and traces. Programs execute in parallel under a per-turn time budget. The environment determines rewards and transitions from realized execution; an LLM judge only structures feedback. Turn-level traces record symbolic states, recommendations when available, intended and completed actions and quantities, programs, validity, durations, rewards, and successor states. These traces support diagnosis of decision and execution failures.

Evaluation protocol. We evaluate a discounted infinite-horizon objective through geometric stopping. After each completed turn, a rollout continues with probability γ and terminates with probability $1 - \gamma$. The undiscounted return before termination is an unbiased estimator of the discounted infinite-horizon return (Appendix B.1), and the run is the unit of evaluation.

Evaluation metrics. **Social Welfare** measures total realized reward. **Physical Success (PS)** measures whether attempted macro-actions complete, **Count Fulfillment (CF)** measures the fraction of requested quantities that are realized, and **Joint Feasibility (JF)** measures whether both selected actions satisfy the current action constraints. **Recommendation Match** measures agreement between the selected joint action and the VSI recommendation. Thus, Welfare and Match primarily characterize coordination, while PS, CF, and JF characterize execution reliability. Full metric definitions are provided in Appendix B.

5 EXPERIMENTS

We evaluate SEGUE from four complementary perspectives and address the following questions:

- **RQ1:** Does SEGUE improve end-to-end embodied social welfare?
- **RQ2:** Does symbolic coordination survive embodied execution?
- **RQ3:** Can SEGUE adapt when the symbolic and embodied action spaces mismatch?
- **RQ4:** What drives the performance gain, and is it robust across LLM backbones?

5.1 EXPERIMENTAL SETUP

We consider two initial resource regimes: Balanced, $s_0 = (6, 6)$, and Scarce, $s_0 = (2, 2)$. After each completed turn, an episode terminates with probability $1/3$, matching the default VSI discount factor $\gamma = 2/3$. Each condition contains five paired runs, with methods sharing the same termination, recommendation, and, when applicable, perturbation seeds. DeepSeek Chat is used as the default backbone, while GPT-4o and Qwen3.6 Plus are used for cross-backbone evaluation. We report means in the main paper and provide standard deviations and additional diagnostics in Appendix A.

We evaluate two execution protocols. The *free-action* protocol evaluates coordination and execution end to end by letting each agent independently select an action type, quantity, and executable program. The *controlled protocol* fixes one macro-action per agent per turn to isolate coordination transfer through embodied execution. All embodied LLM methods use the same Voyager planning and execution interface (Wang et al., 2024).

We compare SEGUE with four LLM-based baselines. *LLM-Context* removes the coordination recommendation while retaining the task rules, current state, and execution interface. *Myopic Grounding* replaces long-horizon coordination with immediate feasible reward. *Negotiated Contract* follows the negotiation-based coordination strategy of Yocum et al. (2023), while *LLM Mediator* uses a centralized LLM planner following Chen et al. (2024). Controlled experiments additionally include *Symbolic VSI*, which directly executes the VSI policy (Gan & Majumdar, 2025).

5.2 FREE-ACTION EMBODIED COORDINATION

RQ1: Does SEGUE improve welfare without sacrificing execution reliability? We first evaluate the least constrained and most demanding setting, in which agents independently select their action types, quantities, and executable programs. Table 2 reports the result.

Table 2: Free-action embodied results. PS and CF are percentages. Bold marks the best result in each resource regime, including ties. Uncited baselines are controlled variants defined in this work.

	Balanced $(6, 6)$			Scarce $(2, 2)$		
Method	Welfare ↑	PS ↑	CF ↑	Welfare ↑	PS ↑	CF ↑
LLM-Context	7.80	94.2	92.5	**5.80**	83.9	82.8
Myopic Grounding	8.20	**100.0**	**100.0**	**5.80**	**100.0**	**100.0**
Negotiated Contract (Yocum et al., 2023)	8.00	96.0	95.5	5.40	79.7	79.6
LLM Mediator (Chen et al., 2024)	7.40	93.5	93.1	5.60	96.8	97.2
SEGUE	**8.80**	**100.0**	**100.0**	**5.80**	**100.0**	**100.0**

In the Balanced state, SEGUE obtains the highest welfare, improving over the strongest baseline by 0.60 and over LLM-Context by 1.00. It also completes every requested action and quantity. This

combination matters: the welfare gain does not come from proposing ambitious actions that fail during execution. Instead, SEGUE turns long-horizon guidance into choices that remain executable.

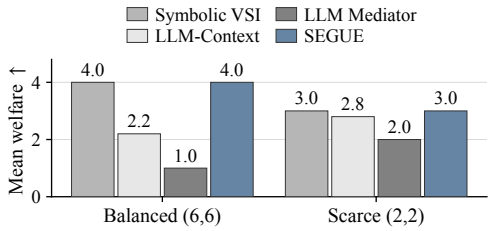
Under scarcity, SEGUE reaches the best observed welfare, matching LLM-Context and Myopic Grounding, while maintaining 100% PS and CF as shown in Table 2. Its execution is also more reliable than LLM-Context and the two communication-based baselines. Taken together with the Balanced results, these findings show that SEGUE combines high welfare with reliable embodied execution across resource regimes, with its clearest welfare gains emerging when multiple feasible short-term choices differ in their downstream value.

5.3 Controlled transfer

RQ2: Does symbolic coordination survive embodied execution? We next isolate the symbolic-to-embodied transfer problem by fixing one macro-action per agent at each turn. Table 3(a) compares SEGUE with Symbolic VSI, which directly executes the symbolic policy.

Table 3: Controlled transfer and action-set mismatch (five paired runs). SEGUE matches Symbolic VSI welfare under aligned dynamics and preserves welfare while improving execution reliability under mismatch. PS and CF coincide in (b); execution metrics are percentages.

(a) Aligned dynamics



(b) Action-set mismatch

Scarce $(2, 2)$, $\rho = 0.25$

Method	Welfare ↑	PS/CF ↑	JF ↑
Symbolic VSI	**4.00**	88.5	95.0
LLM-Context	3.60	92.2	**100.0**
LLM Mediator	2.80	**100.0**	**100.0**
SEGUE	**4.00**	100.0	**100.0**

Under aligned dynamics, SEGUE matches Symbolic VSI in mean welfare and recommendation match across the two resource regimes (Table 3(a) and Appendix Table 8). Relative to LLM-Context, SEGUE improves welfare by 1.80 in Balanced and 0.20 in Scarce. Because SEGUE and LLM-Context share the same embodied planning and execution interface, this comparison isolates the contribution of the current VSI guidance. The exact agreement with Symbolic VSI further indicates that the long-horizon coordination signal can survive program generation and embodied execution without measurable welfare loss when the symbolic and embodied dynamics are aligned.

5.4 Robustness to Action-Set Mismatch

RQ3: Can SEGUE adapt when the embodied action space differs from the symbolic model? We next introduce action-set perturbations that are invisible to the symbolic coordinator. At each turn, with probability ρ, one uniformly sampled agent–action pair is made unavailable. The current constraint is visible to embodied LLM agents, while Symbolic VSI continues to select its original recommendation; if the disabled pair is selected, execution becomes a hard no-op with zero executed count. Table 3(b) reports results for $\rho = 0.25$ in the Scarce regime.

SEGUE achieves welfare of 4.00, matching Symbolic VSI while improving PS and CF from 88.5% to 100.0% and JF from 95.0% to 100%. The welfare tie does not mean that the perturbation failed: Symbolic VSI requests 25 harvesting units and completes 20, whereas SEGUE requests 22 and also completes 20. SEGUE therefore preserves the welfare of symbolic coordination while improving execution reliability under action-set mismatch. Moreover, the separate Balanced rate sweep in Appendix Table 14 shows a welfare advantage over LLM Mediator across the tested perturbation probabilities.

LLM Mediator achieves 100% PS, CF, and JF but welfare of only 2.80, completing 14 harvesting and 16 cleaning actions. Feasible, fully completed actions therefore do not guarantee high welfare.

5.5 Ablations and Cross-Backbone Robustness

RQ4: What drives the coordination gain? We first remove or modify the coordination information provided to the embodied agents (Table 4(a)). In the aligned Balanced setting, rules, state,

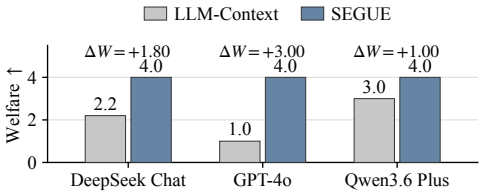
and history alone achieve welfare between 2.00 and 2.60, while a stale recommendation provides little additional benefit. In contrast, the current VSI recommendation increases welfare to 4.00 and achieves 100% recommendation match. These results indicate that the gain comes from a state-aligned coordination signal rather than simply from providing additional context.

Table 4: Aligned Balanced ablations (five paired runs). (a) DeepSeek Chat information. (b) Cross-backbone welfare. SEGUE has 100% PS and Match across backbones.

(a) Coordination information

Condition	Welfare	Match
Rules only	2.00	26.7%
State only	2.20	36.7%
History	2.60	43.3%
Stale recommendation	2.20	13.3%
Current SEGUE	**4.00**	**100.0%**

(b) LLM backbone



Is the gain backbone-independent? We evaluate SEGUE with DeepSeek Chat, GPT-4o, and Qwen3.6 Plus. SEGUE achieves welfare of 4.00 across all three backbones (Table 4(b)), improving over LLM-Context by 1.80, 3.00, and 1.00, respectively. This consistency suggests that the benefit arises primarily from the coordination interface rather than model-specific action preferences.

5.6 Long-Horizon Returns under Continuing Interaction

We evaluate SEGUE with continuous resource regeneration over 100 turns as a finite-window approximation to continuing deployment. We compare the full continuous configuration with variants removing continuation value or duration awareness, using five paired runs per method in each resource regime.

Table 5: Continuous-regeneration results. W_{60s}: harvest in the first 60 active seconds; W_{100t}: harvest over 100 turns; Min.: mean per-run minimum individual 100-turn harvest. Bold marks the best mean among the reported variants.

	Scarce $(2, 2)$			Balanced $(6, 6)$		
Method	W_{60s}	W_{100t}	Min.	W_{60s}	W_{100t}	Min.
SEGUE w/o continuation value	34.0	133.0	64.0	31.2	130.6	61.8
SEGUE w/o duration awareness	38.2	157.2	**73.8**	34.8	162.4	70.4
SEGUE	**39.2**	**160.2**	72.4	**43.2**	**174.0**	**81.0**

SEGUE achieves the highest mean aggregate harvest under both budgets in both regimes. Its 100-turn return exceeds the continuation-free variant by 20.5% and 33.2%, and the duration-unaware variant by 1.9% and 7.1%, in Scarce and Balanced, respectively. This supports combining future resource value with execution-aware grounding. The duration-unaware variant retains a slight advantage in minimum individual harvest in Scarce.

6 Conclusion

We propose SEGUE, a framework that bridges long-horizon symbolic coordination with context-grounded embodied execution, and ESD-Bench, a closed-loop benchmark spanning program generation, physical execution, and resource transitions. Across free-action, aligned-transfer, and action-set perturbation settings, SEGUE achieves strong welfare while maintaining reliable execution, attaining the best welfare in Balanced, matching the best result in Scarce, preserving Symbolic VSI welfare under aligned dynamics, and maintaining 100% Physical Success, Count Fulfillment, and Joint Feasibility across all tested perturbation rates. Ablations further show that these gains stem from current, state-aligned coordination guidance rather than additional context or a specific LLM backbone. Future work will explore richer multimodal state abstractions, online correction of reward and transition mismatch from execution feedback, and coordination objectives that jointly optimize welfare, resource preservation, and execution robustness, as well as larger teams and more diverse sequential social dilemmas.

AI USE STATEMENT

During manuscript preparation, we used generative AI tools for minor language editing of selected passages to improve grammar, clarity, and conciseness. The authors reviewed all AI-assisted edits and take full responsibility for the final manuscript.

REFERENCES

Yongchao Chen, Jacob Arkin, Yang Zhang, Nicholas Roy, and Chuchu Fan. Scalable multi-robot collaboration with large language models: Centralized or decentralized systems? In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4311–4317. IEEE, 2024. doi: 10.1109/ICRA57147.2024.10610676. URL https://doi.org/10.1109/ICRA57147.2024.10610676.

Robyn M. Dawes. Social dilemmas. *Annual Review of Psychology*, 31:169–193, 1980. doi: 10.1146/annurev.ps.31.020180.001125. URL https://doi.org/10.1146/annurev.ps.31.020180.001125.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. MineDojo: Building open-ended embodied agents with internet-scale knowledge. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18343–18362, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/74a67268c5cc5910f64938cac4526a90-Abstract-Datasets_and_Benchmarks.html. Datasets and Benchmarks Track.

Víctor Gallego. Cooperation and exploitation in llm policy synthesis for sequential social dilemmas. *arXiv preprint arXiv:2603.19453*, 2026.

Jiarui Gan and Rupak Majumdar. Value-set iteration: Computing optimal correlated equilibria in infinite-horizon multi-player stochastic games. *arXiv preprint arXiv:2506.07186*, 2025. doi: 10.48550/arXiv.2506.07186. URL https://arxiv.org/abs/2506.07186.

Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968. doi: 10.1126/science.162.3859.1243. URL https://doi.org/10.1126/science.162.3859.1243.

Edward Hughes, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, Heather Roff, and Thore Graepel. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/7fea637fd6d02b8f0adf6f7dc36aed93-Abstract.html.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z. Leibo, and Nando de Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3040–3049. PMLR, 2019. URL https://proceedings.mlr.press/v97/jaques19a.html.

Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 464–473. International Foundation for Autonomous Agents and Multiagent Systems, 2017. doi: 10.5555/3091125.3091194. URL https://www.ifaamas.org/Proceedings/aamas2017/pdfs/p464.pdf.

Joel Z. Leibo, Edgar A. Dueñez-Guzman, Alexander Vezhnevets, John P. Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with melting pot. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6187–6199. PMLR, 2021. URL https://proceedings.mlr.press/v139/leibo21a.html.

Qingyun Liu, Jiwen Zhang, Jingyi Hu, Siyuan Wang, and Zhongyu Wei. MECoBench: A systematic study of multimodal agent collaboration in embodied environments. *arXiv preprint arXiv:2606.31966*, 2026.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zAdUB0aCTQ.

Qian Long, Zhi Li, Ran Gong, Ying Nian Wu, Demetri Terzopoulos, and Xiaofeng Gao. Teamcraft: A benchmark for multi-modal multi-agent systems in minecraft. *arXiv preprint arXiv:2412.05255*, 2024.

Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, 1990. doi: 10.1017/CBO9780511807763. URL https://doi.org/10.1017/CBO9780511807763.

Julien Pérolat, Joel Z. Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/2b0f658cbffd284984fb11d90254081f-Abstract.html.

Emanuel Tewolde, Xiao Zhang, David Guzman Piedrahita, Vincent Conitzer, and Zhijing Jin. Coopeval: Benchmarking cooperation-sustaining mechanisms and llm agents in social dilemmas. *arXiv preprint arXiv:2604.15267*, 2026.

Bernhard von Stengel and Françoise Forges. Extensive-form correlated equilibrium: Definition and computational complexity. *Mathematics of Operations Research*, 33(4):1002–1022, 2008. doi: 10.1287/moor.1080.0340. URL https://doi.org/10.1287/moor.1080.0340.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=nfx5IutEed.

Isadora White, Kolby Nottingham, Ayush Maniar, Max Robinson, Hansen Lillemark, Mehul Maheshwari, Lianhui Qin, and Prithviraj Ammanabrolu. Collaborating action by action: A multi-agent llm framework for embodied reasoning. *arXiv preprint arXiv:2504.17950*, 2025.

Julian Yocum, Phillip J. K. Christoffersen, Mehul Damani, Justin Svegliato, Dylan Hadfield-Menell, and Stuart Russell. Mitigating generative agent social dilemmas. In *NeurIPS 2023 Workshop on Foundation Models for Decision Making*, 2023. URL https://openreview.net/forum?id=5TIdOk7XQ6.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. MultiAgentBench: Evaluating the collaboration and competition of LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8580–8622. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.421. URL https://aclanthology.org/2025.acl-long.421/.

A EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

A.1 PROTOCOL AND RUN INTEGRITY

We evaluate two initial resource regimes: Balanced, $s_0 = (6, 6)$, and Scarce, $s_0 = (2, 2)$. Each method–configuration–state condition contains five paired runs. Within a geometric-stopping comparison, methods share termination and recommendation seeds and, where applicable, perturbation seeds. After each completed joint turn, the rollout terminates with probability $1/3$, corresponding to the default VSI discount factor $\gamma = 2/3$.

Table 6 summarizes 215 unique reported runs. The reported set comprises 165 runs from the original geometric-stopping archive, 20 runs from the separate Scarce perturbation extension, and 30 runs from the continuous-regeneration cohort. Rows reused across result tables contribute only once to this total.

Table 6: Reported experiment coverage. B and S denote Balanced and Scarce. Run counts include all listed states and count each reused run once.

Study	States	Conditions counted	Runs
Free action	B, S	5 methods $\times$ 5 runs per state	50
Aligned transfer	B, S	4 methods $\times$ 5 runs per state	40
Context ablation	B	3 added conditions $\times$ 5 runs	15
Cross-backbone	B	2 added backbones $\times$ 2 methods $\times$ 5 runs	20
Perturbation comparison	B	4 methods at $\rho = 0.25 \times 5$ runs	20
Additional rate-sweep conditions	B	2 methods $\times$ 2 rates $\times$ 5 runs	20
Scarce perturbation extension	S	4 methods at $\rho = 0.25 \times 5$ runs	20
Continuous regeneration	B, S	3 variants $\times$ 5 runs per state	30
Total reported			215

Archive accounting. The original geometric-stopping archive contains 230 validated runs and 695 completed turns; 165 of these runs enter the reported comparisons. Its audit found no missing or duplicate valid jobs. Two invalid candidate attempts remain archived but are excluded from aggregation; their conditions were completed using the original seeds and configurations. The continuous-regeneration results report three variants, accounting for 30 runs from a separate 40-run cohort.

Backbones and private guidance. DeepSeek Chat (`deepseek-chat`) is the default backbone. The cross-backbone evaluation additionally uses GPT-4o (`gpt-4o`) and Qwen3.6 Plus (`qwen3.6-plus`). In private-guidance conditions, each agent receives only its own component of the VSI recommendation. All embodied LLM methods use the same Voyager planning and execution interface described in the main text.

Action selection and parallel execution. The free-action protocol lets each agent independently select an action type, requested quantity, and executable program. The controlled protocol fixes one macro-action per agent per turn to isolate coordination transfer through embodied execution. Both agents make their decisions from the same pre-turn state, and their committed programs execute in parallel in the shared environment under a 60-second per-turn budget. Rewards and successor states depend on completed actions and realized quantities. The independent LLM judge structures execution feedback; it does not assign rewards or determine resource transitions. The geometric-stopping draw occurs only after a completed joint turn. Continuous-regeneration evaluation also uses parallel execution, but applies the active-time and fixed-turn budgets described in Appendix A.6.

A.2 FREE-ACTION RESULTS WITH DISPERSION

Table 7 supplements Table 2 with welfare dispersion and execution diagnostics. PS and CF are computed within each run and then averaged equally across the five runs, following Appendix B.

In Balanced, SEGUE achieves the highest mean welfare of 8.80, compared with 8.20 for Myopic Grounding and 7.80 for LLM-Context, while attaining 100% PS and CF. The paired welfare differences between SEGUE and Myopic Grounding are $[1, 1, 0, 0, 1]$, yielding a mean gain of 0.60,

Table 7: Free-action results over five paired runs per method and regime. Welfare is mean ± sample standard deviation; PS and CF are mean run-level percentages. Negotiated Contract follows Yocum et al. (2023), and LLM Mediator follows Chen et al. (2024).

	Balanced $(6, 6)$			Scarce $(2, 2)$		
Method	Welfare	PS	CF	Welfare	PS	CF
LLM-Context	7.80 ± 1.48	94.2%	92.5%	5.80 ± 1.92	83.9%	82.8%
Myopic Grounding	8.20 ± 1.92	100.0%	100.0%	5.80 ± 1.92	100.0%	100.0%
Negotiated Contract	8.00 ± 1.73	96.0%	95.5%	5.40 ± 1.52	79.7%	79.6%
LLM Mediator	7.40 ± 2.07	93.5%	93.1%	5.60 ± 1.95	96.8%	97.2%
SEGUE	8.80 ± 1.92	100.0%	100.0%	5.80 ± 1.92	100.0%	100.0%

three wins, and two ties. In Scarce, SEGUE ties LLM-Context and Myopic Grounding at the best observed mean welfare of 5.80 and maintains 100% PS and CF. The Balanced welfare gain therefore coexists with full action and quantity completion, while the Scarce result combines the highest observed welfare with reliable execution.

A.3 Controlled transfer and perturbation results

Aligned transfer. Table 8 expands the aligned comparison in main-text Table 3(a). SEGUE matches Symbolic VSI in mean welfare and VSI recommendation Match in both regimes: welfare is 4.00 in Balanced and 3.00 in Scarce, with 100% Match. Relative to LLM-Context, the welfare gains are 1.80 and 0.20, respectively. Both SEGUE and Symbolic VSI have PS of 100% in Balanced and 91.7% in Scarce.

Table 8: Aligned transfer over five paired runs per method and regime. Welfare is mean ± sample standard deviation; percentage metrics are run-level means. Match measures agreement with the VSI recommendation. A dash denotes an unreported VSI-referenced value. Symbolic VSI follows Gan & Majumdar (2025); LLM Mediator follows Chen et al. (2024).

	Balanced $(6, 6)$			Scarce $(2, 2)$		
Method	Welfare	PS	Match	Welfare	PS	Match
Symbolic VSI	4.00 ± 2.00	100.0%	100.0%	3.00 ± 1.00	91.7%	100.0%
LLM-Context	2.20 ± 0.84	100.0%	36.7%	2.80 ± 1.30	90.0%	50.0%
LLM Mediator	1.00 ± 1.00	100.0%	–	2.00 ± 1.00	100.0%	–
SEGUE	4.00 ± 2.00	100.0%	100.0%	3.00 ± 1.00	91.7%	100.0%

Match is consistently defined against the VSI recommendation. The available LLM Mediator diagnostic instead measures adherence to its own plan; it is kept separate from VSI Match rather than entered in the same column. Mediator-plan adherence is 100% in both aligned regimes and in the Balanced perturbation comparison below.

Action-set perturbations. At each turn, with probability ρ, one uniformly sampled agent–action pair is made unavailable. This constraint is visible to embodied LLM agents but not to the symbolic coordinator. Symbolic VSI retains its original recommendation; selecting the disabled pair produces a hard no-op with zero executed quantity. JF evaluates whether both selected actions satisfy the current availability constraint.

In Balanced at $\rho = 0.25$, SEGUE ties Symbolic VSI for the highest mean welfare of 4.80 and increases PS, CF, and JF from 95.0% to 100%. Relative to LLM-Context, its paired welfare differences are $[0, 2, 2, 1, 2]$, giving four wins and one tie. Relative to LLM Mediator, the differences are $[1, 2, 3, 1, 4]$, giving five wins. These comparisons distinguish preserving the coordination value of VSI from adhering to every symbolic recommendation when current constraints change.

Scarce perturbation extension. Table 10 supplies dispersion for the Scarce comparison in main-text Table 3(b). The five stopping horizons are $[5, 2, 4, 1, 3]$, and the shared perturbation schedule contains two triggered perturbations across 15 turns per method. Replaying this schedule across

Table 9: Action-set perturbations in Balanced at $\rho = 0.25$ over five paired runs. Welfare is mean $\pm$ sample standard deviation; percentages are run-level means. Match uses the VSI reference, with a dash for an unreported value. JF requires both selected actions to satisfy the current availability constraint.

Method	Welfare	PS	CF	Match	JF
Symbolic VSI	4.80 ± 1.79	95.0%	95.0%	100.0%	95.0%
LLM-Context	3.40 ± 1.95	100.0%	100.0%	49.3%	100.0%
LLM Mediator	2.60 ± 1.52	100.0%	100.0%	–	100.0%
SEGUE	4.80 ± 1.79	100.0%	100.0%	95.0%	100.0%

methods provides paired comparisons under the same disturbances, rather than additional independent perturbation schedules. Aligned runs use different stopping seeds, so aligned and perturbed mean welfare values are not paired before/after measurements.

Table 10: Scarce action-set perturbations at $\rho = 0.25$, supplementing main-text Table 3(b). Welfare is mean $\pm$ sample standard deviation over five paired runs; percentages are equally averaged over runs. PS and CF coincide in this comparison.

Method	Welfare	PS	CF	JF
Symbolic VSI	4.00 ± 1.58	88.5%	88.5%	95.0%
LLM-Context	3.60 ± 1.67	92.2%	92.2%	100.0%
LLM Mediator	2.80 ± 1.79	100.0%	100.0%	100.0%
SEGUE	4.00 ± 1.58	100.0%	100.0%	100.0%

SEGUE attains mean welfare of 4.00, matching Symbolic VSI while increasing PS and CF from 88.5% to 100% and JF from 95.0% to 100%. LLM Mediator also attains 100% PS, CF, and JF, completing all 14 harvesting and 16 cleaning units it requests, but achieves lower mean welfare of 2.80. Feasible, fully completed decisions therefore do not by themselves ensure high welfare; the coordination signal determines which executable actions are selected.

A.4 Context and backbone ablations

Table 11 reports information ablations in the aligned Balanced setting. Rules, state, and history alone yield mean welfare between 2.00 and 2.60, and a stale recommendation yields 2.20. The current VSI recommendation raises mean welfare to 4.00 and achieves 100% Match. These results identify current, state-aligned coordination guidance, rather than additional context alone, as the source of the gain in this comparison.

Table 11: Information ablation in the aligned Balanced setting. Welfare is mean $\pm$ sample standard deviation over five paired runs; PS and Match are mean run-level percentages. Current SEGUE and State only reuse the corresponding rows of Table 8. Match uses the current VSI recommendation.

Condition	Welfare	PS	Match
Rules only	2.00 ± 1.41	100.0%	26.7%
State only	2.20 ± 0.84	100.0%	36.7%
History	2.60 ± 1.52	100.0%	43.3%
Stale recommendation	2.20 ± 1.30	100.0%	13.3%
Current SEGUE	4.00 ± 2.00	100.0%	100.0%

Table 12 evaluates the same coordination interface across three backbones. SEGUE attains mean welfare of 4.00 with DeepSeek Chat, GPT-4o, and Qwen3.6 Plus, improving over LLM-Context by 1.80, 3.00, and 1.00, respectively. PS and VSI recommendation Match are 100% for SEGUE with each tested backbone.

Table 12: Cross-backbone results in the aligned Balanced setting. Welfare is mean ± sample standard deviation over five paired runs. DeepSeek Chat reuses the corresponding aligned runs. SEGUE PS and VSI recommendation Match are mean run-level percentages.

Backbone	LLM-Context	SEGUE	Δ Welfare	SEGUE PS / Match
DeepSeek Chat	2.20 ± 0.84	4.00 ± 2.00	$+1.80$	100.0% / 100.0%
GPT-4o	1.00 ± 1.00	4.00 ± 2.00	$+3.00$	100.0% / 100.0%
Qwen3.6 Plus	3.00 ± 1.00	4.00 ± 2.00	$+1.00$	100.0% / 100.0%

A.5 Perturbation-rate sensitivity

Table 13 reports SEGUE's execution reliability across the Balanced rate sweep, and Table 14 reports the corresponding welfare comparison. SEGUE maintains 100% PS, CF, and JF at every tested rate. Its mean welfare exceeds LLM Mediator by 2.20, 2.20, and 1.60 at $\rho = 0.25$, 0.50, and 0.75, respectively. The $\rho = 0.25$ results reuse the four-method comparison in Table 9; the two higher rates compare SEGUE and LLM Mediator only.

Table 13: SEGUE execution reliability across the Balanced perturbation-rate sweep. Entries are mean run-level percentages over five paired runs at each rate. These are the same SEGUE conditions used in Table 14, not additional runs.

ρ	PS	CF	JF
0.25	100.0%	100.0%	100.0%
0.50	100.0%	100.0%	100.0%
0.75	100.0%	100.0%	100.0%

Table 14: Two-method perturbation-rate sensitivity in Balanced. Welfare is mean ± sample standard deviation over five paired runs. The $\rho = 0.25$ rows reuse Table 9; the $\rho = 0.50$ and 0.75 rows are distinct rate-sweep conditions.

ρ	LLM Mediator	SEGUE	Δ Welfare
0.25	2.60 ± 1.52	4.80 ± 1.79	$+2.20$
0.50	2.20 ± 1.30	4.40 ± 1.52	$+2.20$
0.75	2.40 ± 1.34	4.00 ± 2.00	$+1.60$

A.6 Long-Horizon Returns under Continuing Interaction

Planning variants and execution budget. The continuous-regeneration evaluation in Table 5 observes interaction over 100 turns as a finite-window approximation to continuing deployment. We compare the full continuous configuration, SEGUE, with SEGUE w/o continuation value and SEGUE w/o duration awareness. SEGUE includes both continuation value and duration awareness. The continuation-free variant removes continuation value while retaining duration awareness; the duration-unaware variant removes duration awareness while retaining continuation value. The duration-unaware variant is a planning ablation within the same frozen execution build. We use five paired runs per method in each resource regime, giving 30 reported runs from the 40-run continuous-regeneration cohort.

All runs cover at least 60 seconds of active execution. The active clock excludes planning and paused intervals. Navigation does not dig unassigned blocks, while explicit target digging and resource-attribution checks remain enabled. Historical continuous runs and failed attempts are archived separately and are not pooled into the reported cohort.

Evaluation endpoints. The fixed-time endpoint $W_{60\mathrm{s}}$ counts actual harvest during the first 60 active seconds. The fixed-turn endpoint $W_{100\mathrm{t}}$ sums harvest over the first 100 turns. Min. is the mean per-run minimum individual 100-turn harvest: we take the lower individual harvest within

each run and then average across the five runs. Active execution time and turn count are distinct budgets; the run is the unit of replication, and its 100 turns are not independent samples.

Mean-return comparisons. SEGUE achieves the highest mean aggregate harvest under both budgets in both regimes. Its mean 100-turn return is 160.2 in Scarce and 174.0 in Balanced, exceeding the continuation-free variant (133.0 and 130.6) by 20.5% and 33.2%, and the duration-unaware variant (157.2 and 162.4) by 1.9% and 7.1%, respectively. These observed mean-return gains support combining future resource value with execution-aware grounding. The duration-unaware variant retains a slight advantage in mean minimum individual harvest in Scarce (73.8 versus 72.4 for SEGUE). In Balanced, SEGUE also achieves the highest mean minimum individual harvest (81.0).

Paired uncertainty and scope. For SEGUE minus SEGUE w/o continuation value, the paired mean W_{60s} differences are 5.2 in Scarce and 12.0 in Balanced. The corresponding exploratory Student-t 95% confidence intervals are $[-4.671, 15.071]$ and $[-0.874, 24.874]$, respectively. Both intervals include zero. These intervals concern the fixed-active-time comparison with the continuation-free variant and use five paired runs per regime without multiplicity correction. The results describe observed mean-return gains over the tested finite window; they do not establish a reliable fixed-active-time gain over the continuation-free variant or indefinite sustainability.

B EVALUATION MEASURES

Let T_r be the number of completed turns in run r, $r_{i,t}$ the realized reward of agent i at turn t, $q_{i,t}$ its requested quantity, and $\widehat{q}_{i,t}$ its completed quantity. The notation $\widehat{q}_{i,t}$ distinguishes execution counts from the private VSI recommendation c_t^i. Each condition reports the arithmetic mean of five run-level metrics. Welfare dispersion is the sample standard deviation, using denominator $5-1$. Percentage metrics are not pooled across runs; the run is the unit of replication.

Social welfare. Run-level social welfare is total realized reward:

$$W_r = \sum_{t=0}^{T_r-1} \sum_{i=1}^{n} r_{i,t}. \tag{18}$$

Rewards are determined by the environment from realized execution, not by the proposed program or the LLM judge. Harvesting provides immediate private reward, while cleaning changes pollution and subsequent resource regeneration. The continuous-regeneration endpoints explicitly measure realized harvest under the active-time and fixed-turn budgets specified in Appendix A.6.

Physical Success. Let $\mathcal{A}_r = \{(i,t) : q_{i,t} > 0\}$ be the positive-quantity macro-action attempts in run r. Let $S_{i,t}$ indicate whether an attempted macro-action completes in the execution record. Physical Success is the fraction of completed macro-action attempts, excluding explicit zero-quantity waits:

$$\mathrm{PS}_r = \frac{1}{|\mathcal{A}_r|} \sum_{(i,t)\in\mathcal{A}_r} S_{i,t}. \tag{19}$$

Thus PS evaluates completion at the level of action attempts, while CF evaluates the quantities realized by those attempts.

Count Fulfillment. Count Fulfillment is the fraction of requested quantity that is physically completed:

$$\mathrm{CF}_r = \frac{\sum_{t=0}^{T_r-1} \sum_{i=1}^{n} \widehat{q}_{i,t}}{\sum_{t=0}^{T_r-1} \sum_{i=1}^{n} q_{i,t}}. \tag{20}$$

The PS and CF ratios are defined for runs with positive-quantity attempts. A blocked action with a positive request and zero execution contributes an unsuccessful attempt to PS and an unfulfilled request to CF. The reported Scarce controlled-perturbation comparison has equal PS and CF, as shown in Table 10.

Recommendation Match. For the selected joint action $\mathbf{a}_t$ and current logged VSI recommendation $\mathbf{c}_t$,

$$\mathrm{Match}_r = \frac{1}{T_r} \sum_{t=0}^{T_r-1} \mathbb{I}[\mathbf{a}_t = \mathbf{c}_t]\,. \tag{21}$$

Action order and agent identities are preserved; quantities are not included. All entries labeled Match use this VSI reference. Adherence to an LLM mediator's own plan is a separate diagnostic, not a substitute for agreement with VSI. The corresponding VSI-referenced entries for LLM Mediator are therefore left unreported in Tables 8 and 9.

Joint Feasibility. Let $F_{r,t} = 1$ when both selected actions satisfy the current action constraints and 0 otherwise:

$$\mathrm{JF}_r = \frac{1}{T_r} \sum_{t=0}^{T_r-1} F_{r,t}. \tag{22}$$

In the action-set perturbation evaluations, JF concerns the injected availability constraint; physical completion is evaluated separately by PS and CF. One available action and one blocked action give joint feasibility zero, not the individual-availability average of $1/2$.

B.1 Geometric stopping

Under geometric stopping, an episode always executes its first turn and continues after each completed turn with probability $\gamma = 2/3$. The termination draws are independent of the continuing trajectory, so $\Pr(T_r > t) = \gamma^t$. If R_t denotes the reward at turn t of that trajectory, bounded per-turn rewards give

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T_r-1} R_t\right] &= \sum_{t=0}^{\infty} \mathbb{E}[\mathbb{I}[T_r > t] R_t] \\
&= \sum_{t=0}^{\infty} \Pr(T_r > t)\, \mathbb{E}[R_t] \\
&= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[R_t].
\end{aligned} \tag{23}$$

The undiscounted return before geometric termination is therefore an unbiased estimator of the evaluated policy's discounted infinite-horizon return. No additional within-run discount is applied to this estimator. The 100-turn continuous-regeneration evaluation uses a finite observation window and is evaluated separately from this geometric-stopping identity.

C Proof of Theorem 1

Proof of Theorem 1. Write $\pi = \pi_{\mathrm{VSI}}$ and make the encoded history explicit. At a common history σ and matched states x and $s = \phi(x)$, the symbolic policy and its deployment use the same distribution $\pi(\cdot \mid \sigma, s)$. Define

$$D = \sup_{\sigma, x} \left| V^{\pi}_{\mathrm{real}}(\sigma, x) - V^{\pi}_{\mathrm{VSI}}(\sigma, \phi(x)) \right|,$$

where the supremum ranges over compatible encoded histories and physical states. Both value functions have absolute value at most $R_{\max}/(1-\gamma)$, so D is finite.

Fix $(\sigma, x, \mathbf{a})$, with $s = \phi(x)$. Retaining the recommended and realized action in the continuation history, define

$$f(s') = V^{\pi}_{\mathrm{VSI}}(\sigma, s, \mathbf{a}, \mathbf{a}; s'), \qquad g(x') = V^{\pi}_{\mathrm{real}}(\sigma, s, \mathbf{a}, \mathbf{a}; x').$$

For the same encoded-history update, $|g(x') - f(\phi(x'))| \le D$. Adding and subtracting the real expectation of $f(\phi(x'))$ in the continuation difference yields

$$\begin{aligned}
&\left|\mathbb{E}_{x' \sim p_{\mathrm{real}}(\cdot|x,\mathbf{a})} g(x') - \mathbb{E}_{s' \sim p_{\mathrm{VSI}}(\cdot|s,\mathbf{a})} f(s')\right| \\
&\quad \le D + \left|\mathbb{E}_{s' \sim \widetilde{p}(\cdot|x,\mathbf{a})} f(s') - \mathbb{E}_{s' \sim p_{\mathrm{VSI}}(\cdot|s,\mathbf{a})} f(s')\right| \\
&\quad \le D + 2\|f\|_\infty \,\mathrm{TV}(\widetilde{p}(\cdot \mid x, \mathbf{a}), p_{\mathrm{VSI}}(\cdot \mid s, \mathbf{a})) \\
&\quad \le D + \frac{2R_{\max}}{1-\gamma}\Delta_p.
\end{aligned}$$

The first inequality uses the pushforward definition of $\widetilde{p}$; the second uses total variation. The immediate coordinator rewards differ by at most Δ_r. Averaging over the common recommendation distribution, multiplying the continuation bound by γ, and taking the supremum over (σ, x) gives

$$D \le \Delta_r + \gamma D + \frac{2\gamma R_{\max}}{1-\gamma}\Delta_p.$$

Rearranging yields

$$D \le \frac{\Delta_r}{1-\gamma} + \frac{2\gamma R_{\max}}{(1-\gamma)^2}\Delta_p,$$

which proves Eq. (13), with the history argument suppressed as in the theorem statement. □

The theorem compares the same symbolic policy across the symbolic and real models. As described in the main text, VSI's correlated-equilibrium guarantees apply to the symbolic game G_{VSI}, rather than to the complete embodied policy after context-grounded action revision and execution.

D EXECUTION OUTLINE

Algorithm 1 summarizes the SEGUE loop. Each agent receives private symbolic guidance and independently grounds its action, quantity, and program. The committed programs execute in parallel. Only realized outcomes update rewards, the shared environment, and symbolic history; the LLM judge provides structured feedback for reflection. As in the main text, SEGUE advances the stored continuation when the predicted symbolic successor matches the realized one and otherwise replans from the realized symbolic state.

Algorithm 1 SEGUE: private symbolic guidance and parallel embodied execution

Require: VSI policy π_{VSI}, encoder ϕ, task specification D, shared environment $\mathcal{E}$, evaluation stopping rule

Initialize physical state x_0, symbolic history σ_0, agent memories $\{m_{-1}^i\}_{i=1}^n$, and judge feedback $\{j_{-1}^i\}_{i=1}^n$
for $t = 0, 1, 2, \ldots$ **do**
  Encode the physical state: $s_t \leftarrow \phi(x_t)$
  Sample joint guidance: $\mathbf{c}_t \sim \pi_{\mathrm{VSI}}(\cdot \mid \sigma_t, s_t)$
  for all agents $i \in \{1, \ldots, n\}$ **in parallel do**
    Provide agent i only with its private recommendation c_t^i
    Obtain its current local observation o_t^i and available history h_t^i
    Construct context: $z_t^i \leftarrow (D, o_t^i, j_{t-1}^i, h_t^i, m_{t-1}^i)$
    Ground guidance: $u_t^i = (a_t^i, q_t^i, \psi_t^i) \leftarrow P_{\theta_i}(z_t^i, s_t, c_t^i)$
    The planner may revise the action or quantity to fit the current context.
  end for
  Execute committed programs in parallel:
  $(x_{t+1}, r_t, \tau_t) \sim \mathcal{E}(\cdot \mid x_t, u_t^1, \ldots, u_t^n)$
  Rewards and successor state depend on realized execution.
  for all agents $i \in \{1, \ldots, n\}$ **in parallel do**
    Structure execution feedback: $j_t^i \leftarrow \textsc{Judge}(\tau_t)$
    Update reflection memory: $m_t^i \leftarrow \textsc{Reflect}(m_{t-1}^i, j_t^i, \tau_t)$
  end for
  Encode the realized successor: $\widetilde{s}_{t+1} \leftarrow \phi(x_{t+1})$
  if $\widetilde{s}_{t+1}$ matches the successor predicted by the VSI continuation branch **then**
    $\sigma_{t+1} \leftarrow \textsc{Advance}(\sigma_t, \mathbf{c}_t, \tau_t, \widetilde{s}_{t+1})$
  else
    $\sigma_{t+1} \leftarrow \textsc{ReplanFrom}(\widetilde{s}_{t+1})$
  end if
  if the evaluation stopping rule terminates the rollout after this completed turn **then**
    break
  end if
end for

For geometric stopping, the post-turn termination probability is $1 - \gamma = 1/3$. For the continuous-regeneration evaluation, the rollout is observed for 100 completed turns, and the first 60 active seconds define the fixed-time endpoint. These evaluation budgets do not change the private-guidance and parallel-execution structure of the loop.