Learning, Reasoning and Planning with Neuro-Symbolic Concepts

Jiayuan Mao MIT CSAIL jiayuanm@mit.edu

I aim to build machines that can continually learn new knowledge from their experiences and reason across tasks, modalities, and environments: answer queries, infer human intentions, and make long-horizon plans spanning hours to days. While recent advances in Internet-scale data collection and deep learning tools have made significant progress in many practical applications such as visual recognition, chatbots, and game agents, extending this success to general-purpose agents remains hard. First, collecting data for long-horizon interactions with highly variable objects and environments is prohibitively expensive; second, such agents must be able to continuously acquire new concepts while being able to generalize out-of-distribution to novel states, goals, and preferences — a challenge beyond the reach of current end-to-end neural network systems.

Building upon large-scale data and models, and integrating insights from theories of computation and studies of human cognition, I focus on building intelligent agents that can learn as efficiently and generalize as reliably as humans do. Concretely, these agents should continuously adapt to new concepts (words, objects, skills) quickly, on the order of 1 to 10 examples, and generalize rapidly to novel states, goals, and even embodiments. To achieve this, I work on **understanding and leveraging the principles of representation and computation for reasoning**, to design systems that are efficient during both training and inference and have strong generalization.

In my research, I use robots as the primary testbed. My key insight into robot decision-making in the physical world is the importance of **abstraction** in representation, and **compositionality** in learning and inference. My theoretical work has shown that by introducing spatial and temporal factorizations that break down scenes into entities and trajectories into subgoals, we can obtain provable polynomial circuit complexity bounds for environmental transition rules learning [1, 2, 3], policies learning [4], and planning [5].

Drawing inspiration from these theoretical evidences and also studies in cognitive sciences, my work proposed **neuro-symbolic concept** representations as a **composable abstraction** of the physical world. Concepts, as understood by cognitive scientists and philosophers, are the primitive building blocks of thought, which can be composed to form sophisticated compound thoughts: beliefs, intentions, and plans. Technically, a neuro-symbolic concept is a unit of knowledge with a name, a type declaration, and an implementation. Illustrated in Fig. 1, the color "*orange*" is an object-property concept whose type is a mapping from objects to Boolean values. Its implementation contains a neural network that recognizes the color based on visual inputs. Action concepts such as "*put-left*" are annotated with argument types and additionally, preconditions and postconditions — enabling **automatic** composition of them given novel goals. Their implementations contain neural network-based sensorimotor policies and generative models for contacts, trajectories, and forces. Leveraging the typing, preconditions, and postconditions, these concepts are **combinatorially re-usable** via general-purpose inference and planning mechanisms. Having learned a finite set of concepts, the agents can be recombined with an infinite variety, to automatically solve novel problems that may appear completely dissimilar to problems encountered during training. Fig. 1 illustrates how various concepts can be composed to build a computation graph that generates robot control commands from sensory inputs based on a natural language instruction. The graph is constructed by connecting the inputs and outputs of primitive neural network modules, making it end-to-end differentiable. This allows for gradient-based optimization, enabling the training of all components using diverse data sources — either through fine-grained annotations of primitive concepts or input-output pixel-to-torque trajectories. Compared to alternative end-to-end approaches, this system is more data-efficient: by decomposing the learning problem into the learning of individual concepts; it performs inference faster using symbolic reasoning tools and generalizes compositionally to unseen states and instructions.

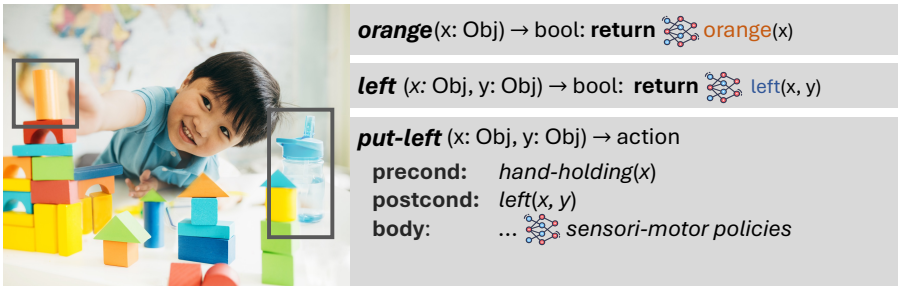

Figure 1: (Left) Illustration of three types of neuro-symbolic concepts: object properties (e.g., "*orange*"), relations (e.g., "*left*"), and actions (e.g. "*put-left*"). (Right) An *end-to-end differentiable* computation graph that composes concepts.

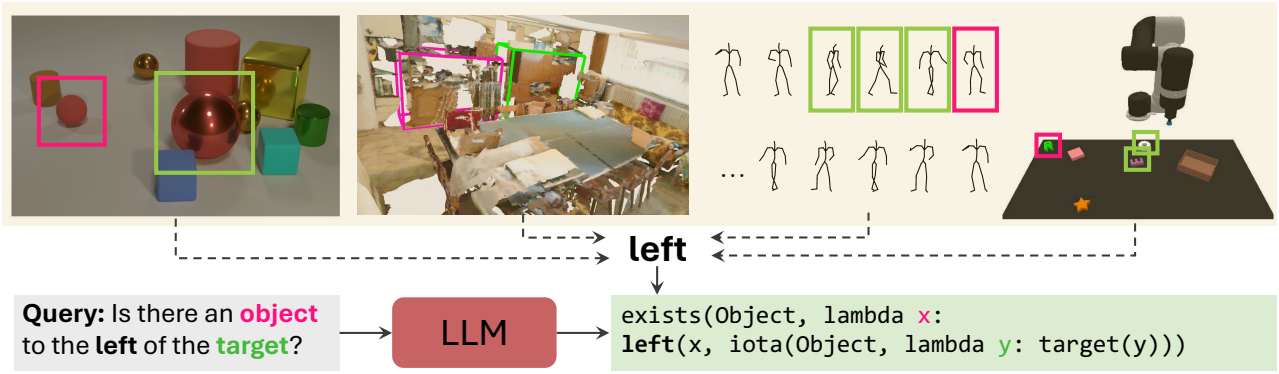

Figure 2: LEFT [6] is a unified concept learning and reasoning framework that grounds modular concepts across domains, and flexibly reasons with concepts across tasks with a foundation model.

Recognizing and Reasoning about Visual Concepts

In 2018, I started working on neuro-symbolic concept learning by building systems that learn to recognize concepts from visual data streams (e.g., images, videos, point clouds, and motion trajectories) for visual reasoning purposes. The Neuro-Symbolic Concept Learner [NS-CL; 7] is such a framework that learns the grounding of visual concepts and semantic parsing of sentences without any explicit supervision; instead, it learns by simply looking at images and reading paired questions and answers. NS-CL builds an object-based scene representation and translates sentences into executable, symbolic programs. It uses a neuro-symbolic reasoning module that executes these programs on the latent scene representation. Like human concept learning, **the perception module learns visual concepts based on the language description of the objects being referred to**. Meanwhile, **the learned visual concepts facilitate learning new words and parsing new sentences.** The NS-CL framework is very data-efficient. Using only 10% of the training data, it achieves more than 90% of the accuracy on CLEVR, a standard visual reasoning benchmark [8], surpassing all end-to-end and hybrid methods by at least 20%. This trend of data efficiency has continued to apply when my co-authors and other researchers extend the system to a broad set of domains: image-caption retrieval [9], metaconcept learning (i.e., relational concepts about concepts) [10], video and counterfactual reasoning [11], 3D scene understanding [12], and reasoning about human motion captures [6]. Joy Hsu, a Ph.D. student at Stanford whom I co-mentored, and I recently proposed logic-enhanced foundation models [6] that can unify visual reasoning across different modalities and domains (Fig. 2), by integrating NS-CL and large language models (LLMs).

Another crucial advantage of the NS-CL framework is its compositional generalization. The learned visual concepts can be recombined to answer more complex questions about scenes involving more objects. It also enables zero-shot transfer of learned concepts across tasks and even domains, such as transferring learned object concepts from the task of image captioning ("an apple") to visual question answering ("how many apples are there?"), and to robotic manipulation ("push the apples").

Connections to (human) language acquisition. The broad idea of joint vision-language learning also opens up new opportunities for visually grounded language acquisition, including learning syntax [13] and joint syntax and meaning representations [14]. In the past, I have studied how constituency structures in language can emerge in the process of learning to align images and captions. Moving towards extremely efficient visual concept learning by leveraging the syntax-semantics correspondences of words, I built systems that can infer the syntax and meaning of novel words that are grounded in vision from just a single example [14, 15], suggesting a promising approach for online concept adaptation and human language acquisition modeling.

Planning for Actions to Realize Concepts in the Physical World

In robotic domains, we want not only to recognize object properties, relations, and layouts but also want to actively produce action sequences to achieve them. This is difficult due to the fundamental interactions between geometric (e.g., reachability and collision), physical (e.g., friction and stability), and other task-related constraints (e.g., human preference). This creates a highly variable set of situations due to the combinatorial nature of objects, states, and goals.

Drawing insights from robotic mechanics, I propose to **learn factorized generative models that can generate control parameters** (e.g., poses, forces, and trajectories) at primitive levels, and **produce long-horizon plans with general-purpose planning and inference algorithms** that find control parameters satisfying all constraints of robots, task goals, and user preferences. I develop algorithms that deeply integrate model and

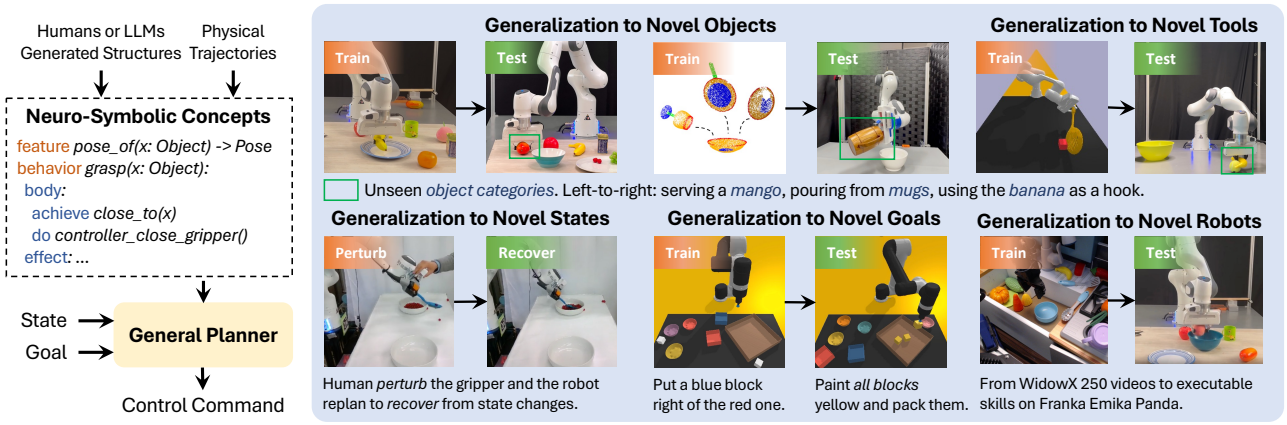

Figure 3: (Left) My neuro-symbolic framework learns various components of a world model and solves novel tasks using a general-purpose planner. (Right) It can achieve various types and levels of generalizations.

policy learning with online planning. I build novel object-centric representations for geometric, physical, and semantic constraints, and develop novel domain-independent planning and inference algorithms [16, 5].

Primitive manipulation concepts. My past research has studied learning generative models of actions grounded on relations among object parts. Working with Weiyu Liu, a Stanford Postdoc that I co-mentored, and other colleagues, I build generative models for goal-achieving trajectories (e.g., pouring, lifting handles, opening doors and drawers) that are subject to various geometric and physical constraints (e.g., kinematic joints, stability, and alignments) that are defined over features among different object parts (e.g., pouring requires the alignment between two rims) [17, 18, 19]. Learning such factorized models is significantly more efficient than learning the entire sensorimotor policy (reducing the required amount of training examples from hundreds to 10). Importantly, factorized models also show stronger generalization to novel object poses, instances, and even categories (e.g., generalizing from pouring from glasses to saucepans) and unseen backgrounds.

Causal action models for planning. With the goal of composing primitive concepts to form long-horizon plans, I worked on learning **causal action models** that support combination in novel situations. On the representational language side, I have developed PDSketch [16] and its successor CDL [5], which are interpretable representation languages that can be produced by humans or machine models, such as large language models (LLMs). In CDL, a state is described by a set of objects and their relational features (e.g., poses and contacts). These features can Boolean values or neural network features. A causal action model describes the set of constraints over states and actions and the effect of a behavior, similar to the pre- and post-conditions of a computer program. CDL is a novel representation unifying both imperative strategies ("policies") that can be followed step-by-step and declarative models for planning. They can be combined seamlessly to solve a single problem. All parts of these representations, including the programmatic structures and neural networks predicting state features, action effects, and trajectories, can be learned from data. At inference time, a domain-independent planner takes the current state and the task goal as input and produces robot control commands.

CDL provides a flexible interface for us to model the sparsity and locality nature of physical interactions. As a concrete example, many manipulation skills can be decomposed into a sequence of subgoals about object interactions. Building on these principles, illustrated in Fig. 3, I have built systems that can acquire manipulation skills such as pushing thin objects to edges for grasping, using hooks to reach distant objects, and hanging various objects on rods, from single demonstrations and some practice [20, 21]. Our key insight lies in interpreting these skills as a sequence of robot-object and object-object contact modes, which provides a natural scaffold for learning generative models for continuous parameters. These learned strategies and models can be seamlessly integrated into the CDL framework to enable their compositional use with other skills. My past work also includes systems that can generalize to unseen combinations of skills and objects [22], systems that are robust to motion and task-level perturbations [18], and systems that can make and execute long-horizon plans in novel states for novel goals [16, 23, 24, 25]. The principles of building composable abstractions around entity and temporal structures extend beyond robotics to general decision-making, such as digital agents capable of performing everyday tasks on computers by cumulatively learning reusable subroutines from experience [26].

Future Research: General-Purpose, Theoretically-Grounded, and Human-Centered Intelligence

Structured, continuously learning foundation models for robotics. The success of integrating general planning and inference algorithms with generative models for continuous parameters in robotics has shown

promise in few-shot learning and robust generalization. This suggests a path forward for scalable and general-purpose robots. Building on my previous work, I envision a "compositional foundation model" for robotics that broadly covers a collection of primitive manipulation concepts and causal action models for everyday and industrial tasks. This approach is totally compatible and complementary with existing efforts on large-scale data-collection and pre-training — they can be used as the data and building blocks to construct the composable representations. More importantly, the adaptation of training "compositional foundation models" would help us to shift our attention from data quantity to the quality and diversity of tasks covered. Additionally, it also enables us to learn physical knowledge from a wider array of data streams other than robot trajectories, from language descriptions of task workflows [27], to image-only annotations for spatial constraints, to human-object interaction videos [28]. This paradigm offers a flexible and scalable foundation for robot learning.

A current limitation of many robot learning systems, including mine, is their reliance on human demonstrations or language-based knowledge. The concept-centric representations provide a basis for more directed and effective exploration that would enable robots to explore and solve new tasks autonomously. Starting with a basic set of human-taught concepts, these systems would use planning and optimization to discover efficient strategies (e.g., tool use) and thus form new action concepts. This approach would also allow new relational concepts about object states to emerge, defining the preconditions under which these strategies can be applied. Ultimately, this would enable robots to continually learn, adapt, and self-improve in the environment.

From physical intelligence to human-centered intelligence. In my previous research, I have focused on building systems capable of robust reasoning and decision-making under explicit human instructions, typically without engaging in social interactions with humans. Moving forward, I aim to extend this to robots that can communicate and collaborate with people. A critical direction for my future work is building robotic systems capable of inferring human beliefs, preferences, and intentions to enable effective collaboration. Unlike purely instruction-following systems, robots need to infer implicit human goals from ambiguous cues and engage in proactive dialogue to resolve uncertainties. Some of my previous work has laid the groundwork for interpreting human intentions in social contexts by collecting new datasets (e.g., are they helping each other?" [29], can you hand that to me?"[30], or "set the dining table for two"[27]). I plan to develop methods for solving these challenges leveraging insights from computation and cognition.

Another future direction that I am excited about is to develop more efficient and intuitive human-computer interfaces leveraging neuro-symbolic representations. Fro example, my prior work on program-based representations of images and videos [31, 32, 33] can serve as inference algorithms in creating interfaces that users to manipulate real images and videos by editing code.

The engineering science of machine intelligence. I plan to study to the "engineering science of intelligence," in particular the hardness of practical reasoning and decision-making problems (e.g., complexity) and the theoretical trade-offs between learning and inference (e.g., policy learning vs. planning), and between structural biases and efficiency. My past work has made several contributions to theoretical understanding: circuit complexity upper bounds for policies in discrete decision-making problems [4], the expressivity and generalization bounds for relational neural networks (such as transformers) [1, 2], as well as the sample complexity and identifiability of Transformers [3] but they are far from explaining the true human-level intelligence and learning. In the short term, I hope to extend these results to continuous state and action spaces, and understand better the learning dynamics and complexity. These theories will have practical consequences in helping us make design choices about representations and algorithms.

Reverse engineering human intelligence. I have been working at the boundary of AI and cognitive science, and I plan to deepen this interdisciplinary approach in the future by applying new computational tools to advance cognitive science. In particular, the neuro-symbolic techniques I have developed can serve as, first, analytic tools for modeling data distributions in a more interpretable manner than purely neural network-based approaches. For instance, examining changes in logographic writing systems using a symbolic program learning approach has allowed us to quantify simplification trends over time [34]. Second, neuro-symbolic models can serve as alternative computational models for human cognition. I am particularly interested in understanding human causal reasoning, language acquisition and learning, and human learning and use of tools, possibly building on top of my past work on related datasets and machine reasoning algorithms [35, 36, 20].

Summary. I plan to continue to design methods for representation and reasoning, understand their theoretical properties, and use them in real robot systems, with the goal of understanding and building systems that can learn as flexibly and generalize as robustly as humans.

References

[1] Honghua Dong*, **Jiayuan Mao***, Tian Lin, Chong Wang, Lihong Li, and Denny Zhou. Neural Logic Machines. In *ICLR*, 2019. 1, 4

[2] Zhezheng Luo, **Jiayuan Mao**, Joshua B. Tenenbaum, and Leslie Pack Kaelbling. On the Expressiveness and Generalization of Hypergraph Neural Networks. In *LoG*, 2022. 1, 4

[3] Morris Yau, Eykin Akyurek, **Jiayuan Mao**, Joshua B. Tenenbaum, Stefanie Jegelka, and Jacob Andreas. Learning Linear Attention in Polynomial Time. *arXiv preprint arXiv:2410.10101*, 2024. 1, 4

[4] **Jiayuan Mao**, Tomás Lozano-Pérez, Joshua B. Tenenbaum, and Leslie Pack Kaelbling. What Planning Problem Can A Relational Neural Network Solve. In *NeurIPS*, 2023. 1, 4

[5] **Jiayuan Mao**, Joshua B. Tenenbaum, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Hybrid Declarative-Imperative Representations for Hybrid Discrete-Continuous Decision-Making. In *WAFR*, 2024. 1, 3

[6] Joy Hsu*, **Jiayuan Mao***, Joshua B. Tenenbaum, and Jiajun Wu. What's Left? Concept Grounding with Logic-Enhanced Foundation Models. In *NeurIPS*, 2023. 2

[7] **Jiayuan Mao**, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *ICLR*, 2019. 2

[8] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *CVPR*, 2017. 2

[9] Hao Wu*, **Jiayuan Mao***, Yufeng Zhang, Weiwei Sun, Yuning Jiang, Lei Li, and Wei-Ying Ma. Unified Visual-Semantic Embeddings: Bridging Vision and Language with Structured Meaning Representations. In *CVPR*, 2019. 2

[10] Chi Han*, **Jiayuan Mao***, Chuang Gan, Joshua B. Tenenbaum, and Jiajun Wu. Visual Concept-Metaconcept Learning. In *NeurIPS*, 2019. 2

[11] Zhenfang Chen, **Jiayuan Mao**, Jiajun Wu, Kwan-Yee K. Wong, Joshua B. Tenenbaum, and Chuang Gan. Grounding Physical Concepts of Objects and Events Through Dynamic Visual Reasoning. In *ICLR*, 2021. 2

[12] Joy Hsu, **Jiayuan Mao**, and Jiajun Wu. NS3D: Neuro-Symbolic Grounding of 3D Objects and Relations. In *CVPR*, 2023. 2

[13] Haoyue Shi*, **Jiayuan Mao***, Kevin Gimpel, and Karen Livescu. Visually Grounded Neural Syntax Acquisition. In *ACL*, 2019. 2

[14] **Jiayuan Mao**, Haoyue Shi, Jiajun Wu, Roger P. Levy, and Joshua B. Tenenbaum. Grammar-Based Grounded Lexicon Learning. In *NeurIPS*, 2021. 2

[15] Lingjie Mei*, **Jiayuan Mao***, Ziqi Wang, Chuang Gan, and Joshua B. Tenenbaum. FALCON: Fast Visual Concept Learning by Integrating Images, Linguistic Descriptions, and Conceptual Relations. In *ICLR*, 2022. 2

[16] **Jiayuan Mao**, Tomás Lozano-Pérez, Joshua B. Tenenbaum, and Leslie Pack Kaelbling. PDSketch: Integrated Domain Programming, Learning, and Planning. In *NeurIPS*, 2022. 3

[17] Weiyu Liu, **Jiayuan Mao**, Joy Hsu, Tucker Hermans, Animesh Garg, and Jiajun Wu. Composable Part-Based Manipulation. In *CoRL*, 2023. 3

[18] Yanwei Wang, Tsun-Hsuan Wang, **Jiayuan Mao**, Michael Hagenow, and Julie Shah. Grounding Language Plans in Demonstrations through Counter-factual Perturbations. In *ICLR*, 2024. 3

[19] Xiaolin Fang*, Bo-Ruei Huang*, **Jiayuan Mao***, Jasmine Shone, Joshua B. Tenenbaum, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Keypoint Abstraction using Large Models for Object-Relative Imitation Learning. In *CoRL Workshop on Language and Robot Learning: Language as an Interface*, 2024. 3

[20] **Jiayuan Mao**, Tomás Lozano-Pérez, Joshua B. Tenenbaum, and Leslie Pack Kaelbling. Learning Reusable Manipulation Strategies. In *CoRL*, 2023. 3, 4

[21] Yuyao Liu*, **Jiayuan Mao***, Joshua Tenenbaum, Tomás Lozano-Pérez, and Leslie Kaelbling. One-Shot Manipulation Strategy Learning by Making Contact Analogies. In *CoRL Workshop on Learning Effective Abstractions for Planning*, 2024. 3

[22] Renhao Wang*, **Jiayuan Mao***, Joy Hsu, Hang Zhao, Jiajun Wu, and Yang Gao. Programmatically Grounded, Compositionally Generalizable Robotic Manipulation. In *ICLR*, 2023. 3

[23] Lio Wong*, **Jiayuan Mao***, Pratyusha Sharma*, Zachary S. Siegel, Jiahai Feng, Noa Korneev, Joshua B. Tenenbaum, and Jacob Andreas. Learning Adaptive Planning Representations with Natural Language Guidance. In *ICLR*, 2024. 3

[24] Weiyu Liu*, Geng Chen*, Joy Hsu, **Jiayuan Mao**†, and Jiajun Wu†. Learning Planning Abstractions from Language. In *ICLR*, 2024. 3

[25] Weiyu Liu*, Neil Nie*, Ruohan Zhang, **Jiayuan Mao**†, and Jiajun Wu†. BLADE: Learning Compositional Behaviors from Demonstration and Language. In *CoRL*, 2024. 3

[26] Zora Zhiruo Wang, **Jiayuan Mao**, Daniel Fried, and Graham Neubig. Agent Workflow Memory. *ArXiv*, 2024. 3

[27] Yiqing Xu, **Jiayuan Mao**, Yilun Du, Tomás Lozano-Pérez, Leslie Pack Kaelbling, and David Hsu. Set It Up!: Functional Object Arrangement with Compositional Generative Models. In *RSS*, 2024. 4

[28] Po-Chen Ko, **Jiayuan Mao**, Yilun Du, Shao-Hua Sun, and Joshua B. Tenenbaum. Learning to Act from Actionless Video through Dense Correspondences. In *ICLR*, 2024. 4

[29] Joy Hsu, **Jiayuan Mao**, Joshua B. Tenenbaum, Noah D. Goodman, and Jiajun Wu. What Makes a Maze Look Like a Maze? In *ECCV (Human-Inspired Computer Vision Workshop)*, 2024. 4

[30] Yanming Wan*, **Jiayuan Mao***, and Joshua B. Tenenbaum. HandMeThat: Human-Robot Communication in Physical and Social Environments. In *NeurIPS*, 2022. 4

[31] **Jiayuan Mao***, Xiuming Zhang*, Yikai Li, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. Program-Guided Image Manipulators. In *ICCV*, 2019. 4

[32] Yikai Li*, **Jiayuan Mao***, Xiuming Zhang, William T. Freeman, Joshua B. Tenenbaum, Noah Snavely, and Jiajun Wu. Multi-Plane Program Induction with 3D Box Priors. In *NeurIPS*, 2020. 4

[33] Sumith Kulal*, **Jiayuan Mao***, Alex Aiken, and Jiajun Wu. Hierarchical Motion Understanding via Motion Programs. In *CVPR*, 2021. 4

[34] Guangyuan Jiang, Matthias Hofer, **Jiayuan Mao**, Lio Wong, Joshua B. Tenenbaum, and Roger P. Levy. Finding Structure in Logographic Writing with Library Learning. In *CogSci*, 2024. 4

[35] **Jiayuan Mao***, Xuelin Yang*, Xikun Zhang, Noah D. Goodman, and Jiajun Wu. CLEVRER-Humans: Describing Physical and Causal Events the Human Way. In *NeurIPS*, 2022. 4

[36] Ruocheng Wang*, **Jiayuan Mao***, Samuel J. Gershman†, and Jiajun Wu†. Language-Mediated, Object-Centric Representation Learning. In *ACL (Findings)*, 2021. 4