中图法分类号:TP242.6　　**文献标识码**:A　　**文章编号**:1006-8961(2026)06-1911-31

论文引用格式:He Y, Lu H C, Wang D, Li S H, Li Z, Liu Y, Zhao J and Ruan S L. 2026. Vision-language-action models: current developments and frontier advances. Journal of Image and Graphics, 31(6):1911-1941(何友, 卢湖川, 王栋, 李勐辉, 李徵, 刘洋, 赵洁, 阮书岚. 2026. 视觉—语言—动作模型发展现状与前沿进展. 中国图象图形学报, 31(6):1911-1941)[DOI:10.11834/jig.260042]

视觉—语言—动作模型发展现状与前沿进展

何友[1],卢湖川[2],王栋[2*],李勐辉[3],李徵[4],刘洋[5],赵洁[2],阮书岚[4]

1. 清华大学电子工程系,北京　100084; 2. 大连理工大学信息与通信工程学院,大连　116024; 3. 浙江大学信息与电子工程学院,杭州　310007; 4. 清华大学深圳国际研究生院,深圳　518055; 5. 大连理工大学未来技术学院,大连　116024

摘　要:视觉—语言—动作(vision-language-action,VLA)模型是近年多模态具身智能的重要研究方向,通过联合建模视觉观测、语言指令与动作决策,推动了机器人感知与控制范式的更新。随着具身智能大模型的快速发展,VLA在泛化性和鲁棒性上相较于传统控制方案展现出显著优势,并在理解现实物理世界及交互效果方面取得突破性进展。本文系统梳理了VLA模型的发展背景、核心机制与最新进展,重点讨论了跨模态对齐、从感知到行动的因果建模、基于语言的任务条件化以及动作生成等关键技术领域。同时,结合具身思维链、高效VLA、强化学习与跨动作学习等研究方向,综合分析了当前在该领域的最新进展和探索效果;并从仿真环境、真实机器人与人类视频3个维度总结了VLA模型的数据集与评测基准。最后,围绕数据集质量、仿真到现实的迁移以及跨机器人适配性等核心瓶颈,深入讨论了VLA领域面临的挑战。

关键词:视觉语言动作(VLA)模型;具身智能;具身思维链;多模态推理;机器人控制

Vision-language-action models: current developments and frontier advances

He You[1], Lu Huchuan[2], Wang Dong[2*], Li Shaohui[3], Li Zhi[4], Liu Yang[5], Zhao Jie[2], Ruan Shulan[4]

1. Department of Electronic Engineering, Tsinghua University, Beijing　100084, China; 2. School of Information and Communication Engineering, Dalian University of Technology, Dalian　116024, China; 3. College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou　310007, China; 4. Shenzhen International Graduate School, Tsinghua University, Shenzhen　518055, China; 5. School of Future Technology, Dalian University of Technology, Dalian　116024, China

Abstract: Vision-language-action (VLA) models represent a paradigm shift in multimodal artificial intelligence (AI) by unifying visual perception, linguistic comprehension, and motor control into a cohesive computational framework. Traditional robotic control frequently relies on decoupled modules for sensing, planning, and execution. However, such modules frequently fail to generalize across unstructured environments or complex semantic instructions. VLA models address these limitations by co-embedding multimodal input into a unified representation space, leveraging the expansive knowledge within large language models and vision-language models to facilitate zero-shot task execution and robust physical interaction. This study provides a systematic review of the VLA landscape, analyzing technical architectures, training methodologies, and empirical evaluation frameworks. By synthesizing the transition from modular robotics to end-to-end generative controllers, the survey elucidates how large-scale pretraining on diverse Internet data can be effectively transferred to downstream physical tasks. The internal mechanisms of VLA systems center on cross-modal alignment and

收稿日期:2026-01-20;**修回日期**:2026-02-19;**预印本日期**:2026-02-26

*通信作者:王栋　wdice@dlut.edu.cn

基金项目:国家自然科学基金项目(62293540,U23A20384)

Supported by:National Natural Science Foundation of China(62293540,U23A20384)

sequence modeling. By discretizing robotic actions into tokens, researchers regard embodied control as a generative task where the model predicts subsequent motor commands based on high-dimensional visual observations and textual goals. This formulation allows agents to capture temporal dependencies between environmental states and linguistic intents through the self-attention mechanisms of Transformer-based backbones. The survey examines how causal modeling enables robots to anticipate the consequences of their actions, while language-conditioned task formulations allow for the interpretation of diverse natural language instructions without task-specific fine-tuning. We specifically analyze the technical implementation of action tokenization, discussing how continuous joint velocities or end effector poses are mapped into discrete vocabularies that the model can process alongside linguistic tokens. This integration ensures that the reasoning capability of the language component directly informs the low-level motor output. Recent research has introduced several optimization strategies to enhance the operational capability of VLA models. Embodied chain-of-thought reasoning improves long-horizon planning by generating intermediate symbolic or natural language subgoals, increasing success rate and system interpretability. To facilitate real-time deployment on edge hardware, studies have focused on efficiency via model quantization, knowledge distillation, and architectural innovations, such as state-space models and mixture of experts. Furthermore, the integration of reinforcement learning allows pretrained VLA models to adapt to specific physical dynamics through environmental interaction, mitigating the limitations of static imitation learning. Cross-action learning techniques further extend these capabilities by enabling skill transfer across heterogeneous robotic platforms and varied degrees of freedom, effectively creating a shared representation for diverse robotic morphologies. This review also explores the role of auxiliary objectives, such as future image prediction or contrastive alignment, in stabilizing the learning process and improving the visual grounding of linguistic concepts. The data ecosystem for VLA development is categorized into three primary domains. Simulation environments provide scalable platforms for automated data generation by using synthetic supervision and physics-based domain randomization. These platforms enable the collection of millions of trajectories without the risk of hardware damage. However, they necessitate sophisticated techniques to bridge the gap to reality. Real robot repositories, including the Open X-Embodiment dataset, offer high-fidelity demonstrations across diverse hardware but face scaling constraints due to the labor-intensive nature of teleoperation. Human video datasets serve as a massive passive learning source for understanding world physics, object affordances, and task hierarchies without explicit action labels. This review evaluates these resources based on their support for manipulation, navigation, and mobile manipulation tasks, providing a comparative analysis of benchmarks, such as robotics Transformer-1 and virtual manipulator, alongside simulation-to-real evaluation frameworks. We also discuss the importance of data diversity, noting that performance correlates strongly with the variety of objects, environments, and camera perspectives present during the pretraining phase. Despite recent progress, significant bottlenecks remain in achieving general purpose-embodied intelligence. Data scarcity for high-quality VLA triplets restricts model scaling compared with pure text or image domains. The simulation-to-real gap, which is driven by discrepancies in physical friction and sensor noise, continues to hinder the direct transfer of simulated policies to physical platforms. In addition, cross-robot adaptability and covariate shift pose ongoing challenges to maintaining performance across different kinematics and long-duration tasks. Safety constraints and the lack of transparency in neural controllers also complicate human-robot collaboration. Current models frequently struggle with fine-grained manipulation that requires high-frequency tactile feedback. This limitation arises from the predominantly vision-centric nature of current datasets. Furthermore, the computational cost of running large-scale vision Transformers at frequencies required for stable control remains a barrier for low-latency applications. This study is concluded by identifying future research directions, emphasizing the need for improved physical common sense and data-efficient adaptation to move toward reliable and autonomous embodied agents. We argue that future VLA systems must move beyond simple imitation to include active exploration and self-correction mechanisms. The integration of multimodal feedback, including haptic and auditory signals, is identified as a necessary step for achieving human-level dexterity. Moreover, the development of standardized evaluation protocols that account for success rate and safety metrics will be essential for the field to progress. By addressing these open problems, the robotics community can transition from specialized task-specific agents to general-purpose robots that are capable of assisting in diverse domestic and industrial environments. This roadmap highlights the convergence of generative AI and physical robotics as the primary path toward artificial general intelligence in the physical world.

Key words: vision-language-action (VLA) model; embodied intelligence; embodied chain-of-thought; multimodal reasoning; robot control

0 引 言

在人工智能诞生之初，具身智能(embodied intelligence)思想的萌芽便已初现——智能不应仅依赖程序员的预先设计，而应通过与物理及社会环境的持续交互逐步构建。这一观点与婴儿通过感知和行动认识世界并改造世界的过程高度契合。具身智能强调机器人感知与交互的能力，旨在通过环境交互获取图像、语言和动作等多模态数据，进而学习执行物理任务的能力。计算机视觉、自然语言处理以及机器人学等多个领域，基于具身智能的理念，提出了大量的研究任务。

单模态建模的成熟为具身智能提供了坚实的理论基础，推动了该领域的发展。在自然语言处理(natural language processing，NLP)方面，分布式表示(Hinton，1986)的提出为词嵌入基础提供了坚实的基础，使得计算机能够将词语表示为连续向量，克服了传统统计语言模型的稀疏性问题。随后，Transformer(Vaswani等，2017)基于这一新的表征方式，提出了全新的架构，显著提高了并行处理效率，并有效解决了长距离上下文的建模问题。这些技术进展为后续大语言模型(large language model，LLM)，如BERT (bidirectional encoder representations from Transformers)(Devlin 等，2019)、GPT(generative pre-trained Transformer)(Radford 等，2018)等的训练提供了重要支持。在视觉处理方面，Vision Transformers(Dosovitskiy 等，2021)将 Transformer 模型应用于图像处理，通过捕捉长距离的上下文信息显著提升了特征建模的能力，并能够通过预训练模型在不同任务与数据集之间进行高效的迁移学习。CLIP (contrastive language-image pre-training)(Radford 等，2021)进一步拓展了视觉 Transformer(vision Transformer，ViT)架构，将大规模图像—文本对映射到共享嵌入空间中，实现了零样本和少样本的高效识别与检索。在此基础之上，大规模预训练和少样本学习技术显著提升了视觉—语言模型(vision-language-model，VLM)在跨模态任务中的适应性和泛化能力，以 Flamingo(Alayrac 等，2022)、BLIP(bootstrapping language-image pre-training)(Li 等，2022b)、GIT(generative image-to-text Transformer)(Wang 等，2022)、LLaVA(large language and vision assistant)(Liu 等，2023b)为代表的模型展现了强大的少样本学习能力，推动了具身智能系统在现实环境中的应用。视觉—语言—动作(vision-language-action，VLA)模型是属于具身智能领域的一类多模态大模型。自2023年首次在机器学习领域被系统性提出(Zitkovich等，2023)以来，VLA就受到了广泛的关注，并且在短短两年时间内经历了高速迭代与演化，成为该领域增长最快的研究方向之一。

传统的机器人系统通常依赖独立的感知设备、高度数学化的物理建模、人工设计的行为逻辑或面向特定任务的强化学习(reinforcement learning，RL)方法，该模式在固定场景和特定任务中能有效提高机器人工作效率。然而，随着机器人结构设计泛化性和鲁棒性方面的提升、电池续航能力的增强，以及各类控制模型的不断成熟，现代机器人在机械结构和控制算法上已经具备适应更加复杂的工作环境的能力。因此，如何将视觉感知、语言理解和动作执行的能力整合到一个连贯统一的系统中，已经成为机器人领域面临的关键挑战之一。VLA模型通过构建一个统一的多模态建模框架，将视觉观测、自然语言指令及其他传感器输入进行联合建模，并通过直接动作生成或调用动作专家模块实现控制决策，从而支持机器人在复杂场景下高效完成任务。

与主要侧重数据建模与语义理解的LLM和VLM不同，VLA不仅处理多模态感知信息，更直接面向动作生成与决策执行。其核心目标在于将感知、语言理解与动作规划紧密耦合，使智能体能在物理或虚拟环境中完成复杂任务，强调“感知—认知—行动”的闭环机制。VLA汇集了计算机视觉、NLP、RL、VLM与机器人学等多个方向的研究成果，是当前多模态人工智能与具身智能深度交叉的典型代表。从发展趋势上看，如图1所示，自2023年起，VLA领域发表的论文数(按arxiv提交/发表日期统计)呈井喷式输出。因此，对VLA模型开展系统而全面的综述，对于把握其在具身智能领域的发展脉络与最新进展具有重要意义。围绕VLA模型的理

论建构与系统发展，国内外学界已开始从不同视角对该领域进行阶段性总结与归纳。张慧等人(2025)通过回顾前VLA时代的技术积淀，梳理模块化、端到端和混合3类主流建模范式，分析其结构特点、能力优势与面临的关键挑战。刘国华等人(2025)总结了“跨模态交互—对齐”分析框架，用于刻画感知、语言理解与规划、动作生成以及环境反馈之间的误差传递路径。

本文对VLA模型的研究进展进行系统综述，依次介绍核心机制、前沿研究方向、数据集与评测标准，并在此基础上总结当前面临的主要挑战与未来发展趋势，以全面呈现该领域的发展现状与技术动向。

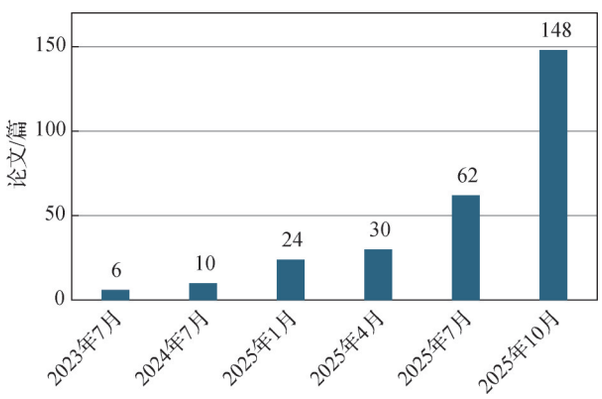

图1　以VLA模型为主题的论文数量

Fig. 1　Number of papers on the VLA model

1　VLA核心机制

本节围绕VLA“感知—认知—行动”的核心闭环，拆解实现多模态协同与动作生成的底层技术支撑。核心机制包含4个相互关联的关键环节：1)跨模态统一表征为不同类型信息建立语义关联，是后续所有处理的基础；2)感知到行动的因果建模构建信息与动作的逻辑链路，保证决策可靠性；3)基于语言的任务条件化将抽象指令转化为可执行目标，明确动作方向；4)动作生成机制则将前期处理结果转化为物理动作，完成与环境的交互。4个环节层层递进、协同作用，共同构成VLA模型的核心功能体系，为后续领域进展的系统梳理奠定理论与方法基础。

1.1　跨模态统一表征

跨模态统一表征是VLA模型实现多模态协同的基础前提，其核心是打破视觉、语言和动作3类异构信息的壁垒，建立稳定的语义关联，为后续因果建模、任务转化与动作生成提供统一的数据支撑。其技术演进围绕“保留模态特异性”与“实现语义通用性”的平衡展开，形成了图2所示的3类典型技术路线，分别适用于不同复杂度的任务场景与资源约束条件。

1.1.1　早期融合

早期融合路线的核心特征是在特征提取阶段对视觉与语言的原始特征进行融合，而不引入复杂的跨模态交互机制。具体而言，视觉特征通常通过卷积神经网络(convolutional neural network，CNN)或视觉Transformer(ViT)模型提取为固定维度向量，语言特征则经双向编码器表示模型(bidirectional encoder representations from transformers，BERT)等编码器完成语义表征。在两者维度对齐后，通过拼接、逐元素相加或加权融合等方式实现信息整合。该直接融合方式在较大程度上保留了各模态原始信息，避免了高层编码可能带来的信息损失；同时，由于无需额外的跨模态交互模块，其计算复杂度较低、推理速度较快，适用于资源受限或模态结构较为简单的基础任

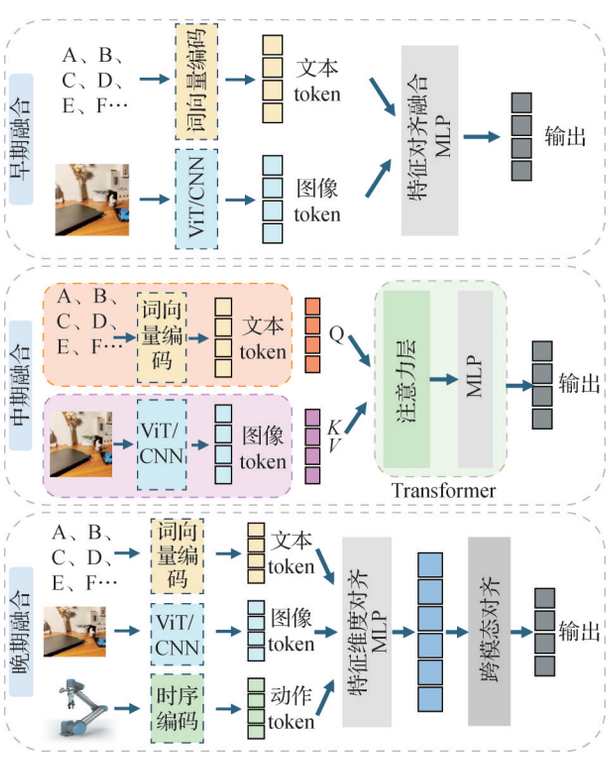

图2　跨模态统一表征的3类典型技术路线

Fig. 2　Three typical technical routes for cross-modal unified characterization

务场景。早期VLA模型如CLIPort(Shridhar等，2022)便采用了这一技术路线，通过CLIP模型提取视觉与语言特征后直接拼接，成功实现了简单指令驱动的物体操作任务，验证了早期融合路线在基础跨模态统一表征场景的可行性。不过，该路线未充分考虑模态间的语义差异，直接融合易产生特征冗余与噪声干扰，因此后续多数VLA模型仅将其作为辅助融合手段，与其他路线配合使用以提升整体性能。

1.1.2　中期融合

中期融合路线以跨模态注意力机制为核心，通过动态交互实现视觉、语言与动作特征的语义对齐与耦合。该路线通常将一种模态特征作为查询(query)，另一种模态特征作为键值(key-value)，通过注意力权重的动态计算，实现语义引导下的特征筛选与加权组合。例如在处理“将书放在左侧书架”的语言指令时，模型会自动强化视觉特征中“左侧书架”区域的权重，同时弱化无关背景信息，使语义意图与视觉场景信息形成精准对应。这种动态交互模式在保留各模态信息完整性的同时，通过语义引导提升了特征表达的针对性，已成为当前VLA模型跨模态统一表征的主流技术路径。跨模态推理框架(cross-modal reasoning architecture，CRA)(Chen等，2025c)设计的跨模态对齐模块(cross-modal alignment module，CMA)通过语言—视觉交叉注意力机制动态调整特征权重，同时结合显式因果干预模块消除模态间的虚假关联，显著提升了弱监督场景下跨模态对齐的鲁棒性；OpenHelix(Cui等，2025)则在中期融合阶段进一步引入三维(three-dimensional，3D)场景表征与本体感受信息，通过多层交叉注意力网络深度整合多模态特征，实现了视觉—语言—动作三者的精细化交互对齐，为后续因果推理与动作生成提供了更可靠的特征基础。

1.1.3　晚期融合

晚期融合路线的核心思路是先保留各模态的特异性特征，再通过细粒度语义匹配对齐到统一语义空间。在具体实现中，视觉特征通常经ViT或DINOv2(distillation with no labels)(Oquab等，2025)等模型完成高层编码，语言特征通过大语言模型提取语义表征，动作特征则经时序编码器处理为结构化特征，3类特征分别完成模态内的表征增强后，通过线性投影层统一维度，再借助跨模态对比学习优化模态间的语义相似度。这种“先独立编码，后统一对齐”的方式有效避免了早期融合的特征冗余问题，同时降低了中期融合的交互复杂度，尤其适用于基于预训练VLM构建的VLA架构，具备更强的抗干扰能力与鲁棒性。FLAVA(foundational language and vision alignment)(Singh等，2022)作为多模态预训练的基础框架，通过跨模态对比学习与掩码重建任务，成功构建了通用多模态语义空间，为VLA模型的晚期融合提供了核心范式，其训练策略被后续众多模型借鉴以提升对齐效率；ChatVLA(Zhou等，2025)则针对晚期融合中可能出现的“虚假遗忘”问题，设计了分阶段对齐训练策略，先在预训练VLM中巩固视觉—语言对齐知识，再通过线性投影层与动作特征进行对齐，从而缓解了动作训练过程对既有对齐信息的覆盖问题，提升了晚期融合的稳定性。

1.2　感知到行动的因果建模

感知到行动的因果建模是连接跨模态表征与实际动作决策的关键环节，其核心是突破传统统计学习对数据相关性的依赖，通过构建结构化因果关系，剥离虚假关联与混杂因素，确保动作生成的可靠性与泛化性。围绕因果关系的建模方式与干预策略，形成了基于因果图的显式建模、基于干预学习的隐式建模以及基于因果链推理的结构化建模3类核心路径，分别从“结构可视化”、“轻量化适配”和“执行落地性”3个维度满足不同场景需求。

1.2.1　基于因果图的显式建模

基于因果图的显式建模路线以结构因果模型(structural causal model，SCM)为理论基础，通过有向无环图清晰表征视觉(V)、语言(L)、动作(A)及混杂变量(confounders，C)之间的因果关系。该路线通常首先利用因果发现算法学习模态间的因果流向，识别核心因果链以及混杂变量的干扰路径，进而针对性地设计多模态去混杂模块，以减弱混杂因素对因果关系的影响。这种显式建模方式使因果结构具备可视化与可解释性，能够帮助模型清晰区分“因果关联”与“相关关联”，为动作生成提供明确的逻辑依据，从而显著提升模型在长尾场景与动态环境中的决策安全性。

1.2.2　基于干预学习的隐式建模

基于干预学习的隐式建模路线不依赖完整因果图的构建，而是通过前门干预、后门干预等因果干预策略，在特征层面或决策层面显式解耦因果信号与

干扰信号。针对视觉模态，该建模方式通过前门干预筛选与动作执行直接相关的核心视觉证据，过滤无关背景噪声与虚假关联特征；针对语言模态，则通过后门干预过滤高频无关词汇与语义偏差，确保语言指令中的核心意图被精准捕捉，最终实现因果效应的无偏估计。这种隐式建模方式无需复杂的因果结构学习过程，计算开销更低，且易于与现有 VLA 架构融合，更易工程化应用，能够快速提升模型对数据偏差的抵抗能力。CRA 提出的显式因果干预模块是该路线的典型代表，其针对视觉—语言推理与动作生成的因果对齐问题，通过前门干预聚焦核心视觉区域，后门干预过滤语言中的高频无关词汇，有效降低了模型对虚假相关性的依赖，并在视频问题定位任务中提升了问答推理与视觉定位之间的因果一致性；该路线通过轻量化的干预策略设计，在不显著增加计算量的前提下，实现了因果建模与跨模态对齐的协同优化，为现有 VLA 模型的性能提升提供了高性价比的解决方案。

1.2.3 基于因果链推理的结构化建模

基于因果链推理的结构化建模路线聚焦于因果关系的可执行化，通过人机混合标注生成与动作对齐的因果推理文本，构建“视觉场景—语言指令—动作输出”的结构化因果链，将抽象的因果关系转化为可执行的推理规则。该路线通常借助 VLM 骨干网络生成与场景适配的因果推理文本，如“前方有行人→减速避让”，并将推理结果作为约束条件参与动作生成过程，从而在一定程度上将动作输出与因果逻辑进行紧密绑定。这种建模方式兼顾了逻辑可解释性与执行一致性，尤其适用于自动驾驶、复杂操纵等高安全需求场景，能够显著提升长尾场景的决策可靠性。Alpamayo-R1（Wang 等，2025b）通过人机混合标注生成驾驶场景的因果推理数据，构建“道路场景特征—驾驶指令—行驶轨迹”的结构化因果链，利用 Cosmos-Reason VLM 骨干网络生成精准的因果推理文本，再通过基于流匹配的动作解码器生成符合因果逻辑的驾驶轨迹，使事故率降低 31.2%；该模型通过多阶段训练策略，先利用监督微调激发 VLM 的因果推理能力，再通过强化学习提升推理结果与动作执行的一致性，充分体现了因果链推理在高风险场景中“逻辑可解释、执行可信赖”的核心优势，为因果建模机制的工程化落地提供了重要参考。

1.3 基于语言的任务条件化

基于语言的任务条件化是连接人类抽象意图与机器具体执行的桥梁，其核心是将自然语言指令转化为模型可执行的结构化条件，为动作生成明确目标导向。该环节以跨模态表征为输入，依托因果建模的逻辑基础，根据任务复杂度与推理需求，形成了直接映射、分层分解和中间语言子目标引导 3 类技术路线，如图 3 所示，分别适配简单任务、复杂长程任务和精细操作任务的处理需求。

1.3.1 直接映射

直接映射型路线充分利用 LLM 强大的语义理解与端到端学习能力，将自然语言指令直接转化为动作生成模块可接收的条件向量，无需额外的任务分解或中间转换过程。该路线的核心实现方式是通过“伪文本词向量”技术将视觉特征转化为 LLM 可处理的输入格式，使视觉、语言与动作信息在 LLM 架构中实现联合建模。模型可直接从语言指令与视觉场景的联合输入中提取任务要素，并输出对应的动作规划。这种端到端的处理流程具备显著的简洁性与高效性，推理速度快，能够快速响应简单或中等复杂度的任务指令，且无需额外的任务分解模块，易于部署在资源受限场景。PaLM-E（pathways language model-embodied）（Driess 等，2023）是该路线的

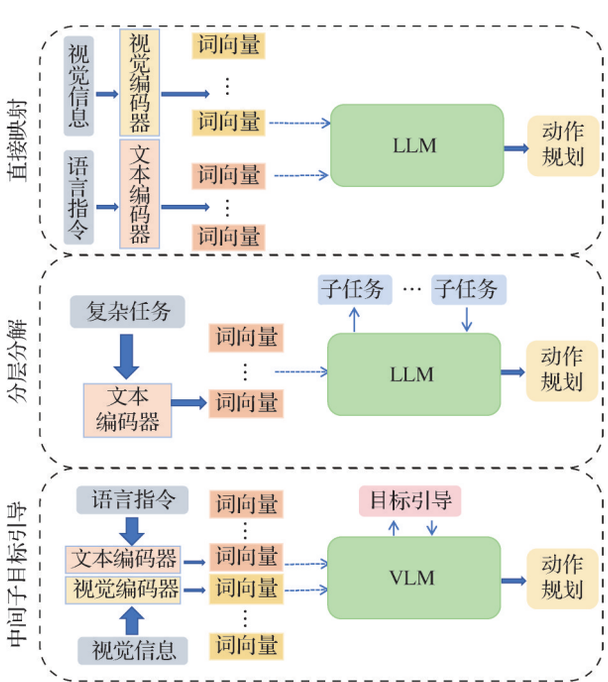

图 3 基于语言的任务条件化 3 类技术路线

Fig. 3 Three typical technical routes for language-based task conditionalization

标志性工作，其以大语言模型为核心架构，通过“伪文本词向量”技术成功将视觉特征融入LLM的输入空间，能够直接从复杂语言指令（如“根据桌上的食谱制作咖啡”）中提取任务要素并生成精准的动作规划，在多物体操纵任务中完成率达82.3%；该模型通过800 k机器人轨迹数据的微调，实现了符号推理与物理交互的深度结合，不仅验证了直接映射路线在端到端VLA架构中的可行性，还为后续模型如何利用LLM的通用能力简化任务条件化流程提供了重要范式。

1.3.2　分层分解

分层分解型路线针对复杂长程任务的解析需求，利用LLM的逻辑推理能力，将原始语言指令分解为可分步执行的子任务序列，每个子任务均包含明确的目标、约束与时序关系，再将这些结构化的子任务逐一转化为对应的动作生成条件。例如面对“整理桌面”这一复杂指令，模型会自动将其分解为“抓取文件→移动至文件夹→抓取杯子→放置至桌面中央”等连续子任务，通过子任务的逐步完成实现整体任务目标，有效降低了复杂任务的执行难度。这种分层处理方式能够显著提升动作序列的逻辑性与连贯性，使模型能够应对“准备早餐”、“修复损坏的玩具”等需要长程规划的任务，拓展了VLA模型的应用场景。ChatGPT for Robotics（Vemprala等，2024）利用LLM的零样本任务分解能力，将开放域语言指令转化为结构化子任务序列，并进一步映射为机器人可执行的动作条件，在无需针对特定场景微调的情况下实现了跨环境任务迁移；π0（Black等，2024）则进一步优化了子任务的分解精度，通过高层VLM将自然语言指令分解为更细粒度的子任务描述，为动作生成提供了更明确的时序约束与空间约束，显著提升了复杂精细操作的任务完成度，深化了分层分解路线在高精度任务中的应用。

1.3.3　中间语言子目标引导

中间语言子目标引导型路线在语言指令与动作生成之间引入了中间语言子目标这一关键环节，通过视觉—语言模型（VLM）生成与当前场景高度适配的细粒度子目标描述，将抽象的原始指令转化为具体、可落地的执行指南，进而指导动作生成过程。这些中间语言子目标通常具备极强的针对性，例如“抓取蓝色方块→移动至绿色平台上方5 cm→释放”，既保留了语言的语义表达能力，又具备动作执行的导向性，能够精准约束动作的空间姿态与时序节奏。该建模方式在一定程度上提升了精细操作任务的执行精度，并增强了动作生成过程的可解释性，使人类能够更直观地理解模型的执行逻辑。VLABench（Zhang等，2025d）构建的大规模评估数据集则专门针对该路线的核心挑战，设计了大量隐含意图指令与长程推理任务，为中间语言子目标的生成质量、场景适配性及任务导向性提供了全面的验证基准，同时也揭示了当前模型在常识转移与复杂指令解析方面的不足，为该路线进一步优化提供了明确方向。

1.4　动作生成机制

动作生成是VLA模型与物理世界交互的最终输出环节，其核心是将经跨模态表征、因果建模和任务条件化处理后的信息，转化为精准、流畅和可行的物理动作，并在性能与实时性之间取得平衡。现有方法围绕时序建模与动作表示展开，形成了自回归预测、基于扩散模型和动作专家3类技术路径，如图4所示，分别从“时序连贯性”、“复杂动作适配性”和“跨平台泛化性”3个维度满足不同场景动作生成需求。

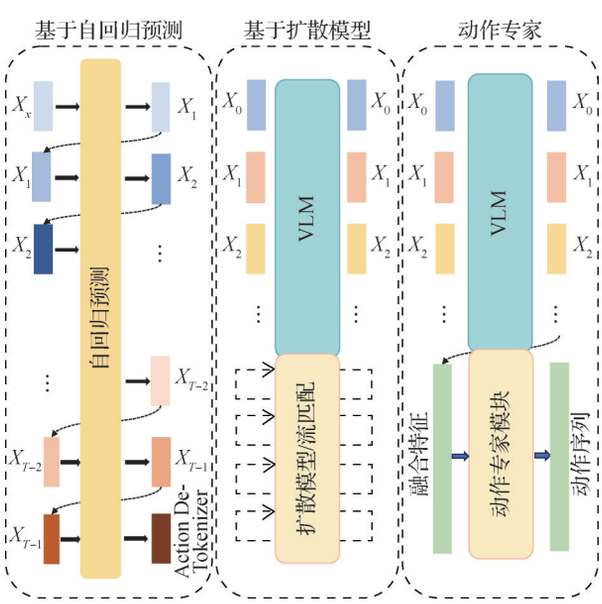

图4　动作生成机制的3类技术路径

Fig. 4　Three typical technical routes for action generation mechanism

1.4.1　自回归预测

自回归预测路线以时序依赖建模为核心，将动作序列编码为离散词向量，并通过Transformer架构逐帧生成动作。该路线通常借助交叉注意力层深度

融合视觉—语言条件信息，以捕捉动作间的长程关联，从而适配连续控制场景的需求。该路线与动作的时序特性较为契合，生成的动作序列在逻辑连贯性与流畅性方面具有一定优势，同时训练与推理流程相对简洁，便于与现有 VLM 架构集成，而无需大规模的模态适配改造。早期奠基性工作 RT-1（Brohan 等，2023）将机器人动作编码为时序词向量序列，通过 Transformer 自回归建模融合多模态条件，在700余种日常操纵任务中实现97%的平均成功率，验证了该路线的基础有效性；RT-2（Zitkovich 等，2023）进一步引入“思维链”机制，通过语言推理增强长程任务的动作规划能力，实现零样本泛化，拓展了应用边界。最新研究 WorldVLA（Cen 等，2025）创新性地将世界模型与动作模型融合，采用独立编码器分别处理图像、文本和动作数据，通过共享词表实现跨模态统一建模。同时，针对自回归生成中的错误累积问题，该方法提出动作注意力掩码策略，在生成当前动作时选择性屏蔽无关历史动作信息。实验结果显示，在动作分块生成任务中，其成功率提升约4%～23%，在一定程度上改善了模型对物理世界动态规律的建模能力与动作生成的鲁棒性。

1.4.2　基于扩散模型

基于扩散策略（diffusion policy）的动作生成路线，通过条件化去噪过程对高维动作分布进行建模，将随机噪声向量逐步优化为符合物理约束的动作序列，无需依赖人工设计的奖励函数，仅通过模仿学习或离线强化学习即可完成训练。该路线的核心优势在于对复杂动作空间的建模能力突出，能够生成多峰动作分布，有效解决行为克隆中的“左右为难”数据冲突问题，同时对环境噪声具备强鲁棒性，适用于机器人灵巧操纵、自动驾驶等高精度场景。Diffusion Policy（Chi 等，2025）首次将扩散模型直接应用于动作分布建模，通过逐步去噪生成高维关节动作序列，在12个任务，4个机器人操作基准上平均提升约46.9%；后续优化工作 Responsive Noise-Relaying Diffusion Policy（Chen 等，2025e）设计噪声中继缓冲器，采用序贯去噪机制，在序列头部生成无噪即时动作、尾部追加带噪动作以保证连贯性，既提升了响应灵敏度，又通过复用去噪步骤加速动作生成，在响应敏感任务中成功率提升18%，同时计算效率较现有加速方法提升6.9%。该路线创新性地通过低步数采样优化（可压缩至5～10步），已逐步突破实时性瓶颈，成为高维动作生成的主流选择。

1.4.3　动作专家

动作专家路线通过在预训练 VLM 骨干网络基础上增设独立的动作头模块，为不同机器人本体（如单臂、双臂、移动操控器等）设计专门的动作生成单元，并结合流匹配技术建模连续动作分布，以提升动作输出的精细度与可控性。该路线的一个重要特征在于将动作生成与多模态理解进行功能解耦：VLM 主要负责语义解析与任务规划，动作专家则专注于物理可行的动作建模。在此框架下，模型在继承 VLM 通用知识的同时，有助于提升动作生成的精度与跨平台泛化能力。典型代表 π0 基于 PaliGemma VLM 构建，通过独立的 Transformer 动作专家模块处理动作生成任务，采用流匹配技术建模连续动作分布，仅需10次去噪迭代即可生成50 Hz的高频动作块，满足灵巧操作需求。其训练过程分为预训练与后训练两阶段，预训练阶段在7种机器人、68个任务的数据上学习通用动作规律，后训练阶段针对特定任务微调，使模型在纸巾更换、抽屉整理等复杂任务中表现优于 OpenVLA（Kim 等，2024）、Octo（Octo Model Team 等，2024）等主流模型；在推理阶段，π0 通过欧拉步进法恢复连续动作序列，并配合三段式注意力掩码策略（图像语言块、自身状态块与动作块的有序交互），以加强动作生成与多模态条件之间的对齐。在跨本体迁移实验中，该方法展现出较好的适应性。

2　VLA 前沿进展

本节概述 VLA 领域前沿进展，涵盖5个主要研究方向，分别是模型架构、具身思维链、高效 VLA、强化学习结合 VLA 以及跨动作学习方法，旨在总结当前 VLA 研究的关键发展与技术趋势。

2.1　VLA 模型架构

视觉—语言—动作（VLA）模型的核心挑战在于如何有效地将视觉感知、语言理解与动作生成3个本质异构的模态统一于一个连贯的计算框架中。这一挑战的复杂性不仅源于各模态在表征空间、时间尺度与信息粒度上的根本差异，更在于机器人操作任务对跨模态推理深度、动作生成精度与实时控制带宽的综合要求。回顾 VLA 架构的演进历程，可以清晰地观察到一条从模块化解耦走向端到端统一、

从浅层融合迈向深度交互的技术主线。

本节系统梳理 VLA 领域的 4 种核心架构范式，它们之间存在着明确的技术演进逻辑与互补关系。后期融合与独立编码（2.1.1 节）代表了 VLA 的技术起点，其通过独立编码器分别处理视觉与语言输入，在策略层进行浅层特征融合。这一范式以其模块化设计与工程简洁性奠定了早期 VLA 研究的基础，但其跨模态交互深度的不足逐渐成为复杂任务中的性能瓶颈。为突破这一限制，规划器—执行器架构（2.1.2 节）引入大型语言模型作为高层语义规划器，将任务理解与动作执行显式解耦，从而在长时序推理与复杂指令理解方面取得显著进展；然而，这种功能分离也带来了规划与执行之间的一致性挑战。针对上述两种范式在跨模态深度融合上的结构性局限，统一多模态序列模型（2.1.3 节）提出了更为彻底的解决方案：将视觉块、语言词向量与动作词向量统一映射至同一序列空间，通过单一 Transformer 实现端到端的联合建模。这一范式在跨模态推理能力与任务泛化性方面展现出显著优势，但自回归解码机制固有的逐步生成特性限制了其在高频控制场景中的适用性。为此，离散动作扩散（2.1.4 节）作为最新涌现的研究方向，通过在离散动作词向量上引入扩散式并行生成机制，在保持统一架构优势的同时有效提升了动作预测的效率与稳定性，代表了 VLA 动作建模从自回归范式向并行化生成的重要演进。

VLA 的 4 种模型架构如图 5 所示。上述 4 种架构并非简单的线性替代关系，而是在不同任务需求与资源约束下各具适用场景。后期融合在资源受限的快速原型开发中仍具价值；规划器—执行器在需要显式可解释性的长任务链中表现优异；统一序列模型在跨任务泛化与复杂语义理解中占据优势；离散扩散则为高精度连续控制提供了新的技术路径。理解这些架构的设计动机、核心机制与适用边界，对于 VLA 系统的合理选型与未来改进至关重要。

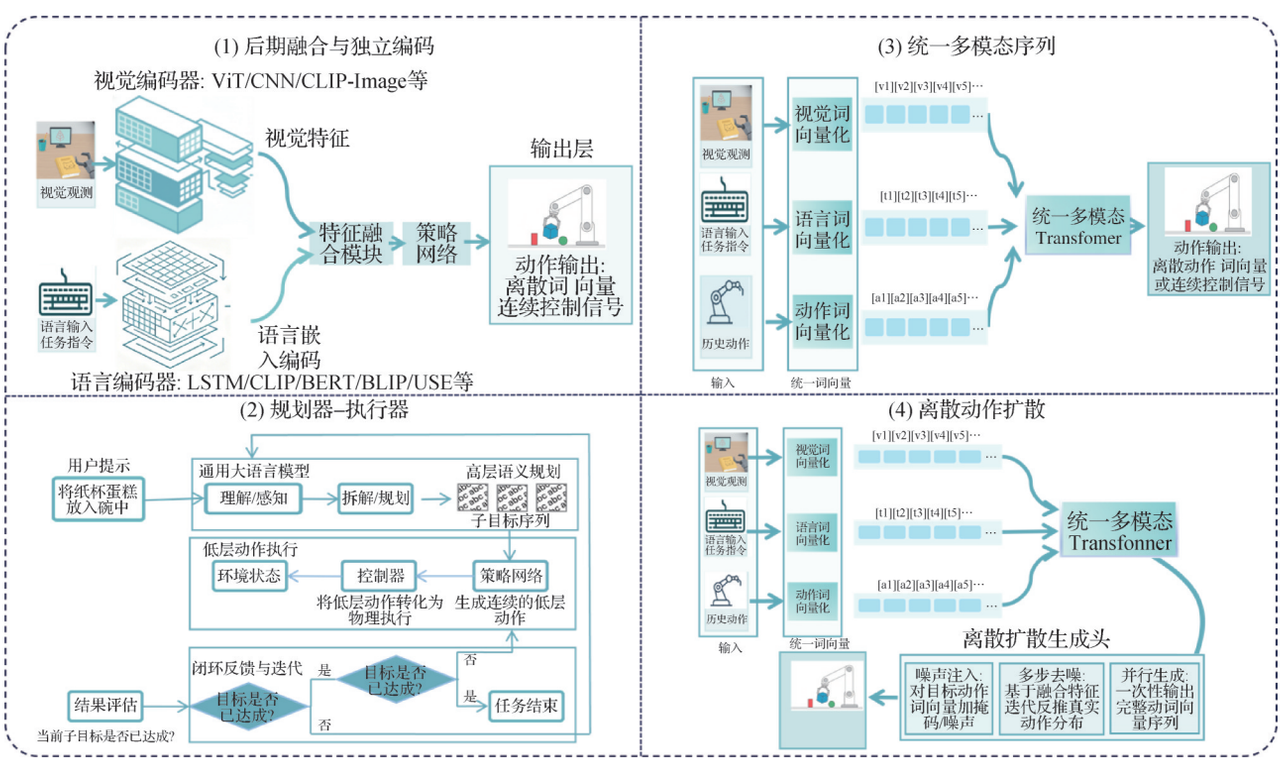

图 5　VLA 的 4 种模型架构示意图

Fig. 5　Illustration of the four VLA model architectures

2.1.1　后期融合与独立编码

在 2020—2023 年间，VLA 模型多采用以后期融合（late fusion）或独立编码为代表的双塔结构。该范式将视觉与语言输入分别通过独立编码器进行处理，例如使用 CNN、ViT 或 CLIP 图像塔提取视觉表征，并使用 Transformer 文本塔进行语言语义建模。两路特征在策略网络或轻量融合层中进行拼接或基于注意力的浅层交互，最终生成动作。这种结构具

有显著的模块化与替换性优势，训练过程稳定且工程实现简单，因此在早期语言条件的机器人操作研究中得到广泛采用。

最早展示该结构有效性的工作包括语言条件化模仿学习（Stepputtis等，2020），其采用“独立视觉编码器 + 文本编码器 + 行为克隆策略融合”的典型流程。随后，大规模预训练视觉—语言模型被逐渐纳入该框架以增强表征能力，其中最具代表性的系统是CLIPort。CLIPort通过CLIP获取稳健的语言嵌入，并构建what/where双路径视觉网络分别负责语义识别与空间定位；二者与语言特征在策略模块中融合，从而实现像素级操作的端到端抓取与放置。其严格区分模态功能、保持数据流向清晰的设计特性，使其成为后期融合范式的经典工作。

在此基础上，PerAct/Perceiver-Actor（Shridhar等，2023）进一步将语言编码与三维视觉体素编码分离处理，再以Perceiver Transformer架构实现跨模态融合并预测离散化6-DoF操作，体现了后期融合在应对高维视觉输入和3D操作任务时的可扩展性。其他工作如BC-Z（Jang等，2022），以及一系列基于CLIP/BLIP特征对齐的机器人策略模型，也普遍遵循“模态独立编码—跨模态对齐—策略融合”的路径。这类方法的结构统一性使研究者能够迅速在不同数据集、机器人平台与任务设置间迁移与复用模块组件，为早期VLA研究提供了极高的实验效率。

随着Transformer在机器人时序建模中的普及，这一范式逐渐与更强的序列化动作学习方法结合，其中RT-1可视为后期融合的成熟形态。在RT-1中，视觉与语言特征仍由独立编码器生成，但随后被统一输入多模态Transformer以预测离散化动作词向量，从而实现了动作序列的系统化词向量化表示，使动作、图像与语言能够在同一序列空间中进行建模。尽管RT-1具备更强的表达能力和大规模训练优势，其视觉与语言的深层交互仍然集中于策略阶段，本质上仍延续了后期融合的设计思路。

随着机器人任务复杂度不断提升，例如开放集物体处理、长时序任务执行以及依赖语言推理的高层决策需求增加，后期融合的结构性限制也愈发明显。由于视觉与语言在编码阶段缺乏深度交互，模型在视觉定位（visual grounding）与语义对齐方面受限；策略网络成为主要的跨模态推理载体，使模型难以解析复杂语言结构或多条件组合；在Open-X Embodiment（O’Neill等，2024）、RT-X等大规模多机器人数据集背景下，后期融合的表征能力和泛化能力开始出现瓶颈。这些局限促使研究开始转向更紧密、更一体化的跨模态融合方式，包括依赖强VLM的互交互融合、统一词向量空间驱动的多模态Transformer，以及近年来快速发展的扩散式动作生成模型。

总体来看，后期融合结构确立了VLA发展的基础范式，在模块化设计、可解释性与工程可行性方面具有重要作用。其方法简洁、组件解耦且易于部署，使其在资源受限或系统搭建初期的场景中仍保持价值。然而，面对更高层次的跨模态理解与复杂任务需求，其在表征融合深度、抽象语言理解能力与长时序控制方面的局限性日益突出。正因如此，后期融合不仅定义了VLA的技术起点，同时也成为后续紧密融合架构的重要对照基线，推动了统一化、多模态深融合模型的快速发展。

2.1.2 规划器—执行器

规划器—执行器（planner-executor）架构是近年来VLA体系中发展最为迅速的核心范式，其基本思想是通过将高层规划与低层执行解耦，以增强复杂任务中的可解释性、泛化能力与长时序推理效率。在这一架构中，大语言模型（LLM）通常负责任务理解、步骤拆解与语义规划等高层功能；具体动作的执行则由独立的策略网络、技能库或控制器等可学习模块承担。与早期后期融合依赖策略网络承担全部跨模态推理的方式不同，规划器—执行器的分工机制有效弥补了其在语言逻辑理解、长任务链构建以及高层抽象推理方面的不足，实现了语言系统与控制系统的互补协作。

该范式的早期系统化体现是SayCan（Ichter等，2022）。该系统提出“LLM生成高层计划 + 价值函数评估可行性”的闭环决策方式，使得自然语言指令能够被解析为候选子目标，并结合训练好的技能库与可行性估计选择可执行操作，最终由底层控制器完成动作。SayCan开创性地将LLM的语义推理能力与机器人操作中基于物理约束的行为选择机制耦合，形成了语义规划到物理执行的可解释桥接路径，使“LLM-as-Planner”成为后续长任务机器人操作的重要方向。同一时期的Inner Monologue（Huang等，2023）将LLM用做“对话式内部规划器”，通过生成行为步骤、解释环境状态并进行自我反思（self-

reflection）不断修正计划，使长时序规划过程以语言的形式保持可解释与可追踪性。另一条代表性路线是 Code-as-Policies（Liang 等，2023），其由 LLM 直接生成结构化的可执行代码（如 Python API 调用）完成任务规划，再由既有视觉与控制模块执行，是“结构化规划—控制接口调用”范式的典型实例。

在更开放的环境与长期自主任务中，Voyager（Wang 等，2023a）展示了显著增强的自主规划能力。该系统以 GPT-4 作为高层策略生成器，能够自动构建任务目标、生成新技能并维护不断扩展的技能库，机器人通过环境接口（application programming interface，API）或现有控制器执行生成的动作。这种“自主技能生长式规划”展现了 LLM-as-Planner 在开放世界场景中的持续学习潜力。面向现实部署的更轻量框架，如 InteLiPlan（Ly 等，2026）则强调将 LLM 规划与本地可部署控制器结合，从而提升规划的实时性与系统的应用性，适用于家庭服务机器人等资源受限的应用场景。

总体而言，规划器—执行器范式的典型特征在于高层 LLM 能够将自然语言解析为结构化任务步骤或子目标，从而在语义层面形成清晰的任务逻辑链；低层控制由已有技能库、行为克隆策略或强化学习控制器承担，保证执行阶段的物理稳定性与行为质量；高层语义规划与低层控制之间通过可行性函数、技能接口或中间代码形式建立明确映射，从而保持跨模态推理的可控性与可解释性。多项研究表明，该范式在多步骤操作与长任务链理解中具有较好的表现，但仍受限于规划可靠性与可行性约束。

尽管如此，该范式仍面临多方面限制。由于规划由 LLM 生成，高层计划可能缺乏物理可行性，因此需要额外的可行性模型或过滤机制；规划与执行的模块化分离可能导致规划与动作不匹配，尤其在动态环境或实时任务中更为突出；多数系统仍采用开环方式执行整段计划，缺乏端到端闭环优化能力；其性能上限往往受限于技能库或策略网络本身的覆盖范围，使其难以在缺乏大量示例的情况下获得统一的动作表达能力。

作为 VLA 的重要阶段性范式，规划器—执行器在本质上解决了后期融合在长序列推理与复杂语言理解上的关键瓶颈，使机器人能够通过自然语言进行高层策略生成。然而，其“规划—执行”割裂的结构也预示着更深层次融合架构的必要性。

2.1.3 统一多模态序列

随着 VLA 任务由单步骤操作与短时控制逐渐走向复杂场景理解、跨任务泛化与长时序动作生成，研究开始从“规划—执行解耦”的策略体系转向在单一模型内部统一视觉、语言与动作的端到端架构，统一多模态序列模型经由这一研究思路提出。该方向的核心思想是将视觉观测、语言指令与机器人动作共同视为序列空间中的元素，通过统一的 Transformer 或生成式模型进行整体建模，使跨模态表示在模型内部形成深度融合。与后期融合的浅层融合和规划器—执行器在结构上的功能割裂相比，统一建模方法尝试在同一框架中完成从感知、语义理解到动作生成的全过程，从而提升跨模态一致性与整体泛化能力。

这一范式的早期代表是 RT-1。RT-1 虽然仍属于后期融合架构，但其动作词向量化思想与大型 Transformer 策略网络为随后真正意义上的统一多模态序列模型提供了前置条件，因此常视为两类范式之间的重要过渡。

在 RT-2 中，统一端到端架构得到了更为完整的实现。RT-2 以大型视觉语言模型（PaLI-X/PaLM-E）作为基础网络结构，将视觉—词向量、语言—词向量与动作—词向量纳入统一的自回归建模框架，使模型在具备机器人操作能力的同时，能够利用互联网视觉知识与跨模态语义进行推理与迁移，从而在一定程度上实现“世界知识—动作能力”的统一建模。RT-2 的出现标志着 VLA 从“机器人专用模型”向“具备通用知识的统一多模态模型”迈出关键一步。随后的研究进一步推动了统一架构在开源与学术领域的成熟化。OpenVLA 基于 LLaMA2、DINOv2 与 SigLIP，将视觉块、语言词向量与动作词向量全部离散化后输入同一序列 Transformer，实现了统一编码、推理与生成，并在 Open-X Embodiment 等大规模数据集上展示了较强的跨机器人泛化能力。UniVLA（Bu 等，2025b）则进一步强化了这一趋势，将视觉块、语言符号与动作词向量完全统一为离散化 symbolic 序列，通过单一 Transformer 进行端到端建模，是“统一词向量编码器 + 单 Transformer 主干”路线的代表性成果，体现了统一序列建模在结构上的极简主义倾向。

在动作生成方面，研究也开始突破纯自回归词向量的局限，将统一模型与扩散式或流匹配式生成

方法结合。Octo在统一视觉与语言编码的基础上采用扩散生成连续控制动作，从而将统一架构扩展到高维、长时的连续控制任务场景。π0则通过流匹配技术实现从文本指令与视觉输入到高频、连续控制流的端到端生成，大幅提升了统一模型在实时控制中的响应速度与执行稳定性。这类方法表明，统一序列模型在机器人动作建模中正逐步从自回归词向量转向更具表达性和实时性的生成机制。

从整体趋势来看，统一端到端架构的发展呈现出3个显著方向：1）模态序列化（tokenization）程度不断提升，使视觉块、语言符号与动作表示能够在同一序列空间中进行统一建模；2）跨模态深度融合成为核心特征，Transformer的全局注意力机制使模型能够在内部形成语义、视觉与动作间的高维交互；3）大规模模型与大规模示范数据成为性能提升的关键来源，RT-2、OpenVLA与UniVLA等系统的显著成功均依赖数十万至百万级别的机器人示范数据。

尽管统一序列模型在跨模态推理能力、任务泛化能力以及复杂语义一致性方面的表现普遍优于规划器—执行器，其局限性同样突出。该类模型高度依赖动作词向量的质量：离散化过粗会限制动作精度，而过细则导致序列长度膨胀并增加训练难度；自回归解码在连续控制场景中存在明显延迟，限制了其在高速、高精度任务中的适用性；模型规模与数据需求巨大，使其训练与部署成本居高不下。此外，统一模型尽管能够捕获长距离时间依赖，但缺乏显式的规划结构，在长任务链中规划一致性仍不及LLM-as-Planner等具有显式结构的体系。

为了缓解这些瓶颈，近期研究开始探索在统一架构中引入离散扩散、流匹配等非自回归动作生成方式，以在保证表达能力的同时提升控制带宽、减少延迟，并进一步强化统一模型在长时序任务中的一致性与稳定性。

2.1.4　离散动作扩散

在统一多模态序列模型逐渐成为VLA的主流框架后，机器人动作建模所面临的瓶颈日益凸显。传统基于自回归的动作预测需要逐词向量生成，不仅解码效率低，而且随着序列长度增加会产生显著的误差累积，使其难以适应高维动作空间、高带宽控制场景和长时任务。为突破自回归的结构性限制，近年来出现了以离散动作扩散（discrete action diffusion）为代表的全新动作生成范式，通过在离散化动作词向量上引入扩散式反推（denoising）过程，以更稳定、更高效的方式建模复杂动作分布。

这一方向的思想可追溯至Diffusion Policy等连续控制中的扩散式策略模型，它们证明扩散模型在多峰策略分布、复杂操作模式以及长时间依赖建模上的显著优势。随后的研究将扩散过程系统化到离散空间，为离散扩散的收敛性质、多模态条件建模机制以及策略优化方法提供了理论基础，使扩散模型能够在动作词向量空间中稳定运行，并适用于组合式动作结构和高维动作语义。

自2024年起，离散扩散开始与VLA的统一架构深度结合，形成了新兴的“扩散式VLA（Diffusion VLA）”研究路线。相关代表性工作展示了离散扩散在机器人动作预测中的加速、稳定与泛化优势。例如，Discrete Diffusion VLA（Liang等，2025）通过对离散动作词向量进行掩码去噪的扩散反推，替代自回归的逐步生成，使动作序列能够一次性并行预测，大幅降低推理延迟并提高长任务的稳定性。在此基础上，dVLA（Wen等，2025a）则进一步将离散扩散语言模型作为统一骨干网络，使视觉、语言与动作词向量在扩散过程中进行联合建模，实现跨模态推理与动作生成的一体化优化。DiffusionVLA（Wen等，2025c）则展示了语言链式推理与扩散动作生成的协同作用，将LLMreasoning用做高层语义引导，使扩散模型能够生成更细粒度、更高精度的动作序列。上述研究从多个角度验证了离散扩散在统一VLA框架中的可行性与应用潜力。

与此同时，VQ-VLA（Wang等，2025c）指出高质量动作离散化（VQ-Action）对扩散式动作生成具有重要影响，动作词向量的压缩率、语义结构与几何一致性会直接影响扩散建模性能，从而表明“动作离散化设计 + 扩散生成”协同优化具有实际必要性。总体来看，离散动作扩散在多个方面展现出相较于自回归方法的显著优势：通过扩散反推实现的并行式动作解码可有效降低推理延迟，使其在连续控制与长任务中具有更好的实时性；其在多峰与高维动作分布建模中具有较好的稳定性，能够处理复杂策略结构；扩散过程在长时任务中对误差累积具有更强的抑制能力，有助于提升长序列执行的稳定性。同时，扩散模型可以嵌入统一Transformer或多模态序列架构中，使扩散头能够与OpenVLA、UniVLA等一体化VLA系统进行集成，成为它们的通用控制组

件。然而，该方向仍面临采样效率较低、动作词向量设计复杂以及训练成本较高等挑战，需要进一步发展高效采样（如 few-step diffusion 与流匹配）、统一的动作离散化方法以及跨模态联合训练策略。

综上，离散动作扩散代表了 VLA 动作建模从逐步自回归到并行化、高精度生成的重要演进方向，并成为 2024—2025 年机器人研究中最活跃的前沿之一。随着高效扩散推理机制与统一多模态训练框架的持续发展，该范式有望在下一代端到端统一 VLA 模型中发挥更加重要的作用，为高维动作理解、长序列控制以及复杂多任务操作提供更具表达性与稳定性的解决方案。

2.2 具身思维链

2.2.1 LLM 中的思维链

思维链（chain-of-thought，CoT）是近年来在 LLM 中取得重要突破的一个核心概念。CoT 通过显式生成一系列中间推理步骤，使模型能够具备执行复杂多步骤推理任务的能力。在 NLP 的早期探索中，由于模型缺乏显式逻辑推理能力，其在面对多步推理类复杂问题时性能往往明显下降。Seq2Seq（Sutskever 等，2014）等早期序列生成模型虽然尚未明确提出 CoT 的概念，但是其通过递归神经网络进行步骤化序列生成的思想，在形式上已具备思维链的初步特征。

2022 年，“思维链提示”（CoT prompting）技术首次正式提出（Wei 等，2022），通过提示模型生成一系列中间推理步骤，显著提升了模型在深层次、多维度等复杂推理任务中的表现，并为其在多模态与具身智能领域的迁移应用提供了重要的方法基础。随着 CoT 技术的不断发展，研究者开始进一步探索思维链生成过程的优化方法，并尝试将其拓展应用到更广泛的任务场景中。Wang 等人（2023b）提出了自一致性（self-consistency）策略，通过多次生成多个思维链来选择最一致的答案，从而提高推理的准确性。Yao 等人（2023）则是基于人类思维可发散、可回溯的特点，允许模型在推理的每一步探索多种不同的可能性，通过对这些“思路”进行评估和搜索，找到最优的推理路径。总体而言，LLM 中的思维链技术经历了从简单提示机制到结构化推理框架的演变。随着技术的不断进步，研究者不仅持续优化思维链的生成方式，还逐步将其扩展应用到多模态建模、强化学习等更复杂的任务领域。

2.2.2 具身思维链与 VLA

具身思维链（embodied chain of thought，ECoT）是经典 CoT 在多模态与具身环境中的延伸与演进，其核心思想在于：一个具身智能体（如机器人）在执行物理任务时，应生成一系列连贯的、基于物理常识的推理步骤，并据此执行动作，同时能根据环境反馈进行动态调整。Zawalski 等人（2025）系统性提出并形式化了 ECoT 的概念，构建了其理论与实践框架。该工作通过结合子任务、边界框预测以及二维运动轨迹等监督信号，使 VLM 学习到更适合具身任务的表征，并在泛化基准测试中取得了更优的性能。

具身思维链将高层次的任务推理和低层次的机器人状态联系起来，通过数据驱动的方式，显式训练 VLA 模型的具身思维链推理能力。如图 6 所示，具身思维链的 5 个关键步骤包括：1）TASK（任务重述）：首先引导模型理解任务并对目标进行重述；2）PLAN（高层计划）：基于任务要求生成实现目标所需的步骤序列；3）SUBTASK（子任务决策）：模型基于当前场景和机器人状态推理出需要执行的子任务；4）MOVE（低层动作）：根据推理的结果，模型生成对应的低层次的动作命令；5）关键空间特征：预测场景中与具身交互相关的空间化特征，例如机械臂的末端位姿，物体位置与边界框等。具身思维链的设计将高层任务规划、中层状态机推理和低层空间感知统一到一个连贯的推理链中，从而支持对复杂任务的分步推理。

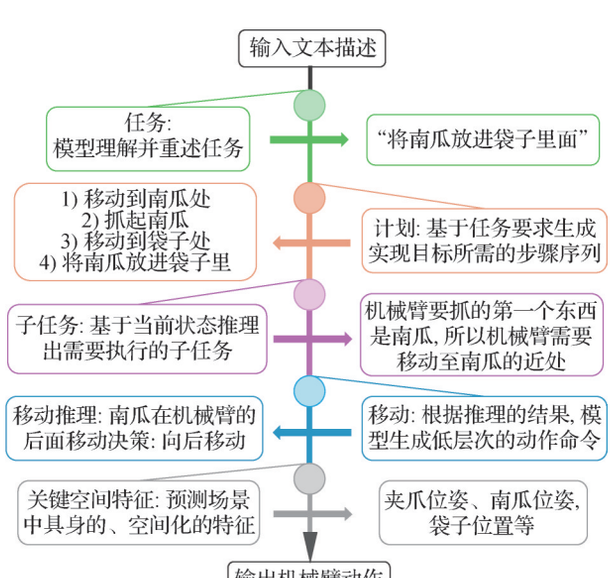

图 6 具身思维链的推理过程

Fig. 6 Reasoning process of the embodied thinking chain

当前，具身思维链的研究呈现出蓬勃发展的态势，并逐步从单一模型探索走向多元化的协同设计。Fast ECoT（Duan等，2025）提出通过识别和缓存重复的推理片段，在保留思维过程的结构和可解释性的前提下，减少冗余计算，并使用并行化策略以提高推理速度。Hybrid VLA（Mazzaglia等，2025）提出将具身思维链的预训练分解为思考、行动与跟随3个子任务，在保持较低推理时延的同时实现了更高的性能。后续工作（Hancock等，2025）通过使用子任务、文本形式的动作描述以及中间层运动规划对机器人数据集进行重新标注，而非采用传统离散动作标记。在仅进行LoRA（Hu等，2021）微调的条件下即可获得有效的动作预测能力，同时不会显著损失VLM的推理性能。但总体而言，如何将具身思维链更有效地应用于VLA体系仍有诸多问题尚待解决，例如多样化数据集的缺乏、实时性与效率之间的平衡以及多模态融合机制的深入建模等。

2.3 高效VLA

高效VLA是应对VLA模型“大规模、高消耗”问题的核心研究方向，其目标是在不损失核心任务性能的前提下，降低模型的计算开销、内存占用与推理延迟，同时满足边缘设备（如移动机械臂、车载终端）的资源约束和实时性需求。该方向围绕模型全生命周期进行优化，与核心机制中的跨模态表征、动作生成等环节深度关联，形成了模型架构、感知特征、动作生成和训练推理四大维度的系统性优化路径，共同实现“高效能、低消耗”的技术目标。

2.3.1 高效模型架构设计

对模型架构的改进通过优化基础架构的结构逻辑与计算分配方式，旨在从根本上减少资源消耗，同时保留VLA模型的多模态语义理解与精准动作生成能力。其核心思路为“按任务复杂度动态分配计算资源”——针对复杂语义任务提供充分的计算资源，而对简单动作任务则简化计算流程。基于这一思路，研究提出了3类主要技术路线：静态骨干轻量化、动态计算路径和双系统设计。

静态骨干轻量化的核心思想是将传统大规模参数的VLM骨干网络（如LLaMA-7B、CLIP-L/14）替换为高效架构或小尺寸模型，从而在结构设计阶段有效控制参数规模与计算复杂度。这样做的优势在于：传统Transformer架构的自注意力机制存在$O(n^2)$的计算复杂度（n为词向量长度），而轻量化架构通过优化时序建模逻辑或精简网络层数，兼顾“小参数+高性能”。RoboMamba（Liu等，2024）采用状态空间模型（Mamba）作为序列建模核心，在2.7 B参数下实现了比传统Transformer更高效的时序建模；TinyVLA（Wen等，2025b）、SmolVLA（Shukor等，2025）直接采用Pythia-1.3B、SmolVLM-2（0.24～2.25 B参数）等轻量化模型，通过结构精简降低部署门槛。

动态计算路径的核心思想是“训练阶段保留大模型的全功能，推理阶段根据输入复杂度和任务需求动态选择有效计算流程”，从而避免了“一刀切”的冗余计算。DEER-VLA（Yue等，2024）引入早期退出机制，在中间层设置轻量策略头，通过输出相似度判断是否提前终止计算；MoLE-VLA（Zhang等，2025c）采用混合专家（mixture of experts，MoE）框架，动态选择参与计算的网络层，结合自蒸馏保证性能。

双系统设计则借鉴认知科学“快慢系统”理论，将VLA的“语义推理”与“动作生成”拆分为两个独立子系统，分别处理复杂决策与快速响应任务，通过异步协作平衡推理深度与实时性。其中慢系统（System2）负责高复杂度的语义理解、长程规划（如“根据食谱制作咖啡”的步骤分解），采用大参数量VLM保证推理能力；快系统（System1）负责低延迟的动作生成（如“移动机械臂至咖啡机上方5 cm”），采用轻量模型降低延迟，两者通过潜向量或特殊token传递信息。例如，在OpenHelix中，LLaVA-7B（慢系统）负责语义推理，3D Diffuser Actor（快系统）则生成动作；RoboDual结合OpenVLA（慢系统）与DiT（快系统），通过异步协作平衡推理深度与实时性。

2.3.2 感知特征高效处理

VLA的感知输入以视觉为主（占总词向量长度的70%～90%），且包含大量冗余信息（如背景区域、静态物体）。感知特征高效处理通过“筛选关键信息+复用有效特征”的方式，在降低前端感知计算负担的同时，尽量保持动作生成与场景之间的关联性。围绕该核心路径，高效感知处理主要集中在单帧特征选择性处理和时序特征复用两种技术路线。

单帧特征选择性处理面向单帧视觉输入，通过“重要性评估+词向量剪枝”的方式剔除与任务无关的视觉词向量（如背景像素、非目标物体），仅保留目标物体、机器人末端执行器等核心信息，从而减少后续LLM需要处理的词向量数量。其核心依据是对

“语义相关性”（与语言指令的关联度）和“空间相关性”（与动作执行区域的关联度）进行量化，从而筛选必要的词向量。例如，SP-VLA（Li 等，2025b）结合注意力语义权重与轮廓边缘的空间权重，从语义与空间两个维度对词向量进行筛选；FlashVLA（Tan 等，2025）通过奇异值分解（singular value decomposition，SVD）与计算信息贡献度（information contribution score，ICS），实现与高效注意力机制兼容的词向量剪枝；LightVLA（Jiang 等，2025a）采用查询驱动的可微分词向量选择机制，使词向量选择过程能够随任务需求动态调整。

时序特征复用则利用连续帧间的视觉相关性（如静态背景、缓慢移动的物体），复用前一帧已计算的特征，避免每帧都进行全流程特征提取，进一步降低计算负担。VLA-Cache（Xu 等，2025）通过复用静态图像块对应的 KV（key-value）缓存，并根据注意力熵动态调整复用比例，以减少重复计算；TTF-VLA（Liu 等，2025）通过像素差异与注意力相关性生成二进制掩码，仅对动态区域或任务关键区域的特征进行更新；Fast ECoT 则通过缓存缓慢变化的高层推理结果，减少重复规划带来的计算开销。

2.3.3　高效动作生成机制

动作生成是 VLA 与物理世界交互的核心环节，传统逐帧生成方式存在“误差累积”、“推理延迟高”等问题，高效动作生成通过优化动作表示形式（如从离散词向量到连续块）与生成流程（如并行解码），在保证控制精度的同时提升效率，主要涵盖动作块生成、动作表示压缩和推理感知型生成三大方向。

动作块生成改变了传统“一帧一动作”的逐帧生成模式，使模型在单次推理中生成多步连续动作（即动作块），从而减少解码次数，并通过时序集成降低单步误差累积。其核心做法是将动作序列按时间维度划分为固定长度的块（如 5 步/块、10 步/块），模型单次输出一个动作块。块内动作通过时序平滑（如滑动平均）进行处理，块间则通过重叠区域约束以保证动作连贯性。例如，RTC（real-time chunking）（Black 等，2025b）将动作块生成转化为序列补全问题，并通过软掩码方式处理块间重叠区域，以增强动作序列的连贯性；VOTE（vision-language-action optimization with trajectory ensemble voting）（Lin 等，2025）引入〈ACT〉特殊词向量，单个词向量映射一整段动作序列，配合集成投票机制以提升鲁棒性。

动作表示压缩通过离散化、频域转换等方式，缩短动作序列的表示长度，降低模型对动作词向量的处理成本，同时保证动作的可恢复性与控制精度，核心是利用连续动作序列的时序冗余（如匀速运动的动作参数变化小），通过数学变换提取关键特征后离散化编码，推理时再通过逆变换恢复原始序列。FAST（frequency-space action sequence tokenization）对离散动作序列应用离散余弦变换（discrete cosine transform，DCT），保留低频频谱系数后再进行字节对编码（byte pair encoding，BPE），实现无损压缩；Omni-SAT（Astruc 等，2024）采用 B 样条时序对齐与残差向量量化（residual vector quantization，RVQ），将连续轨迹转化为紧凑离散词向量。

推理感知型生成则在动作生成流程中融入轻量化推理逻辑（如子目标预测、轨迹修正），既避免传统“纯动作生成”的泛化性不足，又防止“重推理 + 轻动作”的延迟问题，核心是通过“简化推理步骤 + 复用推理结果”，在动作生成中嵌入必要的逻辑判断，而无需每次单独调用大模型进行完整推理。例如，ECoT-Lite（Chen 等，2025b）通过引入推理 dropout 机制，使模型在训练阶段学习推理逻辑，而在推理阶段仅输出动作，从而显著降低推理延迟；DreamVLA（Zhang 等，2025e）通过光流检测动态区域，仅预测与动作相关的视觉子目标，从而减少无效计算。

2.3.4　训练与推理优化

训练与推理优化贯穿 VLA 模型的整个生命周期。在训练阶段，重点关注“降低资源消耗”（如算力与数据量）；在推理阶段，重点关注“加速部署落地”（如降低延迟与内存占用）。通过算法优化，旨在实现“低成本训练 + 高效率部署”，可以分为训练阶段优化与推理阶段优化两个主要方向。

训练阶段优化旨在解决传统 VLA 训练中“参数量大、数据需求高”的问题。优化方法包括参数高效微调、知识蒸馏、量化训练与数据高效训练等，通过这些方式减少训练过程中的资源消耗，同时保证模型的性能。知识蒸馏用于转移大模型的能力，CEED-VLA（Song 等，2025a）采用一致性蒸馏稳定非自回归推理；Vita-VLA（Dong 等，2025）将小尺寸动作专家模型的控制知识蒸馏至大 VLM 骨干；量化训练（quantization-aware training，QAT），如 BitVLA（Wang 等，2025a）通过 1-bit 量化与渐进式蒸馏，将模型内存占用从 15.1 GB 压缩至 1.4 GB。

推理阶段优化针对传统自回归(autoregressive, AR)解码“串行生成、延迟高”的问题,通过非自回归解码、投机解码、并行推理和硬件适配优化等方式,突破串行瓶颈,提升推理速度。如OpenVLA-OFT(Kim等,2025)改用双向注意力与连续回归目标,支持单步并行生成动作序列。

2.4 强化学习结合VLA

随着VLA模型在机器人操作中的广泛应用,其核心训练范式仍主要依赖行为克隆或大规模示范数据的监督学习。然而,纯模仿学习难以覆盖复杂环境中的全部状态空间,模型在分布偏移、稀疏反馈、未知动态与长序列任务中仍面临明显性能瓶颈。因此,结合强化学习以实现自我改进、交互式学习与持续适应,逐渐成为VLA体系发展的重要方向。近期研究从在线RL、离线RL、RL微调、世界模型辅助RL以及安全RL等维度逐步探索VLA与RL的深度融合,形成了多个具有代表性的技术路径。

最直接的路线是将RL引入预训练VLA模型的后训练阶段,以弥补监督学习在未知状态与策略退化问题上的局限。例如,Improving VLA(Guo等,2025)提出iRe-VLA,通过“RL + 行为克隆”交替训练的方式,使大规模VLA模型能够在真实交互中进一步提升性能。该工作指出,直接对VLA进行在线RL会遇到训练不稳定和计算成本极高的问题,因而提出在RL更新中加入示范引导与价值约束,使得大型VLA模型能够实现可控、可收敛的策略优化。类似地,ReinboT(Zhang等,2025b)将RL原则融入VLA的混合数据训练框架,通过引入奖励信号对低质量示范进行再加权,提高策略在噪声示范和偏移轨迹下的鲁棒性。

另一类工作将RL视为VLA的可扩展能力增强器,通过构建专门的RL微调(reinforcement fine-tuning)机制,使VLA在多任务、多机器人平台下获得持续提升。例如,VLA-RL(Lu等,2025)提出在预训练VLA的基础上使用规模化的在线RL提升任务泛化能力,同时提出轨迹级RL表述和语言—视觉奖励模型,缓解了真实机器人任务中的稀疏奖励问题。ConRFT(Chen等,2025d)则探索在离线示范基础上进行RL-style微调,通过将一致性学习与Q-learning结合,使VLA能够在无需大量在线交互的情况下有效更新策略。

此外,一些研究开始在更具结构化的RL框架中重构VLA系统,以形成更稳定、更具解释性的奖励学习机制。例如,VLAC(Zhai等,2025)提出专门为VLA设计的Critic模型,通过progress-delta与done信号构建任务无关的奖励模型,使VLA可以在真实环境中直接学习累积奖励,减少手工reward engineering的成本。ManipLVM-R1(Song等,2025b)则将RL引入大型视觉语言模型的推理与操作任务中,利用“可验证奖励”增强VLA的泛化能力,突破监督学习在长任务链推理中的局限。

在可交互训练难、真实机器人风险高的问题下,基于世界模型辅助的RL也成为重要趋势。VLA-RFT(Li等,2025a)利用世界模型生成未来视觉观测与验证奖励,在模拟环境中对VLA进行RL微调,从而显著减少真实交互样本需求,并提升长时任务与视觉复杂场景中的策略稳定性。这种“世界模型 + RL微调 + VLA”路线为大规模、低成本的安全训练提供了一条可行的路径。

安全性也是RL与VLA结合的重要主题之一。SafeVLA(Zhang等,2025a)将安全强化学习引入VLA训练,通过约束MDP(Markov decision process)、风险检测以及对抗式不安全行为生成等策略,使模型在真实部署时能够规避危险操作,并在环境变化中保持策略稳定。这类方法强调VLA模型在真实世界语义—物理映射中的安全对齐问题,是未来具身大模型不可避免的研究方向。

总体而言,结合强化学习的VLA工作在多个方面展示出显著优势:通过奖励信号补充监督学习的偏差,使模型能够在未知状态中自我改进;通过在线或离线RL的策略优化机制,提升多任务泛化和环境适应能力;通过Critic、世界模型、安全RL等模块,使策略学习更稳定、安全并具备可验证性。然而,这一方向仍面临显著挑战,包括高昂的样本需求、奖励难以设计、真实环境中的风险成本、RL与大型VLA模型融合时的训练不稳定性,以及如何在不破坏预训练能力的前提下进行有效的RL更新。

随着VLA模型规模的进一步扩大、交互式学习需求的增加以及世界模型技术的成熟,RL增强式VLA逐步成为“具身智能体”的关键训练范式之一。其目标不再局限于提升单任务性能,而是通过交互式强化学习使VLA具有自主探索、持续进步以及在人类无法穷尽示范的长尾场景中自我提升的能力。未来数年,基于RL的VLA有望成为通用机器人模

型训练流程中的基础组件，与2.1节中的统一序列模型、扩散式动作生成器等方向共同构成下一代具身大模型的核心技术框架。

2.5 跨动作空间学习VLA

VLA的另一项关键挑战是动作空间的异构性，特别是在跨机器人和跨数据源的场景中。为了支撑更大规模、多平台的具身智能，如何协调不同形态机器人的动作表示和控制接口，并将其纳入统一的模型框架，已成为亟待解决的问题。现实世界机器人数据高度异构：不同平台具有维数各异的关节与控制接口，搭配不同相机布局、传感器与控制频率，而数据又往往来自多个实验室或公司的独立采集，因此在分布上呈现出显著差异。早期如RT-1、RT-2等VLA工作主要在单一或少数机器人平台上验证，展示了“大模型 + 多任务数据”在一定程度上的泛化能力，但其动作接口仍相对统一，尚未触及大规模跨形体、跨数据源的动作空间异构问题。

在VLA之前，小规模强化学习领域已经开始探索跨形体迁移。例如，通过将不同机器人的状态与动作投影到公共潜在空间，在该空间中学习统一的控制策略，再借助各自的编码器与解码器对齐不同形体，从而在抽象潜在空间中统一多种动作空间。然而，这类方法的任务规模与形体多样性通常较为有限。

进入通用VLA阶段后，主流做法先训练一个共享的视觉—语言—状态主干，再在主干后面为不同机器人各自接上动作预测头，由这些“尾部模块”去适配维度和控制方式不同的动作空间。以π0及其后续HiMOE-VLA（Du等，2025）为代表，模型在大规模多机器人、多任务数据上预训练一个统一的视觉—语言—状态主干，再为每个机器人配置独立的动作头或分层混合动作专家，以适配维度和控制模式各异的动作空间，这类方法从工程角度缓解了动作维度对齐的问题，并支持模型在多种机器人形体上的扩展。

并行发展的一条路线是通过“提示”显式编码域信息。早期做法是在输入中加入描述硬件配置的自然语言，使预训练VLM以熟悉的形式感知平台差异，但这种方式依赖人工模板，难以适应硬件配置组合数量快速增长的情况。随后，X-VLA（Zheng等，2025）提出为每个数据源或形体分配一组可学习软提示向量，在冻结或少量调整主干的情况下，通过连续提示即可路由不同数据域，实现跨具身条件下的稳健训练，并在多仿真与多真实机器人上取得领先性能。

另一条重要脉络是“统一潜在动作空间”。UniVLA以大规模跨具身视频为起点，学习任务中心的潜在动作表示，并在该抽象空间中进行规划，再通过轻量解码映射到各机器人动作接口，实现“在统一动作语义上学策略、在末端轻量适配形体”。以潜在动作扩散（latent action diffusion）为代表的做法，则是在一个统一的、由模型自动学习出的抽象动作空间中，通过逐步生成的方式产生动作序列，使不同机器人都能在同一种“动作意图表示”下执行控制。相较于为每种机器人不断增加额外适配模块，这类方法将研究重心转向到如何设计一种能够概括任务意图又不依赖具体关节结构和控制接口的抽象动作表达，从根源上解决不同机器人之间动作形式不一致的问题。

在更大规模的跨具身预训练上，XR-1（Fan等，2025）通过统一视觉—运动编码将视觉动态与机器人运动编码到共享离散表示，并结合分阶段训练，在6种形体、百余真实操作任务上显著超越包括π0、UniVLA与GR00T-N1.5（Bjorck等，2025b）在内的一系列基线取得了更好的实验结果。这表明，统一的视觉—运动离散码可以成为连接多源人类与机器人示范的有效“中间语言”。

综合来看，跨动作空间的研究正在逐步从依赖单一机器人适配的范式，过渡到在共享模型中对不同硬件信息进行统一管理，并进一步探索能够同时描述视觉变化与动作意图的统一表达方式。当前的关键挑战在于：如何在不破坏预训练能力的前提下，使模型稳定地感知不同平台之间的差异，同时仍能在统一表示中捕捉任务共性，并支持多形体扩展。

3 数据集与评估基准

作为构建与评估VLA能力的基础，数据集与基准在推动具身智能发展中发挥核心作用。本节系统梳理现有数据资源与评测体系，为VLA的训练支撑与能力边界提供整体框架。

3.1 数据集

目前，VLA数据集主要可分为仿真环境数据集、真实世界机器人数据集和人类视频数据集3类，

其发展概况如图7所示。

3.1.1 仿真环境数据集

仿真环境长期以来是机器人学习的重要数据来源，能够以可扩展、安全且低成本的方式支持大规模数据生成，在模仿学习、大规模预训练和具身智能体研究中具有重要作用。早期工作多依赖人工远程操控收集演示数据。例如，基于 MuJoCo 的 RoboTurk（Mandlekar 等，2018）使用 Sawyer 等机器人，通过移动端或云端终端采集高质量的6自由度操作轨迹。在真实数据采集受成本与时延限制的背景下，这类仿真方式凸显了其在规模与效率上的优势。

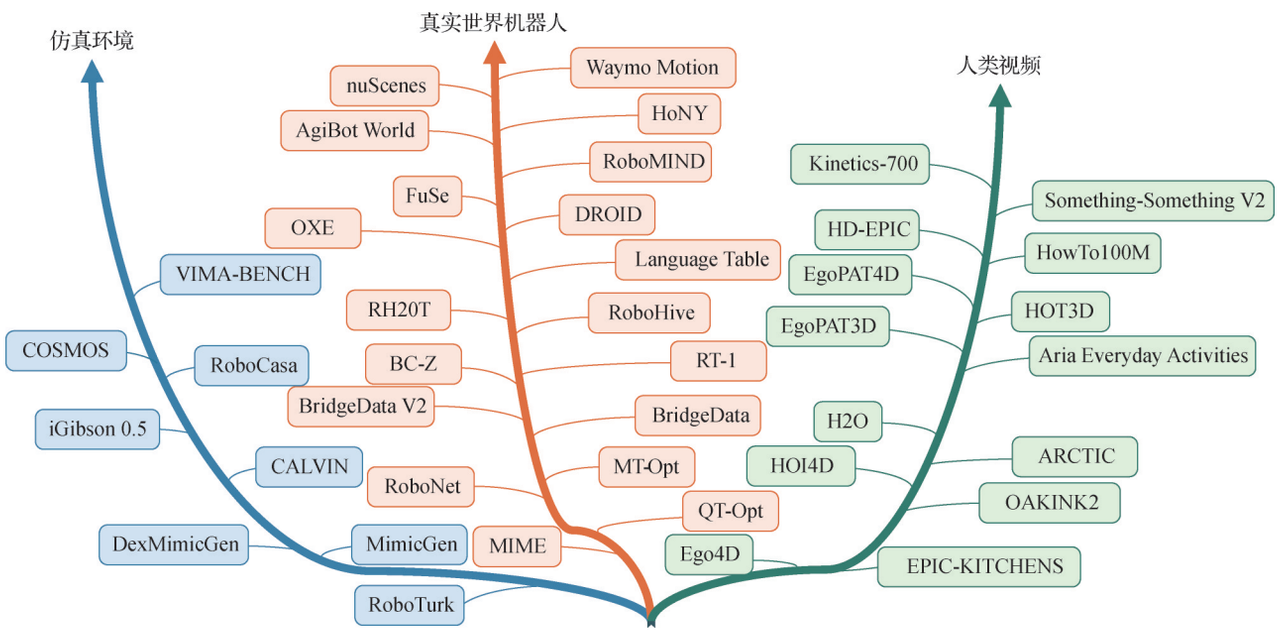

图7 3类VLA数据集的发展现状

Fig. 7 The current development status of three types of VLA datasets

为缓解人工演示采集成本，部分工作转向程序化和模型驱动的合成数据。MimicGen（Mandlekar 等，2023）从少量专家演示出发，将任务分解为以物体为中心的子任务，并通过空间变换与场景重配置生成多样轨迹；DexMimicGen（Jiang 等，2025b）将该范式扩展至双臂和多指等复杂形体；RoboCasa（Nasiriany 等，2024）进一步在家庭场景中大规模合成多种操作任务轨迹，并已被 GR00T N1.5（Bjorck 等，2025a）等工作用做评测场景之一。与之互补，Cosmos（Agarwal 等，2025）等大规模视频世界模型可自动生成多样化的合成视频序列与环境演化过程，为视觉—语言—动作模型预训练提供可扩展的多模态序列。

在任务形式上，仿真数据集已从单一操作拓展到导航、探索和多步骤操作等多种范式。iGibson（Xia 等，2020）提供高保真交互式导航环境，并通过交互式导航评分对路径效率与动力学代价进行量化评估。

3.1.2 真实世界机器人数据集

尽管仿真环境在规模化生成方面具有无可比拟的优势，但其物理逼真度与传感噪声始终与现实存在差距。因此，真实世界采集的数据成为验证策略可执行性、缓解模拟到现实落差的关键补充。真实世界机器人数据集通过在物理平台上采集，记录机器人与环境交互中的传感噪声、接触动力学和不确定性，为学习物理上可执行的策略提供关键监督。相比仿真或人类视频，真实数据更能反映实际操作约束，是缩小模拟—现实差距和训练底层控制策略的重要基础。

早期的大规模具身数据集为该领域的研究奠定了基础。MIME（multiple interactions made easy）（Sharma 等，2018）数据集包含8.2 K条跨20个任务的真实示范，通过机器人本体示教采集关节轨迹，并为每条机器人演示配对采集对应的人类第三人称视频示范；QT-Opt（Kalashnikov 等，2018）数据集基于多台机械臂采集了58万次抓取尝试，并在 MT-Opt

(Kalashnikov 等,2022)中扩展到更广泛的操作技能。RoboNet (Dasari 等 , 2019) 统一了 Sawyer、Baxter、WidowX、Panda、KUKA iiwa、Fetch 与 Google Robot 等多种平台,含约 15 M 帧、约 16 万条机器人交互轨迹,是跨平台泛化研究的代表性资源。

基于遥操作的 BridgeData(Ebert 等,2021)系列将多环境、多任务的真实演示纳入统一框架。BridgeData 包含 7 200 条 VR(virtual reality)操作轨迹,覆盖 10 个环境与 71 项任务;BridgeData V2(Walke 等,2024)扩展至 60 000 条、覆盖 24 个环境,在多步操作学习和跨领域迁移研究中得以广泛采用。Google Robots 采集的 RT-1 数据集包含 13 万条演示轨迹,是 RT 系列 VLA 模型的主要训练数据;BC-Z 数据集提供 25 900 条多任务演示,并提供人类监督的策略执行数据,用于评估场景内组合泛化。

此外,部分数据集建立了配套的软件平台。RoboHive(Kumar 等,2023)提供系统化的具身学习环境,RH20T(Fang 等,2023)数据集涵盖 147 个任务、10 万次操作实验,并包含同步的 RGB-D、力/力矩、关节力矩与音频信号,以及配套语言描述,适用于多模态控制与一次性模仿学习。

随着跨具身泛化与大规模训练的需求增长,社区开始整合多来源真实数据。Open X-Embodiment 数据集(open x-embodiment dataset, OXE)将 RT-1、BC-Z、BridgeData、Language Table 等 22 个数据集统一至 RLDS(reinforcement learning dataset standard)格式,总计覆盖超 527 项技能与 160 266 个任务实例,是目前最全面的真实机器人数据整合集。

与此同时,若干大型统一平台数据集进一步提升了数据规模与一致性。DROID(dataset of robot interactions in the wild)(Khazatsky 等 , 2024) 基于 Franka Emika Panda 机器人采集 76 000 条轨迹,在硬件高度一致的条件下构建跨实验室基准;RoboMIND(multi-embodiment intelligence normative data for robot manipulation)(Wu 等,2025)提供 107 000 条覆盖单臂、双臂、人形与灵巧手的演示;AgiBot World 数据集(Bu 等,2025a)则通过百余台机器人采集百万级轨迹,使大规模真实具身训练成为可能。此外,一些相关工作也基于统一硬件平台构建了规模可观的多模态训练数据,用于验证策略微调与跨模态泛化能力(Jones 等,2025)。

任务或平台特定数据集进一步扩展了真实数据的覆盖,包括无目标探索、灵巧臂操作、电缆布线及多物体重排等。用于长时程规划的 BridgeDataV2 厨房任务子集,以及面向家庭自然场景的 HoNY 数据集,则提升了真实环境复杂度。同样的发展也出现在移动机器人领域,如 nuScenes(Caesar 等,2020)和 WaymoMotion (Ettinger 等 , 2021) , 提供以 LiDAR (light detection and ranging)、RADAR(radio detection and ranging)为核心的真实标注轨迹,用于训练安全关键的移动策略。

尽管真实数据规模不断扩大,其采集仍受成本和操作效率限制。因此当前 VLA 训练通常采用“仿真或网络数据大规模预训练与小规模高质量真实微调”相结合的两阶段策略,以同时保证泛化能力与现实可执行性。总体而言,真实世界机器人数据集的规模化、多样化与标准化正在成为推动 VLA 模型发展的关键动力。

3.1.3　人类视频数据集

除机器人自身采集的数据外,大规模人类行为视频同样为具身智能提供重要的时空先验。相比真实机器人采集成本高昂、场景受限,人类视频能够以更高效率、更自然的方式覆盖丰富的交互模式与任务结构。人类行为视频数据由于采集效率高、无需依赖实体机器人或处于安全关键的交互环境,已成为 VLA 模型预训练和世界建模的重要数据来源。特别是第一人称视频更符合真实机器人(如头戴式相机或类人形机器人)的观测方式,因此在近年来的视觉—语言—动作研究中占据核心地位。代表性大规模数据集如 Ego4D(Grauman 等 , 2022) , 包含逾 3 000 h、来自 9 个国家的自然活动视频,是目前覆盖最广的头戴式 RGB 数据集之一。类似地,EPIC-KITCHENS(Damen 等,2021)系列聚焦厨房场景中的日常操作,HOI4D(human-object interaction 4D)(Liu 等 , 2022) 捕捉丰富的人—物交互过程;而 OAKINK2(Zhan 等,2024)、H2O(hand-object interaction)(Kwon 等,2021)以 RGB-D 与动作捕捉记录双手精细操作,ARCTIC(Fan 等,2023)则面向人类双手操控关节化物体;EgoPAT3D v2(Fang 等,2024)进一步强调动作目标预测。这些第一人称数据为学习时间视觉编码、潜在动作结构和状态变化提供了关键支持。

随着可穿戴设备的发展,第一人称数据呈现更自然、连续和长时程的趋势。Aria Everyday Activi-

ties (Lyu 等, 2024)记录真实日常行为, Ego-Exo4D (Grauman 等, 2024)将第一人称与第三人称视角结合作为3D运动对齐信号, 体现具身迁移能力; HOT3D (Banerjee 等, 2025)专注细粒度手—物体跟踪, HD-EPIC (high-definition egocentric interaction dataset) (Perrett 等, 2025)则扩展了第一人称烹饪活动场景。在可操作表示学习中, 这类时间序列数据常与潜在动作预测方法结合, 用于构建对策略学习友好的潜在动作空间。在模仿学习方向, 高质量演示数据集如Being-H0 (Luo 等, 2025)、MIME通过人类第三人称示范进一步弥补机器人示范的稀缺性, 广泛用于策略初始化与行为克隆。

此外, 大规模视频—语言数据集如HowTo100M (Miech 等, 2019)、Something-Something V2 (Goyal 等, 2017)、Kinetics-700 (Carreira 等, 2022)虽非第一人称视角且缺乏直接可用的动作标签, 但仍提供多样的操作技能、物体交互与物理常识, 在时间视觉编码器预训练和世界模型表征学习中发挥重要作用。近期VLA模型利用弱监督对齐技术从中提取轨迹与潜在状态变化, 例如Magma (Yang 等, 2025)使用动作标记对齐方法, 而Ego-Exo4D的第三人称视角数据进一步提供3D结构化监督。这些数据使模型获得更强的时间一致性、语义结构和体现先验, 为面向类人机器人与具身智能系统的VLA预训练奠定基础。

3.2 评估基准

VLA模型的评估在现实环境中仍受本体差异、安全要求与低可复现性限制, 因此当前研究几乎完全依赖模拟平台。本节聚焦这些平台构建的任务级评估基准, 并依据任务范式分为两类: 以机器人—物体交互为中心的操作任务基准, 以及面向大尺度场景、涉及导航与长时序决策的具身导航与移动操作基准。

3.2.1 操作任务基准

操作任务基准主要围绕机器人在桌面或局部三维空间中的抓取、放置、重排、装配及多指精细操控能力展开, 用于系统评估VLA模型在视觉、语言与控制之间的协同表现。现有基准多依托真实物理引擎, 通过标准化任务定义、初始化策略与终态判定, 量化模型在多情境下的成功率、终态误差、泛化能力与稳定性。表1展示了基于VLA的操作任务基准。

在MuJoCo生态中, robosuite (Zhu 等, 2020)以MJCF描述机器人与场景, 并构建抓取、推拉、插入等9个典型任务, 是最常用的算法验证与评测平台; 其衍生的robomimic (Mandlekar 等, 2022)则提供8个基于仿真和真实机器人的不同任务及高质量轨迹, 用于研究模仿学习中的样本效率与策略鲁棒性。LIBERO (Liu 等, 2023a)系列聚焦语言条件操作, 包含130个任务, 分测空间推理、物体属性、目标解析与组合泛化等能力, 核心指标为语言条件任务成功率; Meta-World (Yu 等, 2021)则通过50个Sawyer任务评估多任务与元学习场景中的跨任务迁移。

在PhysX生态中, IsaacSim/IsaacLab (Mittal 等, 2025)支持高保真GPU (graphics processing unit)并行仿真, 使大规模操作评估成为可能。基于此, LeVERB-Bench (Xue 等, 2025)定义面向人形体的150+语言与视觉—语言任务, 除成功率外还衡量动力学稳定性。Isaac Gym (Makoviychuk 等, 2021)虽然不直接提供任务级基准, 但作为高并行物理后端广泛用于生成连续控制数据, 是IsaacLab等上层基准的关键基础。

表1 基于VLA的操作任务基准

Table 1 Operation task benchmark based on VLA

基准名称	年份	序列数	任务数	说明
robosuite	2020	—	11	标准11个任务
robomimic	2021	15万～20万	8	8个任务, 5种示范
RoboCAS	2024	约20万	约100	复杂物体排列
LIBERO	2023	约13万	130	语言条件任务
Meta-World	2019	—	50/45	常用于多任务RL
LeVERB-Bench	2025	—	约600	人形体语言操作
ManiSkill1	2021	20万～25万	20	多类型物体交互
ManiSkill2	2022	20万～25万	约20	增加关节物体任务
ManiSkill-HAB	2024	—	约10	与Habitat对齐
DexArt	2023	约10万	约30	多指高难动作任务
RoboTwin	2025	约5万	50	双臂协作&迁移任务
Ravens	2020	—	10	桌面操作任务
VIMA-Bench	2023	约5万	17	多模态指令操控
CALVIN	2022	约25万	34	桌面操作任务
LoHo-Ravens	2023	5万	17	多模态指令操控
RLBench	2020	约1万	100	多模态仿真环境

注: "–"表示无数据。

基于同样使用PhysX的SAPIEN，ManiSkill（Xiang等，2020；Mu等，2021）系列构建覆盖关节物体、可变形物体与具运动性物体的综合基准，并提供大规模演示数据，常以视觉条件下的成功率、终态误差及跨实例泛化表现评估策略能力；ManiSkill-HAB（Shukla等，2025）将重排任务与Habitat（Szot等，2021）的Home Assistant Benchmark对齐，通过最终布局与目标布局的一致度细化评估。其他面向特定难点的基准包括：拥挤堆叠场景下精细操控的RoboCAS（Zheng等，2024）、多指机器人高精度关节操控的DexArt（Bao等，2023），以及包含50个双臂协作任务并测试跨本体迁移的RoboTwin2.0（Chen等，2025a）。在PyBullet生态中，Ravens（Huang等，2024）提供10个桌面任务，用于考察拾取、放置、分类与堆叠等基本技能；其扩展VIMA-Bench（Jiang等，2023）通过17个任务与多模态提示接口评估语义对齐与操控能力；LoHo-Ravens（Zhang等，2023）聚焦长时程与分阶段任务的时序推理。CALVIN（composing actions from language and vision）（Mees等，2022）除提供语言条件操控数据集以外，还通过自然语言指令序列驱动Panda执行连续多步任务对模型进行评估，除整体成功率外还统计指令级子任务完成度

在V-REP（virtual robot experimentation platform）（CoppeliaSim）生态中，RLBench（James等，2020）以100个Panda任务构建统一的模仿与强化学习评测框架；其扩展THE COLOSSEUM（Pumacay等，2024）在20个任务上加入14类环境扰动，并通过不同扰动强度下的成功率构建整体鲁棒性指标。操作任务基准构成VLA评估体系中最成熟的资源类型，这些基准依托可控、可复现和可扩展的物理仿真环境，为研究可泛化操作智能提供了关键实验基础。

3.2.2　具身导航与移动操作基准

相较于桌面操作任务，导航类与具身规划基准强调大尺度场景理解、运动生成与跨环境迁移能力，是检验VLA模型是否具备全身级、长时程决策能力的关键组成部分。具身导航与移动操作基准相较局部桌面操作更强调机器人在复杂三维空间中的环境理解、长时程规划以及跨房间、多对象、多阶段的综合决策能力。此类基准通常依托高保真室内场景模拟器，通过真实住宅扫描或程序化建模构建大规模环境，并在其中设置由导航、感知与操作耦合的复合任务。表2为基于VLA的具身导航与移动操作基准。

表2　基于VLA的具身导航与移动操作基准

Table 2　Embodied navigation and mobile operation benchmark based on VLA

基准名称	年份	场景规模	说明
RoboTHOR	2020	89个场景	室内导航与sim2real平台
ProcTHOR-10K	2022	10 000个房屋	程序化生成室内环境
ALFRED	2020	120个场景	语言条件长程家庭任务
Habitat 1.0	2019	数量不固定	照片级3D具身仿真平台
Habitat 2.0/HAB	2021	数量不固定	移动操作与家庭任务
Habitat 3.0	2023	数量不固定	人机协作环境
iGibson 1.0	2021	15个场景	交互式家庭仿真平台
iGibson 2.0	2022	15个场景	对象中心家庭仿真平台

基于Unity的THOR/AI2-THOR（Kolve等，2022）系列是最早的具身平台之一，提供照片级场景和拾取、放置、开关等基础操作，支持导航、模仿学习、强化学习和视觉问答等任务。后续iTHOR、RoboTHOR（Deitke等，2020）、ProcTHOR-10K（Deitke等，2022）与ArchitecTHOR覆盖跨场景泛化、大规模程序化生成及建筑结构设计等需求。依托THOR的ALFRED进一步引入烹饪、收纳等具有不可逆状态变化的长时程家务任务，同时包含高层目标与低层语言指令，使任务在序列长度与动作空间上更接近真实家庭流程。而Habitat 1.0（Savva等，2019）聚焦视觉导航，使用真实建筑扫描构建大规模三维环境；Habitat 2.0（Szot等，2021）将框架扩展为移动操作，引入HAB基准，使智能体在导航过程中完成抓取、搬运等操作；Habitat 3.0（Puig等，2023）则支持模拟的人类与机器人多主体交互。

在提升仿真环境的真实性与有效性方面，iGibson1.0（Shen等，2021）/2.0（Li等，2022a）基于真实住宅扫描构建高保真、多房间及多楼层环境，使导航、搜索与基础操作任务具备更强的现实对应性，并支持跨建筑与跨户型的迁移评估。更广泛的具身平台还支持多智能体交互与RGB-D、IMU、LiDAR等多模态输入，为复杂物理约束下的感知—决策评测提

供基础。类似思想也延伸至自动驾驶模拟器，如CARLA（Dosovitskiy 等，2017）与LGSVL（Rong 等，2020），通过虚拟城市场景与动态交通流评估任务成功率、碰撞率和轨迹偏差等指标，其评测理念与室内导航基准保持一致。

具身导航与移动操作基准通过在逼真大规模场景中定义导航—操作一体化任务，并采用任务成功率、路径效率、碰撞与违规统计、多场景泛化性能及行为稳定性等指标，为系统评估VLA模型的空间理解、长时程规划与环境交互能力提供了统一而全面的基准。

4 主要挑战与未来方向

本节将讨论VLA领域面临的主要挑战及未来发展方向，重点围绕6个关键议题展开，如图8所示，旨在为后续研究提供体系化发展参考。

图8 VLA的主要挑战与未来方向

Fig. 8 Main challenges and future directions of VLA

4.1 数据质量与基准测试

VLA模型的性能在很大程度上依赖于数据质量。数据的准确性、完整性和一致性是影响模型推理可靠性与泛化能力的关键因素。不同模态的数据在VLA任务中各自发挥着重要作用，而如何有效地将这些数据整合并确保其质量，是提升模型泛化能力的核心。然而，由于高质量数据集的构建需要大量资源投入，基于模仿学习获取的数据缺乏统一的量化标准，同时数据多样性与任务适配性难以兼顾，使得“以数据为中心”的研究范式在VLA场景中面临较高实施难度。因此，如何量化和管理数据质量，并结合合成和现实数据有效地提升模型表现，依然是VLA领域中的一个关键挑战。

在科研领域，基准测试往往成为重要的评判标准。一方面，基准测试的评价指标不一致使得VLA的泛化问题难以解决；另一方面，部分基准测试逐渐趋于性能饱和，可能掩盖模型在真实复杂场景中的实际进步。例如，一些开源模型在特定基准测试上已能够达到甚至超过π0.5（Black 等，2025a）等代表性模型的指标水平，但在面对真实零样本任务时，其实际表现仍与前沿封闭模型存在明显差距。因此，通过寻找优化基准测试评价指标的方法，改进基准测试的评价体系，可以更准确地衡量模型的综合能力，促进科研领域在实践中真正实现技术突破。

4.2 仿真到现实的迁移

在数据质量与评测体系之外，VLA真正落地过程中面临的重要现实障碍，是如何让仅在仿真中训练的策略可靠地过渡到真实世界。现阶段视觉—语言—动作模型在真实部署中仍受制于“仿真到现实”迁移的系统性差异，包括视觉感知偏移、动力学不匹配、语言指令分布变化及现实场景中新颖情境的出现。这些差异破坏了多模态语义的一致性，使模型在真实物理系统中的动作输出易偏离预期。特别是视觉域外观变化会导致物体属性与空间结构识别失准；语言域中口语化、非结构化表达难以被仿真中的模板式指令覆盖；而复杂动力学与开放环境更进一步加剧策略失效风险。

现有缓解策略主要包括多模态域随机化与跨域特征适应。前者通过扰动视觉纹理、光照和动力学属性，并引入多样化语言表达，以扩展训练分布并提升鲁棒性。但随机化需在语义与物理结构上保持约束，避免过度扰动导致任务关键语义被削弱。域适应则强调在视觉—语言联合特征空间中对齐跨域分布，通过对抗式判别器、跨域重建和自监督约束等方法学习共享语义表示，从而减少外观差异对动作推理的干扰。未来VLA系统需具备持续在线适应能力，通过少量真实交互实现快速更新，并结合元学习与知识蒸馏保持多模态语义稳定，避免在长期学习

中遗忘语言—动作映射。此外，多机器人平台间传感器噪声与物理差异将进一步放大迁移难度，因此构建具有结构化先验与跨平台一致性的VLA模型很可能成为未来的重要研究方向之一。

4.3 跨机器人适配性

在成功跨越仿真与现实差异之后，另一个更具系统性与工程挑战的问题随之显现，即模型能否在不同机器人平台之间保持一致的操作能力。现阶段的策略在跨机器人迁移时仍需额外适配甚至重新训练，这与实现真正通用的VLA模型存在明显差距，跨机器人适配性因此成为未来重要的研究方向之一。面向未来，相关研究主要体现在4个方面：1）在表征层面，现有研究逐渐探索构建可跨形体共享的抽象动作空间，使模型先预测与任务意图对应的动作表征，再由各机器人根据自身结构解码为可执行指令，从而提升在不同形态下的稳定性。2）在模型架构层面，一类具有代表性的思路是采用“通用主干 + 轻量适配模块”的分工模式，由主干负责视觉与语言理解，适配模块负责各平台的观测与动作映射，以降低跨机器人迁移所需的训练与计算开销。3）在数据与评测层面，已有研究逐步引入多机器人、多形体的大规模数据作为统一预训练基础，评测也从单平台性能扩展到跨平台一致性，以检验模型的跨机器人泛化能力。4）在部署层面，跨机器人适配有望与持续学习和在线调整机制相结合，使模型在少量交互下快速优化适配模块，同时尽可能保持主干能力的稳定性，从而提升在多形体场景下的适用性。

4.4 多任务干扰与遗忘问题

随着VLA模型在更大规模、多场景和多任务体系中的应用不断扩展，其训练与部署过程中愈发明显地暴露出多任务干扰与灾难性遗忘的问题。由于统一多模态架构通常在一个共享参数空间内同时吸收来自视觉、语言与动作三模态的异质信息，不同任务在动力学属性、视觉语义和动作分布上的巨大差异，会在训练过程中造成显著的参数竞争与表示冲突。这种竞争与冲突，使模型在学习新任务时覆盖甚至破坏先前获得的技能。相比纯语言、多模态问答等任务，具身领域的任务分布往往更加多样且跨尺度，导致VLA在面对复杂场景累计技能时面临更严重的稳定性—可塑性矛盾（stability-plasticity dilemma）。

这一问题在长序列任务或开放世界操作中尤为突出：由于VLA需要在同一模型内部持续整合多个任务的语义推理结构与动作控制策略，缺乏显式任务边界和长期记忆机制的统一序列模型常常无法保持跨任务表示的一致性，使得视觉定位、语义分解和动作生成三者在任务迁移中出现退化。此外，当模型在大规模多任务示范数据上进行混合训练时，视觉块与动作词向量之间的耦合关系容易在任务切换中发生漂移，进一步加剧训练不稳定性。

尽管已有研究尝试通过模块化技能结构、参数高效微调（如Adapter、LoRA）、跨任务表示正则化或经验重放等方法缓解多任务遗忘，但在统一Transformer或扩散式VLA中，这些技术仍缺乏系统化验证，尚难支撑真正意义上的长期任务积累。随着VLA规模的持续扩大，如何在统一架构中保持技能的可塑性与知识的长期稳定性，如何设计不破坏既有能力的增量式更新机制，以及如何在多模态空间中确保跨任务语义与动作表示的兼容性，构成了具身智能体迈向“终身学习（lifelong embodied learning）”过程中需要重点关注的问题。

4.5 多机协作与人机协同能力

随着机器人系统在真实世界场景中的部署规模不断扩大，单智能体的决策与执行能力已无法完全满足复杂环境下的需求，多机器人协作与人机协同逐渐成为具身智能的发展重点。然而，现有VLA架构主要面向单体机器人执行，缺乏跨主体的信息共享机制、分布式决策策略以及协作式规划框架，使其难以直接扩展至多主体、多角色和混合人机团队的任务形态。多机器人系统在视觉表征、局部可观性、通信带宽和策略一致性等方面的固有限制，使得传统统一序列模型难以承载多主体间的动态协作关系；而在人机协作场景中，机器人不仅需要理解自然语言指令，还需处理人类隐含意图、实时互动与安全约束，这些均超出了当前VLA所设计的单体范式。

在多机器人场景中，核心困难在于如何为不同主体构建共享的语义空间，使各机器人能够在部分可观测和异构感知条件下实现语义对齐与协同行动；如何在多主体动态策略中保持决策稳定性，避免因策略漂移而导致的非平稳性问题；以及如何让大型视觉—语言模型参与分布式任务分解、角色分配与协作规划，使语言推理能力不再局限于单体行为，而成为整个协作系统的策略生成器。在实际应用场景中，通信受限与实时性约束进一步加剧了多主体

协作的难度，对VLA的推理效率和跨主体信息通道提出了更高要求。

在人机协同方面，VLA虽然具备较强的语言理解与视觉感知能力，但缺乏对人类行为的预测建模、人类意图的深层表征与混合主动性（mixed-initiative）决策机制，使其难以实现与人类的流畅协作。真实环境中，人类的语言指令往往不完整、含糊甚至随任务进展而动态变化，机器人需要在任务执行中不断推理、校正并与人类进行双向沟通。针对复杂动态环境下的意图识别难题，蹇木伟等人（2026）提出了一种具身监测框架，通过多模态特征融合实现了对人类非正常状态的精准捕捉，这为VLA在高安全性要求任务中的意图表征提供了重要参考。同时，在交互的精细度与实时反馈方面，孟启帆（2025）探索了虚拟空间中的具身交互闭环，强调了动作映射与物理反馈的一致性，为VLA实现跨虚实空间的精细人机协同操作提供了技术启发。

未来VLA在协作智能方向的发展需要在统一多模态表达的基础上，构建跨主体共享语义空间，使多个机器人或机器人—人类主体能够在同一表示体系下进行信息交换；需要引入多主体强化学习与协作规划机制，使语言推理能够承担团队级决策的组织角色；并需要结合世界模型预测多个主体的未来状态，以支持前瞻性决策与安全协作。多机协作与人机协同能力将是VLA是否能够从“单体具身智能”迈向“协作型具身系统”的重要因素之一，并且是具身大模型向真实世界部署发展的关键能力。

4.6 欠缺情景学习

情景学习是VLA模型适配真实世界动态任务的核心能力，指模型能够根据任务上下文、环境状态和用户意图差异等情景变量，动态调整感知聚焦、语义理解与动作生成策略的过程——例如同样的“抓取杯子”指令，在桌面杂乱（需避开障碍物）、用户为老人（需轻力度抓取）、环境昏暗（需强化视觉边缘检测）等不同情景下，动作逻辑需完全适配，而非遵循固定“指令—动作”映射。但当前VLA模型在情景学习层面存在显著欠缺，难以突破“静态任务假设”的局限，成为其在复杂动态场景中鲁棒性不足的关键瓶颈。

首先，情景感知多停留在浅层化的“视觉—语言显性特征匹配”，未能捕捉隐性与动态情景变量：现有基于静态语义对齐（cross-modal relational alignment）的方法主要聚焦于对齐物体位置、任务目标等表层信息，忽略了用户特征（老人与儿童的动作力度、高度需求差异）、环境动态变化（中途新增障碍物未实时更新感知）以及任务上下文关联（“拿杯子”作为“倒水”前置步骤时，需预判杯口朝上姿态以适配后续动作），导致生成的动作与实际情景需求脱节。其次，情景知识存在固化问题，跨情景泛化能力薄弱：模型多依赖训练数据中的特定情景模板（如实验室整洁环境的操作逻辑），未将“避开障碍物”、“适配用户属性”等抽象为通用情景知识，一旦进入未见过的场景（如家庭杂乱桌面），动作成功率会大幅下降。此外，情景反馈机制缺失，闭环调整能力不足：多数VLA采用“指令输入—动作输出”的单向生成模式，未将动作执行后的情景反馈（如杯子滑落、用户避让等失败信号）纳入模型更新，无法修正前期偏差，甚至会重复执行错误动作。

这种情景学习能力的欠缺，直接导致VLA在真实动态场景中的鲁棒性不足，未来需通过引入情景感知模块、构建结构化情景知识体系以及设计反馈驱动的闭环学习机制等方向突破这一局限。

5 结 语

视觉—语言—动作模型的发展不仅为多模态具身智能带来了新的技术挑战，也为机器人应用的诸多新兴场景打开了更广阔的空间。当前，该领域仍处于快速演进阶段，尚有大量基础性问题值得深入研究。总体而言，VLA的发展正以统一的模型架构为核心、高质量多模态数据为基础，并以强化推理与行动的一致性为目标。在这一趋势的推动下，具备灵活适应与自主决策能力的具身智能体迈向真实世界的愿景正日益成为可能。

致谢：大连理工大学孔明骏、李梦昕、母炜坤和张梓康同学为本文撰写搜集和整理了资料。本文由中国图象图形学学会机器视觉专业委员会组织撰写，该专业委员会链接为https://www.csig.org.cn/16/201612/49315.html。

参考文献（References）

Agarwal N, Ali A, Bala M, Balaji Y, Barker E, Cai T, et al. 2025. COSMOS world foundation model platform for physical AI [EB/

OL]. [2025-11-21]. https://arxiv.org/pdf/2501.03575.pdf

Alayrac J B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. 2022. Flamingo: a visual language model for few-shot learning//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1723

Astruc G, Gonthier N, Mallet C and Landrieu L. 2024. OmniSat: self-supervised modality fusion for earth observation//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 409-427 [DOI: 10.1007/978-3-031-73390-1_24]

Banerjee P, Shkodrani S, Moulon P, Hampali S, Han S C, Zhang F, et al. 2025. HOT3D: hand and object tracking in 3D from egocentric multi-view videos//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 7061-7071 [DOI: 10.1109/CVPR52734.2025.00662]

Bao C, Xu H L, Qin Y Z and Wang X L. 2023. DexArt: benchmarking generalizable dexterous manipulation with articulated objects//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 21190-21200 [DOI: 10.1109/CVPR52729.2023.02030]

Bjorck J, Castañeda F, Cherniadev N, Da X Y, Ding R Y, Fan L J, et al. 2025a. GR00T N1.5: an improved open foundation model for generalist humanoid robots [EB/OL]. [2025-11-21]. https://research.nvidia.com/labs/gear/gr00t-n1_5/

Bjorck J, Castañeda F, Cherniadev N, Da X Y, Ding R Y, Fan L J, et al. 2025b. GR00T N1: an open foundation model for generalist humanoid robots [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2503.14734.pdf

Black K, Brown N, Darpinian J, Dhabalia K, Driess D, Adnan E, et al. 2025. $\pi_{0.5}$: a vision-language-action model with open-world generalization//Proceedings of the 9th Conference on Robot Learning. Seoul, Korea(South): PMLR: 17-40

Black K, Brown N, Driess D, Esmail A, Equi M, Finn C, et al. 2024. π0: a vision-language-action flow model for general robot control [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2410.24164.pdf

Black K, Galliker M Y and Levine S. 2025. Real-time execution of action chunking flow policies [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2506.07339.pdf

Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, Finn C, et al. 2023. RT-1: robotics transformer for real-world control at scale// Bekris K E, Hauser K, Herbert S L, Yu J J, eds. Robotics: Science and Systems XIX. Daegu, Korea(South): [s.n.]

Bu Q W, Cai J S, Chen L, Cui X Q, Ding Y, Feng S Y, et al. 2025a. AgiBot world Colosseo: a large-scale manipulation platform for scalable and intelligent embodied systems [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2503.06669.pdf

Bu Q W, Yang Y T, Cai J S, Gao S Y, Ren G H, Yao M Q, et al. 2025b. UnivLA: learning to act anywhere with task-centric latent actions [EB/OL]. [2025-11-22]. https://arxiv.org/pdf/2505.06111.pdf

Caesar H, Bankiti V, Lang A H, Vora S, Liong V E, Xu Q, et al. 2020. nuScenes: a multimodal dataset for autonomous driving//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11618-11628 [DOI: 10.1109/CVPR42600.2020.01164]

Carreira J, Noland E, Hillier C and Zisserman A. 2022. A short note on the kinetics-700 human action dataset[EB/OL]. [2025-11-21]. https://arxiv.org/pdf/1907.06987.pdf

Cen J, Yu C H, Yuan H J, Jiang Y M, Huang S T, Guo J Y, et al. 2025. WorldVLA: towards autoregressive action world model [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2506.21539.pdf

Chen T X, Chen Z X, Chen B J, Cai Z J, Liu Y B, Li Z X, et al. 2025a. RoboTwin 2.0: a scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2506.18088.pdf

Chen W, Belkhale S, Mirchandani S, Mees O, Driess D, Pertsch K, et al. 2025b. Training strategies for efficient embodied reasoning [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2505.08243.pdf

Chen W X, Liu Y, Chen B L, Su J D, Zheng Y S and Lin L. 2025c. Cross-modal causal relation alignment for video question grounding//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 24087-24096 [DOI: 10.1109/CVPR52734.2025.02243]

Chen Y H, Tian S, Liu S G, Zhou Y T, Li H R and Zhao D B. 2025d. ConRFT: a reinforced fine-tuning method for VLA models via consistency policy [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2502.05450.pdf

Chen Z Q, Yuan X, Mu T Z and Su H. 2025e. Responsive noise-relaying diffusion policy: responsive and efficient visuomotor control [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2502.12724.pdf

Chi C, Xu Z J, Feng S Y, Cousineau E, Du Y L, Burchfiel B, et al. 2025. Diffusion policy: visuomotor policy learning via action diffusion. The International Journal of Robotics Research, 44(10/11): 1684-1704 [DOI: 10.1177/02783649241273668]

Cui C, Ding P X, Song W X, Bai S H, Tong X Y, Ge Z R, et al. 2025. OpenHelix: a short survey, empirical analysis, and open-source dual-system VLA model for robotic manipulation [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2505.03912.pdf

Damen D, Doughty H, Farinella G M, Fidler S, Furnari A, Kazakos E, et al. 2021. The EPIC-KITCHENS dataset: collection, challenges and baselines. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11): 4125-4141 [DOI: 10.1109/TPAMI.2020.2991965]

Dasari S, Ebert F, Tian S, Nair S, Bucher B, Schmeckpeper K, et al. 2019. RoboNet: large-scale multi-robot learning//Proceedings of the 3rd Annual Conference on Robot Learning. Osaka, Japan: PMLR: 885-897

Deitke M, Han W, Herrasti A, Kembhavi A, Kolve E, Mottaghi R, et al. 2020. RoboTHOR: an open simulation-to-real embodied AI platform//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3161-3171 [DOI: 10.1109/CVPR42600.2020.00323]

Deitke M, Vander Bilt E, Herrasti A, Weihs L, Salvador J, Ehsani K, et al. 2022. ProcTHOR: large-scale embodied AI using procedural generation//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #433

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA: Association for Computational Linguistics: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Dong S Q, Fu C Y, Gao H H, Zhang Y F, Yan C, Wu C, et al. 2025. Vita-VLA: efficiently teaching vision-language models to act via action expert distillation [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2510.09607.pdf

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2021. An image is worth 16×16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. Vienna, Austria: ICLR

Dosovitskiy A, Ros G, Codevilla F, López A and Koltun V. 2017. CARLA: an open urban driving simulator//Proceedings of the 1st Annual Conference on Robot Learning. Mountain View, USA: PMLR: 1-16

Driess D, Xia F, Sajjadi M S, Lynch C, Chowdhery A, Ichter B, et al. 2023. PaLM-E: an embodied multimodal language model//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 340

Du Z Y, Liu B, Liang Y B, Shen Y C, Cao H D, Zheng X Y, et al. 2025. HiMoE-VLA: hierarchical mixture-of-experts for generalist vision-language-action policies [EB/OL]. [2025-11-21]. https://openreview.net/forum?id=TX3oGD99CJ

Duan Z K, Zhang Y, Geng S K, Liu G W, Boedecker J and Lu C X. 2025. Fast ECoT: efficient embodied chain-of-thought via thoughts reuse [EB/OL]. [2025-11-22]. https://arxiv.org/pdf/2506.07639.pdf

Ebert F, Yang Y L, Schmeckpeper K, Bucher B, Georgakis G, Daniilidis K, et al. 2021. Bridge data: boosting generalization of robotic skills with cross-domain datasets//Hauser K, Shell D A, Huang S D, eds. Robotics: Science and Systems XVIII. New York City, USA: [s.n.]

Ettinger S, Cheng S Y, Caine B, Liu C X, Zhao H, Pradhan S, et al. 2021. Large scale interactive motion forecasting for autonomous driving: the Waymo open motion dataset//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9690-9699 [DOI: 10.1109/ICCV48922.2021.00957]

Fan S C, Wu K, Che Z P, Wang X H, Wu D, Liao F, et al. 2025. XR-1: towards versatile vision-language-action models via learning unified vision-motion representations [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2511.02776.pdf

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

Fang H S, Fang H J, Tang Z Y, Liu J R, Wang C X, Wang J B, et al. 2023. RH20T: a comprehensive robotic dataset for learning diverse skills in one-shot//Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE: 653-660 [DOI: 10.1109/ICRA57147.2024.10611615]

Fang I, Chen Y, Wang Y, Zhang J, Zhang Q, Xu J, et al. 2024. EgoPAT3Dv2: predicting 3D action target from 2D egocentric vision for human-robot interaction//Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE: 3036-3043 [DOI: 10.1109/ICRA57147.2024.10610283]

Goyal R, Ebrahimi Kahou S, Michalski V, Materzynska J, Westphal S, Kim H, et al. 2017. The "something something" video database for learning and evaluating visual common sense//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5843-5851 [DOI: 10.1109/ICCV.2017.622]

Grauman K, Westbury A, Byrne E, Chavis Z, Furnari A, Girdhar R, et al. 2022. Ego4D: around the world in 3, 000 hours of egocentric video//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18973-18990 [DOI: 10.1109/CVPR52688.2022.01842]

Grauman K, Westbury A, Torresani L, Kitani K, Malik J, Afouras T, et al. 2024. Ego-Exo4D: understanding skilled human activity from first- and third-person perspectives//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, United States: IEEE: 19383-19400 [DOI: 10.1109/CVPR52733.2024.01834]

Guo Y J, Zhang J K, Chen X Y, Ji X, Wang Y J, Hu Y, et al. 2025. Improving vision-language-action model with online reinforcement learning//Proceedings of 2025 IEEE International Conference on Robotics and Automation (ICRA). Atlanta, USA: IEEE: 15665-15672 [DOI: 10.1109/ICRA55743.2025.11127299]

Hancock A J, Wu X D, Zha L H, Russakovsky O and Majumdar A. 2025. Actions as language: fine-tuning VLMs into VLAs without catastrophic forgetting [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2509.22195.pdf

Hinton G E. 1986. Learning distributed representations of concepts//Pro-

ceedings of the 8th Annual Conference of the Cognitive Science Society. Amherst, USA: Erlbaum Associates: 1-12

Hu E J, Shen Y L, Wallis P, Allen Zhu Z, Li Y Z, Wang S, et al. 2021. LoRA: low-rank adaptation of large language models//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: OpenReview.net

Huang J, Ping W, Xu P, Shoeybi M, Chang K C C and Catanzaro B. 2024. RAVEN: in-context learning with retrieval-augmented encoder-decoder language models [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2308.07922.pdf

Huang W L, Xia F, Xiao T, Chan H, Liang J, Florence P, et al. 2023. Inner monologue: embodied reasoning through planning with language models//Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR: 1769-1782

Ichter B, Brohan A, Chebotar Y, Finn C, Hausman K, Herzog A, et al. 2022. Do as I can, not as I say: grounding language in robotic affordances//Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR: 287-318

James S, Ma Z C, Arrojo D R and Davison A J. 2020. RLBench: the robot learning benchmark and learning environment. IEEE Robotics and Automation Letters, 5(2): 3019-3026 [DOI: 10.1109/LRA.2020.2974707]

Jang E, Irpan A, Khansari M, Kappler D, Ebert F, Lynch C, et al. 2022. BC-Z: Zero-Shot task generalization with robotic imitation learning//Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR: 991-1002

Jian M W, Ling Y K, Zhang H R, Zhang L S and Ma J J. Embodied intelligence-driven distracted driving detection: a framework and research prospects [J/OL]. Journal of Image and Graphics, 2026: 1-16. https://www.cjig.cn/zh/article/doi/10.11834/jig.250514/ (蹇木伟, 凌钰坤, 张昊然, 张琳松, 马嘉骏. 具身智能驱动的分心驾驶检测: 框架研究与前沿展望 [J/OL]. 中国图象图形学报, 2026: 1-16). https://www.cjig.cn/zh/article/doi/10.11834/jig.250514/

Jiang T T, Jiang X F, Ma Y, Wen X, Li B L, Zhan K, et al. 2025a. The better you learn, the smarter you prune: towards efficient vision-language-action models via differentiable token pruning [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2509.12594.pdf

Jiang Y F, Gupta A, Zhang Z C, Wang G Z, Dou Y Q, Chen Y J, et al. 2023. VIMA: robot manipulation with multimodal prompts//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR: 14975-15022

Jiang Z Y, Xie Y Q, Lin K, Xu Z J, Wan W K, Mandlekar A, et al. 2025b. DexMimicGen: automated data generation for bimanual dexterous manipulation via imitation learning//Proceedings of 2025 IEEE International Conference on Robotics and Automation (ICRA). Atlanta, USA: IEEE: 16923-16930 [DOI: 10.1109/ICRA55743.2025.11127809]

Jones J, Mees O, Sferrazza C, Stachowicz K, Abbeel P and Levine S. 2025. Beyond sight: finetuning generalist robot policies with heterogeneous sensors via language grounding//Proceedings of 2025 IEEE International Conference on Robotics and Automation (ICRA). Atlanta, USA: IEEE: 5961-5968 [DOI: 10.1109/ICRA55743.2025.11127987]

Kalashnikov D, Irpan A, Pastor P, Ibarz J, Herzog A, Jang E, et al. 2018. Scalable deep reinforcement learning for vision-based robotic manipulation//Proceedings of the 2nd Conference on Robot Learning. Zürich, Switzerland: PMLR: 651-673

Kalashnikov D, Varley J, Chebotar Y, Swanson B, Jonschkowski R, Finn C, et al. 2022. Scaling up multi-task robotic reinforcement learning//Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR: 557-575

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//Proceedings of Robotics: Science and Systems (RSS 2024. Delft, Netherlands: RSS: #120 [DOI: 10.15607/RSS.2024.XX.120]

Kim M J, Finn C and Liang P. 2025. Fine-tuning vision-language-action models: optimizing speed and success [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2502.19645.pdf

Kim M J, Pertsch K, Karamcheti S, Xiao T, Balakrishna A, Nair S, et al. 2024. OpenVLA: an open-source vision-language-action model//Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR: 2679-2713

Kolve E, Mottaghi R, Han W, VanderBilt E, Weihs L, Herrasti A, et al. 2022. AI2-THOR: an interactive 3D environment for visual AI [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/1712.05474.pdf

Kumar V, Shah R, Zhou G Y, Moens V, Caggiano V, Vakil J, et al. 2023. RoboHive: a unified framework for robot learning//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1918

Kwon T, Tekin B, Stühmer J, Bogo F and Pollefeys M. 2021. H2O: two hands manipulating objects for first person interaction recognition//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 10118-10128 [DOI: 10.1109/ICCV48922.2021.00998]

Li C S, Xia F, Martín-Martín R, Lingelbach M, Srivastava S, Shen B K, et al. 2022a. iGibson 2.0: object-centric simulation for robot learning of everyday household tasks//Proceedings of the 5th Conference on Robot Learning (CoRL 2021). London, UK: PMLR: 455-465

Li H T, Ding P X, Suo R Z, Wang Y H, Ge Z R, Zang D Y, et al. 2025a. VLA-RFT: vision-language-action reinforcement fine-tuning with verified rewards in world simulators [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2510.00406.pdf

Li J N, Li D X, Xiong C M and Hoi S C H. 2022b. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 12888-12900

Li Y, Meng Y, Sun Z W, Ji K Y, Tang C, Fan J J, et al. 2025b. SP-

VLA: a joint model scheduling and token pruning approach for VLA model acceleration [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2506.12723.pdf

Liang J, Huang W L, Xia F, Xu P, Hausman K, Ichter B, et al. 2023. Code as policies: language model programs for embodied control//Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA). London, UK: IEEE: 9493-9500 [DOI: 10.1109/ICRA48891.2023.10160591]

Liang Z X, Li Y Z, Yang T S, Wu C Y, Mao S T, Nian T, et al. 2025. Discrete diffusion VLA: bringing discrete diffusion to action decoding in vision-language-action policies [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2508.20072.pdf

Lin J Y, Taherin A, Akbari A, Akbari A, Lu L, Chen G Y, et al. 2025. VOTE: vision-language-action optimization with trajectory ensemble voting [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2507.05116.pdf

Liu B, Zhu Y F, Gao C K, Feng Y H, Liu Q, Zhu Y K, et al. 2023a. LIBERO: benchmarking knowledge transfer for lifelong robot learning//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1939

Liu C H, Zhang J C, Li C X, Zhou Z M, Wu S X, Huang S F, et al. 2025. TTF-VLA: temporal token fusion via pixel-attention integration for vision-language-action models [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2508.19257.pdf

Liu G H, Cui J Z, Lu Y C, Hu J J, Xie Q, Guo Y J, et al. 2025. Cross-modal interaction-driven embodied intelligence: a review of vision-language-action models, data, and platforms. Journal of Integration Technology (刘国华, 崔纪泽, 陆勇辰, 胡军军, 谢强, 郭媛君, 等. 2025. 跨模态交互驱动的具身智能: 视觉—语言—动作融合模型、数据与平台综述. 集成技术) [DOI: 10.12146/j.issn.2095-3135.20250923001]

Liu H T, Li C Y, Wu Q Y and Lee Y J. 2023b. Visual instruction tuning//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1516

Liu J M, Liu M Z, Wang Z Y, An P J, Li X Q, Zhou K C, et al. 2024. RoboMamba: efficient vision-language-action model for robotic reasoning and manipulation//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #1266

Liu Y Z, Liu Y, Jiang C, Lyu K B, Wan W K, Shen H, et al. 2022. HOI4D: a 4D egocentric dataset for category-level human-object interaction//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 20981-20990 [DOI: 10.1109/CVPR52688.2022.02034]

Lu G X, Guo W K, Zhang C B, Zhou Y H, Jiang H N, Gao Z F, et al. 2025. VLA-RL: towards masterful and general robotic manipulation with scalable reinforcement learning [EB/OL]. [2025-05-24]. https://arxiv.org/pdf/2505.18719.pdf

Luo H, Feng Y C, Zhang W P, Zheng S P, Wang Y, Yuan H Q, et al. 2025. Being-H0: vision-language-action pretraining from large-scale human videos [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2507.15597.pdf

Lyu Z Y, Charron N, Moulon P, Gamino A, Peng C, Sweeney C, et al. 2024. Aria everyday activities dataset [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2402.13349.pdf

Ly K T, Lu K and Havoutis I. 2026. InteLiPlan: an interactive lightweight LLM-based planner for domestic robot autonomy. IEEE Robotics and Automation Letters, 11(3): 3875-3882 [DOI: 10.1109/LRA.2026.3662577]

Makoviychuk V, Wawrzyniak L, Guo Y R, Lu M, Storey K, Macklin M, et al. 2021. Isaac Gym: high performance GPU based physics simulation for robot learning//Proceedings of the 35th Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.

Mandlekar A, Nasiriany S, Wen B W, Akinola I, Narang Y, Fan L X, et al. 2023. MimicGen: a data generation system for scalable robot learning using human demonstrations//Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR: 1820-1864

Mandlekar A, Zhu Y K, Garg A, Booher J, Spero M, Tung A, et al. 2018. ROBOTURK: a crowdsourcing platform for robotic skill learning through imitation//Proceedings of the 2nd Conference on Robot Learning. Zurich, Switzerland: PMLR: 879-893

Mazzaglia P, Sancaktar C, Peschl M and Dijkman D. 2025. Hybrid training for vision-language-action models [EB/OL]. [2025-10-01]. https://arxiv.org/pdf/2510.00600.pdf

Mees O, Hermann L, Rosete-Beas E and Burgard W B. 2022. CALVIN: a benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics and Automation Letters, 7(3): 7327-7334 [DOI: 10.1109/LRA.2022.3180108]

Meng Q F, Zhang Y R, Zhang H W, Hu X Y and Luo Y H. Research on virtual embodied interaction technology for VR physics experiments. Journal of Image and Graphics, 2025: 1-12 (孟启帆, 张怡冉, 张鸿文, 胡晓雁, 骆岩红. 面向VR物理实验的虚拟具身交互技术研究. 中国图象图形学报, 2025: 1-12) [DOI: 10.11834/jig.250425]

Miech A, Zhukov D, Alayrac J B, Tapaswi M, Laptev I and Sivic J. 2019. HowTo100M: learning a text-video embedding by watching hundred million narrated video clips//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE: 2630-2640 [DOI: 10.1109/ICCV.2019.00272]

Mittal M, Roth P, Tigue J, Richard A, Zhang O, Du P, et al. 2025. Isaac Lab: a GPU-accelerated simulation framework for multi-modal robot learning [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2511.04831.pdf

Mu T Z, Ling Z, Xiang F B, Yang D, Li X L, Tao S, et al. 2021. ManiSkill: generalizable manipulation skill benchmark with large-scale

demonstrations//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.

Nasiriany S, Maddukuri A, Zhang LC, Parikh A, Lo A, Joshi A, et al. 2024. RoboCasa: large-scale simulation of everyday tasks for generalist robots//Robotics: Science and Systems 2024. Delft, Netherlands: [s.n.]

O'Neill A, Rehman A, Gupta A, Maddukuri A, Gupta A, Padalkar A, et al. 2023. Open X-Embodiment: robotic learning datasets and RT-X models [EB/OL]. [2025-11-22]. https://arxiv.org/pdf/2310.08864.pdf

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]//Proceedings of 2024 IEEE International Conference on Robotics and Automation. Yokohama, Japan: IEEE: 6892-6903 [DOI: 10.1109/ICRA57147.2024.10611477]

Octo Model Team, Ghosh D, Walke H, Pertsch K, Black K, Mees O, et al. 2024. Octo: an open-source generalist robot policy//Proceedings of Robotics: Science and Systems 2024. Delft, Netherlands: RSS: #90 [DOI: 10.15607/RSS.2024.XX.090]

Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. 2025. DINOv2: learning robust visual features without supervision//Proceedings of the 13th International Conference on Learning Representations (ICLR 2025). Singapore: TMLR: 2835-8856

Perrett T, Darkhalil A, Sinha S, Emara O, Pollard S, Parida K K, et al. 2025. HD-EPIC: a highly-detailed egocentric video dataset//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 23901-23913 [DOI: 10.1109/CVPR52734.2025.02226]

Puig X, Undersander E, Szot A, Dallaire-Cote M, Yang T Y, Partsey R, et al. 2023. Habitat 3.0: a co-habitat for humans, avatars, and robots//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Pumacay W, Singh I, Duan J F, Krishna R, Thomason J and Fox D. 2024. THE COLOSSEUM: a benchmark for evaluating generalization for robotic manipulation//Robotics: Science and Systems 2024. Delft, Netherlands: RSS: #133 [DOI: 10.15607/RSS.2024.XX.133]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8748-8763

Radford A, Narasimhan K, Salimans T and Sutskever I. 2018. Improving language understanding by generative pre-training [EB/OL]. [2025-11-21]. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Rong G D, Shin B H, Tabatabaee H, Lu Q, Lemke S, Možeiko M, et al. 2020. LGSVL simulator: a high fidelity simulator for autonomous driving//Proceedings of the 23rd IEEE International Conference on Intelligent Transportation Systems (ITSC). Rhodes, Greece: IEEE: 1-6 [DOI: 10.1109/ITSC45102.2020.9294422]

Savva M, Kadian A, Maksymets O, Zhao Y L, Wijmans E, Jain B, et al. 2019. Habitat: a platform for embodied AI research//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9338-9346 [DOI: 10.1109/ICCV.2019.00943]

Sharma P, Mohan L, Pinto L and Gupta A. 2018. Multiple interactions made easy (MIME): large scale demonstrations data for imitation//Proceedings of the 2nd Conference on Robot Learning. Zurich, Switzerland: PMLR: 906-915

Shen B K, Xia F, Li C S, Martín-Martín R, Fan L X, Wang G Z, et al. 2021. iGibson 1.0: a simulation environment for interactive tasks in large realistic scenes//Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague, Czech Republic: IEEE: 7520-7527 [DOI: 10.1109/IROS51168.2021.9636667]

Shridhar M, Manuelli L and Fox D. 2022. CLIPort: what and where pathways for robotic manipulation//Proceedings of the 5th Conference on Robot Learning. New Orleans, USA: PMLR: 894-906

Shridhar M, Manuelli L and Fox D. 2023. Perceiver-actor: a multi-task transformer for robotic manipulation//Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR: 785-799

Shukla A, Tao S and Su H. 2025. ManiSkill-HAB: a benchmark for low-level manipulation in home rearrangement tasks//Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net

Shukor M, Aubakirova D, Capuano F, Kooijmans P, Palma S, Zouitine A, et al. 2025. SMOLVLA: a vision-language-action model for affordable and efficient robotics [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2506.01844.pdf

Singh A, Hu R H, Goswami V, Couairon G, Galuba W, Rohrbach M, et al. 2022. FLAVA: a foundational language and vision alignment model//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 15617-15629 [DOI: 10.1109/CVPR52688.2022.01519]

Song W X, Chen J Y, Ding P X, Huang Y X, Zhao H, Wang D L, et al. 2025a. CEED-VLA: consistency vision-language-action model with early-exit decoding [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2506.13725.pdf

Song Z R, Ouyang G X, Li M Z, Ji Y H, Wang C X, Xu Z X, et al. 2025b. ManiPLVM-R1: reinforcement learning for reasoning in embodied manipulation with large vision-language models [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2505.16517.pdf

Stepputtis S, Campbell J, Phielipp M J, Lee S, Baral C and Ben Amor H. 2020. Language-conditioned imitation learning for robot manipulation tasks//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #1102

Sutskever I, Vinyals O and Le Q V. 2014. Sequence to sequence learn-

ing with neural networks//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 3104-3112

Szot A, Clegg A, Undersander E, Wijmans E, Zhao Y L, Turner J, et al. 2021. Habitat 2.0: training home assistants to rearrange their habitat//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.: #20

Tan X D, Yang Y X, Ye P, Zheng J L, Bai B Z, Wang X Y, et al. 2025. Think twice, act once: token-aware compression and action reuse for efficient inference in vision-language-action models [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2505.21200.pdf

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Vemprala S H, Bonatti R, Bucker A and Kapoor A. 2024. ChatGPT for robotics: design principles and model abilities. IEEE Access, 12: 55682-55696 [DOI: 10.1109/ACCESS.2024.3387941]

Walke H R, Black K, Zhao T Z, Vuong Q, Zheng C Y, Hansen-Estruch P, et al. 2024. BridgeData V2: a dataset for robot learning at scale//Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR: 1723-1736

Wang G Z, Xie Y Q, Jiang Y F, Mandlekar A, Xiao C W, Zhu Y K, et al. 2023a. Voyager: an open-ended embodied agent with large language models [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2305.16291.pdf

Wang H Y, Xiong C Y, Wang R P and Chen X L. 2025a. BitVLA: 1-bit vision-language-action models for robotics manipulation [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2506.07530

Wang J F, Yang Z Y, Hu X W, Li L J, Lin K, Gan Z, et al. 2022. GIT: a generative image-to-text transformer for vision and language [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2205.14100.pdf

Wang X Z, Wei J, Schuurmans D, Le Q V, Chi E, Narang S, et al. 2023b. Self-consistency improves chain of thought reasoning in language models//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net

Wang Y, Luo W J, Bai J J, Cao Y L, Che T, Chen K, et al. 2025b. Alpamayo-R1: bridging reasoning and action prediction for generalizable autonomous driving in the long tail [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2511.00088.pdf

Wang Y T, Zhu H Y, Liu M Y, Yang J G, Fang H S and He T. 2025c. VQ-VLA: improving vision-language-action models via scaling vector-quantized action tokenizers [EB/OL]. [2025-11-22]. https://arxiv.org/pdf/2507.01016.pdf

Wei J, Wang X Z, Schuurmans D, Bosma M, Ichter B, Xia F, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1800 [DOI: 10.5555/3600270.3602070]

Wen J J, Zhu M J, Liu J M, Liu Z Y, Yang Y C, Zhang L F, et al. 2025a. dVLA: diffusion vision-language-action model with multimodal chain-of-thought [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2509.25681.pdf

Wen J J, Zhu Y C, Li J M, Zhu M J, Tang Z B, Wu K, et al. 2025b. TinyVLA: toward fast, data-efficient vision-language-action models for robotic manipulation. IEEE Robotics and Automation Letters, 10(4): 3988-3995 [DOI: 10.1109/LRA.2025.3544909]

Wen J J, Zhu Y C, Zhu M J, Tang Z B, Li J M, Zhou Z Y, et al. 2025c. DiffusionVLA: scaling robot foundation models via unified diffusion and autoregression//Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: OpenReview.net

Wu K, Hou C K, Liu J M, Che Z P, Ju X Z, Yang ZZ, et al. 2025. RoboMIND: benchmark on multi-embodiment intelligence normative data for robot manipulation//Proceedings of the Robotics: Science and Systems 2025. Los Angeles, USA: Robotics: Science and Systems Foundation: 152-163

Xia F, Zamir A R, He Z, Sax A, Malik J and Savarese S. 2020. iGibson: a simulation environment for interactive tasks in large realistic scenes//Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas, United States: IEEE

Xiang F B, Qin Y Z, Mo K C, Xia Y K, Zhu H, Liu F C, et al. 2020. SAPIEN: a simulated part-based interactive environment//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11094-11104 [DOI: 10.1109/CVPR42600.2020.01111]

Xu S Y, Wang Y K, Xia C H, Zhu D H, Huang T and Xu C. 2025. VLA-Cache: efficient vision-language-action manipulation via adaptive token caching//Proceedings of the 39th International Conference on Neural Information Processing Systems

Xue H R, Huang X Y, Niu D T, Liao Q Y, Kragerud T, Gravdahl J T, et al. 2025. LeVERB: humanoid whole-body control with latent vision-language instruction [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2506.13751.pdf

Yang J W, Tan R B, Wu Q H, Zheng R J, Peng B L, Liang Y Y, et al. 2025. Magma: a foundation model for multimodal AI agents//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, United States: IEEE: 14203-14214 [DOI: 10.1109/CVPR52734.2025.01325]

Yao S Y, Yu D, Zhao J, Shafran I, Griffiths T L, Cao Y A, et al. 2023. Tree of thoughts: deliberate problem solving with large language models//Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc.: #517

Yu T H, Quillen D, He Z P, Julian R, Narayan A, Shively H, et al. 2021. Meta-World: a benchmark and evaluation for multi-task and meta reinforcement learning//Proceedings of the 3rd Annual Conference on Robot Learning. Osaka, Japan: PMLR: 1094-1100

Yue Y, Wang Y L, Kang B Y, Han Y Z, Wang S Z, Song S J, et al. 2024. DEER-VLA: dynamic inference of multimodal large lan-

guage models for efficient robot execution//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #1803

Zawalski M, Chen W, Pertsch K, Mees O, Finn C and Levine S. 2025. Robotic control via embodied chain-of-thought reasoning//Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR: 3157-3181

Zhai S P, Zhang Q, Zhang T Y, Huang F X, Zhang H R, Zhou M, et al. 2025. A vision-language-action-critic model for robotic real-world reinforcement learning [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2509.15937.pdf

Zhan X Y, Yang L X, Zhao Y F, Mao K R, Xu H L, Lin Z N, et al. 2024. OakInk2: a dataset of bimanual hands-object manipulation in complex task completion//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 445-456 [DOI: 10.1109/CVPR52733.2024.00050]

Zhang B R, Zhang Y H, Ji J M, Lei Y S, Dai J, Chen Y P, et al. 2025a. SafeVLA: towards safety alignment of vision-language-action model via constrained learning [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2503.03480.pdf

Zhang H, Liang S T, Li M X, Tian Y L, Ge J W, Yu H, et al. 2025. 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 Y, Zhuang Z F, Zhao H, Ding P X, Lu H C and Wang D L. 2025b. ReinboT: amplifying robot visual-language manipulation with reinforcement learning//Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: OpenReview.net

Zhang R Y, Dong M H, Zhang Y, Heng L, Chi X W, Dai G L, et al. 2025c. MOLE-VLA: dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2503.20384.pdf

Zhang S D, Xu Z, Liu P J, Yu X P, Li Y, Gao Q H, et al. 2025d. VLABench: a large-scale benchmark for language-conditioned robotic manipulation with long-horizon reasoning tasks//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. Miami, USA: IEEE: 11142-11152

Zhang S Q, Wicke P, Şenel L K, Figueredo L, Naceri A, Haddadin S, et al. 2023. LoHoRavens: a long-horizon language-conditioned benchmark for robotic tabletop manipulation [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2310.12020.pdf

Zhang W Y, Liu H S, Qi Z K, Wang Y N, Yu X Q, Zhang J Z, et al. 2025e. DreamVLA: a vision-language-action model dreamed with comprehensive world knowledge [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2507.04447.pdf

Zheng J L, Li J X, Wang Z H, Liu D X, Kang X R, Feng Y C, et al. 2025. X-VLA: soft-prompted transformer as scalable cross-embodiment vision-language-action model [EB/OL]. [2025-11-21]. https://arxiv.org/pdf /2510.10274.pdf

Zheng L M, Yan F, Liu F F, Feng C J, Kang Z L and Ma L. 2024. RoboCAS: a benchmark for robotic manipulation in complex object arrangement scenarios [EB/OL]. [2025-11-21]. https://arxiv.org/pdf/2407.06951.pdf

Zhou Z Y, Zhu Y C, Zhu M J, Wen J J, Liu N, Xu Z Y, et al. 2025. ChatVLA: unified multimodal understanding and robot control with vision-language-action model//Proceedings of 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: ACL: 5377-5395 [DOI: 10.18653/v1/2025. emnlp-main.273]

Zhu Y K, Wong J, Mandlekar A, Martín-Martín R, Joshi A, Lin K, et al. 2020. Robosuite: a modular simulation framework and benchmark for robot learning [EB/OL]. [2025-11-21]. https://arxiv.org/pdf /2009.12293. pdf

Zitkovich B, Yu T H, Xu S C, Xu P, Xiao T, Xia F, et al. 2023. RT-2: vision-language-action models transfer web knowledge to robotic control//Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR: 2165-2183

作者简介

何友，男，教授，主要研究方向为信号检测、信息融合、智能技术与应用研究。E-mail: heyou@mail.tsinghua.edu.cn

王栋，通信作者，男，教授，主要研究方向为计算机视觉、机器学习和机器人视觉。E-mail: wdice@dlut.edu.cn

卢湖川，男，教授，主要研究方向为计算机视觉、机器学习和模式识别。E-mail: lhchuan@dlut.edu.cn

李劭辉，男，研究员，主要研究方向为信号处理、人工智能、图像/视频压缩，以及多智能体系统。

E-mail: lishaohui@zju.edu.cn

李徵，男，助理教授，主要研究方向为数据挖掘、知识计算与多智能体系统。E-mail: zhilizl@sz.tsinghua.edu.cn

刘洋，男，副教授，主要研究方向为智能机器人感知与导航、智能集群对抗博弈、深度学习与大模型。

E-mail: ly@dlut.edu.cn

赵洁，女，助理研究员，主要研究方向为计算机视觉、视觉目标跟踪和机器人视觉。E-mail: zj982853200@dlut.edu.cn

阮书岚，男，助理研究员，主要研究方向为大模型驱动的多模态理解、内容生成及多智能体系统。

E-mail: slruan@sz.tsinghua.edu.cn