L2M2: A Hierarchical Framework Integrating Large Language Model and Multi-agent Reinforcement Learning

SMU Singapore Management University

IJCAI 2025 MONTREAL

Minghong Geng[1], Shubham Pateria[1], Budhitama Subagdja[1], Lin Li[2], Xin Zhao[3], Ah-Hwee Tan[1]

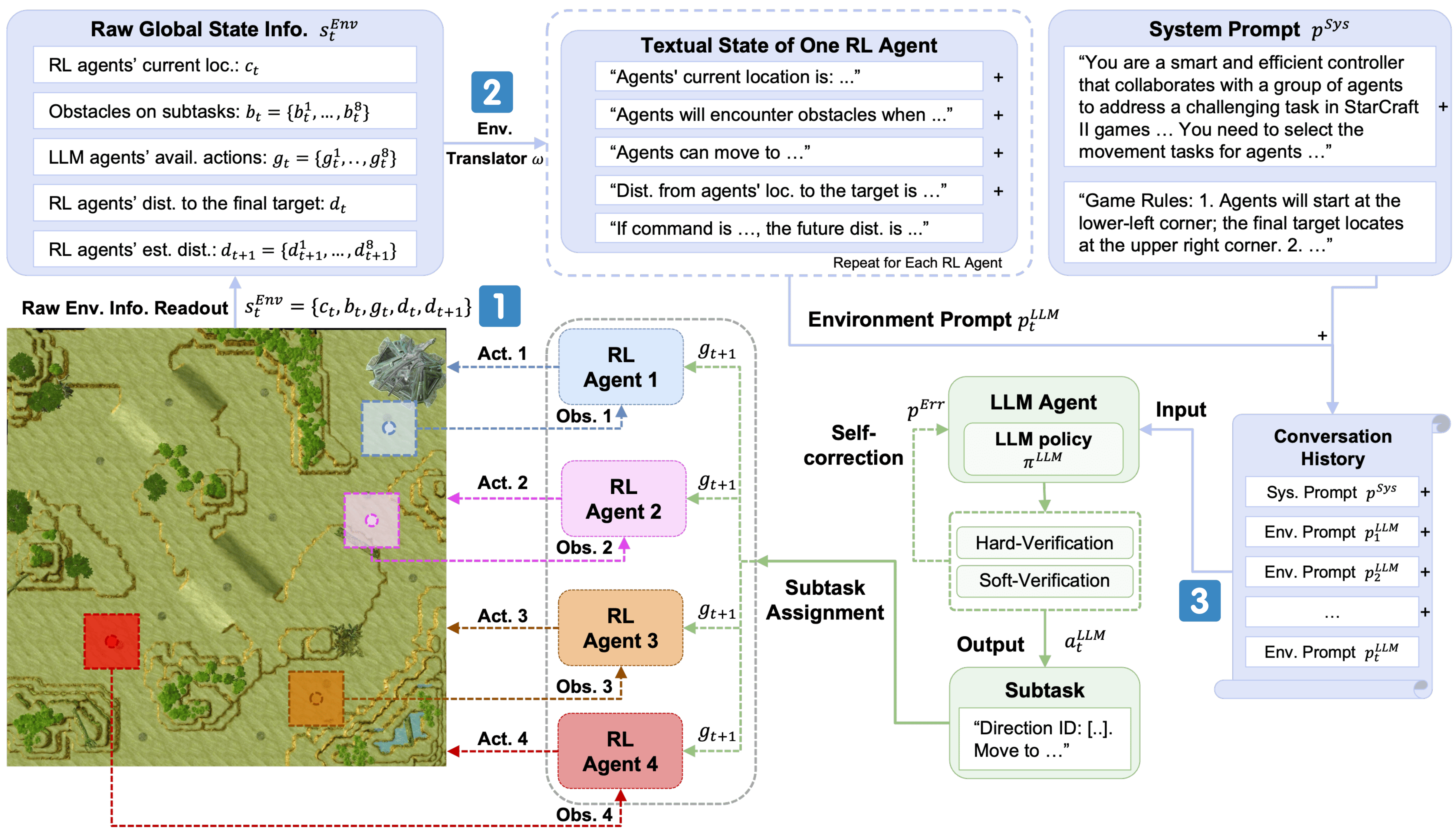
Singapore Management University[1], MIGU Co., Ltd[2], Tsinghua University[3]

GitHub Repo

MOTIVATION

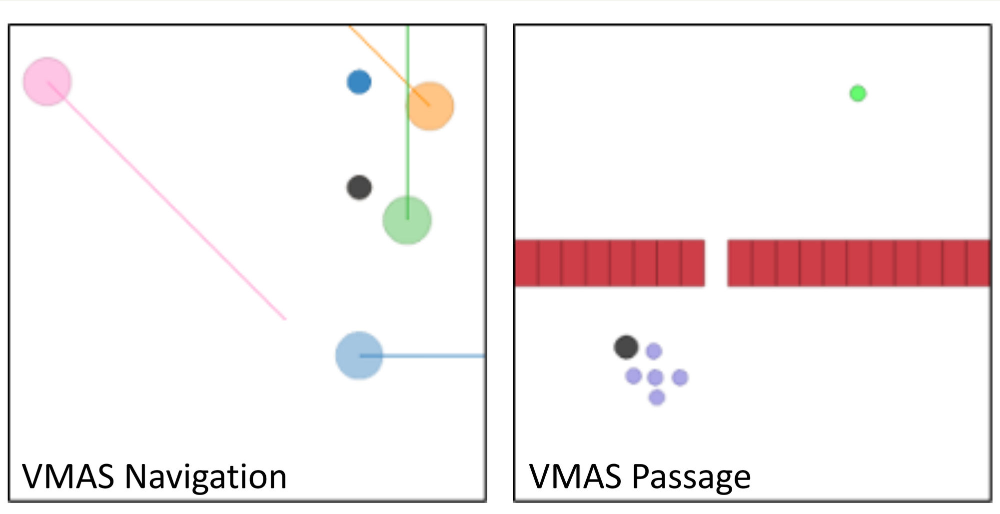
- Multi-agent reinforcement learning (MARL) faces challenges in scaling to complex scenarios w. sustained planning and coordination across long horizons.
- We present L2M2, a novel hierarchical framework that leverages large language models (LLMs) for high-level strategic planning and MARL for low-level execution.
- L2M2 achieves superior performance while requiring less than 20% of the training samples compared to baselines.
- L2M2 Features:
  - Zero-shot RL agent control using LLMs
  - Sample efficient LLM-guided MARL Training
  - Generalizability across different env. and scenarios

An Overview of the L2M2 Framework



LLM Agent: State Representation: Environmental prompts p^{LLM}; Action Space: Discrete subtasks G; Feedback Mechanism: Hard/soft verification
RL Agents: Observation Space: Environment + Subtask info.; Action Space: Primitive actions; Reward: Environment + Subtask rewards

Experiments Settings



The VMAS [1] navigation (four RL agents) and passage (five RL agents) scenarios implemented in this study.

MOSMAC [2] scenarios implemented in this study. In each scenario, four units perform navigate tasks.

Baseline Comparisons

- Non-Hierarchical methods (End-to-end training): QMIX [3]
- Hierarchical methods (end-to-end training and direct integration):
  - Rule-Based Controller + QMIX
  - HiSOMA [4] (FALCON + QMIX)
  - L2M2 (LLM + QMIX)

L2M2 Architecture: LLM + MARL Integration

The LLM Agent: Strategic planning and subtask allocation

1 Raw Environmental Information Readout:

$$s_t^{Env} = (c_t, b_t, g_t, d_t, d_{t+1})$$

To extract key information from the simulation environment as environmental states.

2 Environment Translator ω:

$$\omega: S^{Env} \rightarrow P^{LLM}$$

To map numerical environmental states into environmental prompts.

3 Prompt Construction:

To construct inputs that incorporate system prompts and existing environmental prompts utilized for LLM's inferencing.

LLM's Decision-making:

$$a_t^{LLM} = \{g_{t+1}^i \in G \mid i \in \{1, \dots, n\}\}$$

LLM agent generates temporally abstracted subtasks from the set of available subtasks G for n RL agents.

Verification on output format and action validity. Self-correction with error descriptions if error occurs.

The Reinforcement Learning Agents: Execute primitive actions.

Observation:

$$o_t^i = (o_t^{e,i}, o_t^{g,i})$$

RL agents perceive environments partially, observing general local environment information and subtask-related information.

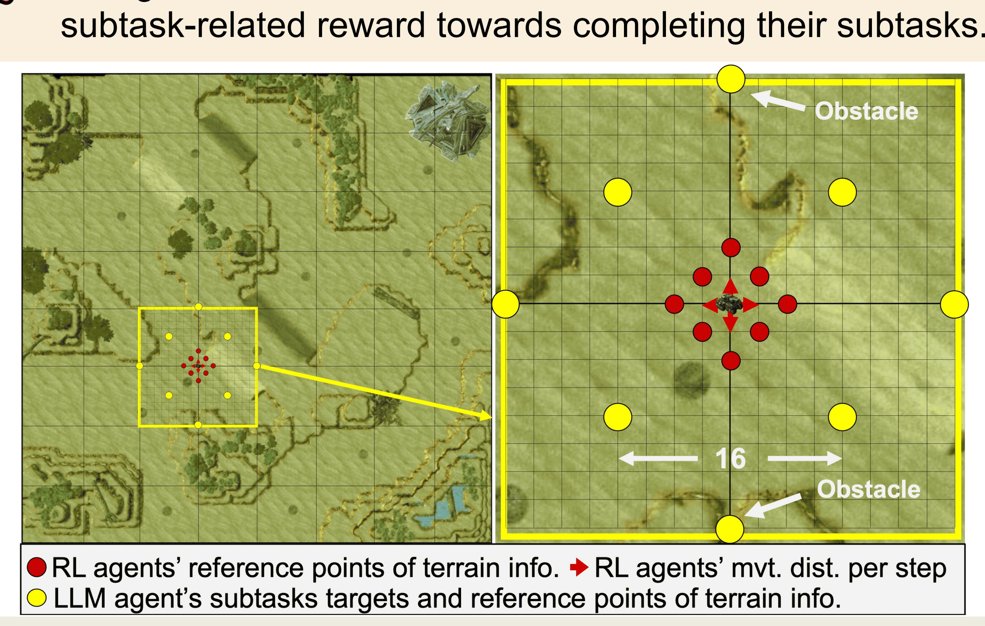
Action:

RL agents take actions follow the default configurations of the benchmark environments. For example, actions in MOSMAC are *no-op*, *movement in four directions* and *stop*.

Reward function:

$$r_t^i = r_t^{e,i} + r_t^{g,i}$$

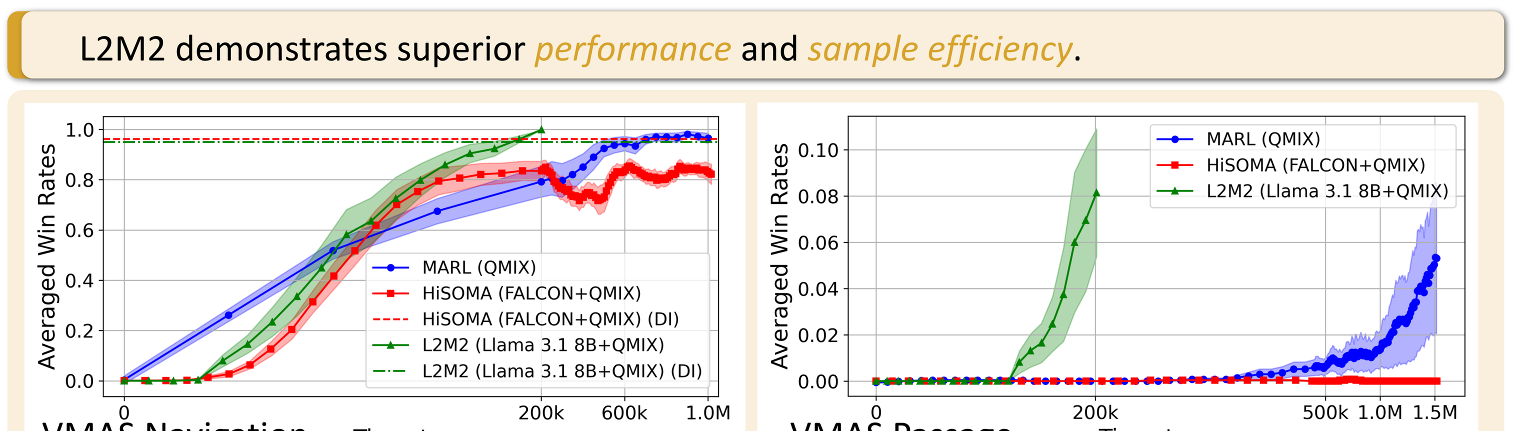
RL agents balance immediate environmental reward with subtask-related reward towards completing their subtasks.



RL agents' reference points of terrain info. ➔ RL agents' mvt. dist. per step
LLM agent's subtasks targets and reference points of terrain info.

Results and Analysis

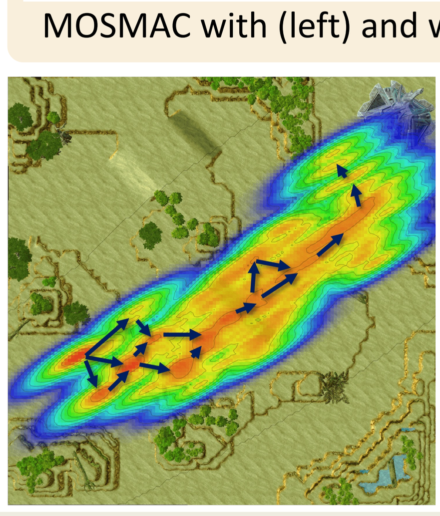
L2M2 demonstrates superior *performance* and *sample efficiency*.



Metrics	Rule-based Control	HiSOMA (DI)	L2M2 (DI)
	RBC w. QMIX	FALCON w. QMIX	Llama 3.1 8B w. QMIX
Avg. Win Rates (%) (mean and std)	83.13 ± 10.96	94.38 ± 1.40	98.75 ± 2.80
Avg. Returns (mean and std)	9.45 ± 0.85	10.82 ± 0.13	9.92 ± 0.21

Metrics	End-to-End	Policy Transfer	
	MARL	HiSOMA (DI)	L2M2 (DI)
Avg. Win Rates (%) (mean and std)	0.00 ± 0.00	14.68 ± 14.14	68.13 ± 25.14
Avg. Returns (mean and std)	0.73 ± 0.09	7.65 ± 2.49	9.39 ± 2.06

MOSMAC with (left) and w/o (right) subgoals



Kernel density estimation reveals that L2M2's LLM agent automatically generates *strategic navigation paths* that avoid challenging terrain features.

LLM Action Density Map shows spatial distribution of LLM's action selections using kernel density estimation, demonstrated on the MOSMAC scenario w/o subgoals.
LLM performs strategic path selection with zero-shot planning, with high density in central regions with short path and low density near cliffs and ramps.

Conclusion

L2M2 is an efficient and novel method for addressing challenging multi-agent problems, benefiting from the power of pre-trained language models.

Key Benefits of L2M2 Framework

- Zero-Shot Planning: Immediate strategic guidance from pre-trained LLMs
- Sample Efficiency: 80-85% reduction in training samples
- Generalizability : Adaptable to different MARL algorithms and LLMs

Future Extensions of L2M2

- Multi-Level Hierarchy: Extend to 3+ level hierarchies for complex task decomposition
- Dynamic Subtask Generation: LLM automatically create new subtasks
- Heterogeneous Agent Teams: Different agent types with specialized capabilities

Acknowledgement

This research was supported in part by Tsinghua University - Migu Xinkong Culture Technology (Xiamen) Co., Ltd. Joint Research Center for Intelligent Light Field and Interactive Technology, Beijing, under its Joint research and development project (R24119F0) - Phase 1, the National Research Foundation, Singapore, under its AI Singapore Programme (AISG Award No: AISG2-RP-2020-019), and the Lee Kong Chian Professorship awarded to Ah-Hwee Tan by Singapore Management University.

References

[1] Matteo Bettini, Ryan Kortvelesy, Jan Blumenkamp, and Amanda Prorok. 2022. VMAS: A Vectorized Multi-agent Simulator for Collective Robot Learning. In *Distributed Autonomous Robotic Systems*, November 28, 2022. Springer Nature Switzerland, 42--56.
[2] Minghong Geng, Shubham Pateria, Budhitama Subagdja, and Ah-Hwee Tan. 2025. MOSMAC: A Multi-agent Reinforcement Learning Benchmark on Sequential Multi-Objective Tasks. In Proceedings of AAMAS '25, June 05, 2025. IFAAMAS, 867–876.
[3] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *Proceedings of ICML 2018*, July 10, 2018. PMLR, 4295–4304.
[4] Minghong Geng, Shubham Pateria, Budhitama Subagdja, and Ah-Hwee Tan. 2024. HiSOMA: A hierarchical multi-agent model integrating self-organizing neural networks with multi-agent deep reinforcement learning. *Expert Systems with Applications* 252, (October 2024), 124117.