HELIX: A VISION–LANGUAGE–ACTION MODEL FOR GENERALIST HUMANOID CONTROL

February 20, 2025

Introducing Helix

We're introducing Helix, a generalist Vision-Language-Action (VLA) model that unifies perception, language understanding, and learned control to overcome multiple longstanding challenges in robotics. Helix is a series of firsts:

- **Full-upper-body control**: Helix is the first VLA to output high-rate continuous control of the entire humanoid upper body, including wrists, torso, head,

and individual fingers.

- **Multi-robot collaboration:** Helix is the first VLA to operate simultaneously on two robots, enabling them to solve a shared, long-horizon manipulation task with items they have never seen before.

- **Pick up anything:** Figure robots equipped with Helix can now pick up virtually any small household object, including thousands of items they have never encountered before, simply by following natural language prompts.

- **One neural network:** Unlike prior approaches, Helix uses a single set of neural network weights to learn all behaviors—picking and placing items, using drawers and refrigerators, and cross-robot interaction—without any task-specific fine-tuning.

- **Commercial-ready:** Helix is the first VLA that runs entirely onboard embedded low-power-consumption GPUs, making it immediately ready for commercial deployment.

0:00

Video 1: Collaborative grocery storage. A single set of Helix neural network weights runs simultaneously on two robots as they work together to put away groceries neither robot has ever seen before.

homes are filled with countless objects–delicate glassware, crumpled clothing, scattered toys–each with unpredictable shapes, sizes, colors, and textures. For robots to be useful in households, they will need to be capable of generating intelligent new behaviors on-demand, especially for objects they've never seen before.

The current state of robotics will not scale to the home without a step change. Teaching robots even a single new behavior currently requires substantial human effort: either hours of PhD-level expert manual programming or thousands of demonstrations. Both are prohibitively expensive when we consider how vast the problem of the home truly is.

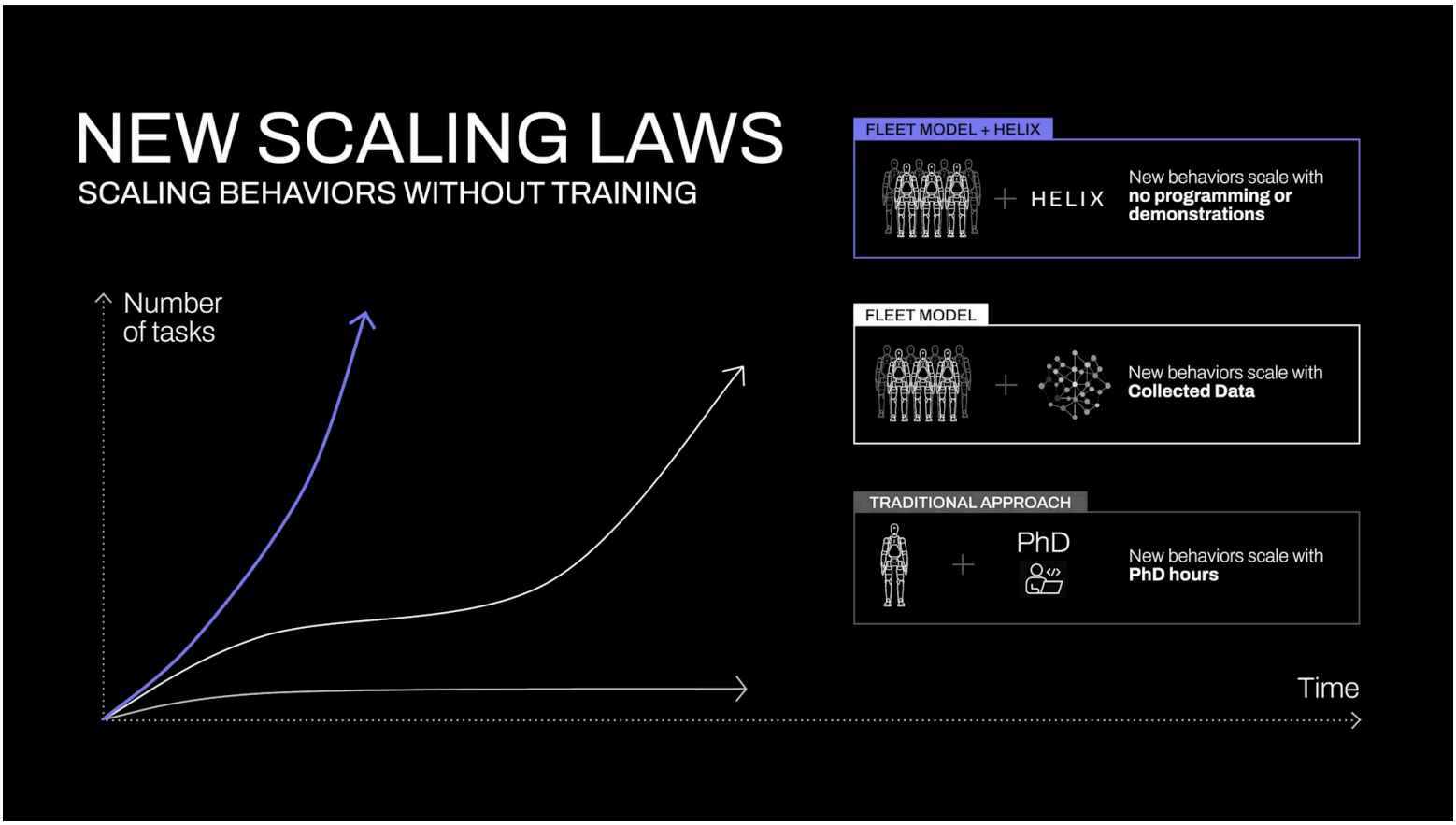

Figure 1: Scaling curves for different approaches to acquiring new robot skills. In conventional heuristic manipulation, skills grow with PhDs who manually script them. In conventional robot imitation learning, skills scale with data collected. With Helix, new skills can be specified on the fly with language.

But other domains of AI have mastered this kind of instant generalization. What if we could simply translate the rich semantic knowledge captured in Vision Language Models (VLMs) directly into robot actions? This new capability would fundamentally alter robotics' scaling trajectory (Figure 1). Suddenly, new skills that once took hundreds of demonstrations could be obtained instantly just by talking to robots in natural language.

The key problem becomes: how do we extract all this common-sense knowledge from

Helix: A "System 1, System 2" VLA for Whole

Helix is a first-of-its-kind "System 1, System 2" VLA model for high-rate, dexterous control of the entire humanoid upper body.

Prior approaches face a fundamental tradeoff: VLM backbones are general, but not fast, and robot visuomotor policies are fast but not general. Helix resolves this tradeoff through two complementary systems, trained end-to-end to communicate:

- System 2 (S2): An onboard internet-pretrained VLM operating at 7-9 Hz for scene understanding and language comprehension, enabling broad generalization across objects and contexts.

- System 1 (S1): A fast reactive visuomotor policy that translates the latent semantic representations produced by S2 into precise continuous robot actions at 200 Hz.

This decoupled architecture allows each system to operate at its optimal timescale. S2 can "think slow" about high-level goals, while S1 can "think fast" to execute and adjust actions in real-time. For example, during collaborative behavior (see Video 2), S1 quickly adapts to the changing motions of a partner robot while maintaining S2's semantic objectives.

Video 2: Helix allows for fast fine grained motor adjustments, necessary when reacting to a collaborative partner, while carrying out novel semantic goals.

Helix's design offers several key advantages over existing approaches:

- **Speed and Generalization**: Helix matches the speed of specialized single-task behavioral cloning policies while generalizing zero-shot to thousands of novel test objects.

- **Scalability**: Helix directly outputs continuous control for high-dimensional action spaces, avoiding complex action tokenization schemes used in prior VLA approaches, which have shown some success in low-dimensional control setups (e.g. binarized parallel grippers) but face scaling challenges with high-dimensional humanoid control.

- **Architectural Simplicity**: Helix uses standard architectures - an open source, open weight VLM for System 2 and a simple transformer-based visuomotor policy for S1.

- **Separation of concerns**: Decoupling S1 and S2 allows us to iterate on each system separately, without constraints of finding unified observation space or action representations.

Model and Training Details

Data

We collect a high quality, multi-robot, multi-operator dataset of diverse teleoperated behaviors, ~500 hours in total. To generate natural language-conditioned training pairs, we use an auto-labeling VLM to generate hindsight instructions. The VLM processes segmented video clips from the onboard robot cameras, prompted with: "What instruction would you have given the robot to get the action seen in this video?" All items handled during training are excluded from evaluations to prevent contamination.

Architecture

Our system comprises two main components: S2, a VLM backbone, and S1, a latent-conditional visuomotor transformer. S2 is built on a 7B-parameter open-source, open-weight VLM pretrained on internet-scale data. It processes monocular robot images and robot state information (consisting of wrist pose and finger positions) after projecting them into vision-language embedding space. Combined with natural language commands specifying desired behaviors, S2 distills all semantic task-relevant information into a single continuous latent vector, passed to S1 to condition its low-level actions.

S1, an 80M parameter cross-attention encoder-decoder transformer, handles low-level control. It relies on a fully convolutional, multi-scale vision backbone for visual processing, initialized from pretraining done entirely in simulation. While S1 receives the same image and state inputs as S2, it processes them at a higher frequency to enable more responsive closed-loop control. The latent vector from S2 is projected into S1's token space and concatenated with visual features from S1's vision backbone along the sequence dimension, providing task conditioning.

S1 outputs full upper body humanoid control at 200hz, including desired wrist poses, finger flexion and abduction control, and torso and head orientation targets. We append to the action space a synthetic "percentage task completion" action, allowing Helix to predict its own termination condition, which makes it easier to sequence multiple learned behaviors.

Training

Helix is trained fully end-to-end, mapping from raw pixels and text commands to conti[illegible] S1 into S2 via the latent communication vector used to condition S1's behavior, allowing joint optimization of both components. Helix requires no task-specific adaptation; it maintains a single training stage and single set of neural network weights without separate action heads or per-task fine-tuning stages.

During training, we add a temporal offset between S1 and S2 inputs. This offset is calibrated to match the gap between S1 and S2's deployed inference latency, ensuring that the real-time control requirements during deployment are accurately reflected in training.

Optimized Streaming Inference

Helix's training design enables efficient model parallel deployment on Figure robots, each equipped with dual low-power-consumption embedded GPUs. The inference pipeline splits across S2 (high-level latent planning) and S1 (low-level control) models, each running on dedicated GPUs. S2 operates as an asynchronous background process, consuming the latest observation (onboard camera and robot state) and natural language commands. It continuously updates a shared memory latent vector that encodes the high-level behavioral intent.

S1 executes as a separate real-time process, maintaining the critical 200Hz control loop required for smooth whole upper body action. It takes both the latest observation and the most recent S2 latent vector. The inherent speed difference between S2 and S1 inference naturally results in S1 operating with higher temporal resolution on robot observations, creating a tighter feedback loop for reactive control.

This deployment strategy deliberately mirrors the temporal offset introduced in training, minimizing the train-inference distribution gap. The asynchronous execution model allows both processes to run at their optimal frequencies, allowing us to run Helix as fast as our fastest single task imitation learning policies.

Results

Video 3: Helix's VLA controls the full humanoid upper body, a first in robot learning.

Fine-grained VLA whole upper body control

Helix coordinates a 35-DoF action space at 200Hz, controlling everything from individual finger movements to end-effector trajectories, head gaze, and torso posture. Head and torso control pose unique challenges—as they move, they change both what the robot can reach and what it can see, creating feedback loops that have historically caused instability. Video 3 demonstrates this coordination in action: the robot smoothly tracks its hands with its head while adjusting its torso for optimal reach, all while maintaining precise finger control for grasping. Historically, achieving this level of precision with such a high-dimensional action space has been considered extremely challenging, even for a single known task. To our knowledge, no prior VLA system has demonstrated this degree of real-time coordination while maintaining the ability to generalize across tasks and objects.

0:00

Video 4: Helix coordinates precise multi-robot manipulation.

Zero-shot multi-robot coordination

We push Helix to the limit in a challenging multi-agent manipulation scenario: collaborative zero-shot grocery storage between two Figure robots. Video 1 showcases two fundamental advances: The robots successfully manipulate entirely novel groceries—items never encountered during training—demonstrating robust generalization across diverse shapes, sizes, and materials. Additionally, both robots operate using identical Helix model weights, eliminating the need for robot-specific training or explicit role assignments. They achieve coordination through natural language prompts like "Hand the bag of cookies to the robot on your right" or "Receive the bag of cookies from the robot on your left and place it in the open drawer" (see Video 4). This marks the first demonstration of flexible, extended collaborative manipulation between multiple robots using a VLA, particularly significant given their successful handling of completely novel objects.

Emergent "Pick up anything"

0:00

We find that Figure robots equipped with Helix can pick up virtually any small household object with a simple "Pick up the [X]" command. In systematic testing, the robots successfully handled thousands of novel items in clutter—from glassware and toys to tools and clothing—without any prior demonstrations or custom programming.

Particularly notable is how Helix bridges the gap between internet-scale language understanding and precise robot control. When prompted to "Pick up the desert item", for instance, Helix not only recognizes that a toy cactus matches this abstract concept, but also selects the closest hand and executes the precise motor commands needed to grasp it securely.

This general-purpose "language-to-action" grasping capability opens new exciting new possibilities for humanoid deployment in unstructured environments.

Video 5: Helix translates high level conceptual commands like "Pick up the desert item" to low-level action.

Discussion

Helix's training is highly efficient

Helix achieves strong object generalization with remarkably few resources. We train Helix with ~500 hours of high quality supervised data in total, a small fraction of the size of previously collected VLA datasets (<5%), and without any dependencies around multi-robot-embodiment collect or multiple stages of training. We note that this is a scale of collect more comparable to modern *single*-task imitation learning datasets. Despite this comparatively small data requirement, Helix scales to the significantly more challenging action space of full upper body humanoid control, with high-rate, high-dimensional outputs.

A single set of weights

Existing VLA systems often require specialized fine-tuning or dedicated action heads to optimize performance across different high-level behaviors. Remarkably, Helix achieves strong performance across diverse tasks with a single unified model. Using just one set of neural network weights (7B for System 2, 80M for System 1), Helix picks and places items

in various containers, operates drawers and refrigerators, coordinates dexterous multi-

Video 6 "Pick up the helix"

Conclusion

We have presented Helix, the first Vision-Language-Action model to directly control an entire humanoid upper body from natural language. Unlike earlier robot systems, Helix is capable of generating long-horizon, collaborative, dexterous manipulation on the fly without any task-specific demonstrations or extensive manual programming. Helix displays strong object generalization, being able to pick up thousands of novel household items with varying shapes, sizes, colors, and material properties never encountered before in training, simply by asking in natural language. This represents a transformative step forward in how Figure scales humanoid robot behaviors—one that we believe will be pivotal as our robots increasingly assist in everyday home environments.

While these early results are truly exciting, we think they only scratch the surface of what is possible. We are eager to see what happens when we scale Helix by 1,000x and beyond. If you're as fascinated by the possibilities of Helix—and the future of dexterous humanoid robotics, we invite you to join us on this journey.

KEEP UP WITH US.

Get news, photos, events, and business updates

Email Address*

Sign Up

Contact Us

01 ABOUT US

02 AI

03 MASTER PLAN

04 CULTURE

05 NEWS

06 CAREERS