

# “Don’t Do That!”: Guiding Embodied Systems through Large Language Model-based Constraint Generation

Aladin Djuhera<sup>1,\*</sup>, Amin Seffo<sup>1,†</sup>, Masataro Asai<sup>2</sup>, Holger Boche<sup>1</sup>

<sup>1</sup>Technical University of Munich, Germany <sup>2</sup>MIT-IBM Watson AI Lab, USA

Emails: {aladin.djuhera, amin.seffo, boche}@tum.de, masataro.asai@ibm.com

\*Lead Author †Core Contributor

**Abstract:** Recent advancements in large language models (LLMs) have spurred interest in robotic navigation that incorporates complex spatial, mathematical, and conditional constraints from natural language into the planning problem. Such constraints can be informal yet highly complex, making it challenging to translate into a formal description that can be passed on to a planning algorithm. In this paper, we propose STPR, a constraint generation framework that uses LLMs to translate constraints (expressed as instructions on “what not to do”) into executable Python functions. STPR leverages the LLM’s strong coding capabilities to shift the problem description from language into structured and transparent code, thus circumventing complex reasoning and avoiding potential hallucinations. We show that these LLM-generated functions accurately describe even complex mathematical constraints, and apply them to point cloud representations with traditional search algorithms. Experiments in a simulated Gazebo environment show that STPR ensures full compliance across several constraints and scenarios, while having short runtimes. We also verify that STPR can be used with smaller, code-specific LLMs, making it applicable to a wide range of compact models at low inference cost.

**Keywords:** Neuro-Symbolic Navigation, Heuristic Search, Path Constraints

## 1 Introduction

Real-world navigation involves not only reaching a goal but also adhering to constraints specified by human operators, which may be non-standardized, vague, implicit, or informal, capturing *semantic information* that is difficult to extract from sensor data only. For example, let us consider a cleaning robot without temperature sensors in Fig. 1. The owner might instruct it to avoid getting close to a fireplace and may provide additional details about its heat dissipation. The challenge for the robot is to incorporate these contextual and potentially complex spatial constraints into its path planning accordingly.

While recent advances in large language models (LLMs) have enabled robots to interpret such natural language instructions [1], purely LLM-based planning, where the language model directly returns a solution plan, has several critical shortcomings: First, LLMs can *hallucinate*, generating seemingly plausible yet incorrect plans that do not align with the robot’s physical constraints or the true environment [2]. Second, non-reasoning models lack *interpretability*, making it difficult to embed contextual constraints such as site-specific hazards, social norms, or dynamically changing no-go

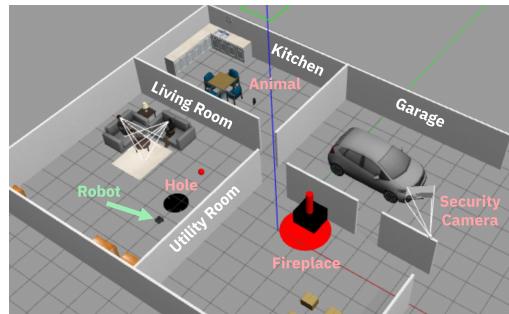


Figure 1: Gazebo environment with a garage, utility room, living room, and kitchen. A dangerous fireplace with a specific heat dissipation radius (red) must be avoided by the cleaning robot.

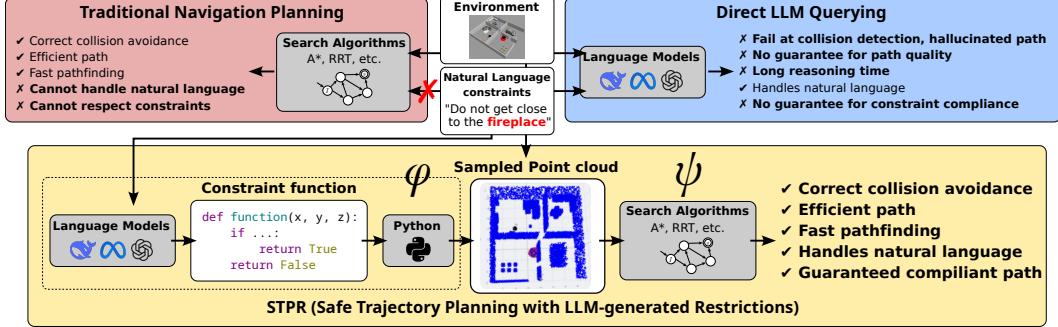


Figure 2: STPR Overview: LLM generates a Python function based on user constraints using a prompt template. The function is then integrated into a point sampling pipeline to generate a constrained representation of the environment. A classical algorithm uses this point cloud for path planning.

zones. Third, even if conditional constraints (e.g., “never enter the kitchen if there is an animal”) are directly encoded in an LLM prompt, it often results in *partial compliance*, i.e., LLMs may arbitrarily ignore or misinterpret constraints, as they have no explicit mechanisms to enforce them reliably [3]. While advanced reasoning models (e.g., OpenAI o-family [4]) may mitigate these weaknesses to some extent, there is no theoretical guarantee that such issues never occur, and yet these models incur significant computational cost and latency, both critical factors in real-time deployments.

To address these limitations, we propose *Safe Trajectory Planning with Restrictions* (STPR, pronounced “stopper”, Fig. 2), a practical and cost-effective neuro-symbolic navigation framework that combines natural language comprehension with traditional pathfinding algorithms. Instead of relying on LLMs for direct plan generation, STPR employs LLMs to translate high-level natural language constraints (“what *not* to do”) into executable Python boolean functions. These are then used to dynamically *prune the robot’s state space* by generating point clouds that act as imaginary obstacles via a form of rejection sampling [5]. A traditional pathfinding algorithm (e.g., A\* [6], RRT\* [7]) then operates within this refined state space, ensuring context-aware and constraint-compliant navigation.

This approach bypasses potentially error prone complex mathematical or spatial reasoning in text form, and instead leverages the LLM’s strong pre-training on code to generate precise, executable, and interpretable constraint functions. It then relies on the rigorous theoretical guarantees of search algorithms, such as optimality, soundness, and completeness. This *separation of concerns*, where LLMs handle constraint translation and classical algorithms handle decision-making, allows STPR to circumvent common LLM pitfalls while maintaining transparency in constraint enforcement.

We conduct a comprehensive evaluation of STPR in multiple *Gazebo* [8] experiments with six different LLMs across four challenging scenarios, ranging from hazard avoidance to conditional safety rules. Our empirical results show that STPR achieves *full compliance* and outperforms naive LLM-based planners in success rate and plan quality. Further, STPR does not require extensive hyperparameter tuning. Thus, default inference settings suffice to produce robust constraint functions. In addition, we verify that *smaller, code-specific LLMs* are equally compatible with STPR and generate reliable constraint functions without requiring larger, more sophisticated models.

## 2 Problem Formulation and Preliminaries

We define the robotic navigation problem as a 4-tuple  $\Pi = \langle X, A, T, s_0, G \rangle$ , where  $X \subseteq \mathbb{R}^3$  is a free space,  $A$  is a set of actions,  $T : X \times A \times X$  is a set of transitions,  $s_0 \in X$  is the initial state, and  $G \subseteq X$  is the set of goal states. A *plan* is a sequence of actions  $\pi = \langle a_1, a_2, \dots, a_N \rangle$  such when executed (i.e.,  $\forall i; (s_i, a_i, s_{i+1}) \in T$ ) leads from  $s_0$  to a goal state  $s_N \in G$ . The corresponding sequence  $\tau = \langle s_0, \dots, s_N \rangle$  is then called a *trajectory* or *path*. In this paper we use the latter terms interchangeably. The quality of a plan is measured by its *cost* and computed as the sum of individual transition costs, i.e.,  $\sum_{t=0}^{N-1} \text{COST}(s_t, a_t, s_{t+1})$ . For simplicity, we assume a Euclidean cost environment and consider

a static, deterministic, and discrete-time setting, though our model is naturally extendable to dynamic, stochastic, or continuous-time settings in the future.

We further model user-specified constraints as a mapping  $\phi : \mathcal{L} \rightarrow 2^X$ , where  $\mathcal{L}$  denotes the set of natural language instructions and  $2^X$  is a power set of  $X$ . For any given instruction  $l \in \mathcal{L}$ ,  $\phi$  produces a *forbidden region*  $C \subseteq X \subseteq \mathbb{R}^3$  that must be avoided by the agent. Herewith, we define the language-constrained robotic navigation problem as  $\Pi_l = \langle X', A, T', s_0, G' \rangle$ , where  $C = \phi(l)$ ,  $X' = X \setminus C$ ,  $T' = \{(s, a, s') \in T \mid s, s' \in X'\}$ , and  $G' = G \setminus C$ . We assume that the initial state  $s_0$  is not within the forbidden region  $C$ . A plan  $\pi$  for  $\Pi_l$  is called *invalid* if a state  $s_t$  within its trajectory is part of  $C$ . In the next section, we show how STPR leverages LLMs as a substitute for  $\phi$  by translating human-specified instructions into Python constraint functions.

### 3 Proposed STPR Method

Our method is formalized as a meta-algorithm  $\text{STPR}(\phi, \psi)$  that takes an LLM-based constraint generator  $\phi$  and a path finding algorithm  $\psi$ . The LLM instantiates  $\phi$  by generating a Python function  $f : X \rightarrow \{\text{True}, \text{False}\}$ , which acts as an indicator function such that  $C = \{\mathbf{x} \in X \mid f(\mathbf{x}) = \text{True}\}$ . Fig. 2 provides an overview of STPR where the constraint functions augment the point-cloud representation of the environment. We then employ  $\psi$  to generate a constraint-compliant path for  $\Pi_l$ .

#### 3.1 LLM-Based Constraint Code Prompting

At the heart of STPR is a carefully engineered *prompt template* that elicits from the LLM a self-contained Python function with a fixed signature (see Fig. 3). Concretely, our template consists of four parts: **(1) System Instruction:** A short directive (e.g., “*You are a robot*”) that establishes the assistant’s role and enforces zero-shot, single-turn behavior. **(2) Environment Block:** A textual representation of the environment. Under simulation (e.g., in Gazebo), such a representation can be obtained by simply parsing the environment definition. In real-world deployments, object-recognition networks such as PointNet [9] or PointRCNN [10] operate directly on raw sensor point clouds to recover a similar symbolic scene representation. **(3) Constraint Block:** A human-readable description of the constraint (e.g., “*Avoid the fireplace’s heat dissipation zone*”) together with all relevant numeric parameters, such as object coordinates and other parameters. **(4) Signature Scaffold:** A fixed Python function signature and doc-string stub, which ensures the LLM never drifts into pseudocode or comments-only answers, and guides it to produce a valid boolean test.

By forcing the LLM to emit *structured code* rather than free-form text, we leverage its extensive pre-training on code to produce syntactically correct, idiomatic Python functions. The rigid signature and minimal in-prompt examples thus leave little room for the model to stray, effectively avoiding hallucinations. Since every constraint can be provided directly by the user, domain experts can audit or tweak it without complex retraining, making this approach highly generalizable. STPR thus reduces the problem of “*what not to do*” in natural language to a transparent and executable Python predicate.

#### 3.2 Point-Cloud Sampling

STPR integrates LLM-generated constraints and other objects into a point cloud representation within a 3D environment, simulated using Gazebo (see setup in App. F). This is done by a form of rejection sampling: Let  $o_i \in O$  be  $i$ -th static object (walls, furniture, etc.) in the environment. To simplify the implementation, each  $o_i$  is over-approximated by a bounding box  $B_i = (\underline{x}_i, \bar{x}_i, \dots, \underline{z}_i, \bar{z}_i)$  obtained by parsing an offline environment file, i.e., a coordinate  $\mathbf{x} = (x, y, z)$  is regarded as colliding with  $o_i$  when  $\underline{x}_i < x < \bar{x}_i$ ,  $\underline{y}_i < y < \bar{y}_i$ , and  $\underline{z}_i < z < \bar{z}_i$ . Each  $B_i$  thus implicitly represents a boolean function  $f_i : X \rightarrow \{\text{True}, \text{False}\}$  that returns True if a state  $\mathbf{x} \in X$  lies inside the bounding box.

Next, given a set of natural language instructions  $\mathcal{L}$ , for each  $i$ -th instruction  $l_i \in \mathcal{L}$ , we produce a boolean function  $g_i$  of the same function signature using the prompt template in Sec. 3.1. For each function  $f_1, \dots, f_{|O|}$  (represented by  $B_1, \dots, B_{|O|}$ ) and  $g_1, \dots, g_{|\mathcal{L}|}$ , we generate a point cloud using rejection sampling: For each  $f_i$ , we use  $B_i$  to uniformly sample coordinates  $x \sim U(\underline{x}_i, \bar{x}_i), \dots, z \sim$

#### STPR Prompt Template for Scenario 4: Fireplace Heat Avoidance

You are a robot placed in a house and need to get to a room with an active fireplace while maintaining safety distance to avoid heat.

The fireplace has the following parameters: x=0.5 y=1.10 z=0 H\_0 = 1000 (heat intensity), alpha = 0.5 (dissipation factor), H\_safe = 50 (intensity threshold), d\_safe = 0.5 (distance threshold).

Generate a safe geometric constraint by using a hemispherical heat radiation model and by following this python syntax without giving any usage example. Rename <OBJ> accordingly:

```
def is_in_constraints_OBJ(x, y, z):
    """
    Check if the input satisfies constraints.
    Parameters:
    - x, y, z (float): Point coordinates.
    Returns:
    - bool: True if the point is forbidden, otherwise False
    """

```

↓ LLM

```
def is_in_constraints_fireplace(x, y, z):
    [...omitted...]
    # Calculate heat intensity using hemispherical heat radiation model
    heat_intensity = H_0 / (4 * 3.14159 * (distance ** 2)) * (1 - alpha)
    # Check if point is within fireplace boundaries
    within_fireplace = (abs(x - fireplace_x) <= fireplace_length / 2 and
                        abs(y - fireplace_y) <= fireplace_width / 2 and
                        abs(z - fireplace_z) <= fireplace_height / 2)
    # Check if heat intensity exceeds safe threshold or distance is less than safe distance
    return within_fireplace or heat_intensity > H_safe or distance < d_safe
```

Figure 3: **Top:** STPR prompt template for constraint code generation, including the System Instruction (“*You are a robot ...*”) in orange, the Environment Block in black, the Constraint Block representing the user-specific instruction (here: *Scenario 4: Fireplace Heat Avoidance*) in blue, and the Python function Signature Scaffold. **Bottom:** Corresponding constraint function generated by the LLM.

$U(\underline{z}_i, \bar{z}_i)$ . For each  $g_i$ , we could uniformly sample points from the entire environment and then reject them if  $g_i(x, y, z) = \text{False}$ . However, such a naive rejection sampling can be inefficient due to the high rejection rate. To address this issue, we additionally query the LLM for a function that returns an over-approximating bounding box for the natural language constraint  $l_i$ , sample some points from the bounding box, and then reject them using  $g_i$ . In future work, we might adapt a sophisticated method such as *adaptive rejection sampling* [11] instead of relying on such an ad-hoc LLM-based method.

For each such function ( $f_i$  or  $g_i$ ), we sample  $K$  points in total (a hyperparameter in our setup) and store them in a *kd-tree* structure [12] ( $T_{f_i}$  or  $T_{g_i}$ ) for fast nearest neighbor queries during the planning step (described in Sec. 3.3). Note that the point cloud density may further affect the accuracy of when a constraint is violated. We thus evaluate its impact in the experimental section later.

### 3.3 Constrained Path Planning

Given the initial state  $s_0$  and the set of goal states  $G$ , represented by a goal condition (e.g., within some Euclidean distance from a target coordinate), STPR( $\phi, \psi$ ) employs a generic path finding algorithm  $\psi$ . In this paper, we apply A\* [6] on an 8-connected grid in the environment, and RRT\* [7] on a continuous space representation of the environment. In general,  $\psi$  can be replaced with more sophisticated and efficient algorithms, such as Informed-RRT\* [13], SST/SST\* [14], or others.

In our experiments, for either search algorithm, each successor state  $x$  is tested against the sampled points in the *kd-trees*. Thus, given  $x$ , we query the nearest neighbor  $c_{\text{nearest}}$ , and, if it is within the robot’s maximum radius  $R$ , i.e.,  $\|x - c_{\text{nearest}}\|_2 < R$ , then  $x$  is pruned as a collision. This means that collisions are only checked against the search nodes and not along the edges because, in the more realistic scenario that involves a full physics simulation, the trajectory generated by  $\psi$  is used as navigation waypoints for low-level local planners (e.g., Dynamic Window Approach [15] in ROS), which are tasked to find an actual collision-free path between waypoints themselves.

The search stops when the expanded node is within  $R$  of the goal, its priority queue becomes empty (in A\*), or when it hits the maximum iteration of 5000 (in RRT\*). Due to the completeness of A\*, the second case proves that no valid path exists on the 8-connected grid. Further, since the search is guided by the Euclidean distance to the goal as the admissible heuristic, the generated path is also guaranteed to be *optimal* with regard to the path length on the 8-connected grid. In RRT\*, due to probabilistic completeness, the chance of false negatives (reporting path non-existence when it exists) decays exponentially to the number of iterations. Also, RRT\* is asymptotically optimal, i.e., the cost of the solution *almost surely* approaches the optimal cost as the number of iterations tends to infinity.

We reiterate that in STPR *constraint generation* is handled by the LLM, while *decision-making* remains within a well-understood planner. This duality avoids pitfalls such as hallucinations or partial compliance by the LLM, since  $\psi$  only traverses states that are deemed valid. As each constraint is an interpretable function, domain experts can modify restrictions without retraining the LLM or altering the path planner. Further, the use of point clouds as constraint representations is a deliberate design choice, enabling STPR to augment existing visual SLAM pipelines, for example, those that can be integrated within ROS’s SLAM Toolbox [16] or NVIDIA Isaac [17].

## 4 Empirical Evaluations

We evaluate STPR across challenging scenarios involving spatial, conditional, and physical constraints. We conduct all simulations in ROS within a Gazebo environment, including a kitchen, living room, garage, and utility room (see Fig. 1). Each constraint is formulated in a “*what not to do*”-style prompt written by the human operator in addition to other relevant parameters. The individual prompts are included in App. B. The robot, based on a Turtlebot 3 Waffle, has *no* sensors that could aid in constraint handling, as the purpose of our experiments is to show that STPR allows for complementing the lack of sensor input with user-specified natural language constraints. For constraint generation, we use *Llama-3.1-70B-Instruct* [18] as our primary model and explore additional models in the further analysis. For point cloud sampling, we generate  $K = 1000$  points for each object  $o_i$  and constraint  $l_i$ . For  $\psi$ , we use STPR with both A\* and RRT\* (see App. E for details), referred to as STPR-A\* and STPR-RRT\*, respectively, demonstrating its compatibility with two widely used algorithms.

Our main interest is *accuracy for path existence* to evaluate two core properties of search algorithms: *Soundness*, i.e., the solution returned by an algorithm is guaranteed to be valid, and *completeness*, i.e., an algorithm should find a solution whenever there exists a solution, or else report its nonexistence. Out of  $N = 10$  runs, we measure the *success ratio* by counting a *success* whenever a method returns a valid solution for a solvable task or reports path-nonexistence for an unsolvable task. Otherwise we count it as a *failure*. In addition, we measure the total runtime until algorithm termination (including path-nonexistence report), as well as the length of the returned path as an indicator of solution quality.

We consider four scenarios (see Fig. 4) that represent various types of constraints and complexities: **(S1) Evading a Security Camera.** The robot must avoid a security camera’s field of view (FOV) defined by its projection parameters (position, yaw, and clip planes). This scenario challenges the system with visibility constraints that require complex mathematical modeling in 3D, relevant for privacy and stealth applications. **(S2) Avoiding a Hole.** The robot needs to avoid falling into an invisible/hidden pit trap. This tests the handling of non-obvious vertical hazards that even sophisticated sensors might miss due to the concealed nature (e.g., the trap could be covered by a carpet or overlooked by LiDAR sensors in occupancy grids, as shown in Fig. 14 in App. F), but it could be easily avoided with a simple verbal warning. **(S3) Animal in the Kitchen.** When an animal is present, the robot should avoid the kitchen. We specifically label this encounter as a *dangerous* event to see whether the LLM will correctly self-infer this conditional constraint from within the context. **(S4) Fireplace Heat Avoidance.** The robot must maintain a safe distance from a fireplace given its heat intensity, a safety threshold, and dissipation range. In this scenario, we test the LLM’s ability to encode physics-based constraints beyond simple spatial boundaries.

To assess STPR’s effectiveness, we deliberately plan paths between points that would otherwise breach the constraints and compare vanilla A\*/RRT\*, STPR, and naive LLM planning outcomes. Here,

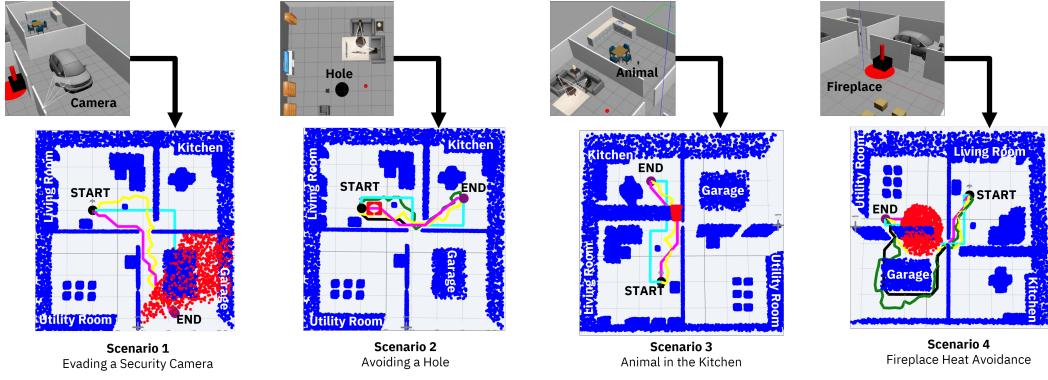


Figure 4: Planning results for STPR and baselines. **Blue:** Point cloud for static objects. **Red:** Point cloud from constraint functions. **Magenta:** Path generated by vanilla A\*. **Yellow:** Path generated by vanilla RRT\*. **Black:** Path generated by STPR-A\*. **Green:** Path generated by STPR-RRT\*. **Cyan:** Path generated by GPT-4o (using annotated image). Only STPR is compliant across all scenarios, refusing paths for (S1) and (S3), and avoiding hazards for (S2) and (S4). Additional results in App. C.

(S1) and (S3) constitute unsolvable tasks where a complete algorithm would report the nonexistence of valid paths while a hallucinating algorithm would return a path that violates the constraints.

#### 4.1 Qualitative Results

**Vanilla A\*/RRT\*.** As a sanity check, we first verify that vanilla A\* and RRT\* (without STPR) indeed produce problematic paths in our scenarios. The results are shown in magenta and yellow in Fig. 4. In fact, all paths generated by both vanilla variants violate  $\Pi_l$  and are thus deemed invalid: For (S1), the security camera is ignored and the agent enters the prohibited region. For (S2), the robot traverses over the hole as it lies on the optimal path to the goal. For (S3), the kitchen is entered even though the animal is present, and for (S4), the trajectory gets too close to the fireplace.

**STPR.** In contrast, for STPR (equipped with either A\* or RRT\*), our experiments show *full compliance* in all scenarios, where A\*/RRT\* either find a compliant path or correctly prove that no such path exists: For (S1), STPR derives a Python function that computes the distance and angular offset from the camera to each sampled point, approximating the FOV (see Fig. 7 in App. B). Specifically, the code checks if a point lies within the camera’s near and far clip boundaries, and if its yaw-based angle remains inside the camera’s horizontal FOV. For (S2), a simple function is generated to check if a point falls into the hole with an added safety radius (visualized by the rectangular shape). For (S3), the constraint function constructs a forbidden region around the entrance, ensuring that A\*/RRT\* cannot plan a path through the kitchen. Because this constraint is self-inferred based on the potentially dangerous encounter, STPR ensures context-aware planning, even though the environment is traversable (open door). This specific example shows effective reasoning where simply “*closing*” the door is enough. For (S4), the constraint function implements a hemispherical heat radiation model, marking high-temperature zones as unsafe (see Fig. 3). This creates a safety radius, requiring A\*/RRT\* to plan a detoured path which enters the utility room from another door. This example not only demonstrates an intelligent workaround but also that the LLM can produce a complex spatial constraint.

**Naive LLM Planning.** Lastly, we evaluate a naive vision LLM (VLM) approach that purely relies on the model (GPT-4o/o3-mini-high) to generate an operational path from an annotated top-down view of the environment (prompt and image provided in App. A). The resulting paths are shown in cyan in Fig. 4, alongside success ratios and other metrics in Table 1. In general, this approach suffers from hallucinations and partial compliance, where the model fails to correctly recognize not only the natural language instructions, but also the static objects or contexts within the environment. Focusing on GPT-4o, while there are some cases, such as (S2), where the model succeeds at generating a valid path that avoids the pit trap, paths for (S1) and (S3) are never valid, either violating the constraint,

Table 1: Success ratio, runtime, and path quality of vanilla A\*/RRT\*, STPR-A\*/-RRT\*, and naive VLMs (GPT 4o/o3-mini-high), averaged over 10 runs. Valid best results in **bold**, invalid results marked by  $\dagger$ , e.g., VLMs often incorrectly report shorter paths than STPR but violate the constraints.

|    | Success Ratio (%) |      |             |               |            |                      | Total Runtime (s) |               |                 |                      |              |               | Path Quality (Length in meters) |                      |  |  |  |  |
|----|-------------------|------|-------------|---------------|------------|----------------------|-------------------|---------------|-----------------|----------------------|--------------|---------------|---------------------------------|----------------------|--|--|--|--|
|    | A*                | RRT* | STPR-<br>A* | STPR-<br>RRT* | GPT<br>4o  | GPT o3-<br>mini-high | STPR-<br>A*       | STPR-<br>RRT* | GPT<br>4o       | GPT o3-<br>mini-high | STPR-<br>A*  | STPR-<br>RRT* | GPT<br>4o                       | GPT o3-<br>mini-high |  |  |  |  |
| S1 | 0                 | 0    | <b>100</b>  | <b>100</b>    | 0          | 0                    | <b>14.17</b>      | 14.63         | 34.10 $\dagger$ | 238.50 $\dagger$     | $\infty$     | $\infty$      | 20.0 $\dagger$                  | 20.4 $\dagger$       |  |  |  |  |
| S2 | 0                 | 0    | <b>100</b>  | <b>100</b>    | <b>100</b> | <b>100</b>           | 12.51             | <b>11.97</b>  | 25.20           | 36.80                | <b>9.95</b>  | 12.39         | 14.0                            | 14.2                 |  |  |  |  |
| S3 | 0                 | 0    | <b>100</b>  | <b>100</b>    | 0          | 10                   | <b>12.91</b>      | 13.58         | 24.30 $\dagger$ | 105.20 $\dagger$     | $\infty$     | $\infty$      | 12.0 $\dagger$                  | 12.3 $\dagger$       |  |  |  |  |
| S4 | 0                 | 0    | <b>100</b>  | <b>100</b>    | 0          | 10                   | 18.01             | <b>17.31</b>  | 32.50 $\dagger$ | 110.00 $\dagger$     | <b>20.49</b> | 25.80         | 14.7 $\dagger$                  | 15.1 $\dagger$       |  |  |  |  |

going through objects (e.g., walls), or both. This is particularly surprising as **(S3)** constitutes a simple conditional constraint that GPT-4o as a reasoning model should be able to adhere to. In addition, for **(S4)**, inconsistent outputs were observed, sometimes denying the existence of a valid path, yet generating (an invalid) one. The results for a high-performance reasoning model (GPT o3-mini-high) are similar, while at significantly longer runtimes (see Table 1). This shows that reasoning, even of advanced models, is brittle and too unreliable for practical use.

## 4.2 Quantitative Results

**Runtime and Quality.** Table 1 compares success ratios, total runtimes, and path lengths, where results marked by  $\dagger$  indicate that the metric belongs to an invalid path due to collisions or violating constraints, and where  $\infty$  path lengths indicate that the method successfully proves that no valid path exists. While vanilla A\* and RRT\* have shorter runtimes (see Table 3 in App. D), none of the paths adhere to the constraints, as indicated by the 0% success rate. Similarly, for either model, the VLM almost never succeeds in generating a valid path (except for **(S2)**), with success rates between 0% and 10%, invalidating end-to-end VLM-based planners. In contrast, STPR returns paths with strong theoretical guarantees: optimality when using A\*, and asymptotic optimality when using RRT\*. Further, paths returned by the VLM approach can be either significantly longer than STPR’s (e.g., 40% for **(S2)** where the constraint is respected) or shorter (e.g., 28% for **(S4)** where the constraint is neglected). However, the returned waypoints are often too sparse and sometimes meters apart. This shows another major weakness in LLM-based planning, namely the lack of theoretical guarantees.

**Model Variations.** We ablated STPR’s generative model across six different LLMs: Llama-3.1-405B, Llama-3.1-1B, Granite-34B-Code [19], GPT o1-pro, o3-mini-high, and 4o. Table 2 (left) shows the runtime for selected models that worked with STPR consistently. The smallest Llama-3.1-1B (not shown in the table) consistently fails at either logic or spatial reasoning, unable to generate constraints for any scenario. A mid-tier, code-specific model such as Granite-34B-Code produces sound functions, though requiring some minor prompt tuning to include missing Python imports. As expected, advanced reasoning (GPT o1-pro/o3/4o) and top-tier (Llama-3.1-405B) models achieve perfect success across all scenarios, however at the cost of higher prompting times. For instance, o1 pro takes *30 times longer* compared to our primary 70B Llama model, and even more compared to the cost-effective Granite alternative. This suggests that smaller, code-specific models can perform equally well. Interestingly, we observe binary behavior such that either the model works every time or never. Thus, we omit success ratios from these results. Note that the GPT entries in Table 1 are for naive VLM-based planners, while Table 2 compares runtimes for different LLM backends in STPR.

**LLM Decoding Parameters.** We varied common LLM decoding parameters to examine the effect on STPR’s performance. In our exploratory testing, STPR was robust against various top- $p$  values and temperature  $\tau$  of nucleus sampling. We observed prompting or sampling failures only when  $(p, \tau) = (1, 1)$ , which introduces excessive output randomness. Further, default decoding settings (e.g.,  $(p, \tau) = (0.7, 0.2)$ ) in most APIs sufficed to produce coherent constraint functions in our scenarios, demonstrating that STPR does not require costly inference hyperparameter tuning.

**Point cloud density.** The density of STPR’s sampled point clouds is a hyperparameter which depends on the number of sampled points  $K$ . A larger  $K$  increases the number of points in  $kd$ -trees and

Table 2: **(Left)** Prompting runtimes across LLMs (Llama-3.1-405/70B, Granite-34B-Code, GPT o1-pro, o3-mini-high, 4o). Llama-3.1-1B (omitted for space) failed on all instances. **(Right)** Accuracy vs. Runtime in STPR-A\* (Llama-3.1-70B) for varying  $K$ . † indicates a run with an invalid path.

|    | GPT<br>o1-pro | GPT o3-<br>mini-high | GPT<br>4o | Llama | Llama | Granite-<br>34B-Code | Success Ratio(%) |      |       | Total Runtime(s) |      |       |
|----|---------------|----------------------|-----------|-------|-------|----------------------|------------------|------|-------|------------------|------|-------|
|    |               |                      |           | 405B  | 70B   |                      | 100              | 1000 | 10000 | 100              | 1000 | 10000 |
| S1 | 5m 40s        | 1m 28s               | 48.7s     | 14.6s | 12.8s | 13.1s                | 0                | 100  | 100   | 12.9†            | 14.1 | 22.6  |
| S2 | 2m 32s        | 1m 8s                | 35.0s     | 27.7s | 11.9s | 16.6s                | 100              | 100  | 100   | 12.0             | 12.5 | 18.6  |
| S3 | 6m 23s        | 1m 35s               | 39.5s     | 23.7s | 11.8s | 16.2s                | 100              | 100  | 100   | 12.0             | 12.9 | 21.2  |
| S4 | 7m 45s        | 1m 55s               | 52.2s     | 21.7s | 16.1s | 11.5s                | 0                | 100  | 100   | 16.3†            | 18.0 | 25.6  |

thereby the average runtime  $O(\log K)$  for each nearest neighbor query. On the other hand, a smaller  $K$  reduces the runtime, though at the cost of potential failure to detect collisions. To investigate this, we evaluated three values of  $K = \{100, 1000, 10000\}$  and measured the performance in Table 2 (right) for STPR-A\*, demonstrating that STPR’s behavior is predictable and tuning  $K$  is easy.

## 5 Related Work

Literature on LLM-based robotic planning can be put in three categories: Low-level motion planning (ours), high-level task planning (classical planning [20]), and TAMP (Task and Motion Planning), which combines both. Similar to our work, some approaches attempt to improve safety, generate code, or use symbolic solvers, but the majority focus on high-level task planning or TAMP.

In high-level task planning and TAMP, Huang et al. [1] proposed using LLMs to generate natural language actions and translate them into commands in a VirtualHome environment. It validates actions only from failures in the environment, i.e., by performing potentially unsafe behaviors. SayCan [21] similarly suggests actions but evaluates their feasibility with a trained value function, which however does not guarantee safety due to its trained nature. This limitation also applies to approaches based on the “LLM-as-a-judge” paradigm. Other works improve speed and theoretical properties (e.g., completeness, soundness) by translating tasks into symbolic formal languages. For example, DELTA [3] generates PDDL domain descriptions and Python-based PDDL problem generators, and LLM+P [22] is a subset of DELTA. Further, Kwon et al. [23] use VLMs to translate images into PDDL problems, and AutoTAMP [24] generates Signal Temporal Logic (STL) specifications, while LLMFP [25] and Thought-of-Search [26] generate SMT formulas and Python-based successor/goal functions, respectively. CaStL [27] similarly encodes user-defined constraints into Python functions, but unlike our work, operates in first-order-logic representations, while ours targets continuous geometries.

In motion planning, PRoC3S [28] verifies LLM-generated motion plans using hand-coded state constraints, for example, through a Pybullet solver that checks joint positions for grasping. Further, chance-constrained path planning (CCPP) [29] focuses on trajectory safety under stochastic dynamics, though, is orthogonal to our approach, as it assumes predefined obstacles. However, integrating STPR’s point clouds with CCPP is a promising direction for future work.

## 6 Conclusion

We proposed STPR, a neuro-symbolic robot navigation framework that leverages LLMs to convert high-level natural language instructions into complex geometric constraints expressed in Python functions. STPR demonstrated that it can quickly and reliably comply to diverse spatial, conditional, and physical constraints, even when using compact, code-specific LLMs. We evaluated STPR using A\* and RRT\*, demonstrating its compatibility with different search algorithms that offer (probabilistic) completeness, soundness, and (asymptotic) optimality. We validated our approach through extensive simulations and ablation studies, covering four challenging scenarios, six different LLMs, and detailed runtime analyses. Our work demonstrates that conversational LLMs can be reliably integrated into robot navigation, while ensuring properties of classical search algorithms.

## 7 Limitations

Since our meta-algorithm relies on LLMs to generate Python constraint functions, there is a chance that it may produce an incorrect constraint, and our current implementation has no automated mechanism for verifying its correctness other than human evaluation. Nevertheless, in such a case, the generated behavior is no less safe than that of vanilla unconstrained or naive-LLM approaches, although it may be potentially less efficient. For example, if STPR were to generate an incorrect set of constraints, we have the following:

1. The resulting path still ensures collision avoidance with static objects, which naive LLM planners often violate (see results in Sec. 4).
2. It may breach some other constraints (e.g., maintaining a minimum distance from obstacles), which unconstrained or naive LLM systems can however also violate.
3. Paths may become longer, as the planner must circumvent the hallucinated constraints.
4. STPR’s embedded search algorithm may report ”no feasible path.” In that case, safety is preserved, since the robot will refuse to act.

As verifying an arbitrary Python program is undecidable and thus equivalent to solving the halting problem, it is widely believed to be impossible to guarantee such correctness in general. Employing an LLM as a judge may reduce the risk of generating incorrect constraints, but it cannot ensure a flawless output. This issue likewise arises in nearly all approaches surveyed in Sec. 5. However, our extensive simulations show promising results. By leveraging the LLM’s strong pre-training on code, STPR consistently generates coherent Python functions using pre-defined function signatures and scaffolds, which help reduce hallucinations and ensure robustness as well as consistency.

Furthermore, while STPR demonstrates planner-agnostic constraint reasoning in our simulated scenarios, our experimental implementation relies on a few simplifying assumptions that are out of scope in this study but not fundamental limitations in practice:

For example, we model obstacles as static, rigid bounding boxes with simplified geometry. In real-world deployments, objects may be deformable or irregular. In such cases, STPR’s rejection sampling strategy can be easily extended to richer mesh and point-cloud representations, or replaced by more advanced rejection sampling strategies (see Sec. 3.2). Another simplification is that our evaluation assumes a deterministic environment without dynamic obstacles or sensor noise. However, integrating probabilistic belief updates and partial observability can be helpful in practice, and is easily achieved, for example, by coupling STPR with existing perception pipelines. Lastly, our current implementation limits spatial reasoning to translations in Euclidean space and does not yet cover rotations or full SE(2)/SE(3) domains. However, extending the prompt template and planner interface is easy and makes it straightforward to introduce rotational constraints in future experimentation.

In this work, we covered extensive simulations in Gazebo and ROS, including detailed ablation results as part of a larger feasibility study, demonstrating that conversationable LLMs can be integrated reliably into robotic navigation with classical search algorithms. We emphasize that all of our simplifying assumptions can be readily addressed in future work to underscore STPR’s practical promise and clear path toward deployment in complex, real-world robotics applications.

### Acknowledgments

Aladin Djuhera, Amin Seffo, and Holger Boche acknowledge support from the German Federal Ministry of Research, Technology and Aerospace (BMFTR) under the program ”Souverän. Digital. Vernetzt.” as part of the research hubs 6G-life (Grant 16KISK002), QD-CamNetz (Grant 16KISQ077), QuaPhySI (Grant 16KIS1598K), and QUIET (Grant 16KISQ093).

## References

- [1] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 9118–9147. PMLR, 2022.
- [2] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.
- [3] Y. Liu, L. Palmieri, S. Koch, I. Georgievski, and M. Aiello. DELTA: Decomposed Efficient Long-Term Robot Task Planning using Large Language Models. *arXiv preprint arXiv:2404.03275*, 2024.
- [4] T. Zhong, Z. Liu, Y. Pan, Y. Zhang, Y. Zhou, S. Liang, Z. Wu, Y. Lyu, P. Shu, X. Yu, et al. Evaluation of OpenAI o1: Opportunities and Challenges of AGI. *arXiv preprint arXiv:2409.18486*, 2024.
- [5] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings Algorithm. *The american statistician*, 49(4):327–335, 1995.
- [6] P. E. Hart, N. J. Nilsson, and B. Raphael. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *Systems Science and Cybernetics, IEEE Transactions on*, 4(2):100–107, 1968.
- [7] S. Karaman and E. Frazzoli. Sampling-Based Algorithms for Optimal Motion Planning. *Int. J. Robot. Res.(IJRR)*, 30(7):846–894, 2011.
- [8] N. Koenig and A. Howard. Design and Use Paradigms for Gazebo, an Open-Source Multi-Robot Simulator. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 3, pages 2149–2154 vol.3, 2004. doi:10.1109/IROS.2004.1389727.
- [9] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017.
- [10] S. Shi, X. Wang, and H. Li. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–779, 2019.
- [11] W. R. Gilks and P. Wild. Adaptive Rejection Sampling for Gibbs Sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):337–348, 1992.
- [12] J. L. Bentley. Multidimensional Binary Search Trees Used for Associative Searching. *Commun. ACM*, 18(9):509–517, Sept. 1975. ISSN 0001-0782. doi:10.1145/361002.361007. URL <https://doi.org/10.1145/361002.361007>.
- [13] J. D. Gammell, S. S. Srinivasa, and T. D. Barfoot. Informed RRT\*: Optimal Sampling-Based Path Planning Focused via Direct Sampling of an Admissible Ellipsoidal Heuristic. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2997–3004. IEEE, 2014.
- [14] Y. Li, Z. Littlefield, and K. E. Bekris. Asymptotically Optimal Sampling-Based Kinodynamic Planning. *Int. J. Robot. Res.(IJRR)*, 35(5):528–564, 2016.
- [15] D. Fox, W. Burgard, and S. Thrun. The Dynamic Window Approach to Collision Avoidance. *IEEE Robotics & Automation Magazine*, 4(1):23–33, 1997.
- [16] S. Macenski and I. Jambrecic. SLAM Toolbox: SLAM for the Dynamic World. *Journal of Open Source Software*, 6:2783, 05 2021. doi:10.21105/joss.02783.
- [17] NVIDIA. *Isaac ROS Visual SLAM*, 2025. URL [https://nvidia-isaac-ros.github.io/repositories\\_and\\_packages/isaac\\_ros\\_visual\\_slam/index.html](https://nvidia-isaac-ros.github.io/repositories_and_packages/isaac_ros_visual_slam/index.html). Accessed: 2025-04-09.
- [18] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.
- [19] M. Mishra, M. Stallone, G. Zhang, Y. Shen, A. Prasad, A. M. Soria, M. Merler, P. Selvam, S. Surendran, S. Singh, et al. Granite Code Models: A Family of Open Foundation Models for Code Intelligence. *arXiv preprint arXiv:2405.04324*, 2024.

- [20] R. E. Fikes, P. E. Hart, and N. J. Nilsson. Learning and Executing Generalized Robot Plans. *Artificial Intelligence*, 3(1-3):251–288, 1972. doi:10.1016/0004-3702(72)90051-3.
- [21] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, et al. Do As I Can and Not As I Say: Grounding Language in Robotic Affordances. In *Proc. of Conference on Robot Learning*, pages 287–318. PMLR, 2023.
- [22] B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, J. Biswas, and P. Stone. LLM+P: Empowering Large Language Models with Optimal Planning Proficiency. *arXiv preprint arXiv:2304.11477*, 2023.
- [23] M. Kwon, Y. Kim, and Y. J. Kim. Fast and Accurate Task Planning using Neuro-Symbolic Language Models and Multi-level Goal Decomposition. *arXiv preprint arXiv:2409.19250*, 2024.
- [24] Y. Chen, J. Arkin, C. Dawson, Y. Zhang, N. Roy, and C. Fan. AutoTAMP: Autoregressive Task and Motion Planning with LLMs as Translators and Checkers. In *Proc. of IEEE International Conference on Robotics and Automaton (ICRA)*, pages 6695–6702. IEEE, 2024.
- [25] Y. Hao, Y. Zhang, and C. Fan. Planning Anything with Rigor: General-Purpose Zero-Shot Planning with LLM-based Formalized Programming. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2025. URL <https://openreview.net/forum?id=OK10aL6XuK>.
- [26] M. Katz, H. Kokel, K. Srinivas, and S. Sohrabi Araghi. Thought of Search: Planning with Language Models Through the Lens of Efficiency. *Proc. of the Advances in Neural Information Processing Systems (Neurips)*, 37:138491–138568, 2024.
- [27] W. Guo, Z. K. Kingston, and L. E. Kavraki. CaStL: Constraints as Specifications through LLM Translation for Long-Horizon Task and Motion Planning. In *Proc. of the Advances in Neural Information Processing Systems (Neurips)*, 2024.
- [28] A. Curtis, N. Kumar, J. Cao, T. Lozano-Pérez, and L. P. Kaelbling. Trust the PRoC3S: Solving Long-Horizon Robotics Problems with LLMs and Constraint Satisfaction. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.
- [29] L. Blackmore, M. Ono, and B. C. Williams. Chance-Constrained Optimal Path Planning with Obstacles. *IEEE Transactions on Robotics*, 27(6):1080–1094, 2011.
- [30] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [31] C. Peng, F. Xia, M. Naseriparsa, and F. Osborne. Knowledge Graphs: Opportunities and Challenges. *Artificial Intelligence Review*, 56(11):13071–13102, 2023.
- [32] S. M. LaValle and J. J. Kuffner. Randomized Kinodynamic Planning. *IJRR*, 20(5):378–400, 2001.
- [33] P. Cheng and S. M. LaValle. Resolution Complete Rapidly-Exploring Random Trees. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 1, pages 267–272. IEEE, 2002.
- [34] O. Salzman and D. Halperin. Asymptotically Near-Optimal RRT for Fast, High-Quality Motion Planning. *IEEE Transactions on Robotics*, 32(3):473–483, 2016.

## A Naive LLM-Based Planning

This section describes the details of naive end-to-end LLM-based path generation. Fig. 5 and Fig. 6 show the annotated image and the corresponding prompt used for scenario (S1) in our experiments with GPT-4o and GPT o3-mini-high, respectively. The setups for other scenarios follow a similar structure. In general, the naive LLM-based approach leads to hallucinations and partial compliance, where the model fails to correctly recognize constraints or to adhere to static obstacles, regardless of the GPT model's complexity and reasoning capabilities.

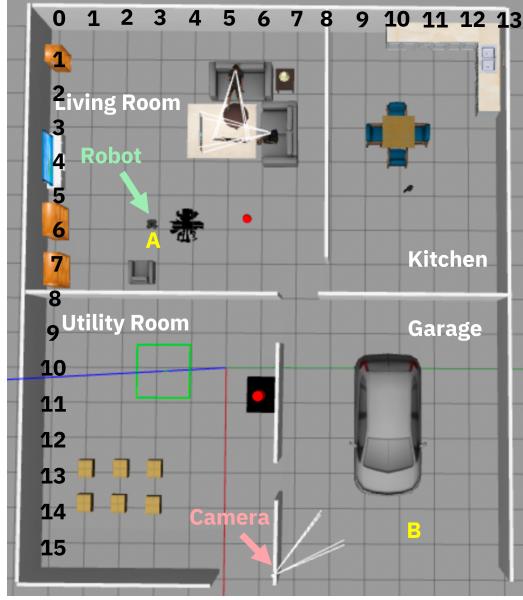


Figure 5: Annotated top-down view of the Gazebo environment for (S1): Evading a Security Camera. Image is attached to the GPT-4o prompt in Fig. 6 for end-to-end naive LLM-based path planning.

### Naive LLM-Based Path Planning: GPT Prompt

I have attached an image that shows the grid layout of a house from top-down, including living room, kitchen, utility room, and garage.

The x-axis is horizontal and the y-axis is vertical; I have added coordinates in BLACK. Each grid is a  $1 \times 1$  square.

You need to plan a path for a robot placed in the living room at point "A" ( $x=3, y=6$ ) on its way to point "B" ( $x=10, y=15$ ).

In the garage is a security camera labeled "Camera" ( $x=7, y=16$ ). You have to avoid the field of view of this camera.

The security camera has the following parameters:  
 roll = 0, pitch = 0, yaw = 2.299585, horizontal fov = 0.8, near clip = 0.1, far clip = 100

### TASK

Please plan a path from "A" to "B" which satisfies this constraint.  
 If no path is possible, return "NO PATH POSSIBLE" and an explanation. If the path is possible, **return a sequence of coordinates from the initial point "A" to "B"**.

Directly answer concisely without generating extra explanation or visualization.

Figure 6: Prompt for end-to-end naive LLM-based path planning with GPT 4o/o3-mini-high for (S1): Evading a Security Camera. Prompt uses the annotated image from Fig. 5.

## B Prompt Templates

For each of the four scenarios—*(S1) Evading a Security Camera, (S2) Avoiding a Hole, (S3) Animal in the Kitchen, (S4) Fireplace Heat Avoidance*—we follow the prompt template introduced in Sec. 3.1, and add additional contextual information about physical, spatial, or conditional constraints. The LLM then interprets this information to generate a constraint function that matches the required Python function signature while adhering to the specified user constraints.

Fig. 7-9 show the prompts and corresponding LLM outputs for scenarios (S1) through (S3). Scenario (S4) is presented in the main body in Fig. 3.

To avoid repeatedly embedding numeric parameters as part of the **Environment Block** of our prompt template, STPR supports a *context grounding* variant in which an environment file (e.g., a JSON containing static objects, such as walls and furniture, each with position, orientation, and shape) can be included as additional context in the system message. In this setting, the LLM retrieves the relevant parameters (e.g., for the fireplace) from the JSON file and generates the same constraint function as it would when the parameters are provided explicitly in the prompt. We provide an example JSON file for the fireplace scenario in Fig. 10. This approach is fully compatible with retrieval-augmented generation (RAG) [30] and other context-injection techniques, such as Knowledge Graphs [31]. This makes STPR highly generalizable and well-suited for real-world deployments, where practitioners often rely on full-stack LLM frameworks that incorporate RAG and other connectors to several databases holding such environment files.

### STPR Prompt Template for Scenario 1: Evading a Security Camera

You are a robot placed in a room where there is a security camera. You have to avoid the field of view of this camera.

The security camera has the following parameters:  $x = 5.45795$ ,  $y = 1.55736$ ,  $z = 2.03216$ , roll = 0, pitch = 0, yaw = 2.299585, horizontal fov = 0.8, near clip = 0.1, far clip = 100. The camera is in a room with the following boundary:  $x = -2$ .

Generate geometric constraints to avoid the field of view of the camera by following this python syntax without giving any usage example. Rename <OBJ> accordingly:

```
def is_in_constraints_OBJ(x, y, z):
    """
    Check if the input satisfies constraints.
    Parameters:
        - x, y, z (float): Point coordinates.
    Returns:
        - bool: True if the point is forbidden, otherwise False
    """

```

↓ LLM

```
def is_in_constraints_security_camera(x, y, z):
    """
    Check if the input satisfies constraints.
    Parameters:
        - x, y, z (float): Point coordinates.
    Returns:
        - bool: True if the point is forbidden, otherwise False
    """

    # Camera parameters
    cam_x, cam_y, cam_z = 5.45795, 1.55736, 2.03216
    cam_yaw = 2.299585
    horizontal_fov = 0.8
    near_clip, far_clip = 0.1, 100

    # Room boundary
    room_boundary = -2

    # Calculate the angle between the camera and the point
    dx, dy = x - cam_x, y - cam_y
    angle = math.atan2(dy, dx) - cam_yaw

    # Check if the point is within the camera's field of view
    if abs(angle) < horizontal_fov / 2:
        # Calculate the distance between the camera and the point
        distance = math.sqrt(dx**2 + dy**2 + (z - cam_z)**2)

        # Check if the point is within the camera's near and far clip
        if near_clip <= distance <= far_clip:
            # Check if the point is within the room boundary
            if x >= room_boundary:
                return True

    return False
```

Figure 7: **Top:** STPR prompt template for constraint code generation, including the System Instruction (“*You are a robot ...*”) in orange, the Environment Block in black, the Constraint Block representing the constraint-specific instruction (here for *Scenario 1: Evading a Security Camera*) in blue, and the Python function Signature Scaffold. The robot is explicitly told to avoid the field of view (FOV) of the security camera. To this end, the prompt provides precise camera parameters, including position, orientation (yaw), horizontal FOV, and near/far clipping planes, along with a room boundary, guiding the LLM to generate constraints that exclude points within the camera’s FOV. **Bottom:** Corresponding constraint function generated by the LLM. The code reconstructs the FOV as a forbidden region by computing the relative yaw angle and Euclidean distance to each point, checking whether the point lies within the horizontal angle and clipping range. Points violating these constraints are marked as invalid, forming a dynamic obstacle for path planning in our experiments.

### STPR Prompt Template for Scenario 2: Avoiding a Hole

You are a robot placed in a room where there is a hole which you have to avoid.

The hole has the following information:  $x = -4.0$ ,  $y = -1.0$ ,  $z = 0$ , radius = 0.5, height = 0.0005.

Generate geometric constraint to avoid falling down by following this python syntax without giving any usage example. Rename <OBJ> accordingly:

```
def is_in_constraints_OBJ(x, y, z):
    """
    Generalized function to check if the point (x, y, z) meets a set of constraints.

    Parameters:
        x, y, z (float): Coordinates of the point in 3D space.

    Returns:
        bool: True if the point meets all specified constraints; False otherwise.
        list: list of x, y, z min and max boundaries
    """

```

↓ LLM

```
def is_in_constraints_Hole(x, y, z):
    """
    Checks if a point (x, y, z) is within the boundaries of the hole.

    Parameters:
        x, y, z (float): Coordinates of the point in 3D space.

    Returns:
        bool: True if the point is within the hole; False otherwise.
        list: list of x,y,z min and max boundaries
    """

    # Define the center and radius of the hole
    hole_center_x = -4.0
    hole_center_y = -1.0
    hole_radius = 0.5

    # Define the height of the hole
    hole_height = 0.0005

    # Check if the point is within the horizontal boundaries of the hole
    is_within_horizontal_boundaries = (
        (x - hole_center_x) ** 2 + (y - hole_center_y) ** 2
        <= hole_radius ** 2
    )

    # Check if the point is within the vertical boundaries of the hole
    is_within_vertical_boundaries = abs(z) <= hole_height / 2

    # Return True if the point is within both horizontal and vertical boundaries, False otherwise
    return is_within_horizontal_boundaries and is_within_vertical_boundaries, [
        hole_center_x - hole_radius, hole_center_x + hole_radius,
        hole_center_y - hole_radius, hole_center_y + hole_radius,
        -hole_height / 2, hole_height / 2
    ]
```

Figure 8: **Top:** STPR prompt template for constraint code generation, including the System Instruction (“You are a robot ...”) in orange, the Environment Block in black, the Constraint Block representing the constraint-specific instruction (here for *Scenario 2: Avoiding a Hole*) in blue, and the Python function Signature Scaffold. The robot is instructed to avoid falling into a hole in the floor. The prompt describes the hole as a cylindrical region using its center coordinates, radius, and height, and then asks the LLM to generate a function that determines whether a given 3D point lies within this region. **Bottom:** Corresponding constraint function generated by the LLM. The code defines the hole as a vertical cylinder and checks whether a point lies within both its horizontal (radius) and vertical (height) bounds. If so, the point is marked as forbidden. In addition to the boolean test, the function provides an axis-aligned bounding box that encloses the cylindrical region to help generate constraints that align better with the environment.

### STPR Prompt Template for Scenario 3: Animal in the Kitchen

You are a robot placed in a house where someone heard that there is a raccoon in the kitchen. It is well-known that raccoons are wild animals.

The kitchen entrance has the following information:  $x = -2.531100$ ,  $y = 2.880299$ ,  $z = 2$ , roll = 0.0, pitch = 0, yaw =  $-1.57$ , dimensions = 0.8, 0.1, 2.

Generate geometric constraint to stay safe by following this python syntax without giving any usage example.  
Rename <OBJ> accordingly:

```
def is_in_constraints_OBJ(x, y, z):
    """
    Check if the input satisfies constraints.
    Parameters:
        - x, y, z (float): Point coordinates.
    Returns:
        - bool: True if the point is forbidden, otherwise False
    """
```

↓ LLM

```
def is_in_constraints_kitchen_entrance(x, y, z):
    """
    Check if the input satisfies constraints.
    Parameters:
        - x, y, z (float): Point coordinates.
    Returns:
        - bool: True if the point is forbidden, otherwise False
    """
    # Define the center of the kitchen entrance
    center_x = -2.531100
    center_y = 2.880299

    # Define the dimensions of the kitchen entrance
    width = 0.8
    depth = 0.1

    # Calculate the boundaries of the kitchen entrance
    min_x = center_x - width / 2
    max_x = center_x + width / 2
    min_y = center_y - depth / 2
    max_y = center_y + depth / 2

    # Check if the point is within the kitchen entrance boundaries
    if min_x <= x <= max_x and min_y <= y <= max_y:
        return True
    else:
        return False
```

**Figure 9: Top:** STPR prompt template for constraint code generation, including the System Instruction (“*You are a robot ...*”) in orange, the Environment Block in black, the Constraint Block representing the constraint-specific instruction (here for *Scenario 3: Animal in the Kitchen*) in blue, and the Python function Signature Scaffold. The robot is warned about the presence of a raccoon in the kitchen, but is not explicitly instructed to avoid the kitchen itself. With that, we want to test the LLM’s ability to infer implicit safety constraints. The prompt specifies the kitchen entrance’s position, orientation (yaw), and dimensions, and instructs the LLM to ensure safe behavior accordingly. **Bottom:** Corresponding constraint function generated by the LLM. In addition to the boolean test, the code defines a rectangular boundary around the kitchen entrance in the  $xy$ -plane and returns True for any point within this region, effectively marking it as forbidden to enforce a safety constraint.

### JSON File Context Grounding Example for Scenario 4: Fireplace Heat Avoidance

```
{
    "map_name": "utility_room",
    "description": "A utility room containing stacked cardboard boxes, enclosing walls, and a fireplace as a heat source.",
    "dimensions": {
        "x_min": -10,
        "x_max": 10,
        "y_min": -10,
        "y_max": 10,
        "z_min": 0,
        "z_max": 3
    },
    "obstacles": [
        {
            "name": "cardboard_box_0",
            "type": "box",
            "dimensions": [0.5, 0.4, 0.3],
            "position": [3.89151, -2.01001, 0.149],
            "orientation": [-6.8e-11, 1.6e-11, -6.4e-16]
        },
        {
            "name": "cardboard_box_1",
            "type": "box",
            "dimensions": [0.5, 0.4, 0.3],
            "position": [3.85544, -2.99685, 0.149],
            "orientation": [-6.8e-11, 1.6e-11, -6.4e-16]
        },
        {
            "name": "wall_1",
            "type": "wall",
            "dimensions": [2.23994, 0.125301, 2],
            "position": [4.62638, 1.59258, 0.940027],
            "orientation": [0, 0, -1.5821]
        },
        {
            "name": "wall_2",
            "type": "wall",
            "dimensions": [3.21559, 0.13633, 2],
            "position": [0.913087, 1.59301, 0.940027],
            "orientation": [0, 0, -1.5821]
        },
        // ... more cardboard boxes and walls omitted for brevity
    ],
    "heat_source": {
        "name": "fireplace",
        "type": "heat_source",
        "dimensions": [1.0, 0.876594, 0.65496],
        "position": [0.5, 1.10, 0],
        "heat_parameters": {
            "H_0": 5000,
            "alpha": 0.5,
            "H_safe": 50,
            "d_safe": 0.5
        }
    }
}
```

Figure 10: Context-grounding JSON description for (S4)’s utility room environment. Defines: (1) global map metadata (map name, description, dimensions); (2) an array of static obstacles, e.g., cardboard boxes and walls, each with type, dimensions, position, and orientation; and (3) a heat source entry for the fireplace, including its geometric dimensions and heat parameters. To avoid repeatedly embedding numeric parameters as part of the **Environment Block** of our prompt template in Sec. 3.1, such a JSON file can be passed to the prompt as part of a system message, e.g., via RAG.

## C Additional Visualizations

To provide additional perspectives on the simulation outcomes, Fig. 11 shows snapshots of the four scenarios (S1)-(S4) from different viewing angles. To reiterate:

**Vanilla A\*/RRT\*.** Vanilla A\*/RRT\* (magenta/yellow in Fig. 4) consistently produce invalid paths across all scenarios by ignoring or violating the constraints, e.g., entering the camera’s FOV in (S1), traversing the hole in (S2), crossing into the kitchen in (S3), and approaching too close to the fireplace in (S4).

**STPR.** In contrast, STPR (with A\* or RRT\*) achieves full compliance, pruning the camera’s FOV in (S1), avoiding the hole in (S2), blocking the kitchen entrance in (S3), and enforcing a hemispherical safety radius around the fireplace in (S4). The employed A\*/RRT\* search algorithms thus correctly yield either valid paths or prove infeasibility.

**Naive LLM Planning.** The naive vision LLMs (GPT 4o/o3-mini-high) produce hallucinated or partial paths, failing in (S1) and (S3), yielding inconsistent results in (S4), and only occasionally succeeding in (S2). GPT o3-mini-high exhibits similar brittleness, though at much higher runtimes.

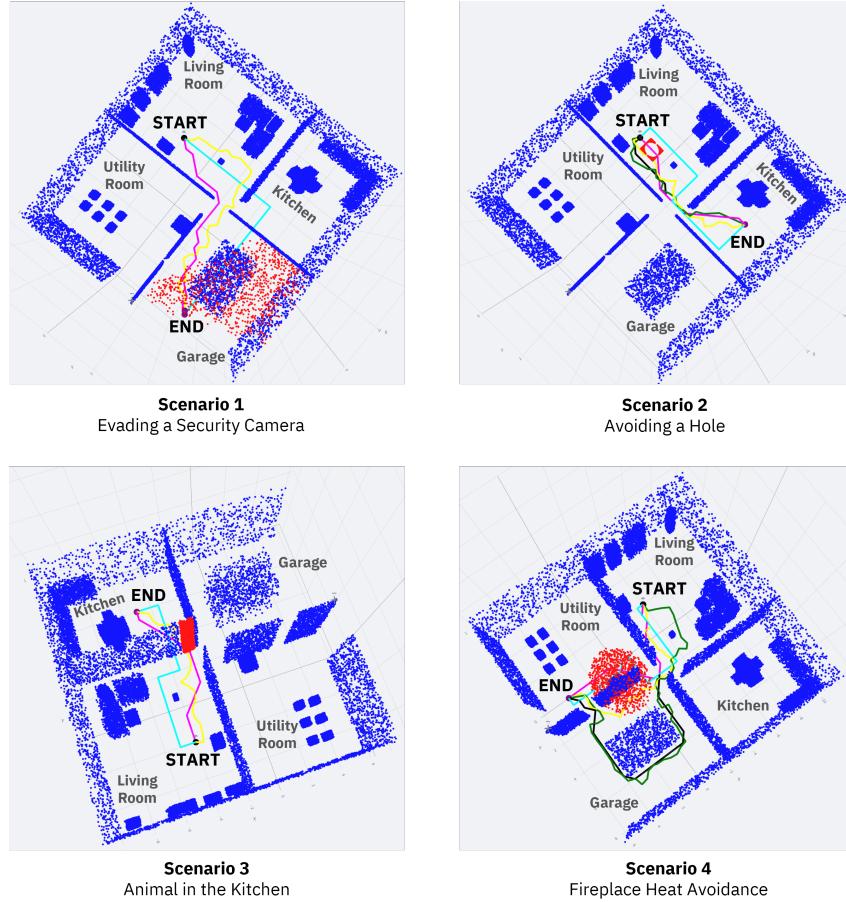


Figure 11: Planning results for STPR and baseline methods in all four scenarios (S1-S4), shown from different viewing angles. **Blue:** Point cloud for static objects. **Red:** Point cloud from constraint functions. **Magenta:** Path generated by vanilla A\*. **Yellow:** Path generated by vanilla RRT\*. **Black:** Path generated by STPR-A\*. **Green:** Path generated by STPR-RRT\*. **Cyan:** Path generated by GPT-4o (using annotated image from Fig. 5). Only STPR is compliant across all scenarios, refusing paths for (S1) and (S3), and avoiding hazards for (S2) and (S4).

## D Runtime Breakdown of STPR

In Table 3, we report a detailed runtime breakdown of the steps in STPR, including prompting, point cloud sampling, and pathfinding. In general, for both A\* and RRT\*, STPR maintains end-to-end latencies between 12 and 18 seconds, where prompting takes some longer time for (S4), in which the LLM engages in detailed physical modeling. For point cloud generation, (S2) and (S3) are faster due to their smaller forbidden regions, while (S1) and (S4) require more time for the 3D spheres. Depending on the goal state, planning varies, where (S1) and (S3) take the longest for the search algorithms to determine that a valid path cannot be found. Notably, for (S2) and (S4), where valid paths do exist, STPR’s planning is significantly faster with RRT\* due to its efficient implementation (see App. E.2). Further, although vanilla A\*/RRT\* implementations are generally faster, none of the generated paths are actually valid (see Sec. C).

Table 3: Runtime breakdown across scenarios in seconds. (**Left**) Vanilla A\*/RRT\* runtimes. All paths are invalid ( $\dagger$ ). (**Right**) STPR-A\*/-RRT\* runtimes, including prompting, sampling, and planning.

|          | Vanilla<br>A* | Vanilla<br>RRT* | Prompting      | Sampling | Planning<br>(STPR-A*) | Planning<br>(STPR-RRT*) | Total<br>(STPR-A*) | Total<br>(STPR-RRT*) |
|----------|---------------|-----------------|----------------|----------|-----------------------|-------------------------|--------------------|----------------------|
| Scenario | S1            | 0.60 $\dagger$  | 0.01 $\dagger$ | 12.84    | 0.13                  | 1.20                    | 1.66               | 14.17                |
|          | S2            | 0.63 $\dagger$  | 0.08 $\dagger$ | 11.88    | 0.05                  | 0.59                    | 0.04               | 12.51                |
|          | S3            | 0.63 $\dagger$  | 0.08 $\dagger$ | 11.77    | 0.03                  | 1.11                    | 1.78               | 12.91                |
|          | S4            | 0.66 $\dagger$  | 0.01 $\dagger$ | 16.13    | 0.98                  | 0.90                    | 0.02               | 18.01                |

## E Search Algorithms Implementation Details

In our experiments, we employ two popular search algorithms, A\* and RRT\*. In what follows, we provide some essential implementation details.

### E.1 A\*

A\* is a classical graph-based search algorithm widely used in robotics for path planning due to its balance between optimality and efficiency. It explores the search space by minimizing a cost-to-go function that combines actual and heuristic distances. A\* is easy to implement and primarily requires tuning two parameters: the grid resolution (i.e., step size) and the goal threshold that determines when the agent is considered to have reached its goal.

In our experiments, we implement A\* as a discrete grid-based planner operating in 2D space. The environment is sampled using an 8-connected neighborhood, and obstacle collision is checked via nearest-neighbor queries using a kd-tree built from obstacle points. The planner maintains open and closed sets and uses the Euclidean distance as the heuristic to evaluate nodes.

We found that a step size of 0.5 meters and a goal threshold of 0.3 meters were sufficient to yield good results across all scenarios. Obstacle clearance is ensured by requiring the nearest obstacle to be at least the threshold distance away before accepting a state as valid.

### E.2 RRT\*

In real-world motion planning applications, A\* is not a standard choice due to the more complex requirements from the task, including non-holonomic constraints and the fact that the true runtime bottleneck in motion planning lies in collision detection for path connectivity. To demonstrate the generalizability of STPR on more realistic scenarios, we implemented STPR with the RRT\* algorithm [7], a probabilistically complete, asymptotically optimal continuous space search algorithm.

While our implementation follows the standard algorithm proposed by Karaman and Frazzoli [7], we adopted *goal-biased* sampling of new nodes, a method originally proposed in the RRT paper [32, Section 4.2, page 387, left column] and later adopted as an optional enhancement in several newer methods [33, 34].

Given an initial state  $x_{\text{init}}$  and a goal state  $x_{\text{goal}}$ , RRT operates by repeating the following procedure:

1. Select a random state  $x_{\text{rand}}$  from the entire free space  $\mathcal{X}$ ,
2. Select the nearest neighbor  $x_{\text{nearest}}$  of  $x_{\text{rand}}$  from the set of nodes of the tree,

3. Generate a new state  $x_{\text{new}}$  by generating a collision-free path from  $x_{\text{nearest}}$  to  $x_{\text{rand}}$  using a local planner.
4. Add an edge  $(x_{\text{nearest}}, x_{\text{new}})$  to the tree.
5. Terminate if  $x_{\text{new}}$  satisfies a goal criteria (e.g., proximity to a goal state  $x_{\text{goal}}$ ).

In RRT\*, Step 4 instead adds an edge from one of the nodes already in the tree that are close to  $x_{\text{new}}$  (within a dynamically shrinking radius as a function of  $|V|$ ) and has the smallest cost from the root, if the cost is smaller than that of  $x_{\text{nearest}}$ .

Goal-biased sampling with hyperparameter  $r_{\text{GB}} \in [0, 1]$  alters Step 1: It uses  $x_{\text{goal}}$  instead of  $x_{\text{rand}}$  with probability  $r_{\text{GB}}$ . This biases the tree growth toward the target and increases the likelihood of early goal connection, thereby improving the runtime for finding the goal. Another adjustable parameter for RRT\* is the maximum number of iterations  $n_{\text{iters}}$ , which determines the maximum number of node expansion attempts allowed per planning run, thereby controlling the search budget.

In our experiments, we conduct multiple trials with  $n_{\text{iters}}$  ranging from 1000 to 5000 and  $r_{\text{GB}}$  set between 0% and 10%. Each configuration is evaluated over 10 random seeds. The results are shown in Table 4.

In general, goal-biased RRT\* finds a path more frequently and efficiently than the non-goal-biased RRT\*, especially under low iteration budgets. For example, with just 2000 iterations, the addition of 10% goal biasing increases the path completion rate from 50% to 90%, while reducing the average path length and runtime. At higher iterations, both biased and unbiased versions reach 100% success, but goal-biased RRT\* consistently generates shorter trajectories faster, demonstrating its advantage under constrained computational resources. Based on these results, we report RRT\* and STPR-RRT\* for  $n_{\text{iters}} = 5000$  iterations and  $r_{\text{GB}} = 0.1$  in all tables of the main body.

Table 4: Goal-Based RRT\* Performance Analysis (10 seeds per configuration) for Scenario 4: Fireplace Heat Avoidance. Configurations with 5000 iterations and 10% goal bias yields best results.

| <b>Max Iter</b> | <b><math>r_{\text{GB}}</math></b> | <b>Path Found (%)</b> | <b>Avg Length</b> | <b>Avg Time</b> |
|-----------------|-----------------------------------|-----------------------|-------------------|-----------------|
| 1000            | 0%                                | 20%                   | 22.50             | 0.04s           |
|                 | 10%                               | 40%                   | 24.38             | 0.06s           |
| 2000            | 0%                                | 50%                   | 25.41             | 0.11s           |
|                 | 10%                               | 90%                   | 24.11             | 0.12s           |
| 3000            | 0%                                | 60%                   | 26.39             | 0.18s           |
|                 | 10%                               | 90%                   | 24.11             | 0.12s           |
| 4000            | 0%                                | 90%                   | 27.51             | 0.47s           |
|                 | 10%                               | 100%                  | 24.40             | 0.15s           |
| 5000            | 0%                                | 100%                  | 27.56             | 0.54s           |
|                 | 10%                               | 100%                  | 24.40             | 0.15s           |

## F Simulation and Experimental setup

### F.1 Software in the Loop

To evaluate the STPR framework under realistic conditions, we deploy it in a Software in the Loop (SIL) simulation using ROS and Gazebo. Natural language constraints are processed by the LLM, which generates Python functions representing forbidden regions. These constraints guide a rejection sampling process that creates a point cloud of invalid areas, which is then used by a classical planner (such as A\* or RRT\*) to compute a valid trajectory (see Sec. 3).

The computed path is transmitted from the planning module via a ROS topic. A subscriber node on the robot listens to this topic and receives the trajectory data. The motion controller then interprets the incoming path and continuously publishes velocity commands to the robot's command topic, allowing the simulated *TurtleBot 3 Waffle* to follow the trajectory in the Gazebo environment. This setup enables real-time, closed-loop evaluation of constraint-aware motion planning in a fully virtual environment.

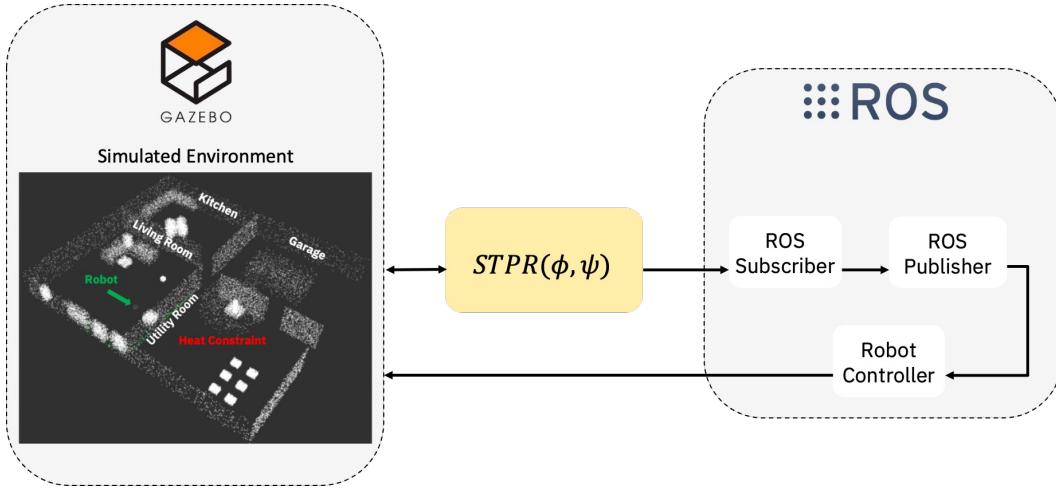


Figure 12: Software-in-the-Loop architecture for evaluating STPR in simulation. Natural language constraints are transformed into constraint functions, used to generate obstacle-aware point clouds for classical path planning. Planned trajectories are subscribed to by the robot, which then executes the motion via published velocity commands in the Gazebo simulation.

### F.2 Experimental Setup

The experimental pipeline is illustrated in Figure 13. The operator provides a prompt template to the LLM, including all necessary information and parameters.

The UI communicates with an NVIDIA Inference Microservice (NIM) container, which hosts the primary *LLaMA-3.2-70B* model on a local server. The LLM interprets the provided prompt and returns a Python function that evaluates whether a point in 3D space violates the specified constraint.

This generated Python function is subsequently used in a point-cloud sampling module, where points that violate the constraint are rejected via sampling and stored in a *kd*-tree-based database. This enhances the map representation with constraint-violating regions that reflect both physical obstacles (e.g., furniture) and instruction-based constraints (e.g., presence of an animal in the kitchen).

The augmented point-cloud map is then used for constraint-aware path planning. Classical search algorithms such as A\* or RRT\* are applied to find valid trajectories through the constrained space. These algorithms thus only operate on feasible regions, ensuring that the robot avoids violating any static or instruction-induced constraints during navigation.

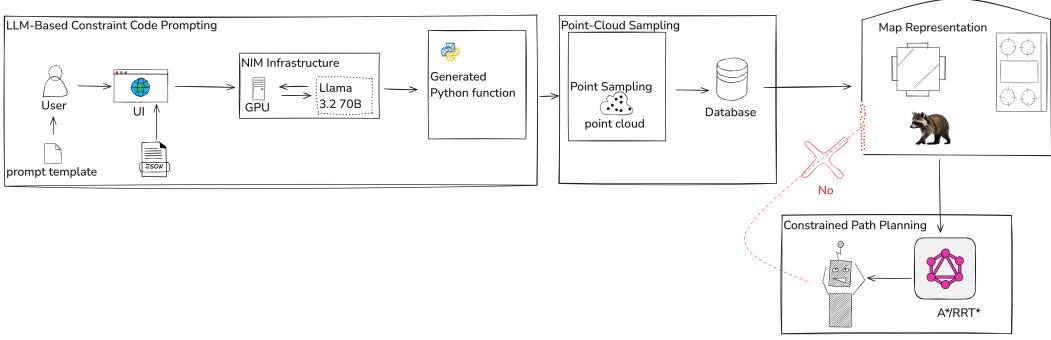


Figure 13: Flow diagram of the LLM-based constraint prompting, point cloud sampling, and constrained path planning process using NIM infrastructure and a Streamlit UI.

### E.3 Constraint Aware Navigation using STPR

While our framework is motivated by robotic agents that may lack the necessary sensory capabilities to identify potential conflicts from their surroundings, we demonstrate in the SLAM map in Fig. 14 that even in case of advanced sensory equipment, e.g., LiDAR, some hazards may still remain undetected.

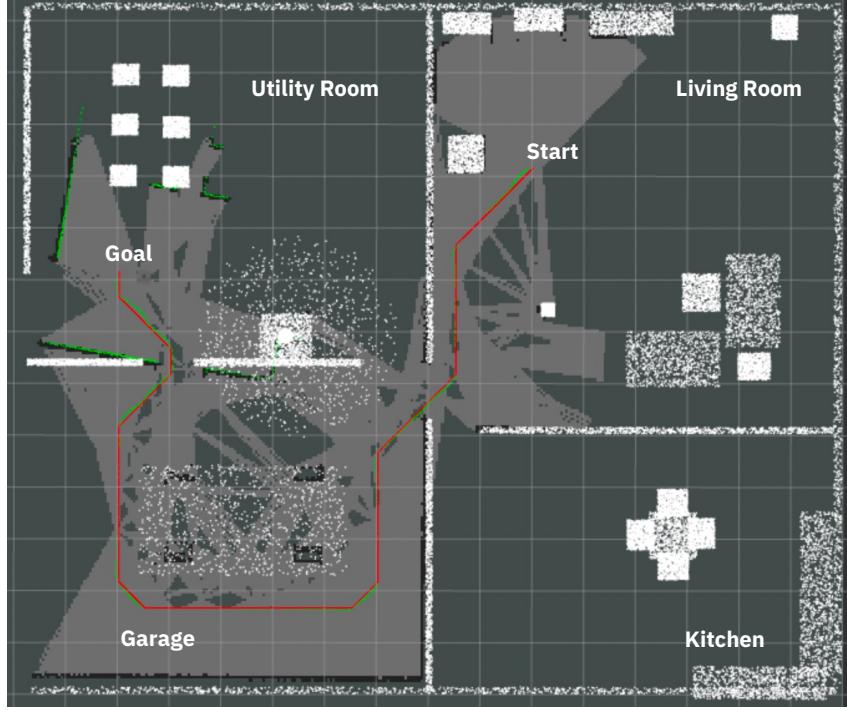


Figure 14: SLAM-generated occupancy grid map using LiDAR data, visualized in RVIZ. Static structures such as walls and furniture are detected by the LiDAR, but potential hazards like open holes or heat zones cannot be captured by the sensor. The white points represent the point cloud generated by STPR. The red path represents the constraint-aware plan from STPR-A\*, while the green line shows the executed trajectory that avoids the unsafe region.

For example, in scenario (S2): *Avoiding a Hole*, the robot must avoid a concealed pit trap. However, as shown in Fig. 14, a LiDAR sensor still cannot capture the hole due to lack of sufficient surface features. Similarly, in scenario (S4): *Fireplace Heat Avoidance*, the robot is instructed to stay clear of an active fireplace. While the SLAM map correctly captures walls and objects, it does not entail the necessary

information of potentially hazardous heat dissipation. Thus, without additional information or the human operator's input, planners might choose unsafe paths.

In contrast, our proposed STPR framework enables constraint-aware planning by translating natural language like "Avoid the hole in ..." or "Do not get close to the fireplace ..." into Python functions that define forbidden zones during point cloud sampling, which act as virtual obstacles. Thus, as shown in Fig. 14, STPR leads to safe and valid paths in both scenarios, where constraints are respected and hazards that could not be detected by the LiDAR equipment are successfully avoided. STPR therefore bridges the gap between high-level instructions in natural language and classical planning while ensuring algorithmic properties of classical planning algorithms.