

Verifiably Following Complex Robot Instructions with Foundation Models Supplementary

Benedict Quartey^{†*}, Eric Rosen*, Stefanie Tellex, George Konidaris
Department of Computer Science, Brown University

CONTENTS

A1	Appendix Summary	1
A2	Extended Related Works	1
A2-A	Foundation Models in Robotics	1
A2-B	Planning Models in Robotics	2
A3	Language Instruction Module	2
A3-A	Interactive Symbol Verification	3
A4	Spatial Grounding Module	4
A4-A	3D Spatial Comparators	4
A5	Task and Motion Planning Module	4
A6	Robot Skills	4
A7	Evaluation and Baseline Details	5
A7-A	NLMap-Saycan Implementation Prompt	6
A7-B	Code-as-Policies Implementation Prompt	7
A7-C	Instruction set	7
References		9

A1. APPENDIX SUMMARY

These sections presents additional details on our approach for leveraging foundation models and temporal logics to verifiably follow expressive natural language instructions with complex spatiotemporal constraints without prebuilt semantic maps. We encourage readers to visit our website robotlimp.github.io for project summary and demonstration videos.

A2. EXTENDED RELATED WORKS

A. *Foundation Models in Robotics*

Grounding language referents to entities and actions in the world [1], [2], [3], [4] is challenging in part due to the fact that complex perceptual and behavioral meaning can be constructed from the composition of a wide-range of open-vocabulary components[5], [6], [7], [8], [2]. To address this problem, foundation models have recently garnered interest as an approach for generating perceptual representations that are aligned with language [9], [10], [11], [12], [13], [14]. Because there are an ever-expanding number of ways foundation models are being leveraged for instruction following

in robotics (e.g: generating plans [14], code [15], etc.), we focus our review on the most related approaches in two relevant application areas: 1) generating natural language queryable scene representations and 2) generating robot plans for following natural language instruction [16], [17].

Visual scene understanding: The most similar approach for visual scene understanding to ours is NLMap [11], a scene representation that affords grounding open-vocabulary language queries to spatial locations via pre-trained visual language models. Given a sequence of calibrated RGB-D camera images and pre-trained visual-language models, NLMap supports language-queries by 1) segmenting out the most likely objects in the 2D RGB images based on the language queries, and 2) estimating the most likely 3D positions via back-projection of the 2D segmentation masks using the depth data and camera pose. While NLMap is suitable for handling complex descriptions of individual objects (e.g: “green plush toy”), it is fundamentally unable to handle instructions involving complex constraints between multiple objects since it has no way to account for object-object relationships (e.g: “the green plush toy that is between the toy box and door”). LIMP handles these more complicated language instructions by using a novel spatial grounding module to easily incorporate a wide-variety of complex spatial relationships between objects. In addition, our scene representation is compatible with both LLM planners as well as TAMP solvers, whereas NLMap is only compatible with LLM planners.

While NLMap is the most relevant approach to ours, there are other approaches for visual scene understanding and task planning with foundation models which are worth highlighting. VoxPoser [18] leverages the abilities of LLMs to identify affordances and write code for manipulation tasks, along with VLMs complementary abilities to identify open-vocabulary entities in the environment. SayPlan [19] integrates 3D scene graphs with LLM-based planners to bridge the gap between complex, heirarchical scene representations and scalable task planning with open-ended task specifications. Generalizable Feature Fields (GeFF) [20] use an implicit scene representation to support open-world manipulation and navigation via an extension of Neural Radiance Fields (NeRFs) and feature distillation in NeRFs. OK-Robot [21] adopts a system-first approach to solving structured mobile pick-and-place tasks with foundation models by offering an integrated solution to object detection, mapping, navigation and grasp generation. While these methods are related, none of them have all the features of LIMP: 1) Explicit support for both LLM-

*Equal Contribution

[†]Corresponding Author (Email: benedict_quartey@brown.edu)

based planning and off-the-shelf task and motion planning approaches, 2) Verifiable representations for following complex natural language instructions in mobile manipulation domains that involve object-object relationships, and 3) The ability to dynamically generate task-relevant state abstractions (semantic maps) for individual instructions.

Language instruction for robots: Our approach to handling complex natural language instructions involves translating the command into a temporal logic expression. This problem framing allows us to leverage state-of-the-art techniques from machine translation, such as instruction-tuned large language models. Most similar to our approach in this regard is [16], which uses a multi-stage LLM-based approach and finetuning to perform entity-extraction and replacement to translate natural language instructions into temporal logic expressions. However, [16] relies on a prebuilt semantic map that grounds expression symbols, limiting the scope of instructions it can operate since landmarks are predetermined. Instead, our approach interfaces with a novel scene representation that supports open-vocabulary language and generates the relevant landmarks based on the open-vocabulary instruction. Additionally, the symbols in our temporal logic translation correspond to parameterized task relevant robot skills as opposed to propositions of referent entities extracted from instructions.

B. Planning Models in Robotics

Semantic Maps: Semantic maps [22] are a class of scene representations that capture semantic (and typically geometric) information about the world, and can be used in conjunction with planners to generate certain types of complex robot behavior like collision-free navigation with spatial constraints [23], [24]. However, leveraging semantic maps for task planning with mobile manipulators has been challenging since the modeling information needed may highly depend on the robot’s particular skills and embodiment. [25] recently proposed Action-Oriented Semantic Maps (AOSMs), which are a class of semantic maps that include additional models of the regions of space where the robot can execute manipulation skills (represented as symbols). [25] demonstrated that AOSMs can be used as a state representation that supports TAMP solvers in mobile manipulation domains. Our scene representation is similar to an AOSM since it captures spatial information about semantic regions of interest, and is compatible with TAMP solvers, but largely differs in that AOSMs require learning via online interaction with the scene. Instead, our approach leverages foundation models and requires no online learning. Also, once an AOSM is generated for a scene, there is only a closed-set of goals that can be planned for, whereas our approach can handle open-vocabulary task specifications.

Task and Motion Planning: Task and Motion planning approaches are hierarchical planning methods that involve high-level task planning (with a discrete state space) [26] and low-level motion planning (with a continuous state space) [27]. The low-level motion planning problem involves generating paths to goal sets through continuous spaces (e.g: configu-

ration space, cartesian space) with constraints on infeasible regions. When the constraints and dynamics can change, it is referred to as multi-modal motion planning, which naturally induces a high-level planning problem that involves choosing which sequence of modes to plan through, and a low-level planning problem that involves moving through a particular mode. Finding high-level plan skeletons and satisfying low-level assignment values for parameters to achieve goals is a challenging bi-level planning problem[27]. LIMP contains sufficient information to produce a problem and domain description augmented with geometric information for bi-level TAMP solvers like [28], [29].

A3. LANGUAGE INSTRUCTION MODULE

We implement a two-stage prompting strategy in our language instruction module to translate natural language instructions into LTL specifications. The first stage translates a given instruction into a conventional LTL formula, where propositions refer to open-vocabulary objects. For any given instruction, we dynamically generate K in-context translation examples from a standard dataset [30] of natural language and LTL pairs, based on cosine similarity with the given instruction. Here is the exact text prompt used:

```

1 You are a LLM that understands operators involved with
  Linear Temporal Logic (LTL), such as F, G, U, &, |, ~
  , etc. Your goal is translate language input to LTL
  output.
2 Input:<generated_example_instruction>
3 Output:<generated_example_LTL>
4 ...
5 Input:<given_instruction>
6 Output:

```

Listing 1: Base prompt used to obtain a conventional LTL formula from a natural language query

The second stage takes the given instruction and the LTL response from the first stage as input to generate a new LTL formula with predicate functions that correspond to parameterized robot skills. Skill parameters are instruction referent objects expressed in our novel Composable Referent Descriptor (CRD) syntax. CRDs enable referent disambiguation by chaining comparators that encode descriptive spatial information. We define eight spatial comparators and provide their descriptions as part of the second stage prompt. We find that LLMs conditioned on this information and a few examples are able translate arbitrarily complex instructions with appropriate comparator choices. Here is the exact prompt used:

```

1 You are an LLM for robot planning that understands
  operators involved with Linear Temporal Logic (LTL),
  such as F, G, U, &, |, ~ , etc. You have a finite set
  of robot predicates and spatial predicates, given a
  language instruction and an LTL formula that
  represents the given instruction, your goal is to
  translate the ltl formula into one that uses
  appropriate composition of robot and spatial
  predicates in place of propositions with relevant
  details from original instruction as arguments.
2 Robot predicate set (near,pick,release).
3 Usage:
4 near[referent_1]:returns true if the desired spatial
  relationship is for robot to be near referent_1.
5 pick[referent_1]:can only execute picking skill on
  referent_1 and return True when near[referent_1].

```

```

6 release[referent_1,referent_2]:can only execute release
  skill on referent_1 and return True when near[
  referent_2].
7 Spatial predicate set (isbetween,isabove,isbelow,isleftof,
  isrightof,isnextto,isinfrontof,isbehind).
8 Usage:
9 referent_1::isbetween(referent_2,referent_3):returns true
  if referent_1 is between referent_2 and referent_3.
10 referent_1::isabove(referent_2):returns True if referent_1
  is above referent_2.
11 referent_1::isbelow(referent_2):returns True if referent_1
  is below referent_2.
12 referent_1::isleftof(referent_2):returns True if
  referent_1 is left of referent_2.
13 referent_1::isrightof(referent_2):returns True if
  referent_1 is right of referent_2.
14 referent_1::isnextto(referent_2):returns True if
  referent_1 is close to referent_2.
15 referent_1::isinfrontof(referent_2):returns True if
  referent_1 is in front of referent_2.
16 referent_1::isbehind(referent_2):returns True if
  referent_1 is behind referent_2.
17 Rules:
18 Strictly only use the finite set of robot and spatial
  predicates!
19 Strictly stick to the usage format!
20 Compose spatial predicates where necessary!
21 You are allowed to modify the structure of Input_ltl for
  the final Output if it does not match the intended
  Input_instruction!
22 You should strictly only stick to mentioned objects,
  however you are allowed to propose and include
  plausible objects if and only if not mentioned in
  instruction but required based on context of
  instruction!
23 Pay attention to instructions that require performing
  certain actions multiple times in generating and
  sequencing the predicates for the final Output
  formula!
24 Example:
25 Input_instruction: Go to the orange building but before
  that pass by the coffee shop, then go to the parking
  sign.
26 Input_ltl: F (coffee_shop & F (orange_building & F
  parking_sign ) )
27 Output: F ( near [coffee_shop] & F ( near [orange_building]
  & F near [parking_sign] ) )
28 Input_instruction: Go to the blue sofa then the laptop,
  after that bring me the brown bag between the
  television and the kettle on the left of the green
  seat, I am standing by the sink.
29 Input_ltl: F ( blue_sofa & F ( laptop & F ( brown_bag & F
  ( sink ) ) ) )
30 Output: F ( near [blue_sofa] & F ( near [laptop] & F ( near [
  brown_bag::isbetween (television,kettle::isleftof (
  green_seat))] & F (pick [brown_bag::isbetween (
  television,kettle::isleftof (green_seat))] & F ( near [
  sink] & F ( release [brown_bag,sink] ) ) ) ) )
31 Input_instruction: Hey need you to pass by chair between
  the sofa and bag, pick up the bag and go to the
  orange napkin on the right of the sofa.
32 Input_ltl: F ( chair & F ( bag & F ( orange_napkin ) ) )
33 Output: F ( near [chair::isbetween (sofa,bag)] & F ( near [
  bag] & F ( pick [bag] & F ( near [orange_napkin::
  isrightof (sofa)] ) ) ) )
34 Input_instruction: Go to the chair between the green
  laptop and the yellow box underneath the play toy
35 Input_ltl: F ( green_laptop & F ( yellow_box & F (
  play_toy & F ( chair ) ) ) )
36 Output: F ( near [chair::isbetween (green_laptop,yellow_box
  ::isbelow (play_toy))] )
37 Input_instruction: Check the table behind the fridge and
  bring two beers to the couch one after the other
38 Input_ltl: F ( check_table & F ( bring_beer1 ) & F (
  bring_beer2 ) & F ( couch ) )
39 Output: F ( near [table::isbehind (fridge)] & F ( pick [beer]
  & F ( near [couch] & F ( release [beer,couch] & F (
  near [table::isbehind (fridge)] & F ( pick [beer] & F (
  near [couch] & F ( release [beer,couch] ) ) ) ) ) ) ) )
40 Input_instruction: <given_instruction>
41 Input_ltl: <stage1_ltl_response>
42 Output:

```

Listing 2: Second stage prompt to output our LTL syntax with CRD parameterized robot skills

A. Interactive Symbol Verification

Verifying sampled LTL formulas is essential, as such we implement an interactive dialog system that presents users with extracted referent composable referent descriptors (CRDs) in sampled formulas as well as the implied task structure—encoded in the sequence of state-machine transition expressions that must hold to progressively solve the task. We translate the task structure into English statements via a simple deterministic strategy that replaces logical connectives and skill predicates from the formula with equivalent English phrases. Users can verify a formula as correct or provide corrective statements which are used to reprompt the LLM to obtain new formulas. Below is the exact prompt used for regenerating formulas.

```

1 There was a mistake with your output LTL formula: Error
  with <verification_type>! Consider the clarification
  feedback and regenerate the correct output for the
  Input_instruction. Make sure to adhere to all rules
  and instructions in your original prompt!
2 previous_output:<last_response>
3 error_clarification: <given_error_clarification>
4 correct_output:

```

Listing 3: Corrective reprompting prompt used to obtain new LTL formulas

To illustrate, the instruction “Bring the green plush toy to the whiteboard in front of it” yields the interactive Referent and Task Structure Verification dialog below:

```

1 *****
2 Instruction Following
3 *****
4 Input_instruction: "Bring the green plush toy to the
  whiteboard in front of it"
5 Sampled LTL formula: F(A & F(B & F(C & FD)))
6 A: near [green_plush_toy]
7 B: pick [green_plush_toy]
8 C: near [whiteboard::isinfrontof (green_plush_toy)]
9 D: release [green_plush_toy, whiteboard::isinfrontof (
  green_plush_toy)]
10 *****
11 Referent Verification
12 *****
13 I extracted this list of relevant objects based on your
  instruction:
14 * whiteboard::isinfrontof (green_plush_toy)
15 * green_plush_toy
16 Does this match your intention? (y/n)
17 *****
18 Task Structure Verification
19 *****
20 Based on my understanding here is the sequence of subgoal
  objectives needed to satisfy the task:
21 Subgoal_1:
22 Logical Expression: A&B
23 English translation: I should be near the [
  green_plush_toy] and not have picked up the [
  green_plush_toy]
24 Subgoal_2:
25 Logical Expression: B&C
26 English translation: I should have picked up the [
  green_plush_toy] and not be near the [whiteboard::
  isinfrontof (green_plush_toy)]
27 Subgoal_3:
28 Logical Expression: C&D
29 English translation: I should be near the [whiteboard
  ::isinfrontof (green_plush_toy)] and not have released
  the [green_plush_toy] at the [whiteboard::
  isinfrontof (green_plush_toy)]
30 Subgoal_4:
31 Logical Expression: D
32 English translation: I should have released the [
  green_plush_toy] at the [whiteboard::isinfrontof (
  green_plush_toy)]

```

Listing 4: Interactive referent and task structure verification dialog.

A4. SPATIAL GROUNDING MODULE

The spatial grounding module detects and localizes specific instances of objects referenced in a given instruction by first detecting, segmenting and back-projecting all referent occurrences and then filtering based on the descriptive spatial details captured by each referent’s composable referent descriptor (CRD). We use the Owl-Vit model [31] to detect bounding boxes of open-vocabulary referents and SAM [32] to generate masks from detected bounding boxes. To illustrate referent filtering via spatial information, consider an example scenario where the goal is to resolve the composable referent descriptor below:

whiteboard :: *isinfrontof*(green_plush_toy). (A.1)

Let $W = \{w_1, w_2, \dots, w_n\}$ and $G = \{g_1, g_2, \dots, g_m\}$ represent the set of representative 3D positions of detected whiteboards and green_plush_toys respectively. The cartesian product of these sets enumerates all possible pairs (w, g) for comparison.

$$W \times G = \{w, g \mid w \in W, g \in G\} \quad (\text{A.2})$$

The ‘*isinfrontof*(w, g)’ comparator is applied to each pair, yielding a subset S that contains only those ‘whiteboards’ that satisfy the ‘*isinfronto*’ condition with at least one ‘green_plush_toy’.

$$S = \{w \in W \mid \exists g \in G \text{ such that } \text{isinfrontof}(w, g) \text{ is true}\} \quad (\text{A.3})$$

A. 3D Spatial Comparators

Our 3D spatial comparators enable Relative Frame of Reference (FoR) spatial reasoning between referents, based on their backprojected 3D positions. Threshold values in the spatial comparators give users the ability to specify the sensitivity or resolution at which spatial relationships are resolved, we keep all threshold values fixed across all experiments. Below is a description of each spatial comparator.

```

1 1. isbetween(referent_1_pos, referent_2_pos,
   referent_3_pos, threshold): Returns true if
   referent_1 is within 'threshold' distance from the
   line segment connecting referent_2 to referent_3,
   ensuring it lies in the directional path between them
   without extending beyond.
2 2. isabove(referent_1_pos, referent_2_pos, threshold):
   Returns true if the z-coordinate of referent_1
   exceeds that of referent_2 by at least 'threshold'.
3 3. isbelow(referent_1_pos, referent_2_pos, threshold):
   Returns true if the z-coordinate of referent_1 is
   less than that of referent_2 by more than 'threshold'.
4 4. isleftof(referent_1_pos, referent_2_pos, threshold):
   Returns true if the y-coordinate of referent_1
   exceeds that of referent_2 by at least 'threshold',
   indicating referent_1 is to the left of referent_2.
5 5. isrightof(referent_1_pos, referent_2_pos, threshold):
   Returns true if the y-coordinate of referent_1 is
   less than that of referent_2 by more than 'threshold',
   indicating referent_1 is to the right of
   referent_2.

```

```

6 6. isnextto(referent_1_pos, referent_2_pos, threshold):
   Returns true if the Euclidean distance between
   referent_1 and referent_2 is less than 'threshold',
   indicating they are next to each other.
7 7. isinfrontof(referent_1_pos, referent_2_pos, threshold):
   Returns true if the x-coordinate of referent_1 is
   less than that of referent_2 by more than 'threshold',
   indicating referent_1 is in front of referent_2.
8 8. isbehind(referent_1_pos, referent_2_pos, threshold):
   Returns true if the x-coordinate of referent_1
   exceeds that of referent_2 by at least 'threshold',
   indicating referent_1 is behind referent_2.

```

Listing 5: Implementation description of 3D spatial comparators

A5. TASK AND MOTION PLANNING MODULE

We present pseudocode for our Progressive Motion Planner (Alg.1) and our algorithm for generating Task Progression Semantic Maps (Alg.2). Alg.2 generates a TPSM $\mathcal{M}_{\text{tpsm}}$ by integrating an environment map (\mathcal{M}) and a referent semantic map (\mathcal{M}_{rsm}) given a logical transition expression (\mathcal{T}), a desired automaton state (S'), and a nearness threshold (θ). The algorithm first initializes $\mathcal{M}_{\text{tpsm}}$ with a copy of \mathcal{M} and extracts relevant instruction predicates from \mathcal{T} . For each predicate (parameterized skill), the algorithm identifies satisfying referent positions in \mathcal{M}_{rsm} , generates a spherical grid of surrounding points within a radius θ , and assesses how these points affect the progression of the task automaton towards S' . These points demarcate regions of interest, and are assigned a value of 1 if they cause the automaton to transition to the desired state, -1 if they lead to a different automaton state or violate the automaton, and 0 if they do not affect the automaton. The points are then integrated into $\mathcal{M}_{\text{tpsm}}$, yielding a semantic map that identifies goal and constraint violating regions.

A6. ROBOT SKILLS

We define three predicate functions: **near**, **pick** and **release** for the navigation, picking and placing skills required for multi-object goal navigation and mobile pick-and-place. As highlighted in the main paper, we formalize navigation as continuous path planning problems and manipulation as object parameterized options. We discuss navigation at length in the paper, so here we focus on the pick and place manipulation skills.

Pick Skill: Once the robot has executed the near skill and is at the object to be manipulated, it takes a photo of the current environment to detect the object using the Owl-Vit model. The robot is guaranteed to be facing the object as the computed path plan uses the backprojected object 3D position to compute yaw angles for the robot. After detecting the object in the picture, we obtain a segmentation mask from detected boundary box using the Segment Anything model, and compute the center pixel of this mask. We feed this center pixel to the Boston dynamics grasping API to compute a motion plan to grasp the object.

Release Skill: We implement a simple routine for the release skill which takes two parameters: the object to be placed and the place receptacle. Once a navigation skill

Algorithm 1 Progressive Motion Planning Algorithm

```
1: procedure PMP( $X_{start}, \varphi, \mathcal{M}, \mathcal{M}_{rsm}, \theta$ )
  Input:
     $X_{start}$ : Start position in the environment.
     $\varphi$ : CRD syntax LTL formula specifying task objectives.
     $\mathcal{M}$ : Environment map.
     $\mathcal{M}_{rsm}$ : Referent semantic map.
     $\theta$ : Nearness threshold.
  Output:
     $\Pi$ : Generated task and motion plan.
2:  $\mathcal{A} \leftarrow \text{ConstructAutomaton}(\varphi)$ 
3:  $\text{path} \leftarrow \text{SelectAutomatonPath}(\mathcal{A})$   $\triangleright$  Task plan
4: while  $\Pi$ .status is active do
5:   while  $\mathcal{A}$ .state  $\neq$  path.acceptingState do
6:      $S, \mathcal{T}, S' \leftarrow \mathcal{A}.$ GetTransition(path.currentStep)
7:     objective  $\leftarrow$  NextObjectiveType( $\mathcal{T}$ )
8:     if objective = "skill" then
9:        $\Pi$ .UpdateWithSkill( $\mathcal{T}$ )
10:       $\mathcal{A}.$ UpdateAutomatonState( $S'$ )
11:     else if objective = "navigation" then
12:        $\mathcal{M}_{params} \leftarrow \mathcal{M}, \mathcal{M}_{rsm}, \mathcal{T}, \mathcal{A}, S', \theta$ 
13:        $\mathcal{M}_{tpsm} \leftarrow \text{GENERATETPSM}(\mathcal{M}_{params})$ 
14:        $\mathcal{O} \leftarrow \text{GenerateObstacleMap}(\mathcal{M}_{tpsm})$ 
15:        $\text{plan} \leftarrow \text{FMT}^*(X_{start}, \mathcal{O})$   $\triangleright$  Path plan
16:       if plan.exists then
17:          $\Pi$ .UpdateWithPlan(plan)
18:          $X_{start} \leftarrow$  plan.endPosition
19:          $\mathcal{A}.$ UpdateAutomatonState( $S'$ )
20:       else
21:          $\Pi, \mathcal{A}, \text{path} \leftarrow \text{Backtrack}(\Pi, \mathcal{A}, \text{path})$ 
22:       end if
23:     end if
24:   end while
25: end while
26: return  $\Pi$ 
27: end procedure
```

gets the robot to the place receptacle, the robot gently moves its arm up or down to release the grasped object, based on the place receptable 3D position. Future work will implement more complex semantic placement strategies to better leverage LIMP’s awareness and spatial grounding of instruction specific place receptacles. Kindly, visit our [website](#) to see demonstrations of these skills.

A7. EVALUATION AND BASELINE DETAILS

All computation including planning, loading and running pretrained visual language models was done on a single computer equipped with one NVIDIA GeForce RTX 3090 GPU. We leverage GPT-4-0613 as the underlying LLM for our instruction understanding module and all our baselines. In all experiments we set the LLM temperature to 0, however since deterministic greedy token decoding is not guaranteed

Algorithm 2 Task Progression Semantic Mapping Algorithm

```
1: procedure GENERATETPSM( $\mathcal{M}, \mathcal{M}_{rsm}, \mathcal{T}, \mathcal{A}, S', \theta$ )
  Input:
     $\mathcal{M}$ : Environment map.
     $\mathcal{M}_{rsm}$ : Referent semantic map.
     $\mathcal{T}$ : Automaton transition expression.
     $\mathcal{A}$ : Task Automaton.
     $S'$ : Desired State.
     $\theta$ : Nearness threshold.
  Output:
     $\mathcal{M}_{tpsm}$ : Task Progression Semantic Map.
2:  $\mathcal{M}_{tpsm} \leftarrow \text{Copy}(\mathcal{M})$ 
3:  $\mathcal{P} \leftarrow \text{ExtractRelevantPredicates}(\mathcal{T})$ 
4: for  $p$  in  $\mathcal{P}$  do
5:    $\mathcal{R} \leftarrow \text{QueryPositions}(\mathcal{M}_{rsm}, p)$ 
6:   for  $r$  in  $\mathcal{R}$  do
7:      $G \leftarrow \{g \mid g = r + \delta, \|\delta\| \leq \theta\}$   $\triangleright$  spherical grid of surrounding points
8:     for  $g$  in  $G$  do
9:        $\mathcal{Q} \leftarrow \text{TruePredicatesAt}(g, \mathcal{M}_{rsm}, \theta)$ 
10:       $S_{next} \leftarrow \text{ProgressAutomaton}(\mathcal{A}, \mathcal{Q})$ 
11:      if  $S_{next} = S'$  then
12:         $g.value \leftarrow 1$   $\triangleright$  Goal value
13:      else if  $\text{IsUndesired}(S_{next})$  then
14:         $g.value \leftarrow -1$   $\triangleright$  Avoidance value
15:      else
16:         $g.value \leftarrow 0$ 
17:      end if
18:    end for
19:     $\text{AddPoints}(\mathcal{M}_{tpsm}, G)$ 
20:  end for
21: end for
22: return  $\mathcal{M}_{tpsm}$ 
23: end procedure
```

with GPT4, we perform three (3) queries for each instruction and evaluate on the most recurring response (mode response).

We compare LIMP with baseline implementations of NLMap-Saycan [11] and Code-as-policies [15]. Both baselines use the same GPT-4 LLM, prompting structure, and in-context learning examples as our language understanding module. We integrate our composable referent descriptor syntax, spatial grounding module and low-level robot control into these baselines as APIs. This enables baselines to execute plans by querying relevant object positions, using our FMT* path planner to find paths to said positions and executing manipulation options.

We visualize some qualitative results of LIMP from our experiments in Figure A.1. We also highlight results in Figure A.2 that illustrates how our interactive symbol verification and reprompting strategy A3-A improves instruction satisfaction with minimal chat turns for different instruction sets.

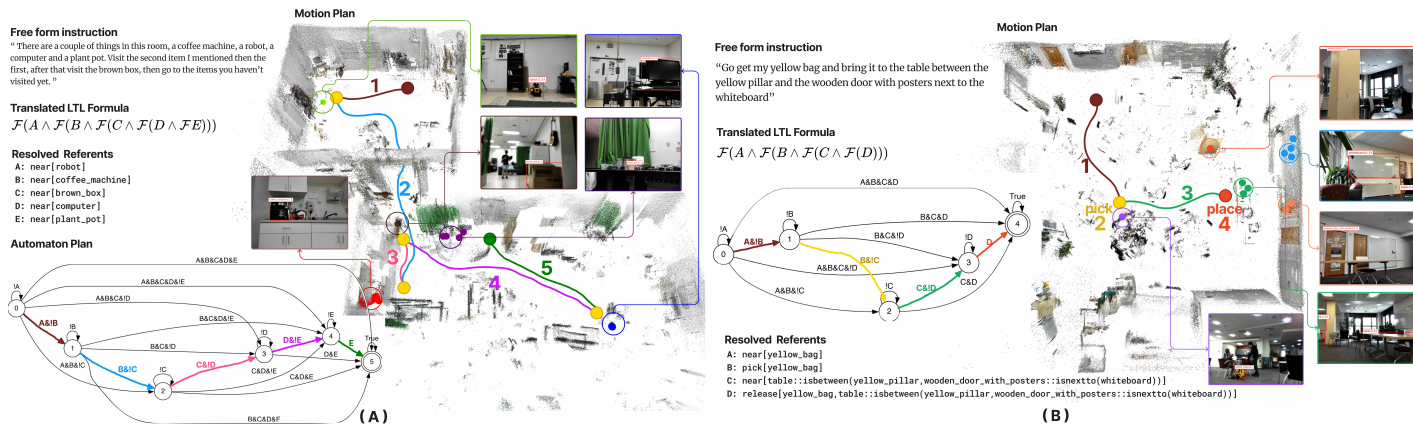


Fig. A.1: [A] Sample generated plan for a multi object-goal navigation task. [B] Sample generated plan for a mobile pick-and-place task.

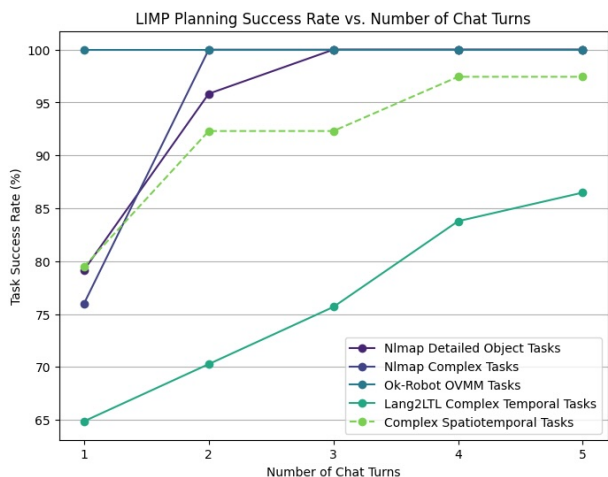


Fig. A.2: Our interactive reprompting strategy implemented in the symbol verification node regenerates corrective formulas that improve plan success rates with minimal chat turns.

A. NLMaP-Saycan Implementation Prompt

1 You are an LLM for robot planning that understands logical operators such as `&`, `|`, `~`, etc. You have a finite set of robot predicates and spatial predicates, given a language instruction, your goal is to generate a sequence of actions that uses appropriate composition of robot and spatial predicates with relevant details from the instruction as arguments.

2 Robot predicate set (`near`, `pick`, `release`).

3 Usage:

4 `near[referent_1]:` returns true if the desired spatial relationship is for robot to be `near` referent_1.

5 `pick[referent_1]:` can only execute picking skill on referent_1 and return True when `near[referent_1]`.

6 `release[referent_1, referent_2]:` can only execute `release` skill on referent_1 and return True when `near[referent_2]`.

7 Spatial predicate set (`isbetween`, `isabove`, `isbelow`, `isleftof`, `isrightof`, `isnextto`, `isinfrontof`, `isbehind`).

8 Usage:

9 `referent_1::isbetween(referent_2, referent_3):` returns true if referent_1 is between referent_2 and referent_3.

10 `referent_1::isabove(referent_2):` returns True if referent_1 is above referent_2.

11 `referent_1::isbelow(referent_2):` returns True if referent_1 is below referent_2.

12 `referent_1::isleftof(referent_2):` returns True if referent_1 is left of referent_2.

13 `referent_1::isrightof(referent_2):` returns True if referent_1 is right of referent_2.

14 `referent_1::isnextto(referent_2):` returns True if referent_1 is close to referent_2.

15 `referent_1::isinfrontof(referent_2):` returns True if referent_1 is in front of referent_2.

16 `referent_1::isbehind(referent_2):` returns True if referent_1 is behind referent_2.

17 Rules:

18 Strictly only use the finite set of robot and spatial predicates!

19 Strictly stick to the usage format!

20 Compose spatial predicates where necessary!

21 You should strictly stick to mentioned objects, however you are allowed to propose and include plausible objects if and only if not mentioned in instruction but required based on context of instruction!

22 Pay attention to instructions that require performing certain actions multiple times in generating and sequencing the predicates for the final **Output!**

23 Example:

24 **Input_instruction:** Go to the orange building but before that pass by the coffee shop, then go to the parking sign.

25 **Output:**

26 1. `near[coffee_shop]`

27 2. `near[orange_building]`

28 3. `near[parking_sign]`

29 **Input_instruction:** Go to the blue sofa then the laptop, after that bring me the brown bag between the television and the kettle on the left of the green seat, I am standing by the sink.

30 **Output:**

31 1. `near[blue_sofa]`

32 2. `near[laptop]`

33 3. `near[brown_bag::isbetween(television, kettle::isleftof(green_seat))]`

34 4. `pick[brown_bag::isbetween(television, kettle::isleftof(green_seat))]`

35 5. `near[sink]`

36 6. `release[brown_bag, sink]`

37 **Input_instruction:** Hey need you to pass by chair between the sofa and bag, `pick` up the bag and go to the orange napkin on the right of the sofa.

38 **Output:**

39 1. `near[chair::isbetween(sofa, bag)]`

40 2. `near[bag]`

41 3. `pick[bag]`

42 4. `near[orange_napkin::isrightof(sofa)]`

43 **Input_instruction:** Go to the chair between the green laptop and the yellow box underneath the play toy

44 **Output:**

45 1. `near[chair::isbetween(green_laptop, yellow_box::isbelow(play_toy))]`

46 **Input_instruction:** Check the table behind the fridge and bring two beers to the couch one after the other

47 **Output:**

48 1. `near[table::isbehind(fridge)]`

49 2. `pick[beer]`

50 3. `near[couch]`

51 4. `release[beer, couch]`

52 5. `near[table::isbehind(fridge)]`

53 6. `pick[beer]`

54 7. `near[couch]`

55 8. `release[beer, couch]`

```
56 Input_instruction: <given_instruction>
57 Output:
```

Listing 6: Exact prompt to implement NIMap-Saycan LLM planner

B. Code-as-Policies Implementation Prompt

```
1 ##Python robot planning script
2 from robotactions import near, pick, release
3 spatial_relationships = [
4 "isbetween", #referent_1::isbetween(referent_2,referent_3)
   :returns true if referent_1 is between referent_2 and
   referent_3.
5 "isabove", #referent_1::isabove(referent_2):returns True
   if referent_1 is above referent_2.
6 "isbelow", #referent_1::isbelow(referent_2):returns True
   if referent_1 is below referent_2.
7 "isleftof", #referent_1::isleftof(referent_2):returns
   True if referent_1 is left of referent_2.
8 "isrightof", #referent_1::isrightof(referent_2):returns
   True if referent_1 is right of referent_2.
9 "isnextto", #referent_1::isnextto(referent_2):returns
   True if referent_1 is close to referent_2.
10 "isinfrontof", #referent_1::isinfrontof(referent_2):
   returns True if referent_1 is in front of referent_2.
11 "isbehind" #referent_1::isbehind(referent_2):returns True
   if referent_1 is behind referent_2.]
12 ##Rules:
13 ##Strictly only use the finite set of robot and spatial
   predicates!
14 ##Strictly stick to the usage format!
15 ##Compose spatial predicates where necessary!
16 ##You should strictly stick to mentioned objects, however
   you are allowed to propose and include plausible
   objects if and only if not mentioned in instruction
   but required based on context of instruction!
17 ##Pay attention to instructions that require performing
   certain actions multiple times in generating and
   sequencing the predicates for the final Output!
18 # Go to the orange building but before that pass by the
   coffee shop, then go to the parking sign.
19 ordered_navigation_goal_referents = ["coffee_shop", "
   orange_building", "parking_sign"]
20 for referent in ordered_navigation_goal_referents:
21     near(referent)
22 # Go to the blue sofa then the laptop, after that bring me
   the brown bag between the television and the kettle
   on the left of the green seat, I am standing by the
   sink.
23 ordered_navigation_goal_referents = ["blue_sofa", "laptop
   ", "brown_bag::isbetween(television,kettle::isleftof(
   green_seat)", "sink"]
24 referents_to_pick= ["brown_bag::isbetween(television,
   kettle::isleftof(green_seat))"]
25 release_location_referents = ["sink"]
26 picked_item = None
27 for referent in ordered_navigation_goal_referents:
28     near(referent)
29     if referent in referents_to_pick:
30         pick(referent)
31         picked_item = referent
32     if referent in release_location_referents:
33         release(picked_item, referent)
34 #Hey need you to pass by chair between the sofa and bag,
   pick up the bag and go to the orange napkin on the
   right of the sofa.
35 ordered_navigation_goal_referents = ["chair::isbetween(
   sofa,bag)", "bag", "orange_napkin::isrightof(sofa)"]
36 referents_to_pick= ["bag"]
37 picked_item = None
38 for referent in ordered_navigation_goal_referents:
39     near(referent)
40     if referent in referents_to_pick:
41         pick(referent)
42         picked_item = referent
43 #Go to the chair between the green laptop and the yellow
   box underneath the play toy
44 near("chair::isbetween(green_laptop,yellow_box::isbelow(
   play_toy))")
45 #Check the table behind the fridge and bring two beers to
   the couch one after the other
46 for i in range(2):
47     near("table::isbehind(fridge)")
```

```
48 pick("beer")
49 near("couch")
50 release("beer", "couch")
51
52 #<given_instruction>
```

Listing 7: Exact prompt to implement Code-as-policies planner

C. Instruction set

We perform a large-scale evaluation on 150 instructions across five real-world environments. The taskset includes 24 tasks with fine-grained object descriptions (NLMD) from [11], 25 tasks with complex language (NLMD) from [11], 25 tasks with simple structured phrasing (OKRB) from [21], 37 tasks with complex temporal structures (CT) from [16], and an additional 39 tasks we propose that have descriptive spatial constraints and temporal structures (CST).

```
1 1: put the red can in the trash bin
2 2: put the brown multigrain chip bag in the woven basket
3 3: find the succulent plant
4 4: pick up the up side down mug
5 5: put the apple on the macbook with yellow stickers
6 6: use the dyson vacuum cleaner"
7 7: bring the kosher salt to the kitchen counter
8 8: put the used towels in washing machine
9 9: move the used mug to the dish washer
10 10: place the pickled cucumbers on the shelf
11 11: find my mug with the shape of a donut
12 12: put the almonds in the almond jar
13 13: fill the zisha tea pot with a coke from the cabinet
14 14: take the slippery floor sign with you
15 15: take the slippers that have holes on them to the shoe
   rack
16 16: find the mug on the mini fridge
17 17: bring the mint flavor gum to the small table
18 18: find some n95 masks
19 19: grab the banana with most black spots
20 20: fill the empty bottle with lemon juice
21 21: throw away the apple that's about to rot
22 22: throw away the rotting banana
23 23: take the box of organic blueberries out of the fridge
24 24: give a can of diet coke to the toy cat
```

Listing 8: Nlmap Detailed Object Tasks (NLMD)

```
1 1: I opened a pepsi earlier, bring an open can to the
   orange table
2 2: I spilled my coke, can you put a replacement on the
   kitchen counter
3 3: I spilled some coke on the television, go and bring
   something to clean it up
4 4: I accidentally dropped that jalapeno chips after eating
   it. Would you mind throwing it away
5 5: I like fruits, can you put something I would like on
   the yellow sofa for me
6 6: There is a green counter, a yellow counter, and a table
   . visit all the locations
7 7: There is a green counter, a trash can, and a table.
   visit all the locations
8 8: Redbull is my favoriorite drink, can you put one on the
   desk please
9 9: Would you bring a coke can to the door for me
10 10: Please, move the pepsi to the red counter
11 11: Can you move the coke can to the orange counter
12 12: Would you throw away the bag of chips for me
13 13: Put an energy bar and water bottle on the table
14 14: Bring a lime soda and a bag of chips to the sofa
15 15: Can you throw away the apple and bring a coke to the
   bed
16 16: Bring a 7up can and a tea to the office desk
17 17: Move the multigrain chips to the table and an apple to
   the yellow counter
18 18: Move the lime soda, the sponge, and the water bottle
   to the table
19 19: Bring two sodas to the table
20 20: Move three cokes to the trash can
21 21: Throw away two cokes from the counter
```

- 22: Bring two different sodas to the cabinet, there is a coke, pepsi, soup, tea and 7up in the fridge
- 23: Bring an apple, a coke, and water bottle to the sofa
- 24: I spilled my coke on the table, throw it away and then bring something to help clean
- 25: I just worked out, can you bring me a drink and a snack to recover, i am on the sof

Listing 9: Nlmap Complex Language Understanding Tasks (NLMC)

- 1: Move the Takis on the desk to the nightstand
- 2: Move the soda can to the box
- 3: Move the purple shampoo to the red bag
- 4: Move the white meds box to the trash bin
- 5: Move the power adapter to the chair
- 6: Move the blue gloves to the sink
- 7: Move the McDonalds paper bag to the stove
- 8: Move the herbal tea can to the box
- 9: Move the cooking oil bottle to the marble surface
- 10: Move the milk bottle to the chair
- 11: Move the purple shampoo to the white rack
- 12: Move the purple lightbulb box to the sofa chair
- 13: Pick up purple medicine, drop it on chair
- 14: Pick up eyeglass case, drop it on chair
- 15: Pick up grey rag, drop it in sink
- 16: Pick up golden can rag, drop it on table
- 17: Pick up red navaratna oil, drop it on table
- 18: Pick up purple shampoo, drop it on green rack
- 19: Pick up taki chips, drop it on countertop
- 20: Pick up bandage box, drop it in dustbin
- 21: Pick up white aerosol, drop it in trash can
- 22: Pick up peanut butter, drop it on countertop
- 23: Pick up blue gloves, drop it in sink
- 24: Pick up brown box, drop it on chair
- 25: Pick up axe body spray, drop it on shel

Listing 10: Ok-Robot Tasks (OKRB)

- 1: go to brown bookshelf, metal desk, wooden desk, kitchen counter, and the blue couch in any order
- 2: move to grey door, then bookrack, then go to the brown desk, then counter, then white desk
- 3: visit brown wooden desk but only after bookshelf
- 4: go from brown bookshelf to white metal desk and only visit each landmark one time
- 5: go to brown wooden desk exactly once and do not visit brown desk before bookshelf
- 6: go to brown desk only after visiting bookshelf, in addition go to brown desk only after visiting white desk
- 7: visit the blue IKEA couch, in addition never go to the big steel door
- 8: visit white kitchen counter then go to brown desk, in addition never visit white table
- 9: go to the grey door, and only then go to the bookshelf, in addition always avoid the table
- 10: go to kitchen counter then wooden desk, in addition after going to counter, you must avoid white table
- 11: Go to bookshelf, alternatively go to metal desk
- 12: Go to counter, alternatively go to metal desk
- 13: Go to the counter, but never visit the counter
- 14: do not go to the wooden desk until bookshelf, and do not go to bookshelf until wooden desk
- 15: go to brown desk exactly once, in addition go to brown desk at least twice
- 16: move to couch exactly twice, in addition pass by counter at most once
- 17: navigate to the counter then the brown desk, in addition after going to the counter, you must avoid doorway
- 18: visit counter at least six times
- 19: either go to bookshelf then the brown desk, or go to couch
- 20: navigate to the wooden door, the glass door and the table, kitchen counter, and the blue couch in any order
- 21: go to the painting, then find the kitchen table, then front desk, then staircase
- 22: navigate to classroom but do not visit classroom before the white table
- 23: only visit classroom once, and do not visit classroom until you visit elevator first

- 24: Go to the staircase, front desk and the white table in that exact order. You are not permitted to revisit any of these locations
- 25: go to the front desk then the yellow office door, in addition do not visit the glass door
- 26: go to the stairs then the front desk, in addition avoid purple elevator
- 27: move to elevator then front desk, in addition avoid staircase
- 28: go to front desk then the cabinet, always avoid the elevator
- 29: Go to elevator, alternatively go to staircase
- 30: Visit the elevator exactly once, in addition visit the front desk on at least 2 separate occasions
- 31: Go to the office, in addition avoid visiting the elevator and the classroom
- 32: Visit the front desk, in addition you are not permitted to visit elevator and staircase
- 33: Visit the purple door elevator, then go to the front desk and then go to the kitchen table, in addition you can never go to the elevator once you have seen the front desk
- 34: Visit the front desk then the sofa then the white table, in addition if you visit the sofa you must avoid the television after that
- 35: Go to the glass door, but never visit the glass door
- 36: do not go to the white table until classroom, and do not go to the classroom until white table
- 37: find the office, in addition avoid visiting the front desk and the classroom and the table

Listing 11: Lang2LTL Complex temporal Tasks (CT)

- 1: Go to the red sofa but dont pass by any door then go to the computer, the sofa is behind the pillar and on the right side
- 2: Find the door with posters in front of the yellow pillar then go stand by the tree but avoid any sofa
- 3: There is a cabinet, a television and a tree in this room. I want you to go to the last item I mentioned then pass by the second item but dont go close to the first when you are going to thsecond item.",
- 4: Go to the trash bin, after that go to the large television. Actually can you go to the laptop on the table before doing all that?
- 5: Try to go to the tree without going near the trash bin.
- 6: Find the yellow trash bin and go to it, then go by the white table. But before doing the first thing visit where the fridge is then after the last thing, stop by the red sofa but avoid the wooden door with posters",
- 7: Go to the television but avoid the tree.
- 8: Go to any trash bin but not the yellow one
- 9: Visit one of the televisions in this room then go to the fridge to the left of the cabinet
- 10: I need you to first go to the table between the pillar and the door with posters, then go to the tree. You know what, ignore the last thing I asked you to go to , after the first thing go tthe computer next to the green box instead.
- 11: Go to the red sofa then the blue one next to the cabinet but first pass by the book shelf
- 12: Go to water filter, dont pass by any computer to the left of the red sofa
- 13: Visit these things in the following order the microwave, then the black chair infront of the yellow robot then the fridge, but before doing any of this go to the robot
- 14: Go to the cabinet between the blue sofa and the brown box.
- 15: I want you to go the cabinet but dont go near any black chair. Hold on, actually can you pass by the whiteboard before doing all that?
- 16: I need you to go to the red sofa but dont go anywhere near the sink
- 17: You can either go as close as you can to the sink or the plant pot but dont go near the yellow robot.
- 18: Go to the green curtain after that the large television but try not to go near the red sofa
- 19: Go to the sink then the red sofa, then the guitar. After that I want you to return to the first thing you visited.
- 20: I need you to go to the sink then any computer on a table but before you start that can you first go to

the bookshelf, when you are done with all that, return to brown box

21 21: There are a couple of things in this room, a coffee machine, a robot, a computer and a plant pot. Visit the second item I mentioned then the first, after that visit the brown box, then go the items you haven't visited yet.

22 22: Avoid any cabinet but go get the mug on the blue sofa and take it to red one

23 23: Please get me my plant pot from between the sofa and the sink, bring it to the white cabinet

24 24: Can you pick up my book from the left side of the television and bring it to me? I am sitting on the couch which is to the right of the door with posters. Make sure to never go near the trhb in while doing all this, it on the yellow table.

25 25: Bring the green plush toy to the whiteboard in front of it, watch out for the robot in front of the toy

26 26: Go get my yellow bag and bring it to the table between the yellow pillar and the wooden door with posters next to the whiteboard

27 27: Find the bottle that is on the table to the left of the computer and bring it to the wardrobe that is next to the glass door

28 28: Go to the cabinet between the blue sofa and the yellow robot or the cabinet to the left of the blue sofa.

29 29: Go take the green toy that is next to the sofa under the poster and bring it to the bookshelf

30 30: I have a white cabinet, a green toy, a bookshelf and a red chair around here somewhere. Take the second item I mentioned from between the first item and the third. Bring it the cabinet but void the last item at all costs.

31 31: Hey I am standing by the whiteboard in front of the bookshelf, can you bring me the mug from the table to the left of the fridge?

32 32: Never pass by a robot, but I need you to bring the bag on the table to the cabinet

33 33: Take a soda can to the fridge, you can find one on the table to the left of the blue sofa

34 34: Hey, can you pick up my bag for me? It is under the table in front of the glass door, bring it to the cabinet. Actually you know what, forget what I asked for and bring the green toy instead its at the same place",

35 35: Go to the sink and take the mug beside the coffee maker, drop it off at the red sofa

36 36: Go to the sofa behind the chair next to the orange counter in front of the pillar

37 37: I want you to go visit the chair behind the sofa then the green bag, do that 3 times but during the second time avoid the fridge

38 38: Go bring the plush toy between the sofa and the television to the couch, my cat is on the brown table so please don't pass near the table when you are returning with the toy

39 39: I think I left my wallet on the kitchen counter, go get it I will meet you at the bed. There is a lantern near the bed, make sure you don't hit it

Listing 12: Complex Spatiotemporal Tasks (CST)

REFERENCES

- [1] S. Tellex, N. Gopalan, H. Kress-Gazit, and C. Matuszek, "Robots that use language," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 25–55, 2020.
- [2] V. Blukis, R. Knepper, and Y. Artzi, "Few-shot Object Grounding and Mapping for Natural Language Robot Instruction Following," in *Proceedings of the 2020 Conference on Robot Learning*. PMLR, Oct. 2021, pp. 1829–1854, iSSN: 2640-3498. [Online]. Available: <https://proceedings.mlr.press/v155/blukis21a.html>
- [3] R. Patel, E. Pavlick, and S. Tellex, "Grounding Language to Non-Markovian Tasks with No Supervision of Task Specifications," in *Robotics: Science and Systems XVI*. Robotics: Science and Systems Foundation, July 2020. [Online]. Available: <http://www.roboticsproceedings.org/rss16/p016.pdf>
- [4] C. Wang, C. Ross, Y.-L. Kuo, B. Katz, and A. Barbu, "Learning a natural-language to LTL executable semantic parser for grounded robotics," in *Proceedings of the 2020 Conference on Robot Learning*. PMLR, Oct. 2021, pp. 1706–1718, iSSN: 2640-3498. [Online]. Available: <https://proceedings.mlr.press/v155/wang21g.html>
- [5] K. Zheng, D. Bayazit, R. Mathew, E. Pavlick, and S. Tellex, "Spatial Language Understanding for Object Search in Partially Observed City-scale Environments," *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pp. 315–322, Aug. 2021, conference Name: 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN) ISBN: 9781665404921 Place: Vancouver, BC, Canada Publisher: IEEE. [Online]. Available: <https://ieeexplore.ieee.org/document/9515426/>
- [6] M. Berg, D. Bayazit, R. Mathew, A. Rotter-Aboyoun, E. Pavlick, and S. Tellex, "Grounding Language to Landmarks in Arbitrary Outdoor Environments," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, May 2020, pp. 208–215, iSSN: 2577-087X. [Online]. Available: <https://ieeexplore.ieee.org/document/9197068>
- [7] M. Cosler, C. Hahn, D. Mendoza, F. Schmitt, and C. Trippel, "nl2spec: Interactively Translating Unstructured Natural Language to Temporal Logics with Large Language Models," in *Computer Aided Verification*, C. Enea and A. Lal, Eds. Cham: Springer Nature Switzerland, 2023, pp. 383–396.
- [8] X. Wang, W. Wang, J. Shao, and Y. Yang, "LANA: A Language-Capable Navigator for Instruction Following and Generation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, June 2023, pp. 19048–19058. [Online]. Available: <https://ieeexplore.ieee.org/document/10203301/>
- [9] S.-M. Park and Y.-G. Kim, "Visual language navigation: a survey and open challenges," *Artificial Intelligence Review*, vol. 56, no. 1, pp. 365–427, Jan. 2023. [Online]. Available: <https://doi.org/10.1007/s10462-022-10174-9>
- [10] D. Shah, B. Osifski, B. Ichter, and S. Levine, "LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action," in *Proceedings of The 6th Conference on Robot Learning*. PMLR, Mar. 2023, pp. 492–504, iSSN: 2640-3498. [Online]. Available: <https://proceedings.mlr.press/v205/shah23b.html>
- [11] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, M. S. Ryoo, A. Stone, and D. Kappler, "Open-vocabulary Queryable Scene Representations for Real World Planning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023, pp. 11 509–11 522. [Online]. Available: <https://ieeexplore.ieee.org/document/10161534>
- [12] B. Yu, H. Kasaei, and M. Cao, "L3MVN: Leveraging Large Language Models for Visual Target Navigation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2023, pp. 3554–3560, iSSN: 2153-0866. [Online]. Available: <https://ieeexplore.ieee.org/document/10342512>
- [13] C. H. Song, B. M. Sadler, J. Wu, W.-L. Chao, C. Washington, and Y. Su, "LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, pp. 2986–2997. [Online]. Available: <https://ieeexplore.ieee.org/document/10378628/>
- [14] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual Language Maps for Robot Navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023, pp. 10 608–10 615. [Online]. Available: <https://ieeexplore.ieee.org/document/10160969>
- [15] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as Policies: Language Model Programs for Embodied Control," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023, pp. 9493–9500. [Online]. Available: <https://ieeexplore.ieee.org/document/10160591>
- [16] J. X. Liu, Z. Yang, I. Idrees, S. Liang, B. Schornstein, S. Tellex, and A. Shah, "Grounding Complex Natural Language Commands for Temporal Tasks in Unseen Environments," in *Proceedings of The 7th Conference on Robot Learning*. PMLR, Dec. 2023, pp. 1084–1110, iSSN: 2640-3498. [Online]. Available: <https://proceedings.mlr.press/v229/liu23d.html>
- [17] E. Hsiung, H. Mehta, J. Chu, X. Liu, R. Patel, S. Tellex, and G. Konidaris, "Generalizing to New Domains by Mapping Natural Language to Lifted LTL," in *2022 International Conference on Robotics and Automation (ICRA)*. Philadelphia, PA, USA: IEEE Press, May 2022, pp. 3624–3630. [Online]. Available: <https://doi.org/10.1109/ICRA46639.2022.9812169>
- [18] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models," in *Proceedings of The 7th Conference on*

- Robot Learning*. PMLR, Dec. 2023, pp. 540–562, iISSN: 2640-3498. [Online]. Available: <https://proceedings.mlr.press/v229/huang23b.html>
- [19] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suenderhauf, “SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Robot Task Planning,” in *Proceedings of The 7th Conference on Robot Learning*. PMLR, Dec. 2023, pp. 23–72, iISSN: 2640-3498. [Online]. Available: <https://proceedings.mlr.press/v229/rana23a.html>
- [20] R.-Z. Qiu, Y. Hu, G. Yang, Y. Song, Y. Fu, J. Ye, J. Mu, R. Yang, N. Atanasov, S. A. Scherer, and X. Wang, “Learning Generalizable Feature Fields for Mobile Manipulation,” *CoRR*, vol. abs/2403.07563, 2024, arXiv: 2403.07563. [Online]. Available: <https://doi.org/10.48550/arXiv.2403.07563>
- [21] P. Liu, Y. Orru, J. Vakil, C. Paxton, N. Shafiullah, and L. Pinto, “Demonstrating OK-Robot: What Really Matters in Integrating Open-Knowledge Models for Robotics,” in *Robotics: Science and Systems*. Robotics: Science and Systems Foundation, July 2024. [Online]. Available: <http://www.roboticsproceedings.org/rss20/p091.pdf>
- [22] I. Kostavelis and A. Gasteratos, “Semantic mapping for mobile robotics tasks: A survey,” *Robotics and Autonomous Systems*, vol. 66, pp. 86–103, Apr. 2015. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0921889014003030>
- [23] J. Crespo, J. C. Castillo, O. M. Mozos, and R. Barber, “Semantic Information for Robot Navigation: A Survey,” *Applied Sciences*, vol. 10, no. 2, p. 497, Jan. 2020, number: 2 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: <https://www.mdpi.com/2076-3417/10/2/497>
- [24] A. Pronobis, P. Jensfelt, and J. Little, *Semantic Mapping with Mobile Robots*. Stockholm: KTH Royal Institute of Technology, 2011.
- [25] E. Rosen, S. James, S. Orozco, V. Gupta, M. Merlin, S. Tellex, and G. Konidaris, “Synthesizing Navigation Abstractions for Planning with Portable Manipulation Skills,” in *Proceedings of The 7th Conference on Robot Learning*. PMLR, Dec. 2023, pp. 2278–2287, iISSN: 2640-3498. [Online]. Available: <https://proceedings.mlr.press/v229/rosen23a.html>
- [26] R. E. Fikes and N. J. Nilsson, “Strips: A new approach to the application of theorem proving to problem solving,” *Artificial intelligence*, vol. 2, no. 3-4, pp. 189–208, 1971.
- [27] C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, and T. Lozano-Pérez, “Integrated Task and Motion Planning,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, no. Volume 4, 2021, pp. 265–293, May 2021, publisher: Annual Reviews. [Online]. Available: <https://www.annualreviews.org/content/journals/10.1146/annurev-control-091420-084139>
- [28] C. R. Garrett, T. Lozano-Pérez, and L. P. Kaelbling, “PDDLStream: Integrating Symbolic Planners and Blackbox Samplers via Optimistic Adaptive Planning,” *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 30, pp. 440–448, June 2020. [Online]. Available: <https://ojs.aaai.org/index.php/ICAPS/article/view/6739>
- [29] R. Holladay, T. Lozano-Pérez, and A. Rodriguez, “Planning for Multi-stage Forceful Manipulation,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, May 2021, pp. 6556–6562, iISSN: 2577-087X. [Online]. Available: <https://ieeexplore.ieee.org/document/9561233>
- [30] J. Pan, G. Chou, and D. Berenson, “Data-Efficient Learning of Natural Language to Linear Temporal Logic Translators for Robot Task Specification,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023, pp. 11 554–11 561. [Online]. Available: <https://ieeexplore.ieee.org/document/10161125>
- [31] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, and N. Houlsby, “Simple Open-Vocabulary Object Detection,” in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 728–755.
- [32] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment Anything,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 3992–4003, iISSN: 2380-7504. [Online]. Available: <https://ieeexplore.ieee.org/document/10378323>