Planning with Vision-Language Models and a Use Case in Robot-Assisted Teaching

Xuzhe Dang[1], Lada Kudláčková[2], Stefan Edelkamp[1,2]

[1] Computer Science Department, Artificial Intelligence Center, Faculty of Electrical Engineering, Czech Technical University, Prague, Czech Republic

[2]Department of Theoretical Computer Science and Mathematical Logic, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

dangxuzh@fel.cvut.cz, kudlackova@ktiml.cuni.cz, stefan.edelkamp@gmail.com

Abstract

Automating the generation of Planning Domain Definition Language (PDDL) with Large Language Model (LLM) opens new research topic in AI planning, particularly for complex real-world tasks. This paper introduces Image2PDDL, a novel framework that leverages Vision-Language Models (VLMs) to automatically convert images of initial states and descriptions of goal states into PDDL problems. By providing a PDDL domain alongside visual inputs, Imasge2PDDL addresses key challenges in bridging perceptual understanding with symbolic planning, reducing the expertise required to create structured problem instances, and improving scalability across tasks of varying complexity. We evaluate the framework on various domains, including standard planning domains like blocksworld and sliding tile puzzles, using datasets with multiple difficulty levels. Performance is assessed on syntax correctness, ensuring grammar and executability, and content correctness, verifying accurate state representation in generated PDDL problems. The proposed approach demonstrates promising results across diverse task complexities, suggesting its potential for broader applications in AI planning. We will discuss a potential use case in robot-assisted teaching of students with Autism Spectrum Disorder.

Introduction

Automating the generation of Planning Domain Definition Language (PDDL) problems has long been a challenge in AI planning, particularly for applications involving complex, real-world tasks. Traditional approaches to PDDL generation often require domain-specific knowledge and substantial manual effort to accurately structure problem instances. This barrier limits the scalability and accessibility of AI planning, as defining object relationships, spatial configurations, and task goals is labor-intensive and demands expertise.

Recent advances in Large Language Models (LLMs) have opened new possibilities by enabling models to interpret and translate visual and textual data into structured, symbolic formats, bridging the gap between perceptual understanding and symbolic reasoning (Liu et al. 2023; Xie et al. 2023; Shirai et al. 2024). However, each approach has its limitations. LLM+P (Liu et al. 2023) requires text descriptions for both initial and goal states, which, in practical applications, depends on human input or additional tools to convert images to text. Xie's work (Xie et al. 2023) focuses solely on translating text descriptions of goal states, lacking a mechanism to interpret visual data. Meanwhile, ViLaIn (Shirai et al. 2024) can generate PDDL problems from images but relies on an object detection model and a captioning model to process images, adding complexity to the pipeline. These limitations highlight the need for a more streamlined approach that can directly process both visual and textual inputs for automated PDDL problem generation.

To address these challenges, we present Image2PDDL, a novel framework that leverages Vision-Language Models (VLMs) to automatically generate PDDL problems from both visual and textual inputs. Image2PDDL is designed to handle a wide range of input formats, including images of initial and goal states as well as textual descriptions, allowing it to adapt seamlessly across diverse domains and task complexities. By directly interpreting spatial and categorical relationships within visual data and integrating them with goal-oriented descriptions, Image2PDDL provides a flexible solution for PDDL problem generation. We evaluated the framework across traditional planning domains—Blocksworld and Sliding-Tile Puzzle—as well as a 3D world domain, Kitchen, assessing both syntax correctness and content correctness of the generated PDDL problems. In all domains, Image2PDDL demonstrated promising results, effectively reducing the need for domain-specific expertise and making AI planning more accessible, scalable, and applicable to real-world scenarios.

Image2PDDL makes several key contributions to the field of AI planning. Its adaptability across both traditional planning domains and complex 3D environments underscores its versatility and scalability, positioning it as a valuable tool for a wide range of AI planning applications. By reducing dependency on domain-specific knowledge, Image2PDDL broadens accessibility and opens new possibilities for applying AI planning to real-world tasks. Future work could focus on enhancing the model's capacity to interpret more intricate object relationships and dynamic scenarios, further expanding the potential of automated planning across diverse and complex environments.

We will first introduce Image2PDDL together with experiments in several benchmark domains, and then discuss a

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

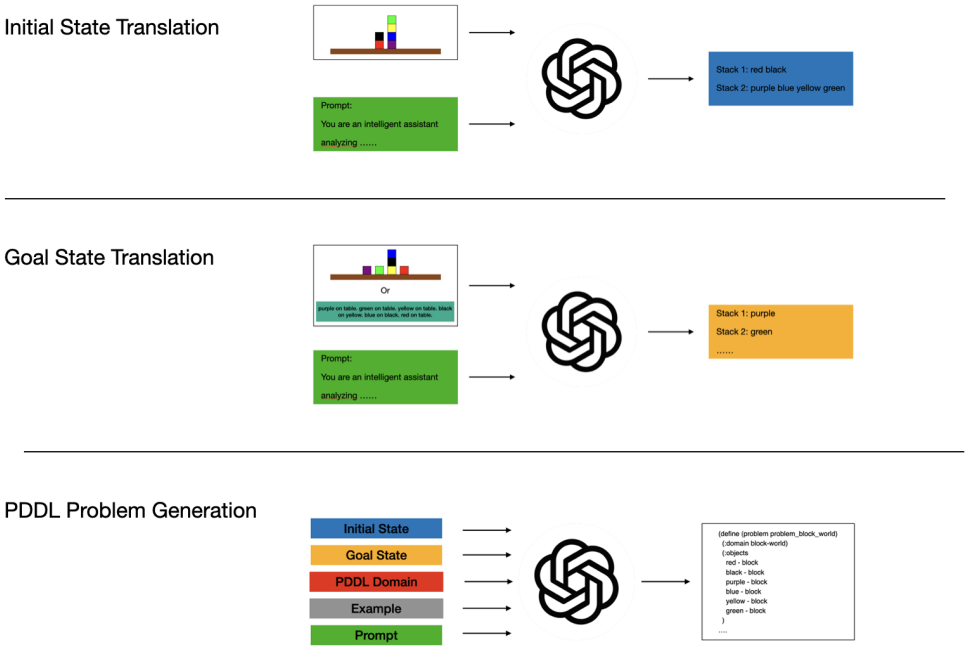

Figure 1: Image2PDDL operates in three main steps. First, the image of the initial state is translated into a predefined state format. Next, either an image or text description of the goal state is similarly converted to the same state format. Finally, both states are used to generate a PDDL problem based on predefined domains and examples.

case in automated teaching of students with Autism Spectrum Disorders (ASD).

Related Work

Recent advances in LLMs and VLMs have spurred interest in their applications to automated planning. Most current work in this area focuses on using LLMs and VLMs as planners, where the models are tasked with generating plans directly based on high-level descriptions or visual inputs (Huang et al. 2022; Raman et al. 2022; Zhang et al. 2023; Dagan, Keller, and Lascarides 2023; Guan et al. 2023). In these approaches, the models produce a sequence of symbolic actions (Lin et al. 2023), code (Liang et al. 2023), or predefined skills (Singh et al. 2023; Brohan et al. 2023) to complete a task, acting as planners that bridge the gap between high-level goals and executable actions. A significant drawback of these methods is that the generated plans are not always guaranteed to be correct, as both LLMs and VLMs may produce sequences that overlook domain constraints or violate task feasibility. Additionally, the planning process is inherently opaque, making it difficult to verify or explain the rationale behind the generated plans.

Recent research has also explored the potential of LLMs and VLMs in directly generating PDDL domains or PDDL problems, shifting the focus from plan generation to the formulation of structured problem instances (Liu et al. 2023; Xie et al. 2023; Shirai et al. 2024; Smirnov et al. 2024; Oswald et al. 2024). These approaches aim to leverage the language and visual understanding capabilities of LLMs and VLMs to automate the creation of PDDL problem definitions, which include initial and goal states and domain-specific constraints.

Method

The Image2PDDL framework leverages VLM to parse visual and textual inputs, transforming them into Planning Domain Definition Language (PDDL) problem representations based on pre-defined domains. The core idea is to utilize the spatial relationship understanding that VLMs offer, enabling the translation of complex images and descriptions into structured PDDL problems, which traditionally require extensive domain expertise to manually create.

Framework Overview

Image2PDDL consists of a pipeline, as shown in Figure 1, that translates an image of an initial state and either an image or text description of a goal state into PDDL problem format. This is achieved in in the following three main steps, each using ChatGPT4o for structured output generation.

Initial State Translation: First, we input the image of the initial state into ChatGPT4o, which is guided by a structured prompt. This prompt includes an example of the desired output format to encourage accurate parsing of spatial relationships between objects. The model generates a pre-defined textual representation of the spatial relationships observed in the initial state, such as object locations and their relative positions.

Goal State Translation: For the goal state, we offer either an image or a text description. Similarly to the initial state, ChatGPT4o is prompted to parse this goal state in-

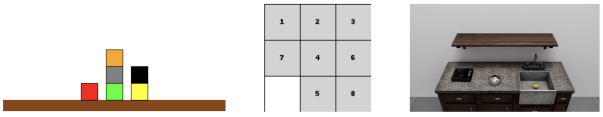

Figure 2: Images of example scenarios of each domain.

formation into the same structured format, ensuring consistency between the initial and goal state representations. This format standardization is crucial for correctly translating the data into PDDL syntax.

PDDL Problem Generation: With structured text representations of both the initial and goal states, we proceed to PDDL problem generation. ChatGPT4o receives the initial and goal states, the corresponding PDDL domain definition, and an example format to guide output structure. Additional information, such as predicate rules and detailed state structure definitions, is also included to refine output quality and enhance syntax correctness and executability.

This pipeline allows Image2PDDL to leverage the spatial reasoning capabilities of VLMs, providing an automated, scalable solution to generate PDDL problems for diverse planning domains. The iterative prompt refinements and examples enable ChatGPT4o to achieve high accuracy in syntax and content representation, making the framework effective for complex task definitions across different difficulty levels.

Experiments and Evaluation

Dataset

To evaluate the Image2PDDL framework, we prepared datasets across three distinct domains: Blocksworld, Sliding-Tile Puzzle, and Kitchen Domain. Each domain includes scenarios categorized by difficulty levels—Easy, Medium, and Hard—allowing us to test the framework's scalability and accuracy across a range of task complexities. For each difficulty level, we prepared 50 unique scenarios. Table 1 shows object types, predicates, and actions of each domains. Figure 2 shows image of an scenario from each domain.

The Blocksworld domain assesses Image2PDDL's ability to interpret spatial relationships by arranging blocks, with difficulties based on the number of blocks (5 for Easy, 6 for Medium, and 7 for Hard). In the Sliding-Tile Puzzle domain, the framework rearranges tiles to match a goal configuration, testing sequential spatial handling across three puzzle sizes: 8 tiles (Easy), 15 tiles (Medium), and 24 tiles (Hard). The Kitchen Domain, created in IsaacSim, evaluates real-world scenario comprehension by identifying and categorizing objects. Task difficulty increases from identifying a single item's location (Easy) to distinguishing between two items by attributes (Medium) and recognizing three items with finer detail, such as brand or color (Hard).

For each scenario, the dataset provides an initial and goal image alongside a textual goal state description, enabling Image2PDDL to process diverse data formats and effectively translate them into structured PDDL problems. This setup allows for a systematic evaluation of the framework's ability to handle different levels of task complexity across varied planning domains.

Results

Performance was assessed based on two metrics: Syntax Correctness and Content Correctness. Syntax correctness was verified by passing the generated PDDL problems to a Fast Downward planner (Helmert 2006) to ensure each problem could be successfully parsed and executed without syntax errors. Content correctness was evaluated by comparing the initial and goal states described in the generated PDDL problem against the true states, verifying accurate representation of object locations and relationships. To examine the flexibility of Image2PDDL, we used both images and text descriptions to represent goal states, analyzing the framework's ability to interpret both visual and textual inputs accurately. Tables 2, 3, and 4 provide detailed error counts across domains and difficulty levels, showcasing the method's performance across varying task complexities.

Syntax Correctness: Overall, Image2PDDL demonstrated strong syntax correctness across all domains, with zero syntax errors recorded across both input modalities—images for both initial and goal states, and images for the initial state paired with a text description for the goal state. The generated PDDL problems were consistently verified as syntactically valid by the Fast Downward planner, showing that the framework adheres effectively to PDDL syntax rules. Occasionally, the VLM returned PDDL problems with syntax highlighting, but these highlights were easily removed without affecting the problem structure. This consistency suggests that Image2PDDL reliably constructs grammatically correct and executable PDDL statements, regardless of input type or task complexity within each domain.

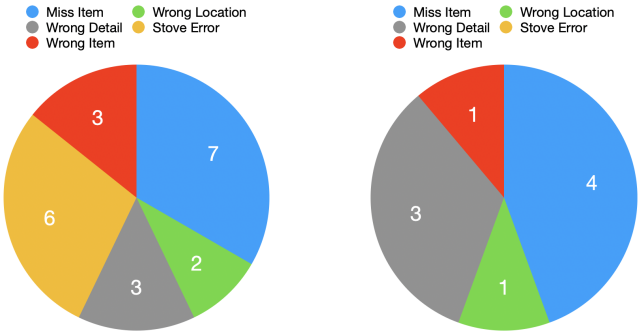

Figure 3: Types of Errors in the Kitchen Domain: The left chart represents errors when using an image to describe the goal state, while the right chart represents errors when using a text to describe the goal state.

Content Correctness: Our analysis shows that as domain complexity increases, Image2PDDL is more prone to generating incorrect states in PDDL problems. We reviewed errors within each domain to identify recurring patterns and domain-specific challenges.

Domain	Object Types	Predicates	Actions
Blocksworld	block	on, ontable, clear, holding, arm-empty	pick-up, put-down, stack, unstack
Sliding-Tile Puzzle	tile, position	tile, position, at, blank, inc, dec	move-down, move-up, move-right, move-left
Kitchen	item (fruit: lemon, apple; cutting_board; mug: black_mug, green_mug; kettle; wine; soda: cola, fanta), location (counter, sink, shelf, stove)	at, clear	move-item

Table 1: Defined Object Types, Predicates, and Actions in the Domain Descriptions

Input Type	Difficulty	Syntax Errors	Content Errors
Images (Initial & Goal)	Easy	0	1/50
	Medium	0	3/50
	Hard	0	3/50
Image (Initial), Text (Goal)	Easy	0	1/50
	Medium	0	3/50
	Hard	0	3/50

Table 2: Syntax and Content Errors in Blocksworld Domain

Input Type	Difficulty	Syntax Errors	Content Errors
Images (Initial & Goal)	Easy	0	2/50
	Medium	0	1/50
	Hard	0	1/50
Image (Initial), Text (Goal)	Easy	0	2/50
	Medium	0	1/50
	Hard	0	2/50

Table 3: Syntax and Content Errors in Sliding-Tile Puzzle Domain

In the Blocksworld domain, Image2PDDL consistently recognized the correct number of objects across both image and text-based goal states, demonstrating reliable object identification. However, the same errors appeared across scenarios, revealing a recurring issue in goal state interpretation. These errors typically involved incorrect stacking orders: blocks intended to be on the table were sometimes placed atop stacks, while blocks meant to be on top were placed on the table. This reversal in stacking sequences indicates a systematic misinterpretation of positional relationships in the generated PDDL problem.

In the Sliding-Tile Puzzle domain, errors primarily involved switched tile locations, where specific tiles were incorrectly swapped. Although Image2PDDL generally captured the spatial configuration well, it occasionally struggled with precise tile placement, particularly in more complex tile arrangements.

In the Kitchen domain, the visually intricate environment introduced additional sources of error. We classified these errors into five categories: (1) Stove Error, where the LLM mistakenly identifies the stove as an item instead of a location; (2) Missing Item, where certain items are omitted from the generated PDDL problem; (3) Wrong Item, where items are misclassified, leading to inaccuracies in object identification; (4) Wrong Detail, where items are assigned incorrect attributes (e.g., brand or color); and (5) Wrong Location, where items are placed in incorrect positions. Figure 3 presents the distribution of these error types. Across goal state representation formats, the most common error was missing items, typically smaller objects like fruit or soda cans. Additionally, using images to represent the goal state often led to stove misidentifications, where the stove was treated as an item rather than a location, a recurring issue when defining goal states.

Input Type	Difficulty	Syntax Errors	Content Errors
Images (Initial & Goal)	Easy	0	7/50
	Medium	0	5/50
	Hard	0	9/50
Image (Initial), Text (Goal)	Easy	0	3/50
	Medium	0	3/50
	Hard	0	3/50

Table 4: Syntax and Content Errors in Kitchen Domain

Image2PDDL Use Case

Robot-assisted teaching of students with Autism Spectrum Disorder Teachers utilize different types of prompts to support student learning. In robot-assisted education, robots can deliver physical, verbal, visual or gestural prompts, and can teach problem-solving through demonstration. Research indicates that students with Autism Spectrum Disorder (ASD) often respond positively to interactions with robots (van den Berk-Smeekens et al. 2020).

Robot-assisted teaching shows potential for significant enhancements in educational outcomes of students with ASD. Educators note that children with ASD may benefit from robot predictability and consistency, especially with humanoid robots. However, further research is needed to verify these observations. It is essential to identify specific use cases and conditions under which skills acquired during robot-assisted sessions can be effectively transferred to the child's daily life. Additionally, we should consider the effects of robot morphology on children's responses and behavior. Collaboration with ASD experts is crucial for developing systems that are both effective and appropriate for use in special education (Alcorn et al. 2019). Students with ASD often show a strong interest in technological devices, including robots.

The Codey Rocky robot is an educational tool designed to teach programming through block-based and text-based code. Studies involving this robot indicate that students with ASD are motivated to engage in programming activities, and that robot-mediated interventions can enhance peer cooperation (Gkiolnta, Zygopoulou, and Syriopoulou-Delli 2023). Robots can also support the development of communication skills and help reduce maladaptive behaviors in students with ASD (Dubois-Sage et al. 2024). Humanoid robots, such as NAO, have shown substantial benefits in teaching children with ASD and supporting the acquisition of new skills. Focus and attention span were improved in robot-assisted activities (Qidwai, Kashem, and Conor 2019).

Some researchers focus on challenges specific to individuals with ASD, such as joint attention and imitation skills, utilizing methods based on Computer-Assisted Therapies (CATs) and Robot-Assisted Therapies (RATs) to improve student skills (Zahid et al. 2024). However, some contradictory results were published on the effects of robots on joint attention (Sani-Bozkurt and Bozkus-Genc 2021) and the overall skill development. For instance, while some students found interactions with robots engaging, they did not demonstrate significant learning progress (Abu-Amara, Mohammad, and Bensefia 2024). Earlier studies have emphasized the use of video models to improve fine and gross motor task perfor-mance in individuals with ASD (Mechling and Swindle 2012). Mobile applications have also shown potential as effective tools for students with ASD; for example, a mathematical problem-solving app was well received by participants, who demonstrated the ability to complete tasks independently (Blanco 2024).

Recently, interventions based on virtual reality have been investigated to explore effects on enhancing social skills of children (Sun, Huang, and Ho 2024) and to support remote team collaboration of autistic adults (Amat 2023). In an industrial context, experiments examining the task of robotic collaborative assembly have implied that some persons with ASD can collaborate very effectively with the robot. However, behavioral differences in gaze, gestures, etc. suggest that solutions designed for neurotypical participants may not align with the needs of the ASD groups (Mondellini 2023).

TEACCH® Autism Program

The TEACCH® Autism Program was developed at the University of North Carolina (Us 2024). It was created in the 1970s by Eric Schopler who disproved the hypothesis that autism is a mental illness caused by emotionally cold parents. His experiments imply that autism is not primarily a disorder of emotions but a disorder of processing sensory information (Mesibov 2004).

The TEACCH approach exploits the premise that people with ASD are predominantly visual learners and prefer visually cued instructions in the educational process (Quill 1995). It's the recommended methodology in the Czech Republic for the education of people with ASD and is used worldwide. The TEACCH Structured Work Session consists of structured tasks. Educators often use a specific type of structured task, the shoe-box task (see Figure 4), to be solved by students in the class.

Automated assessment and planning of shoe-box tasks

There is a visual structure that allows to recognize how the shoe-box task should be completed. Vision-based machine learning methods seems to be suitable for exploring tasks and automatically evaluating them. A system that provides automatic assessment of structured tasks extended by actions planning could be used in structured teaching, with a real robot in the role of teacher. To the best of our knowledge, there is no existing research on automated assessment of shoe-box tasks, or on automated action planning for robot-assisted shoe-box activities.

For learning new shoe-box tasks, students typically follow teacher guidance, or use picture-based or text-based instructional systems (Bryan and Gast 2000). Additional experiments have explored alternative types of prompts, including video, computer, and other devices. For example, a Personal Digital Assistant was used to assist students in completing new tasks and to facilitate smoother transitions between tasks (Mechling and Savidge 2010).

Some research addresses action planning for structured tasks (not for shoe-box tasks), employing a mathematical framework with an agent in the role of a teacher. In experiments, a humanoid NAO robot was utilized in a storytelling scenario to guide the child's gaze toward target screens, using a sequence of planned actions (Baraka et al. 2020, 2022).

Experiments with real robots can be performed with the humanoid robots and robot manipulators. Depth cameras can be used as main sensors to interpret the environment.

Objective

The main goal of a use case for Image2PDDL is to implement a computerized system for the automatic assessment and planning of structured shoe-box tasks in robot-assisted education.

This system provides methods for creating models of the shoe-box tasks, automatically identifying the current state of a task instance and planning actions for it. It allows a robot to assist a student in completing the shoe-box task, which will be demonstrated by the real robot. The robot responds to the student's requests for help and provides instructions to complete the task. There are implemented multiple interaction strategies that represent different teaching approaches. Each strategy contains, for each task instance, a sequence of robot actions that lead to successful completion of the task, if the student follows robot instructions.

Methods of Research

The TEACCH Structured Work Session consists of *structured tasks*. The shoebox task is a structured task with a box that serves as a workstation and contains all the materials for the task (puzzles, cards, blocks, ...). There is a visual structure that hints how the task should be completed.

Structured task assessment are answers to the following questions: What is the goal? Was the goal achieved in this instance? What needs to be done to achieve the goal?

We can represent the structured task as abstract structure containing set of given elements with pre-defined states, ac-

Figure 4: Shoebox tasks as a use case for robot-assisted teaching.

tions and goal. For this structure, logic programming could be used to search for sequence of actions (plan) to reach the goal. Then, for a given task instance (i.e., configuration of elements in the box), the vision-based method Image2PDDL can be used to decide whether the task has been completed or what state it is in.

In the structured teaching sessions, there is typically a teacher that supervises a student while he is working on the task. System for automated task assessment could be used as a tool to provide basic feedback to a student, for example to confirm that the task is completed. The teacher also assists the student by giving prompts or by demonstration. This part of the teaching process can be performed by a real robot that would help the student to solve the task, using the action plan to reach the goal from current state.

Another approach might be to try to infer the abstract structure and to create a logic model automatically using machine learning methods, e.g., from demonstration (video) or from annotated samples of completed and incomplete task instances (images, simulation snapshots). Planning would use some predefined actions to manipulate objects.

Ideally, the system would not require any programming skills to be used: for a new task, the teacher creates the model just by demonstration of the task. Then, for individual students and task instances, s/he choose an educational strategy which would be performed by the robot.

Representation of shoe-box tasks Each task type is represented by its model that contains elements, locations, initial state, goal states and possible actions. This PDDL model can be provided with the shoe-box task, or created automatically using vision-based machine learning. For each task instance the Image2PDDL model then allows to identify automatically (from camera data) the current state of the instance and to plan actions for the current state.

Initial experiments We conducted initial tests with a small dataset of shoe-box tasks to verify that the Image2PDD model can be used to identify shoe-box task states and to plan actions.

For simple tasks such as distributing balls into labeled bowls, ChatGPT can easily derive the goal of the task and

Figure 5: Instance of the distribution task that requires single snapshot to obtain specification from ChatGPT.

Figure 6: Instance of the distribution task with unambiguous assignment: labels are treated as bowl names, not as number of balls.

plan actions. In some cases, a single snapshot is sufficient to obtain a full task description. However, we usually need to provide snapshots of multiple configurations to avoid ambiguous conclusions.

The Image2PDD framework provides tools for specifying a task using an image or text description. This allows us to work with more complex shoe-box tasks models and also makes the application more intuitive. In robot-assisted teaching, teachers should have a choice of how to work with the robots: some would prefer to use only images, others would be interested in advanced methods that allow to describe difficult tasks with many objects.

Robot as a teacher The robot will assist the student in completing the shoe-box task e.g., by manually moving elements, pointing, or giving verbal prompts. The student can ask the robot for help, e.g., by using a button, gesture, or voice commands.

There are many interaction strategies to decide what the robot should do and when. We will consult the strategies with experts in ASD and implement multiple recommended strategies, as options for the user (teacher). Planned actions will be translated into abstract robot actions using the selected strategy; abstract robot actions will be translated into commands for the real robot.

Collaboration with experts in ASD In the preparation phase, we will discuss basic practical questions related to structure teaching, e.g., which structured tasks types are currently used in education (in Czech Republic and in world), are there any standard for shoe-box tasks in Czech schools (for example, products from ShoeboxTasks®) or are they typically designed and made by teachers, what are physical parameters of the boxes.

We also need to consult educational strategies and methodology specific to ASD, especially the safety recom-

mendations for experiments with students and the real robot. For testing, annotated data from the multi-modal database of autistic children's interactions are available (DE-ENIGMA 2019).

Conclusion

In this work, we introduced Image2PDDL, a framework that leverages VLM to automate the generation of PDDL problems. By utilizing both images and textual descriptions for initial and goal states, Image2PDDL bridges the gap between visual perception and symbolic planning, allowing for a streamlined, accessible process of problem generation across diverse domains. Our evaluation across three domains—Blocksworld, Sliding-Tile Puzzle, and Kitchen—demonstrated that Image2PDDL achieves high syntax correctness, reliably producing grammatically valid PDDL outputs across various task complexities and input types. While content correctness varied depending on domain difficulty and input modality, results indicated that the framework performs well with structured tasks and benefits from textual input in visually complex scenarios.

Image2PDDL's promising performance suggests broader applications in AI planning, particularly for complex real-world tasks requiring visual and symbolic integration. Future work could enhance the framework's capacity to manage intricate object relationships and multi-step transformations in dynamic environments, such as the autistic shoebox task. Overall, this framework lays the groundwork for more accessible and scalable AI planning solutions, creating new opportunities for integrating vision and language models into automated planning domains.

We discussed an apparent application case for Image2PDDL in robot-assisted teaching of students with ASD. Some possible research avenues include

- automatically generating 3D models of shoebox tasks (to provide the visual feedback or to practice the task in a simulation).
- using vision language model to generate verbal instructions.
- automating the entire process: use the robot to prepare the shoebox task or the entire work session.
- using the system as tool for education of robot programming for students with ASD.

References

Abu-Amara, F.; Mohammad, H.; and Bensefia, A. 2024. Robot-based therapy for improving academic skills of children with autism. *International Journal of Information Technology*.

Alcorn, A. M.; Ainger, E.; Charisi, V.; Mantinioti, S.; Petrović, S.; Schadenberg, B. R.; Tavassoli, T.; and Pellicano, E. 2019. Educators' Views on Using Humanoid Robots With Autistic Learners in Special Education Settings in England. *Frontiers in Robotics and AI*, 6.

Amat, A. Z. 2023. Design of a Desktop Virtual Reality-Based Collaborative Activities Simulator (ViRCAS) to Support Teamwork in Workplace Settings for Autistic Adults. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31: 2184–2194.

Baraka, K.; Couto, M.; Melo, F. S.; Paiva, A.; and Veloso, M. 2022. Sequencing Matters: Investigating Suitable Action Sequences in Robot-Assisted Autism Therapy. *Frontiers in Robotics and AI*, 9.

Baraka, K.; Melo, F. S.; Couto, M.; and Veloso, M. 2020. Optimal action sequence generation for assistive agents in fixed horizon tasks. *Autonomous Agents and Multi-Agent Systems*, 34(2).

Blanco, R. 2024. Design of a Mathematical Problem-Solving Application for Students with Autism Spectrum Disorder. *Educational Technology & Society*, 27(2): 220–242.

Brohan, A.; Chebotar, Y.; Finn, C.; Hausman, K.; Herzog, A.; Ho, D.; Ibarz, J.; Irpan, A.; Jang, E.; Julian, R.; et al. 2023. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning*, 287–318. PMLR.

Bryan, L. C.; and Gast, D. L. 2000. Teaching On-Task and On-Schedule Behaviors to High-Functioning Children with Autism Via Picture Activity Schedules. *Journal of Autism and Developmental Disorders*, 30(6): 553–567.

Dagan, G.; Keller, F.; and Lascarides, A. 2023. Dynamic planning with a llm. *arXiv preprint arXiv:2308.06391*.

DE-ENIGMA. 2019. The DE-ENIGMA database. Available: https://de-enigma.eu/database.

Dubois-Sage, M.; Jacquet, B.; Jamet, F.; and Baratgin, J. 2024. People with Autism Spectrum Disorder Could Interact More Easily with a Robot than with a Human: Reasons and Limits. *Behavioral Sciences*, 14(2): 131.

Gkiolnta, E.; Zygopoulou, M.; and Syriopoulou-Delli, C. K. 2023. Robot programming for a child with autism spectrum disorder: a pilot study. *International Journal of Developmental Disabilities*, 69(3): 424–431.

Guan, L.; Valmeekam, K.; Sreedharan, S.; and Kambhampati, S. 2023. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems*, 36: 79081–79094.

Helmert, M. 2006. The fast downward planning system. *Journal of Artificial Intelligence Research*, 26: 191–246.

Huang, W.; Abbeel, P.; Pathak, D.; and Mordatch, I. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, 9118–9147. PMLR.

Liang, J.; Huang, W.; Xia, F.; Xu, P.; Hausman, K.; Ichter, B.; Florence, P.; and Zeng, A. 2023. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 9493–9500. IEEE.

Lin, K.; Agia, C.; Migimatsu, T.; Pavone, M.; and Bohg, J. 2023. Text2motion: From natural language instructions to feasible plans. *Autonomous Robots*, 47(8): 1345–1365.

Liu, B.; Jiang, Y.; Zhang, X.; Liu, Q.; Zhang, S.; Biswas, J.; and Stone, P. 2023. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*.

Mechling, L. C.; and Savidge, E. J. 2010. Using a Personal Digital Assistant to Increase Completion of Novel Tasks and Independent Transitioning by Students with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, 41(6): 687–704.

Mechling, L. C.; and Swindle, C. O. 2012. Fine and Gross Motor Task Performance When Using Computer-Based Video Models by Students With Autism and Moderate Intellectual Disability. *The Journal of Special Education*, 47(3): 135–147.

Mesibov, G. B. 2004. The Origins and History of the TEACCH Program. *Springer eBooks*, 1–12.

Mondellini, M. 2023. Behavioral patterns in robotic collaborative assembly: comparing neurotypical and Autism Spectrum Disorder participants. *Frontiers in Psychology*, 14.

Oswald, J.; Srinivas, K.; Kokel, H.; Lee, J.; Katz, M.; and Sohrabi, S. 2024. Large Language Models as Planning Domain Generators. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 34, 423–431.

Qidwai, U.; Kashem, S. B. A.; and Conor, O. 2019. Humanoid Robot as a Teacher's Assistant: Helping Children with Autism to Learn Social and Academic Skills. *Journal of Intelligent & Robotic Systems*.

Quill, K. A. 1995. Visually Cued Instruction for Children with Autism and Pervasive Developmental Disorders. *Focus on Autistic Behavior*, 10(3): 10–20.

Raman, S. S.; Cohen, V.; Rosen, E.; Idrees, I.; Paulius, D.; and Tellex, S. 2022. Planning with large language models via corrective re-prompting. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Sani-Bozkurt, S.; and Bozkus-Genc, G. 2021. Social Robots for Joint Attention Development in Autism Spectrum Disorder: A Systematic Review. *International Journal of Disability, Development and Education*, 1–19.

Shirai, K.; Beltran-Hernandez, C. C.; Hamaya, M.; Hashimoto, A.; Tanaka, S.; Kawaharazuka, K.; Tanaka, K.; Ushiku, Y.; and Mori, S. 2024. Vision-language interpreter for robot task planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2051–2058. IEEE.

Singh, I.; Blukis, V.; Mousavian, A.; Goyal, A.; Xu, D.; Tremblay, J.; Fox, D.; Thomason, J.; and Garg, A. 2023. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 11523–11530. IEEE.

Smirnov, P.; Joublin, F.; Ceravola, A.; and Gienger, M. 2024. Generating consistent PDDL domains with Large Language Models. *arXiv preprint arXiv:2404.07751*.

Sun, S.-J.; Huang, A.-C.; and Ho, W.-S. 2024. Enhancing Social Skills in Autism Students with Augmented-Reality Picturebooks. *Applied Sciences*, 14(11): 4907–4907.

Us, A. 2024. TEACCH® Autism Program.

van den Berk-Smeekens, I.; van Dongen-Boomsma, M.; De Korte, M. W. P.; Den Boer, J. C.; Oosterling, I. J.; Peters-Scheffer, N. C.; Buitelaar, J. K.; Barakova, E. I.; Lourens, T.; Staal, W. G.; and Glennon, J. C. 2020. Adherence and acceptability of a robot-assisted Pivotal Response Treatment protocol for children with autism spectrum disorder. *Scientific Reports*, 10(1).

Xie, Y.; Yu, C.; Zhu, T.; Bai, J.; Gong, Z.; and Soh, H. 2023. Translating natural language to planning goals with large-language models. *arXiv preprint arXiv:2302.05128*.

Zahid, Z.; S. Ali, S. S.; Y. Ayaz, N. N.; and Yaseen, I. 2024. RoboCA3T: A Robot Inspired - Computer Assisted adaptive autism therapy for improving joint attention and imitation skills through - learning and computing innovations. *Journal of Computer Assisted Learning*, 40(5): 2031–2048.

Zhang, X.; Ding, Y.; Amiri, S.; Yang, H.; Kaminski, A.; Esselink, C.; and Zhang, S. 2023. Grounding classical task planners via vision-language models. *arXiv preprint arXiv:2304.08587*.