Keystone Arch Architecture for Embodied Intelligence: Toward General-Purpose Robots

Yingdong Hu

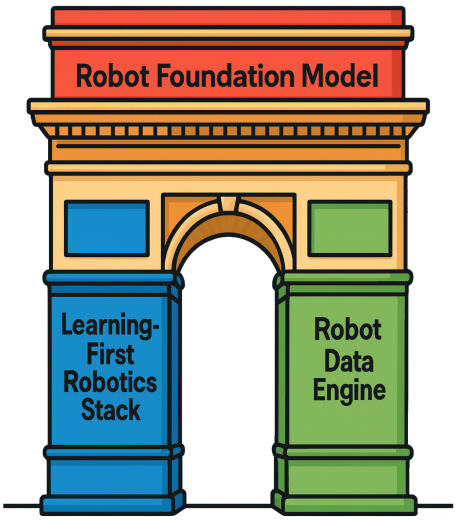
October 2025

We are witnessing the rapid advancement of digital AI, exemplified by large language models (LLMs) that are reshaping how humans create, communicate, and reason. Yet in the physical world, robots have not entered our daily lives at scale. In open environments, they remain conservative and brittle, struggling with rich contact dynamics, visual occlusions, cluttered scenes, and long-horizon decision making.

My research aims to develop general-purpose robots endowed with physical intelligence. Such robots must combine two traits: (1) human-like **agility and dexterity** when interacting with objects, and (2) the ability to **generalize** to novel tasks and unseen environments.

To achieve this, my past and ongoing work builds toward what I call the **Keystone Arch Architecture**:

- **The left pillar** is a **learning-first robotics stack**, which renovates the traditional perception → planning → control pipeline by embedding scalable learning-based methods at every layer.
- **The right pillar** is a **robot data engine**, which investigates what types of data are most effective and how to scale them properly to fuel learning-based methods.
- **The keystone** that locks these two pillars into a unified arch is a generalist **robot foundation model**—a model capable of perform diverse and complex tasks across different robot embodiments in a generalizable manner.



The two pillars form the essential foundation of this system—each indispensable on its own—while the keystone bridges and unifies them into a cohesive whole. In the following sections, I introduce each of these three elements in detail.

Learning-First Robotics Stack

Over the past decade, a common thread across the major advances in AI has been the **replacement of hand-crafted designs with end-to-end learning-based methods**. For example, in computer vision, convolutional neural networks eliminated the need for engineered features such as SIFT and HOG. A similar trend is taking place in robotics, where learning-based methods are increasingly used to replace or augment traditional modules. Classical robotics pipelines are typically divided into three major components—***perception***, ***planning***, and ***control***. My research explores how learning-based methods can make each of these components more scalable and generalizable.

Perception. The central challenge of perception lies in extracting representations from raw observations (e.g., RGB images, depth maps) that effectively support downstream robotic tasks.

To address this, my collaborators and I have developed a series of works—**SFC** (Semantic-Aware Fine-Grained Correspondence) [1], **SGR** (Semantic-Geometric Representation) [2], and **SGRv2** [3]—unified by a core principle inspired by the hierarchical structure of the human visual system: humans naturally perceive high-level semantic information (e.g., scene categories) while simultaneously attending to low-level, fine-grained cues (e.g., precise object geometry). Our works leverage 2D pre-trained vision models to provide the high-level information, and each explores a different strategy for integrating low-level cues. Specifically, **SFC** employs pixel-level self-supervised learning, **SGR** incorporates 3D geometric information from point clouds, and **SGRv2** further exploits the inductive bias of action locality. By combining these complementary signals, we construct more general visual representations, yielding improved performance in object tracking and robotic manipulation tasks. In another work [4], I conducted a **large-scale benchmarking study** evaluating diverse pre-trained vision models for robotic control. A key finding is that the effectiveness of visual pre-training is highly dependent on the choice of the downstream policy learning algorithm. This work provides practical guidance for researchers on when and how to incorporate pre-trained vision models into robot learning pipelines.

Planning. In early 2023, as large vision-language models (VLMs) such as GPT-4V began to demonstrate rich world knowledge and commonsense reasoning, I set out to harness these capabilities for robotic planning. Planning is a natural interface to such knowledge: it requires a deep understanding of both the task and the environment to decide how a robot should act over time. I began with ***high-level*** planning and proposed **ViLa** (Robotic Vision-Language Planning) [5], which converts complex, long-horizon instructions into a closed-loop, step-by-step plan, where each step invokes an atomic skill. To our knowledge, this was the **first approach to use vision-language models for robotic task planning**, and it drew considerable attention on social media, including a like from the CEO of Figure. We then introduced **CoPa** [6], which leverages knowledge from foundation models to push planning to a ***lower level*** by generating directly executable actions—specifically, robot end-effector poses.

Control. In robotics, control is one of the most challenging problems—particularly for *high-dimensional, non-smooth systems* such as quadrupeds and humanoids. Conventional model-based approaches often struggle to handle the discontinuous contact dynamics inherent in these systems. Recently, data-driven methods, especially reinforcement learning (RL), have made remarkable advances: robots can now learn to walk, run, and recover from perturbations purely through interaction and experience. Our work **HuB** [7] takes this a step further by enabling humanoids to perform tasks that require **precise balance control**, such as long-duration single-leg standing. Yet this is only the beginning. In more complex loco-manipulation tasks, humanoids must coordinate their entire body to interact with objects and environments involving rich contact dynamics. Achieving such general, robust control demands a more powerful and unified whole-body controller—the focus of one of my ongoing research projects.

Future Work 1.1. The Design Space of End-to-End Policy Learning. I aim to advance end-to-end policy learning for robotics by systematically exploring its vast and under-explored design space. For manipulation tasks, most existing approaches rely heavily on imitation learning. One key question is how to enable robots to learn from *action-label-free videos*—such as the massive amount of unlabeled data available on the internet—to acquire manipulation skills [8, 9]. Beyond imitation, can *reinforcement learning* be leveraged to push the upper bound of policy performance, achieving near-perfect success rates? Another promising direction is to develop *action representations* analogous to visual representations learned through self-supervised pre-training. By pre-training on large-scale robotic action data, we may uncover structured representations that facilitate downstream policy learning. Moreover, most current policies rely only on short-horizon observations. Introducing *memory mechanisms* is crucial for solving long-horizon tasks that require temporal reasoning. Finally, our recent findings reveal a surprising phenomenon: current policy learning methods often overfit to proprioceptive inputs [10], and

in some cases, removing these inputs improves generalization [11]. This highlights that many **subtle but critical design choices** remain to be systematically explored.

Future Work 1.2. General Whole-Body-Control. Humanoid hardware has advanced rapidly in recent years, but what is still missing is a general whole-body controller. Such a controller can unlock two complementary kinds of value. 1) It enables *human-like, expressive, and personable motion*, empowering humanoids to perform activities such as competitive sports (e.g., table tennis, badminton) or artistic performances (e.g., dance). These capabilities can foster joy, engagement, and emotional connection in human–robot interaction. 2) It provides *practical productivity gains*. A robust whole-body controller must coordinate the entire body to perform loco-manipulation tasks—demonstrating fluid, robust, and adaptive behavior across both static and dynamic settings. Building upon such a foundation, we can further develop **low-latency, long-distance teleoperation systems** that enable remote work in manufacturing or service industries, while simultaneously collecting large-scale, high-quality robotic data—fueling the next generation of learning-based robot models.

Robot Data Engine

All learning-based methods are inherently *data-driven*. The success of LLMs has been fueled by training on massive amounts of internet-scale data. However, unlike the digital world, we do not yet have an "Internet for Robots." Collecting large-scale, high-quality robotic data remains difficult. This raises fundamental, million-dollar questions in robotics: *What data should we use to train robots?* And *how much data is needed to achieve general-purpose physical intelligence?*

My work **"Data Scaling Laws"** [12] seeks to provide partial answers to these questions. In this study, I investigate whether data scaling laws—similar to those observed in other domains—also exist in robotics, particularly in robotic manipulation, and whether appropriate data scaling can yield single-task robot policies that can be deployed zero-shot for any object within the same category in any environment. Throughout our research, we collect over *40,000 demonstrations* and execute more than 15,000 real-world robot rollouts under a rigorous evaluation protocol.

Our findings reveal several intriguing insights: First, the generalization performance of the policy follows a roughly **power-law relationship** with the number of environments and objects. Second, the **diversity** of environments and objects is far more important than the absolute number of demonstrations. Based on these insights, we propose an efficient data collection strategy. With four data collectors working for one afternoon, we collect sufficient data to enable the policies for two tasks to achieve approximately **90% success rates** in **8 novel environments** with **unseen objects**.

This work has received broad recognition from the research community, winning the **CoRL 2024 X-Embodiment Best Paper Award** and being selected as an **ICLR 2025 Oral**. All robotic data were collected using UMI (a hand-held gripper) in this work. In a related line of work, MotionTrans [13], we further explore how to leverage human motion data captured via VR devices—which is abundant, easy to collect, and rich in diverse manipulation behaviors—to directly learn new motions for task completion.

Future Work 2.1. Dexterous In-the-Wild Data Collection. In our Data Scaling Laws project, we employed a hand-held parallel-jaw gripper (UMI) to collect large-scale demonstration data. UMI enables in-the-wild data collection, which greatly enhances data diversity—a key factor in scaling—but such systems still lack the dexterity and expressiveness of the *human hand*. Recently, hardware for dexterous hands has advanced rapidly, with the emergence of affordable, robust, and anthropomorphic designs. This opens exciting opportunities for future research to develop **in-the-wild data collection systems** specifically tailored for **dexterous**

manipulation. The design challenge lies in creating collection hardware that is *portable*, *intuitive*, and *low-cost*, yet capable of capturing rich multimodal information—including visual observations, tactile signals, and action trajectories. On the algorithmic side, policy learning methods must evolve to fully leverage dexterous data for more general and adaptive manipulation capabilities. Building on these foundations, an important open question remains: *Do data scaling laws extend to dexterous manipulation tasks?* Answering this could provide fundamental insights into how large-scale data can unlock richer, more human-like robotic skills.

Future Work 2.2. The Data Pyramid for Robots. The data pyramid in robotics consists of three distinct layers. At the base lies a massive amount of unstructured web data, in the middle sits synthetic data from simulation, and at the top rests high-quality but expensive real-world data. To train a generalist robot policy, a key question arises: *Can we rely on only one type of data, or must we integrate all three?* If the latter, how should we design training curricula that leverage data of varying fidelity and cost—for instance, pretraining on large-scale web data, intermediate adaptation in simulation, and fine-tuning with limited real-world demonstrations? Moreover, we must consider how to weight or balance contributions from different layers so that lower-fidelity data helps rather than misleads the learning process. Ultimately, a promising direction is to develop adaptive systems that can dynamically determine which samples, from which layer, are most beneficial at each stage of training—automatically shaping the learning process to make the most of diverse data sources.

Robot Foundation Model

While the left pillar (Learning-First Robotics Stack) and right pillar (Data Engine) provide the essential supports, what ultimately completes the **Keystone Arch Architecture** is the keystone itself—a generalist **robot foundation model**. This model serves as a large-scale robot policy that unifies the key modules of the Robotics Stack within a *single architecture*, leveraging the *diverse, large-scale* data provided by the Data Engine to achieve broad generalization and dexterous physical manipulation.

My first work in this direction is **OneTwoVLA** [14], a unified vision-language-action (VLA) model capable of both *acting* ("System One") and *reasoning* ("System Two"). A central innovation of OneTwoVLA is its ability to adaptively switch between these two modes—explicitly reasoning at critical moments during task execution and generating actions based on its most recent reasoning at other times. Gemini Robotics 1.5 later cited our work and described this capability as **Embodied Thinking**—the ability for a robot to "think" before acting.

During my internship at ByteDance, I contributed to the **GR-3** project [15], where I participated in the training and evaluation of large-scale VLA models. With access to the company's substantial computational resources, we co-trained the VLA with web-scale vision-language data, significantly improving its ability to generalize to novel objects, environments, and instructions.

Future Work 3.1. Scalable and Reliable Evaluation. In the second half of AI, evaluation becomes more important than training. I fully agree with this view—and believe that evaluation in robotics is even more challenging. Real-world evaluation introduces uncontrollable physical variability, task diversity, safety constraints, and a lack of standardized hardware setups and reproducible metrics. These challenges will only intensify as the field moves toward robot foundation models, which demand rigorous, scalable, and standardized evaluation frameworks. A promising direction is to develop **scalable simulation-based evaluation**. The first challenge is minimizing the *sim-to-real gap*: contact dynamics, friction, sensor noise, and actuation latency must all be modeled with high fidelity. The second challenge is efficiently generating *assets and scenes*, a task that is becoming increasingly tractable with generative AI models. Finally, we need mechanisms to automatically generate *diverse evaluation tasks* grounded in real-world

goals, ensuring a strong correlation between simulated and real-world performance. Building such evaluation infrastructure will be essential for driving progress in general-purpose robot learning.

Future Work 3.2. World Model. While VLA models represent one promising path toward a robot foundation model, an equally compelling alternative lies in action-conditioned world models that can **predict the future**. This direction raises two fundamental questions. The first concerns *how to construct the world model itself.* Some studies treat video generation models as a proxy for world models, but our physical world is intrinsically three-dimensional, requiring richer representations. Moreover, even state-of-the-art models such as Genie 3 accept only low-dimensional discrete actions as input conditions, while handling high-dimensional continuous robot actions remains an open challenge. The second question is *how should the world model be used in robotics?* One approach is to generate abundant neural trajectories from the world model to augment real robot trajectories. Another is to use the world model for planning, in conjunction with Model Predictive Control (MPC)—raising questions such as how to specify planning goals (e.g., via goal images or language descriptions). Ultimately, my goal is to enable *dexterous manipulation tasks* beyond simple pick-and-place using only the world model. Achieving this would mark a critical step toward making world models truly useful for robotics, bridging predictive modeling and real-world control.

Publications

[1] **Yingdong Hu**, Renhao Wang, Kaifeng Zhang, and Yang Gao. Semantic-aware fine-grained correspondence. In *European Conference on Computer Vision*, 2022.

[2] Tong Zhang*, **Yingdong Hu***, Hanchen Cui, Hang Zhao, and Yang Gao. A universal semantic-geometric representation for robotic manipulation. In *Conference on Robot Learning*, 2023.

[3] Tong Zhang, **Yingdong Hu**, Jiacheng You, and Yang Gao. Leveraging locality to boost sample efficiency in robotic manipulation. In *Conference on Robot Learning*, 2024.

[4] **Yingdong Hu**, Renhao Wang, Li Erran Li, and Yang Gao. For pre-trained vision models in motor control, not all policy learning methods are created equal. In *International Conference on Machine Learning*, 2023.

[5] **Yingdong Hu***, Fanqi Lin*, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*, 2023.

[6] Haoxu Huang, Fanqi Lin, **Yingdong Hu**, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[7] Tong Zhang, Boyuan Zheng, Ruiqian Nai, **Yingdong Hu**, Yen-Jen Wang, Geng Chen, Fanqi Lin, Jiongye Li, Chuye Hong, Koushil Sreenath, et al. Hub: Learning extreme humanoid balance. In *Conference on Robot Learning*, 2023.

[8] Yuyang Liu, Weijun Dong, **Yingdong Hu**, Chuan Wen, Zhao-Heng Yin, Chongjie Zhang, and Yang Gao. Imitation learning from observation with automatic discount scheduling. In *International Conference on Learning Representations*, 2024.

[9] Jialei Huang, Zhao-Heng Yin, **Yingdong Hu**, and Yang Gao. Policy contrastive imitation learning. In *International Conference on Machine Learning*, 2023.

[10] Fuhang Kuang, Jiacheng You, **Yingdong Hu**, Tong Zhang, Chuan Wen, and Yang Gao. Adapt your body: Mitigating proprioception shifts in imitation learning. *arXiv preprint arXiv:2506.23944*, 2025.

[11] Juntu Zhao, Wenbo Lu, Di Zhang, Yufeng Liu, Yushen Liang, Tianluo Zhang, Yifeng Cao, Junyuan Xie, **Yingdong Hu**, Shengjie Wang, et al. Do you need proprioceptive states in visuomotor policies? *arXiv preprint arXiv:2509.18644*, 2025.

[12] **Yingdong Hu***, Fanqi Lin*, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. In *International Conference on Learning Representations*, 2025.

[13] Chengbo Yuan, Rui Zhou, Mengzhen Liu, **Yingdong Hu**, Shengjie Wang, Li Yi, Chuan Wen, Shanghang Zhang, and Yang Gao. Motiontrans: Human vr data enable motion-level learning for robotic manipulation policies. *arXiv preprint arXiv:2509.17759*, 2025.

[14] Fanqi Lin*, Ruiqian Nai*, **Yingdong Hu***, Jiacheng You, Junming Zhao, and Yang Gao. Onetwovla: A unified vision-language-action model with adaptive reasoning. *arXiv preprint arXiv:2505.11917*, 2025.

[15] Chilam Cheang, Sijin Chen, Zhongren Cui, **Yingdong Hu**, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, et al. Gr-3 technical report. *arXiv preprint arXiv:2507.15493*, 2025.