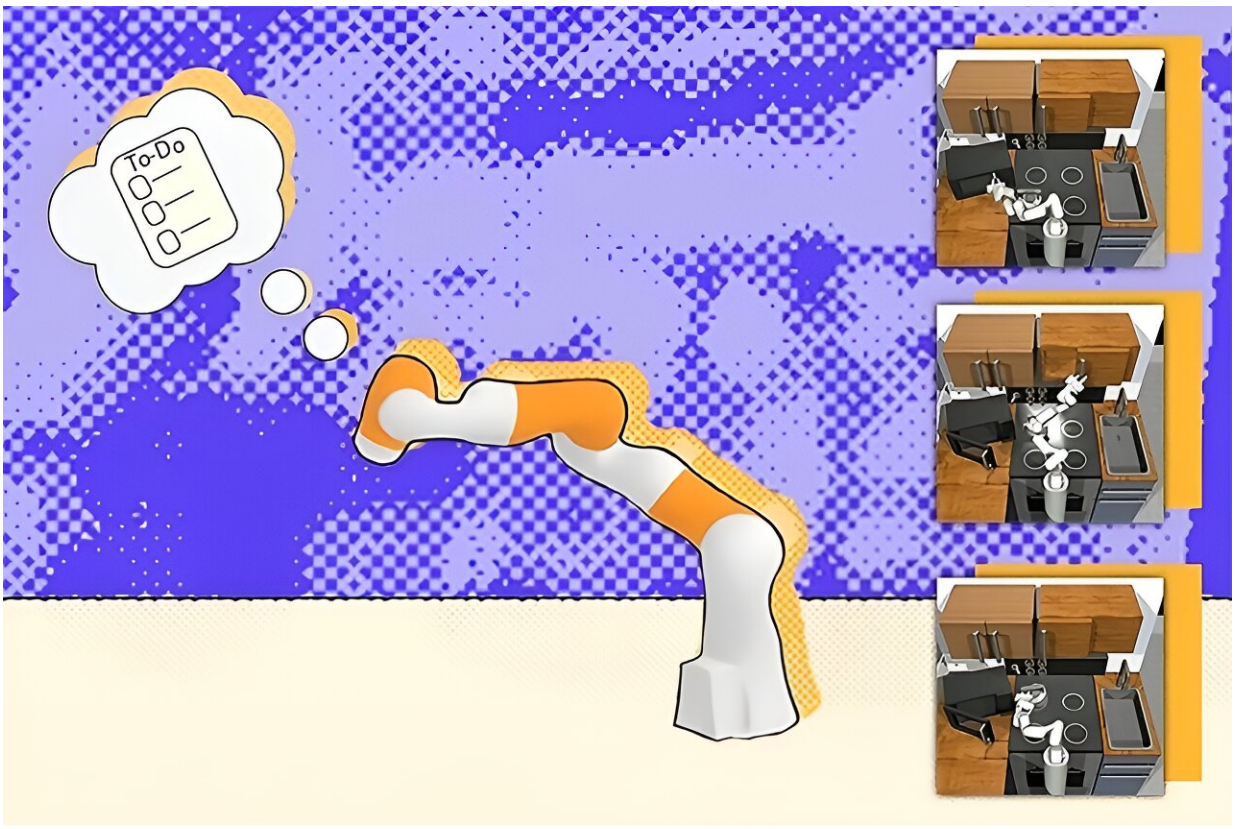
Multiple AI models help robots execute complex plans more transparently

January 8 2024, by Alex Shipps



The HiP framework developed at MIT CSAIL develops detailed plans for robots using the expertise of three different foundation models, helping it execute tasks in households, factories, and construction that require multiple steps. Credit: Alex Shipps/MIT CSAIL

Your daily to-do list is likely pretty straightforward: wash the dishes, buy

groceries, and other minutiae. It's unlikely you wrote out "pick up the first dirty dish," or "wash that plate with a sponge," because each of these miniature steps within the chore feels intuitive. While we can routinely complete each step without much thought, a robot requires a complex plan that involves more detailed outlines.

MIT's Improbable AI Lab, a group within the Computer Science and Artificial Intelligence Laboratory (CSAIL), has offered these machines a helping hand with a new multimodal framework: Compositional Foundation Models for Hierarchical Planning (HiP), which develops detailed, feasible plans with the expertise of three different foundation models. Like OpenAI's GPT-4, the foundation model upon which ChatGPT and Bing Chat were built, these foundation models are trained on massive quantities of data for applications like generating images, translating text, and robotics.

The work is published on the *arXiv* preprint server.

Unlike RT2 and other multimodal models that are trained on paired vision, language, and action data, HiP uses three different foundation models each trained on different data modalities. Each foundation model captures a different part of the decision-making process and then works together when it's time to make decisions. HiP removes the need for access to paired vision, language, and action data, which is difficult to obtain. HiP also makes the reasoning process more transparent.

What's considered a daily chore for a human can be a robot's "long-horizon goal"—an overarching objective that involves completing many smaller steps first—requiring sufficient data to plan, understand, and execute objectives. While computer vision researchers have attempted to build monolithic foundation models for this problem, pairing language, visual, and action data is expensive. Instead, HiP represents a different, multimodal recipe: a trio that cheaply incorporates linguistic, physical,

and environmental intelligence into a robot.

"Foundation models do not have to be monolithic," says NVIDIA AI researcher Jim Fan, who was not involved in the paper. "This work decomposes the complex task of embodied agent planning into three constituent models: a language reasoner, a visual world model, and an action planner. It makes a difficult decision-making problem more tractable and transparent."

The team believes that their system could help these machines accomplish household chores, such as putting away a book or placing a bowl in the dishwasher. Additionally, HiP could assist with multistep construction and manufacturing tasks, like stacking and placing different materials in specific sequences.

Evaluating HiP

The CSAIL team tested HiP's acuity on three manipulation tasks, outperforming comparable frameworks. The system reasoned by developing intelligent plans that adapt to new information.

First, the researchers requested that it stack different-colored blocks on each other and then place others nearby. The catch: Some of the correct colors weren't present, so the robot had to place white blocks in a color bowl to paint them. HiP often adjusted to these changes accurately, especially compared to state-of-the-art task planning systems like Transformer BC and Action Diffuser, by adjusting its plans to stack and place each square as needed.

Another test: arranging objects such as candy and a hammer in a brown box while ignoring other items. Some of the objects it needed to move were dirty, so HiP adjusted its plans to place them in a cleaning box, and then into the brown container. In a third demonstration, the bot was able

to ignore unnecessary objects to complete kitchen sub-goals such as opening a microwave, clearing a kettle out of the way, and turning on a light. Some of the prompted steps had already been completed, so the robot adapted by skipping those directions.

A three-pronged hierarchy

HiP's three-pronged planning process operates as a hierarchy, with the ability to pre-train each of its components on different sets of data, including information outside of robotics. At the bottom of that order is a large language model (LLM), which starts to ideate by capturing all the symbolic information needed and developing an abstract task plan. Applying the common sense knowledge it finds on the internet, the model breaks its objective into sub-goals. For example, "making a cup of tea" turns into "filling a pot with water," "boiling the pot," and the subsequent actions required.

"All we want to do is take existing pre-trained models and have them successfully interface with each other," says Anurag Ajay, a Ph.D. student in the MIT Department of Electrical Engineering and Computer Science (EECS) and a CSAIL affiliate. "Instead of pushing for one model to do everything, we combine multiple ones that leverage different modalities of internet data. When used in tandem, they help with robotic decision-making and can potentially aid with tasks in homes, factories, and construction sites."

These models also need some form of "eyes" to understand the environment in which they're operating and correctly execute each sub-goal. The team used a large video diffusion model to augment the initial planning completed by the LLM, which collects geometric and physical information about the world from footage on the internet. In turn, the video model generates an observation trajectory plan, refining the LLM's outline to incorporate new physical knowledge.

This process, known as iterative refinement, allows HiP to reason about its ideas, taking in feedback at each stage to generate a more practical outline. The flow of feedback is similar to writing an article, where an author may send their draft to an editor, and with revisions incorporated, the publisher reviews for any last changes and finalizes.

In this case, the top of the hierarchy is an egocentric action model, or a sequence of first-person images that infer which actions should take place based on its surroundings. During this stage, the observation plan from the video model is mapped over the space visible to the robot, helping the machine decide how to execute each task within the long-horizon goal. If a robot uses HiP to make tea, this means it will have mapped out exactly where the pot, sink, and other key visual elements are, and begin completing each sub-goal.

Still, the multimodal work is limited by the lack of high-quality video foundation models. Once available, they could interface with HiP's small-scale video models to further enhance visual sequence prediction and robot action generation. A higher-quality version would also reduce the current data requirements of the video models.

That being said, the CSAIL team's approach only used a tiny bit of data overall. Moreover, HiP was cheap to train and demonstrated the potential of using readily available foundation models to complete long-horizon tasks.

"What Anurag has demonstrated is proof-of-concept of how we can take models trained on separate tasks and data modalities and combine them into models for robotic planning. In the future, HiP could be augmented with pre-trained models that can process touch and sound to make better plans," says senior author Pulkit Agrawal, MIT assistant professor in EECS and director of the Improbable AI Lab. The group is also considering applying HiP to solving real-world long-horizon tasks in

robotics.

Ajay and Agrawal are lead authors on a paper describing the work. They are joined by MIT professors and CSAIL principal investigators Tommi Jaakkola, Joshua Tenenbaum, and Leslie Pack Kaelbling; CSAIL research affiliate and MIT-IBM AI Lab research manager Akash Srivastava; graduate students Seungwook Han and Yilun Du; former postdoc Abhishek Gupta, who is now assistant professor at University of Washington; and former graduate student Shuang Li, Ph.D.

More information: Anurag Ajay et al, Compositional Foundation Models for Hierarchical Planning, *arXiv* (2023). DOI: 10.48550/arxiv.2309.08587

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology