

LEVERAGING LARGE LANGUAGE MODELS TO REPAIR HIGH-LEVEL ROBOT CONTROLLERS FROM ASSUMPTION VIOLATIONS

Qian Meng, Jin Peng Zhou, Kilian Q. Weinberger & Hadas Kress-Gazit *

Cornell University

Ithaca, NY 14850, USA

{qm34, jz563, kqw4, hadaskg}@cornell.edu

ABSTRACT

This paper presents INPROVF, an automatic framework that combines large language models (LLMs) and formal methods to speed up the repair process of high-level robot controllers. Previous approaches based solely on formal methods are computationally expensive and cannot scale to large state spaces. In contrast, INPROVF uses LLMs to generate repair candidates, and formal methods to verify their correctness. To improve the quality of these candidates, our framework first translates the symbolic representations of the environment and controllers into natural language descriptions. If a candidate fails the verification, INPROVF provides feedback on potential unsafe behaviors or unsatisfied tasks, and iteratively prompts LLMs to generate improved solutions. We demonstrate INPROVF through 12 violations with various workspaces, tasks, and state space sizes.

1 INTRODUCTION

Safety-critical robotic applications, such as autonomous driving (Yurtsever et al., 2020), search and rescue missions (Lyu et al., 2023), and assistive caregiving robots (e.g. Madan et al. (2024)), benefit from formal guarantees to ensure safe operations and task completions, the lack of which may result in harm to users and even loss of life in catastrophic events. One way to provide such guarantees is the use of formal synthesis—an automatic technique that transforms temporal logic specifications into correct-by-construction systems—to create provably correct high-level robot controllers that compose low-level robot skills to satisfy temporally extended tasks in complex environments.

Among formal synthesis methods, Generalized Reactivity(1) (GR(1)) synthesis (Bloem et al., 2012) is used in robotics applications due to its relatively low polynomial-time complexity with respect to the state space and its expressivity covering many robotic tasks (Menghi et al., 2019). A GR(1) specification takes the form of an implication from *assumptions* about the environment’s behaviors to *guarantees* on the robot’s behaviors. Specifically, the guarantees depend on the validity of the assumptions; if any assumption is violated at runtime, the synthesized controller is no longer guaranteed to be correct, leading to potential undesirable behaviors.

To address assumption violations, our previous work (Meng & Kress-Gazit, 2024) leverages formal methods-based approaches (Pacheck & Kress-Gazit, 2023) to repair robot behaviors by generating new skills necessary for task satisfaction following assumption violations. However, our previous repair approach fails to scale to real-world settings with large state spaces at runtime due to the state explosion problem that has challenged formal methods for decades (Valmari, 1996).

Recent advances in large language models (LLMs) present a unique opportunity to overcome the scalability problem of formal methods-based repair. Modern LLMs, trained on Internet-scale data across diverse domains, have strong priors and can address complex reasoning tasks—ranging from competitive mathematics and coding problems to PhD-level science questions (OpenAI, 2024).

*The first two authors contribute equally. This paper is also submitted to 2025 IEEE 21th International Conference on Automation Science and Engineering (CASE).

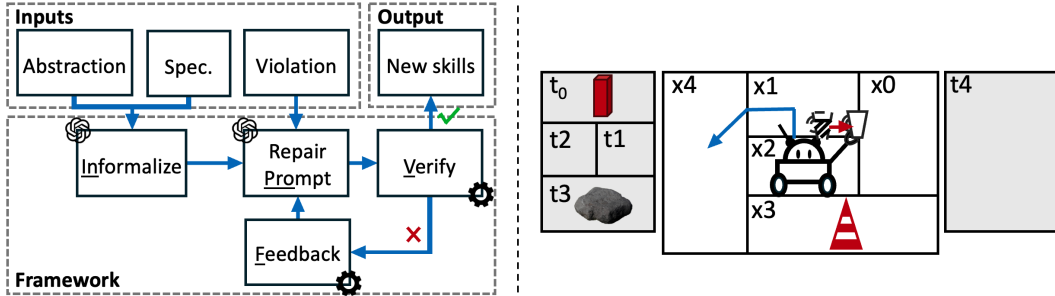


Figure 1: Left: INPROVF overview. Right: Example 1 workspace.

Since repairing high-level robot controllers under assumption violation is itself a complex reasoning task, we aim to leverage the reasoning capacities of LLMs to scale up repair.

In this work, we propose INPROVF, a hybrid framework that combines the strengths of LLMs and formal methods to repair assumption violations in high-level robot controllers at runtime. We use LLMs to generate repair candidates, and then verify them through formal methods. If the verification fails, we employ formal analysis to provide feedback for LLMs to iteratively refine the candidates.

As illustrated in Fig. 1 (left), INPROVF consists of four key steps: *Informalization*, *Repair Prompt*, *Verification*, and *Feedback*. In *Informalization*, we convert the symbolic abstractions of both the environment and the synthesized high-level controller into natural languages (NL) descriptions that capture their underlying semantics—specifically, the physical meaning of the abstraction and the behavior description of the controller (Sec. 5.1). Our intuition is that LLMs, primarily trained on natural language data and further fine-tuned through reinforcement learning from human feedback (Ouyang et al., 2022), are suitable for processing these informalized NL representations. Next, in the *Repair Prompt* step, we assemble the NL descriptions of the abstractions, the robot’s behaviors, the task specifications, and the assumption violations into a prompt for LLMs to generate repair candidates—new skills that potentially enable the robot to fulfill its tasks under the violation (Sec. 5.2). To ensure the soundness of the candidates, the *Verification* step first checks their syntax against a predefined grammar, and then check whether the specification, updated by the violation and the new skills, is realizable (Sec. 5.3). If the verification step fails, the *Feedback* step automatically generates NL feedback, through syntax and counterstrategy analyses, for LLMs to iteratively generate more promising repair candidates (Sec. 5.4).

Contributions. To our knowledge, INPROVF is the first framework that leverages LLMs to accelerate the repair process of assumption violations in large state spaces. By integrating LLM-based skill generation with verification and feedback via formal methods, INPROVF enables provably-correct robots to remain robust in complex, real-world environments with prohibitively large state spaces.

2 RELATED WORK

Synthesis for robots. Synthesis from Linear Temporal Logic (LTL) and GR(1) specifications operates on a discrete abstraction of the environment and the robot’s skills (Kress-Gazit et al., 2018). The abstraction can be automatically generated from sensor inputs using, e.g. Konidaris et al. (2018). Synthesis takes as input a temporal logic (Clarke et al., 2000) specification encoding the safety and task requirements for the robot. While traditionally provided by the user, specifications can also be automatically generated from natural language task descriptions using LLMs (Chen et al., 2023). Synthesis algorithms (e.g. Bloem et al. (2012)) automatically generate a high-level controller that composes robot skills to satisfy the given specification, thereby providing correctness guarantees, assuming the abstraction matches the physical world.

Assumption violations. To address temporary environment assumption violations, Wong et al. (2018) synthesizes robust controllers that can maintain safety constraints during the violations, while Ehlers & Topcu (2014) tolerates intermittent violations that occur only a fixed number of times consecutively, separated by violation-free intervals. Cooperative synthesis generates controllers that do not attempt to violate assumptions (Ehlers et al., 2015). For long-lasting violations that make the specifications unrealizable, Alur et al. (2013) leverages counterstrategies (Könighofer et al., 2009) to

generate additional assumptions that restore the realizability. Similarly, Wong et al. (2018) employs unrealizability analysis (Raman & Kress-Gazit, 2013) to guide the user in adding new liveness assumptions. Our previous work (Meng & Kress-Gazit, 2024) utilizes synthesis-based repair (Pacheco & Kress-Gazit, 2023) to create new robot skills to ensure task success given the violations.

LLMs for formal methods. The formal methods community has leveraged LLMs to enhance scalability. Jiang et al. (2023); Yang et al. (2023) employ LLMs for automated theorem proving, tackling formal mathematical problems that otherwise are computationally intractable and require human inputs. Similarly, Wu et al. (2024); Pirzada et al. (2024); Liu et al. (2024) leverage LLMs to generate program loop invariants, which are crucial for unrolling and reasoning about loops in program verification, and then verify the invariants using formal methods such as bounded model checking, theorem prover, or symbolic execution. In symbolic execution, Wang et al. (2024) utilize LLMs to translate path constraints of real-world Python programs into SMT constraints, improving the scalability of symbolic execution for complex programs. The software engineering community proposes to combine LLMs and static analysis to identify correctness and security vulnerabilities in open-source software and operating systems (Li et al., 2024a;b).

3 PRELIMINARIES

Example 1. Consider a mobile manipulator in a factory-like workspace in Fig. 1 (right). The robot is tasked to move the cup between the loading table t_2 and the assembly table t_4 depending on the status of the cup, with the assumption that the status of the cup can only change when it is on the tables, and the constraint that an empty cup is not allowed to appear in x_0 . However, when the robot holds the cup in x_2 while executing a skill y_{original} that moves the robot from x_4 to x_0 via x_2 , the cup status changes from full to empty. The robot should create new skills, e.g., the new skill y_{new} indicated by the blue arrow in Fig. 1 (right), that drives the robot back to x_4 . The robot then places the cup to t_2 , waits for the cup to be filled up again, and continues the execution.

Abstractions. We model the physical state space with a discrete abstraction. An abstraction comprises a set of atomic propositions AP partitioned into environment inputs $\mathcal{X}$ and system outputs $\mathcal{Y}$ ($AP = \mathcal{X} \cup \mathcal{Y}$). The set of inputs $\mathcal{X}$, representing the environment states, is further divided into sets of controllable inputs $\mathcal{X}_c$ and uncontrollable inputs $\mathcal{X}_u$ ($\mathcal{X} = \mathcal{X}_c \cup \mathcal{X}_u$). Controllable inputs $\mathcal{X}_c$ represent the part of the environment that the robot can control indirectly, e.g. the position of the robot base. Uncontrollable inputs $\mathcal{X}_u$ represent the part of the environment that the robot has no control over, e.g. the position of an obstacle. An input state $\sigma_{\mathcal{X}} \subseteq \mathcal{X}$ and a controllable input state $\sigma_c \subseteq \mathcal{X}_c$ are subsets of inputs and controllable inputs, respectively. Let the physical state space be $X \subseteq \mathbb{R}^n$. A grounding function $G : \mathcal{X} \rightarrow 2^X$ maps each input $\pi \in \mathcal{X}$ into a set of physical states $G(\pi) \subseteq X$. Intuitively, the grounding function captures the physical meaning of each input.

In Example 1, the inputs are $\pi_o^r \in \mathcal{X}$, where o is an object and r is a region. The grounding function maps each input π_o^r to the set of physical states where the object o locates in the region r . An uncontrollable input π_{empty} represents that the cup is empty. The input state in Fig. 1 right is $\sigma_{\mathcal{X}} := \{\pi_{\text{base}}^{x_2}, \pi_{\text{cup}}^{ee}, \pi_{\text{empty}}, \pi_{\text{cone}}^{x_3}, \pi_{\text{block}}^{t_0}, \pi_{\text{stone}}^{t_3}\}$, meaning the robot base is in x_2 , the cup is in the end-effector and is empty, the cone is in x_3 , the block is in t_0 , and the stone is in t_3 .

The system outputs $\mathcal{Y}$ represent the robot skills. Each skill has a set of preconditions, may have intermediate states, and results in postconditions. Each state is a controllable input state. In Example 1, the precondition of the new skill y_{new} is $\{\pi_{\text{base}}^{x_2}, \pi_{\text{cup}}^{ee}, \pi_{\text{block}}^{t_0}\}$ (stone and cone are uncontrollable), the intermediate state is $\{\pi_{\text{base}}^{x_1}, \pi_{\text{cup}}^{ee}, \pi_{\text{block}}^{t_0}\}$, and the postcondition is $\{\pi_{\text{base}}^{x_4}, \pi_{\text{cup}}^{ee}, \pi_{\text{block}}^{t_0}\}$.

Specifications. We use the GR(1) fragment of LTL for its synthesis algorithm and expressivity. The syntax of LTL is $\varphi := \pi \mid \neg\varphi \mid \varphi \vee \varphi \mid \bigcirc\varphi \mid \varphi \mathcal{U}\varphi$ where a formula φ is constructed recursively from atomic propositions $\pi \in \mathcal{AP}$, logical operators ($\neg, \vee$), and temporal operators *next* ($\bigcirc$) and *until* ($\mathcal{U}$). Additional logical operators ($\wedge, \rightarrow, \leftrightarrow$) and temporal operators *eventually* ($\Diamond$) and *always* ($\Box$) can be derived from previous operators. The semantics of LTL is defined over infinite traces $\tau = (\sigma_0, \sigma_1, \dots)$, where each $\sigma_i \subseteq \mathcal{AP}$ denotes the set of propositions that are true at step i . When a trace suffix starting at position i satisfies a formula φ , we denote this as $\tau^i \models \varphi$. By convention, we say that a trace τ satisfies a formula φ , denoted as $\tau \models \varphi$, if the formula is satisfied from the beginning of the trace (i.e., $\tau^0 \models \varphi$). The satisfaction relation is defined recursively as follows: $\tau^i \models \pi$ if $\pi \in \sigma_i$ (π is true at step i); $\tau^i \models \bigcirc\varphi$ if $\tau^{i+1} \models \varphi$ (φ holds at the next step); $\tau^i \models \Diamond\varphi$ if

$\exists j \geq i$ such that $\tau^j \models \varphi$ (φ eventually holds); and $\tau^i \models \Box \varphi$ if $\forall j \geq i, \tau^j \models \varphi$ (φ always holds). A more detailed introduction of LTL semantics can be found in Clarke et al. (2000).

Specifications in the GR(1) fragment of LTL take the form $\varphi = \varphi_e \rightarrow \varphi_s$ where $\varphi_e = \varphi_e^i \wedge \varphi_e^t \wedge \varphi_e^g$ represents the environment assumptions and $\varphi_s = \varphi_s^i \wedge \varphi_s^t \wedge \varphi_s^g$ denotes the system guarantees. The initial constraints φ_e^i and φ_s^i , defined over $\mathcal{X}$ and $\mathcal{Y}$, specify the initial conditions of the environment and the system, respectively. The safety constraints, φ_e^t and φ_s^t , capture conditions that must hold at all times and take the form of $\varphi_a^t = \bigwedge_i \Box \varphi_{a,i}^t$, for $a \in \{e, s\}$, where $\varphi_{e,i}^t$ is defined over $\mathcal{AP} \cup \bigcirc \mathcal{X}$ and $\varphi_{s,i}^t$ over $\mathcal{AP} \cup \bigcirc \mathcal{AP}$. Lastly, we have the liveness constraints φ_e^g and φ_s^g which encode environment assumptions and system goals that must be satisfied repeatedly. For $a \in \{e, s\}$, $\varphi_a^g = \bigwedge_i \Box \Diamond \varphi_{a,i}^g$ where each $\varphi_{a,i}^g$ is defined over $\mathcal{AP}$.

Writing GR(1) specification can be partially automated from the abstraction (e.g., from topological constraints, skills definitions, etc), allowing the user to provide the task specification φ_{task} that captures the essence of the task. We define such φ_{task} as a *sub-specification* of φ , denoted $\varphi_{\text{task}} \preceq \varphi$, such that φ_{task} only contains a subset of safety constraints of φ . In Example 1, an example constraint not included in φ_{task} but in φ is: one object cannot appear in two places simultaneously.

Synthesis. GR(1) synthesis (Bloem et al., 2012) first transforms the GR(1) specification into a two-player game between the environment and the system, where the transition rules for each player are defined by the safety constraints φ_e^t and φ_s^t , and the winning conditions are defined by the liveness goals φ_e^g and φ_s^g , respectively. Next, synthesis computes a set of winning states from which the system is guaranteed to win the game, a process called *realizability check*. Finally, synthesis extracts a finite automaton from the winning states as the robot *strategy*.

Definition 1. A *strategy* is a deterministic finite automaton $\mathcal{A}_s = (\Sigma, \Sigma_0, \mathcal{X}, \mathcal{Y}, \delta, \gamma_{\mathcal{X}}, \gamma_{\mathcal{Y}})$ where (i) Σ is a set of states and $\Sigma_0 \subseteq \Sigma$ is the set of initial states. (ii) $\mathcal{X}$ and $\mathcal{Y}$ are sets of inputs and outputs. (iii) $\delta : \Sigma \times 2^{\mathcal{X}} \rightarrow \Sigma$ maps a state and an input state to the next state. (iv) $\gamma_{\mathcal{X}} : \Sigma \rightarrow 2^{\mathcal{X}}$ maps a state to its labeled input state. (v) $\gamma_{\mathcal{Y}} : \Sigma \rightarrow 2^{\mathcal{Y}}$ maps a state to its labeled output state.

A *strategy description* $\mathcal{D}$ for the strategy $\mathcal{A}_s$ is a natural language description of how $\mathcal{A}_s$ behaves.

If the set of winning states does not overlap with all system liveness goals or does not cover the initial states, a winning strategy for the system does not exist, i.e. the specification is unrealizable. Conversely, a winning strategy for the environment, or a *counterstrategy*, exists.

Definition 2. A *counterstrategy* is a finite automaton $\mathcal{A}_e = (\Sigma, \Sigma_0, \mathcal{X}, \mathcal{Y}, \delta_e, \delta_s, \gamma_{\mathcal{X}}, \gamma_{\mathcal{Y}}, \gamma_g)$ where (i) $\Sigma, \Sigma_0, \mathcal{X}, \mathcal{Y}, \gamma_{\mathcal{X}}, \gamma_{\mathcal{Y}}$ are defined in Definition 1. (ii) $\delta_e : \Sigma \rightarrow 2^{\mathcal{X}}$ maps a state to the next input state. (iii) $\delta_s : \Sigma \times 2^{\mathcal{X}} \rightarrow 2^{\Sigma}$ maps a state σ and the next input state $\sigma_{\mathcal{X}}$ to possible next states. (iv) $\gamma_g : \Sigma \rightarrow \mathbb{N}$ maps a state σ to the index of a system liveness goal the state attempts to satisfy.

When synthesis produces a strategy, we deploy the strategy on robots as the high-level controller. During the robot execution, environment safety assumptions φ_e^t may be *violated*.

Definition 3. An *assumption violation* is a triplet $t = (\sigma_{\mathcal{X}}, \sigma_{\mathcal{Y}}, \sigma'_{\mathcal{X}})$ such that $\exists i, \sigma_{\mathcal{X}} \cup \sigma_{\mathcal{Y}} \cup \bigcirc \sigma'_{\mathcal{X}} \not\models \varphi_{e,i}^t$. In Example 1, $\sigma_{\mathcal{X}} := \{\pi_{base}^{x_2}, \pi_{cup}^{ee}, \pi_{cone}^{x_3}, \pi_{block}^{t_0}, \pi_{stone}^{t_3}\}$, $\sigma_{\mathcal{Y}} := \{y_{original}\}$, and $\sigma'_{\mathcal{X}} := \sigma_{\mathcal{X}} \cup \{\pi_{empty}\}$.

4 PROBLEM STATEMENT

We tackle the runtime repair problem of high-level robot controllers under assumption violations.

Problem 1. Given (i) an abstraction of the environment $\mathcal{X}$ and skills $\mathcal{Y}$, (ii) specifications consisting of a task specification φ_{task} and a GR(1) specification φ such that $\varphi_{\text{task}} \preceq \varphi$, (iii) an assumption violation $t = (\sigma_{\mathcal{X}}, \sigma_{\mathcal{Y}}, \sigma'_{\mathcal{X}})$, and (iv) a set of example strategies and their descriptions $\{(\mathcal{A}_s^i, \mathcal{D}^i)\}_{i=1}^k$; find a set of new skills $\mathcal{Y}_{\text{new}}$ that makes the specification φ realizable under the violation t , at runtime.

5 APPROACH

To address Problem 1 in large state spaces, we propose a hybrid framework INPROVE, as depicted in Fig. 2, that leverages LLMs to efficiently produce new skill candidates, and formal methods to verify

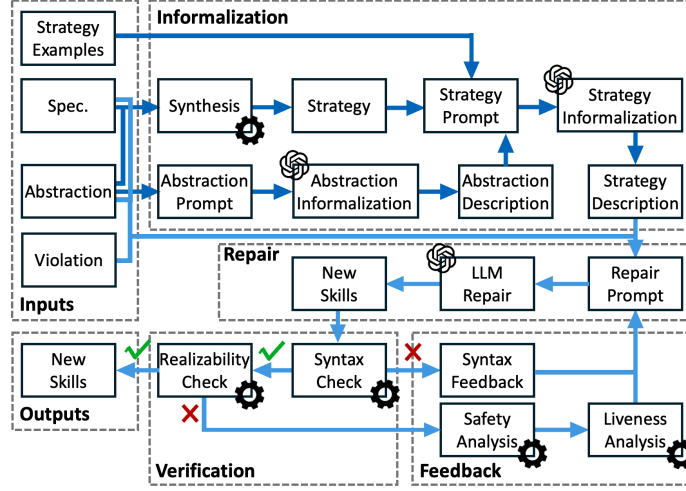


Figure 2: INPROVF system details.

them. We first informalize the abstraction into the NL description of its physical meaning. We then synthesize a strategy from the specification and use few-shot prompting to obtain the NL description of the strategy behavior (Sec. 5.1). Next, we leverage LLMs to generate a set of new skill candidates that potentially repair the behavior of the strategy under the violation (Sec. 5.2). We verify the candidates through syntax and realizability checking (Sec. 5.3). If the candidates cannot repair the violation, we provide feedback on syntax and counterstrategy analyses to the LLMs (Sec. 5.4). Appendix C presents a case study that walks through our framework with Example 1.

5.1 INFORMALIZATION

We begin by leveraging LLMs to translate the environment abstraction and the synthesized strategy into informal natural language descriptions. Since LLMs are primarily trained on natural language data and optimized for a wide range of tasks on NL traces, we hypothesize that LLMs better utilize the abstraction and strategy behaviors in natural languages than in the formal representations.

Abstraction informalization. We take in an abstraction consisting of the environment inputs $\mathcal{X}$ and a grounding function G that maps each input to a set of physical states. We prompt LLMs to describe the physical meaning of the abstraction. Appendix A.1 presents the prompt, and Appendix C shows the informalized abstraction description for Example 1.

Strategy informalization. Next, we obtain the behavior description of the strategy $\mathcal{A}_s$, synthesized from the GR(1) specification φ . Since each state of the strategy consists of environment input propositions, we leverage the previously generated environment abstraction description to provide the context for LLMs to ingest the strategy behavior. Our preliminary study found that this is a relatively challenging task due to the complexity of the strategy, which is a finite automaton as described in Definition 1. So, we use one-shot prompting—including one example translation in the prompt—to help steer LLMs towards generating accurate strategy behavior. Appendix A.2 describes the one-shot prompt and Appendix C.2 shows the resulting strategy behavior for Example 1. Appendix A.2.1 shows the example translation for Example 1.

5.2 REPAIR

The repair process takes in the description of the abstraction $\mathcal{X}$, the existing robot skills $\mathcal{Y}$, the task specification $\varphi_{\text{task}} \preceq \varphi$, provided by the user, that only contains the essential constraints of the full GR(1) specification φ , the assumption violation $t = (\sigma_{\mathcal{X}}, \sigma_{\mathcal{Y}}, \sigma'_{\mathcal{X}})$, and the behavior description $\mathcal{D}$ of the strategy $\mathcal{A}_s$ synthesized from φ . We form a prompt that combines the inputs and asks LLMs to only produce a set of new skill candidates $\mathcal{Y}_{\text{new}}$ in a predefined domain-specific language (DSL). We also require the new skills to be formatted as JSON files only, and we perform regular expression matching to extract $\mathcal{Y}_{\text{new}}$ from the responses. Appendix A.3 presents the repair prompt, and Appendices C.3 and C.6 show two example repair candidates for Example 1.

Algorithm 1: Safety Analysis**Input:** Counterstrategy $\mathcal{A}_e$, System safety constraints $\varphi_s^t = \bigwedge_i \Box \varphi_{s,i}^t$ **Output:** NL feedback on violated safety constraints

```

1 for  $\sigma, \sigma' \in \Sigma$  such that  $\sigma' \in \delta_s(\sigma, \delta_e(\sigma))$  and  $\delta_s(\sigma', \delta_e(\sigma')) = \emptyset$  do
2   for  $\Box \varphi_{s,i}^t$  in  $\varphi_s^t$  do
3     if  $\gamma_{\mathcal{X}}(\sigma) \cup \gamma_{\mathcal{Y}}(\sigma) \cup \bigcirc(\gamma_{\mathcal{X}}(\sigma') \cup \gamma_{\mathcal{Y}}(\sigma')) \not\models \varphi_{s,i}^t$  then
4       yield to feedback: “skills in  $\gamma_{\mathcal{Y}}(\sigma)$  violate the safety constraints  $\Box \varphi_{s,i}^t$ .”

```

5.3 VERIFICATION

After obtaining a set of new skill candidates $\mathcal{Y}_{\text{new}}$, we verify the correctness of $\mathcal{Y}_{\text{new}}$ through syntax and realizability checking, ensuring that $\mathcal{Y}_{\text{new}}$ indeed repairs the violation.

Syntax check. We check the syntax of $\mathcal{Y}_{\text{new}}$ against the DSL grammar. We leverage a parser to automatically transform $\mathcal{Y}_{\text{new}}$ into a set of abstract syntax trees (ASTs). If parsing is successful, the new skills are in the correct syntax. We also ensure type correctness by recursively traversing each node of the ASTs, verifying each component of $\mathcal{Y}_{\text{new}}$ against its expected type.

Realizability check. Once $\mathcal{Y}_{\text{new}}$ is in the correct syntax, we check whether $\mathcal{Y}_{\text{new}}$ can repair the assumption violation. We use a compiler from Meng & Kress-Gazit (2024) to add $\mathcal{Y}_{\text{new}}$ to φ and relax the assumptions in φ that are violated by the violation $t = (\sigma_{\mathcal{X}}, \sigma_{\mathcal{Y}}, \sigma'_{\mathcal{X}})$, resulting in an updated specification φ' . We then check the realizability of φ' . If φ' is realizable, then $\mathcal{Y}_{\text{new}}$ is verified to be sufficient to solve the violation.

5.4 FEEDBACK

If the new skills $\mathcal{Y}_{\text{new}}$ do not pass the verification step, we provide three types of feedback, each of which targets a different type of error, to ensure that we have informative and yet concise NL feedback to LLMs for iterative prompting.

Syntax feedback. For syntax errors detected in the syntax-checking procedure, we provide feedback based on the nodes of the ASTs that violate the grammar and the grammar rules that are violated. For instance, for a new skill $y_{\text{new}} \in \mathcal{Y}_{\text{new}}$ that contains an uncontrollable input $\pi_u \in \mathcal{X}_u$ in its precondition, we automatically provide the following feedback: “ π_u in the precondition of y_{new} is not a controllable input.”

Counterstrategy. Next, if the new skills $\mathcal{Y}_{\text{new}}$ are syntactically correct but the updated specification φ' remains unrealizable, meaning skills in $\mathcal{Y}_{\text{new}}$ do not repair the violation, we compute a counterstrategy $\mathcal{A}_e$ to provide feedback. While the counterstrategy contains information about the cause of the unrealizability of φ' , LLMs may not be able to utilize the formal representation of the counterstrategy (as shown in Definition 2). We thus perform formal analysis on the counterstrategy to transform it into concise NL feedback.

Safety analysis. Given the counterstrategy $\mathcal{A}_e$ and the system safety constraints $\varphi_s^t = \bigwedge_i \Box \varphi_{s,i}^t$, we provide feedback on if and which skill violates any safety constraint $\Box \varphi_{s,i}^t$. As shown in Algorithm 1, We first identify any transition (σ, σ') in $\mathcal{A}_e$ such that σ' is a sink state, i.e. there are no outgoing edges from σ' (Line 1). We are interested in sink states because those are the states where the system has no valid next move, indicating states that violate safety constraints. We then iterate through all system safety constraints $\Box \varphi_{s,i}^t$ (Line 2), and check whether the transition violates any safety constraint (Line 3). We perform this check by formulating a conjunction of the labels of σ , the labels of σ' , and the safety constraint of interest, and then leveraging an SMT solver Z3 De Moura & Bjørner (2008) to check the satisfiability of the conjunction, as done in Raman & Kress-Gazit (2013). If the constraint $\Box \varphi_{s,i}^t$ is violated, we provide the following NL feedback: “the skills in $\gamma_{\mathcal{Y}}(\sigma)$ violate the hard constraint $\Box \varphi_{s,i}^t$ ” (Line 4). Appendix C.5 presents an example safety analysis feedback for Example 1.

Liveness analysis. Given the counterstrategy $\mathcal{A}_e$ and the system liveness goals $\varphi_s^g = \bigwedge_{i=1}^n \Box \Diamond \varphi_{s,i}^g$, we provide feedback on which liveness goals are not satisfied given the new skills. Algorithm 2 describes our approach. We first leverage Tarjan’s algorithm Tarjan (1972) to identify a set of strongly

Algorithm 2: Liveness Analysis**Input:** Counterstrategy $\mathcal{A}_e$, Liveness goals $\varphi_s^g = \bigwedge_{i=1}^n \Box\Diamond\varphi_{s,i}^g$ **Output:** NL feedback on unsatisfied liveness goals

```

1  $SCCs := \text{find\_strongly\_connected\_components}(\mathcal{A}_e);$ 
2 for  $S \in SCCs$  do
3   if  $(\forall \sigma \in S, \nexists \sigma' \notin S. \sigma' \in \delta_s(\sigma, \delta_e(\sigma))) \wedge (\exists \sigma \in S. \delta_s(\sigma, \delta_e(\sigma)) \neq \emptyset)$  then
4      $L := \{i \mid \exists \sigma, \sigma' \in S. \sigma' \in \delta_s(\sigma, \delta_e(\sigma)), \gamma_g(\sigma) \leq i < \gamma_g(\sigma') \pmod n\};$ 
5      $L' := \{i \mid \exists \sigma \in S. \gamma_g(\sigma) = i\} \setminus L;$ 
6     return “The new skills cannot satisfy liveness goals  $\{\Box\Diamond\varphi_{s,i}^g \mid i \in L'\}$  after satisfying
       liveness goals  $\{\Box\Diamond\varphi_{s,i}^g \mid i \in L\}$ .”

```

connected sets in the nodes Σ of $\mathcal{A}_e$ (Line 1). We then only consider the sinking ones, i.e. there is no outgoing edge towards other components, and are not singletons without transitions (Line 3). Thus, we identify a sinking strongly connected component that can trap the robot within it once any state of the component is reached. We then identify the indices of a set of liveness goals L satisfied in the component (Line 4). We note that the inequality comparisons are modulo n (where $n \in \mathbb{N}$ is the total number of system liveness goals arranged in a cyclic order); that is, the condition $\gamma_g(\sigma) \leq i < \gamma_g(\sigma') \pmod n$ in Line 4 means that $\gamma_g(\sigma) \leq i < \gamma_g(\sigma')$ if $\gamma_g(\sigma) \leq \gamma_g(\sigma')$, and $(\gamma_g(\sigma) \leq i \leq n) \vee (1 \leq i < \gamma_g(\sigma'))$ if $\gamma_g(\sigma) > \gamma_g(\sigma')$. Next, we obtain the indices of another set of liveness goals L' that the robot attempts to, but cannot, satisfy in the component (Line 5). Lastly, we provide feedback which states that the robot cannot satisfy the liveness goals in L' after satisfying L (Line 6). Appendix C.8 presents an example feedback from liveness analysis for Example 1.

Both the safety and liveness analyses have computation complexity $O(|\Sigma|^2)$ since they only access each state and transition of the counterstrategy at most a finite number of times. Our feedback allows LLMs to correct their mistakes and generate more promising repair candidates.

6 EXAMPLES

Our examples aim to explore the following questions: **Q1.** Does INPROVF repair violations more efficiently than the state-of-the-art formal-methods-based approach? **Q2.** How does each component (informalization, feedback) contribute to the overall performance of INPROVF?

Examples. We evaluate 12 violations in four workspaces over two abstraction types (object-centric and swarm). Detail descriptions of the examples can be found in Appendix B.

Baseline and Ablations. Our previous work Meng & Kress-Gazit (2024) provides a synthesis-based baseline that modifies the pre and post-conditions of existing skills to suggest new skills that repair the violations. We also consider two ablations of INPROVF: *no-inform* where we do not perform informalization and instead directly include the abstraction and strategy in the repair prompt; and *no-feedback*, where we perform the iterative prompting without feedback from formal analyses.

Setup. We run our experiments on a Linux machine with a 2.3GHz 8-core AMD RyzenTM-7 CPU. We use OpenAI o1-preview-2024-09-12 as the LLM for INPROVF. We measure runtime in seconds. We set the runtime limit to 3000 seconds. For INPROVF and its ablations, we set a maximum limit for feedback iteration to be 5 for monetary cost control of LLMs. We run 5 trials with different random seeds for the baseline and 3 trials for INPROVF and its ablations for each violation.

Results. Fig. 3 presents the success rates and the runtime comparison for the successful trails. For INPROVF and its ablations, we also show the detailed runtime breakdown of each key component. We do not include the runtime for informalization in Fig. 3 because we only perform informalization once offline for each example; other violations under the same example can reuse the informalization results. Table 1 shows the average runtime and standard deviation for each informalization component over three trials per example.

Analysis. In the *factory* example with a small state space (37 variables), the average runtime of formal-methods-based baseline (144.05 s) is comparable to that of INPROVF (92.36 s). However, for the three larger examples (64-79 variables), INPROVF significantly outperforms the baseline. In

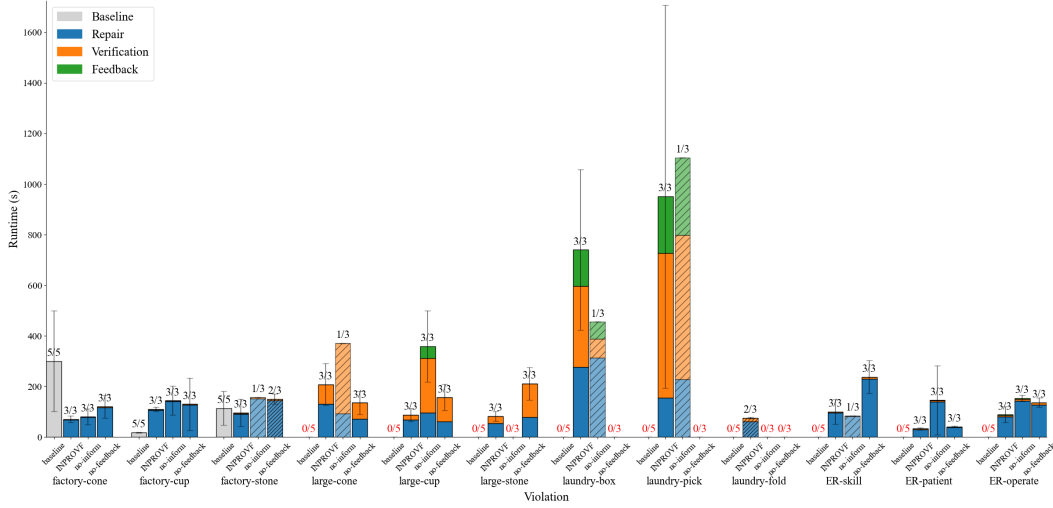


Figure 3: Comparison of success rates (n/m) and runtime breakdowns (with std error bars for total runtime) for baseline, INPROVF, and its ablations (no-inform, no-feedback) on successful runs.

Table 1: Informalization average runtime with standard deviation (s)

	Abstraction-inform	Synthesis	Strategy-inform
Factory	16.32 ± 6.91	1.94 ± 0.58	65.79 ± 29.11
Large	16.14 ± 10.78	90.43 ± 3.48	44.46 ± 12.09
Laundry	17.83 ± 7.78	24.81 ± 2.83	42.99 ± 12.03
ER	10.17 ± 4.64	1.79 ± 0.03	27.63 ± 8.60

these larger cases, the baseline fails to complete within the 3000-second time limit. We also remove the time limit for the *larger-cup* violation, but the baseline does not terminate after 18523.93 s, eventually causing a machine out-of-memory crash. In contrast, for these larger examples, within five iterations, INPROVF successfully repairs every violation except one *in a single run*; the failed run is for the *laundry-fold* violation. These results indicate that INPROVF is more scalable and efficient than the baseline in large state space (Q1). Our ablation studies show that INPROVF achieves a higher success rate (97.22%) than *no-inform* (55.56%) and *no-feedback* (72.22%). Additionally, in the successful runs, the runtime for *no-inform* is on average 7.43% higher than that of INPROVF. These results suggest both informalization and feedback are critical for INPROVF to achieve high performance in runtime repair (Q2).

7 LIMITATIONS AND CONCLUSION

We present INPROVF, a novel hybrid framework that combines the scalability of LLMs and the guarantees of formal methods to efficiently repair high-level robot controllers under assumption violations at runtime with prohibitively large state space. Our experiments across various state space sizes indicate that our approach outperforms the formal methods-based approach in settings with large state spaces. Nonetheless, our evaluation faces several limitations. First, since there is no standard benchmark for runtime repair, our evaluation instances are limited. In future work, we plan to leverage LLMs to procedurally generate a diverse range of repair instances across various abstraction types for a more comprehensive evaluation. Moreover, the formal-methods-based repair executes on CPUs, and its implementation has not been optimized, while the LLM components of INPROVF run on GPUs and benefit from extensive engineering optimizations. Thus, a direct runtime comparison may not be fair. Future work will focus on developing evaluation metrics that better account for the hardware disparities. Finally, we currently assess the quality of informalization indirectly through the final repair correctness, which is neither sound nor complete. Future work will explore automatic evaluation methods that directly measure the quality of informalization. Additionally, we will work on further improving the scalability of formal methods components within our framework, such as the realizability check.

ACKNOWLEDGMENTS

We thank Lin Luan for her invaluable assistance in enhancing the visual presentation of Figs. 1, 2, and 5, Kevin Guan for his insightful discussion on the design of the experiments in Sec. 6, Bo-Ruei Huang and Andrew Violette for their insightful discussion on the ER example (Fig. 5 right), and the anonymous reviewers from ICLR 2025 VerifAI Workshop for their extremely helpful suggestions.

REFERENCES

- Rajeev Alur, Salar Moarref, and Ufuk Topcu. Counter-strategy guided refinement of gr (1) temporal logic specifications. In *2013 Formal Methods in Computer-Aided Design*, pp. 26–33. IEEE, 2013.
- Roderick Bloem, Barbara Jobstmann, Nir Piterman, Amir Pnueli, and Yaniv Sa’ar. Synthesis of reactive (1) designs. *Journal of Computer and System Sciences*, 78(3):911–938, 2012.
- Yongchao Chen, Rujul Gandhi, Yang Zhang, and Chuchu Fan. Nl2tl: Transforming natural languages to temporal logics using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15880–15903, 2023.
- E Clarke, O Grumberg, and D Peled. *Model Checking*. The MIT Press, 2000.
- Leonardo De Moura and Nikolaj Bjørner. Z3: An efficient SMT solver. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pp. 337–340. Springer, 2008.
- Rüdiger Ehlers and Vasumathi Raman. Slugs: Extensible GR(1) synthesis. In *Computer Aided Verification: 28th International Conference, CAV 2016, Toronto, ON, Canada, July 17-23, 2016, Proceedings, Part II* 28, pp. 333–339. Springer, 2016.
- Rüdiger Ehlers and Ufuk Topcu. Resilience to intermittent assumption violations in reactive synthesis. In *Proceedings of the 17th International Conference on Hybrid Systems: Computation and Control (HSCC)*, pp. 203–212, 2014.
- Rüdiger Ehlers, Robert Könighofer, and Roderick Bloem. Synthesizing cooperative reactive mission plans. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3478–3485. IEEE, 2015.
- Albert Qiaochu Jiang, Sean Welleck, Jin Peng Zhou, Wenda Li, Jiacheng Liu, Mateja Jamnik, Timothée Lacroix, Yuhuai Wu, and Guillaume Lample. Draft, Sketch, and Prove: Guiding formal theorem provers with informal proofs. In *International Conference on Learning Representations*, 2023. URL <https://doi.org/10.48550/arXiv.2210.12283>.
- George Konidaris, Leslie Pack Kaelbling, and Tomas Lozano-Perez. From skills to symbols: Learning symbolic representations for abstract high-level planning. *Journal of Artificial Intelligence Research*, 61:215–289, 2018.
- Robert Könighofer, Georg Hofferek, and Roderick Bloem. Debugging formal specifications using simple counterstrategies. In *2009 Formal Methods in Computer-Aided Design*, pp. 152–159. IEEE, 2009.
- Hadas Kress-Gazit, Morteza Lahijanian, and Vasumathi Raman. Synthesis for robots: Guarantees and feedback for robot behavior. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:211–236, 2018.
- Haonan Li, Yu Hao, Yizhuo Zhai, and Zhiyun Qian. Enhancing static analysis for practical bug detection: An llm-integrated approach. *Proceedings of the ACM on Programming Languages*, 8 (OOPSLA1):474–499, 2024a.
- Ziyang Li, Saikat Dutta, and Mayur Naik. Llm-assisted static analysis for detecting security vulnerabilities. *arXiv preprint arXiv:2405.17238*, 2024b.
- Chang Liu, Xiwei Wu, Yuan Feng, Qinxiang Cao, and Junchi Yan. Towards general loop invariant generation: A benchmark of programs with memory manipulation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

- Mingyang Lyu, Yibo Zhao, Chao Huang, and Hailong Huang. Unmanned aerial vehicles for search and rescue: A survey. *Remote Sensing*, 15(13):3266, 2023.
- Rishabh Madan, Skyler Valdez, David Kim, Sujie Fang, Luoyan Zhong, Diego T Virtue, and Tapomayukh Bhattacharjee. Rabbit: A robot-assisted bed bathing system with multimodal perception and integrated compliance. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 472–481, 2024.
- Qian Meng and Hadas Kress-Gazit. Automated Robot Recovery from Assumption Violations of High-Level Specifications. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pp. 4154–4161, Bari, Italy, August 2024. IEEE.
- Claudio Menghi, Christos Tsigkanos, Patrizio Pelliccione, Carlo Ghezzi, and Thorsten Berger. Specification patterns for robotic missions. *IEEE Transactions on Software Engineering*, 47(10):2208–2224, 2019.
- Salar Moarref and Hadas Kress-Gazit. Automated synthesis of decentralized controllers for robot swarms from high-level temporal logic specifications. *Autonomous Robots*, 44(3):585–600, 2020.
- OpenAI. Learning to reason with LLMs, 2024. URL <https://openai.com/index/learning-to-reason-with-llms/>.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.
- Adam Pacheck and Hadas Kress-Gazit. Physically feasible repair of reactive, linear temporal logic-based, high-level tasks. *IEEE Transactions on Robotics*, 2023.
- Muhammad AA Pirzada, Giles Reger, Ahmed Bhayat, and Lucas C Cordeiro. Llm-generated invariants for bounded model checking without loop unrolling. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pp. 1395–1407, 2024.
- Vasumathi Raman and Hadas Kress-Gazit. Towards minimal explanations of unsynthesizability for high-level robot behaviors. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 757–762. IEEE, 2013.
- Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1(2): 146–160, 1972.
- Antti Valmari. The state explosion problem. In *Advanced Course on Petri Nets*, pp. 429–528. Springer, 1996.
- Wenhan Wang, Kaibo Liu, An Ran Chen, Ge Li, Zhi Jin, Gang Huang, and Lei Ma. Python symbolic execution with LLM-powered code generation. *arXiv preprint arXiv:2409.09271*, 2024.
- Kai Weng Wong, Rüdiger Ehlers, and Hadas Kress-Gazit. Resilient, provably-correct, and high-level robot behaviors. *IEEE Transactions on Robotics*, 34(4):936–952, 2018.
- Guangyuan Wu, Weining Cao, Yuan Yao, Hengfeng Wei, Taolue Chen, and Xiaoxing Ma. Llm meets bounded model checking: Neuro-symbolic loop invariant inference. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pp. 406–417, 2024.
- Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan J Prenger, and Animashree Anandkumar. Leandojo: Theorem proving with retrieval-augmented language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 21573–21612. Curran Associates, Inc., 2023.
- Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

APPENDIX

CONTENTS

A Prompts	12
A.1 Abstraction Informalization	12
A.2 Strategy Informalization	12
A.3 Repair	14
B Example descriptions	15
B.1 Overview	15
B.2 Factory and Large	15
B.3 Laundry	17
B.4 ER	18
C Case study	19
C.1 Abstraction informalization	19
C.2 Strategy informalization	19
C.3 Repair prompt	20
C.4 Verification	20
C.5 Safety analysis	20
C.6 Repair prompt - second iteration	21
C.7 Repair prompt (a different run)	22
C.8 Liveness analysis	22

A PROMPTS

We describe the prompts used in INPROVF.

A.1 ABSTRACTION INFORMALIZATION

System prompt: You are a researcher in robotics and an expert in formal methods for robotics.

User prompt: You are given an abstraction that contains a set of input propositions X , a physical state space $PS \subseteq R^n$, and a grounding function $G: X \rightarrow 2^{PS}$ that maps each input proposition to a set of physical states. You are asked to explain the physical meaning of the input propositions.

Input propositions: **<Inserted input propositions $\mathcal{X}$ >**

Grounding: **<Inserted grounding function G >**

Please only provide a concise, general explanation covering all input propositions' physical meaning. Do not provide any examples.

<Informalized abstraction to be returned here>

A.2 STRATEGY INFORMALIZATION

We leverage one-shot prompts for strategy formalization.

System prompt: You are a researcher in robotics and an expert in formal methods for robotics.

User prompt: You are given an abstraction of the environment, a set of robot skills, a GR(1) specification, and its corresponding synthesized strategy. You are asked to describe the behavior of the strategy. I will first give an example of a strategy, its behavior description, and its explanation. Then, I will ask you to describe the behavior of a different strategy.

Example: **<Inserted strategy informalization example $(\mathcal{A}_s^1, \mathcal{D}^1)$ >**

Problem: Now, I will ask you to describe the behavior of a different strategy. You are given: (1) an abstraction of the physical workspace and objects in the workspace, (2) a set of robot skills, (3) a task specification, and (4) a synthesized robot controller.

Input propositions: **<Inserted inputs $\mathcal{X}$ and informalized abstraction>**

Output propositions: **<Inserted output propositions $\mathcal{Y}$ >**

Task specification: We use the GR(1) fragment of Linear Temporal Logic to encode the task specification. The task specification is given in the following JSON file: **<Inserted task specification φ_{task} >**

Strategy: We represent the robot strategy synthesized from the specification as a finite state automaton. For each node, 'rank' represents which liveness goal the node is trying to satisfy, 'state' represents the value of each variable in the state, in the order of the field 'variables', and 'trans' represents the possible successor states. The strategy is in the JSON file: **<Inserted strategy $\mathcal{A}_s$ >**

Please describe the behavior of the synthesized strategy only. Do not provide any explanation: **<Informalized strategy behavior to be returned here>**

A.2.1 STRATEGY INFORMALIZATION EXAMPLE FOR OBJECT-CENTRIC ABSTRACTION

We use the following strategy informalization example for object-centric abstraction including the following examples: *factory*, *large*, and *laundry*.

Inputs: ["p_base_x0", "p_base_x1", "p_base_x2", "p_base_x3"]

Outputs: ["skill0", "skill1"]

Task specification and strategy: <omitted for space >

Behavior: Initially, the robot's base is in x0, and the liveness goal 'p_base_x0' which requires the base to repeatedly reach x0 is satisfied. The robot then executes skill0 to move to x1 and then x3, satisfying the liveness goal 'p_base_x3' which requires the base to repeatedly reach x3. Next, the robot executes skill1 to move back to x0 via x1, satisfying the liveness goal 'p_base_x0' again. This behavior continues indefinitely.

Explanation: "Initially, the robot's base is in x0, and the liveness goal 'p_base_x0' which requires the base to repeatedly reach x0 is satisfied" corresponds to node 0. "The robot then executes skill0 to move to x1 and then x3, satisfying the liveness goal 'p_base_x3' which requires the base to repeatedly reach x3" corresponds to the transitions from node 0 to node 1, node 2, node 3, and then node 4. "Next, the robot executes skill1 to move back to x0 via x1, satisfying the goal 'p_base_x0' again" corresponds to the transitions from node 4 to node 5, node 6, node 7, and then back to node 0.

A.2.2 STRATEGY INFORMALIZATION EXAMPLE FOR SWARM ABSTRACTION

We use the following strategy informalization example for swarm abstraction including the example *ER*.

Inputs: ["p_swarm_x0", "p_swarm_x1", "p_swarm_x2", "p_swarm_x3"]

Outputs: ["skill0", "skill1", "skill2", "skill3"]

Task specification and strategy: <omitted for space >

Behavior: Initially, the robot swarm is in x0 and the liveness goal 'p_swarm_x0 & !p_swarm_x1 & !p_swarm_x2 & !p_swarm_x3' which requires the swarm to repeatedly stay in x0 only is satisfied. The swarm then executes skill0 to reach x1 and x2, and then executes skill1 to reach x3, satisfying the goal 'p_swarm_x3 & !p_swarm_x0 & !p_swarm_x1 & !p_swarm_x2' which requires the swarm to repeatedly stay in x3 only. Next, the robot executes skill2 to move to x1 and x2, and finally executes skill3 to reach x0, satisfying the first goal again. This behavior continues indefinitely.

Explanation: "Initially, the robot swarm is in x0 and the liveness goal 'p_swarm_x0 & !p_swarm_x1 & !p_swarm_x2 & !p_swarm_x3' which requires the swarm to repeatedly stay in x0 only is satisfied" corresponds to node 0 and 1. "The swarm then executes skill0 to reach x1 and x2" corresponds to the transitions from node 1 to node 2, node 2 to node 3 or node 4, and node 3 or node 4 to node 5 and then node 6. "and then executes skill1 to reach x3, satisfying the goal 'p_swarm_x3 & !p_swarm_x0 & !p_swarm_x1 & !p_swarm_x2' which requires the swarm to repeatedly stay in x3 only" corresponds to the transitions from node 6 to node 7, and then node 8 or node 9, and then node 10. "Next, the robot executes skill2 to move to x1 and x2" corresponds to the transitions from node 10 to node 11 or node 12, and then to node 13, node 14, and node 15. "and finally executes skill3 to reach x0, satisfying the first goal again" corresponds to the transitions from node 15 to node 16 or node 17, and then to node 0.

A.3 REPAIR

System prompt: You are a researcher in robotics and an expert in formal methods for robotics.

User prompt: You are given: (1) an abstraction of the physical workspace and objects in the workspace, (2) a set of robot skills, (3) a task specification, (4) the original behavior of the synthesized strategy, and (5) a violated observation during the robot execution. You are asked to provide a set of new skills that along with the original skills can allow the robot to recover from the violation and satisfy the task specification.

Abstraction:

Input propositions: <Inserted inputs and informalized abstraction description>

Output propositions: <Inserted output propositions>

Task specification:

We use the GR(1) fragment of Linear Temporal Logic to encode the task specification. The task specification is given in the following JSON file: <Inserted task specification>

Strategy behavior: <Inserted informalized strategy behavior description>

Assumption Violation: During the robot execution, we detect that an observed transition violates some environment safety assumptions.

The violated observation is in the following JSON file:

<Inserted assumption violation $t = (\sigma_x, \sigma_y, \sigma'_x)$ >

The violated environment safety assumptions are as follows:

<Inserted violated assumptions>

Returns: After the assumption violation, we relax the violated assumption to include the observed violation transition. You are asked to provide a set of new skills that, along with the original skills, allow the robot to continue satisfying the task after the assumption violation. Please respond with new skills only in the following format:

```
““JSON
{
  "new_skill_0": [
    [ <PRECONDITION>,
      [<POSTCONDITION>, ...] ],
    [ <PRECONDITION>,
      [<POSTCONDITION>, ...] ],
    ... ],
  "new_skill_1": [
    [ <PRECONDITION>,
      [<POSTCONDITION>, ...] ],
    [ <PRECONDITION>,
      [<POSTCONDITION>, ...] ],
    ... ],
  ...
}
““
```

The response follows the grammar:

```
<new_skill>: [<intermediate_transition>+]
<intermediate_transition> = [<precondition>, [<postcondition>+]]
<precondition> = [<controllable_input>+]
<postcondition> = [<controllable_input>+]
```

Each pre or post-condition in the new skills should be a complete controllable input state. Do not include original skills in the new skills. Do not include any comments in skills.

<New skills to be returned here>

B EXAMPLE DESCRIPTIONS

B.1 OVERVIEW

We evaluate 12 violations in four workspaces over two abstraction types (object-centric and swarm).

- **Factory.** A mobile manipulator operates in a factory-like workspace (Fig. 4). During the execution, the robot encounters three violations: *factory-cone* where an obstacle *cone* moves to the front of the robot, *factory-cup* where the status of a *cup* changes (Example 1), and *factory-stone* where a static obstacle *stone* unexpectedly moves. This example is from Sec. VC of Meng & Kress-Gazit (2024).
- **Large.** A larger version of the factory example with additional regions surrounding the workspace in Fig. 1, increasing the number of input propositions from 29 to 71. The three violations, *large-cone*, *large-cup*, and *large-stone*, are the same as those in the factory example.
- **Laundry.** A mobile manipulator is responsible for laundry tasks in an apartment-like workspace (Fig. 5 left). In the violation *laundry-box*, an obstacle *box* blocks the path to a folding table. In the violation *laundry-pick*, the skill that picks up a cloth fails, and the robot should attempt to pick up another cloth instead. In the violation *laundry-fold*, a folding table is malfunctioning, and the robot should use another folding table instead.
- **ER.** A swarm of robots is in an emergency room (ER)-like workspace (Fig. 5 right). In the violation *ER-skill*, a low-level controller failure causes some robots to enter an undesired region. In the violation *ER-patient*, a patient enters the ER unexpectedly, requiring the swarm to visit the ER in response. In the violation *ER-operate*, a patient appears in an operation room unexpectedly, forcing the swarm to patrol other operation rooms instead.

Informalization examples. For strategy informalization (Sec. 5.1), we use an example in Appendix A.2.1 for object-centric abstraction (*factory*, *large*, and *laundry*) and the one in Appendix A.2.2 for swarm abstraction (*ER*).

B.2 FACTORY AND LARGE

Setup. As shown in Fig. 4, a mobile manipulator operates in a factory-like environment with five objects (including the robot base) and 11 regions (including the robot’s end-effector). The resulting abstraction contains 29 input propositions and eight output propositions (robot skills).

Abstractions. The input propositions are of the form π_o^r , representing that the object o is in the region r . Additionally, a user-controlled input π_{empty} represents the status of the cup (whether it is empty or not). The robot is initially given two mobile skills that allow the robot to move between x_4 and x_0 via x_2 , and six manipulation skills. Among them, three pick-up skills include picking up the controllable obstacle *block* from t_1 , picking up *cup* from t_2 and t_4 . The other three skills place *block* to t_3 , *cup* to t_2 , and to t_4 .

Tasks. We task the robot to move the object *cup* to a loading table t_2 whenever the *cup* is empty, and move it to an assembly table t_4 whenever the *cup* is full. Formally, $(\Box \Diamond \pi_{\text{empty}} \rightarrow \pi_{\text{cup}}^{t_2}) \wedge (\Box \Diamond \neg \pi_{\text{empty}} \rightarrow \pi_{\text{cup}}^{t_4})$.

Safety guarantees. The robot is asked to satisfy the following safety guarantees.

1. **Collision avoidance.** For any different objects o_1 and o_2 , and for any region r , it must hold that $\Box(\neg(\pi_{o_1}^r \wedge \pi_{o_2}^r) \wedge \neg(\bigcirc \pi_{o_1}^r \wedge \bigcirc \pi_{o_2}^r))$.
2. **Blocking constraint.** The robot cannot pick the *cup* if *block* is in front of the cup.
3. **Entrance restrictions.** The robot is not allowed to bring an empty *cup* to enter the assembly region x_0 , or bring a full *cup* to the loading region x_4 .

Safety assumptions. The environment behaviors are assumed to obey the following constraints:

1. **Static obstacles.** Uncontrollable obstacles *stone* and *cone* are assumed to be static.

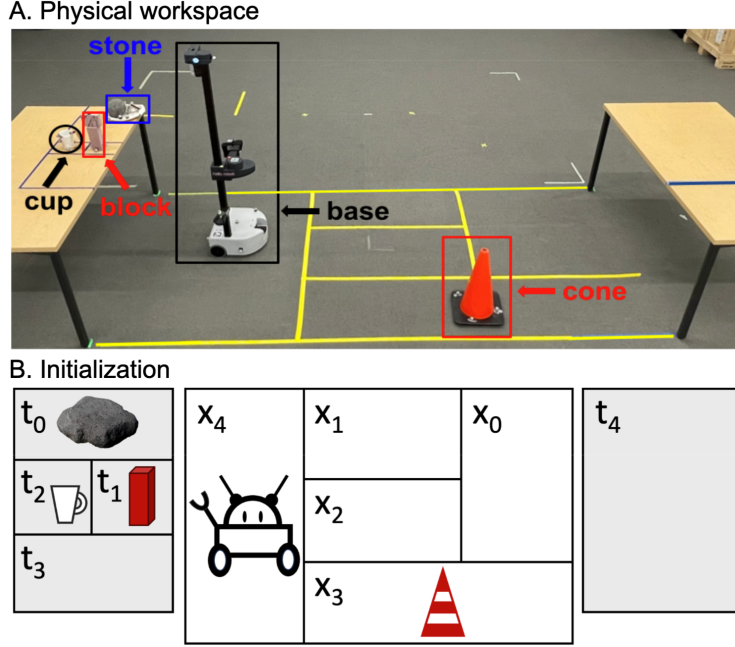


Figure 4: Workspace from Meng & Kress-Gazit (2024). (A) Physical workspace of the setup. (B) Initial position of each object in the workspace.

2. **Cup status.** The robot assumes that the status of *cup* only changes in the loading table t_2 and the assembly table t_4 .

Violations. We consider the following violations of environment safety assumptions.

- **Factory-stone.** When the robot starts to pick up *block* from t_1 , the uncontrollable obstacle *stone* moves from t_0 to t_3 , violating the assumption that *stone* should be static.
- **Factory-cup.** When the robot *base* is about to enter x_0 from x_2 , with *cup* in the robot’s end-effector *ee*, the status of *cup* changes from full to empty, violating the assumption that the status change should only occur in the loading or assembly table.
- **Factory-cone.** When the robot starts to move from x_4 to x_2 , the uncontrollable obstacle *cone* moves from x_3 to x_2 , violating the assumption that *cone* should be static.

The three violations in the *large* example – *large-cone*, *large-cup*, and *large-stone* – are the same as *factory-cone*, *factory-*, and *factory-*, respectively, except that the sizes of their state space are different: we increase the number of regions from 11 to 30, and consequently the number of input propositions is increased from 29 to 71.

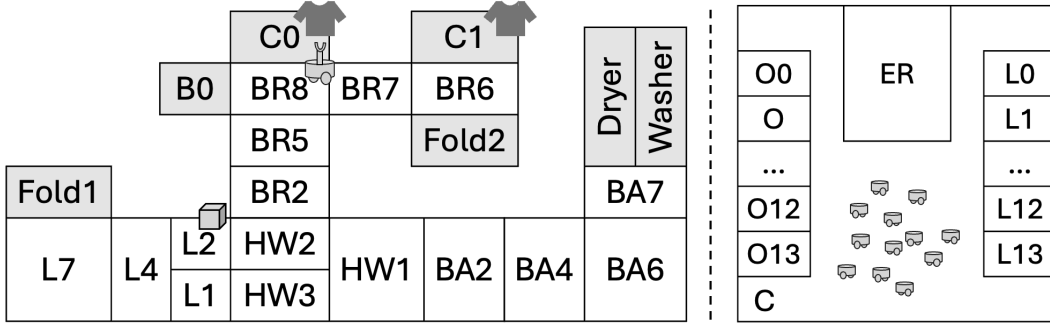


Figure 5: **Example workspace.** Left: An apartment-like workspace where a mobile manipulator performs laundry tasks. Right: An emergency room-like workspace where a robot swarm performs patrolling tasks.

B.3 LAUNDRY

Setup. As shown in Fig. 5 left, a mobile manipulator operates in an apartment-like environment with four objects (including the robot base) and 23 regions (including the robot’s end-effector). The resulting abstraction contains 55 input propositions and nine output propositions (robot skills).

Abstractions. We use the object-centric abstraction, where the input propositions are of the form π_o^r , representing that the object o is in the region r . We have six user-controlled inputs, $\pi_{\text{clean}}^0, \pi_{\text{clean}}^1, \pi_{\text{dry}}^0, \pi_{\text{dry}}^1, \pi_{\text{fold}}^0$, and π_{fold}^1 , that indicate the statuses of two clothes in the workspace. Additionally, a user-controlled input π_{wear} represents whether the user requests to wear a cloth or not. The robot is initially given six mobile skills to move between HW_2 and BR_8 , HW_2 and BA_7 , and HW_2 and L_7 , and three manipulation skills to pick up and drop off the clothes.

Tasks. We task the robot to move a cloth, cloth_0 or cloth_1 , to the bed area B_0 whenever the user requests to wear a cloth, and move them back to the closets C_0 or C_1 if the user does not request. Formally, $(\Box\Diamond\neg\pi_{\text{wear}} \rightarrow (\pi_{\text{cloth}_0}^{C_0} \vee \pi_{\text{cloth}_0}^{C_1}) \wedge (\pi_{\text{cloth}_1}^{C_0} \vee \pi_{\text{cloth}_1}^{C_1})) \wedge (\Box\Diamond\pi_{\text{wear}} \rightarrow (\pi_{\text{cloth}_0}^{B_0} \vee \pi_{\text{cloth}_1}^{B_0}))$.

Safety guarantees. The robot is required to satisfy the following safety guarantees.

1. **Cloth status.** The robot cannot bring a dirty or wet cloth to the bed area B_0 ; or a dirty, wet, or unfolded cloth to the closets C_0 and C_1 ; or a dirty or wet cloth to the folding tables Fold_1 or Fold_2 ; or a dirty cloth to the dryer.
2. **Collision avoidance.** Any two objects should not be in the same region (same as the constraint in *factory* example in Appendix B.2).
3. **Topology constraints.** The robot should follow the transition relations specified by the workspace (Fig. 5 left).
4. **Pickup skill constraint.** The robot cannot pick up the cloth from C_0 if it does not execute the pickup skill skill_6 : $\Box\pi_{\text{cloth}_0}^{C_0} \wedge \neg\text{skill}_6 \rightarrow \bigcirc\pi_{\text{cloth}_0}^{C_0}$.

Safety Assumptions. The environment behaviors are assumed to obey the following constraints:

1. **Static obstacles.** The uncontrollable obstacle *box* is assumed to be static (same as the constraint in *factory* example in Appendix B.2).
2. **Cloth statuses.** The cloth statuses can and will change in some regions. Specifically, a cloth can only change from clean to dirty if it is in the bed area B_0 ; from dry to wet if in the washer; from folded to unfolded if in B_0 , washer, or dryer. Moreover, a cloth will be cleaned in the next step if it is in the washer; dried in the next step if in the dryer; folded in the next step if in a folding table.
3. **Request status.** The user request status π_{wear} can only change if a cloth is in the bed area B_0 or both clothes are in the closets C_0 and C_1 . Specifically, π_{wear} can only change from True to False if a cloth is in B_0 ; from False to True if both clothes are in the closets.

Violations. We consider the following violations of environment safety assumptions.

- **Laundry-box.** When the robot holds the cleaned, dried, but unfolded $cloth_0$ in region HW_2 , the uncontrollable box moves from L_2 to L_4 , blocking the path towards the folding table $Fold_1$.
- **Laundry-pick.** When the robot is in region BR_8 executing $skill_6$ to pick up $cloth_0$, the pickup skill is faulty, and as an unexpected postcondition of the skill, $cloth_0$ remains in the closet $C0$.
- **Laundry-fold.** When the robot is in region L_7 placing $cloth_0$ to the folding table $Fold_1$, the folding table is malfunctioning and as a consequence, $cloth_0$ remains unfolded in the next step.

B.4 ER

Setup. As shown in Fig. 5 right, a swarm of mobile robots (the exact number of robots is unknown) is in an emergency room-like workspace in a hospital with 30 regions. The resulting abstraction contains 60 input propositions and four output propositions (robot skills).

Abstractions. We use the swarm abstraction following Moarref & Kress-Gazit (2020). We consider two groups of objects: *swarm* which represents the group of robots, and *patients* which represents a group of patients in the hospital. Then for any region r , the input proposition π_{swarm}^r represents that there exists a robot in region r , and similarly, π_{patient}^r represents that there exists a patient in region r . The robot is initially given four mobile skills: $skill_0$ moves the swarm from the corridor C to L_0 and L_9 ; $skill_1$ moves the swarm from L_0 and L_9 back to C ; $skill_2$ moves the swarm from C to O_0 and O_9 ; $skill_3$ moves the swarm from O_0 and O_9 back to C .

Tasks. The robot is asked to patrol between the operation rooms $O_0 \dots O_{13}$ and the lab $L_0 \dots L_{13}$. Specifically, for room type $\alpha \in \{L, O\}$, we partition the rooms into two subsets, $\{\alpha_0, \dots, \alpha_6\}$ and $\alpha_7, \dots, \alpha_{13}$. A robot in any room of the subset suffices to patrol all the rooms in the subset. Formally, we have $\Box \Diamond (\bigvee_{i \in \{0, \dots, 6\}} \pi_{\text{swarm}}^{O_i}) \wedge (\bigvee_{i \in \{7, \dots, 13\}} \pi_{\text{swarm}}^{O_i}) \wedge \neg (\bigvee_{i \in \{0, \dots, 13\}} \pi_{\text{swarm}}^{L_i}) \wedge \neg \pi_{\text{swarm}}^{ER} \wedge \Box \Diamond (\bigvee_{i \in \{0, \dots, 6\}} \pi_{\text{swarm}}^{L_i}) \wedge (\bigvee_{i \in \{7, \dots, 13\}} \pi_{\text{swarm}}^{L_i}) \wedge \neg (\bigvee_{i \in \{0, \dots, 13\}} \pi_{\text{swarm}}^{O_i}) \wedge \neg \pi_{\text{swarm}}^{ER}$. Moreover, we require some robots (not need to be all) to visit ER if there is a patient in ER : $\Box \Diamond \pi_{\text{patient}}^{ER} \rightarrow \pi_{\text{swarm}}^{ER}$.

Safety guarantees. The robot swarm is required to satisfy the following safety guarantees.

1. **Mutual exclusion in operation rooms.** The robot cannot interrupt the operation if a patient is in the operation room: $\forall i \in \{0, \dots, 13\}, \Box (\neg (\pi_{\text{swarm}}^{O_i} \wedge \pi_{\text{patient}}^{O_i})) \wedge \Box (\neg (\bigcirc \pi_{\text{swarm}}^{O_i} \wedge \bigcirc \pi_{\text{patient}}^{O_i}))$
2. **Topology constraints.** The robot transition should follow the ER workshop as shown in Fig. 5 right.

Safety Assumptions. The environment behaviors are assumed to obey the following constraints:

1. **Static patients.** We assume that the patient status for each region r does not change: $\Box \pi_{\text{patient}}^r \leftrightarrow \bigcirc \pi_{\text{patient}}^r$.

Violations. We consider the following violations of environment safety assumptions.

- **ER-skill.** During the execution of $skill_3$ when the robot swarm moves from O_0 and O_9 to C , some robots accidentally enter ER due to low-level controller failure.
- **ER-patient.** During the execution of $skill_0$ when the robot swarm is still in C , a patient unexpectedly appears in ER .
- **ER-operate.** During the execution of $skill_0$ when the robot swarm is still in C , a patient unexpectedly appears in O_0 .

C CASE STUDY

We provide a case study that walks through INPROVF for the *factory-cup* violation.

C.1 ABSTRACTION INFORMALIZATION

This step takes as input (i) the set of input propositions $\mathcal{X} = \{\pi_o^r \mid o \in \mathcal{O}, r \in \mathcal{R}\} \cup \{\pi_{empty}\}$, where $\mathcal{O}$ and $\mathcal{R}$ are the sets of objects and regions as described in Appendix B, and (ii) the grounding function $G(\pi_o^r) := \{x \mid \text{pose}(x, o) \in \text{region}(r)\}$ and $G(\pi_{empty}) = \{x \mid \text{status_of_cup}(x) = \text{empty}\}$. We prompt LLMs using the prompt described in Appendix A.1. We use `p_object_region` to represent the input π_{object}^{region} and prime (') for the next operator ($\odot$), following the syntax of GR(1) formulas in Slugs (Ehlers & Raman, 2016). We obtain the following abstraction description.

LLM response: Each input proposition $\backslash(p_{\{o_r\}} \backslash)$ represents the condition that object $\backslash(o \backslash)$ is located within region $\backslash(r \backslash)$. The grounding function $\backslash(G \backslash)$ maps each proposition $\backslash(p_{\{o_r\}} \backslash)$ to the set of physical states where the pose of object $\backslash(o \backslash)$ is in region $\backslash(r \backslash)$. The proposition "empty" signifies the status of the cup, indicating whether it is empty.

C.2 STRATEGY INFORMALIZATION

Next, we take as inputs the informalized abstraction and a strategy (high-level controller) synthesized from the GR(1) specification. We leverage LLMs to generate a NL description of the behavior of the strategy. We use the one-shot prompt as described in Appendix A.2, where the strategy informalization example is shown in Appendix A.2.1. We obtain the following informalized abstraction:

LLM response: The synthesized strategy behaves as follows:
Initially, the robot's base is at position x4, the cup is at location k2, the block is at location k1, the cone is at x3, the stone is at k0, and the cup is empty.
When the cup becomes ****not empty****, the robot executes ****skill4**** to pick up the block from k1 and then ****skill7**** to place the block at k3. Next, the robot uses ****skill2**** to pick up the cup from k2. It then executes ****skill1**** to move from x4 via x2 to x0 while holding the cup. At x0, the robot uses ****skill6**** to place the cup on the table.
When the cup becomes ****empty**** again, the robot executes ****skill3**** to pick up the cup from the table. It then uses ****skill0**** to move from x0 via x2 to x4 while holding the cup. At x4, the robot executes ****skill5**** to place the cup back at k2.

C.3 REPAIR PROMPT

We incorporate the informalized abstraction and strategy description, along with the task specification and the assumption violation, to form a prompt for LLMs to repair the violation. Appendix A.3 shows the prompt. We obtain the following response:

LLM response:

```

“JSON
{
  "new_skill_0": [
    [
      ["p_base_x2", "p_cup_ee"],
      ["p_base_x3", "p_cup_ee"]
    ],
    [
      ["p_base_x3", "p_cup_ee"],
      ["p_base_x4", "p_cup_ee"]
    ]
  ]
}
“
```

C.4 VERIFICATION

While the repair candidate is syntactically correct, the updated specification is still unrealizable, meaning the new skills still cannot overcome the violation.

C.5 SAFETY ANALYSIS

After verification, we compute a counterstrategy and perform our safety analysis. Note that the new skill brings the robot *base* to x_3 , the same region as the obstacle *cone*, thus violating the collision avoidance safety constraint. The analysis obtains the following feedback:

Safety analysis feedback: new_skill_0 violates the hard constraints $!(p_base_x3' \ \& \ p_cone_x3')$

C.6 REPAIR PROMPT - SECOND ITERATION

After providing the safety feedback for iterative prompting, the repair prompt successfully generates the correct repair. The generated new skills are as follows:

LLM response:

```

“JSON
{
  "new_skill_0": [
    [
      ["p_base_x2", "p_cup_ee"],
      ["p_base_x1", "p_cup_ee"]
    ],
    [
      ["p_base_x1", "p_cup_ee"],
      ["p_base_x4", "p_cup_ee"]
    ]
  ],
  "new_skill_1": [
    [
      ["p_base_x4", "p_cup_ee"],
      ["p_base_x1", "p_cup_ee"]
    ],
    [
      ["p_base_x1", "p_cup_ee"],
      ["p_base_x0", "p_cup_ee"]
    ]
  ]
}
“““

```

C.7 REPAIR PROMPT (A DIFFERENT RUN)

We now provide the repair result from a different run that triggers liveness analysis:

LLM response:

```
“ “JSON
{
  "new_skill_0": [
    [
      ["p_base_x2", "p_cup_ee", "p_block_k3"],
      ["p_base_x4", "p_cup_ee", "p_block_k3"]
    ]
  ]
} “ “
```

C.8 LIVENESS ANALYSIS

While the new skill brings the robot *base* from x_2 to x_4 , we note that the position of *block* in the new skill is in k_3 , but the current position of *block* is in k_0 . Therefore, while this new skill is syntactically correct and does not violate any safety constraint, it is not helpful for the robot to satisfy the liveness goals under the violation. The liveness analysis provides the following feedback:

Feedback: The new skills cannot satisfy the liveness goals (empty $\rightarrow$ p_cup.t2)