SKILL MACHINES: TEMPORAL LOGIC COMPOSITION IN REINFORCEMENT LEARNING

Geraud Nangue Tasse, Devon Jarvis, Steven James, Benjamin Rosman
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
geraudnt@gmail.com
{devon.jarvis, steven.james, benjamin.rosman1}@wits.ac.za

ABSTRACT

A major challenge in reinforcement learning is specifying tasks in a manner that is both interpretable and verifiable. One common approach is to specify tasks through reward machines—finite state machines that encode the task to be solved. We introduce skill machines, a representation that can be learned directly from these reward machines that encode the solution to such tasks. We propose a framework where an agent first learns a set of base skills in a reward-free setting, and then combines these skills with the learned skill machine to produce composite behaviours specified by any regular language, such as linear temporal logics. This provides the agent with the ability to map from complex logical task specifications to near-optimal behaviours zero-shot. We demonstrate our approach in both a tabular and high-dimensional video game environment, where an agent is faced with several of these complex, long-horizon tasks. Our results indicate that the agent is capable of satisfying extremely complex task specifications, producing near optimal performance with no further learning. Finally, we demonstrate that the performance of skill machines can be improved with regular off-policy reinforcement learning algorithms when optimal behaviours are desired.

1 INTRODUCTION

Reinforcement learning (RL) is a promising framework for developing truly general agents capable of acting autonomously in the real world. Despite recent successes in the field, ranging from video games (Badia et al., 2020) to robotics (Levine et al., 2016), there are several shortcomings to existing approaches that hinder RL's real-world applicability. One issue is that of sample efficiency—while it is possible to collect millions of data points in a simulated environment, it is simply not feasible to do so in the real world. This inefficiency is exacerbated when a single agent is required to solve multiple tasks (as we would expect of a generally intelligent agent). Another issue arises when an agent is required to solve a long horizon task in the presence of a sparse learning signal. In this case, it is often near impossible for the agent to solve the task, regardless of how much data it collects, since the sequence of actions to execute before a learning signal is received is too large (Arjona-Medina et al., 2019). However, this can be mitigated by leveraging higher-order skills, which shorten the planning horizon (Sutton et al., 1999).

A desirable characteristic of generally intelligent agents is their ability to reuse learned behaviours to solve new tasks (Taylor & Stone, 2009), preferably without further learning. One approach to overcoming this challenge is to rely on *composition*, where an agent first learns individual skills and then combines them to produce novel behaviours. There are several notions of compositionality in the literature, such as temporal composition, where skills are invoked one after the other (Sutton et al., 1999; Barreto et al., 2019), and concurrent composition, where skills are combined to produce a new behaviour to be executed (Todorov, 2009; Saxe et al., 2017; Van Niekerk et al., 2019; Alver & Precup, 2022).

Notably, work by Nangue Tasse et al. (2020) has demonstrated how an agent can learn skills that can be combined using Boolean operators, such as negation and conjunction, to produce semantically

meaningful behaviours without further learning. An important benefit of this compositional approach is that it provides a way to address another key issue with RL: tasks, as defined by reward functions, can be notoriously difficult to specify. This may lead to undesired behaviours that are not easily interpretable and verifiable. Composition that enables simpler task specifications and produces reliable behaviours thus represents a major step towards safe AI (Cohen et al., 2021).

Unfortunately, these compositions are strictly concurrent and cannot be chained to solve temporally-specified tasks. One solution to this issue is *reward machines*—finite state machines that encode the tasks to solve (Icarte et al., 2018). While this obviates the sparse reward problem, the agent is still required to learn how to solve a given task through environment interaction, and the subsequent solution is monolithic, restricting its applicability to new tasks and limiting the reliability of resulting behaviours.

In this work, we combine these two approaches to develop an agent capable of zero-shot concurrent *and* temporal composition. We particularly focus on temporal logic composition, such as linear temporal logic (LTL) (Pnueli, 1977), allowing agents to sequentially chain and order their skills while ensuring certain conditions are always or never met. We make the following contributions: (a) we propose *skill machines*, a finite state machine that can be autonomously learned by a compositional agent, and which can be used to solve any task expressible as a finite state machine without further learning; (b) we prove that these skill machines are *satisficing*—given a task specification, an agent can successfully solve it while adhering to any constraints; and (c) we demonstrate our approach in several environments, including a high-dimensional video game domain. Having learned a set of base skills in a reward-free setting, our results indicate that our method is capable of producing near-optimal behaviour for a variety of long-horizon tasks without further learning.

2 Background

We model the agent's interaction with the world as a Markov Decision Process (MDP), given by $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where (i) $\mathcal{S} \subseteq \mathbb{R}^n$ is the n-dimensional state space; (ii) $\mathcal{A}$ is the set of (possibly continuous) actions available to the agent; (iii) $P(s'|s,a)$ is the dynamics of the world, representing the probability of the agent reaching state s' after executing action a in state s; (iv) R is a reward function bounded by $[R_{\text{MIN}}, R_{\text{MAX}}]$ that represents the task the agent needs to solve; and (v) $\gamma \in [0, 1]$ is a discount factor.

The aim of the agent is to compute a Markov policy π from $\mathcal{S}$ to $\mathcal{A}$ that optimally solves a given task. Instead of directly learning a policy, an agent will often instead learn a value function that represents the expected return following policy π from state s: $V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$. A more useful form of value function is the action-value function $Q^\pi(s,a)$, which represents the expected return obtained by executing a from s, and then following π. The optimal action-value function is given by $Q^*(s,a) = \max_\pi Q^\pi(s,a)$ for all states s and actions a, and the optimal policy follows by acting greedily with respect to Q^* at each state.

2.1 Logical Composition in the Multitask Setting

We are interested in the multitask setting, where an agent is required to reach a set of goals in some goal space $\mathcal{G} \subseteq \mathcal{S}$. We assume that all tasks share the same state space, action space and dynamics, but differ in their reward functions. We model this setting by defining a background MDP $M_0 = \langle \mathcal{S}_0, \mathcal{A}_0, P_0, R_0 \rangle$ with its own state space, action space, transition dynamics and background reward function. Any individual task τ is then specified by a task-specific reward function R_τ that is non-zero only for states in $\mathcal{G}$. The reward function for the resulting MDP is then simply $R_0 + R_\tau$.

Nangue Tasse et al. (2020) consider the case where $R_\tau \in \{R_{\text{MIN}}, R_{\text{MAX}}\}$ and develop a framework that allows agents to apply the Boolean operations of conjunction ($\wedge$), disjunction ($\vee$) and negation ($\neg$) over the space of tasks and value functions. This is achieved by first defining the goal-oriented reward function $\bar{R}$ which extends the task rewards r to penalise an agent for achieving goals different from the one it wished to achieve:

$$\bar{R}(s,g,a) \coloneqq \begin{cases} \bar{R}_{\text{MIN}} & \text{if } g \neq s \text{ and } s \text{ is absorbing} \\ R(s,a) & \text{otherwise,} \end{cases} \tag{1}$$

where $\bar{R}_{\text{MIN}}$ is a large negative penalty that can be derived from the bounds of the reward function.

Using Equation 1, we can define the related goal-oriented value function as:

$$\bar{Q}(s, g, a) = \bar{R}(s, g, a) + \gamma \int_{\mathcal{S}} \bar{V}^{\bar{\pi}}(s', g) P_{(s,a)}(ds'), \tag{2}$$

where $\bar{V}^{\bar{\pi}}(s, g) = \mathbb{E}_{\bar{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \bar{R}(s_t, g, a_t)\right]$.

If a new task can be represented as the logical expression of previously learned tasks, Nangue Tasse et al. (2020) prove that the optimal policy can immediately be obtained by composing the learned goal-oriented value functions using the same expression. For example, the union ($\vee$), intersection ($\wedge$), and negation ($\neg$) of two goal-reaching tasks A and B can be solved as follows (we omit the value functions' parameters for readability):

$$\bar{Q}^*_{A \vee B} = \bar{Q}^*_A \vee \bar{Q}^*_B := \max\{\bar{Q}^*_A, \bar{Q}^*_B\}$$
$$\bar{Q}^*_{A \wedge B} = \bar{Q}^*_A \wedge \bar{Q}^*_B := \min\{\bar{Q}^*_A, \bar{Q}^*_B\}$$
$$\bar{Q}^*_{\neg A} = \neg \bar{Q}^*_A := \left(\bar{Q}^*_{SUP} + \bar{Q}^*_{INF}\right) - \bar{Q}^*_A,$$

where $\bar{Q}^*_{SUP}$ and $\bar{Q}^*_{INF}$ are the goal-oriented value functions for the maximum task ($r = R_{\text{MAX}}$ for all $\mathcal{G}$) and minimum task ($r = R_{\text{MIN}}$ for all $\mathcal{G}$), respectively . Following Nangue Tasse et al. (2022), we refer to these goal-oriented value functions as *world value functions* (WVFs).

2.2 Reward Machines

One difficulty with the standard MDP formulation is that the agent is often required to solve a complex long-horizon task using only a scalar reward signal as feedback from which to learn. To overcome this, Icarte et al. (2018) propose *reward machines* (RMs), which provide structured feedback to the agent in the form of a finite state machine. RMs encode a reward function using a set of propositional symbols $\mathcal{P}$ that represent abstract environment features as follows:

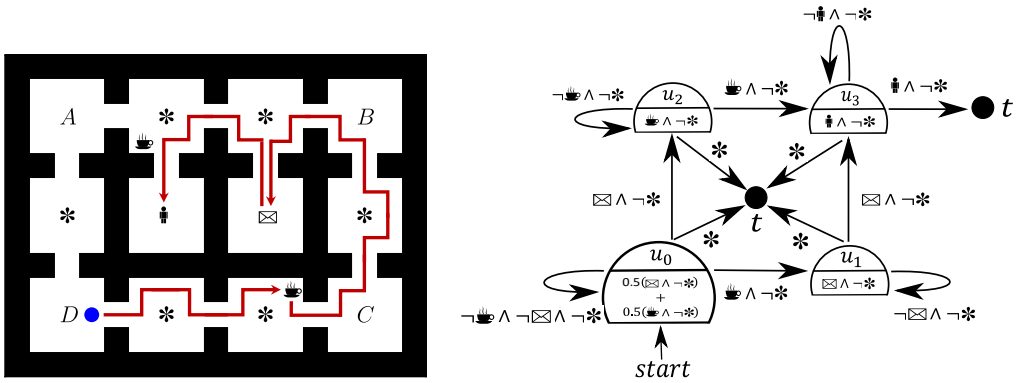
Definition 1 (Reward Machine)**.** *Given a set of propositional symbols $\mathcal{P}$, states $\mathcal{S}$ and actions $\mathcal{A}$, a reward machine is a tuple $R_{\mathcal{PSA}} = \langle U, u_0, F, \delta_u, \delta_r \rangle$ where (i) U is a finite set of states; (ii) $u_0 \in U$ is an initial state; (iii) F is a finite set of terminal states; (iv) $\delta_u : U \times 2^{\mathcal{P}} \to U \cup F$ is the state-transition function; and (v) $\delta_r : U \to [\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}]$ is the state-reward function.*

RMs consist of a finite set of states U, each of which represents a set of propositions that are true at a given environment state. Transitions between RM states are governed by δ_u, and the RM emits a reward function according to δ_r. A particular instantiation of an RM that is used in practice is a *simple reward machine* (SRM), which restricts the form of the state-reward function to be $\delta_r : U \times 2^{\mathcal{P}} \to \mathcal{R}$ (Icarte et al., 2018). In other words, when a transition between $u, u' \in U$ is made, the SRM emits a single scalar instead of a function (as in the case of RMs).

To incorporate RMs into the RL framework, the agent must be able to determine which abstract propositions are true at any given state. To achieve this, the agent is equipped with a labelling function $L : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to 2^{\mathcal{P}}$ that assigns truth values to the propositions based on the agent's interaction with its environment. The agent can then learn a policy in a new decision process where the reward function in the original MDP is replaced with the RM, which is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, P, \gamma, \mathcal{P}, L, U, u_0, F, \delta_u, \delta_r \rangle$. The agent's aim is now to learn a policy over the joint MDP and RM state space $\pi : \mathcal{S} \times U \to \mathcal{A}$, which can be achieved with standard algorithms such as Q-learning (Icarte et al., 2018).

3 Leveraging Skill Composition for Temporal Logic Tasks

To describe our approach to temporal composition, we use the *Office Gridworld* (Icarte et al., 2018) as a running example. In the environment, illustrated by Figure 1a, an agent (blue circle) can move to adjacent cells in any of the cardinal directions. It can also pick up coffee or mail at locations ☕ or ✉ respectively, and it can deliver them to the office at location 🚹. Cells marked ✻ indicate decorations that are broken if the agent collides with them, and cells marked A–D indicate the centres of the corner rooms. The reward machines that specify tasks in this environment are defined over 10 propositions: $\mathcal{P} = \{A, B, C, D, ✻, ☕, ✉, 🚹, ✉^+, 🚹^+\}$, where the first 8 propositions are true when the agent is at their respective locations, $✉^+$ is true when the agent is at ✉ and there is mail to be collected, and $🚹^+$ is true when the agent is at 🚹 and there is someone in the office.



(a) Office Gridworld

(b) Skill machine

Figure 1: Illustration of (a) the office gridworld where the blue circle represents the agent and (b) the skill machine for the task "deliver coffee and mail to the office without breaking any decoration" where the black dots labeled t represent terminal states.

3.1 Task Space

We now define the set of tasks to be considered. We first introduce the concept of *constraints* $C \subseteq \mathcal{P}$, which are the set of propositions that an agent should avoid setting to true and corresponds to the global operator G in a linear temporal logic (LTL) specification. An example of a constraint might be that the agent should complete a task, but avoid breaking any decorations while doing so. We can now define the notion of base (primitive) tasks, which will later be composed.

Definition 2 (Task Primitive)**.** *Let $M = \langle \mathcal{S}_0, \mathcal{A}_0, P_0, R_0 \rangle$ be a background MDP. We define a set of task primitives in this domain as $\mathcal{M}_\mathcal{G} = \{\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle\}$ with absorbing goal space $\mathcal{G} = 2^\mathcal{P}$ and labelling function L, where*

$$\mathcal{S} := (\mathcal{S}_0 \times 2^C) \cup 2^\mathcal{P}, \textit{where } C \textit{ is the set of constraints};$$

$$\mathcal{A} := \mathcal{A}_0 \times \mathcal{A}_\tau, \textit{ where } \mathcal{A}_\tau = \{0, 1\} \textit{ represents whether or not to terminate a task};$$

$$P(\langle s_0, c \rangle, \langle a_0, a_\tau \rangle) := \begin{cases} L(s_0, a_0, s_0') & \textit{if } a_\tau = 1 \\ \langle s_0', c' \rangle & \textit{otherwise} \end{cases},$$

$$\textit{where } s_0' \sim P_0(\cdot | s_0, a_0) \textit{ and } c' = c \cup (c \cap L(s_0, a_0, s_0'));$$

$$R(\langle s_0, c \rangle, \langle a_0, a_\tau \rangle) := \begin{cases} R_\tau \in \{R_{MIN}, R_{MAX}\} & \textit{if } a_\tau = 1 \\ R_0(s_0, a_0) & \textit{otherwise}. \end{cases}$$

The above defines the tasks' state space to be the product of the environment state and the set of constraints, incorporating the set of propositions that are currently true. The action space is augmented with a terminating action following Barreto et al. (2019) and Nangue Tasse et al. (2020), which indicates that the agent wishes to achieve the goal it is currently at, and is similar to an option's termination condition (Sutton et al., 1999). The transition dynamics update the environment state and constraints set to true when a regular action is taken, and use the labelling function to return the set of propositions achieved when the agent decides to terminate. Finally, the agent receives the regular environment reward when taking an action, but a task-specific goal reward when it terminates. We will assume that the environment and task-specific rewards are such that the optimal policies for all tasks are guaranteed to attain desired reachable goal states—a common example is to have a reward of 1 at desired goal states and no rewards everywhere else.

Equipped with this definition, we can now define the set of all tasks under consideration:

Definition 3 (Task space)**.** *The set of all tasks $\mathcal{M}$ is all linear preferences over task primitives:*

$$\mathcal{M} = \left\{ \langle \mathcal{S}, \mathcal{A}, P, R_\mathbf{w}, \gamma \rangle : R_\mathbf{w}(s, a) = \sum_{i=1}^{|\mathcal{M}_\mathcal{G}|} w_i R_i(s, a) \textit{ where } \sum_{i=1}^{|\mathcal{M}_\mathcal{G}|} w_i = 1 \textit{ and } \mathbf{w} \in \mathbb{R}^{|\mathcal{M}_\mathcal{G}|} \right\}$$

This definition of task space provides a general notion of tasks that are still grounded in achieving goals. It also subsumes most definitions considered in the composition literature, including Boolean algebra tasks (Nangue Tasse et al., 2020) and linear preference tasks (Barreto et al., 2020). Hence, we will restrict our attention to reward machines whose rewards per state originate from the defined task space (instead of arbitrary real-valued functions that are not grounded in achieving goals in an environment). We denote such reward machines as $R_{\mathcal{MSA}} = \langle U, u_0, F, \delta_u, \delta_{\mathcal{M}} \rangle$ where $\delta_{\mathcal{M}}$ maps from RM states to task rewards.

3.2 Skill Machines

The goal space of a primitive task is defined by a set of Boolean propositions. We can leverage prior work to solve each individual task using a set of base primitive skills (Nangue Tasse et al., 2020). We therefore only need concern ourselves with how to solve any task expressed as a linear combination of the primitive tasks. Fortunately, Theorem 1 below demonstrates that a linear combination of base skills does just this (proofs of all theorems are presented in the Appendix).

Theorem 1. *Let $\mathbf{R}_{\mathcal{M}_\mathcal{G}}$ be a vector of rewards for each primitive task, and $\bar{\mathbf{Q}}^*_\mathcal{G}$ be the corresponding vector of optimal WVFs. Then, for a task $m \in \mathcal{M}$ with reward function $R_\mathbf{w} = \mathbf{w} \cdot \mathbf{R}_{\mathcal{M}_\mathcal{G}}$, we have*

$$\bar{Q}^*_m = \mathbf{w} \cdot \bar{\mathbf{Q}}^*_\mathcal{G}.$$

We now have agents capable of solving any logical and linear composition of tasks by learning a finite set of base skills $\bar{\mathbf{Q}}^*_\mathcal{G}$ for task primitives $\mathcal{M}_\mathcal{G}$. We refer to this basis set of skills as *skill primitives*. Given this compositional ability over skills, and reward machines that expose the structure of tasks, agents can solve temporally extended tasks with little or no further learning. To achieve this, we define a skill machine (SM) as a representation of logical and temporal knowledge over skills.

Definition 4 (Skill Machine)**.** *Given a set of propositional symbols $\mathcal{P}$ with constraints $\mathcal{C} \subseteq \mathcal{P}$, their corresponding skill primitives $\bar{\mathbf{Q}}^*_\mathcal{G}$ and task space $\mathcal{M}$, states $\mathcal{S}$ and actions $\mathcal{A}$, a skill machine is a tuple $\bar{\mathcal{Q}}^*_{\mathcal{SA}} = \langle U, u_0, F, \delta_u, \delta_Q, \mathbf{w}_U, \mathbf{w}_\mathcal{G} \rangle$ where (i) U, u_0, F, δ_u are defined as for reward machines; (ii) $\mathbf{w}_U : U \times U \to \mathbb{R}$ is a preference function over state transitions; (iii) $\mathbf{w}_\mathcal{G} : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ is a preference function over goals; and (v) $\delta_Q : \mathcal{S} \times U \times \mathcal{A} \to [\mathcal{S} \times \mathcal{A} \to \mathbb{R}]$ is the state-skill function defined by:*

$$\delta_Q(s, u, a) \mapsto \sum_{g \in \mathcal{G}} \sum_{u' \in U} \mathbf{w}_\mathcal{G}(s, u, g) \mathbf{w}_U(u, u') \bar{Q}^*_{u,u'}(s, g, a),$$

where $\bar{Q}^*_{u,u'}$ is the WVF obtained by composing the skill primitives $\bar{\mathbf{Q}}^*_\mathcal{G}$ according to the Boolean expression for the transition $\delta_u = u'$.

For a given state s in the environment and state u in the skill machine, the skill machine uses its preference over transitions $\mathbf{w}_U$ and goals $\mathbf{w}_\mathcal{G}$ to compute a skill $Q(s, a) \coloneqq \delta_Q(s, u, a)$ that an agent can use to take an action a. The environment then transitions to the next state $s' \leftarrow P(s, a)$ and the skill machine transitions to $u' \leftarrow \delta_u(u, L(s, a, s'))$. $\mathbf{w}_U$ represents cases where there is not necessarily a single desirable transition to follow given the current SM state. This is illustrated by the SM in Figure 1b, where mail and coffee are equally desirable at the initial state. Similarly, $\mathbf{w}_\mathcal{G}$ represents cases where there may be a single desirable task, but its goals are not necessarily equally desirable given the environment state—for example when the agent needs to first pick up coffee but there are two coffee locations. Interestingly, there always exists a choice for $\mathbf{w}_U$ and $\mathbf{w}_\mathcal{G}$ that is optimal with respect to the corresponding reward machine, as shown in Theorem 2.

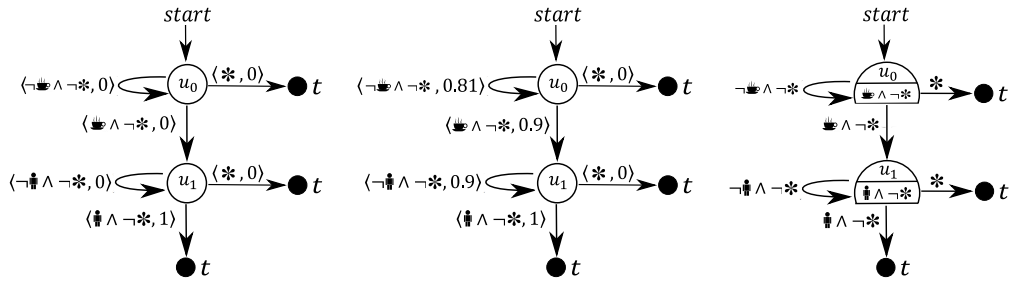
Theorem 2. *Let $\pi^*(s, u)$ be the optimal policy for the cross-product MDP between a reward machine and a task space $\mathcal{M}$, with $\mathcal{C} = \mathcal{P}$. Then there exists a corresponding skill machine with a $\mathbf{w}_\mathcal{G}$ and $\mathbf{w}_U$ such that*

$$\pi^*(s, u, a) \in \arg\max_a \delta_Q(s, u, a),$$

where δ_Q is given by $\mathbf{w}_\mathcal{G}$ and $\mathbf{w}_U$ as per Definition 4.

Theorem 2 shows that skill machines can be used to solve tasks without having to relearn action level policies. The next section shows how an agent can approximate a skill machine by planning over simple reward machines.

3.3 FROM REWARD MACHINES TO SKILL MACHINES



(a) Reward machine (b) Value iterated RM (c) Skill machine

Figure 2: The reward machine, value iterated RM and skill machine for the task "Deliver coffee to the office without breaking any decoration". This task is specified using LTL as $(F(☕ \wedge X(F\ 🚹))) \wedge (G\ \neg✻))$, where $F = Finally, X = neXt, G = Globally$ are LTL operators. The corresponding RM is obtained by converting the LTL into a finite state machine and then giving a reward of 1 for accepting transitions and 0 otherwise. The black dots labeled t represent terminal states.

In the previous section, we introduced skill machines and showed that they can be used to represent the logical and temporal composition of skills needed to solve reward machines. We now show how for simple RMs their approximate SM can be obtained zero-shot without further learning. To achieve this, we first plan over the reward machine (using value iteration, for example) to obtain Q-values for each transition. We then select the skills for each SM state greedily. This process is illustrated in Figure 2. While this only holds for cases where the greedy skills are always satisfiable from any environment state, this still covers many tasks of interest. In particular, this holds for any RM with non-zero rewards of R_{MAX} at accepting transitions,[1] as shown in Theorem 3.

Theorem 3. *Let $R_{\mathcal{MSA}} = \langle U, u_0, F, \delta_u, \delta_{\mathcal{M}} \rangle$ be a satisfiable simple reward machine with non-zero rewards R_{MAX} only for accepting transitions, and for which all valid transitions (u, u') are achievable from any state $s \in \mathcal{S}$. Define the skill machine $\bar{\mathcal{Q}}^*_{\mathcal{SA}} = \langle U, u_0, F, \delta_u, \delta_Q, \mathbf{w}_U, \mathbf{w}_{\mathcal{G}} \rangle$ with*

$$\mathbf{w}_U(u, u') := \begin{cases} 1 & \text{if } u' = \arg\max_{u''} Q^*(u, u''), \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{w}_{\mathcal{G}}(s, u, g) := \begin{cases} 1 & \text{if } g = \arg\max_{g'} \max_a \sum_{u'} \mathbf{w}_U(u, u') \bar{Q}^*_{u,u'}(s, g', a), \\ 0 & \text{otherwise} \end{cases}$$

where Q^* is the optimal transition-value function for $R_{\mathcal{PSA}}$. Then following the policy $\pi^*(s, u) \in \arg\max_a \delta_Q(s, u, a)$ will reach an accepting transition.

Theorem 3 is critical as it provides soundness guarantees, ensuring that the policy derived from the skill machine will always satisfy the task requirements when it is possible to do so. Finally, in cases where the composed skill δ_Q obtained from the approximate SM is not sufficiently optimal, we can use any off-policy algorithm to learn a new skill Q_{new} few-shot. This is achieved by using the maximising Q-values $\max\{\beta Q_{new}, (1-\beta)\delta_Q\}$ in the behaviour policy during learning. Here, $\beta \in [0, 1]$ is a parameter that determines how much of the composed policy to use. It can also be seen as decreasing the potentially overestimated values of δ_Q, since δ_Q is greedy with respect to both goals and RM transitions. Algorithm 1 illustrates this process with Q-learning where $\beta = \gamma$, which guarantees convergence since δ_Q will never dominate the optimal Q-values in the $\max$.

4 EXPERIMENTS

4.1 ZERO-SHOT TEMPORAL LOGICS

We consider the Office Gridworld domain presented by Icarte et al. (2018) and depicted in Figure 1a. This environment is used as a multitask domain, consisting of the four tasks described in Table 1.

[1] Accepting transitions are transitions at which the high level task—described, for example, by linear temporal logics—is satisfied.

Algorithm 1: Few-shot Q-learning using skill machines

Input : $\gamma, \alpha, \mathcal{P}, L, U, u_0, F, \delta_u, \delta_Q$
Initialise : $Q(s, u, a)$
foreach *episode* **do**
  Observe initial state $s \in \mathcal{S}$ and get initial $u \leftarrow u_0$
  while *episode is not done* **do**
    `/* Using the composed skill` δ_Q `in the behaviour policy */`
    $a \leftarrow \begin{cases} \arg\max_{a \in \mathcal{A}} \left(\max\{\gamma Q(s,u,a), (1-\gamma)\delta_Q(s,u,a)\}\right) & \text{if } Bernoulli(1-\epsilon) = 1 \\ \text{a random action} & \text{otherwise} \end{cases}$
    Take action a and observe next state s'
    Get reward $r \leftarrow \delta_r(u)(s, a, s')$ and the next RM state $u' \leftarrow \delta_u(u, L(s, a, s'))$
    $Q(s, u, a) \xleftarrow{\alpha} r$ **if** s' is terminal or $u' \in F$ **else** $\left[r + \gamma \max_{a'} Q(s', u', a')\right]$
    $s \leftarrow s'$

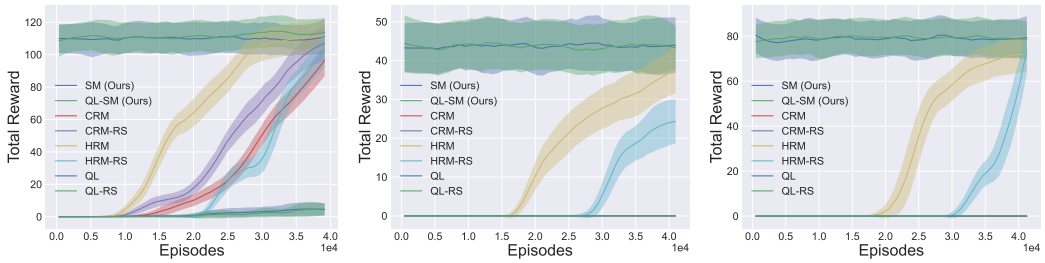
Task	Description
1	Deliver coffee to the office without breaking any decoration
2	Patrol rooms A, B, C, and D without breaking any decoration
3	Deliver coffee and mail to the office without breaking any decoration
4	Deliver mail to the office until there is no mail left, then deliver coffee to office while there are people in the office, then patrol rooms A-B-C-D-A, and never break a decoration

Table 1: Tasks in the Office Gridworld.

We begin by evaluating how long it takes an agent to learn a policy that can solve all four tasks. The agent iterates through the tasks, changing from one to the next after each episode. In all of our experiments, we compare the performance of skill machines with that of state-of-the-art RM-based learning approaches like counterfactual RMs (CRM)—where the Q-functions are updated with respect to all possible RM transitions from a given environment state—and hierarchical RMs (HRM)—where an agent learns options per RM state that are grounded in the environment states (Icarte et al., 2018). Note that CRM and HRM are theoretically capable of solving the multi-task problem setup because they can use experience from solving one task to update the policies for solving other tasks. However, skill machines can additionally share both experience between tasks when learning the skill primitives, as well as use the composition of these primitives to generalise to more difficult tasks without further learning.

We run 80 independent trials and report the average reward per step across the four tasks in Table 1. In addition to learning all four tasks, we also experiment with Tasks 3 and 4 in isolation. For these experiments, 80 independent trials are run and the average reward per step computed. In the single task domains, the difference between CRM, HRM, skill machines and Q-learning should be less pronounced, since CRM, HRM and skill machines now cannot leverage prior knowledge. Thus, the comparison between multi-task and single-task learning in this setting will evaluate the benefit of the compositionality afforded by skill machines.

The results of these three experiments are shown in Figure 3. Regular Q-learning struggles to learn Task 3 and completely fails to learn the hardest task (Task 4). Additionally, notice that while QL and CRM can theoretically learn the tasks optimally given infinite time, only HRM and SM are able to learn hard long horizon tasks in practice. It is important to note that we train all algorithms for the same amount of time during these experiments and previous work (Nangue Tasse et al., 2020) has shown that learning the WVFs takes longer than learning task-specific skills. In addition, the skill machines are being used to zero-shot generalise to the office tasks using skill primitives. Thus using the skill machines in isolation (labelled SM and shown in blue on Figure 3) may provide sub-optimal performance compared to the task-specific agents, since the skill machines have not been trained to optimality and are not specialised to the domain. Even under these conditions, we observe that skill machines perform near-optimally in terms of final performance, and due to the amortised nature of learning the WVF will achieve its final rewards from the first epoch.



(a) Task 3 (b) Task 4 (c) Multiple tasks (1-4)

Figure 3: Average returns during training in the Office Gridworld.

4.2 Few-shot Temporal Logics

It is possible to pair the skill machines with a learning algorithm such as Q-learning to achieve few-shot generalisation. From the results shown in Figure 3, it is apparent that skill machines paired with Q-learning (labelled QL-SM and shown in orange on Figure 3) achieves the best performance for both the single-task and multi-task setting. While it is not clear from the rewards that adding Q-learning provides significant improvements to the skill machine, their trajectories show that Q-learning does indeed improve on the skill machine policies when they are not optimal (Appendix 7). Additionally, skill machines with Q-learning always begin with a significantly higher reward and converge on their final performance faster than all benchmarks—except the zero-shot one which is (near) optimal in all cases. The speed of learning is due to the compositionality of the skill primitives with skill machines, and the high final performance is due to the generality of the learned primitives being paired with the domain specific Q-learner. In sum, skill machines provide fast composition of skills and achieve optimal performance compared to all benchmarks when paired with a learning algorithm.

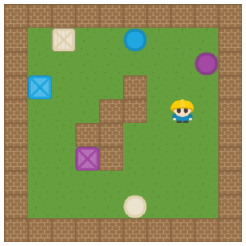
4.3 Moving Targets Domain

We now demonstrate our temporal logic composition approach in a canonical object collection domain with high dimensional pixel observations (Nangue Tasse et al., 2020) (Figure 4a). The agent here needs to pick up objects of various shapes and colors; picked objects respawn at random empty positions similarly to previous object collection domains (Barreto et al., 2020). There are 3 object colours—beige, blue, purple—and 2 object shapes—squares, circles. To learn the WVF for a given task primitive, we use the goal oriented Q-learning method of Nangue Tasse et al. (2020) where the agent keeps track of reached goals and uses deep Q-learning (Mnih et al., 2015) to update the WVF with respect to all seen goals at every time step.

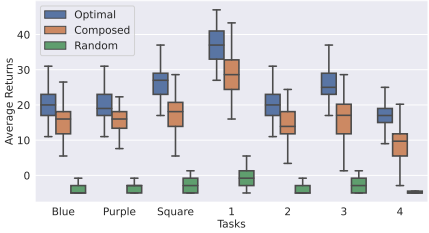
We first train the agent on three base task primitives: *pick up blue objects*, *pick up purple objects*, and *pick up squares*. We then use the learned skill primitives to solve multiple temporal logic tasks. Figure 4b shows the average returns of the optimal policies and SM policies for the four tasks described in Table 2 with a maximum of 50 steps per episode. Our results show that even when using function approximation with sub-optimal skill primitives, the zero-shot policies obtained from skill machines are very close to optimal on average. We also observe that for very challenging tasks like Tasks 3 and 4 (where the agent must satisfy difficult temporal constraints), the compounding effect of the sub-optimal policies sometimes leads to failures. In such cases, learning new skills few-shot by leveraging the SM would guarantee convergence to optimal policies as demonstrated in Section 4.2.

Task	Description
1	Pick up any object. Repeat this forever.
2	Pick up blue then purple objects, then objects that are neither blue nor purple. Repeat this forever.
3	Pick up blue objects or squares, but never blue squares. Repeat this forever.
4	Pick up non-square blue objects, then non-blue squares in that order. Repeat this forever.

Table 2: Tasks in the Moving Targets domain. Objects respawn in random positions when picked.



(a) Moving Targets domain



(b) Average returns over 100 runs for tasks in Table 2.

Figure 4: Illustration of the domain and plots of the average returns over 100 runs for tasks in Table 2.

5 Related Work

One family of approaches to concurrent composition leverages forms of regularisation to achieve semantically meaningful disjunction (Todorov, 2009; Van Niekerk et al., 2019) or conjunction (Haarnoja et al., 2018; Hunt et al., 2019). Weighted composition has also been demonstrated; for example, Peng et al. (2019) learn weights to compose existing policies multiplicatively to solve new tasks. Approaches that leverage the successor feature (SF) framework (Barreto et al., 2017) are capable of solving tasks defined by linear preferences over features (Barreto et al., 2020). Alver & Precup (2022) show that an SF basis can be learned that is sufficient to span the space of tasks under consideration, while Nemecek & Parr (2021) determine which policies should be stored in limited memory so as to maximise performance on future tasks. In contrast to these approaches, our framework allows for both concurrent composition (including operators such as negation that other approaches do not support) and temporal composition such as LTL.

A popular way of achieving temporal composition is through the options framework (Sutton et al., 1999; Bacon et al., 2017). Here, high-level skills are first discovered and then executed sequentially to solve a task (Konidaris & Barto, 2009; Bagaria & Konidaris, 2019). Barreto et al. (2019) leverage the SF and options framework and learn how to linearly combine skills, chaining them sequentially to solve temporal tasks. However, these options-based approaches offer a relatively simple form of temporal composition. By contrast, we are able to solve tasks expressed through regular languages zero-shot, while providing soundness guarantees.

Work has also centred on approaches to defining tasks using human-readable logic operators. For example, Li et al. (2017) and Littman et al. (2017) specify tasks using LTL, which is then used to generate a standard reward signal for an RL agent. Camacho et al. (2019) show how to perform reward shaping given LTL specifications, while Jothimurugan et al. (2019) develop a formal language that encodes tasks as sequences, conjunctions and disjunctions of subtasks. This is then used to obtain a shaped reward function that can be used for learning. All of these approaches focus on how an agent can improve learning given such specifications or structure, but we show how an explicitly compositional agent can immediately solve such tasks using WVFs without further learning.

6 Conclusion

We proposed skill machines—finite state machines that can be learned from reward machines—that allow agents to solve extremely complex tasks involving temporal and concurrent composition. We demonstrated how skills can be learned and encoded in a specific form of goal-oriented value function that, when combined with the learned skill machines, are sufficient for solving subsequent tasks without further learning. Our approach guarantees that the resulting policy adheres to the logical task specification, which provides assurances of safety and verifiability to the agent's decision making, important characteristics that are necessary if we are to ever deploy RL agents in the real world. While the resulting behaviour is provably satisficing, empirical results demonstrate that the agent's performance is near optimal; further fine-tuning can be performed should optimality be required, which greatly improves the sample efficiency. We see this approach as a step towards truly generally intelligent agents, capable of immediately solving human-specifiable tasks in the real world with no further learning.

REFERENCES

Safa Alver and Doina Precup. Constructing a good behavior basis for transfer using generalized policy updates. In *International Conference on Learning Representations*, 2022.

Jose Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32, 2019.

P. Bacon, J. Harb, and D. Precup. The option-critic architecture. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. Agent57: Outperforming the Atari human benchmark. In *International Conference on Machine Learning*, pp. 507–517, 2020.

Akhil Bagaria and George Konidaris. Option discovery using deep skill chaining. In *International Conference on Learning Representations*, 2019.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 4055–4065, 2017.

André Barreto, Diana Borsa, Shaobo Hou, Gheorghe Comanici, Eser Aygün, Philippe Hamel, Daniel Toyama, Shibl Mourad, David Silver, Doina Precup, et al. The option keyboard: Combining skills in reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

André Barreto, Shaobo Hou, Diana Borsa, David Silver, and Doina Precup. Fast reinforcement learning with generalized policy updates. *Proceedings of the National Academy of Sciences*, 117 (48):30079–30087, 2020.

Alberto Camacho, Rodrigo Toro Icarte, Toryn Klassen, Richard Valenzano, and Sheila McIlraith. LTL and beyond: Formal languages for reward function specification in reinforcement learning. In *International Joint Conferences on Artificial Intelligence*, volume 19, pp. 6065–6073, 2019.

Vanya Cohen, Geraud Nangue Tasse, Nakul Gopalan, Steven James, Matthew Gombolay, and Benjamin Rosman. Learning to follow language instructions with compositional policies. In *AAAI Fall Symposium Series*, 2021.

Tuomas Haarnoja, Vitchyr Pong, Aurick Zhou, Murtaza Dalal, Pieter Abbeel, and Sergey Levine. Composable deep reinforcement learning for robotic manipulation. In *2018 IEEE International Conference on Robotics and Automation*, pp. 6244–6251, 2018.

Jonathan Hunt, Andre Barreto, Timothy Lillicrap, and Nicolas Heess. Composing entropic policies using divergence correction. In *International Conference on Machine Learning*, pp. 2911–2920, 2019.

Rodrigo Toro Icarte, Toryn Klassen, Richard Valenzano, and Sheila McIlraith. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *International Conference on Machine Learning*, pp. 2107–2116, 2018.

Kishor Jothimurugan, Rajeev Alur, and Osbert Bastani. A composable specification language for reinforcement learning tasks. *Advances in Neural Information Processing Systems*, 32, 2019.

George Konidaris and Andrew Barto. Skill discovery in continuous reinforcement learning domains using skill chaining. *Advances in Neural Information Processing Systems*, 22, 2009.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Xiao Li, Cristian-Ioan Vasile, and Calin Belta. Reinforcement learning with temporal logic rewards. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3834–3839, 2017.

Michael Littman, Ufuk Topcu, Jie Fu, Charles Isbell, Min Wen, and James MacGlashan. Environment-independent task specifications via GLTL. *arXiv preprint arXiv:1704.04341*, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei Rusu, Joel Veness, Marc Bellemare, Alex Graves, Martin Riedmiller, Andreas Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Geraud Nangue Tasse, Steven James, and Benjamin Rosman. A Boolean task algebra for reinforcement learning. *Advances in Neural Information Processing Systems*, 33:9497–9507, 2020.

Geraud Nangue Tasse, Steven James, and Benjamin Rosman. World value functions: Knowledge representation for multitask reinforcement learning. In *The 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*, 2022.

Mark Nemecek and Ron Parr. Policy caches with successor features. In *International Conference on Machine Learning*, pp. 8025–8033, 2021.

Xue Bin Peng, Michael Chang, Grace Zhang, Pieter Abbeel, and Sergey Levine. MCP: Learning composable hierarchical control with multiplicative compositional policies. In *Advances in Neural Information Processing Systems*, pp. 3686–3697, 2019.

Amir Pnueli. The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science*, pp. 46–57. IEEE, 1977.

Andrew Saxe, Adam Earle, and Benjamin Rosman. Hierarchy through composition with multitask LMDPs. *International Conference on Machine Learning*, pp. 3017–3026, 2017.

Richard Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, 1999.

Matthew. Taylor and Peter Stone. Transfer learning for reinforcement learning domains: a survey. *Journal of Machine Learning Research*, 10:1633–1685, 2009.

Emanuel Todorov. Compositionality of optimal control laws. In *Advances in Neural Information Processing Systems*, pp. 1856–1864, 2009.

Benjamin Van Niekerk, Steven James, Adam Earle, and Benjamin Rosman. Composing value functions in reinforcement learning. In *International Conference on Machine Learning*, pp. 6401–6409, 2019.

A APPENDIX

A.1 PROOFS OF THEORETICAL RESULTS

Theorem 1. *Let $\mathbf{R}_{\mathcal{M}_\mathcal{G}}$ be a vector of rewards for each primitive task, and $\bar{\mathbf{Q}}^*_\mathcal{G}$ be the corresponding vector of optimal WVFs. Then, for a task $m \in \mathcal{M}$ with reward function $R_\mathbf{w} = \mathbf{w} \cdot \mathbf{R}_{\mathcal{M}_\mathcal{G}}$, we have*

$$\bar{Q}^*_m = \mathbf{w} \cdot \bar{\mathbf{Q}}^*_\mathcal{G}.$$

Proof.

$$\begin{aligned}\bar{Q}^*_m(s,g,a) &= \mathbb{E}_{\bar{\pi}^*}\left[\sum_{t=0}^{\infty}\gamma^t \mathbf{w}\cdot\bar{\mathbf{R}}_{\mathcal{M}_\mathcal{G}}(s_t,g,a_t)\right]\\ &= \mathbf{w}\cdot\mathbb{E}_{\bar{\pi}^*}\left[\sum_{t=0}^{\infty}\gamma^t \bar{\mathbf{R}}_{\mathcal{M}_\mathcal{G}}(s_t,g,a_t)\right];\\ &\quad \text{since the world policies are independent of task Nangue Tasse et al. (2020)[Lemma 2].}\\ &= \mathbf{w}\cdot\bar{\mathbf{Q}}^*_\mathcal{G}\end{aligned}$$

□

Theorem 2. *Let $\pi^*(s,u)$ be the optimal policy for the cross-product MDP between a reward machine and a task space $\mathcal{M}$, with $\mathcal{C} = \mathcal{P}$. Then there exists a corresponding skill machine with a $\mathbf{w}_\mathcal{G}$ and $\mathbf{w}_U$ such that*

$$\pi^*(s,u) \in \arg\max_a \delta_Q(s,u),$$

where δ_Q is given by $\mathbf{w}_\mathcal{G}$ and $\mathbf{w}_U$ as per Definition 4.

Proof. Let $\mathbf{w}_U(u,\cdot) = \frac{1}{N_{\delta_u}}$ where N_{δ_u} is the number of possible RM transitions from u. Also let $\mathbf{w}_\mathcal{G}(s,u,\cdot)$ be 1 for the set of propositions $g \in 2^\mathcal{C}$ that are satisfied when following $\pi^*(s,u)$, and zero everywhere else. Then $\pi^*(s,u) \in \arg\max_a \delta_Q(s,u)$ since $\mathbf{w}_U(u,u')\bar{Q}^*_{u,u'}(s,g,a)$ is optimal using Theorem 1 and optimal policies are assumed to reach task goals. □

Theorem 3. *Let $R_{\mathcal{MSA}} = \langle U, u_0, F, \delta_u, \delta_\mathcal{M}\rangle$ be a satisfiable simple reward machine with non-zero rewards R_{MAX} only for accepting transitions, and for which all valid transitions (u,u') are achievable from any state $s \in \mathcal{S}$. Define the skill machine $\bar{\mathcal{Q}}^*_{\mathcal{SA}} = \langle U, u_0, F, \delta_u, \delta_Q, \mathbf{w}_U, \mathbf{w}_\mathcal{G}\rangle$ with*

$$\mathbf{w}_U(u,u') := \begin{cases}1 & \text{if } u' = \arg\max_{u''} Q^*(u,u''),\\ 0 & \text{otherwise}\end{cases}$$

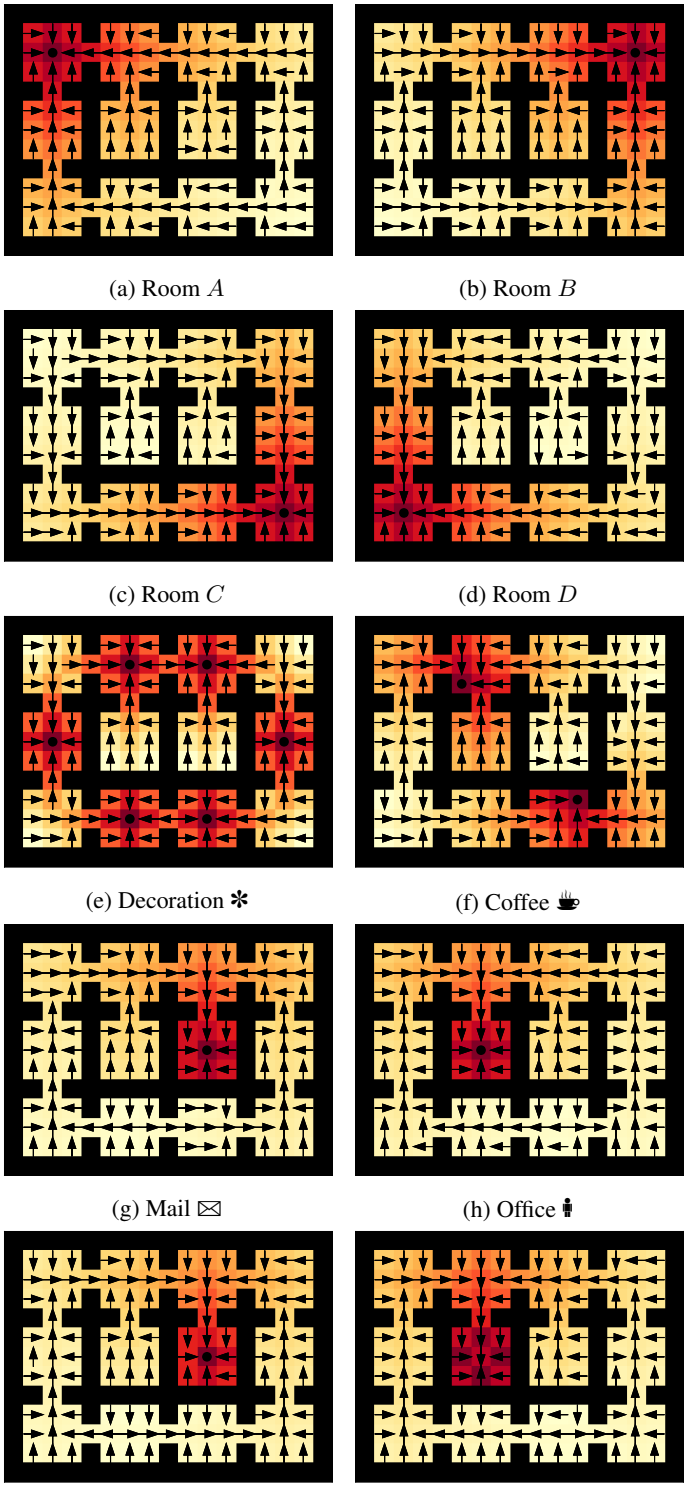
$$\mathbf{w}_\mathcal{G}(s,u,g) := \begin{cases}1 & \text{if } g = \arg\max_{g'}\max_a \sum_{u'}\mathbf{w}_U(u,u')\bar{Q}^*_{u,u'}(s,g',a),\\ 0 & \text{otherwise}\end{cases}$$

where Q^* is the optimal transition-value function for $R_{\mathcal{PSA}}$. Then following the policy $\pi^*(s,u) \in \arg\max_a \delta_Q(s,u)$ will reach an accepting transition.

Proof. This follows from the optimality of $\pi^*(s,u)$ and Q^*, since each transition of the RM is satisfiable from any environment state.

□

A.2 Office World Additional Figures



(a) Room A (b) Room B

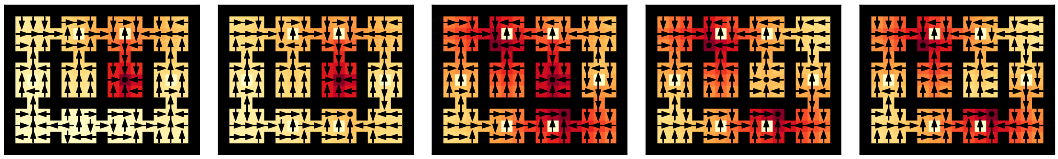
(c) Room C (d) Room D

(e) Decoration ✻ (f) Coffee ☕

(g) Mail ✉ (h) Office 🚹

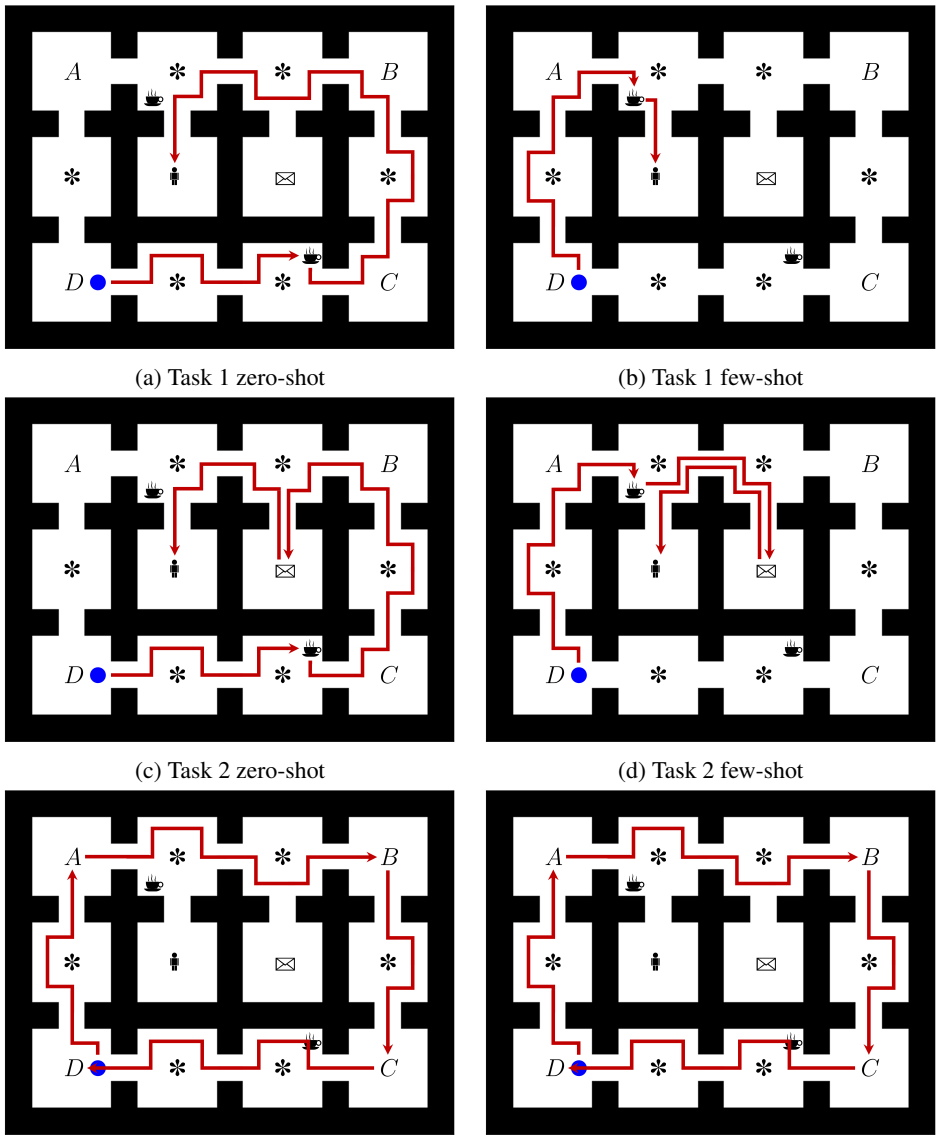
(i) Mails present ✉$^+$ (j) People present 🚹$^+$

Figure 5: The policies (arrows) and value functions (heat map) of the base primitive tasks in the Office Gridworld. These are obtained by maximising over the goals of the learned WVFs. All errors in the figures are due to training the WVFs for 200000 time steps, hence not to convergence.



(a) $1(✉ \wedge \neg ✻) + 0(☕ \wedge \neg ✻)$ (b) $0.75(✉ \wedge \neg ✻) + 0.25(☕ \wedge \neg ✻)$ (c) $0.5(✉ \wedge \neg ✻) + 0.5(☕ \wedge \neg ✻)$ (d) $0.25(✉ \wedge \neg ✻) + 0.75(☕ \wedge \neg ✻)$ (e) $0(✉ \wedge \neg ✻) + 1(☕ \wedge \neg ✻)$

Figure 6: The policies (arrows) and value functions (heat map) for various preferences over the tasks "get a mail without breaking decorations" ($✉ \wedge \neg ✻$) and "get coffee without breaking decorations" ($☕ \wedge \neg ✻$). These are obtained by first doing the weighted sum of the composed WVFs according to the preferences, and then maximising over the goals and actions. All errors in the figures are due to training the base WVFs for 200000 time steps, hence not to convergence.



(a) Task 1 zero-shot (b) Task 1 few-shot

(c) Task 2 zero-shot (d) Task 2 few-shot

(e) Task 3 zero-shot (f) Task 3 few-shot

Figure 7: Agent trajectories for various tasks in the Office Gridworld (Table 1) using the skill machine without further learning (left) and with further learning (right).