中图法分类号： **文献标识码：** A **文章编号：** 1006-8961(XXXX)XX-0001-29

论文引用格式： Liang Shutong, Xie Dongjin, Li Dong, Zhang Hui, Jia Xiaofeng, Wang Fei-Yue, Li Yidong, Li Lingxi. Dexterous Robotic Hands: Enabling General-Purpose Manipulation[J/OL]. Journal of Image and Graphics, XXXX:1-29. DOI: 10.11834/jig.260100.(梁姝彤, 谢东锦, 李东, 张慧, 贾晓丰, 王飞跃, 李浥东, 李灵犀. 机器人灵巧手:迈向通用操作的关键技术[J/OL]. 中国图象图形学报,XXXX:1-29. DOI: 10.11834/jig.260100.)[DOI:10.11834/jig.260100]

机器人灵巧手:迈向通用操作的关键技术

梁姝彤[1†],谢东锦[2†],李东[3],张慧[1*],贾晓丰[4],王飞跃[3,5*],李浥东[1],李灵犀[6]

1. 北京交通大学,北京 100044; 2. 新疆大学,乌鲁木齐 830049; 3. 澳门科技大学,澳门 999078; 4. 北京大数据中心,北京 101117; 5. 中国科学院自动化研究所,北京 100190; 6. 普渡大学,印第安纳州西拉法叶 47907,美国

摘 要： 灵巧手是人形机器人实现高维度、精细化物理交互的关键末端执行器,其高自由度、强接触非线性与多模态反馈耦合,使灵巧操作成为具身智能最具代表性的挑战任务之一。近年来,视觉-语言-动作模型与大语言模型等基础模型范式的兴起,结合扩散/流匹配等连续控制建模、强化学习与模仿学习的融合训练,以及高分辨率触觉、可变刚度与刚柔混合结构的发展,正推动灵巧手从“刚性高精度”的机械决定论走向“感知-学习-执行”闭环驱动的柔性智能体系。本文首先从历史视角系统回顾灵巧手机械结构与硬件范式的演进脉络,涵盖多指全驱动、欠驱动柔顺、腱绳传动以及软体与变刚度等代表性路线,并讨论其在尺寸重量、可靠性与可控性之间的权衡。其次,提出以感知能力演进为主线的五级灵巧智能分级框架(H1-H5),归纳各层级的关键使能技术、典型方法与能力边界,为评估“从可重复执行到开放世界任务规划,再到自主进化”的能力跃迁提供统一参照。进一步地,本文从真实交互与高保真仿真两个维度梳理训练数据来源与评测基准,强调数据管线与可诊断评估标准对任务泛化与可部署性的基础作用。最后,总结灵巧手走向通用化部署仍面临的机械可靠性与成本、实时推理与安全性、仿真可信化与标准化评测等关键挑战,并展望软硬件协同设计、多模态自监督预训练与世界模型驱动的长时序决策等研究方向。

关键词： 灵巧手;具身智能;人形机器人;多模态触觉;视觉-语言-动作模型;任务泛化

Dexterous Robotic Hands: Enabling General-Purpose Manipulation

Liang Shutong[1†], Xie Dongjin[2†], Li Dong[3], Zhang Hui[1*], Jia Xiaofeng[4], Wang Fei-Yue[3,5*], Li Yidong[1], Li Lingxi[6]

1. Beijing Jiaotong University, Beijing 100044, China; 2. Xinjiang University, Urumqi 830049, China; 3. Macau University of Science and Technology, Macau 999078, China; 4. Beijing Big Data Centre, Beijing 101117, China; 5. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; 6. Purdue University, West Lafayette IN 47907, United States

Abstract: In recent years, the rapid advancement of foundation models, including large language models (LLMs), vision-language models (VLMs) and world models, has introduced a paradigm shift that enables humanoid robots to transition from laboratory demonstrations to open-world applications such as household services, industrial manufacturing, and medical assistance. As the primary end-effector responsible for high-dimensional and fine-grained physical interaction, multi-fingered dexterous hands represent one of the most challenging and emblematic platforms in embodied intelligence, due to their high degrees of freedom, strongly nonlinear contact dynamics, and tightly coupled multimodal feedback mechanisms. The emergence of vision-language-action (VLA) models and large-scale foundation architectures, the breakthrough appli-

收稿日期: 2026-02-14;**修回日期:** 2026-03-09

* **通信作者:** 王飞跃 feiyue.wang@ia.ac.cn;张慧 huizhang1@bjtu.edu.cn

基金项目: 国家自然科学基金项目(项目编号:62203040);中央高校基本科研业务费(项目编号:2024JBMC045)

Supported by: National Natural Science Foundation of China under Grants (62203040); Fundamental Research Funds for the Central Universities (2024JBMC045)

cation of diffusion models and flow matching in continuous control policy generation, hybrid reinforcement-imitation learning frameworks, and advances in high-resolution tactile sensing, variable-stiffness mechanisms, and rigid-soft hybrid materials are collectively driving a fundamental transition in dexterous hands—from a paradigm of "rigid high-precision" mechanical determinism toward an integrated, perception-learning-execution-centered closed-loop intelligent system. This paper presents a comprehensive review of robotic dexterous hands across four dimensions: mechanical structures, intelligence capability grading, data resources, and benchmarking methodologies. First, from a historical perspective, we systematically trace the evolution of mechanical architectures and hardware paradigms, summarizing representative technical routes including fully actuated multi-finger designs, underactuated compliant mechanisms, tendon-driven systems, soft robotic hands, and rigid-soft hybrid structures. Our analysis indicates that the evolution of dexterous hand mechanisms is not merely an accumulation of degrees of freedom, but rather a gradual shift toward engineering-oriented paradigms characterized by underactuated coupling, material compliance, and hybrid structural design. By embedding adaptive coordination mechanisms into the mechanical body through passive responses, these approaches effectively reduce actuation and control dimensionality while physically enhancing robustness against object diversity and contact uncertainty. Building upon this foundation, we propose a systematic five-level taxonomy of dexterous intelligence (H1-H5) centered on the evolution of perceptual capability. H1 (Perception-Free) is characterized by open-loop program execution and teleoperation, where the system lacks environmental modeling and policy generation capabilities. H2 (Single-Modal Perception) introduces either vision or tactile feedback to enable perception-driven grasping and basic stability regulation. H3 (Multimodal Perception) integrates vision, tactile, and force sensing through deep multimodal collaboration, supporting complex fine manipulation tasks such as precision assembly, deformable object manipulation, and tool use. At this stage, systematic methodologies emerge across three technical directions: hierarchical task planning, multimodal servo control, and data-driven policy learning. H4 (Open Perception) centers on vision-language-action models and addresses perceptual generalization, long-horizon task planning, and deep multimodal fusion to enable language-guided open-world task understanding and zero-shot manipulation. H5 (Dynamic Perception) envisions autonomous, evolving general manipulation capabilities supported by deep multimodal dynamic perception and real-time coordination mechanisms, representing a historical leap from robots as "tools" to embodied "symbiotic agents." This taxonomy provides a unified reference framework for evaluating the technological transition of dexterous hands from repetitive execution to open-world task planning and ultimately toward autonomous evolution. Furthermore, we systematically review the key data resources and evaluation benchmarks that support dexterous intelligence from two complementary dimensions: real-world interaction and high-fidelity simulation. At the data level, real-world datasets offer ecological validity but suffer from high collection costs, limited scalability, and safety risks. Synthetic datasets and simulation platforms enable large-scale and diverse data generation at controllable costs but remain constrained by simplified contact models and the simulation-to-reality gap. We outline the evolution of synthetic datasets from static grasp poses to dynamic manipulation sequences and analyze representative resources in terms of their contributions to grasp generation, cross-hand generalization, articulated object manipulation, and long-horizon modeling. We further summarize the technological progression of simulation platforms from basic physical validation to high-fidelity interaction and cross-domain transfer. In terms of evaluation, we categorize performance metrics into outcome-oriented and process-oriented dimensions, including task success rate, grasp cycle time, target pose error, normalized task error, contact region error, stability and drop rate, as well as efficiency and robustness. Benchmark tasks are organized into five families: stable grasping and transport; re-grasping and contact transition; in-hand manipulation and reorientation; constrained operation and assembly; and tool use and functional manipulation. Together, these constructs form a systematic two-dimensional evaluation spectrum spanning contact complexity and temporal depth, emphasizing reproducible and diagnostically meaningful standards for assessing generalization capability and deployment readiness. Finally, we summarize the core challenges and future directions toward the general-purpose deployment of dexterous hands. From a data perspective, the scarcity of real interaction data and the persistent simulation-to-reality gap remain fundamental bottlenecks for effective policy transfer. From a modeling perspective, efficient and robust multimodal joint representations, 3D foundation model construction, and interpretable decision-making mechanisms have yet to converge into a unified theoretical framework, while inherent tensions persist between model scale and real-time inference requirements. From a hardware

perspective, long-standing engineering trade-offs exist between high degrees of freedom and low cost, reliability, and lightweight design, as well as between precision force–tactile control and structural simplicity. Looking ahead, deep integration of perception, decision-making, and execution; incorporation of physical commonsense and causal reasoning through world models and embodied foundation models; generative AI–driven data-efficient learning and simulation credibility enhancement; biomimetic variable-stiffness mechanisms and endogenous tactile sensing through soft–hard co-design; and long-term real-world deployment with closed-loop optimization in high-value scenarios such as intelligent manufacturing, domestic service, and specialized operations will be critical pathways. These efforts will drive dexterous hands from laboratory prototypes toward reliable real-world applications, ultimately achieving the goal of general embodied intelligence capable of perceiving, reasoning, and manipulating "like a human hand." This work provides a unified capability framework and systematic reference for understanding and tracking the frontier of robotic dexterous hands, offering theoretical guidance and practical insights for future research in hardware paradigm evolution, intelligence capability transition, and data and benchmarking system construction.

Key words: dexterous hand; embodied AI; humanoid robot; multimodal tactile perception; vision-language-action model; task generalization

0 引　言

近年来，大型语言模型、视觉语言模型与世界模型等基础模型快速发展，使机器人在“理解指令—感知环境—生成动作”的统一建模上出现了新的范式转折。人形机器人作为通用具身智能平台的重要载体，正从实验室演示走向家庭、工业与医疗辅助等开放场景；而在所有末端能力中，灵巧手承担着机器人对物体进行抓取、在手操作、力控装配与工具使用等精细交互的核心职责，其性能直接决定机器人能否稳定完成非结构化环境下的复杂任务。

与传统两指夹爪不同，多指灵巧手面对的是高维动作空间、频繁接触切换与复杂摩擦接触动力学带来的强非线性耦合问题；同时，精细操作往往需要视觉、触觉与力觉多模态信息的闭环融合，并在长时序任务中具备错误恢复与在线重规划能力。过去很长一段时间，灵巧手研究更多依赖精密加工与刚性传动，通过增加自由度与提高定位精度来提升能力，但其代价是驱动与传动链路复杂、系统维护困难、控制建模与鲁棒性不足。随着高分辨触觉、先进材料与增材制造、紧凑高功率密度致动器的成熟，以及模仿学习、强化学习与视觉-语言-动作模型等方法的发展，灵巧手的

设计与控制正在从“机械决定论”转向“感知-学习-执行”一体化的智能体系：通过更丰富的多模态反馈与更强的策略学习能力，以较低的先验依赖适应物体多样性与环境不确定性。

在这一背景下，一个关键问题逐渐凸显：如何建立可对齐的能力刻画与评价体系，系统理解灵巧手从“能抓”到“会做事”、再到“能在开放世界持续进化”的技术路径，并明确每一阶段的核心瓶颈与研究重点。为此，本文从硬件范式与智能能力两条主线出发，系统综述机器人灵巧手的研究进展与发展趋势，主要贡献如下：

1）从历史演进视角梳理灵巧手机械结构与硬件范式的发展脉络，总结多指全驱动、欠驱动柔顺、腱绳传动、软体与刚柔混合等路线的设计动机与工程权衡；

2）提出以感知能力演进为主线的五级灵巧智能分级框架（H1-H5, H-hand），覆盖从无感知开环执行到开放感知下语言引导任务规划，再到动态感知下自主进化的通用操作，并归纳各层级的关键技术与能力边界；

3）从真实交互与高保真仿真两个维度总结训练数据来源与评测基准，讨论仿真-现实差距、数据成本与可复现评估标准对泛化与部署的影响；

4）结合当前趋势，分析通用灵巧操作走向实

用化仍需突破的机械可靠性与低成本化、多模态闭环的高效可解释策略、实时推理与安全约束、仿真可信化与标准化评测等关键技术挑战，并展望未来研究方向。

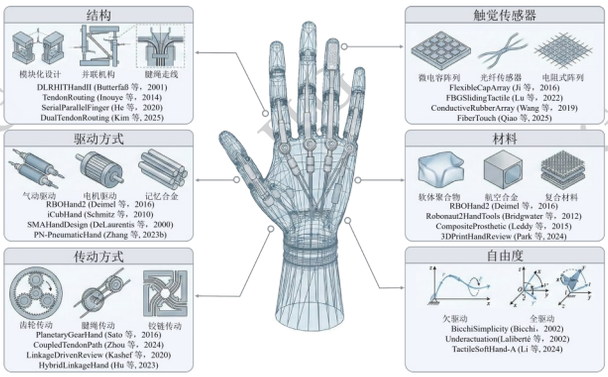


图 1　多指灵巧手系统分类及关键技术概览，包括：结构、驱动方式、传动方式、触觉传感器、材料及自由度分类

Fig. 1　Classification and key technologies of multi-finger dexterous hand systems: structures, actuation, transmission, tactile sensing, materials, and degrees of freedom

1　机械结构与硬件范式演进

1.1　从功能性钩爪到灵巧需求的提出

在灵巧手概念尚未被明确提出的早期阶段，机械手的应用场景主要集中在假肢康复和工业搬运等领域，其核心目标是“恢复或替代基本抓取功能”，而非实现类似人手的精细操控。典型的假肢钩爪如 Dorrance Split Hook（Belter 等，2014）、Hosmer Model 5XA（Smit 等，2012）等均采用两指对置结构，依靠钢缆、弹簧或橡皮筋等简易机构实现开闭，多数为体驱动或简化电驱方式。Smit 等人（2012）的研究结果表明，体驱动钩爪相比仿生手通常具有更低的启动力度和更高的夹持力，其通过钢缆传递肩部、躯干或残余肢体动作以完成张合的设计，使用户能够在较低结构复杂度与能量输入条件下稳定实现“张合-夹持-搬运-释放”的基础操作闭环，体现出该类末端装置在工程实现与日常使用层面的明确可行性，但同时也存在钢缆易磨损与动作模式受限等潜在问题。该研究还进一步显示，Hosmer Model 5XA 等经典钩爪在 Box and Blocks Test（盒块测试）等功能性测试中往往表现突出，但在更依赖指尖精细调姿、物体重定位与连续操控的任务中仍存在显著局限。从操作形式来看，它们更接近“单一夹具”，主要支持少数几种固定抓取模式，难以完成指尖旋转、物体重定位、精细调姿等复杂操作，在接触调节和力分配层面与人手存在本质差距。与假肢领域的“功能性抓取”需求类似，早期工业机械手末端执行器同样以结构简单的刚性夹具为主。Birglen 等人（2018）在其对工业机器人夹爪的统计性回顾中指出，工业领域长期主流的末端执行器形态集中在少数成熟结构上，尤其是气动驱动的平行两指夹爪最为常见；从厂商产品规格的统计分布来看，这类夹爪的设计与选型通常围绕行程、夹持力与重量等基础指标展开，以优先满足产线节拍、重复定位与稳定夹持等工程诉求。在工程实践中，常通过针对特定工件和工位设计专用夹具，结合少自由度机械臂完成重复性强、环境高度结构化的装配、上下料等任务。为了降低系统复杂度并提高可靠性，这一阶段的设计普遍将“抓得牢、放得准”置于首位，对多接触状态、掌内操控以及人与机器人近距离协作等问题关注有限。随着服务机器人、康复机器人以及家用机器人等新型应用的出现，传统假肢钩爪和工业夹具在面对形状多样、材质各异、姿态不确定的日常物体时，其操作能力不足的矛盾日益凸显。正是在这一系列现实需求和技术瓶颈的共同推动下，研究者逐渐意识到，仅依靠“能抓住”的简单夹持已无法满足未来人机共融场景的要求，“灵巧性”作为涉及多指协同、掌内操作与接触调节的综合能力，开始被系统性地提出并成为后续仿人灵巧手设计的出发点。

为从硬件视角系统梳理多指灵巧手的发展脉络，本文将多指灵巧手系统的关键硬件技术归纳为六个维度：结构、驱动、传动、触觉传感、材料体系与自由度配置（见图 1 所示）。

1.2　高自由度灵巧手的仿生演化

多指高自由度仿人灵巧手的核心目标，是在机

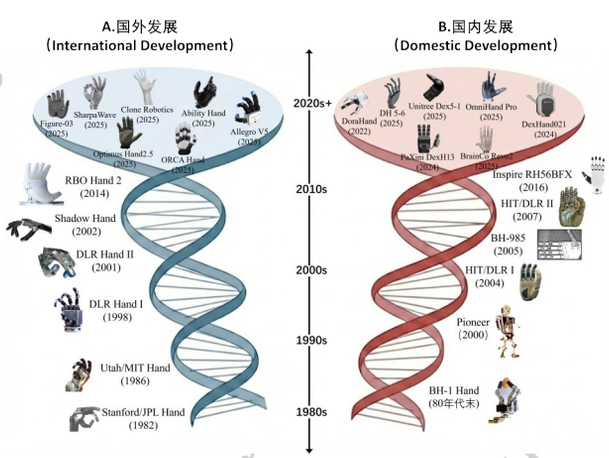


图 2　国内外不同的灵巧手的对比

Fig. 2　Comparison of domestic and international dexterous hands

器手末端执行器层面获得更大的运动范围与更强的抓取适应性，以支持精细抓取与复杂操作任务。相较于二指夹持器，该类系统通常采用三至五指构型，在单指内部设置多关节串联结构，并通过拇指对掌与一定的掌部弯曲设计，增强指间协同与对物体几何形状的适配能力。Gifu Hand III（Mouri 等，2002）采用“拇指+四指”的五指拓扑，具备 20 个关节与 16 个自由度，可覆盖更接近人手的指间协同与构型变化；DLR-Hand II（Butterfaß 等，2001；Borst 等，2003）作为多传感灵巧手研究平台，通过可重构手掌与多传感设计面向抓取/操作任务开展实验验证；Shadow Dexterous Hand（Walck 等，2014；Li 等，2019）以五指、24 自由度的腱驱结构提供了较大的关节可动范围，常被用作在手操作与学习控制的实验载体。上述仿生设计虽大幅拓展了构型空间与抓取模式，但执行器数量增多、传动链路复杂化、布线与散热压力加剧，导致系统难以在体积、重量与可靠性间实现平衡。然而，以“堆砌自由度”为特征的仿生设计也暴露出一系列难以忽视的工程和控制问题。一方面，随着手指数量和关节自由度的增加，系统维度快速上升并伴随频繁的接触切换与强非线性耦合，Li 等人（2022）在多指灵巧手的综述中指出，这类高维、多接触系统使得基于解析建模的轨迹规划与逐关节控制更难在同一框架下同时兼顾实时性、稳定性与鲁棒性。在冗余自由度条件下，如何选择合理的关节构型、协调多指接触力分布，并在摩擦模型不完备与接触不确定性存在的情况下保持抓取稳定，仍是高自由度灵巧手走向可靠应用的关键瓶颈；Bicchi（2002）在其关于“灵巧操控与稳健抓取”的典型工作中强调，多接触操控本质上受到摩擦锥约束、接触状态切换与不确定性影响，导致稳定性分析与控制设计难以简化为低维的解析问题。另一方面，大量微型执行器、减速器与传动机构被集成在紧凑的手掌与指骨内部，导致系统在体积与质量压缩方面受到明显限制，制造与维护成本随复杂度上升而显著增加；Gama Melo 等人（2014）在其对拟人灵巧手的系统回顾中总结指出，为获得更接近人手的运动能力，拟人灵巧手往往需要更复杂的驱动与传动布局，这会显著抬升机构实现难度与工程代价，并在可靠性与可维护性层面带来持续挑战。研究者逐渐认识到，在追求灵巧性的同时必须在自由度数量、驱动方式与控制复杂度之间进行全局权衡，这也直接推动了欠驱动灵巧手、软体灵巧手以及刚柔混合结构等更具工程可行性的方案发展。

1.3 从欠驱动到刚柔混合的结构演进

面对多指高自由度仿人灵巧手在成本、可靠性与控制复杂度上的挑战，欠驱动灵巧手提出了一种“以结构换控制”的折中方案。其核心思想是让系统“自由度多于驱动数”，通过连杆、滑轮、差动齿轮、弹簧等机械耦合元件，将多个关节绑定在少量驱动源下协同运动。当手指接触物体时，耦合链路会依据接触约束与受力变化自动重分配关节转角与输出力矩，使手指在无需逐关节精确轨迹控制的条件下形成对不同形状物体的自适应包覆抓取。Catalano 等人（2014）提出的 Pisa/IIT SoftHand 以“synergy”为结构设计原则，通过欠驱动耦合将多关节协同固化在机构本体中，实现少驱动条件下的自适应抓取；Odhner 等人（2014）提出的 iHY Hand 强调以较少驱动通道获得鲁棒抓取能力，面向实际任务兼顾抓取多样性与耐用性。总体而言，欠驱动通过机构耦合将部分协调与适配过程内化为被动机械响应，在降低驱动与控制维度的同时维持了可观的抓取适应性与稳定性。软体灵巧手进一步将“形变”本身纳入功能设计。通过采用硅胶、弹性体、织物骨架及纤维增强气动网络等柔性材料，软体手指可以在接触物体时产生连续、平滑的大幅弯曲与扭转，从而在物体几何形状未知或存在较大姿态误差时，仍能通过被动形变实现贴合和包覆。软体灵巧手在面对易碎物体、柔性物体以及与人体的近距离交互时具有天然优势：一方面，柔软结构可以显著降低单点接触应力，提高安全性；另一方面，柔顺性能够吸收定位误差和环境扰动，使抓取任务对高精度传感和控制的依赖程度显著降低。Ilievski 等人（2011）提出并系统展示了嵌入式气动网络（PneuNets）软执行器的结构范式，表明软材料在简单压力输入下即可产生大幅连续形变，从而支撑软体抓取与安全交互等任务。Polygerinos 等人（2015）面向可穿戴辅助与康复场景，基于纤维增强流体驱动软执行器实现了沿手指方向的柔顺施力与可控弯曲，展示了软体结构在贴合抓取与人机接触安全性方面的优势。然而，欠驱动结构固有的自由度耦合，使得系统在执行高精度指尖操作、独立调整某一关节姿态或精确控制各接触点力分布时存在明显限制。软体结构虽然形变能力突出，却在动力学建模、有限元仿真和高频控制方面面临更大

困难,同时其响应速度、寿命和环境稳定性也需要通过材料和结

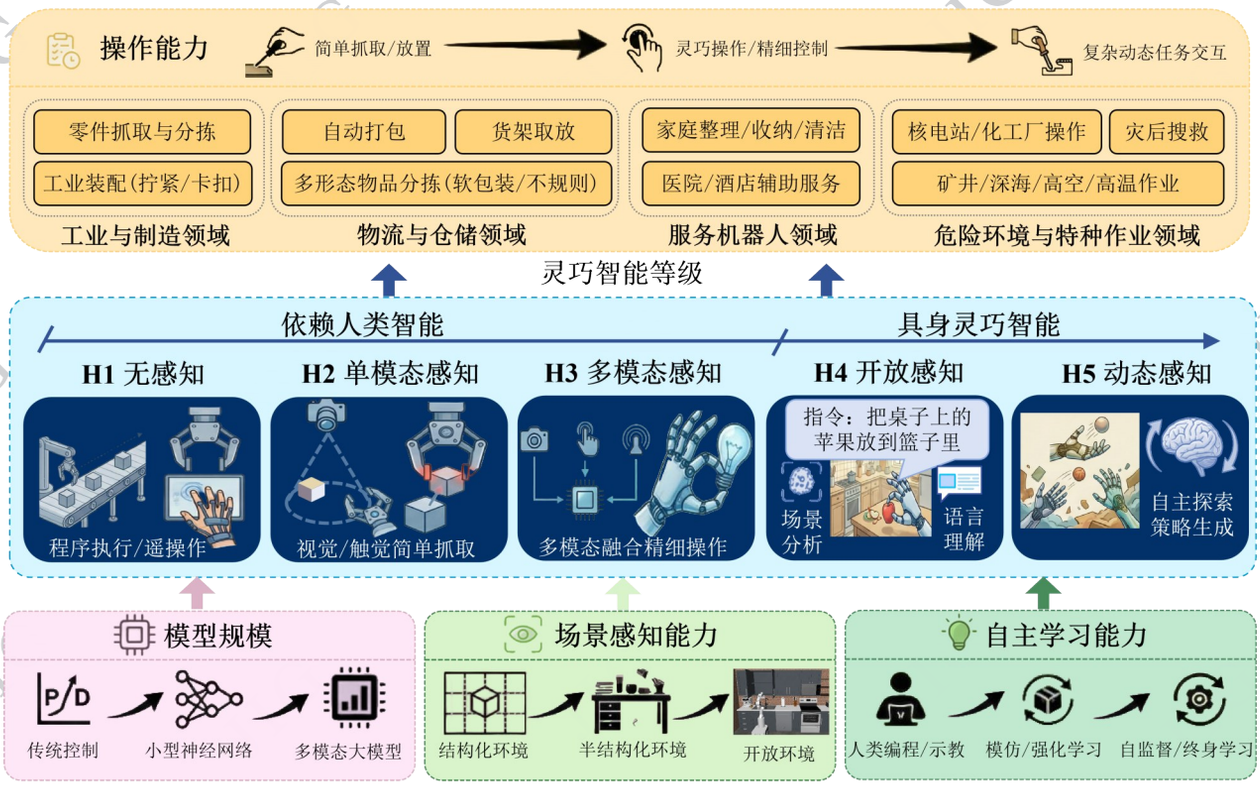


图 3　灵巧智能能力分级

Fig. 3　Classification of dexterous intelligence capabilities

构优化加以保证。因此,当前越来越多的工作倾向于采用“刚柔混合”的设计,即在关键关节上保留主动可控自由度,在指腹和指尖区域引入柔顺结构或软体材料,并结合少量欠驱动约束,期望在灵巧性、结构简洁性与控制可行性之间寻找新的平衡点。

综上,灵巧手机械结构的演进并非简单的自由度叠加,而是逐步转向以欠驱动、软体与刚柔混合为代表的工程化路线。该路线通过结构耦合与材料柔顺性将原本复杂的适应与协调机制内化为机械本体的被动响应,从而在有效降低驱动与控制维度的同时,从物理层面确保了系统面对多样物体与不确定接触时的鲁棒操作能力。

2　灵巧智能能力分级与关键技术

真正的智能并非预先编程或孤立存在于大脑之中,而是通过在物理世界、社会环境及语言交互的多模态具身体验中逐步“涌现”出来的(Smith 等,2005)。Bao 等人(2025)从认知模型与结构视角进一步深化了这一观点,即“智能为什么能涌现出来”以及“机器如何像人一样通过身体理解世界”。这一具身智能理念正深刻影响机器人学的发展,推动机器人操作走向与物理世界更深层次的交互与适应。在这一进程中,灵巧手作为实现精细操作与物理交互的关键载体,承担着重要作用。其操作能力正从依赖规则与模型的控制方式,加速迈向面向真实世界任务的自主智能阶段。这一演化不仅源于多模态感知、控制策略与机器人硬件的快速发展,更体现人类社会对高自主性、高通用性操作能力的迫切需求。

为此,本文从四个关键维度对机器人灵巧手的能力进行分类与分级。1)场景感知与泛化能力:从依赖结构化先验知识,向开放环境的主动感知与零样本泛化演进;2)操作任务复杂度:从简单抓取逐步发展为多指协同、精细操控、动态接触等复杂物理交互任务;3)模型规模与算力需求:从轻量化的专用简单控制模型,逐步过渡到支持多任务泛化与在线推理的大规模多模态策略模型;4)自主学习能力:从完全依赖人工编程和示教学习逐步迈向自驱动探索与自主进化能力。基于上述维度,本文以感知能力的演进为主线,提出并系统阐述灵巧手从 H1 到 H5 的

能力分级模型，如图3所示：从无感知的固定重复操作（H1），到具备单模态感知的简单抓取（H2），再到多模态感知融合的精细操作（H3），进而发展为开放感知下语言引导的任务规划（H4），最终实现动态感知中自主进化的通用操作（H5），涵盖灵巧手从基础机械操作向高自主性、多任务泛化能力的技术演进路径。

2.1 H1 无感知：开环程序执行与遥操作

H1级灵巧手代表机器人操作能力的初始形态。在这一阶段，灵巧手系统尚未形成任何意义上的环境感知与智能决策能力，其操作行为在算法层面表现为对外界状态完全“无感知”的执行过程。该等级的控制范式不涉及对物体属性、位姿变化或接触状态的在线估计，系统内部也不存在显式的状态表示或策略函数，灵巧手本体仅承担动作执行功能，而不具备环境建模与策略生成能力。在这一能力等级下，灵巧手的操作主要通过两类方式实现：一类是基于固定轨迹的程序化执行方法；另一类是通过遥操作实现的人机协同操作。前者依赖预先设计的机械结构与固定控制逻辑来实现对既定动作的稳定复现；后者将环境感知、任务理解与决策过程完全交由人类操作者完成，灵巧手自身仍不参与任何形式的状态感知与策略推理。

2.1.1 基于固定轨迹的程序化执行方法

基于固定轨迹的程序化执行方法在离线阶段由工程人员根据经验或规则，直接在软件或控制系统中预先设定灵巧手的运动路径、动作顺序及控制参数，在实际运行过程中严格按照既定程序重复执行拾取、放置、装配等操作任务（Fahlman，1974；Will等，1975；Lieberman等，1977）。由于该方法假定环境状态在执行阶段保持不变，其控制策略依赖目标物体位姿、几何模型及执行条件的高度一致性。然而，在实际操作过程中不可避免地存在多种不确定性因素，例如目标物体的实际位置与姿态偏离预期，环境与物体的几何模型存在建模误差，以及由机械结构非理想性（如关节摩擦、弹性效应）和控制延迟引入的执行偏差。这些不确定性会导致实际运动轨迹与预期轨迹产生偏离，并直接影响操作成功率。Bicchi（1995）通过对抓取过程中接触力的力学特性进行深度建模与分析，明确接触力的产生、约束及补偿机制，并借助这些物理特性来抵消抓取过程中的不确定性。Brost（1985，1988）通过分析挤压与推动操作的物理特性，设计挤压抓取、偏移抓取和推送抓取三种特定抓取策略，使规划的抓取动作能够在物体位置与姿态存在有界不确定性的情况下实现成功抓取。Peshkin等人（1986）和Wolter等人（1985）通过构建几何化的配置空间模型明确安全操作区域，将物体位置、机械臂运动等不确定性转化为几何空间中的约束边界，从而确保操作在界定几何范围内的执行鲁棒性。

由于H1级基于固定轨迹的程序化执行方法在执行过程中不引入任何视觉或触觉等感知反馈，系统无法对目标物体位姿偏差、环境变化或接触状态等关键因素进行在线感知、评估或补偿，也无法对既定运动轨迹进行动态修正。因此，该类方法的操作性能高度依赖于执行精度的长期稳定性以及环境条件的高度一致性，其适用范围被严格限制在高度结构化、低不确定性的操作场景中。

2.1.2 遥操作驱动的人机协同执行方法

在基于遥操作的人机协同执行方法中，灵巧手本体并不具备自主决策能力，其动作完全由人类操作者通过主从控制系统直接驱动（Li等，2019；Handa等，2020；Qin等，2023；Chi等，2025；Huang等，2025）。尽管在系统层面引入人类的感知与判断，能够在短期内高效完成复杂操作任务，但从灵巧手自身的算法能力来看，其仍处于无感知状态：环境理解、任务规划与策略调整均由人类操作者完成，灵巧手仅作为动作执行终端被动响应控制指令。因而，遥操作并未改变灵巧手对环境“无建模、无推理”的基本特性，其智能性并未内化至系统本体，整体仍处于无感知、无决策的H1级。

2.2 H2 单模态感知：视觉/触觉驱动的基础抓取

H2级灵巧手标志着机器人操作从“完全无感知执行”向“感知驱动操作”的初步过渡。在H1级，操作系统通常采用先规划后执行的开环控制范式，缺乏视觉或触觉反馈控制回路，使其对执行精度和环境一致性高度敏感，易因目标位姿偏差或执行误差导致操作失败。与H1阶段相比，该级别的核心特征在于：系统开始引入单一模态的感知信息（主要为视觉或触觉），用于支撑简单抓取任务中的状态估计与动作调整，但其感知范围、推理深度与决策能力仍然有限，尚未形成完整的多模态环境理解与自主规划能力。

2.2.1　视觉感知驱动的闭环抓取

H1级的固定装配线通常假设零件以确定且一致的姿态到达装配工位，一旦位姿发生变化，往往需要对机械结构或工艺流程进行重新设计，灵活性较低。为缓解装配系统对零件初始位姿一致性的强依赖，研究者开始在装配流程中引入视觉感知系统，使系统能够在执行阶段对零件的实际位姿进行在线感知与识别。Carlisle等人(1994)通过相机获取零件图像并估计其初始空间姿态；Mirtich等人(1996)通过对零件在随机掉落过程中的运动与接触行为建模，预测其最终稳定姿态的统计分布；Causey等人(1997)通过在水平传送带末端设置透光窗口，并利用背光照明生成零件的轮廓图像，实现对零件姿态的快速、可靠识别。通过将视觉感知结果引入装配流程，系统能够在不改变机械结构的前提下，根据零件的实际位姿动态调整后续抓取与装配动作，从而支持多品种、小批量生产场景下的高效运行。

然而，上述基于视觉感知的研究都遵循“规划然后控制”方案，其成功执行依赖于操作轨迹的完美规划，当环境存在未建模扰动或执行误差时，操作鲁棒性仍然受限。Munoz(1998)提出基于视觉伺服的灵巧操作方法，通过将视觉信息直接融入控制回路动态调整手指运动，从而实现对物体姿态的控制。视觉伺服通常可分为传统视觉伺服与基于学习的视觉伺服两大类。传统视觉伺服方法依赖精确的解析模型建立视觉反馈与机器人运动之间的控制律，实现闭环控制。根据视觉反馈信息的不同，可分为基于位置的视觉伺服、基于图像的视觉伺服以及混合型视觉伺服。然而，传统视觉伺服方法的性能高度依赖系统标定精度、物体几何模型以及手工设计的鲁棒特征，在面对复杂场景和非线性条件下适应性有限。随着机器学习，深度神经网络的发展，基于学习的视觉伺服通过从大量数据中学习高维视觉观测到机器人控制指令的映射，从而减少对显式模型与人工特征的依赖，在非结构化环境和高自由度灵巧操作中展现出更强的泛化能力。

1)基于位置的视觉伺服。基于位置的视觉伺服通过估计目标或末端执行器的三维位姿，并以二者相对位姿误差作为控制变量，在笛卡尔空间中进行运动规划，驱动机器人沿规划轨迹逐步接近目标，从而实现对目标的控制。为提升位姿估计与伺服控制的鲁棒性，Westmore等人(1991)、Wilson(1993)和Wilson等人(1993，1996)通过扩展卡尔曼滤波将二维图像特征递归融合为三维相对位姿，并将估计得到的笛卡尔空间误差直接映射至关节控制，实现了在特征噪声干扰或部分特征丢失情况下的稳定视觉伺服。Yuan(1989)提出无需初值的通用摄影测量封闭解法，仅依赖少量单目视觉特征即可直接恢复目标的六自由度位姿，为快速、稳定的三维定位提供了理论支持。Thuilot等人(2002)提出在线轨迹生成的位置式视觉伺服方法，通过在图像平面中实时规划目标的期望运动路径并同步调节相机位姿，使目标在整个伺服过程中始终保持在视野内，实现了无需离线路径规划的鲁棒定位任务。Allen等人(2002)通过实时估计移动物体的三维运动参数，并结合预测控制与非线性滤波，实现了对动态目标的稳定跟踪与精准抓取。

2)基于图像的视觉伺服。基于图像的视觉伺服不显式估计目标的三维位姿，而是直接在图像空间中选取二维视觉特征(如点、线、轮廓或区域)，以当前特征与期望特征之间的图像误差作为反馈量，通过图像雅可比矩阵将该误差映射为机器人控制输入，从而实现对灵巧手或末端执行器运动的直接控制。针对视觉延迟与动态目标带来的控制不稳定问题，Feddema等人(2002)通过在图像空间内实时生成平滑特征轨迹，实现视觉延迟补偿与连续控制，从而提升眼在手机器人对动态目标的跟踪鲁棒性。Martinet等人(1996a，1996b)通过将几何特征与光学流构建为增广视觉信号，并结合交互矩阵与扩展卡尔曼滤波器，实现在延迟和特征缺失情况下的稳定控制。Papanikolopoulos等人(1993)提出基于SSD光流与多窗口置信度融合的机器人视觉跟踪方法，将图像平面上的目标位移误差直接作为反馈，并结合多种控制方法，实现眼在手上构型的机械臂的实时闭环跟踪。Khadraoui等人(2002)基于相机-激光条纹传感器提取激光投射在球体表面形成的椭圆特征并构建其交互矩阵，将图像误差直接映射为机器人速度控制量，实现无需三维位姿估计的精确定位任务。

3)混合型视觉伺服。混合型视觉伺服综合利用图像空间与三维空间的信息，在统一的控制框架中同时引入图像特征误差与位姿误差。该方法通常通过图像误差建立相机与目标之间的相对位置关系，同时通过三维位姿误差描述二者之间的姿态偏差，

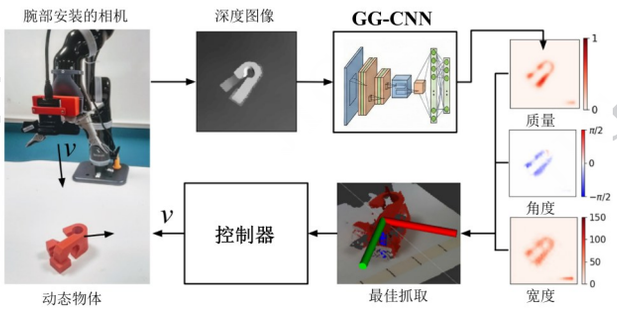
并在此基础上构建混合控制律，驱动机器人向期望位姿收敛。Mehta等人(2016)提出面向农业采摘的鲁棒混合视觉伺服框架，融合图像与位姿空间误差，在标定参数不确定及末端-相机错位条件下，实现稳定趋近与采摘。Hafez等人(2008)提出在线增强的混合视觉伺服方法，将基于图像与基于位置的视觉伺服作为弱算法，通过误差函数动态调节权重，在保证特征可见性的同时兼顾路径可行性，实现复杂运动条件下的鲁棒定位控制。

4)基于学习的视觉伺服。基于学习的视觉伺服方法进一步弱化甚至摆脱了显式的几何建模假设，通过机器学习或深度神经网络从数据中直接学习视觉观测与机器人控制命令之间的映射关系(Ramachandram等，2003)。相比传统方法，该类方法在复杂场景和强不确定性条件下具有更强的适应能力。在抓取任务中，抓取矩形预测通过直接回归图像中的抓取参数估计夹爪位姿。Jiang等人(2011)与Lenz等人(2015)采用两步级联学习策略，先利用轻量级特征快速筛选候选抓取矩形，再通过高级特征进行精确排序，实现了效率与鲁棒性的有效平衡。为进一步提升实时性，Redmon等人(2015)提出单阶段抓取检测网络，直接从RGB-D图像回归抓取参数，显著提升了检测速度。受通用目标检测框架Faster R-CNN(Ren等，2017)的影响，Zhou等人(2018)构建两阶段抓取检测框架，在提高精度的同时增强了鲁棒性。为避免离散候选框带来的搜索复杂性，生成式逐像素抓取方法将每个像素视为潜在抓取点并直接回归抓取质量与参数。Morrison等人(2018)和Gu等人(2019)利用轻量级全卷积网络从图像中逐像素预测抓取参数，实现高频闭环控制和鲁棒抓取。Morrison等人(2020)通过引入反对称抓取位姿表示，实现对整幅图像的密集抓取参数预测，在无需候选采样的情况下完成实时误差补偿，提高了动态抓取稳定性。在复杂非结构化环境中，由于接触动力学的非线性与不连续性，传统视觉伺服难以准确建模抓取过程，因此基于学习的抓取质量预测方法通过数据驱动方式直接评估抓取成功概率，并将其作为反馈信号引入闭环控制。Levine等人(2018)利用大规模自监督数据训练抓取成功预测网络，并结合在线视觉伺服优化夹爪运动，实现无需精确手眼标定的鲁棒闭环抓取。Wang等人(2019)提出基于单应矩阵的平面深度视觉伺服方法，结合轻量级抓取质量网络与并行采样策略，在无相机标定、无显式动作输入的条件下实现实时闭环抓取。此外，端到端方法将视觉感知、策略学习与控制执行统一于单一学习框架中，实现从像素观测到机器人动作的直接映射。Zhang等人(2015)通过深度Q网络(Deep Q-Network)在仿真环境中自主学习目标到达策略，并首次验证了无需任何先验配置即可将像素级观测直接映射为机器人关节控制的端到端可行性。Lampe等人(2013)通过分级Q网络与抓取成功预测器，在零标定、零模型条件下实现从粗定位到精抓取的闭环学习控制。

2.2.2 触觉感知驱动的鲁棒抓取

早在1961年，Ernst(1962)就证明了具备触觉感知能力的计算机控制机械手能够在无视觉条件下找到和堆叠立方体，揭示了触觉在机器人操作中的重要潜力。在早期缺乏视觉或触觉反馈的传统装配系统中，通常要求零件被严格预定向、精确定位，然而实际生产中存在位置偏移、姿态误差及模型失配等不确定性。触觉感知的引入，为应对上述不确定性提供了重要途径。Grossman等人(1975)提出结合倾斜-振动装置与触觉探测的姿态判定方法，实现复杂零件在无视觉条件下的重定向。Fearing(1984)通过触觉感知估计物体表面的法向量信息，从而确定手指的运动方向和抓取位置。Craig等人(1979)通过腕部六维力传感器将细微位姿误差转化为可测可控的力信号，使机器人在无视觉及非高精度定位条件下完成精密装配任务(销-孔、螺母-螺栓及薄片拾取等)，显著提升了触觉在工业操作中的实用价值。近年来研究拓展至基于触觉感知的物体模型重建、基于触觉反馈的伺服控制与基于学习的触觉感知抓取等方向，为机器人在不确定环境中的自主感知与操作能力奠定了关键技术基础。

1)基于触觉感知的物体模型重建。该类方法通过对未知物体进行主动探索在线生成高密度三维触觉点云或概率几何信息，从而重建准确的物体模型，为后续的抓取规划与操控任务提供精确的几何基础。Bierbaum等人(2008)提出基于动态势场的触觉探索方法，通过驱动指尖沿未知物体表面连续滑动实现完整三维形状重建。Sommer等人(2014)建立双臂顺应触觉扫描-抓取框架，通过双臂协同探索与识别实现未知物体的触觉建模与稳定抓取。Jamali等人(2016)提出基于高斯过程分类的主动触觉探索

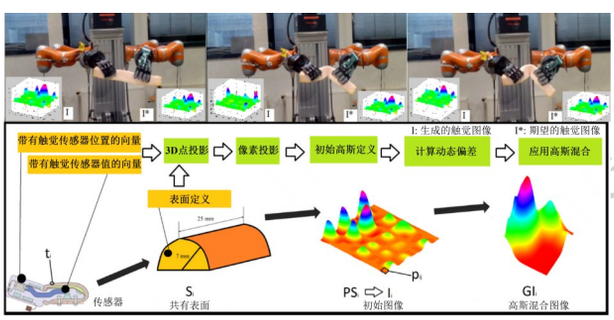


(Morrison等,2018)

图4 基于GG-CNN的实时闭环机器人抓取

Fig. 4 Real-time closed-loop robotic grasping based on GG-CNN (Morrison et al., 2018)

策略,以更少触点实现物体三维形状的高效重建。Pezzementi等人(2011)通过触觉几何建模将识别问题转化为联合"身份-位姿"估计,实现在有限触点条件下未知物体的准确识别与定位。Park等人(2021)提出基于深度神经网络的电阻抗断层触觉重建框架,通过端到端非线性映射与空间敏感度感知损失,在稀疏电极硬件条件下实现高分辨率、高灵敏度与强泛化能力的大面积实时触觉成像。



(Delgado等,2017)

图5 基于高斯图像的可变形物体触觉控制方法

Fig. 5 Tactile control method for deformable objects based on gaussian images (Delgado et al., 2017)

2)基于触觉反馈的伺服控制。基于触觉反馈的伺服控制通过将实时触觉信号直接嵌入机器人控制回路,构建触觉特征(如接触位置、力矢量或图像矩)与机器人运动之间的映射关系,实现对接触状态或交互任务的自适应调控。Zhang等人(2000)系统化地提出了触觉伺服框架,通过建立触觉传感器数据与接触状态之间的映射关系,利用触觉传感器图像的矩特征来计算接触位置、接触力和接触几何形状,并通过显式逆模型和触觉雅可比矩阵两种方法,实现从触觉图像到接触状态的映射,从而实现对机器人接触任务的精确控制。Schill等人(2012)通过支持向量机分类器输出的时间滤波,实现了在抓取尝试过程中对抓取稳定性的连续在线估计与快速稳定。随着传感技术与学习方法的演进,触觉伺服的控制精度、鲁棒性与泛化能力得到显著提升。Lepora等人(2021)将深度学习训练的触觉感知模型直接嵌入控制回路,提出了基于姿态的触觉伺服方法,实现了在复杂三维表面上的精确受控运动。如图5所示,Delgado等人(2017)通过动态高斯混合将异构触觉数据统一为"触觉图像",并以力质心/总力为特征构建指节级PID伺服控制器,实现双手10指并行闭环下的抓握点稳定与滑移抑制。She等人(2021)通过融合高分辨率触觉反馈与最优控制,实现了对多种线径、材质电缆的实时跟踪与精密插入。

3)基于学习的触觉感知抓取。基于学习的触觉感知抓取方法通过数据驱动的方式直接从高维、时序的触觉信号中提取特征,并学习生成适应物体属性与环境变化的抓取策略,从而减少对精确几何模型或复杂物理建模的依赖。Argall等人(2010)通过在线触觉反馈对示范策略进行连续位姿微调,并将修正轨迹加权融入高斯混合回归,实现示范策略的精化与跨任务复用。Romano等人(2011)通过触觉事件驱动的抓取状态机实现未知物体的稳定抓取与放置。Chebotar等人(2014)提出基于触觉反馈的模仿-强化学习框架,通过动态运动基元记录人类示范的期望触觉轨迹,并用相对熵策略搜索在线优化触觉耦合权重,使机器人能在环境变化时凭触觉自适应调整力与姿态。Van Hoof等人(2015)提出面向欠驱动柔顺手的手内触觉滚动技能强化学习框架,通过将任务形式化为马尔可夫决策过程并以非参数相对熵策略搜索直接优化触觉-关节混合状态到手指速度的映射,无需物体模型即可在训练后泛化至新圆柱物体。Tian等人(2019)提出深度触觉模型预测控制框架,通过高分辨率触觉传感器采集接触图像,以无监督方式训练深度循环网络预测"动作-未来触觉图",并在线采用交叉熵MPC优化,使得预测图与用户给定的目标触觉图匹配,从而实现物体的精确定位与重定向。Sundaralingam等人(2019)利用三维卷积网络实现高精度接触力估计,并展现出良好的跨任务泛化能力。Zhang等人(2020)利用长短期记忆网络对时序触觉图像进行分类,并设计像素运动网络预测触觉图像序列的未来帧,显著提升了抓取

系统在动态干扰下的鲁棒性。

2.3 H3多模态感知：融合感知的精细操作

H3级灵巧手标志着机器人操作能力从“有限感知的简单抓取”向“融合感知的复杂操作”的深刻跃迁。在H2阶段，系统依赖单一模态（视觉或触觉）感知以支撑简单的抓取或调整任务，其感知闭环通常作用于局部动作层面，对任务的理解与分解能力有限。相比之下，H3级系统的核心特征在于，将视觉、触觉等多源感知信息在不同层级与时间尺度上进行深度协同与融合，以此支撑对复杂、多阶段且具有强接触特性的操作任务（如精密装配、非刚性物体操控、工具使用等）的自主规划与鲁棒执行。在这一范式下，多模态感知不仅提供任务初始化的状态估计与执行过程中的即时反馈，更被系统性地嵌入到任务分解、运动规划与实时控制的整体框架中，使灵巧手能够应对动态环境、模型不确定性以及执行过程中的意外扰动。基于多模态感知在系统中所处的层级与功能，现有H3级方法可分为多模态感知驱动的分层任务规划，视触融合的多模态伺服控制和数据驱动的多模态策略学习三类。

图6 机器人通过单次示范快速学习1000个日常任务（覆盖402种真实世界物体、31类核心技能）（Dreczkowski等，2025）

Fig. 6 Robots rapidly learn 1,000 everyday tasks from a single demonstration, covering 402 real-world objects and 31 core skill categories (Dreczkowski et al., 2025)

2.3.1 多模态驱动的分层任务规划

多模态驱动的分层任务规划通过融合视觉的全局语义理解与触觉的局部物理验证，实现复杂操作任务的分解、规划与闭环执行。这类方法通常利用不同感知模态在不同层级提供互补信息，从而实现从任务意图到可执行动作序列的自动化生成。Nguyen等人（2013）将苹果采摘任务分解为识别、接近、分离等子任务，并协调移动平台和机械臂完成自主作业。Kragic等人（2003）通过在任务的不同阶段（如运输、对准、抓取）按需切换或融合视觉跟踪、位姿估计与触觉闭环等模块，实现了全流程的模态协同与鲁棒控制。Petersson等人（2002）提出“粗到精”模块化集成框架，将定位-识别-跟踪-伺服-抓取等能力按序组合，并在事件驱动的分布式控制架构下实现真实环境中移动抓取操作。Mao等人（2024）提出基于热敏多模态柔性触觉传感器的触觉-视觉分层融合框架，实现家务场景中易碎/易滑物体的精确定位、稳定抓取与分类归置。Cheng等人（2023）提出“接触模式-物体运动-机器人触点”三级分层搜索框架HiDex，通过多层规划实现高效灵巧操作，并且无需训练即可零样本迁移真实机器人。Zhang等人（2025d）采用“身体-手部技能先验+高层潜码调度”的分层强化学习框架，通过调动身体与手部技能实现“走近-抓取-搬运”等多阶段整机灵巧操作。

2.3.2 视触融合的多模态伺服控制

视触融合的多模态伺服控制通过构建视觉与触觉的耦合反馈回路，在接触不确定性高、精度要求严苛的精细操作任务中实现鲁棒、自适应的闭环执行。其核心在于利用视觉提供全局引导与初始位姿估计，同时依赖触觉进行局部微调、接触状态监控与力交互调节，从而补偿单一模态的感知局限与执行误差。Li等人（2015）构建包含关节层、触觉伺服层与视觉伺服层的三层控制架构，通过任务依赖的投影矩阵灵活整合视触信息，实现了从对齐接近、优化接触到抓取操作的全流程自主执行。Ilonen等人（2013）通过迭代扩展卡尔曼滤波器将视觉点云与触

觉接触点进行最优融合，实现了在抓取过程中在线重建更完整、准确的三维物体模型。Taunyazov等人(2020)提出事件驱动的视觉-触觉脉冲神经网络系统，以高能效的方式处理异步传感事件，提升了容器分类与滑移检测的实时性。Li等人(2018)提出基于视触交互的在线参数估计方法，即使在工具活动部件被遮挡时，也能通过紧凑的闭环控制准确估计其运动学参数并完成翻转任务。在抓取稳定性预测方面，Siddiqui等人(2021)结合深度视觉的初始位姿估计与贝叶斯优化驱动的主动触觉探索，通过有限次探索快速预测安全抓取的置信度，并验证了其在复杂形状物体上的有效性。Jara等人(2014)提出基于动态视觉伺服与触觉反馈的最优控制框架，在考虑多指手动力学模型的基础上，实现了对物体位姿与指尖接触力的同时精确控制，完成了需要高精度力位协同的复杂灵巧操作任务。

2.3.3　数据驱动的多模态策略学习

数据驱动的多模态策略学习方法通过统一表征融合视觉、触觉等异构感知信息，基于模仿学习(Zhang等，2023a)或强化学习从机器人交互数据或人类示教数据中学习复杂交互任务的操作策略。该类方法通过构建有效的多模态联合表征进行策略优化，从而在减少对显式物理模型依赖的同时，提升系统在非结构化环境中的泛化与自适应能力。

多模态统一表征学习是多模态策略学习的基础，其目标是将不同模态的高维原始数据映射到统一的低维语义空间，以提取对下游任务有效的共享特征。Yuan等人(2024)提出一种名为“机器人通感”的基于点云的触觉表示方法，将触觉接触点云与视觉点云融合至同一3D空间，实现了跨模态信息的几何对齐，为强化学习策略提供了丰富的空间推理依据。Heng等人(2025)通过交叉注意力编码器融合高分辨率视觉与触觉信号，并通过自回归触觉预测头来学习对未来接触信号的预测，结合课程学习策略逐步细化视觉-触觉的联合潜在表示，显著提升了表征的准确性与鲁棒性。为提升学习效率与泛化性能，Sferrazza等人(2024)提出了掩码多模态学习框架，通过联合训练多模态掩码自编码器与近端策略优化算法，以掩码自监督的方式同时重建视觉像素与触觉力场，从而学习跨模态共享的紧凑表征。该方法使得策略在测试时即使仅部署视觉模态，仍能保留触觉模态带来的性能增益，实现了对新物体与场景的零样本泛化。

多指灵巧手具有高自由度、多关节协同、多个接触点以及非线性摩擦等特点，在操作任务中表现为接触频繁、约束复杂且动力学高度耦合。基于模仿学习的多模态策略能够通过从人类专家演示中学习视觉、触觉等多模态感知与动作之间的映射关系，快速获取复杂操作技能。Li等人(2025)通过沉浸式演示平台收集包含视觉图像、手指关节角度和关节扭矩的人类抓取数据，并利用多头注意力机制融合这些多模态特征，采用行为克隆方法实现了对不同大小、形状和硬度物体的自适应抓取。Hausman等人(2017)提出了一种能够从无标注、非结构化演示中自动分割与模仿技能的生成对抗网络框架，通过联合学习技能分割与策略模仿，提升了机器人在复杂多阶段任务中的学习效率与适应性。

基于强化学习的多模态策略学习方法通过与环境进行持续的自主交互与试错，在不依赖显式物理模型的前提下，以最大化长期累积回报为目标优化控制策略。在面对复杂接触模式与高度非线性动力学时，通过强化学习能够逐步形成鲁棒且自适应的策略，从而有效应对非结构化环境、不确定感知以及任务目标的变化。Wang等人(2022)提出多模态感知引导的强化学习框架D3Grasp，通过构建统一的视觉-触觉多模态表示，并采用训练时可利用特权信息、部署时仅依赖真实感知输入的非对称强化学习策略，使机器人能够以更高的样本效率和更强的泛化能力完成对一般物体及可变形物体的灵巧抓取。Van Hoof等人(2016)提出“自编码器-强化学习”耦合框架：先用去噪/变分自编码器及其动力学变体将高维触觉/视觉观测压缩到3-5维潜空间并强制线性转移，再以潜在特征为状态输入，并采用非参数相对熵策略搜索进行小批量同策略更新，实现了稳定、抗噪的连续非线性策略学习。Garcia-Hernando等人(2020)提出基于残差强化学习与对抗模仿学习的物理感知灵巧操作框架，通过将含噪声的在线手部姿态估计映射为物理可行的虚拟姿态，并以残差方式微调关节角度，实现仅依赖深度传感器的精细手-物交互与野外动作重建。

2.4　H4开放感知：语言引导的任务规划

虽然深度学习驱动的端到端抓取方法显著提升了感知模块的泛化能力，然而其在复杂开放场景下的物体感知与理解仍面临较大局限性，主要表现为

对未见过物体、严重遮挡、光照剧烈变化以及功能性交互语义的捕捉能力不足。近年来，随着大型语言模型（Large Language Models，LLM）、视觉-语言模型（Vision-Language Models，VLM）、多模态大语言模型（Multimodal Large Language Models，MLLMs）和世界模型（World Model）等基础模型所带来的强大先验知识，以及模仿学习、扩散模型（Diffusion Model）在实时端到端控制策略生成方面的快速发展，机器人系统开始逐步实现对未知物体在非结构化环境中的开放式感知与操作能力。在此背景下，形成了以视觉-语言-动作（Vision-Language-Action，VLA）模型（Zhang等，2025a；Zhang等，2025b）为代表的通用场景操作的方案，灵巧操作技术进入H4等级，即开放感知阶段。本节以视觉-语言-动作模型为核心，从开放感知视角，围绕开放感知的三个核心维度展开论述：感知泛化，长序列感知与操作和多模态融合。

2.4.1　面向强泛化的VLA模型

VLA模型的发展经历了从任务特定模型向大规模基础模型的显著演进，这一路径从感知泛化的视角清晰体现了具身智能从封闭场景专用感知向开放世界通用感知的核心转变。早期VLA模型（如RT-1（Brohan等，2023）及早期Diffusion Policy变体（Chi等，2023））通常依赖特定任务或小规模演示轨迹（数百至数千条）进行端到端训练。这些模型在已知分布的封闭环境中能够实现较为可靠的物体识别与场景理解，但感知泛化能力受限，对未见物体、新颖形状、光照剧烈变化、复杂背景干扰、视角偏移及部分遮挡等开放因素高度敏感，容易出现视觉过拟合与语义理解缺失，导致分布外场景下性能急剧下降，难以支撑真实开放环境的鲁棒感知。随着预训练VLM模型（如CLIP、SigLIP、PaliGemma、InternVL系列）的规模化与开源，研究者开始将VLM作为感知骨干，构建大规模VLA模型。这一转变显著提升了感知泛化能力：通过海量互联网规模视觉-语言数据联合预训练，模型继承强大的开放世界视觉表示与跨模态语义对齐能力，从而在零样本或少样本条件下实现对未见物体、复杂背景、动态光照及指令变异的鲁棒理解与适应。代表性工作如OpenVLA（Kim等，2025）基于Prismatic框架构建，并通过大规模轨迹微调，在跨任务、跨场景零样本迁移中展现较强的物体泛化与视觉扰动鲁棒性；π0（Black等，2024）采用Physical Intelligence提出的流匹配架构，并在异构机器人平台上进行训练，显著提升对开放世界视觉变异与分布偏移的感知适应性；以及英伟达灵巧操作模型GR00T N1模型（Bjorck等，2025），借助Open X-Embodiment（O'Neill等，2024）、DROID（Khazatsky等，2024）、BridgeData V2（Walke等，2023）等多模态轨迹与网络知识联合预训练，实现从网络视觉-语言知识到具身感知控制有效迁移，在未见物体交互、动态环境理解、多视角鲁棒性及跨具身泛化方面表现突出。

增强泛化能力的关键驱动因素体现在感知层面包括数据规模与多样性（从单一平台扩展至多具身、多模态、多任务数据集，显著丰富视觉分布覆盖）、高效训练策略（参数高效微调如LoRA/QLoRA、混合专家MoE等机制，在保留预训练视觉表示前提下最小化微调破坏）以及架构优化（从纯MLP视觉头向保留预训练表示的混合设计演进，如部分冻结视觉编码器，有效缓解微调过程中的视觉退化，提升对背景变化、视角偏移及分布偏移的泛化鲁棒性）。尽管大规模VLA在模拟与部分真实场景中已实现感知层面的显著性能提升，但视觉过拟合、实时推理下的分布偏移、长时序感知累积误差以及极端开放场景下的鲁棒性仍是当前主要瓶颈。

2.4.2　面向长期复杂任务规划的VLA模型

复杂场景下的长期任务规划是开放感知能力的重要体现。早期灵巧手操作主要聚焦短时序任务，如单步抓取、姿态调整或物体重定向，通常依赖端到端模仿学习或强化学习策略，任务时长一般在几秒至数十秒，成功率高度依赖训练分布的覆盖度。随着以基础模型为核心的VLA方法的引入（以RT-2（Zitkovich等，2023）、Octo（Ghosh等，2024）、OpenVLA（Kim等，2025）等为代表），机器人开始具备将自然语言指令直接映射为连续动作序列的能力。然而，早期VLA仍以短时序动作为主，难以有效处理需要多步推理、技能链和环境反馈闭环的长时序任务。近年来，通过引入分层规划、相位感知机制、世界模型辅助搜索（如FLIP框架）以及强化学习后训练，VLA模型逐步突破这一瓶颈。例如，Long-VLA（Fan等，2025）通过阶段掩码和子任务分解机制，显著提升了8-15步甚至更长序列任务的成功率；在真实家庭环境中实现了清理厨房、整理床铺等需10分钟以上、涉及多物体交互与全身协调的长时序操作，

标志着从“短动作执行”向“长程任务规划”的本质跃升。进一步地，针对长时序任务固有的非马尔可夫性和历史依赖问题，近期工作引入记忆机制以增强VLA的时序上下文建模能力，典型代表包括MemoryVLA(Shi等，2025)(感知-认知记忆框架，在长时序基准上显著超越基线)、MAP-VLA(记忆增强提示，最高提升25%性能)以及MindExplore等分层架构(支持实时反馈与重规划)。这些进展共同推动VLA从依赖当前观测的短视行为，向具备持久记忆、动态回忆与自适应规划的长期复杂任务能力演进，为具身智能系统在真实开放场景中的可靠部署奠定基础。

2.4.3　多模态融合的VLA模型

VLA模型的模态演进呈现出从早期以视觉为主导的单/双模态控制，向真正多模态深度融合的快速转型。这一演进显著提升了机器人对物理世界的感知精度、鲁棒性以及语义理解能力，并为开放世界感知与跨任务、跨场景泛化提供了坚实基础。

早期VLA模型(如RT-1(Brohan等，2023)、RT-2(Zitkovich等，2023)、Octo(Ghosh等，2024))主要依赖RGB图像与自然语言指令的双模态输入，通过视觉特征与语言token的简单拼接或跨模态注意力生成动作。该架构在静态或低动态场景中表现良好，但在接触密集、严重遮挡或需精确力控的灵巧操作中明显受限，主要因缺失触觉、力觉及本体感知，导致滑动、掉落或过度施力等问题频发。

近年来，多模态深度融合成为VLA主流范式，显著提升开放世界感知与泛化能力。具体而言，早期融合通过独立模态编码器(视觉如SigLIP/DINOv2、触觉专用CNN/Transformer、本体感知MLP等)将图像、深度、触觉阵列(GelSight、BioTac)、指尖力/扭矩、6D位姿等多源输入投影至共享嵌入空间，并在token级别紧密整合，有助于底层跨模态一致性表征与物理交互理解。中间/晚期融合保留模态专用专家，通过跨模态注意力、MoE或动态路由实现按需融合，代表工作包括ChatVLA-2(Zhou等，2025)以及Being-H0/ Being-H0.5(Luo等，2025)系列(借助大规模人类视频预训练，支持跨具身泛化与精细手部建模)。针对灵巧手复杂在手操作任务，系统引入高频触觉皮肤、力/扭矩反馈及指尖6D跟踪，形成“视觉+触觉+力觉”三重闭环，支持实时滑动预测、接触状态辨识、自适应抓取与在手内精细操作(如旋转、对齐、插入)，大幅增强对非结构化动态环境的响应能力。多模态融合的核心优势在于强化开放感知与泛化：触觉-力觉反馈可实时校正视觉漂移，提升物理推理精度；在接触丰富、长时序任务中，展现更强错误恢复、鲁棒性与少样本适应性。这种提升源于更完整的世界模型表征，使VLA从“看懂指令”向“理解并物理交互世界”实现关键跃升，从而在开放环境下的零/少样本泛化表现更为突出。

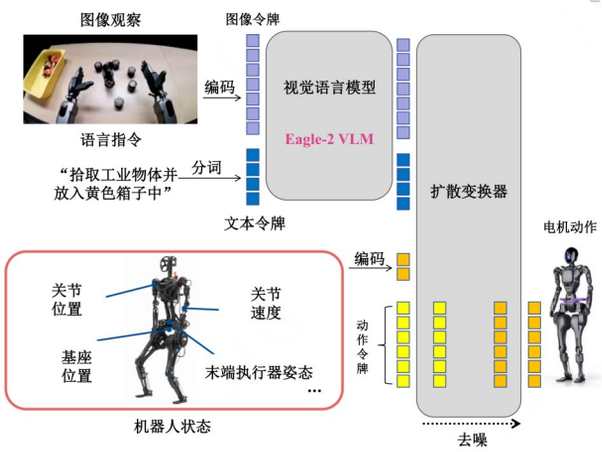


图7　面向灵巧操作的VLA框架：GR00T N1模型架构示意图(Bjorck等，2025)

Fig. 7　VLA framework for dexterous manipulation: GR00T N1 model architecture (Bjorck et al., 2025)

尽管如此，多模态融合亦引入模态对齐难度增大、训练数据异质性、计算资源开销以及噪声/缺失模态敏感性等新挑战。未来研究方向包括自监督多模态预训练以缓解标注依赖、模态dropout与噪声注入增强鲁棒性、统一多传感器世界模型构建，以及探索更高效的跨模态知识蒸馏与参数高效微调机制，以进一步推动VLA向具身智能通用体的演进。

2.5　H5动态感知：自主进化的通用操作

在最高等级的H5灵巧手操作中，系统通过深度集成的多模态动态感知与实时协同机制，实现了与人类操作者之间近乎无缝的物理-认知交互与共同进化。该模式超越传统机器人控制的静态范式，利用高带宽的多源感知融合(视觉、触觉、力反馈、本体觉知等)即时捕捉物理世界的瞬态变化与人类微妙意图，从而使灵巧手能够在复杂、非结构化环境中自适应响应、在线优化并持续演化其行为策略。

这种自主进化能力的核心价值在于：它不再局限于预设任务或特定场景，而是从每一次人与机器

的物理共处中汲取经验，构建出更接近人类直觉的感知-决策-执行闭环。这种渐进式、数据驱动的“类人成长”路径，大幅提升了操作的鲁棒性、灵活性与精细度，从根本上打破了传统机器人“刚性、可预测但缺乏适应性”的局限，使具身智能系统具备了在开放世界中长期自主学习与泛化进化的潜力。

从更大的视野来看，这一技术方向代表着具身智能从“工具”向“伙伴”乃至“共生智能体”的历史性跃迁。它为人类在极端环境（灾后救援、深空探索、微创外科）、日常生活（居家助老助残、精细手工协作）以及未来生产方式（高度个性化、柔性制造）中，提供了全新的物理交互范式。最终，当最高等级的灵巧手能够像人类一样感知、学习、预判并与我们共同创造时，它将深刻重塑人机关系，推动人类从“操控机器”走向“与智能共创世界”，从而为构建更安全、更高效、更具人文温度的未来社会注入决定性动力，真正改变我们与物理世界交互的方式。

3 数据资源与评测基准

灵巧手研究面临的一个共性难题，是高自由度、接触丰富与任务长时序共同带来的数据与评测困境：数据层面，真实交互虽最贴近部署场景并包含噪声、摩擦与机构误差，但采集成本高、效率低且伴随安全与硬件磨损风险；合成数据与仿真平台可在低边际成本下扩展规模与覆盖面，却易受接触模型简化与仿真-现实域差距影响，导致策略迁移性能不稳定。评测层面，不同平台的观测模态、任务定义与成功判据往往不一致，使得结果可比性与误差归因（失败模式诊断）较弱。因此，构建覆盖真实采集、真实-仿真混合扩增与纯合成/仿真的体系化数据管线，并建立可复现、可诊断的评估标准，是推动灵巧手从“可演示”走向“可泛化、可部署”的关键基础工作。按数据来源与构成方式表1汇总了为真实、混合与合成三类典型数据集。

3.1 真实数据集

真实环境灵巧手数据集是指利用物理多指灵巧

表1 灵巧手相关数据集汇总

Table 1 Overview of datasets related to dexterous hands

数据集	年份	数据域	平台 / 模态	数据规模
真实数据集				
ContactDB（Brahmbhatt等，2019）	2019	接触/抓取	RGB-D + 热成像；物体网格；接触区域标注	375K 帧；3,750 网格；50 物体
ContactPose（Brahmbhatt等，2020）	2020	接触+姿态	RGB-D；手姿态；物体姿态；接触标注	2,306 抓取；25 物体；2.9M + 图像
GRAB（Taheri等，2020）	2020	全身 HOI	动捕；3D 身体+手+物体网格；时序接触	10 人；51 物体；4 类意图
InterHand2.6M（Moon等，2020）	2020	手姿态	RGB；3D 双手关键点与姿态	2.6M 标注帧
HO3D v3（Hampali等，2021）	2021	单手 HOI	RGB-D；3D 手姿态；6D 物体姿态	103,462 标注图像
DexYCB（Chao等，2021）	2021	抓取基准	多视角 RGB-D；6D 物体姿态；3D 手姿态	582K 帧；1,000 序列
H2O-3D（Hampali等，2022）	2022	双手 HOI	5 视角 RGB-D；3D 手+物体网格/姿态	571,645 RGB-D 帧
ARCTIC（Fan等，2023）	2023	双手操控	多视角视频；3D 手+物体网格；动态接触	2.1M 视频帧
Open X-Embodiment（O’Neill 等，2024）	2024	跨形态操控	多机器人(含灵巧手)；RGB-D；状态-动作	1M+ 轨迹；160K 任务
TACO（Liu等，2024）	2024	行为语义	多视角视频；手-物 3D 网格/姿态；语义标注	2.5k 动作序列
RealDex（Liu等，2024）	2024	灵巧轨迹	多视角 RGB-D；状态；操作轨迹	2.6k 序列；约 955k 帧

表 1续表

数据集	年份	数据域	平台 / 模态	数据规模
RoboMIND(Wu等,2024)	2024	遥操作	多机器人;观测+状态+语言指令	107k 轨迹;479 任务
OAKINK2(Zhan等,2024)	2024	长时序	多视角图像;手/物体姿态;层级可供性	627 序列;4.01M 帧
DexWild(Tao等,2025)	2025	野外操控	真实环境演示;多视角 RGB-D	9,290 演示;93 环境
AgiBot World(Bu等,2025)	2025	大规模轨迹	双臂操作平台;RGB-D/状态/力觉	1M+ 轨迹;217 任务
混合数据集				
DexMV(Qin等,2022)	2022	示范学习	真实示范 + 仿真扩增;多视角;示范轨迹	5,000 演示
OakInk(Yang等,2022)	2022	HOI 知识	真实多视角采集 + 3D 手-物;可供性知识库	230K 图像;1,800 物体
合成 / 仿真数据集				
D4RL Adroit(Fu等,2020)	2020	动态操控	MuJoCo;状态-动作轨迹;多来源子集	多任务多子集
DexGraspNet(Wang等,2022)	2022	抓取姿态	灵巧手;物体网格;抓取构型	1.32M 抓取;5,355 物体
MultiDex(Li等,2022)	2022	多手型抓取	多手型;抓取姿态/关节角	436k 抓取;58 物体
DexGraspNet 2.0(Zhang等,2024)	2024	杂乱抓取	杂乱场景几何;仿真标注	427M 抓取;1,319 物体
DexGraspAnything(Zhong等,2025)	2025	抓取姿态库	物体网格;物理约束过滤	3.4M+ 抓取;15k+ 物体
GraspXL(Zhang等,2024)	2024	抓取动作	多手型;逐帧手-物轨迹	10M+ 样本
RP1M(Zhao等,2025)	2025	长时序	仿真;状态-动作	约 1M 轨迹
Dex1B(Ye等,2025)	2025	大规模示范	多手型;仿真演示	10^9 frames
DexMimicGen(Jiang等,2025)	2025	双手灵巧	仿真;人示教→生成	21K 轨迹
Dexonomy(Chen等,2025)	2025	抓取体系	taxonomy 标注;仿真抓取	——
BODex(Chen等,2025)	2025	合成抓取	几何 / 接触约束仿真	——
CEDex(Wu等,2025)	2025	跨本体	多本体;统一协议	——
SeqGrasp(Lu等,2025)	2025	序列抓取	多目标;连续任务	——

手在现实场景中采集的交互数据集合,通常包含 RGB/RGB-D 视觉观测、手部关节状态与位姿、物体6D位姿或网格重建信息,并在部分数据集中进一步提供接触区域、力/力矩、触觉阵列或视触觉信号,以及底层控制指令与任务语义标注等。与仿真数据相比,真实数据天然包含传感噪声、硬件公差、摩擦变化与遮挡干扰等复杂因素,采集过程伴随安全风险且标注成本高昂;然而,灵巧手任务高度依赖接触细节与闭环稳定性,真实数据所具备的生态有效性贴近实际部署条件,是评估并提升模型鲁棒性、可迁移性与跨场景泛化能力不可或缺的支撑。

早期真实数据资源主要围绕“手-物几何关系与位姿标注”构建。DexYCB(Chao等,2021)依托多视角RGB-D采集人手抓取桌面物体过程,为多模态感知与状态估计建立统一测试基准;随着研究关注点由静态抓取扩展到时序交互与复杂结构对象,ARCTIC(Fan等,2023)面向双手操控铰接物体提供多视角视频、手-物三维网格及动态接触标注,使接触一致性学习与长时序建模能够在更贴近真实交互的标注体系下开展;TACO(Liu等,2024)将“语义层面的行为结构”引入数据组织与评测协议,以“工具-动作-对象”的组合关系为核心,通过行为语义标注与几何观测的对齐,为任务理解、行为分解及跨组合泛化评测提供更具层次化的基准。与此同时,为降

低高质量示范数据的获取门槛并支撑从示范到策略的学习闭环，DexCap（Wang等，2024）通过可移植的动捕采集与配套模仿学习管线获取人类示范轨迹，使得策略复现与技能迁移能够在更接近真实人类操作分布的条件下进行验证。

综上，真实灵巧手数据集的发展呈现三大趋势：数据维度从单一视觉向多模态融合演进，交互场景从静态抓取向动态操控、双手协作扩展，标注目标从“状态描述”向“行为理解与技能示范”升级。这些数据集不仅为灵巧手感知与控制模型提供了生态有效的测试基准，也为数据驱动的技能迁移与泛化能力提升奠定了核心基础。

3.2 合成数据集

合成数据集是灵巧手研究中与真实采集互补的关键数据来源。通过物理仿真、三维资产建模与程序化场景生成构造手-物交互样本，能够在可控成本下显著扩展物体类别与接触几何的覆盖范围，缓解真实数据在安全风险与硬件磨损等方面的瓶颈。与真实数据相比，合成数据的核心价值在于其高度的可控性与可监督性。研究者可显式设定质量、摩擦、关节柔顺性等生成条件，为策略学习与鲁棒性建模提供可复现的实验基础。依据时序特征，合成数据通常分为静态抓取姿态与动态操作序列两类。

3.2.1 静态抓取姿态

静态抓取姿态数据集聚焦于抓取发生时刻或稳态阶段的手-物相对构型，主要服务于抓取姿态生成、抓取可行性判别、手部姿态估计与接触稳定性分析等任务。此类数据集的共同特征在于：规模可扩展、姿态多样性强、监督信号精确且一致。其生成流程通常依赖于物理仿真与优化搜索，在满足几何可达、碰撞约束与接触稳定性等条件下，批量采样并筛选高质量抓取构型，从而形成覆盖多物体、多握型、多接触模式的标准化样本库。

在代表性数据资源中，DexGraspNet（Wang等，2022）以单一高自由度灵巧手为对象，通过可加速的稳定性估计与仿真验证生成百万级稳定抓取构型库，覆盖大规模日用物体与多类别几何形态，成为“通用物体静态多指抓取”的常用基准之一。围绕“点云观测-抓取生成-抓取执行”的端到端学习链路，UniDexGrasp（Xu等，2023）构建并使用大规模合成抓取姿态作为提案学习与目标条件策略学习的监督来源，推动静态抓取从纯几何采样进一步走向“可执行性导向”的数据组织。针对跨手型泛化难题，GenDexGrasp（Li等，2023）进一步引入 MultiDex 这类“多手型统一标注”的合成抓取资源，在多个常见机械手上以一致的接触稳定性准则生成抓取，从数据层面对“跨具身形态”的可比性与迁移学习提供支撑。

除“桌面单物体”设定外，静态抓取姿态数据向更接近真实部署的分布扩展。DexGraspNet 2.0（Zhang等，2024）将生成对象从孤立物体拓展到杂乱堆叠场景，以场景级局部几何为条件生成并标注海量抓取标签，从而将“多指抓取数据稀缺”问题推进到 cluttered grasping 的可训练尺度。进一步地，DexGrasp Anything（Zhong等，2025）通过大规模对象集合与物理一致性约束构建更广覆盖的抓取姿态库，强调在训练与采样阶段引入物理约束以提升抓取的可行性与稳健性；DexVLG（He等，2025）在大规模合成多指抓取数据与语言标注的支撑下，将抓取构型与对象语义部件及自然语言描述进行联合建模，使静态抓取数据从传统的“几何-稳定性”表征进一步扩展到“部件-功能-指令对齐”的多模态标注空间，从而为语言条件抓取以及 VLA 范式下的可控抓取生成与评测提供了数据与模型基础。与此同时，面向可微接触模拟的工作也开始以“可学习的接触动力学”为目标构建数据资源，Grasp′D-1M（Turpin等，2023）以可微抓取仿真加速优化生成百万级多指抓取样本，为利用梯度信息进行抓取搜索与生成提供了数据支撑；而 MultiGripperGrasp（Casas等，2024）等资源以多抓取器统一验证协议组织大规模抓取样本，体现了“跨末端形态统一评测”的数据化趋势。

3.2.2 动态操作序列

动态操作序列合成数据集以“时序交互过程”为组织核心，显式刻画灵巧手在任务执行过程中的关节/控制序列、物体 6D 位姿与速度演化、接触集合切换以及力/力矩响应等信息，面向长时序闭环稳定性、接触一致性维持与失败恢复等关键问题提供训练与评测基础。与静态抓取相比，动态序列的难点在于交互过程对接触模型、摩擦不确定性与观测噪声更为敏感，数据集不仅需要给出“结果是否成功”，更需要在统一任务定义下提供可复现实验条件，使研究者能够比较不同方法在阶段切换、扰动恢复与长程依赖建模上的能力差异。

动态操作序列数据的可比对研究通常依赖统一

的任务定义与交互接口：研究者在仿真环境中固定一组具有代表性的灵巧操控任务，并在一致的动力学、观测与控制设定下收集大规模交互轨迹，从而使不同学习算法能够在相同条件下开展复现实验与性能比较。Adroit/hand-dapg系列在MuJoCo中定义了Door、Hammer、Pen、Relocate等典型灵巧操控任务，并围绕各任务提供人类示范、专家策略与行为克隆策略等多来源轨迹集合；D4RL（Fu等，2020）进一步将其纳入统一的数据驱动强化学习评测框架，规范数据格式与评测接口，从而支持离线强化学习与模仿学习方法在动态操控场景中的系统比较。RoboHive（Kumar等，2023）在此类思路上更强调环境与控制接口的工程一致性，通过持续维护的任务与数据组件实现跨任务复现实验与可比对评测，为动态序列资源的标准化组织与扩展提供支撑。

面向更复杂对象结构与更强泛化需求，DexArt（Bao等，2023）构建以铰接物体灵巧操控为核心任务基准，在物理仿真中定义多类交互任务并采用类别级泛化设定进行评估。Dex1B（Ye等，2025）面向多种灵巧手形态与大规模对象集合生成十亿量级演示数据，为长时序行为建模与生成式策略学习提供规模化轨迹支撑；DexCanvas（Xu等，2025）以真实动捕示范为种子，在物理一致性约束下进行real-to-sim扩展，构建千小时量级的混合真实-合成交互序列，并提供与接触点与受力剖面一致的标注，使动态序列不仅包含状态-动作过程记录，还具备面向接触与受力机制分析的可用监督信号。

3.3 仿真平台

灵巧手仿真平台是指面向多指灵巧手感知与操控任务构建的虚拟实验环境，通过集成三维场景与对象资产、物理引擎、多指灵巧手/机械臂的运动学与动力学模型，以及数据采集与评估模块，形成可复现、可调控的交互闭环。平台通常显式参数化物体与环境属性（如质量、摩擦系数、碰撞几何、重力等），并支持多源观测与状态输出，包括视觉观测（RGB/RGB-D）、手部关节状态与位姿、物体6D位姿以及接触/力或力矩等仿真信号与底层控制指令，从而为策略学习、对比评测与误差归因提供统一的数据接口。与真实物理环境相比，仿真平台能够减少灵巧手硬件损坏与安全风险、显著降低实验成本，并凭借参数可控、实验可重复与自动标注等优势快速生成大规模多样化训练数据；但由于物理参数近似、接触求解简化与视觉渲染差异等因素，仿真与真实之间仍普遍存在“域差距”，易导致模型或策略迁移性能下降，因此“提升物理接触一致性与场景保真度、弥合域差距”逐渐成为灵巧手仿真平台能力演进的关键目标（Du等，2024）。

早期灵巧手仿真平台主要围绕“基础物理交互与运动学验证”构建，核心目标是为手部姿态控制与简单抓取任务提供低成本测试载体。PyBullet（Coumans等，2016）凭借轻量型物理引擎与实时碰撞检测成为算法验证的高效工具，CoppeliaSim（Rohmer等，2013）通过模块化建模与多模态信号同步为感知控制融合提供了基础支撑。随着研究向高保真交互与跨域迁移演进，仿真平台逐渐从单纯的“实验工具”向“数据生成载体”演进，Isaac Sim（Bonetto等，2023）依托GPU加速渲染与物理仿真技术，实现了大规模并行仿真与多传感器数据同步，显著提升环境与物理建模的真实感；在此基础上，RoboGSim（Li等，2024）创新性地引入3D高斯泼溅（3DGS）技术进行场景重建，有效缩小虚实域差距并实现全流程闭环评估。此外，针对精细操作的接触细节建模需求，DexMOTS（Srinivasan等，2024）优化接触力仿真算法以精准复现复杂接触状态；与此同时，基于仿真平台构建的DexGraspNet（Wang等，2022）等大规模抓取数据集，通过自动化标注与场景多样化配置，为数据驱动的策略学习与跨域迁移提供了高质量的训练资源。

3.4 评估指标与典型基准任务

灵巧手评测应回答两个问题：1）如何度量能力（评估指标），以及2）用什么任务承载能力（典型基准任务）。前者提供跨平台可比的度量坐标，后者提供代表性的任务覆盖，从而避免仅以单一成功率或单一任务定义对方法优劣做出片面判断。

3.4.1 评估指标

为度量灵巧手在不同任务场景下的作业能力，本文将评估指标按“结果完成度”与“过程质量”归纳为若干关键维度。下文对主要评估指标作概述：

任务成功率：衡量“任务到底有没有完成”。把一次次测试（回合/试次）里满足成功条件的比例统计出来。它是最直观的结果型指标。

抓取周期时间：衡量“完成一次抓取动作的节拍有多快”。通常指从手张开-闭合抓住-再回到可进行下一次抓取的状态所需时间。它反映系统的速度

与吞吐能力。

目标位姿误差：衡量“到目标状态有多准”。比较最终达到的位置和姿态与目标值的差距：位置差多少、角度差多少。误差越小，几何精度越高。

归一化任务误差：衡量“在不同尺度/不同任务之间，误差的相对程度”。把原始误差按任务容差、目标尺度或初始—目标距离等做无量纲化，使1cm的偏差在“小物体精密装配”和“大物体搬运”里能放到同一“相对难度/相对偏差”尺度下比较。

接触区域误差：衡量“接触发生在对不对的位置”。看手和物体实际接触的区域，是否与期望/真值接触区域一致。

稳定性与掉落率：稳定性：衡量“抓住以后能不能稳稳保持”，包括是否滑移、是否抖动、是否出现接触丢失/重新抓取等。掉落率：衡量“操作过程中物体掉下来的比例”。它是稳定性最直观、最严重的一类失败结果。

效率与鲁棒性：效率：衡量“用更少代价完成同样任务”，代价可以是时间、步数、能量、动作幅度、接触切换次数等。鲁棒性：衡量“在变化和扰动下还稳不稳”，例如物体质量/摩擦变化、初始位置偏一点、外力干扰、传感噪声变大时，成功率和误差是否明显退化。

3.4.2　典型基准任务

为全面评估灵巧手的感知、规划与控制能力，现有研究与评测基准通常选取一系列具有代表性的具体任务进行测试。这些任务按操作复杂度与接触模态的变化，主要包括以下几类典型基准任务：

稳定抓取与抗扰保持：这是灵巧手最基础的能力基准，旨在验证机械手对多样化对象的适应性。任务通常要求灵巧手对不同几何形状（如YCB物体集中的球体、柱体、不规则物体）、不同材质（刚性、易碎、柔性）及尺寸的对象形成稳定包络或精密捏持。该任务常引入外部冲击、振动或载荷变化等干扰条件，以测试抓取的鲁棒性。

在手重定向与旋转：该任务被视为衡量灵巧操作能力的分水岭，要求手指在不借助外部环境辅助的情况下，仅通过指间协调运动改变物体在手掌中的姿态。这需要灵巧手具备独立的驱动自由度与复杂的多指协同规划能力，能够在保持物体稳定的同时，通过手指的细微运动不断调整物体的三维朝向。

精确拾取与放置：灵巧手需完成从杂乱或非结构化环境中识别并抓取目标、在机械臂运动过程中保持物体稳定搬运、最终在指定位置进行高精度释放的全过程。该任务强调视觉感知与运动控制的闭环性能，要求在抓取完成后能够准确控制释放时的位置与姿态。

接触受限的装配操作：此类任务涉及复杂的接触状态切换与严格的几何约束。灵巧手需在极小的公差范围内精确调整末端姿态，并根据接触力反馈实时调整自身的阻抗特性，以防止在插入或旋拧过程中出现卡死或损坏工件的情况。

工具使用与功能性操作：该任务要求灵巧手不仅能稳定抓取物体，还能根据工具的功能特性（如剪刀的开合、锤子的挥动、笔的书写或电钻的按压）施加特定的操作力。其要求抓取构型必须满足“功能性约束”，即抓取位置不能阻挡工具的工作部位，且必须能够有效地将手部的驱动力传递至工具末端以改变环境状态。

4　挑战与未来方向

4.1　挑战

尽管机器人灵巧手在机械设计、感知融合与学习算法方面已取得显著进展，但其走向通用化、实用化部署仍面临一系列关键挑战。

4.1.1　**数据层面挑战**：真实交互数据的稀缺与仿真-现实鸿沟

高质量、大规模且多样化的真实交互数据获取成本极高，而仿真环境与真实世界之间长期存在的“现实鸿沟”（sim-to-real gap），主要体现在接触力学建模误差、摩擦特性差异、传感器噪声分布偏差以及材料变形表征不足等物理真实性差异，这些问题严重制约了从仿真训练策略向真实硬件的有效迁移。如何构建更高物理保真度的仿真平台，并发展高效的仿真数据生成-真实数据闭环扩增机制，以显著提升数据利用效率并降低昂贵的真实试错成本，成为亟待突破的关键。

4.1.2　**模型层面挑战**：高效鲁棒的多模态联合表征与3D基础模型构建

设计高效、鲁棒且具备一定可解释性的多模态感知-动作联合表征与策略模型，以有效应对真实非结构化环境中物体形态的无限多样性、复杂的接触动力学以及长时序任务规划需求，仍是当前最核心

的理论与技术瓶颈。以视觉-语言-动作为代表的大模型范式虽在任务理解与跨场景泛化上展现出显著潜力，但其决策过程普遍缺乏透明度与可解释性；在动态、高实时性要求的精细灵巧操作场景中，模型规模、推理延迟与实际任务成功率之间存在难以调和的内在矛盾，导致现有架构难以同时满足亚秒级响应、鲁棒物理交互与高精度精细操控的要求。特别值得关注的是，尽管以世界模型为代表的3D基础模型在学习可预测的3D场景动态表征与物理因果推理方面显示出潜力，但当前此类模型仍面临训练数据需求巨大、泛化边界模糊、物理约束难以严格内化以及对复杂多体接触与变形交互的建模精度不足等突出问题，尚未形成足以支撑开放世界灵巧操作的稳定认知基础。

4.1.3 **硬件层面挑战**：高自由度结构与低成本、高可靠性的根本矛盾

灵巧手的机械可靠性、制造成本、功耗以及与上游机械臂-机身的协同控制问题，仍是制约其大规模产业化部署的最主要现实障碍。当前高性能灵巧手普遍追求高自由度（通常20自由度以上）以逼近人类手部的运动灵活性与精细操控能力，但这一设计路径直接导致驱动系统高度复杂化、整体重量显著增加、制造与装配精度要求急剧上升，从而使成本呈指数级飙升；同时，轻量化设计与高精度力/触觉控制之间存在难以兼顾的工程矛盾，高自由度结构往往伴随刚度不足、背隙增大、热累积与疲劳失效风险加剧等问题，进一步削弱系统的长期可靠性和实际可用性。在现有材料、驱动器与集成传感技术水平下，高自由度、高性能与低成本、高可靠性的多目标优化仍处于难以突破的工程死锁状态，成为灵巧手从实验室原型走向真实世界应用的根本性制约因素。

4.2 未来方向

展望未来，机器人灵巧手研究将朝着更高自主性、更强泛化能力以及更紧密人机协同的方向加速演进。其一，感知-决策-执行的深度融合：以世界模型与具身大模型为核心，构建能够内化物理常识、进行因果推理并预测动作长期后果的灵巧操作智能体，实现从单纯技能执行向真正“任务级理解与自主规划”的范式跃迁。其二，数据高效学习与仿真可信化：借助生成式AI大规模合成高质量、多样化训练数据，发展物理约束引导的仿真方法以有效缩减sim-to-real（Zhang等，2022；Zhang等，2025c）差距，并建立覆盖广泛任务与故障模式的标准化评测基准，加速能力迭代与可靠验证。其三，软硬件协同创新与低成本路径：借鉴仿生学原理，设计新一代可变刚度、触觉感知内生化的柔性-刚性混合灵巧手结构；同时，推动开源硬件生态、模块化设计与规模化制造，显著降低研发与部署门槛。其四，面向开放场景的实用化突破：聚焦智能制造（精密装配、柔性生产）、家庭服务（复杂家务、助老助残）、特种作业（太空/深海维修、灾后救援）等高价值应用场景，开展长期、系统的实地部署测试与闭环优化，最终推动灵巧手从实验室原型走向真实世界可靠应用，实现“像人类手一样感知、思考与操作”的通用具身智能目标。

5 结 语

本文从机械结构与硬件范式演进的角度，系统回顾了机器人灵巧手从早期机械夹具到高自由度仿人结构，再到欠驱动、软体及刚柔混合范式，最终向感知-学习一体化智能体的发展脉络。随后，本文提出了以感知能力演进为主线的五级灵巧智能能力分级模型（H1-H5）：从无感知的开环执行（H1），到单模态感知驱动的基础抓取（H2），再到多模态融合的精细操作（H3），进而发展至开放感知下语言引导的任务规划（H4），最终展望了在动态感知中自主进化的通用操作（H5），系统梳理了各层级对应的关键技术、典型方法与应用边界，为理解与评价灵巧操作能力的演进提供了系统性框架。最后，围绕数据驱动的研究范式，本文综述了支撑灵巧手智能发展的关键数据资源与评测基准，涵盖真实与合成数据集、高保真仿真平台以及标准化评估。灵巧手的研究正逐步跨越硬件与控制的传统边界，迈向感知、学习与决策深度融合的新阶段，其进一步发展不仅依赖硬件创新与算法突破，也亟须构建开放、可复现的数据与评测生态。随着多模态感知、仿真计算与自主学习技术的持续进步，机器人灵巧操作有望在更开放、动态的环境中实现真正的智能适应与通用能力，为智能制造、居家服务、医疗辅助等领域带来深远影响。

参考文献（References）

Allen P K, Timcenko A, Yoshimi B and Michelman P. 2002. Automated

tracking and grasping of a moving object with a robotic hand-eye system. IEEE Transactions on Robotics and Automation, 9(2): 152-165[DOI:10.1109/70.238279]

Argall B D, Sauser E L and Billard A G. 2010. Tactile guidance for policy refinement and reuse//Proceedings of the 2010 IEEE 9th International Conference on Development and Learning. Ann Arbor, MI, USA: IEEE: 7-12 [DOI: 10.1109/DEVLRN. 2010. 5578872]

Bao C, Xu H, Qin Y and Wang X. 2023. DexArt: Benchmarking generalizable dexterous manipulation with articulated objects// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 21190-21200[DOI: 10. 1109/CVPR52729.2023.02030]

Bao H, Zheng Y and Liang T J. 2025. Research progress and trends on models and structures of cognitive machines. Journal of Image and Graphics, 30(4):0895-0921 (鲍泓,郑颖,梁天骄. 2025. 认知机器的模型与结构研究进展. 中国图象图形学报, 30(4): 0895-0921)[DOI:10.11834/jig.240108]

Belter J T, Reynolds B C and Dollar A M. 2014. Grasp and force based taxonomy of split-hook prosthetic terminal devices// Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering on Medicine and Biology Society. Chicago, IL, USA: IEEE: 6613-6618[DOI:10.1109/EMBC. 2014.6945144]

Bicchi A. 2002. Hands for dexterous manipulation and robust grasping: A difficult road toward simplicity. IEEE Transactions on Robotics and Automation, 16(6): 652-662[DOI:10.1109/ 70.897777]

Bicchi A. 1995. On the closure properties of robotic grasping. The International Journal of Robotics Research, 14(4): 319-334 [DOI: 10. 1177/027836499501400402]

Bierbaum A, Rambow M, Asfour T and Dillmann R. 2008. A potential field approach to dexterous tactile exploration of unknown objects// Conference on Humanoid Robots. Daejeon, Korea: IEEE: 360-366 [DOI:10.1109/ICHR.2008.4756005]

Birglen L and Schlicht T. 2018. A statistical review of industrial robotic grippers. Robotics and Computer-Integrated Manufacturing, 49: 88-97[DOI:10.1016/j.rcim.2017.05.007]

Bjorck J, Castañeda F, Cherniadev N, Da X, Ding R and Fan L, et al. 2025. GR00T N1: An open foundation model for generalist humanoid robots[EB/OL]. [2026-02-13]. https://arxiv.org/abs/ 2503.14734.pdf

Black K, Brown N, Driess D, Esmail A, Equi M and Finn C. 2024. π0: A vision-language-action flow model for general robot control[EB/OL].[2026-01-25]. https://arxiv.org/abs/2410.24164.pdf

Bonetto E, Xu C and Ahmad A. 2023. GRADE: Generating realistic and dynamic environments for robotics research with Isaac Sim. The International Journal of Robotics Research, 45(2):[DOI: 10. 1177/02783649251346211]

Borst C, Fischer M, Haidacher S, Liu H and Hirzinger G. 2003. DLR hand II: Experiments and experience with an anthropomorphic hand//Proceedings of the 2003 IEEE International Conference on Robotics and Automation. Taipei, Taiwan: IEEE: 702-707[DOI: 10.1109/ROBOT.2003.1241676]

Brahmbhatt S, Ham C, Kemp C C and Hays J. 2019. ContactDB: analyzing and predicting grasp contact via thermal imaging// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE/CVF: 8709-8719 [DOI:10.1109/CVPR.2019.00891]

Brahmbhatt S, Tang C, Twigg C D, Kemp C C and Hays J. 2020. ContactPose: A dataset of grasps with object contact and hand pose// Proceedings of the European Conference on Computer Vision. Glasgow, United Kingdom: Springer: 361-378[DOI:10. 1007/978-3-030-58601-0_22]

Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J and Finn C, et al. 2023. RT-1: Robotics transformer for real-world control at scale//Robotics: Science and Systems XIX. Daegu, Republic of Korea[DOI:10.15607/RSS.2023.XIX.025]

Brost R C. 1988. Automatic grasp planning in the presence of uncertainty. The International Journal of Robotics Research, 7(1): 3-17 [DOI:10.1177/027836498800700101]

Brost R C. 1985. Planning robot grasping motions in the presence of uncertainty. Carnegie-Mellon University, The Robotics Institute.

Bridgwater L B, Ihrke C A, Diftler M A, Abdallah M E, Radford N A and Rogers J M, et al. 2012. The robonaut 2 hand-designed to do work with tools//Conference on Robotics and Automation. Minnesota, USA: IEEE: 3425-3430 [DOI: 10.1109/ICRA. 2012. 6224772]

Butterfaß J, Grebenstein M, Liu H and Hirzinger G. 2001. DLR-Hand II: next generation of a dextrous robot hand// Conference on Robotics and Automation. Seoul, Korea: IEEE: 109-114 [DOI:10.1109/ ROBOT.2001.932538]

Bu Q, Cai J, Chen L, Cui X, Ding Y and Feng S, et al. 2025. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems[EB/OL]. [2026-01-20]. https://arxiv.org/abs/2503.06669.pdf

Carlisle B, Goldberg K, Rao A and Wiegley J. 1994. A pivoting gripper for feeding industrial parts//Proceedings of the 1994 IEEE International Conference on Robotics and Automation. San Diego, CA, USA: IEEE: 1650-1655[DOI:10.1109/ROBOT. 1994.351354]

Casas L F, Khargonkar N, Prabhakaran B and Xiang Y. 2024. MultiGripperGrasp: A dataset for robotic grasping from parallel jaw grippers to dexterous hands//Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems. Abu Dhabi, United Arab Emirates: IEEE/RSJ: 2978-2984 [DOI: 10.1109/ IROS58592.2024.10801708]

Catalano M G, Grioli G, Farnioli E, Serio A, Piazza C and Bicchi A. 2014. Adaptive synergies for the design and control of the Pisa/IIT SoftHand. The International Journal of Robotics Research, 33(5):

768-782[DOI:10.1177/0278364913518998]

Causey G C, Quinn R D, Barendt N A, Sargent D M and Newman W S. 1997. Design of a flexible parts feeding system// Proceedings of the 1997 IEEE International Conference on Robotics and Automation. Albuquerque, NM, USA: IEEE: 1235-1240 [DOI: 10.1109/ROBOT.1997.614306]

Chao Y W, Yang W, Xiang Y, Molchanov P, Handa A and Tremblay J, et al. 2021. DexYCB: a benchmark for capturing hand grasping of objects//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE/CVF: 9044-9053[DOI: 10.1109/CVPR46437.2021.00893]

Chebotar Y, Kroemer O and Peters J. 2014. Learning robot tactile sensing for object manipulation//Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. Chicago, IL, USA: IEEE/RSJ: 3368-3375 [DOI:10.1109/IROS.2014.6943031]

Cheng X, Patil S, Temel Z, Kroemer O and Mason M T. 2024. Enhancing dexterity in robotic manipulation via hierarchical contact exploration. IEEE Robotics and Automation Letters, 9(1): 390-397 [DOI:10.1109/LRA.2023.3333699]

Chen J, Ke Y, Peng L and Wang H. 2025. Dexonomy: Synthesizing All Dexterous Grasp Types in a Grasp Taxonomy[EB/OL]. [2025-12-20]. https://arxiv.org/abs/2504.18829.pdf

Chen J, Ke Y and Wang H. 2025. BODex: Scalable and Efficient Robotic Dexterous Grasp Synthesis Using Bilevel Optimization// International Conference on Robotics and Automation. Piscataway, NJ, USA: IEEE: 1-8[DOI:10.1109/ ICRA55743.2025.11127930]

Chi C, Xu Z, Feng S, Cousineau E, Du Y and Burchfiel B, et al. 2023. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. The International Journal of Robotics Research, 44: 1684-1704 [DOI:10.1177/02783649241273668]

Chi X, Zhang C, Su Y, Dou L, Yang F and Zhao J, et al. 2025. Open TeleDex: A Hardware-Agnostic Teleoperation System for Imitation Learning based Dexterous Manipulation[EB/OL]. [2026-2-10] https://doi.org/10.48550/arXiv.2510.14771

Coumans E and Bai Y. Pybullet, a python module for physics simulation for games, robotics and machine learning[EB/OL]. [2026-2-10]. http://pybullet.org

Craig J J and Raibert M H. 1979. A systematic method of hybrid position/force control of a manipulator//COMPSAC 79. Proceedings. Computer Software and The IEEE Computer Society's Third International Applications Conference. Chicago, IL, USA: IEEE: 446-451 [DOI:10.1109/CMPSAC.1979. 762539]

Dreczkowski K, Vitiello P, Vosylius V and Johns E. 2025. Learning a thousand tasks in a day. Science Robotics, 10(108): eadv7594 [DOI:10.1126/scirobotics.adv7594]

Deimel R and Brock O. 2016. A novel type of compliant and underactuated robotic hand for dexterous grasping. The International Journal of Robotics Research, 2016, 35(1-3): 161-185[DOI: 10.1177/0278364915592961]

DeLaurentis K J, Mavroidis C and Pfeiffer C. 2000. Development of a shape memory alloy actuated robotic hand//Proc. 7th International Conference on New Actuators. Bremen, Germany: 281-284.

Delgado A, Corrales J A, Mezouar Y, Lequievre L, Jara C and Torres F. 2017. Tactile control based on gaussian images and its application in bi-manual manipulation of deformable objects. Robotics and Autonomous Systems, 94: 148-161

Du T, Hu R Z, Liu L B, Yi L and Zhao H. 2024. Research progress in human-like indoor scene interaction. Journal of Image and Graphics, 29(06):1575-1606(杜韬, 胡瑞珍, 刘利斌, 弋力, 赵昊. 2024. 室内场景拟人交互研究进展. 中国图象图形学报, 29(06):1575-1606)[DOI:10.11834/jig.240004]

Ernst H A. 1962. MH-1, A computer-operated mechanical hand// Proceedings of the Spring Joint Computer Conference. New York, NY, USA: ACM: 39-51[DOI:10.1145/1460833.1460839]

Fahlman S E. 1974. A planning system for robot construction tasks. Artificial Intelligence, 5(1): 1-49 [DOI: https://doi. org/10.1016/0004-3702(74)90008-3]

Fan Y, Ding P, Bai S, Tong X, Zhu Y and Lu H, et al. 2025. Long-VLA: unleashing long-horizon capability of vision language action model for robot manipulation//Conference on Robot Learning. Seoul, Korea: PMLR: 2018-2037 [DOI: 10. 48550/arXiv. 2508.19958]

Fan Z, Taheri O, Tzionas D, Kocabas M, Kaufmann M and Black M J, et al. 2023. ARCTIC: A dataset for dexterous bimanual hand-object manipulation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE: 12943-12954[DOI:10.1109/ CVPR52729.2023.01244]

Fearing R S. 1984. Simplified Grasping with Dextrous Robot Hands// American Control Conference, San Diego, CA, USA: IEEE: 32-38[DOI:10.23919/ACC.1984.4788350]

Feddema J T and Mitchell O R. 1989. Vision-guided servoing with feature-based trajectory generation. IEEE Transactions on Robotics and Automation, 5(5): 691-700[DOI:10.1109/70. 88086]

Fu J, Kumar A, Nachum O, Tucker G and Levine S. 2020. D4RL: datasets for deep data-driven reinforcement learning[EB/OL]. [2026-02-01]. https://arxiv.org/abs/2004.07219.pdf

Gama Melo E N, Aviles Sanchez O F and Amaya Hurtado D. 2014. Anthropomorphic robotic hands: a review. Ingeniería y Desarrollo, 32(2): 279-313[DOI:10.14482/inde.32.2.4715]

Garcia-Hernando G, Johns E and Kim T K. 2020. Physics-based dexterous manipulations with estimated hand poses and residual reinforcement learning//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas, NV, USA: IEEE: 9561-9568 [DOI:10.1109/IROS45743.2020.9340 947]

Ghosh D, Walke H R, Pertsch K, Black K, Mees O, Dasari S, et al.

2024. Octo: an open-source generalist robot policy [EB/OL]. [2025-11-23]. https://roboticsproceedings.org/rss20/p090.pdf

Grossman D D and Blasgen M W. 1975. Orienting mechanical parts by computer-controlled manipulator. IEEE Transactions on Systems, Man, And Cybernetics, 5(5): 561-565 [DOI: 10.1109/ TSMC. 1975.5408381]

Gu Q, Su J and Bi X. 2019. Attention Grasping Network: A real-time approach to generating grasp synthesis//2019 IEEE International Conference on Robotics and Biomimetics. Dali, China: IEEE: 3036-3041 [DOI: 10.1109/ROBIO49542.2019.89 61828]

Hafez A H A, Cervera E and Jawahar C V. 2008. Hybrid visual servoing by boosting IBVS and PBVS//2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications. Damascus, Syria: IEEE: 1-6 [DOI: 10.1109/ICTTA. 2008.4530116]

Hampali S, Sarkar S D and Lepetit V. 2021. HO-3D_v3: improving the accuracy of hand-object annotations of the ho-3d dataset [EB/OL]. [2026-01-01]. https://arxiv.org/abs/2107.0088 7.pdf

Hampali S, Sarkar S D, Rad M and Lepetit V. 2022. Keypoint transformer: solving joint identification in challenging hands and object interactions for accurate 3d pose estimation//Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE: 11080-11090 [DOI: 10.1109/ CVPR52688.2022.01 081]

Handa A, Van Wyk K, Yang W, Liang J, Chao Y W and Wan Q, et al. 2020. DexPilot: Vision-based teleoperation of dexterous robotic hand-arm system//2020 IEEE International Conference on Robotics and Automation. Paris, France: IEEE: 9164-9170 [DOI: 10.1109/ ICRA40945.2020.9197124]

Hausman K, Chebotar Y, Schaal S, Sukhatme G and Lim J J. 2017. Multi-modal imitation learning from unstructured demonstrations using generative adversarial nets. Proceedings of the 31st International Conference on Neural Information Processing Systems, Montreal, Canada: NeurIPS: 1235-1245 [DOI: 10.48550/arXiv. 1705. 10479]

Heng L, Geng H, Zhang K, Abbeel P and Malik J. 2025. Vitacformer: learning cross-modal representation for visuo-tactile dexterous manipulation [EB/OL]. [2026-01-10]. https://arxiv.org/abs/2506. 15953.pdf

He J, Li D, Yu X, Qi Z, Zhang W and Chen J, et al. 2025. DexVLG: dexterous vision-language-grasp model at scale [EB/OL]. [2026-01-11]. https://arxiv.org/abs/2507.02747.pdf

He J, Li J, Sun Z, Gao F, Wu Y and Wang Z. 2020. Kinematic design of a serial-parallel hybrid finger mechanism actuated by twisted-and-coiled polymer. Mechanism and Machine Theory, 152: 103951 [DOI: 10.1016/j.mechmachtheory.2020.103951]

Huang J, Chen K, Zhou J, Lin X, Abbeel P and Dou Q, et al. 2025. DIH-Tele: Dexterous in-hand teleoperation framework for learning multiobjects manipulation with tactile sensing. IEEE/ASME Transactions on Mechatronics, 30(5): 3840-3851 [DOI: 10.1109/ TMECH.2025.3532653]

Hu Z, Zhou C, Li J and Hu Q. 2023. Design of a compact anthropomorphic robotic hand with hybrid linkage and direct actuation//International Conference on Intelligent Robotics and Applications. Hangzhou, China: Springer Nature Singapore: 322-332 [DOI: 10.1007/ 978-981-99-6492-5_28]

Ilievski F, Mazzeo A D, Shepherd R F, Chen X and Whitesides G. 2011. Soft robotics for chemists. Angewandte Chemie [DOI: 10. 1002/anie.201006464]

Ilonen J, Bohg J and Kyrki V. 2013. Fusing visual and tactile sensing for 3-D object reconstruction while grasping//2013 IEEE International Conference on Robotics and Automation. Karlsruhe, Germany: IEEE: 3547-3554 [DOI: 10.1109/ICRA. 2013.6631074]

Inouye J M, Kutch J J and Valero-Cuevas F J. 2014. Optimizing the topology of tendon-driven fingers: Rationale, predictions and implementation//The Human Hand as an Inspiration for Robot Hand Development. Cham: Springer International Publishing: 247-266 [DOI: 10.1007/978-3-319-03017-3_12]

Jamali N, Ciliberto C, Rosasco L and Natale L. 2016. Active perception: Building objects' models using tactile exploration//2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids). Cancun, Mexico: IEEE: 179-185 [DOI: 10.1109/ HUMANOIDS.2016.7803275]

Jara C A, Pomares J, Candelas F A and Torres F. 2014. Control framework for dexterous manipulation using dynamic visual servoing and tactile sensors' feedback. Sensors, 14(1): 1787-18 04 [DOI: 10. 3390/s140101787]

Jiang Y, Moseson S and Saxena A. 2011. Efficient grasping from RGBD images: Learning using a new rectangle representation//2011 IEEE International Conference on Robotics and Automation. Shanghai, China: IEEE: 3304-3311 [DOI: 10.1109/ICRA.2011.5980145]

Jiang Z, Xie Y, Lin K, Xu Z, Wan W, Mandlekar A, et al. 2025. DexMimicGen: Automated data generation for bimanual dexterous manipulation via imitation learning//2025 IEEE International Conference on Robotics and Automation. Atlanta, GA, USA: IEEE: 16923-16930 [DOI: 10.1109/ICRA55743.2025. 11127809]

Ji Z, Zhu H, Liu H, Liu N, Chen T, Yang Z, et al. 2016. The design and characterization of a flexible tactile sensing array for robot skin. Sensors, 16(12): 2001 [DOI: 10.3390/s16122001]

Kashef S R, Amini S and Akbarzadeh A. 2020. Robotic hand: a review on linkage-driven finger mechanisms of prosthetic hands and evaluation of the performance criteria. Mechanism and Machine Theory, 145: 103677 [DOI: 10.1016/j. mechmachtheory.2019.103677]

Khadraoui D, Motyl G, Martinet P, Gallice J and Chaumette F. 1996. Visual servoing in robotics scheme using a camera/laser-stripe sen-

sor. IEEE Transactions on Robotics and Automation, 12(5): 743-750[DOI:10.1109/70.538978]

Khazatsky A, Pertsch K, Nair S, Balakrishna A, Dasari S, Karamcheti S, et al. 2024. DROID: a large-scale in-the-wild robot manipulation dataset[EB/OL]. [2026-02-12]. https://arxiv. org/abs/2403.12945.pdf

Kim B, Jeong U and Cho K J. 2025. Dual-tendon routing: Tendon routing for under-actuated tendon-driven soft hand-wearable robot. IEEE Robotics and Automation Letters, 10(4): 3612- 3619[DOI: 10.1109/LRA.2025.3544056]

Kim M J, Pertsch K, Karamcheti S, Xiao T, Balakrishna A and Nair S, et al. 2025. OpenVLA: An open-source Vision-Language-Action model//Conference on Robot Learning. Seoul, Korea: PMLR: 2679-2713 [DOI:10.48550/ arXiv.2406.09246]

Kragic D and Christensen H I. 2003. A framework for visual servoing// International Conference on Computer Vision Systems. Berlin, Heidelberg: Springer Berlin Heidelberg: 345-354 [DOI: 10.1007/3-540-36592-3_33]

Kumar V, Shah R, Zhou G, Moens V, Caggiano V, Gupta A, et al. 2023. RoboHive: A unified framework for robot learning// Advances In Neural Information Processing Systems, 36: 44323-44340[DOI:10.48550/arXiv.2310.06828]

Laliberté T, Birglen L and Gosselin C. 2002. Underactuation in robotic grasping hands. Machine Intelligence & Robotic Control, 4(3): 1-11.

Lampe T and Riedmiller M. 2013. Acquiring visual servoing reaching and grasping skills using neural reinforcement learning//The 2013 International Joint Conference on Neural Networks. Dallas, TX, USA: IEEE: 1-8[DOI:10.1109/IJCNN. 2013.6707053]

Leddy M T, Belter J T, Gemmell K D and Dollar A M. 2015. Lightweight custom composite prosthetic components using an additive manufacturing-based molding technique//2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Milan, Italy: IEEE: 4797-4802 [DOI: 10.1109/EMBC.2015.7319467]

Lenz I, Lee H and Saxena A. 2015. Deep learning for detecting robotic grasps. The International Journal of Robotics Research, 34(4-5): 705-724[DOI:https://doi.org/10.1177/0278364914549607]

Lepora N F and Lloyd J. 2021. Pose-based tactile servoing: controlled soft touch using deep learning. IEEE Robotics & Automation Magazine, 28(4): 43-55 [DOI:10.1109/MRA.2021. 3096141]

Levine S, Pastor P, Krizhevsky A, Ibarz J and Quillen D. 2018. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. The International Journal of Robotics Research, 37(4-5): 421-436[DOI: https:// doi.org/10.1177/0278364917710318:]

Lieberman L I and Wesley M A. 1977. AUTOPASS: An automatic programming system for computer controlled mechanical assembly. IBM Journal of Research and Development, 21(4): 321-333 [DOI: 10.1147/rd.214.0321]

Liu Y, Yang H, Si X, Liu L, Li Z and Zhang Y, et al. 2024. TACO: Benchmarking generalizable bimanual tool-action-object understanding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, United States: IEEE: 21740-21751[DOI:10.1109/CVPR52733.2024. 02054]

Liu Y, Yang Y, Wang Y, Wu X, Wang J and Yao Y, et al. 2024. RealDex: towards human-like grasping for robotic dexterous hand// Thirty-Third International Joint Conference on Artificial Intelligence. Jeju, Korea: IJCAI: 6859-6867 [DOI: 10.24963/ ijcai. 2024/758]

Li H, Ford C J, Lu C, Lin Y, Bianchi M and Catalano M G, et al. 2024. Tactile SoftHand-A: 3D-printed, tactile, highly underactuated, anthropomorphic robot hand with an antagonistic tendon mechanism. The International Journal of Robotics Research[DOI: 10.48550/arXiv.2406.12731]

Li P, Liu T, Li Y, Geng Y, Zhu Y and Yang Y, et al. 2023. GenDexGrasp: generalizable dexterous grasping//International Conference on Robotics and Automation. London, UK: IEEE: 8068-8074 [DOI:10.1109/ICRA48891.2023.10160667]

Li Q, Haschke R and Ritter H. 2015. A visuo-tactile control framework for manipulation and exploration of unknown objects//2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids). Seoul, Korea (South): IEEE: 610-615 [DOI: 10.1109/HUMANOIDS.2015.7363434]

Li Q, Ückermann A, Haschke R and Ritter H. 2018. Estimating an articulated tool′s kinematics via visuo-tactile based robotic interactive manipulation//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, Spain: IEEE: 6938-6944 [DOI: 10.1109/IROS.2018.8594295]

Li S, Ma X, Liang H, Görner M, Ruppel P and Fang B, et al. 2019. Vision-based teleoperation of shadow dexterous hand using end-to-end deep neural network//2019 International Conference on Robotics and Automation. Montreal, QC, Canada: IEEE: 416-422 [DOI: 10.1109/ICRA.2019.8794277]

Li X, Li J, Zhang Z, Zhang R, Jia F and Wang T, et al. 2024. RoboGSim: a Real2Sim2Real robotic gaussian splatting simulator[EB/OL]. [2026-02-12]. https://arxiv.org/abs/2411. 11839.pdf

Li Y, Guo C, Ren J, Chen B, Chen C and Zhang H, et al. 2025. SoftGrasp: adaptive grasping for dexterous hand based on multimodal imitation learning. Biomimetic Intelligence and Robotics, 5(2): 100217[DOI:https://doi.org/10.1016/j.birob. 2025.100217]

Li Y, Wang P, Li R, Tao M, Liu Z and Qiao H. 2022. A survey of multifingered robotic manipulation: Biological results, structural evolvements, and learning methods. Frontiers in Neurorobotics, 16: 843267 [DOI:https://doi.org/10.3389/fnbot. 2022.843267]

Luo H, Feng Y, Zhang W, Zheng S, Wang Y and Yuan H, et al. 2025. Being-H0: vision-language-action pretraining from large-scale

human videos[EB/OL]. [2026-01-12]. https://arxiv.org/abs/ 2507.15597.pdf

Lu G, Fu S and Xu Y. 2022. Design and experimental research of robot finger sliding tactile sensor based on FBG. Sensors, 22(21): 8390 [DOI:https://doi.org/10.3390/s22218390]

Lu H, Dong Y, Weng Z, Pokorny F, Lundell J and Kragic D. 2025. Grasping a Handful: Sequential Multi-Object Dexterous Grasp Generation. IEEE Robotics and Automation Letters, 10(11): 11880-11887[DOI: 10.1109/LRA.2025.3614051]

Mao Q, Liao Z, Yuan J and Zhu R. 2024. Multimodal tactile sensing fused with vision for dexterous robotic housekeeping. Nature Communications, 15(1): 6871[DOI:10.1038/s41467-024-5126 1-5]

Martinet P, Berry F and Gallice J. 1996a. Use of first derivative of geometric features in visual servoing//Conference on Robotics and Automation. Minneapolis, MN, USA: IEEE: 4: 3413-3419[DOI: 10.1109/ ROBOT.1996.509232]

Martinet P, Gallice J and Djamel K. 1996b. Vision based control law using 3d visual features//World Automation Congress, WAC'96, Robotics and Manufacturing Systems, Montpellier, France: 3: 497-502.

Mehta S S, MacKunis W and Burks T F. 2016. Robust visual servo control in the presence of fruit motion for robotic citrus harvesting. Computers and Electronics in Agriculture, 123: 362-375[DOI:10.1016/j.compag.2016.03.007]

Mirtich B, Zhuang Y, Goldberg K, Craig J, Zanutta R and Carlisle B, et al. 1996. Estimating pose statistics for robotic part feeders//Proceedings of IEEE International Conference on Robotics and Automation. Minneapolis, MN, USA: IEEE: 2: 1140-1146[DOI: 10.1109/ROBOT.1996.506861]

Moon G, Yu S-I, Wen H, Shiratori T and Lee K M. 2020. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image//European Conference on Computer Vision. Glasgow, UK: Springer: 548-564 [DOI: 10.1007/978-3-030- 58565-5_33]

Morrison D, Corke P and Leitner J. 2018. Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach//Robotics: Science and Systems XIV. McKeesport, Pennsylvania, USA: MIT Press: 1-10[DOI:10.48550/arXiv. 1804.05172]

Morrison D, Corke P and Leitner J. 2020. Learning robust, real-time, reactive robotic grasping. The International Journal Of Robotics Research, 39(2-3): 183-201[DOI:10.1177/02783649 19859066]

Mouri T, Kawasaki H, Yoshikawa K, Takai J and Ito S. 2002. Anthropomorphic robot hand: Gifu hand III//Proc. Int. Conf. ICCAS. 1288-1293.

Munoz L A. 1998. Robust dexterous manipulation: a methodology using visual servoing//Conference on Intelligent Robots and Systems: Innovations in Theory, Practice and Applications. Victoria, BC, Canada: IEEE: 1: 292-297[DOI: 10.1109/IROS.1998.724634]

Nguyen T T, Kayacan E, De Baedemaeker J and Saeys W. 2013. Task and motion planning for apple harvesting robot. IFAC Proceedings Volumes, 46 (18): 247-252 [DOI: https://doi. org/ 10.3182/20130828-2-SF-3019.00063]

Odhner L U, Jentoft L P, Claffee M R, Corson N, Tenzer Y and Ma R R, et al. 2014. A compliant, underactuated hand for robust manipulation. The International Journal of Robotics Research, 33 (5): 736-752[DOI: https://doi.org/10.1177/027836491351 4466]

O'Neill A, Rehman A, Maddukuri A, Gupta A, Padalkar A, Lee A, et al. 2024. Open X-Embodiment: Robotic learning datasets and RT-X models: Open X-Embodiment collaboration 0//Conference on Robotics and Automation. Yokohama, Japan: IEEE: 6892-6903 [DOI:10.1109/ICRA57147. 2024.10611477]

Papanikolopoulos N P, Khosla P K and Kanade T. 1993. Visual tracking of a moving target by a camera mounted on a robot: a combination of control and vision. IEEE Transactions on Robotics and Automation, 9(1): 14-35[DOI: 10.1109/70. 210792]

Park H, Park K, Mo S and Kim J. 2021. Deep neural network based electrical impedance tomographic sensing methodology for large-area robotic tactile sensing. IEEE Transactions on Robotics, 37 (5): 1570-1583[DOI:10.1109/TRO.2021. 3060342]

Park J, Chang M, Jung I, Lee H and Cho K. 2024. 3D Printing in the design and fabrication of anthropomorphic hands: a review. Advanced Intelligent Systems, 6 (5): 2300607 [DOI: 10.1002/aisy. 202300607]

Peshkin M and Sanderson A. 1986. Reachable grasps on a polygon: The convex rope algorithm. IEEE Journal on Robotics and Automation, 2(1): 53-58 [DOI: 10.1109/JRA.1986.1087030]

Petersson L, Jensfelt P, Tell D, Strandberg M, Kragic D and Christensen H I. 2002. Systems integration for real-world manipulation tasks//Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292). Washington, DC, USA: IEEE: 3: 2500-2505 [DOI: 10.1109/ROBOT. 2002. 1013607]

Pezzementi Z, Reyda C and Hager G D. 2011. Object mapping, recognition, and localization from tactile geometry//Conference on Robotics and Automation. Shanghai, China: IEEE: 5942-5948[DOI:10.1109/ICRA.2011. 5980363]

Polygerinos P, Wang Z, Galloway K C, Wood R J and Walsh C J. 2015. Soft robotic glove for combined assistance and at-home rehabilitation. Robotics and Autonomous Systems, 73: 135-143 [DOI: https://doi.org/10.1016/j.robot.2014.08.014]

Qiao X, Cheng J, Liu H, Xiao X, Liu Q and Zhou J, et al. 2025. Fibertouch: A novel fiber-optic tactile sensor with deep learning demodulation for dexterous robotic hands. Mechanical Systems and Signal Processing, 238: 113212 [DOI: 10.1016/j. ymssp. 2025. 113212]

Qin Y, Wu Y H, Liu S, Jiang H, Yang R and Fu Y, et al. 2022. DexMV: Imitation learning for dexterous manipulation from human videos//European Conference on Computer Vision. Cham: Springer

Nature Switzerland: 570-587[DOI: 10.1007/ 978-3- 03119842-7_33]

Qin Y, Yang W, Huang B, Van Wyk K, Su H and Wang X, et al. 2023. AnyTeleop: A general vision-based dexterous robot arm-hand teleoperation system//Robotics: Science and Systems. Daegu, Korea: RSS: 5184-5203[DOI:10.15607/RSS.2023.XIX.015]

Ramachandram D and Rajeswari M. 2003. A short review of neural network techniques in visual servoing of robotic manipulators// Malaysia-Japan Seminar on Artificial Intelligence Applications in Industry. Kuala Lumpur, Malaysia: 24-25.

Redmon J and Angelova A. 2015. Real-time grasp detection using convolutional neural networks//2015 IEEE International Conference on Robotics and Automation. Seattle, WA, USA: IEEE: 1316-1322[DOI: 10.1109/ICRA.2015.7139361]

Ren S, He K, Girshick R and Sun J. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, (39) 6: 1137-1149[DOI:10.1109/TPAMI.2016.2577031]

Rohmer E, Singh S P N and Freese M. 2013. V-REP: A versatile and scalable robot simulation framework//Conference on Intelligent Robots and Systems. Tokyo, Japan: IEEE: 1321-1326[DOI: 10.1109/IROS.2013. 6696520]

Romano J M, Hsiao K, Niemeyer G, Chitta S and Kuchenbecker K J. 2011. Human-inspired robotic grasp control with tactile sensing. IEEE Transactions on Robotics, 27(6): 1067-1079[DOI: 10.1109/TRO.2011.2162271]

Sato Y and Koganezawa K. 2016. Five finger robot hand with a planetary gear system//The Proceedings of the Asian Conference on Multi-body Dynamics. The Japan Society of Mechanical Engineers: 38_1289936[DOI:10.1299/jsmeacmd. 2016.8.38_1289936]

Schill J, Laaksonen J, Przybylski M, Kyrki V, Asfour T and Dillmann R, et al. 2012. Learning continuous grasp stability for a humanoid robot hand based on tactile sensing//2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics. Rome, Italy: IEEE: 1901-1906[DOI: 10.1109/BioRob. 2012.6290749]

Schmitz A, Pattacini U, Nori F, Natale L, Metta G and Sandini G. 2010. Design, realization and sensorization of the dexterous iCub hand//Conference on Humanoid Robots. Nashville, TN, USA: IEEE: 186-191[DOI: 10.1109/ICHR.2010.5686825]

Sferrazza C, Seo Y, Liu H, Lee Y and Abbeel P. 2024. The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems. Abu Dhabi, United Arab Emirates: IEEE: 9698-9705[DOI: 10.1109/IROS58592.2024.10802719]

She Y, Wang S, Dong S, Sunil N, Rodriguez A and Adelson E. 2021. Cable manipulation with a tactile-reactive gripper. The International Journal of Robotics Research, 40(12-14): 1385-1401[DOI:https://doi.org/10.1177/02783649211027233]

Shi H, Xie B, Liu Y, Sun L, Liu F and Wang T, et al. 2025. MemoryVLA: Perceptual-cognitive memory in vision-language-action models for robotic manipulation[EB/OL].[2026-02-12]. https://arxiv. org/abs/2508.19236.pdf

Siddiqui M S, Coppola C, Solak G and Jamone L. 2021. Grasp stability prediction for a dexterous robotic hand combining depth vision and haptic bayesian exploration. Frontiers in Robotics and AI, 8: 703869[DOI:https://doi.org/10.3389/frobt. 2021.703869]

Smit G, Bongers R M, Van der Sluis C K and Plettenburg D H. 2012. Efficiency of voluntary opening hand and hook prosthetic devices: 24 years of development? . Journal of Rehabilitation Research and Development, 49(4): 523-534[DOI:10.1682/jrrd. 2011.07.0125]

Smith L and Gasser M. 2005. The development of embodied cognition: Six lessons from babies. Artificial Life, 11(1-2): 13-29[DOI: 10.1162/1064546053278973]

Sommer N, Li M and Billard A. 2014. Bimanual compliant tactile exploration for grasping unknown objects//Conference on Robotics and Automation. Hong Kong, China: IEEE: 6400-6407[DOI:10.1109/ICRA.2014. 6907804]

Srinivasan K, Collins J, Heiden E, Ng I, Bohg J and Garg A. 2024. DexMOTS: Dexterous manipulation with differentiable simulation. International Symposium of Robotics Research.

Sundaralingam B, Lambert A S, Handa A, Boots B, Hermans T and Birchfield S, et al. 2019. Robust learning of tactile force estimation through robot interaction//2019 International Conference on Robotics and Automation. Montreal, QC, Canada: IEEE: 9035-9042[DOI: 10.1109/ICRA.2019. 8793502]

Taheri O, Ghorbani N, Black M J and Tzionas D. 2020. GRAB: A dataset of whole-body human grasping of objects//European Conference on Computer Vision. Cham: Springer International Publishing: 581-600[DOI:10.48550/arXiv.2008.11200]

Tao T, Srirama M K, Liu J J, Shaw K and Pathak D. 2025. DexWild: Dexterous human interactions for in-the-wild robot policies[EB/OL].[2026-02-12]. https://arxiv.org/abs/2505.078 13.pdf

Taunyazov T, Sng W, See H H, Lim B, Kuan J and Ansari A F, et al. 2020. Event-driven visual-tactile sensing and learning for robots[EB/OL].[2026-02-12]. https://arxiv.org/abs/2009.07083. pdf

Thuilot B, Martinet P, Cordesses L and Gallice J. 2002. Position based visual servoing: keeping the object in the field of vision//Proceedings 2002 IEEE International Conference on Robotics and Automation. Washington, DC, USA: IEEE: 2: 1624-1629[DOI: 10.1109/ROBOT.2002.1014775]

Tian S, Ebert F, Jayaraman D, Mudigonda M, Finn C and Calandra R, et al. 2019. Manipulation by Feel: Touch-based control with deep predictive models//2019 International Conference on Robotics and Automation. Montreal, QC, Canada: IEEE: 818-824[DOI: 10.

1109/ICRA.2019.8794219]

Turpin D, Zhong T, Zhang S, Zhu G, Liu J and Singh R, et al. 2023. Fast-Grasp'D: Dexterous multi-finger grasp generation through differentiable simulation//2023 IEEE International Conference on Robotics and Automation. London, UK: IEEE: 8082-8089[DOI: 10.48550/arXiv.2306.08132]

Van Hoof H, Chen N, Karl M, Van Der Smagt P and Peters J. 2016. Stable reinforcement learning with autoencoders for tactile and visual data//2016 IEEE/RSJ International Conference on Intelligent Robots and Systems. Daejeon, Korea (South): IEEE: 3928-3934 [DOI: 10.1109/IROS.2016.7759578]

Van Hoof H, Hermans T, Neumann G, and Peters J. 2015. Learning robot in-hand manipulation with tactile features//Conference on Humanoid Robots. Seoul, Korea (South): IEEE: 121-127[DOI: 10.1109/ HUMANOIDS.2015.7363524]

Walck G, Cupcic U, Duran T O and Perdereau V. 2014. A case study of ROS software re-usability for dexterous in-hand manipulation. J. Software Eng. Robot, 5: 36-47.

Walke H R, Black K, Zhao T Z, Vuong Q, Zheng C and Hansen-Estruch P, et al. 2023. BridgeData V2: A dataset for robot learning at scale//Conference on Robot Learning. Atlanta, Georgia USA: PMLR: 1723-1736[DOI:10.48550/arXiv.2308. 12952]

Wang A S, Zhang W, Troniak D, Liang J and Kroemer O. 2019. Homography-based deep visual servoing methods for planar grasps// 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau, China: IEEE: 6570-6577 [DOI: 10.1109/ IROS40897.2019.8968160]

Wang C, Shi H, Wang W, Zhang R, Fei-Fei L and Liu C K. 2024. DexCap: Scalable and portable mocap data collection system for dexterous manipulation[EB/OL]. [2026-02-12]. https://arxiv. org/abs/2403.07788.pdf

Wang R, Zhang J, Chen J, Xu Y, Li P and Liu T, et al. 2022. DexGraspNet: A large-scale robotic dexterous grasp dataset for general objects based on simulation//Conference on Robotics and Automation. London, United Kingdom: IEEE: 11359-11366 [DOI: 10.1109/ICRA48891.2023. 10160982]

Wang Y, Held D and Erickson Z. 2022. Visual Haptic Reasoning: Estimating contact forces by observing deformable object interactions. IEEE Robotics and Automation Letters, 7(4): 11426-11433[DOI: 10.1109/LRA.2022.3199684]

Wang Y, Wu X, Mei D, Zhu L and Chen J. 2019. Flexible tactile sensor array for distributed tactile sensing and slip detection in robotic hand grasping. Sensors and Actuators A: Physical, 297: 111512 [DOI:10.1016/j.sna.2019.07.036]

Westmore D B and Wilson W J. 1991. Direct dynamic control of a robot using an end-point mounted camera and Kalman filter position estimation//Proceedings. 1991 IEEE International Conference on Robotics and Automation. Sacramento, CA, USA: IEEE Computer Society: 2376-2384[DOI:10.1109/ ROBOT.1991.131759]

Wilson W J, Hulls C C W and Bell G S. 1996. Relative end-effector control using cartesian position based visual servoing. IEEE Transactions on Robotics and Automation, 12(5): 684-696[DOI: 10.1109/70.538974]

Wilson W J. 1993. Visual servo control of robots using Kalman filter estimates of relative pose. IFAC Proceedings Volumes, 26(2): 633-638[DOI:10.1016/S1474-6670(17)48804-5]

Will P M and Grossman D D. 2006. An experimental system for computer controlled mechanical assembly. IEEE Transactions on Computers, 100(9): 879-888[DOI:10.1109/T-C.1975. 224333]

Wolter J D, Volz R A and Woo A C. 1985. Automatic generation of gripping positions. IEEE Transactions on Systems, Man, and Cybernetics, (2): 204-213[DOI:10.1109/TSMC.1985.6313350]

Wu K, Hou C, Liu J, Che Z, Ju X and Yang Z, et al. 2025. RoboMIND: Benchmark on multi-embodiment intelligence normative data for robot manipulation[EB/OL]. [2026-2-10]. https://doi.org/10.48550/arXiv.2412.13877

Wu Z, Potamias R A, Zhang X, Zhang Z, Deng J and Luo S. 2025. CEDex: Cross-Embodiment dexterous grasp generation at scale from human-like contact representations[EB/OL]. [2026-2-10]. https://doi.org/10.48550/arXiv.2509.24661

Xu X, Sun J, Chen S, Ma L, Sun K and Zhao B, et al. 2025. DexCanvas: Bridging Human Demonstrations and Robot Learning for Dexterous Manipulation[EB/OL]. [2026-2-10] https://doi.org/10.48550/arXiv.2510.15786

Xu Y, Wan W, Zhang J, Liu H, Shan Z and Shen H, et al. 2023. UniDexGrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, British Columbia, Canada: IEEE: 4737-4746[DOI:10.48550/arXiv.2303.00938]

Yang L, Li K, Zhan X, Wu F, Xu A and Liu L, et al. 2022. OakInk: A large-scale knowledge repository for understanding hand-object interaction//Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE: 20953-20962[DOI:10.1109/ CVPR 52688.2022.02028]

Ye J, Wang K, Yuan C, Yang R, Li Y and Zhu J, et al. 2025. Dex1B: learning with 1B demonstrations for dexterous manipulation [EB/OL].[2026-2-10]. https://doi.org/10.48550/arXiv.2506.17198

Yuan J S C. 1989. A general photogrammetric method for determining object position and orientation. IEEE Transactions on Robotics and Automation, 5(2): 129-142[DOI:10.1109/ 70.88034]

Yuan Y, Che H, Qin Y, Huang B, Yin Z H and Lee K W, et al. 2024. Robot Synesthesia: In-hand manipulation with visuotactile sensing// 2024 IEEE International Conference on Robotics and Automation. Yokohama, Japan: IEEE, 2024: 6558-6565 [DOI: 10.1109/ ICRA57147.2024.10610532]

Zhang C, Bai W S, Du X, Liu W J, Zhou C H and Qian H. 2023a. Sur-

vey of imitation learning: tradition and new advances. Journal of Image and Graphics, 28(06): 1585-1607 (张超, 白文松, 杜歆, 柳伟杰, 周晨浩, 钱徽. 2023. 模仿学习综述:传统与新进展. 中国图象图形学报, 28(06): 1585-1607) [DOI: 10.11834/jig.230028]

Zhang C, Li M, Chen Y, Yang Z, He B and Li X, et al. 2023b. An anthropomorphic robotic hand with a soft-rigid hybrid structure and positive-negative pneumatic actuation. IEEE Robotics and Automation Letters, 8(7): 4346-4353 [DOI: 10.1109/LRA. 2023. 3280829]

Zhang F, Leitner J, Milford M, Upcroft B and Corke P. 2015. Towards vision-based deep reinforcement learning for robotic motion control//Conference on Robotics and Automation. Canberra, Australia: Australian Robotics and Automation Association: 1-8[DOI: arXiv: 1511.03791]

Zhang H and Chen N N. Control of contact via tactile sensing. 2000. IEEE Transactions on Robotics and Automation, 16(5): 482- 495 [DOI:10.1109/70.880799]

Zhang H, Christen S, Fan Z, Hilliges O and Song J. 2024. GraspXL: Generating grasping motions for diverse objects at scale//European Conference on Computer Vision. Cham: Springer Nature Switzerland: Springer: 386-403[DOI:10.1007/ 978-3-031-73347-5_22]

Zhang H, Liang S, Li M, Tian Y, Ge J, Yu H, et al. 2025a. Vision-Language-Action models: From the early foundations to the state-of-the-art. Acta Automatica Sinica, 51(9): 1922−1950 (张慧, 梁姝彤, 李明轩, 田永林, 葛经纬, 于慧, 等. 2025. 视觉－语言－动作模型综述: 从前史到前沿. 自动化学报), 51(9): 1922-1950 [DOI:10.16383/j.aas.c250417]

Zhang H, Luo G, Li Y and Wang F-Y. 2022. Parallel vision for intelligent transportation systems in metaverse: Challenges, solutions, and potential applications. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 53(6): 3400-3413 [DOI: 10.1109/TSMC.2022.3228314]

Zhang H, Xie D, Liang S, Li M, Jia X and Tian Y, et al. 2025b. Agent evolution under the VLA architecture: From mechanistic construction to application expansion. Science & Technology Review, 43(20): 48-61 (张慧, 谢东锦, 梁姝彤, 李明轩, 贾晓丰, 田永林, 等. 2025. VLA 架构下的智能体演化: 从机理建构到应用拓展. 科技导报), 43(20): 48-61[DOI:10.3981/j.issn. 1000-7857. 2025.10.00077]

Zhang H, Xu Y, Tian Y, Li Y, Falk T H and Wang F-Y. 2025c. Selective Shift: Towards personalized pomain adaptation in multi-agent collaborative perception//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin, Ireland: ACM: 886-895 [DOI:10.1145/3746027.3754723]

Zhang J, Liu H, Li D and Yu X, Geng H and Ding Y, et al. 2024. DexGraspNet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes//Conference on Robot Learning. Munich, Germany, PMLR: 5106-5133 [DOI: 10. 48550/arXiv. 2410.23004]

Zhang L, Tang L and Liu L. 2025d. Hierarchical policy learning for humanoid robots whole-body dexterous manipulation. IFAC-PapersOnLine: 59(20): 2543-2548[DOI: 10.1016/j.ifacol. 2025. 11.541]

Zhang Y, Yuan W, Kan Z and Wang M Y. 2020. Towards learning to detect and predict contact events on vision-based tactile sensors// Conference on Robot Learning. Cambridge, MA, USA: PMLR: 1395-1404[DOI:10.48550/arXiv.1910.03973]

Zhan X, Yang L, Zhao Y, Mao K, Xu H and Lin Z, et al. 2024. OAKINK2: A dataset of bimanual hands-object manipulation in complex task completion//Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE: 445-456 [DOI: 10.1109/CVPR52733. 2024.00050]

Zhao Y, Chen L, Schneider J, Gao Q, Kannala J and Schölkopf B, et al. 2025. RP1M: A large-scale motion dataset for piano playing with bi-manual dexterous robot hands. Conference on Robot Learning. Seoul, Korea: PMLR: 5184-5203 [DOI: 10.48550/ arXiv. 2408.11048]

Zhong Y, Jiang Q, Yu J and Ma Y. 2025. DexGrasp Anything: Towards universal robotic dexterous grasping with physics awareness//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, TN, USA: IEEE: 22584 -22594[DOI:10.1109/CVPR52734.2025.02103]

Zhou X, Fu H, Shentu B, Wang W, Cai S and Bao G. 2024. Design and control of a tendon-driven robotic finger based on grasping task analysis. Biomimetics, 9(6): 370 [DOI: 10.3390/biomimetics9060370]

Zhou X, Lan X, Zhang H, Tian Z, Zhang Y and Zheng N. 2018. Fully convolutional grasp detection network with oriented anchor box// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, Spain: IEEE: 7223-7230[DOI: 10.1109/IROS. 2018.8594116]

Zhou Z, Zhu Y, Liu X, Tang Z, Wen J and Peng Y, et al. 2025. ChatVLA-2: Vision-language-action model with open-world reasoning// The Thirty-ninth Annual Conference on Neural Information Processing Systems. San Diego, CA: NeurIPS [DOI:10.48550/arXiv.2505. 21906]

Zitkovich B, Yu T, Xu S, Xu P, Xiao T and Xia F, et al. 2023. RT-2: Vision-language-action models transfer web knowledge to robotic control//Conference on Robot Learning. Atlanta, Georgia USA: PMLR: 2165-2183[DOI: 10.48550/ARXIV.2307. 15818]

作者简介

梁姝彤,女,硕士研究生,研究方向为多智能体协同和具身智能。E-mail: 24140062@bjtu.edu.cn

谢东锦,男,本科生,研究方向为具身智能。E-mail: 20232501487@stu.xju.edu.cn

李东，男，硕士研究生，主要研究方向为机器人感知与操作。E-mail：doongli@ieee.org
张慧，女，副教授，主要研究方向为多智能体协同、多模态感知、具身智能和平行智能。E-mail：huizhang1@bjtu.edu.cn
贾晓丰，男，教授级高工，研究方向为复杂系统下的数据治理与数据智能。E-mail：jiaxf@jxj.beijing.gov.cn
王飞跃，男，研究员，主要研究方向为智能系统和复杂系统的建模、分析与控制。E-mail：feiyue.wang@ia.ac.cn
李浥东，男，教授，主要研究方向为大数据智能、先进计算、智能交通系统、数据治理与隐私保护。E-mail：ydli@bjtu.edu.cn
李灵犀，男，教授，主要研究方向为复杂系统的建模与控制优化、智能交通系统、平行智能以及人机交互。E-mail：lingxili@purdue.edu