中图法分类号:TP18;TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2026)08-2731-43

论文引用格式:Liu R, Wang W G, Wang J and Yang Y. 2026. Survey on embodied intelligence: advances in vision-language-action models. Journal of Image and Graphics, 31(8):2731-2773(刘锐, 王文冠, 王俊, 杨易. 2026. 面向具身智能的视觉—语言—动作模型研究进展. 中国图象图形学报, 31(8):2731-2773)[DOI:10.11834/jig.250544]

面向具身智能的视觉—语言—动作模型研究进展

刘锐[1,2],王文冠[1,2],王俊[3],杨易[1,2*]

1. 浙江大学脑机智能全国重点实验室,杭州　310027; 2. 浙江大学人工智能学院,杭州　310027; 3. OPPO研究院,深圳　518000

摘　要:具身智能作为人工智能与机器人学交叉的前沿领域,旨在使智能体通过与物理世界的动态交互来感知、推理并执行任务。然而,传统基于深度学习的级联式感知—控制模型在开放、动态环境下泛化能力不足,且高度依赖大规模标注数据。近年来,视觉—语言—动作模型(vision-language-action models, VLA)通过融合视觉感知、语言理解与动作生成,为具身智能的研究与应用提供了新的推动力。本文系统梳理了基于VLA的具身智能研究进展,从发展历程、模型架构、系统分类、训练与评估等方面展开综述。首先,本文追溯了视觉与语言基础模型的演进脉络,并阐述VLA概念的提出背景;随后,本文深入剖析VLA的关键技术模块,包括视觉编码、语言表征及动作词元化与解码机制;在此基础上,引入系统架构分类法,将现有工作归纳为单系统、双系统与层次化3类,并分析其设计权衡与适用场景;此外,本文总结了模型的预训练与后训练策略,并梳理了仿真及真实环境下的主流评测基准;最后,分析了VLA在实时推理效率、数据质量、环境泛化性与安全伦理等维度面临的挑战,并展望从被动感知到主动推理、持续学习、场景泛化与可靠部署等未来方向。本文旨在为相关研究者提供系统的技术参考,推动VLA在开放世界具身任务中的理论发展和应用落地。本文提及的算法、数据集和评估指标已汇总至https://github.com/DefaultRui/vision-language-action-models-for-embodied-AI。

关键词:具身智能;多模态大模型;视觉—语言—动作模型(VLA);强化学习;机器人感知与决策

Survey on embodied intelligence: advances in vision-language-action models

Liu Rui[1,2], Wang Wenguan[1,2], Wang Jun[3], Yang Yi[1,2*]

1. The State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou　310027, China; 2. College of Artificial Intelligence, Zhejiang University, Hangzhou　310027, China; 3. OPPO Research Institute, Shenzhen　518000, China

Abstract: Embodied artificial intelligence aims to endow agents with the ability to perceive, reason, and act through continuous interaction with the physical world. In contrast with disembodied intelligence, which mostly operates on static data and symbolic tasks, embodied intelligence must handle sensorimotor coupling, environmental uncertainty, long-horizon decision-making, and real-time control in open-world settings. These requirements expose the limitations of traditional perception-planning-control pipelines. Conventional robotic systems are often built from cascaded modules for visual perception, language understanding, state estimation, task planning, and low-level control. Although such pipelines benefit from modular engineering and interpretability, they frequently suffer from error accumulation across stages, brittle generalization to novel tasks and scenes, and heavy dependence on manually designed interfaces, reward functions, and domain-

收稿日期:2025-11-17;**修回日期**:2026-04-03;**预印本日期**:2026-04-10

*通信作者:杨易　yangyics@zju.edu.cn

基金项目:教育部基础学科与交叉学科突破计划(JYB2025XDXM103)

Supported by: Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM103)

specific annotations. Recent advances in large-scale foundation models have opened a new path for embodied intelligence by enabling the integration of visual grounding, language reasoning, and action generation within a shared model space. This trend has led to the rapid emergence of vision-language-action (VLA) models, which aim to translate multimodal observations and task instructions directly into executable actions or action-conditioned policies, narrowing the gap between semantic understanding and physical execution. This survey presents a systematic review of VLA models in embodied intelligence from the perspectives of historical development, technical architecture, system design, training strategies, evaluation protocols, and future trends. We first revisit the technological trajectory that made VLA possible. The evolution of large vision models and large language models has substantially improved semantic representation, multimodal alignment, and knowledge transfer, while progress in robot learning has highlighted the importance of scalable policy learning from heterogeneous data sources. Within this context, VLA models can be considered the convergence of three lines of research: multimodal foundation models for perception and reasoning, imitation and reinforcement learning for embodied decision-making, and robot action modeling for continuous control. Their significance lies not merely in adding language prompts to robot policies, but also in establishing a unified interface through which high-level task intent, environmental perception, and low-level action generation can be jointly optimized. A central objective of this survey is to clarify the internal structure of VLA systems. We analyze the key technical modules that constitute current VLA pipelines. On the perception side, VLA models rely on visual encoders to transform raw observations into compact and task-relevant representations. These observations may include RGB images, depth maps, multi-view input, temporal image sequences, or structured spatial representations, such as 3D or bird's-eye-view features. The choice of visual representation directly influences the model's capability to capture object semantics, spatial relations, affordances, and scene dynamics. On the language side, textual encoders provide instruction understanding, goal specification, contextual reasoning, and cross-modal alignment. Language not only describes tasks but also offers a flexible interface for compositional generalization, abstract planning, and human-robot interaction. The crucial problem of action representation lies between perception and control. Compared with visual and textual tokens, action signals are continuous, temporally dense, and strongly constrained by embodiment. Therefore, we pay special attention to action tokenization and decoding, reviewing how actions can be represented as discrete tokens, continuous vectors, chunked trajectories, latent action codes, or generative distributions. We further compare the major decoding paradigms, including direct regression, classification after discretization, autoregressive sequence modeling, diffusion-based policy generation, and flow-based generation, and then discuss their respective trade-offs in precision, expressivity, stability, and inference speed. Beyond module-level analysis, this survey proposes a system-level taxonomy that categorizes existing VLA models into three broad classes: single systems, dual-system architectures, and hierarchical systems. Single systems integrate perception, reasoning, and action generation within a unified model, and they are attractive for end-to-end deployment, parameter sharing, and simplified optimization. However, they may experience difficulties in balancing high-level deliberation with fast control. Dual-system architectures address this issue by explicitly separating slow reasoning from fast acting, frequently coupling a deliberative module for semantic planning with a lightweight controller for reactive execution. This design is particularly relevant to robotics because reasoning and motor control operate under different temporal constraints. Hierarchical systems introduce structured intermediate abstractions, such as subgoals, skills, waypoints, object-centric plans, or symbolic task graphs, to improve modularity, interpretability, and reuse across tasks. We argue that this taxonomy provides a useful perspective for understanding the design space of VLA models and analyzing how different systems trade off generalization, efficiency, and robustness. The survey also reviews the training and evaluation ecosystem that supports VLA research. Current VLA models are typically trained through multi-stage pipelines that involve internet-scale visual-language corpora, simulated interactive data, teleoperated demonstrations, and real robot trajectories. Pretraining equips the model with broad perceptual and linguistic priors, while post-training and task-specific adaptation align those priors with embodiment, action semantics, and control requirements. We summarize commonly used strategies, such as supervised imitation learning, behavior cloning, instruction tuning, parameter-efficient fine-tuning, model distillation, reinforcement learning augmentation, and hybrid training paradigms that combine offline and online data. We further discuss the role of simulator-based benchmarks and real-world evaluations. Simulation enables scalable and reproducible testing under controlled settings, while real robot experiments reveal

embodiment gaps, latency constraints, and deployment failures that are not fully captured by simulation alone. In reviewing the literature, we emphasize that cross-paper comparisons must be interpreted with caution because reported results are often influenced by differences in robot morphology, sensor configurations, action horizons, training data quality, evaluation metrics, hardware platforms, and implementation details. Despite their remarkable progress, VLA models still face several fundamental challenges. First, real-time inference remains a major bottleneck. Many powerful multimodal models rely on autoregressive decoding or iterative generative sampling, which may be too slow for high-frequency control. Second, action learning suffers from limited robotic data, inconsistent annotation quality, and poor cross-platform transfer, especially when embodiments differ substantially. Third, environmental generalization is far from solved. Models trained on curated datasets may degrade sharply in cluttered scenes, long-horizon tasks, dynamic environments, or safety-critical settings where perception is ambiguous and recovery from failure is required. Fourth, current VLA systems still exhibit limited causal reasoning and world modeling capabilities. Many methods align observations and instructions effectively but lack deep predictive understanding of object dynamics, interaction consequences, and long-term task structure. Fifth, evaluation remains fragmented. Existing studies often emphasize task success rate while underreporting execution frequency, safety violations, intervention cost, recovery capability, and robustness under distribution shift. Finally, broader issues of trustworthiness, including interpretability, failure diagnosis, ethical risk, and human-centered deployment, have become increasingly important as VLA systems move from laboratory demonstrations toward real applications. In the future, we argue that the next stage of VLA research will be defined by several converging directions. One is the shift from passive perception-conditioned action generation to active reasoning-driven embodied intelligence, in which agents can query, explore, verify, and adapt during execution. Another is the integration of richer world models that support prediction, imagination, and planning across time and modality. Lifelong and continual learning will also be essential for agents that are operating in evolving environments. In addition, future VLA systems are likely to expand beyond fixed single-robot manipulation toward integrated mobile manipulation, long-horizon household assistance, multi-agent collaboration, and broader open-world autonomy. Achieving these objectives will require not only stronger models, but also better data curation, more principled action representations, more realistic evaluation standards, and more reliable safety mechanisms. By synthesizing the development of VLA models from architecture to deployment, this survey aims to provide a structured reference for researchers and practitioners, and to support the advancement of embodied intelligence toward more general, robust, and trustworthy interactions in the physical world. Furthermore, a comprehensively curated list of open-source VLA algorithms, datasets, and simulation benchmarks discussed in this survey is updated at https://github.com/DefaultRui/vision-language-action-models-for-embodied-AI.

Key words: embodied intelligence; multimodal large model; vision-language-action (VLA) model; reinforcement learning; robotic perception and decision-making

0 引　言

具身智能(embodied intelligence)旨在通过智能体与物理世界的持续交互闭环,实现从高级指令理解到具体物理操作的转化。从技术演进与架构范式来看,当前具身智能的研究路径主要分为3类:第1类是“专用策略学习”范式,该路径通常采用轻量级卷积神经网络(convolutional neural network, CNN)等,主要依赖针对特定任务采集的领域数据,核心手段是通过强化学习或模仿学习建立从感知到动作的端到端映射(Firoozi等,2025),这类方法在特定封闭场景下表现优异,但受限于模型容量和数据规模,往往缺乏跨任务的泛化能力;第2类是“基于模型的预测与规划”范式,该路径侧重于显式或隐式地构建环境动力学模型(如世界模型),通过模拟环境状态的演变来进行长程预测,并结合模型预测控制或规划算法实现决策(Feng等,2025),这种方法具备较强的物理一致性和因果推理能力,但对环境建模的精度要求极高;第3类则是“基础模型驱动”范式,视觉—语言—动作模型(vision-language-action model, VLA)正是这一路径的典型代表,VLA通过引入互联网规模的预训练知识,在统一的表示空间中重构了感知与决策过程。它将视觉观测与自然语言指令作

为联合输入，直接输出可执行的动作序列，这一结构不仅继承了基础模型在开放词汇理解与复杂场景推理上的优势，而且补全了从感知理解到物理执行的闭环。值得注意的是，VLA与前两条路径并非割裂，而是呈现出融合趋势：在训练方法上，VLA广泛吸纳了模仿学习与强化学习作为预训练或微调手段；在架构演进上，前沿的VLA研究也开始内嵌世界模型以增强物理预测能力（Cen等，2025）。相比于专用策略学习，VLA利用通用知识缓解了数据稀缺与泛化难题；相比于传统的分层规划，VLA实现了语义理解与动作生成的端到端协同。

传统具身智能系统多以端到端策略学习或模型驱动控制为主要实现形态，常受限于环境泛化性弱、任务多样性有限及高质量数据匮乏等瓶颈，难以构建通用的具身智能体。近年来，基于互联网规模数据训练的基础模型成为推动人工智能发展的重要驱动力，也为具身智能研究注入新的活力。这类大规模神经网络通过对海量语料与视觉数据的统一建模，掌握了丰富的世界知识与强大的表征能力。以GPT-4o（generative pre-trained Transformer-4）（Hurst等，2024）和DeepSeek-R1为代表的大语言模型（large language model，LLM）在自然语言理解、推理等任务中表现出色，支撑了广泛的文本应用；在视觉领域，CLIP（contrastive language-image pre-training）、SAM（segment anything model）等基础模型在多种视觉任务中展现了强大的泛化性能（Radford等，2021）；而GPT、Gemini、Qwen2.5-VL等视觉—语言模型（vision-language model，VLM）实现了不同模态之间的深度融合。

然而，这些模型的感知与生成能力局限于数字空间，缺乏与物理世界直接交互的桥梁，难以直接应用于物理世界交互任务，因此学术界与工业界开始探索将基础模型的能力引入具身智能与机器人控制领域，逐步形成如图1所示的VLA概念。VLA将视觉观测与自然语言指令作为联合输入，直接输出可执行的动作序列，这一结构不仅继承了基础模型在开放词汇理解与复杂场景推理上的优势，而且补全了从感知理解到物理执行的闭环。

VLA的早期探索呈现出多样化的技术路线。一类是解耦式规划与控制，即利用LLM/VLM作为高级任务规划器生成符号化步骤或可执行代码，再由底层控制器解释执行，典型工作如SayCan（Ichter等，

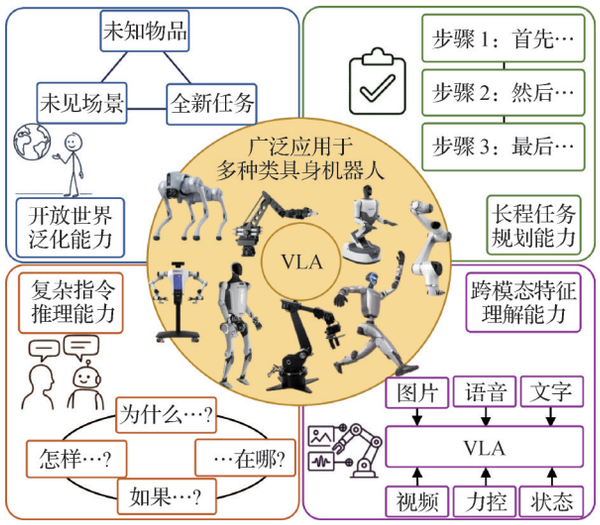

图1　VLA的概念及功能

Fig. 1　Concept and capabilities of VLA

2022）和PaLM-E（Driess等，2023），这实质上是VLA路径与分层符号规划路径的结合。另一类是知识蒸馏与引导，即从基础模型中提取可操作的先验知识，如物体的可供性（affordance）预测（Shridhar等，2021）或场景级轨迹规划，以引导下游基于强化学习或模仿学习的控制策略。同时，也有研究通过专门的预训练构建具身动作的隐变量空间，再微调VLM来预测并由策略网络解码执行。受深度学习领域规模法则（scaling law）的启发，研究者进一步构建了大规模具身演示数据集，并开始探索基于VLM的端到端（end-to-end）训练范式。这一路线在机器人操作、自主导航及自动驾驶等方向涌现出一系列创新成果，展现出强大的多任务学习、长程任务执行与零样本泛化能力（O'Neill等，2024；Bu等，2025a）。总体而言，VLA借助基础模型的通用知识与泛化能力，为破解具身智能领域长期存在的数据稀缺与跨平台迁移难题，提供具有前景的技术路径。与以强化学习和模仿学习为主的端到端控制路径相比，VLA在统一词元空间中显式对齐视觉、语言与动作，减弱了对精细奖励设计和任务专用网络结构的依赖；与以世界模型为核心的预测—规划—控制路径相比，VLA更强调从自然语言和多模态观测出发的一体化建模，同时也可以在解码端引入世界模型或规划模块，与之形成互补；相较于传统感知—建图—规划—控制的分阶段流程，VLA以可微分的大模型框架替代大量人工设计接口，将环境理解、任务分解与动作生成纳入统一的优化体系。近年来，一些前沿工作开

始尝试将VLA与上述路径结合，例如利用世界模型提供长时预测（Cen等，2025），借助强化学习实现安全约束与策略精调，体现出具身智能不同技术路线之间走向协同融合的整体趋势。

从技术发展脉络来看，视觉、语言与动作控制在早期多沿着相互独立的技术路线发展。视觉模型从CNN发展到Transformer架构，主要聚焦于识别与检测任务；语言模型侧重于文本的理解与生成；动作控制则多依赖手工策略或强化学习，通常局限于特定场景。尽管VLM显著缩小了视觉和语言之间的语义鸿沟，但感知—理解—执行的统一机制仍未建立。VLA的提出弥合了这一关键缺失，以RT-2（robotic Transformer 2）（Zitkovich等，2023）为代表的工作较早地确立了相关概念，通过在视觉编码与语言理解的基础上引入动作层面的建模，使智能体能够在动态环境中将高级复杂指令直接映射为相应的底层动作。与传统的级联式感知—控制架构相比，VLA在同一学习框架内直接对接机器人控制接口，在统一表征空间中联合优化任务目标、物理约束和执行细节。这既继承了LLM/VLM在开放世界语义知识理解方面的优势（杨易 等，2022；Wang等，2025e），也补全将理解转化为物理执行的关键一环，从而推动机器人从封闭场景的预编程操作，向开放环境下的泛化技能实现质的飞跃。为统一描述多样化的VLA架构，近期研究（Zhong等，2025a）将VLA系统视为若干可微子网络/功能模块串联的框架，其中视觉与语言信号进行迭代处理，逐步形成动作词元，最终转化为具体的动作输出。此处的动作词元既可指模型的显式输出，也可表示语义明确的中间表征，如动作隐变量或目标图像，而现有VLA体系的主要差异体现在动作词元的设计与组织方式上。

当前，针对具身智能领域VLA系统的中文综述相对较少（李浩然 等，2026），相关工作聚焦于VLM相关的基础模型类型划分与评测，或从更宏观的视角回顾机器人操作的技术发展（Shao等，2025；Zhong等，2025a）。如图2所示，自2022年以来VLA相关研究成果在主要学术会议和期刊中呈现快速增长态势，已形成规模可观的研究与应用版图。

本文旨在从概念、方法与系统3个层面提供统一的分析视角，与现有综述形成有效互补：1）系统回顾视觉与语言基础模型的演进历程，深入阐释规模法则与能力边界；2）全面地整理并构建VLA在状态编码与动作词元层面的分类体系，并深入比较各类代表性方法的技术特性；3）总结VLA的预训练与后训练策略，以及与强化学习结合的框架及主流评测基准，并剖析系统层面的架构权衡；最后，整理VLA在效率、泛化、安全与伦理等维度面临的关键挑战，并提出面向下一代系统的研究方向与可行路径。通过这一系统性的梳理与分析，本文期望为VLA领域的研究者提供清晰的技术发展图谱和前沿进展参考。

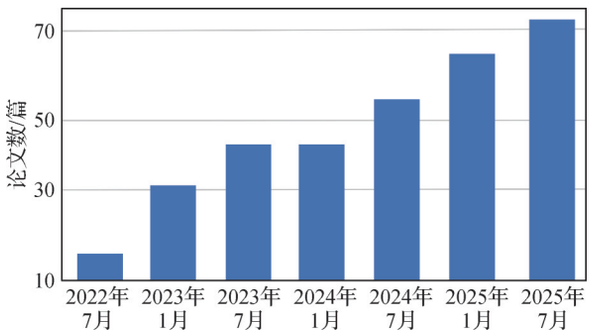

图2　VLA论文数量趋势

Fig. 2　Trend in the number of VLA papers

1　发展历程

1.1　早期探索

视觉编码方面，以ResNet（residual network）（He等，2016）、视觉Transformer（vision Transformer，ViT）（Dosovitskiy等，2021）为代表的模型实现了从像素级识别到语义级理解的跨越，SAM（Kirillov等，2023）则进一步提升分割精度。MAE（masked autoencoder）（He等，2022）、DINOv2（self-distillation with no labels v2）（Oquab等，2024）等自监督学习方法降低了对任务特异性标注的依赖（张燚钧 等，2025）。文本编码方面，从BERT（bidirectional encoder representations from Transformers）（Devlin等，2019）的语义理解到GPT（Ouyang等，2022）的生成与推理能力，为复杂指令解析与任务规划提供了关键引擎。VLM的兴起成为重要转折点，CLIP（Radford等，2021）通过大规模图文对比学习实现跨模态对齐，为语义—感知关联提供了统一空间；后续BLIP-2（bootstrapping language-image pre-training）（Li 等，2023）、LLaVA（large language and vision assistant）（Liu等，2023b）引入LLM增强语义推理，但尚未涉及动作生成，仍停留在感知—理解阶段。CLIPort（Shridhar

等,2021)首次将CLIP的视觉嵌入与运动原语结合,实现了语言描述—视觉语义—决策动作映射的闭环。此阶段的主要瓶颈在于视觉、语言与动作系统通过人工接口衔接,缺乏端到端优化,导致泛化能力弱、工程成本高,难以应对动态非结构化环境,从而产生对VLA一体化框架的需求。

1.2 中期发展

该阶段的核心是将视觉—语言表征与动作生成纳入统一框架,通过"动作词元化"实现跨模态融合(Cui等,2025)。如图3所示,VLA从概念走向技术实践,将"图像、文本、状态、动作"等统一为上下文序列,端到端生成动作,RT-1(Brohan等,2023)采用Transformer处理多模态输入,通过模仿学习将视觉图像、语言指令与动作轨迹编码为统一词元序列,在多个任务上达成较高成功率;SayCan(Ichter等,2022)将语言模型的任务知识与可供性评估相结合,实现更加可靠的动作执行;PaLM-E(Driess等,2023)则将连续视觉状态信号直接注入语言嵌入空间,是具身多模态语言模型的雏形。Google DeepMind的RT-2(Zitkovich等,2023)通过联合微调VLM与机器人轨迹数据,在未见物体与指令上验证了零样本泛化能力,其"视觉—语言—动作"联合词元设计成为后续架构的基准。与此同时,VIMA(Jiang等,2022)通过时序Transformer强化多步操作的空间推理,ACT(action chunking with Transformers)(Zhao等,2023)通过动作分块预测缓解高精度模仿学习中的误差累积,Diffusion Policy(Chi等,2025)将扩散建模用于动作生成,增强了噪声环境下的鲁棒性。该阶段完成VLA的基础功能闭环,但仍存在动作词元类型单一、泛化能力有限以及实时性不足等问题。

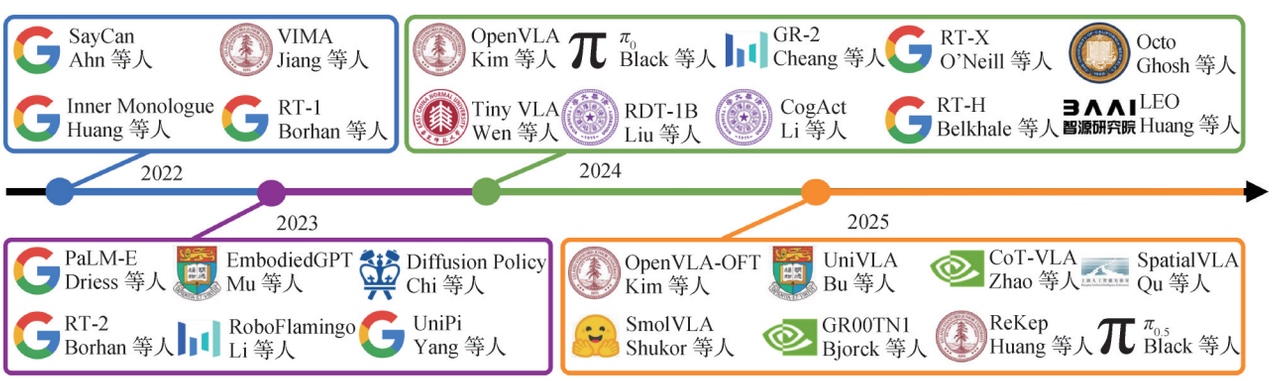

图3 VLA系统发展历程

Fig. 3 Timeline of VLA systems

1.3 落地应用

自2024年起,研究重心由通用框架构建转向领域适配与推理增强,通过引入领域归纳偏置与高级推理机制提升真实场景可用性。动作词元从早期的执行轨迹扩展至代码标记、可供性掩码和目标状态等多种类型:ProgPrompt(Singh等,2023)以代码方式调用机器人接口提升控制精度;ReKep(Huang等,2024b)通过可供性词元预测可操作区域,显著提高抓取成功率。LLM的角色由指令解析器演化为分层规划器:Huang等人(2022)引入行动前规划和行动后反思闭环,Hi Robot(Shi等,2025)采用高层任务分解和底层动作执行的分层结构,使复杂长周期任务的成功率明显提高;CoT-VLA(Zhao等,2025a)则将思维链机制融入动作生成。在领域适配方面,面向导航、操作和自动驾驶分别形成了专用模型:Uni-NaVid(Zhang等,2024b)以三维场景图优化路径规划;MoManipVLA(Wu等,2025c)面向嵌入式部署优化效率;Occllama(Wei等,2024)引入物理感知注意力以应对自动驾驶场景中的遮挡问题。2025年以来多项研究聚焦泛化增强与安全部署,RT-X(robotics Transformer X)(O'Neill等,2024)所整合的Open X-Embodiment数据集汇聚了21个机构的527种技能,为跨场景泛化提供数据支撑;SpatialVLA(Qu等,2025)以4D空间感知促进仿真到现实迁移;GR-1(Wu等,2024b)和GR-2(Cheang等,2024)将视频生成预训练迁移为机器人世界模型,提升环境动态预测能力;OpenVLA(Kim等,2024)构建跨平台数据使模型可适配多种具身本体;SafeVLA(Zhang等,2025a)引入形式化验证实现风险感知决策;GR00T-N1(Bjorck等,2025)融合双系统架构以提升可追溯

性与可解释性。当前趋势主要体现在多种动作词元的联合优化、数据高效学习和通用智能体构建。

2 VLA 系统架构

2.1 视觉特征编码

视觉特征编码是VLA的基础组件，其核心任务是从原始图像或视频观测中提取稳定、可迁移且具备语义可解释性的表征，为后续的跨模态对齐与动作生成提供支撑。当前主流技术路径主要包括预训练视觉模型的迁移适配、视觉信息词元化以及三维时空表征融合。

1）预训练视觉模型的迁移与适配。早期研究多采用基于卷积的编码器（He 等，2016；Tan 和 Le，2019）从单帧图像中提取空间特征，并通过多层感知机（multilayer perceptron，MLP）或 Transformer 将其映射至语义空间。随着对比学习与自监督学习的发展，主流方法逐步转向大规模预训练模型。CLIP（Radford 等，2021）通过图文对比学习构建了视觉与文本的共享嵌入空间，已成为 VLA 视觉编码的重要基础；DINOv2（Oquab 等，2024）则验证了无监督视觉表征在机器人任务中的强泛化能力。OpenCLIP（Cherti 等，2023）、EVA-CLIP（Sun 等，2023）等衍生模型也广泛用于跨场景感知，进一步增强了域外适应性。在视频层面，R3M（reusable representations for robotic manipulation）（Nair 等，2022）和 VIP（value-implicit pre-training）（Ma 等，2023）等策略引入时间对比学习与帧间一致性约束，显式建模动态场景中的语义关联与时序关系，从而有效提升了环境理解与因果推理能力。

2）视觉信息词元化。为实现更精确的跨模态对齐与指令理解，近年研究将视觉特征离散化或压缩为词元形式，以便与文本知识协同建模（Wang 等，2025f；Feng 等，2025）。VQ-VAE（vector quantised variational autoencoder）（van den Oord 等，2017）和 VQ-GAN（vector quantised generative adversarial network）（Esser 等，2021）通过向量量化将图像映射为离散词元序列，使视觉输入能够自然地与语言模型接口耦合。Flamingo（Alayrac 等，2022）通过跨注意力机制将视觉信息压缩为固定长度隐变量，有效控制词元规模；BLIP-2（Li 等，2023）的 Q-Former 与 TokenLearner（Ryoo 等，2021）则通过自注意力与空间聚合筛选任务相关特征，在保持语义完整性的同时显著降低了计算开销。与此同时，借助目标检测、分割与跟踪技术，如 Mask R-CNN（He 等，2017）、SAM（Kirillov 等，2023）和 Grounding DINO（Liu 等，2025b）等，显式提取实体区域特征，在视觉层面增强对可操控物体的空间定位与交互一致性，特别适用于抓取、放置与装配等精细操作场景。

3）三维时空表征融合。随着机器人视觉从二维感知迈向空间理解，三维表征成为 VLA 发展的重要方向（冯明涛 等，2025）。由于二维特征难以完整表达几何与深度信息，研究者引入多种三维感知形态以提升空间推理能力。基于深度图的方法（Yang 等，2024b）利用单目深度估计从 RGB 图像预测深度信息，结合相机参数恢复三维坐标，并通过位置编码增强几何一致性与显式空间感知；SpatialVLA（Qu 等，2025）融合 SigLIP（Zhai 等，2023）与深度编码以提升复杂场景中的空间定位与交互精度。3D-VLA（Zhen 等，2024）扩展 Q-Former 以处理多视角输入，Evo-0（Lin 等，2025c）与 RoboMM（Yan 等，2024）提取多视图几何特征并生成占据网格，为语义理解与动作规划提供统一的体素表征。OpenDriveVLA（Zhou 等，2026）则将体素网格投影为鸟瞰图特征以兼容卷积结构，SOFAR（semantic orientation for autonomous robots）（Qi 等，2025b）和 PointVLA（Li 等，2026）等直接利用 PointNet 系列（Qi 等，2017a）处理点云数据，实现稀疏空间编码与语义定位的结合。部分最新工作进一步将点云与神经辐射场（neural radiance field，NeRF）（Mildenhall 等，2022）或三维高斯表征（Kerbl 等，2023）结合，构建可渲染的连续三维场景，在统一空间内同时捕获几何、语义与可操作信息。这些方法显著提升了模型在复杂物理环境中的泛化能力与鲁棒性，为 VLA 的空间推理与动作理解提供坚实的支撑。

2.2 语言特征编码

语言特征编码不仅涉及任务解析与语义抽取，也是连接视觉与动作模块、确保逻辑一致性的关键环节。随着 LLM/VLM 的快速发展，语言编码已从静态语义表征演进为动态的跨模态条件建模，其目标是在统一表征空间内实现语言、视觉与动作的高效协同。VLA 通常依赖预训练语言模型构建词表与向量嵌入，主流方法沿用 T5（Raffel 等，2020）、LLaMA（Touvron 等，2023）、Qwen（Bai 等，2023）和 Phi（Abdin

等，2024）等模型自带的分词器与词表，以保持与预训练语义空间的一致性，增强对复杂自然指令的泛化与迁移能力。面向实时执行或资源受限平台，可采用DistilBERT（Sanh等，2019）、Sentence-BERT（Reimers和Gurevych，2019）等轻量编码器，以平衡表达力与计算开销。由此获得的语言嵌入可有效捕获任务意图、对象关系与语义逻辑，为后续多模态融合与动作生成提供先验。

近期的VLA普遍基于注意力机制实现跨模态语义细粒度对齐。FiLM通过通道级线性调制，使语言嵌入动态调节视觉特征的激活，跨注意力机制则令语言特征充当查询，直接加权聚合视觉词元，实现显式语义关联。Perceiver Resampler（Alayrac等，2022）与Q-Former进一步优化了交互效率，前者以固定长度潜变量压缩冗余信息，后者使用可学习查询词元提取任务相关特征，显著降低计算成本。这些机制确保语言在视觉处理阶段发挥主动调控作用，从而在复杂语义指令下实现更强的目标理解、区域聚焦与动作引导。BLIP-2（Li等，2023）借助Q-Former或跨模态注意力实现视觉与语言词元的深度交互，使语言特征直接参与视觉理解与推理过程；PaliGemma（Beyer等，2024）和Chameleon（Chameleon Team，2025）代表了图文统一建模方向，为后续VLA借用VLM骨干提供基础。HybridVLA（Liu等，2025a）和CogACT（Li等，2024a）引入扩散式动作生成，使语言从指令载体转变为策略推理。

2.3 其他模态特征编码

1）点云特征编码。三维点云直接表征环境的几何结构（龚靖渝 等，2023），是具身感知与抓取规划中的重要信息源。常见的处理方法采用预训练的点云Transformer进行分词或嵌入表征，其中PointNet系列（Qi等，2017a）与Uni3D ViT（Zhou等，2024）等骨干网络已广泛应用于SOFAR（Qi等，2025b）、LEO（Huang等，2024a）、LMM-3DP（Li等，2025f）、General-Flow（Yuan等，2024a）与DexTOG（Zhang等，2025i）等模型。此外，部分研究将重建式三维表征与语义特征相融合，如LERF-TOGO（Rashid等，2023）与Splat-MOVER（Shorinwa等，2024）分别将NeRF或Gaussian Splatting重建的三维显式场与CLIP语义特征对齐，再结合GraspNet（Asif等，2018）等抓取候选生成器实现语义引导的三维抓取规划。总体而言，点云编码在通用预训练骨干与任务专用模块之间存在重要权衡，前者有利于模型迁移与复用，后者则有望在特定任务上实现更高的几何与语义一致性。

2）音频特征编码。部分研究将环境声学线索与语音信息引入决策输入（Lu等，2024；Jiang等，2025a），主流处理方法是将频谱图作为输入，再借助图像或序列编码器将其转换为音频词元。常用的编码骨干包括VQ-GAN量化模块（Esser等，2021），以及RVQ-VAE体系的SpeechTokenizer（Zhang等，2024d）等预训练模型，此类编码器能够保留难以完整转写为文本的声学细节，为具身决策提供辅助信息。在解码端，通常采用SoundStorm（Monache等，2024）或VQ-GAN解码器以重建或对齐音频表征。在工程实现上，RoboNurseVLA（Li等，2025c）先通过自动语音识别将音频转写为文本，再与语言或视觉特征融合。

3）触觉与力觉模态编码。在操作任务中，触觉与力觉信息为精细操作提供关键物理约束。FuSe（Jones等，2025）、TLA（tactile-language-action model）（Hao等，2025）、VTLA（vision-tactile-language-action model）（Zhang等，2025b）、Tactile-VLA（Huang等，2025b）等模型均将触觉信号作为输入模态之一。在硬件层面，常用的DIGIT与GelStereo 2.0等视觉式触觉传感器可输出图像化的触觉观测；在编码层面，可直接使用ViT将触觉图像嵌入为词元，或借助预训练的触觉—视觉—语言模型实现跨模态对齐。此类表征便于与视觉特征进行联合学习，有助于在估计接触状态与轨迹约束等方面进一步提升任务执行的稳定性与精度。

2.4 上下文状态编码

环境场景（单帧或视频）通常通过ViT、ConvNeXt等视觉编码器压缩为紧凑的视觉嵌入；自然语言指令则通过T5、LLaMA等大语言模型转化为语义嵌入（Liu等，2022）。二者与机器人本体及交互状态，包括关节位置/速度/力矩、触觉与受力线索、里程计、LiDAR（light detection and ranging）扫描、惯性测量等，共同构成连续的上下文状态词元序列，对解码器起到关键的对齐作用。统一后的序列输入使策略在动作生成过程中能够实时调整控制指令，形成对外部环境与内部构型的协同感知与响应。此类上下文状态编码特别强调跨时间的一致性与多源对齐。一方面通过时序位置编码或跨帧注意力保留动态演变信息；另一方面借助多模态对齐与归一化词元预算

控制计算开销，从而实现语义理解与动力学约束在同一序列内由弱耦合向强耦合的渐进融合。基于这一架构，VLA 能够在单一自回归解码器中统一处理环境语境、任务意图与执行约束，输出与情境高度相关的精细动作序列，显著提升在接触丰富操作与非结构化地形导航中的鲁棒性与泛化能力。

2.5 动作词元与解码

面向具身任务，VLA 的核心目标是将多模态输入映射为可执行的动作输出。为实现从感知到控制的完整闭环，系统需依次完成场景理解、语义落地、任务规划、运动预测与动作生成等环节。考虑到具身任务的复杂性与开放性，这些能力需要在执行过程中实现动态切换、复用与递归调用。基于 Transformer等自回归生成架构，视觉、语言、状态与动作等多模态信号被离散化并嵌入统一表征空间，使模型能够在感知（Black 等，2025）、语言理解与动作控制（Yu 等，2023）等多个空间进行整体推理，并以端到端、可组合的方式逐步细化至可直接下发的控制量（Bruce 等，2024）。Hi Robot（Shi 等，2025）通过微调的多模态模型生成自然语言形式的子任务计划，随后由与 pi-0 策略相近的模块输出底层控制指令；VoxPoser（Huang 等，2023c）则将语言指令分解为子任务，进一步生成可供性表征并调用运动规划模块得到原始动作。在实践中，VLA 系统将上述能力分解到多个功能模块中，并将这些模块输出的中间或最终表征统一称为动作词元，可视为动作域的广义表征。

根据现有研究（Zhong 等，2025a），动作词元可分为以下几种类型，且可在同一系统中组合使用：语言描述（以自然语言描述目标动作序列，涵盖从高层规划到细粒度指令）、代码表征（以可执行代码或伪代码表达完整程序或原子操作）、可供性表征（通过关键点、边界框、分割掩码或热图等空间化表征）、运动轨迹（使用时序状态序列描述物体或执行器的动态演化）、目标状态（以未来图像、点云等形式呈现期望任务状态）和推理过程（以自然语言显式表达决策过程）。在统一框架下，不同类型的动作词元在表达能力、数据可获取性、推理时延、可解释性与跨载体迁移性等方面存在显著权衡，这些特性直接影响解码器与控制模块的选择与组合。

在解码与动作建模层面，现有方法主要呈现以下几类技术路线与上述词元类型协同配合，表 1 展示了几种主流的动作词元解码建模方式的对比。

表 1　不同动作词元解码建模方式对比

Table 1　Comparison of different action token decoding methods

建模方式	优势	劣势	适用场景
离散化词元	泛化性强，零样本泛化能力突出，推理较快	存在精度误差，长序列影响输出频率	指令跟随、移动导航
连续回归	输出轨迹平滑，精度较高，推理高效	难以建模复杂动作分布，泛化能力弱	精确轨迹跟踪、动态的连续控制任务
扩散模型	能建模复杂动作分布，策略鲁棒，训练稳定	推理慢，计算开销大	复杂抓取操控、精细装配
流匹配	推理效率高，平衡表达能力与效率	训练数据质量要求高	对实时性有要求的操作任务

1）基于分箱的离散动作词元。将动作空间各维度离散化为若干区间，以区间索引作为离散词元，采用交叉熵损失进行训练，并基于自回归方式进行解码，其形式与 LLM 生成文本类似。代表性工作有：RT-2（Zitkovich 等，2023）结合 PaLI-X 直接输出数值词元；PaLM-E 与 OpenVLA 在词表中预留 256 个词元作为动作槽位（Kim 等，2024）。此外也存在非自回归变体，例如并行生成所有动作词元（Zhao 等，2025c），或将序列末尾若干词元固定映射为手臂或底盘动作（Brohan 等，2023）。标准分箱方法的主要问题在于词元序列较长，可能影响控制频率。FAST（frequency-space action sequence tokenization）（Pertsch 等，2025）通过沿时间轴进行离散余弦变换，并对频域分量量化压缩以显著缩短序列，从而提升推理效率；SpatialVLA（Qu 等，2025）对动作空间进行统计离散化，为高频运动分配更高分辨率，从而减少空间类动作词元的数量并提升系统效率。

2）词元到连续动作的回归解码。以 Transformer 输出的词元序列作为输入，通过多层感知机或循环神经网络将其映射为连续控制量，通常采用 L2 或 L1

损失进行优化，使用均值或最大池化将多词元压缩为单步动作表征（Li等，2024a）以及进一步预测多步动作块以生成更平滑的轨迹（Kim等，2025b）。

3）基于扩散或流匹配的连续动作生成。扩散模型与流匹配正逐渐成为VLA中生成连续动作的有效手段，其中Octo（Ghosh等，2024）与pi-0（Black等，2024）展示了非自回归、平滑且可扩展的控制能力。相较于传统扩散方法，流匹配因所需推理步数更少，更有利于实时控制。pi-0采用流匹配机制通过将动作生成建模为连续归一化流，在保持动作分布多样性的同时，实现了比传统扩散模型更高效的推理。相比于自回归模型的离散化量化误差和扩散模型昂贵的推理步数，流匹配能够在线性轨迹上更高效地建模连续动作空间，训练效率显著提升。在实现上，既可在Transformer之后外接扩散动作模块，也可将其内嵌至Diffusion Transformer架构中（Song等，2021），并通过在各时间步保持去噪一致性，以提升生成稳定性与效率。

从当前技术发展态势来看，3种主流解码机制呈现出明显的分化与融合趋势：自回归方法在零样本泛化和复杂语义理解方面保持优势，但其序列生成的固有特性导致推理延迟较高；扩散策略在建模复杂动作分布和处理多模态任务方面表现优异，输出平滑性和鲁棒性强，但多步迭代的采样过程带来较大计算开销；流匹配方法作为扩散模型的演进，通过更少的推理步数在表达能力与效率之间取得了更好的平衡，成为实时控制场景的常见方案。

表2提供了基于扩散策略（Chi等，2025）和流匹配策略（Zhang等，2024）的VLA在Push-T（Chi等，2025）和RoboMimic（Mandlekar等，2021）数据集上的部分性能对比，表中的数值结果来源于论文*Streaming Flow Policy*（Jiang等，2025b）中对扩散策略与流匹配策略进行对比实验所报告的结果，该对比实验在相同的任务设置、数据集划分与评估协议下进行，确保了比较结果的内部一致性。本文仅对其进行归纳呈现，并保持原文的评价指标与实验设置不变。

表2　基于扩散策略/流匹配的VLA性能对比（Jiang等，2025b）

Table 2　Comparison of performance of diffusion-policy-based and flow-matching-based VLA (Jiang et al., 2025b)

方法	Push-T		RoboMimic Square	
	动作模仿平均分数/%	延迟/ms	动作模仿平均分数/%	延迟/ms
扩散策略（Chi等，2025）	**87.0**	127.2	**77.2**	53.4
流匹配策略（Zhang和Gienger等，2024）	71.0	**12.9**	54.0	**4.8**

注：加粗字体表示对应模型性能在该指标最优。

4）从大规模视频学习隐式动作表征。当缺乏显式动作标注时，可采用世界模型式的隐变量学习框架，从视频数据中提取动作相关的隐式表征，从而显著扩大训练数据规模并提升模型泛化能力。代表性工作包括LAPA（Ye等，2025）、Moto（Chen等，2024）、UniVLA（Bu等，2025b）与UniSkill（Kim等，2025a）等，可作为跨任务、跨本体学习的中间表征层。

5）跨本体动作表征。跨机器人形态的动作与本体感知差异是训练通用型VLA的关键挑战。Open-X Embodiment（O'Neill等，2024）首次系统性提出跨本体标准化方案，在单相机输入、语言指令与多自由度动作的统一格式下，聚合不同机器人的数据，其性能显著优于单载体训练。另一类研究将异构本体的观测与动作标准化至第一人称视角并对齐，从而仅基于观测与目标图像实现多机种统一控制（Yang等，2024a）；然而，当观测模态或控制接口差异过大时，统一表征仍面临挑战。为此，CrossFormer（Doshi等，2024）首先利用模态专用分词器对视觉、本体和任务规约等异构传感信号进行词元化，拼接为统一序列，再由共享的解码器通过读出词元提取任务相关表征，最后接入本体专用动作模块生成动作。Uni-Act（Zheng等，2025a）引入跨本体共享的离散动作词元，由载体专用解码器还原为连续控制量，显式促进知识迁移与复用。此外，LangToMo与ATM等框架利用中间表征实现跨本体学习（Wen等，2024a；Ranasinghe等，2025），规避直接对齐动作空间的复杂性。

现有的词元化方法在本质上是信息压缩与精度损失之间的权衡，其在信息论层面的表达能力差异及对任务性能的影响如下：从信号处理角度看，动作离散化引入了量化噪声，任务成功率并非随量化精

度线性提升，而是存在瓶颈效应。当离散化位深低于某一量级时，量化偏差会沿时序累积，导致闭环控制中的动力学不稳定性；然而过高的离散化精度会导致词元表征空间稀疏，增加模型在训练时的交叉熵损失收敛难度。因此，最优动作词元化是在保持最小有效控制分辨率与模型学习效率之间的折中。

从信息论视角审视，这一过程可建模为率失真理论框架下的有损压缩问题（Davisson，1972）。这一理论框架揭示了动作词元化的核心权衡：词表大小（码率）与动作重构精度（失真）之间的制约关系。增大词表尺寸可降低量化误差，但会带来两个问题：词元序列变长，增加自回归解码的计算复杂度；词元在词汇表中的分布更稀疏，导致模型学习难度增大。反之，过小的词表虽然能提升效率，但量化误差会引入不可忽略的控制偏差，在精细操作任务中可能导致任务失败。这一理论关系为理解不同词元化方法的性能差异提供了统一视角。例如 RT-2 等模型采用的方案，其隐式假设动作服从均匀分布，在信息论层面，其表达能力受限于固定比特率，导致其在对“高频小幅度”等精细操作进行建模时存在信息瓶颈。

2.6 控制策略

2.6.1 任务规划模块

在 VLA 控制策略中，任务规划模块将抽象语义目标转化为可执行步骤的关键职能。该模块不仅需要将复杂任务拆解为子任务序列，还需在长时交互过程中根据环境反馈进行动态调整与规划，从而提升系统的适应性与任务完成率（Ma 等，2024）。

当前任务规划的主流思路依托 LLM 的端到端推理能力，实现从感知输入到高层规划的直接映射。例如，PaLM-E（Driess 等，2023）通过融合 ViT 与 PaLM 的架构，将语言指令转化为可执行技能序列，并具备基于环境变化的在线重规划能力；EmbodiedGPT（Mu 等，2023）则通过专门的模块整合视觉编码与语言规划信息，生成驱动具体动作的实例级特征。为增强空间推理的准确性，多项研究致力于将三维感知与规划能力深度融合。LEO（Huang 等，2024a）采用两阶段训练策略，先进行三维视觉—语言对齐预训练，再进行视觉、语言和动作的端到端微调，有效提升点云编码与语言理解的协同效能。3D-LLM（Hong 等，2023）通过注入点云、gradSLAM（Jatavallabhula 等，2019）和体素场等三维特征，显著增强对话与导航任务中的空间理解能力。ShapeLLM（Qi 等，2025a）则基于改进的三维编码器，融合多视图图像、文本监督和点云表征，在三维基准测试中展现出优越的具身交互性能。SayCan（Ichter 等，2022）将语言模型的语义理解能力与底层策略的可行性评估相结合，通过概率乘积的方式筛选最优技能，确保规划结果在具体环境中的可实施性。

在系统架构设计方面，部分研究采用模块化思路进行构建。Inner Monologue（Huang 等，2022）通过自然语言连接高层规划与底层执行，基于成功/失败状态、场景变化和人工指令等反馈动态调整后续步骤。LLM-Planner（Song 等，2023）通过显式的分层策略设计，由高层生成语言描述的子目标，底层负责将其转化为具体的原子动作序列；LID（Li 等，2022b）则通过对失败轨迹的分析与重标注，主动收集训练数据，并将环境信息统一转化为文本描述，显著提升系统的组合泛化能力；ProgPrompt（Singh 等，2023）利用程序化提示定义动作与对象规范，通过少量示例生成结构化的伪代码计划，并在代码中嵌入断言语句实现运行时的自动校验与修正；CaP（code as policies）（Liang 等，2023）充分发挥大模型的代码生成能力，直接输出可调用感知与控制端口的策略代码。COME-robot（Zhi 等，2025）进一步借助多模态大模型的感知推理能力，直接从环境输入生成执行代码，降低对专用感知模块的依赖，为系统提供了可控可审计的安全保障。

2.6.2 底层控制策略

底层控制策略负责将高层规划转化为具体的执行动作，其设计直接影响系统的响应速度、控制精度和任务成功率。当前主流的控制策略主要包括基于扩散模型的方法和基于自回归模型的方法，两者在控制精度、推理效率和任务适应性方面各具特色。

基于扩散模型的策略将动作生成建模为条件去噪过程，通过逐步去除噪声恢复出符合任务要求的动作分布。这种方法能够有效建模不同的动作分布，避免行为克隆中常见的均值化问题，特别适合需要高精度控制的长时序任务。Diffusion Policy（Chi 等，2025）通过引入视觉条件的时序扩散模型，结合滑动窗口的闭环执行机制，在多个操纵任务中显著超越了传统行为克隆的性能；DP3（3D diffusion policy）（Ze 等，2024）进一步引入稀疏点云的三维表征，增强对视角变化和外观变化的鲁棒性，在少量真实演示数据下就能取得良好的控制效果；MDT（mul-

timodal diffusion Transformer)(Reuss 等,2024)通过多模态扩散架构,实现了图像与语言目标在隐空间的对齐,为弱监督下的长程任务提供有效解决方案;RDT-1B(Liu 等,2025c)作为大规模扩散基础模型,通过统一的动作空间设计和可扩展的Transformer架构,实现了多机器人平台和多模态输入的兼容处理。

自回归策略则将控制问题转化为序列生成任务,通过链式条件概率建模实现动作的逐步预测。这种方法具有推理速度快、支持闭环控制的特点,便于与预训练的多模态大模型集成。RT-1(Brohan 等,2023)通过动作离散化和序列分类的方式,验证大规模演示数据与模型容量对泛化能力的促进作用;RT-2(Zitkovich 等,2023)进一步将动作词元纳入语言词表,实现真正的视觉—语言—动作统一建模,展现从网络知识到机器人任务的迁移能力。OpenVLA(Kim 等,2024)通过融合双视觉编码器与语言模型骨干,在保持参数效率的同时实现多机械臂的统一控制,并通过量化技术满足了实际部署的需求。UniVLA(Bu 等,2025b)则通过全面的词元化设计和视频世界模型预训练,进一步增强长时规划能力,在多个测试基准上取得领先性能。

从系统特性角度看,这两种策略呈现出明显的互补优势。扩散策略在动作分布建模的完整性和输出平滑性方面表现优异,能够处理复杂的多模态控制任务,但其多步迭代的采样过程带来了较高的计算开销;自回归策略则以其高效的单步推理和便捷的多模态集成能力见长,特别适合对实时性要求高的应用场景,但在处理复杂动作分布时可能趋向于保守策略。在实际应用中,两种策略的选择需要综合考虑任务复杂度、实时性要求和计算资源约束。对于需要精细控制和高可靠性的任务,如精密装配或柔顺操作,扩散策略通常能提供更优的性能;而在需要快速响应和强语义理解的任务中,如交互式导航或开放式指令跟随,自回归策略往往更具优势。

3 VLA系统架构分类

为了统一描述多样化的VLA架构,本文根据模型内部的推理—控制耦合程度及运行频率机制,将现有工作归纳为单系统、双系统与层次化3类(Zhang等,2025c;Xu等,2024b)。其中,单系统与双系统的核心区别在于是否在架构设计上显式解耦了高级语义推理与底层高频控制的执行频率,层次化系统结构的特点是存在显式符号化接口,并且中间表征为人类可读的符号(如代码、路径点、语言指令),强调逻辑的可解释性与模块的独立替换性。

3.1 单系统VLA

单系统VLA采用单体式架构,将感知—理解—控制整合于同一计算图中,形成统一的端到端控制闭环。其核心特征在于时序上的同步性与计算的整体性:无论模型内部采用何种解码策略(如自回归头或扩散头),视觉观测、语言指令与动作生成均在同一控制频率下顺序执行。在此架构下,模型通过多模态骨干网络将输入映射为统一的序列表征,并在同一解码器中直接生成动作词元,从而实现从互联网规模的知识到具身操作任务的迁移。单系统模型VLA方法及其特点如表3所示。例如RT-1将Transformer引入VLA,并使用离散词元对连续动作进行编码(Brohan 等,2023);RT-2在更大规模VLM上联合微调互联网视觉—语言数据与机器人轨迹数据,实现了从网络知识到操作技能的迁移(Zitkovich 等,2023);RT-2-X 进一步基于Open X-Embodiment数据集进行协同微调,提升跨机器人技能转移能力(O'Neill 等,2024);而OpenVLA采用SigLIP与DINOv2替代大参数视觉编码器,并在大规模真实操作数据上微调,以更少参数实现更高性能,成为开源领域的重要基线(Zhai 等,2023;Oquab 等,2024)。尽管RDT-1B(Liu 等,2025c)、CogACT(Li 等,2024a)等工作引入了扩散模块来处理复杂的动作分布,但其语义编码与动作生成仍作为紧密耦合的整体进行联合推理,未在系统层面通过异步机制分离"慢思考"与"快行动"过程,因此仍归类为单系统。

在统一序列建模的范式下,模型通过对环境、指令和状态的联合表征,实现从任务目标到执行细节的逐步细化。为增强场景表征能力,研究在输入侧持续进行扩展。在三维结构建模方面,LEO Agent直接从仿真环境中提取物体中心的三维点云,经PointNet++编码后通过空间变换器构建空间信息(Qi 等,2017b);SpatialVLA(Qu 等,2025)将单目深度回投至机器人坐标系,并与像素级特征融合,构建基于本体坐标的空间表征。

在时空动态捕捉方面,TraceVLA在当前图像上叠加采样的运动点引入轨迹线索(Zheng 等,2025b);4D-VLA将三维坐标注入视觉特征,并利用

表 3　单系统模型 VLA 方法及其特点

Table 3　Single-system VLA methods and their characteristics

方法	语言模型	视觉编码器	学习方式	特点
RT-2(Zitkovich 等,2023)	PaLM-E	–	模仿学习(连续动作)+微调(离散动作)	将动作表征为 VLM 词元以实现泛化
RT-2-X(O'Neill 等,2024)		ViT-22B	模仿学习(连续动作)+微调(离散动作)	跨机器人数据微调以实现模型迁移
OpenVLA(Kim 等,2024)	LLaMA2-7B	DINOv2 + SigLIP	模仿学习(连续动作)	开源 7B 参数 VLA 实现通用机器人控制
LEO(Huang 等,2024a)	Vicuna-7B	ConvNext	模仿学习(连续动作)+微调(离散动作)	结合三维特征与 LLM 用于动作决策
ReVLA(Dey 等,2025)	LLaMA2-7B	DINOv2 + SigLIP	模仿学习(连续动作)	反向回溯到策略执行提升可解释性
TraceVLA(Zheng 等,2025b)	LLaMA2-7B	DINOv2 + SigLIP	模仿学习(连续动作)	利用视觉轨迹提示实现跨模态时间感知
FuSe(Jones 等,2025)	PaLM-Gemma-3B	–	模仿学习(连续动作)+微调(离散动作)	使用自然语言用于跨模态微调
UniAct(Zheng 等,2025a)	LLaVA-0.5B	SigLIP ViT-L	策略蒸馏(连续动作)	提出通用动作空间实现自适应控制
SpatialVLA(Qu 等,2025)	Gemini	–	模仿学习(连续动作)	通过三维语义和动作网格改进泛化
RoboFlamingo(Li 等,2024b)	MPT-1B	ViT	模仿学习(连续动作)+微调(离散动作)	适配开源 VLM 用于机器人控制
RoboMamba(Liu 等,2024a)	Mamba-2.8B	CLIP/SigLIP ViT-L	模仿学习(连续动作)+微调(离散动作)	将 Mamba 架构引入 VLA 领域
OpenVLA-OFT(Kim 等,2025b)	LLaMA2-7B	DINOv2 + SigLIP	策略蒸馏(连续动作)	通过 OpenVLA 蒸馏提升性能
Spec-VLA(Wang 等,2025c)	LLaMA2-7B	DINOv2 + SigLIP	策略蒸馏(连续动作)	在解码阶段进行加速,实现无损推理
FlashVLA(Tan 等,2025b)	LLaMA2-7B	DINOv2 + SigLIP	模仿学习(连续动作)	训练加速并提升 VLA 推理效率
CogVLA(Li 等,2025g)	LLaMA2-7B	DINOv2 + SigLIP	策略蒸馏(连续动作)	提出认知对齐框架以实现高效操作
UnifiedVLA(Wang 等,2025f)	Emu	–	模仿学习(连续动作)+微调(离散动作)	将输入信号转换为词元构建统一模型
PD-VLA(Song 等,2025a)	Vicuna-1.5B	CLIP ViT-L	策略蒸馏(连续动作)	增强基于蒸馏的快速机器人控制并行
4D-VLA(Zhang 等,2025f)	InternVL-1.4B	–	策略蒸馏(连续动作)	整合时空线索实现高效 VLA 预训练
OE-VLA(Zhao 等,2025c)	Qwen1.5-5B	–	模仿学习(连续动作)+微调(离散动作)	为开放末端任务引入跨模态交互
ST-VLA(Patratskiy 等,2025)	PalGemma2	–	模仿学习(连续动作)+微调(离散动作)	将视觉轨迹投影到深度图增强理解
VLAS(Zhao 等,2025b)	Vicuna-7B	CLIP ViT-L	模仿学习(连续动作)+微调(离散动作)	引入双模态自编码器并构建配对数据集
DeeR-VLA(Yue 等,2024)	MPT-1B / 7B	CLIP ViT-L	策略蒸馏(连续动作)	提出动态早退出以减少推理开销
BitVLA(Wang 等,2025b)	BitNet1.5 8B	SigLIP ViT-L	策略蒸馏(连续动作)+微调(离散动作)	设计权重模型以减少部署内存成本
MoLe-VLA(Zhang 等,2026)	LLaMA2-7B	DINOv2 + SigLIP	策略蒸馏(连续动作)	通过动态模型层激活减少计算量

注:"–"表示未显示采用独立的视觉编码器,或未说明该特定结构。

关键帧记忆缓解时间歧义(Zhang等,2025f);ST-VLA融合历史轨迹与深度信息,以增强时空推理能力(Patratskiy等,2025)。在多模态触觉与听觉集成方面,VTLA(vision-tactile-language-action model)使用视觉编码器对触觉图像进行建模,将其与视觉、文本词元共同输入LLM(Zhang等,2025b);VLAS(vision-language-action model with speech instructions)利用Whisper编码语音信号,并投影至语言嵌入空间(Zhao等,2025b);FuSe在微调阶段对齐触觉/音频—语言概念,减少对模态专属标注的依赖(Jones等,2025);OE-VLA结合多模态数据与两阶段课程学习,将指令类型扩展至图像和视频(Zhao等,2025c)。为提升决策的可解释性与鲁棒性,ECoT生成链式推理,将高层任务规划与视觉落地逐步结合(Zawalski等,2024);CoT-VLA通过预测像素空间的子目标观测,实现视觉链式推理(Zhao等,2025a);LoHoVLA采用分层闭环机制处理规划误差与外部扰动,以支持长程任务(Yang等,2025b);ReFineVLA则采用选择性迁移微调,仅更新模型上层参数以增强多模态理解能力(Van等,2025)。

在跨任务与跨平台泛化方面,UniAct利用通用动作码本,以统一异构机器人的动作空间,促进跨本体学习与推理(Zheng等,2025a);ReVLA通过可逆训练逐步恢复视觉编码器的预训练状态,缓解灾难性遗忘并改善分布外视觉泛化(Dey等,2025);HybridVLA在同一模型内融合扩散与自回归解码,并通过协同集成自适应选择生成策略(Liu等,2025a);VOTE(vision-language-action optimization with trajectory ensemble voting)则基于历史动作的相似性进行分组与平均,形成更鲁棒的输出(Lin等,2025b)。为实现对物理规律的内化,WorldVLA在单一自回归框架下联合建模世界模型与动作模型(Cen等,2025);UnifiedVLA将世界建模作为后训练任务,从大规模无标注视频中学习因果结构(Wang等,2025f);UP-VLA通过预测下一帧图像,并据此生成更符合物理约束的动作(Zhang等,2025h)。

单系统模型需满足真实机器人高频控制的时延约束,推理效率优化通常与骨干结构及解码模块一体化设计密切相关。在动态计算与路由方面,MoLe-VLA在视觉—语言联合投影空间选择最相关LLM层实现动态推理(Zhang等,2026);DeeR-VLA通过多处提前退出机制,基于一致性判定提前终止深层计算(Yue等,2024);CogVLA采用指令驱动稀疏化路由,自上而下压缩视觉词元(Li等,2025g);RoboMamba引入Mamba序列建模,以线性复杂度实现推理加速(Liu等,2024a)。在模型轻量化方面,BitVLA采用蒸馏感知量化视觉编码器,并利用BitNet b1.58作为语言骨干(Wang等,2025b)。在解码机制方面,RoboFlamingo使用独立动作模块将VLM输出映射为动作序列(Li等,2024b);PD-VLA将自回归解码重构为定点迭代的非线性方程,利用双向注意力迭代更新词元以实现并行推理(Song等,2025a);Flash-VLA则评估动作与环境稳定性,动态决定是否跳过本轮推理,以进一步降低端到端时延(Tan等,2025b)。

从表3可以看出,单系统VLA在设计上呈现出较为明确的分化趋势。一方面,以RT-2(Zitkovich等,2023)及其后续工作为代表的路线,依托大规模VLM骨干网络和互联网数据,主要通过模仿学习和微调的范式,强化跨任务与跨本体的零样本泛化能力,但模型参数规模庞大,训练与推理开销均较高,对算例和硬件资源具有较高要求;另一方面,以OpenVLA(Kim等,2024)、LEO(Huang等,2024a)、SpatialVLA(Qu等,2025)等为代表的开源体系,则在参数规模、视觉编码形式与动作词元设计等方面进行权衡,更加关注在通用性和可部署性之间取得合理平衡。此外,ReVLA(Dey等,2025)、TraceVLA(Zheng等,2025b)、FuSe(Jones等,2025)和UniAct(Zheng等,2025a)等工作分别在思维链引入、策略可解释性、时间轨迹提示以及统一动作空间等方面提出改进,使单系统架构在推理透明度、跨模态对齐质量和控制平滑性等维度上各有侧重。

总体来看,单系统架构的核心优势在于其结构简洁与端到端联合优化的潜力。通过将多种功能集成于单一模型之中,降低模块间的信息损失与集成误差,有利于实现从互联网知识到机器人技能的直接迁移,并在计算资源相对充足的条件下展现出突出的零样本泛化能力。然而,这种高度耦合的设计也带来一系列挑战:1)训练与推理的计算负担较重,难以满足高频、低时延控制需求;2)决策过程高度黑箱化,可解释性不足,不利于安全性验证与故障诊断;3)对任务或平台的轻微变动往往需要重新微调甚至重新训练整个模型,灵活性有限。因此,单系统VLA更适用于对实时性要求相对不高、更侧重强语

义泛化能力的研究型平台或离线任务规划等应用场景。

3.2 双系统VLA

双系统VLA采用慢推理和快控制的协同架构，其与单系统的本质区别在于推理—控制频率的显式解耦与异步协同。表4展示了双系统模型VLA方法及其特点。该体系并行部署两个功能各异的子系统：系统2（慢系统）负责低频的高层语义理解与任务编排，系统1（快系统）负责高频的底层实时控制（Ouyang 等，2022；Zitkovich 等，2023）。以 pi-0（Black 等，2024）和 GR00T N1（Bjorck 等，2025）为代表的模型，虽然在组件上同样使用了VLM与流匹配/扩散策略的组合，但其架构设计允许VLM以较低频率更新高层语义条件或潜变量，而底层策略网络则以较高频率基于最新观测进行闭环控制。这种频率解耦机制解决了大模型推理延迟与机器人高频控制需求之间的矛盾，是双系统架构判定的核心标准。该体系旨在兼顾多模态深层推理能力与机器人高频控制需求。系统2通常为规模较大的 MLLM（multimodal large language model）/VLA，依赖逻辑推理与多源信息整合，承担长时目标分解、技能组合与安全约束等高层功能；系统1则为轻量级策略网络，响应

表4 双系统模型VLA方法及其特点

Table 4 Dual-system VLA methods and their characteristics

方法	System2 骨干网络	System1 学习方式	特点
DP-VLA（Han等，2024）	OpenVLA	回归	提出双系统架构用于机器人操作
RoboDual（Bu等，2024）	OpenVLA	扩散式	引入用于控制的DiT专家模型
LCB（Shentu等，2024）	LLaVA	扩散式	新增特殊词元编码并作为策略条件
GR00T N1（Bjorck等，2025）	Eagle-2	流匹配	结合VLM和DiT用于人形机器人操作
CogACT（Li等，2024a）	OpenVLA	扩散式	将动作扩散过程融入VLA
HiRT（Zhang等，2024a）	InstructCLIP	回归	提出双系统模型，System 2以较低频率运行
Fast-in-Slow（Chen等，2025a）	Prismatic	扩散式、自回归	在基于VLM的推理器中提高执行效率
OpenHelix（Cui等，2025）	LLaVA	扩散式	对连接VLM和策略的词元进行辅助训练
ChatVLA-1（Zhou等，2025c）	Qwen2-VL	扩散式	用于独立感知和控制的共享注意力机制
ChatVLA-2（Zhou等，2025b）	Qwen2-VL	扩散式	动态路由和开放世界机器人推理
Diffusion-VLA（Wen等，2025a）	Qwen2-VL	扩散式	通过FiLM将Qwen2-VL推理与扩散动作融合
TriVLA（Liu等，2025d）	Eagle-2	扩散式	将动力学感知模块作为System 3补充静态感知
GF-VLA（Li等，2025b）	LLaVA-v1.5	回归	通过人类视频的信息实现可解释的双手操作
RationalVLA（Song等，2025b）	LLaMA2	扩散式	引入可学习的隐式接口，实现鲁棒操作
VQ-VLA（Wang等，2025g）	OpenVLA	变分自编码器	引入向量量化的动作词元，实现平滑的控制
TinyVLA（Wen等，2024b）	LLaVA	扩散式	证明高性能VLA无需大规模机器人预训练
pi-0（Black等，2024）	PaliGemma	流匹配	结合VLM和基于流匹配的动作专家模型
pi-0-FAST（Pertsch等，2025）	pi-0	自回归	提出基于DCT的动作词元，高效自回归训练
pi-0.5（Black等，2025）	PaliGemma	流匹配	将高层提示转换为更细粒度的子任务预测
ForceVLA（Yu等，2025）	pi-0	扩散式	通过混合专家模型处理力传感模态信息
SmolVLA（Shukor等，2025）	SmolVLM-2	流匹配	采用SmolVLM-2和流匹配构建轻量级VLA
GR-3（Cheang等，2025）	Qwen2.5-VL	流匹配	结合VLM数据和少量轨迹标注，实现鲁棒操作
Villa-X（Chen等，2025d）	PaliGemma	流匹配	在扩散过程中集成本体感知隐变量信息
GraspVLA（Deng等，2025）	InternLM2	流匹配	利用合成数据实现仿真到开放世界的抓取

快速、自动化程度高，负责末端稳定控制、抓取自适应等底层动作生成，可实现毫秒级延迟，从而满足实时控制要求。两个子系统并行运行但更新频率不同：慢系统以较低频率在高层语义层面进行决策；快系统则以高频率输出连续控制指令。由于慢系统输出存在时间延迟，需通过跨系统接口将其输出的条件编码传递给快系统，由动作模块快速解码为可执行动作。与单系统架构相比，双系统的关键区别在于动作生成通道的显式解耦，相较传统层次化架构，它不强求中间结果为人类可读形式，而是借助可学习接口实现语义到控制的高效映射，从而具备更快的响应速度、更平滑的部署流程以及更强的底层控制定制能力。

早期工作LCB（latent codes bridges）已提出类似结构，而DP-VLA（Han等，2024）首次引入双过程理论，在统一模型框架下，上层由VLM处理多模态输入、执行语义落地与子任务提炼，下层由专门的动作模块在高频率下解码执行。动作模块既可设计为与主干交互的独立模块，也可复用主干部分参数形成一体化实现。在独立动作专家设计中，CogACT与GR00T N1采用扩散模型作为控制模块，通过端到端耦合加速策略步进（Bjorck等，2025；Li等，2024a）。在时序协同与多模态扩展方面，HiRT通过VLM在低频阶段提取场景语义信息，由轻量视觉策略在高频阶段执行具体动作（Zhang等，2024a）；TriVLA将未来帧视频生成、指令理解与扩散动作专家组合为联动架构（Liu等，2025d）；GF-VLA从人类示教数据中提取“手—物”场景图，并与LLM融合输出可解释的行为树与双臂笛卡尔动作（Li等，2025b）；RationalVLA设计可学习的隐式接口，使VLM能够向控制器发出生成/拒绝词元（Song等，2025b）；VQVLA采用卷积残差VQ-VAE替代传统分箱表征，随模拟数据规模扩大性能呈线性提升，并有效缩小仿真与现实间的差距（Wang等，2025g）。在一体化双速设计中，Fast-in-Slow将动作专家内嵌于VLM主干末端模块，以不同频率分别驱动慢推理与快控制过程（Chen等，2025a）。依据本节提出的分类标准，GR00T N1可归类为双系统架构，系统2的VLM与系统1的扩散模块具有独立的网络结构和参数，可分别优化，并且系统2与系统1的运行频率存在较大差异，符合快慢系统的核心定义，系统2的输出以紧凑的隐式编码形式传递给系统1，作为扩散过程的条件输入。

在跨系统信息交互方面，模型可在共享自注意力层中实现视觉/语言词元与噪声、机器人状态等信息的交互，或由主干输出键值对供动作模块通过交叉注意力进行读取。如pi-0（Black等，2024）使用预训练VLM初始化骨干网络，配合流匹配动作模块，处理机器人专用输入并生成动作；ForceVLA（Yu等，2025）在pi-0基础上，引入基于专家混合结构的力觉融合模块集成力觉信息；OneTwoVLA支持在显式推理与动作生成两种模式间切换，实现异步协同（Lin等，2025a）。pi-0.5首先将高层提示细化为具体子任务，再交由基于pi-0的双系统处理；pi-0、FAST采用离散余弦变换驱动的动作分词策略，提升自回归训练效率（Pertsch等，2025）。GraspVLA将自回归感知与流匹配控制统一为渐进式动作生成的链式思维过程，并实现仿真到现实的直接迁移（Deng等，2025）；Villa-X在机器人状态上锚定隐式动作，通过联合扩散模型同时建模实体动作，实现结构化视觉—动作融合（Chen等，2025d）；GR-3则使用流匹配技术统一视觉语言理解与轨迹学习过程（Cheang等，2025）。双系统架构在人形机器人领域的代表性工作为Figure AI推出的Helix，其中系统2采用7B规模的VLM，以7～9 Hz的频率运行，将视觉观测与语言指令压缩为一个连续的潜向量，提炼出任务相关的语义信息；系统1采用轻量级策略模块，接收来自系统2的隐变量作为条件约束，结合高频视觉特征，以200 Hz的频率直接输出连续动作值。与依赖复杂动作词元化的模型不同，Helix实现了35自由度的高维控制，并展现出较强的零样本泛化能力，验证了双系统架构在处理复杂物理交互任务中的优势。

与单系统VLA相比，双系统架构在能力分工与性能取舍方面更加清晰：慢系统侧重多源信息融合和长程任务规划，能够充分利用大模型在语义推理与知识整合方面的优势；快系统则通常由轻量级策略网络或专用控制模块构成，在控制频率、响应延迟和稳定性等指标上更接近传统机器人控制器。双系统VLA适用于同时设计复杂语义推理与高频闭环控制的任务场景，如交互式导航、动态环境下的操作等。由表4可见，不同工作主要在3个方面体现差异：1）慢、快系统是否共享部分骨干参数，共享有助于降低整体模型规模，但在任务迁移时可能引入额外的耦合约束；2）跨系统接口采用何种形式的条件编码（如中间语言描述、可供性表示或动作隐变量），

直接影响系统的可解释性以及端到端优化的难度；3)快系统是否支持在线更新与自适应控制，这在很大程度上决定了系统在真实环境中应对外界扰动和模型偏差时的鲁棒性。

总体而言，双系统VLA通过结构解耦在一定程度上实现了全局深度推理与局部快速响应之间的平衡，有效缓解了单系统在实时性与部署上的瓶颈，提升了动态环境中的响应速度和整体控制稳定性，但也引入了接口设计与协同训练的附加复杂度。慢、快系统之间的接口设计容易演化为新的性能瓶颈与工程难点，一旦接口信息传递不充分或存在显著延迟，便可能导致快系统的执行行为偏离慢系统的高层意图。

3.3 层次化系统VLA

从模型设计的角度，层次化系统VLA以可解释的中间表征作为关键接口，将高层任务规划与底层运动控制在同一体系内进行松耦合对接：规划器接收环境观测与语言指令，生成人类可读、可验证的中间表征，策略模块则基于该表征输出可直接下发的动作或控制代码。该范式既保留了大型VLM的语义理解与复杂推理能力，又兼顾了平台侧的控制精度与系统可移植性，并支持规划端与控制端的独立优化与灵活组合(Sapkota等，2025)。表5展示了层次化系统模型VLA方法及其特点。

根据中间表征的形态，现有研究主要形成3条技术路径：程序化表征、关键点表征、子任务序列表征。程序化表征通过将任务目标转化为可执行代码或辅助程序，实现从自然语言到控制指令的显式映射。例如，Chain-of-Modality通过多轮多模态提示生成Python程序，提升了推理链的可解释性(Wang等，2025a)；Instruct2Act直接调用程序接口驱动硬件(Huang等，2023b)；ReLEP结合记忆库将复杂任务拆解为技能库中的原子技能，实现多步规划的复用(Liu等，2024b)。关键点表征以离散关键点、路径点或可供性区域作为子目标的空间载体，将语言语义落地为可操作的空间约束。如MoManipVLA在每一步预测关键位姿并通过双层轨迹优化细化控制量(Wu等，2025c)；RoboPoint将语言指令转化为可交互目标的视觉关键点(Yuan等，2024b)；ManipLVLM-R1结合分组相对策略优化，同时预测抓取区域与目标轨迹(Song等，2026)；RoboBrain融合LLaVA规划与A-LoRA、T-LoRA模块，实现可供性叠加关键点表达(Ji等，2025)。子任务序列表征以可读的自然语言指令链作为衔接规划与控制的中间层。例如Embodied-Reasoner(Zhang等，2025k)采用观测、思维和动作轨迹链，显式引入空间分析、反思与校验机制，提升分步规划的可靠性；Reinforced Planning采用监督训练与对比优化结合的两阶段方法，提高分布外任务的泛化(Wu等，2025a) ViLA将GPT-4V作为外部规划器，仅执行首条候选指令以确保控制稳定(Hu等，2023)。

当高层规划器与底层策略共同建模时，层次化架构可在保持可解释性的同时实现更强闭环性能。典型工作包括：HAMSTER通过关键点预测路径执行(Li等，2025d)；ReKep将关键点转化为代价函数并优化求解(Huang等，2024b)；A0引入本体无关的可供性表征以实现跨机器人迁移(Xu等，2025)；DexVLA结合规划器与扩散策略协同处理长时操作(Wen等，2025b)；SkillDiffuser利用技能集合预测与扩散执行实现动作细化(Liang等，2024)。此外，RoBridge (Zhang 等，2025j)、RoboMatrix (Mao 等，2024)、HiBerNAC (Wu等，2025b)与MALMM(Singh等，2024)分别从原子动作生成、结构和多智能体等角度扩展层次化实现。还有研究在层次化系统在组合与协同方式上呈现多样性。RT-H通过语言化运动实现动态切换(Belkhale等，2024)；pi-0.5将离散动作词元与流匹配控制结合(Black 等，2025)；GR00T N1整合多种具身技术以实现跨本体泛化(Bjorck 等，2025)； OpenHelix(Cui 等，2025)与DP-VLA(Han等，2024)采用潜在接口对齐规划与控制。

层次化架构的突出优势体现在可解释性与模块化两个方面。一方面，通过引入中间表征(如程序、关键点、子任务序列)，将智能体的决策过程白盒化，便于人工调试、安全校验和系统集成；另一方面，高层规划器与底层策略模块可以相对独立地演化和优化，提升系统的灵活性与可维护性。然而这种解耦也可能导致高层规划的微小偏差在底层执行时被放大，导致性能退化甚至任务失败。如何设计能够被机器精准执行、又能保持语义丰富性的中间表征也是一项关键挑战。总体来看，层次化系统在需要与人类协同、对安全性和可解释性具有严格要求的应用场景(如工业装配、医疗辅助)中具有不可替代的优势，并为集成经典规划与控制算法提供天然接口。

综合表3—表5所示方法可以看出，3类VLA系

表5　层次化系统模型VLA方法及其特点

Table 5　Hierarchical-system VLA methods and their characteristics

方法	骨干网络	学习方式	特点
MoManipVLA（Wu等，2025c）	OpenVLA-7B	模仿学习（IM）	预测路径点并优化全身轨迹
ManipLVM-R1（Song等，2026）	Qwen2.5-VL-3B	强化学习（RL）	对轨迹进行GRPO调优，分布外场景中表现鲁棒
PaLM-E（Driess等，2023）	PaLM	监督微调（SFT）	训练具备通用机器人操作指令生成能力的VLM
Embodied-Reasoner（Zhang等，2025k）	Qwen2-VL-7B	监督微调（SFT）	构建VLM和开源数据集用于规划、推理和反思
RoboPoint（Yuan等，2024b）	Vicuna-v1.5-13B	监督微调（SFT）	以路径点形式实现VLM空间感知决策
Reinforced Planning（Wu等，2025a）	Qwen2.5-VL-7B	监督微调（SFT）、强化学习（RL）	在微调模型上进行GRPO，提升泛化能力
CoM（Wang等，2025a）	Gemini 1.5 Pro	API接口	从演示中通过序列多模态提示提取力感知操作
RoVI（Li等，2025e）	GPT-4 /LLaVA-13B	监督微调（SFT）	通过层次化结构实现精确操作
ReLEP（Liu等，2024b）	LLaVA-1.6-7B	监督微调（SFT）	具有隐式逻辑推理和幻觉缓解的规划框架
ViLA（Hu等，2023）	GPT-4V	API接口	提供感知、推理集成的规划器
RoboBrain（Ji等，2025）	LLaVA	监督微调（SFT）	设计规划、可操作性和轨迹的层次化VLA
HiRobot（Shi等，2025）	PaliGemma-3B/pi-0	监督微调（SFT）、模仿学习（IM）	具有可解释性和复杂任务处理能力
DexVLA（Wen等，2025b）	Qwen2-VL	监督微调（SFT）、模仿学习（IM）	结合VLM和扩散模块，采用3阶段训练方案
Instruct2Act（Huang等，2023b）	ChatGPT	监督微调（SFT）	生成调用API的程序，实现从指令到动作的映射
RoboMatrix（Mao等，2024）	Vicuna-v1.5	监督微调（SFT）、模仿学习（IM）	具有模块化调度层、技能层和硬件层的VLA
PointVLA（Li等，2026）	Qwen2-VL	API接口	附加点云编码器以获得空间感知
A0（Xu等，2025）	Qwen2.5-7B	监督微调（SFT）、API	具有关键点预测的分层扩散模块
RoBridge（Zhang等，2025j）	GPT-4o	模仿学习（IM）、强化学习（RL）	通过VLM与RL动作预测，提高模型迁移性
RT-H（Belkhale等，2024）	PaLI-X 55B	监督微调（SFT）、模仿学习（IM）	动作层次结构，使用语言驱动的动作表征
ReKep（Huang等，2024b）	GPT-4o	API接口	通过关键点约束实现无训练的操作轨迹生成
HiBerNAC（Wu等，2025b）	PaLM2	监督微调（SFT）	提出模仿大脑分层结构的异步模型
LLARVA（Niu等，2024）	LLaMa2 7B	监督微调（SFT）、模仿学习（IM）	视觉—动作指令调优，结合大型指令调优数据集
VLA-Touch（Bi等，2025）	GPT-4o	模仿学习（IM）	基于扩散的模仿学习将触觉传感集成到VLA中

统架构在适用场景与研究价值上各具特点：单系统架构强调整体统一与端到端优化，便于系统性分析动作词元设计及多模态对齐机制，有利于探究规模法则与跨任务泛化规律；双系统架构在具身智能的语义推理与底层控制之间建立起较为清晰的职能分工，更接近人类系统1和系统2形式的认知模式，适合研究推理—控制协同机制及高频物理交互问题；层次化架构通过引入中间可解释表征，在延续传统

感知—建图—规划—控制流程优势的同时，保留了大模型在语义理解和知识迁移方面的能力，为后续引入形式化验证和安全约束提供了接口。

4 VLA 训练与评估

4.1 训练数据

VLA 训练的目标是学习具备可执行性与可组合性的动作词元，使其能够在视觉观测和语言的基础上实现语义落地，并与具体本体的感知运动系统有效对齐。由此形成的底层时序语义—中层任务结构—顶层物理落地 3 层监督机制，共同构建了可迁移、可扩展的 VLA 预训练基础。由于单一数据源通常难以覆盖所有需求，当前主流实践形成了按具身特异性递增、数据规模递减的分层多源范式（Bjorck 等，2025）：底层利用海量网页与人类视频提供视觉—语言对齐和时序语义先验；中层借助仿真与合成数据承载任务结构与技能组合；顶层则通过真实机器人数据注入物理真实感与执行可行性。

底层的视频数据如 Ego4D（Grauman 等，2025）、EPIC-KITCHENS-100（Damen 等，2022）、Something-Something V2（Goyal 等，2017），虽然缺乏直接的动作标注，但蕴含丰富的人物交互模式与物理常识，可用于预训练时序视觉编码器与动作先验模型，支持轨迹提取、状态迁移与隐式动作建模。通过与时序对齐等弱监督手段，可进一步得到基于轨迹或目标状态的动作词元，并结合第三人称视角增强三维语义落地能力（Zhong 等，2025a）。

中层的合成数据与环境仿真，如 MimicGen（Mandlekar 等，2023）、DexMimicGen（Jiang 等，2025c）、RoboCasa（Nasiriany 等，2024）、robosuite（Zhu 等，2020）、Habitat（Savva 等，2019）、Isaac Gym（Makoviychuk 等，2021）等，在低成本条件下显著扩展任务多样性与环境随机性。通过远程操控、算法规划与强化学习自主收集，可生成覆盖广泛技能的高质量数据，有利于在安全关键场景与复杂接触任务中学习策略。

顶层的真实机器人数据为底层控制提供可靠的物理监督，是支撑跨本体泛化的关键。该类数据不仅包含失败样本与多模态传感信息，如力矩、音频等，还强化了对接触力学与声学环境的推理能力（Shafiullah 等，2023）。面向长程任务规划与开放室内场景的专用数据集，以及自动驾驶中的轨迹标注，共同构成该层级的重要支柱（Wu 等，2024b）。

4.2 预训练

1）单一领域数据训练。VLA 预训练依赖大规模机器人轨迹的有监督学习，典型做法是在公开的多机器人数据集上进行预训练，再使用特定任务的小规模数据进行微调。例如 Octo（Ghosh 等，2024）基于 Open X-Embodiment 数据预训练后，再针对当前任务进行适配；RDT（Liu 等，2025c）则构建大规模双臂操作数据进行预训练，并在本地采集的双臂数据上微调，取得了优异的操控性能。然而与互联网级别的预训练语料相比，现有机械臂数据在规模与多样性方面明显不足，导致此类模型在视觉与任务泛化方面存在局限，容易面对新物体、新场景或新指令出现性能退化。

2）分阶段跨领域数据训练。为突破单一数据源的局限性，提升模型在语义理解、时序推理与物理执行之间的协同能力，跨领域数据训练已成为 VLA 预训练的重要方向。分阶段跨领域数据训练遵循由通用到专用的渐进思路，通过预训练、微调的方式将通用基础模型的表征能力逐步迁移至具身决策任务。例如 OpenVLA（Kim 等，2024）直接沿用 LLaMA 的语言模型参数，并将视觉编码与动作序列进行对齐训练；RT-2（Zitkovich 等，2023）则先在互联网规模的图文数据上进行视觉语言对齐，再通过机器人轨迹微调注入操控能力，从而兼顾语义泛化与物理可行性。为进一步增强模型对动态场景与时序关系的建模能力，视频数据被广泛引入作为中间桥梁。GR-1（Wu 等，2024a）将视频预测与动作生成统一于同一架构，通过视频预训练和轨迹微调的方式提升长程任务性能；LAPA 与 GO-1 则借助 VQ-VAE 等技术从视频中学习离散的隐式动作表征，有效链接视觉观测与连续控制（Ye 等，2025）。针对人机行为分布差异，EgoVLA（Yang 等，2025a）等研究利用第一人称视角的人类活动视频进行预训练，再以少量机器人数据实现高效技能迁移。此类方法通过分层次、分阶段地引入不同抽象级别的监督信号，在模型通用性与任务适应性之间取得良好平衡。

3）联合多任务跨领域数据训练。通过在统一模型中实现对多源异构数据的协同学习，以提升跨任务与跨场景的泛化能力，同时缓解分阶段训练中可能出现的灾难性遗忘问题。GATO（Reed 等，2022）

较早在包含图像、文本、游戏与机器人控制等多种任务的混合数据集上进行训练，验证跨模态统一学习的可行性。后续工作如UP-VLA（Zhang等，2025h）将多模态理解与未来帧预测联合建模，在强化物理认知的同时保留语义解析能力。为应对多任务间的参数竞争与优化冲突，模型结构设计尤为重要，ChatVLA系列（Zhou等，2025c）采用混合专家架构，对机器人控制、环境理解等不同功能进行模块化隔离，并进一步通过动态路由机制实现任务自适应，显著降低任务间的负面干扰。联合训练的成功依赖于合理的数据配比、有效的路由策略以及任务间的结构化解耦，其优势在于能够构建一个兼具通用知识响应能力与精确执行能力的具身智能体。

4.3 后训练

4.3.1 自监督对齐

在预训练表征的基础上，自监督对齐通过进一步缩小多模态语义差距，实现在有限标注条件下提升下游任务的收敛效率。首先构建视觉—文本联合嵌入空间，随后基于该对齐表征对动作模块进行任务适配，从而显著减少对手工动作标注的依赖，实验表明该策略能够有效提升基础操作任务的训练速度与稳定性。为进一步增强模型对复杂场景的适应能力，可引入弱监督视频信号并结合高质量场景增强技术，在后训练阶段注入时空先验与长尾场景覆盖。此类方法在应对杂乱环境、遮挡及动态光照等挑战时，可一定程度上提升鲁棒性（Yang等，2023）。需要指出的是，自监督对齐并非孤立过程，其与后续监督微调协同作用，能够在保持语义一致性的同时，大幅降低对任务专用数据量的需求。

4.3.2 参数高效监督微调

监督微调是当前主流的后训练路径，其核心是在预训练VLA基础上，使用少量任务相关的机器人轨迹进行监督学习，以适配目标平台与任务需求（Chen等，2025f）。轨迹数据通常通过遥操作方式采集，常见方案包括主从臂映射、基于人体运动捕捉的虚拟动作映射，以及末端空间直接采集等方式。近年来也涌现出不依赖实体机械臂的低成本采集方案（Sharshar等，2025），进一步提升了数据采集的灵活性与效率。尽管监督微调具备实现简单、训练稳定的优势，但其性能上限受到数据一致性与累积误差的制约（Sameni等，2024）。为降低微调成本并保留模型的通用能力，参数高效微调技术得到广泛应用。其中，低秩适配通过在冻结的Transformer层中插入小规模可训练矩阵，可将可训练参数量缩减至全参数微调的约30%，同时保持性能基本不变（Hu等，2022）。针对高频控制场景，轻量适配器与静态骨干的组合可实现200 Hz的连续控制（Pertsch等，2025）。

模型压缩方面，量化可将模型体积减半并近似实现吞吐量翻倍。采用逐通道校准的后训练量化技术有助于缓解高动态范围输入引起的精度损失。在结构化剪枝方面，对视觉编码器与Transformer层的冗余注意力模块或网络层进行剪枝，如通过扩散策略可减少约20%的参数量，而抓取稳定性基本不受影响（Chi等，2025），可在成功率损失低于2%的前提下将内存占用压缩约25%，为资源受限的设备部署创造条件。

为满足高频控制与低延迟需求，后训练阶段常采用动作表征压缩技术。FAST（Pertsch等，2025）方法将长时间窗口的连续控制信号投影至频域并离散化，显著缩短动作词元序列，降低解码开销。例如，FAST将1 000 ms控制窗口压缩为16个词元，在扩散模块支持下实现约15倍推理加速与200 Hz的策略更新频率。在双系统架构中，并行解码技术可同步生成时空词元，使端到端延迟降低约2.5倍；GR00T N1实现了单步延迟低于5 ms，且在7-DoF、100 Hz控制下位姿误差增量小于3 mm。动作分块技术进一步将多步操作流程抽象为单一词元，在长程厨房任务中可减少约40%的解码步数（Song等，2025a）。结合硬件感知优化，如张量核量化、内核融合与注意力机制（Kim等，2024），可将峰值显存需求控制在8 GB以下，并在RTX A2000等设备上实现30%的延迟降低与25%的能效提升，支持50 W级边缘GPU达到30 Hz的闭环控制。

4.3.3 强化学习微调

强化学习为VLA的后训练提供了超越纯监督方法的泛化能力与任务适应性。通过引入任务回报信号，强化学习能够有效缓解模仿学习对专家数据规模与质量的依赖，提升模型在分布外场景的鲁棒性。当前强化学习微调主要围绕3个技术方向展开：端到端强化学习微调、强化学习底层策略优化以及强化学习用于数据生成。

1）端到端强化学习微调。该方向直接对VLA或其关键组件进行强化学习优化，以提升整体策略

性能。ConRFT(Chen等,2025e)提出两阶段微调框架,首先利用离线数据对策略与值函数进行初始化,再通过人在回路的在线强化学习实现快速适配,显著提升了部署安全性。针对奖励稀疏的问题,VLA-RL(Lu等,2025)引入过程奖励模型,通过分析动作序列的中间进展构建稠密伪奖励,有效稳定了基于PPO (proximal policy optimization) (Schulman 等, 2017)的训练过程;ReWiND(Zhang等,2025g)则利用视觉相似度度量任务进度,形成连续的进度奖励;RIPT-VLA采用留一法估计PPO的值函数,结合动态拒绝采样机制,增强稀疏奖励下的训练稳定性;在训练架构方面,iRe-VLA(Guo等,2025)通过监督微调、在线强化学习以及监督微调的循环框架,实现策略性能的持续提升;TGRPO(Chen等,2025f)通过在线交互中的成败轨迹隐式建模奖励,提出轨迹级偏好对齐机制;ReinboT(Zhang等,2025e)将累积回报目标显式融入端到端训练,使模型能从混合质量轨迹中学习鲁棒策略。

2)强化学习底层策略。该路径建立明确的功能分工架构,由VLA负责高层语义理解与任务规划,强化学习策略负责底层精细控制。在机器人运动控制中,Humanoid-VLA(Ding等,2025a)由VLA生成日志级运动指令,强化学习训练的全身控制器执行人形机器人的精细运动;NaVILA(Cheng等,2024)将VLA的速度级命令转化为足式机器人的扭矩控制;在移动操作系统层面,SLIM (sim-to-real legged instructive manipulation)(Zhang等,2025d)采用教师—学生蒸馏框架,首先使用强化学习教师策略生成运动轨迹,再通过模仿学习将其蒸馏至学生VLA;RPD(refined policy distillation)(Jülg等,2025)则反向利用VLA作为“教师模型”引导强化学习探索过程,通过语义先验塑造策略学习方向;对于特定技能优化,DexTOG(Zhang等,2025i)首先利用扩散模型生成多样化抓取姿态,再通过强化学习评估任务成功率,迭代优化得到物体特定的高质量抓取策略;LUMOS(Nematollahi等,2025)在世界模型的隐空间中进行模仿学习与强化学习,通过内在奖励度量专家轨迹的偏离程度。

3)强化学习用于数据生成。强化学习作为高质量数据生成引擎,能够显著提升后续监督训练的效果。强化学习首先通过人在回路的强化学习将任务成功率提升至专家水平,再通过监督微调将策略知识蒸馏到基础模型中,实现多专家策略的一体化整合(Xu等,2024a);iRe-VLA(Guo等,2025)采用交替训练策略,在在线强化学习探索收集新成功轨迹与基于扩展数据集的监督微调之间循环,形成“模仿—探索—再模仿”的自我改进闭环。该范式能够与参数高效微调、动作词元压缩等技术协同,在不显著增大模型规模的前提下持续提升性能。

4.4 性能评估

4.4.1 仿真环境和基准测试

为克服真实世界数据采集成本高、安全性低及可重复性差等局限,基于物理引擎的仿真环境广泛用于生成大规模可控的交互数据(Ehsani等,2024)。此类环境支持多模态感知、部分长程任务规划与自动化评估,契合VLA的训练需求。

在复杂任务仿真方面,BEHAVIOR(Li等,2022a)支持家庭杂乱场景中的多步骤语义控制;ALFRED(Shridhar等,2020)专注于以第一视角语言指令描述的长视野任务;RLBench(James等,2020)提供了基于RGB-D的桌面精细操作任务,并进一步扩展至双手语言条件操作。为提升多技能泛化能力,Meta-World(Yu等,2019)与LIBERO(Liu等,2023a)等基准引入了多场景语言标注与视觉—语言奖励机制。

表6展示了多种代表性的VLA工作在LIBERO的性能对比,数据综合整理自多篇原始研究论文在LIBERO基准上报告的性能。主要来源包括VLA-Adapter(Wang等,2026)、UniVLA(Bu等,2025b)及相关基线方法在论文中公布的结果。

表6中,Spatial测试对物体空间关系的记忆与区分能力,Object评估对新物体类别的识别与操作能力,Goal要求机器人根据不断变化的任务目标执行相应的动作序列,Long专门用于评测复杂多步骤任务的执行能力。CALVIN(Mees等,2022)支持无约束语言指令下的多阶段操作,适用于长序列评估;MIKASA-Robo(Cherepanov等,2025)则针对部分可观测环境设计考察记忆依赖型智能体表现的桌面任务。本文仅对其进行归纳呈现,并保持原文的评价指标与实验设置不变。受限于不同文献在模型超参数、训练步数及随机种子等底层实现细节上的差异,表6的数据横向比较具有一定的可比性边界,仅供技术路线潜力的定性参考,不宜作为绝对的基准优劣判定依据。读者在参考这些数据时,应结合各原

表6　不同VLA方法在LIBERO数据集的性能

Table 6　Performance of different VLA methods on LIBERO dataset

模型	参数量	成功率/%			
		Spatial	Object	Goal	Long
Diffusion Policy(Chi等,2025)	–	78.3	92.5	68.3	50.5
MaIL(Jia等,2024)	24.7 M	74.3	90.1	81.8	78.6
MDT(Reuss等,2024)	75 M	78.5	87.5	73.5	64.8
Octo(Ghosh等,2024)	93 M	78.9	85.7	84.6	51.1
VLA-OS(Gao等,2025)	0.5 B	87.0	96.5	92.7	66.0
VLA-Adapter(Wang等,2026)	0.5 B	**97.8**	**99.2**	**97.2**	**95.0**
GraspVLA(Deng等,2025)	1.8 B	–	94.1	91.2	82.0
GR00T N1(Bjorck等,2025)	2 B	94.4	97.6	93.0	90.6
SmolVLA(Shukor等,2025)	2.2 B	93.0	94.0	91.0	77.0
NORA(Hung等,2025)	3 B	92.2	95.4	89.4	74.6
pi-0-FAST(Pertsch等,2025)	3 B	96.4	96.8	88.6	60.2
pi-0(Black等,2024)	3 B	96.8	98.8	95.8	85.2
SpatialVLA(Qu等,2025)	4 B	88.2	89.9	78.6	55.5
4D-VLA(Zhang等,2025f)	4 B	88.9	95.2	90.9	79.1
PD-VLA(Song等,2025a)	7 B	95.5	96.7	94.9	91.7
ThinkAct(Huang等,2025a)	7 B	88.3	91.4	87.1	70.9
MolmoAct(Lee等,2025)	7 B	87.0	95.4	87.6	77.2
TraceVLA(Zheng等,2025b)	7 B	84.6	85.2	75.1	54.1
WorldVLA(Cen等,2025)	7 B	87.6	96.2	83.4	60.0
CoT-VLA(Zhao等,2025a)	7 B	87.5	91.6	87.6	69.0
UniVLA(Bu等,2025b)	7 B	96.5	96.8	95.6	92.0
OpenVLA(Kim等,2024)	7 B	84.7	88.4	79.2	53.7
LAPA(Ye等,2025)	7 B	73.8	74.6	58.8	55.4
UnifiedVLA(Wang等,2025f)	8.5 B	95.4	98.8	93.6	94.0
FlowVLA(Zhong等,2025b)	8.5 B	93.2	95.0	91.6	72.6

注:加粗字体表示该列指标下的最优结果。“–”表示原文未说明具体参数量或对应指标的测试结果。

论文的具体实验设置进行综合评估。

在仿真真实性提升方面,SIMPLER(Li等,2024c)通过校准仿真环境与现实机器人配置,有效减小感知与控制层面的领域差异。如图4所示,高保真仿真平台如Habitat(Savva等,2019)与SAPIEN(Xiang等,2020)在真实三维环境中支持空间推理与操作学习,为VLA研究提供可扩展基础设施。此外,COLOSSEUM(Pumacay等,2024)与VLABench(Zhang等,2024c)分别针对分布偏移鲁棒性与通用语言条件操作进行了专门优化。

在仿真数据生成方法上,RoboTurk(Mandlekar等,2018)基于MuJoCo引擎通过云端遥操作收集任务演示;MimicGen(Mandlekar等,2023)提出从少量专家演示中分解重组为新轨迹的方法,显著提高数据生成的效率。此外,COSMOS(Agarwal等,2025)等大规模视频世界模型通过生成多样化轨迹,为VLA训练提供丰富数据源。

在评估指标体系方面,iGibson(Xia等,2020)提

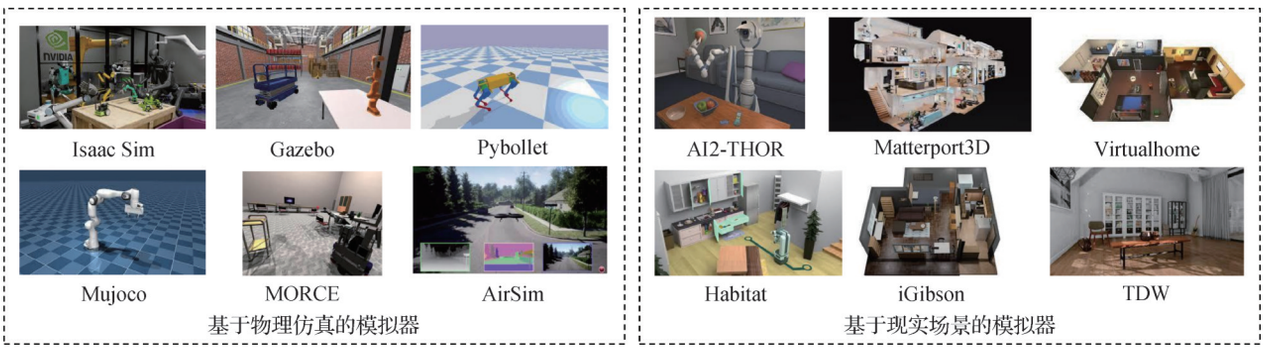

图 4　具身智能领域的模拟器

Fig. 4　Simulators of embodied AI

出包含路径效率与努力效率的交互导航评分，用于衡量导航过程中物理交互的有效性。VIMA-BENCH（Jiang 等，2022）设计了 4 层泛化能力评估协议，从物体位置随机化到任务结构全新组合，逐级测试模型性能。CALVIN（Mees 等，2022）与 LOTA-Bench（Choi 等，2024）则针对语言条件长程任务，以任务成功率作为核心评估指标。

尽管仿真环境存在物理建模不精确、视觉渲染失真等固有局限，其仍为模型高效训练、系统性能验证与现实场景泛化提供了不可替代的实验平台。

4.4.2　真实数据集和基准测试

真实世界数据集通过捕捉动态光照、背景干扰以及物体属性多样性等复杂环境因素，为 VLA 的落地与泛化能力评估提供了关键数据基础。近年来，多个大规模多模态真实机器人数据集的发布，显著促进了语言—动作对齐任务的发展（Dang 等，2025；Luo 等，2025）。

如图 5 所示，在数据集演进方面，早期代表性工作 MIME（Sharma 等，2018）收集了 8.2 万条包含人类演示与机器人演示轨迹的任务数据；RoboNet（Dasari 等，2019）跨机器人平台收录 16.2 万条轨迹，支持跨平台策略泛化研究。为提升任务覆盖度，BC-Z（Jang 等，2021）提供上百种任务的专家演示；RT-1（Brohan 等，2023）及其扩展版本覆盖 700 多项日常活动，累计包含 13 万条轨迹；RT-2（Zitkovich 等，2023）则进一步融合网络规模的视觉—语言数据。

在多模态与跨域泛化方面，BridgeData V2（Ebert 等，2022）与 RH20T（Fang 等，2024）通过大规模语言标注演示支持跨域泛化研究，其中 RH20T 提供 RGB-D、力矩和音频等多模态信号，适用于单次模仿学习场景。DROID（Khazatsky 等，2024）通过标准化 Franka 机器人平台收集数万条遥操作轨迹，确保了环境与硬件的一致性。Open-X Embodiment（O'Neill 等，2024）作为迄今规模最大的跨平台数据集，整合了 22 种机器人、500 余技能的 100 万条多模态演示，采用统一数据格式，极大促进了跨本体泛化研究。此外还要一些专项数据集，FuSe（Jones 等，2025）集成视觉、触觉与听觉模态，提供 2.7 万条机器人轨迹；RoboMIND（Wu 等，2024b）收录 10.7 万条涵盖单臂、双臂、人形及灵巧手等多种具身本体的轨迹；AgiBot World Dataset（Bu 等，2025a）上百台机器人采集百万级轨迹，为极端多样性条件下的 VLA 训练提供支持。此外，Language Table（Lynch 等，2022）提供了 60 万条积木操作轨迹及其众包语言标注；导航领域数据集如 SACSoN（Hirose 等，2024）、BDD100K（Yu 等，2020）等则专注于移动平台的指令遵循与目标导向行为评估。

在基准测试方法上，研究者通常采用任务成功率作为核心指标，部分工作（Black 等，2024；Kim 等，2024）引入指令遵循率以评估指令理解准确性。尽管现有数据集规模显著扩大，开放世界中的长尾物体、场景与技能仍面临表征不足的挑战，亟需更广泛、多样化的真实数据以支撑 VLA 的持续发展。

VLA 系统的性能评估是一个多维度问题，单一的成功率指标难以全面反映模型在实际部署中的综合能力。理想的评估体系应至少包含以下 4 个核心维度：1）任务成功率：衡量模型在给定指令下完成指定任务的比例，是基础性指标；2）执行效率：以控制频率或端到端推理延迟度量，反映模型在实时控制场景中的可用性（表 7）；3）安全性：以安全违规率（如碰撞次数、越界次数、力超限次数等）量化，评估

图5　VLA相关数据集

Fig. 5　Related datasets of VLA

模型在动态环境中规避风险的能力;4)鲁棒性与错误恢复能力:通过首次失败后的恢复成功率、平均恢复时间/步数等指标,衡量模型在面对意外扰动、执行失误时的自我修正能力。

4.4.3　长程任务评估的开放性挑战

尽管评测基准在推动VLA发展方面取得显著成效,但其评估重点仍主要集中在短程、原子性任务,例如单一物体抓取、点对点导航,或虽然涉及多步操作但流程高度固定的任务序列。相较之下,真正能够体现通用具身智能能力的,是开放、长时跨度且具强组合性的复杂任务(Liu等,2023a)。围绕这类长程任务的评估协议目前仍处于起步阶段,在任务成功粒度、鲁棒性度量、组合泛化以及仿真—现实一致性等方面存在一系列挑战。

1)任务成功粒度与部分奖励设计。长程任务通常由数十甚至上百个原子动作组成,单一的成功/失败指标过于粗糙,难以定位智能体在任务执行过程中的具体失效环节。长程任务通常仅有最终的二元奖励,导致评价反馈极度稀疏。这要求评估体系从仅关注最终结果,扩展到对过程与阶段性进展的综合刻画,引入细粒度、过程化的奖励或评分标准。难点在于如何在不过度依赖大规模人工标注的前提下,构建既能合理反映任务进度和子目标完成度又具备可扩展性与客观性的度量机制,以避免评估成本随任务复杂度指数级增长。近期研究如LOTA-Bench(Choi等,2024)尝试引入子目标完成率指标,但其泛化性和自动化程度仍有待提升。

2)鲁棒性度量。在长程任务中,智能体难免会遭遇各种可恢复的失败情形,例如夹持失败导致物体掉落、导航过程中误入死胡同、局部观测被暂时遮挡等。在长程控制中,每一帧动作预测的微小偏差都会随时间推移在物理环境中累积,导致智能体逐渐偏离目标分布。对于具备真实应用价值的系统而言,关键在于能否及时检测异常、进行状态诊断,并通过重规划实现有效恢复。因此,理想的评估协议不应在首次失败时立即终止任务,而应考察系统的错误恢复能力。这要求基准能够支持带有重置/干预机制的交互式动态评估,并设计相应指标,如“在允许多次恢复条件下的最终成功率”、“平均恢复调整时间/步数”等。当前多数基准仍主要采用一次性、刚性成功标准,对这类容错型、可恢复执行过程的衡量明显不足。此外,长程任务中“失败即终止”与“允许恢复”的设置会显著影响成功率、效率与安全统计。因此需明确终止条件、最大步数、恢复窗口以及外部干预类型,并区分“自然恢复”与“干预恢复”,以提高不同工作之间的可比性。

3)组合泛化与指令遵循的精确检验。长程任务的本质难点之一在于组合泛化:智能体需要将已学

习的基础技能在新情境下进行重组与组合，在满足复杂约束的前提下完成新任务。尽管已有部分基准引入了多样化语言指令，但在以下两方面仍显不足：(1)对指令中隐含的复杂约束与条件逻辑缺乏系统化评估；(2)对动态上下文与多备选方案理解与执行能力缺乏针对性测试。未来的长程任务基准需要能够以程序化方式生成并控制指令复杂度，系统性构造层次化目标、多重约束与条件分支，并结合精细的行为日志与语义匹配度量，对 VLA 在语义落地、指令遵循与组合泛化方面的真实能力进行全面检验。LIBERO 的 Long 套件虽然涉及多步任务，但其任务序列仍相对固定，尚未达到真正组合泛化的测试要求。

4)仿真到现实的评估鸿沟。长程任务往往伴随复杂的物理交互过程，例如装配场景中的柔性材料接触、形变、粘附等。这些现象在仿真环境中的建模误差，通常会沿着多步决策过程被逐步放大，导致仿真中表现良好的策略在现实世界中失效。通过校准仿真环境与现实机器人配置，有效减小了感知层面的领域差异，但物理层面的建模误差(如非刚体形变、摩擦系数不确定性、接触动力学等)仍未得到根本解决。因此，构建高保真、物理上更为准确且支持长时间交互的仿真环境，并建立仿真与真实评估结果之间的一致性与相关性分析框架，是长程任务评估中的另一项关键难题。这不仅涉及物理引擎与材质建模的精度提升，还涉及如何在可控成本下采集和标注真实世界长程任务数据，用于校准和修正仿真环境。

综上所述，要推动 VLA 长程任务评估体系的成熟，评估框架一方面需要支持复杂状态追踪、阶段性奖励计算与容错式交互评估，另一方面还应面向组合泛化与仿真—现实迁移提供可扩展的测试工具。未来的长程基准平台应在任务生成、过程监控、结果判定与指标统计等多个层面进行统一设计，使评估体系既能够更准确地刻画 VLA 的真实能力，又能够反向引导模型沿着可恢复、可泛化和可部署的方向持续演进。

5 存在挑战

5.1 实时推理与计算效率挑战

实时推理是 VLA 在机器人操控、自动驾驶等延迟敏感场景应用中的关键制约因素。当前主流 VLA 普遍采用自回归解码范式，逐步生成动作指令，导致推理速度通常仅为 5 Hz 左右，远低于机器人高频控制(>100 Hz)的实际需求(Bjorck 等，2025)。如 OpenVLA(Kim 等，2024)与 pi-0(Black 等，2024)在动态环境中的性能因序列生成的延迟而受限。为提升推理效率，并行解码等加速技术相继提出，例如 NVIDIA GR00T N1(Bjorck 等，2025)通过并行预测多词元实现推理加速，但此类方法常以牺牲轨迹平滑度为代价，使其在手术机器人等高精度场景下的适用性有限。硬件约束进一步加剧这一挑战，处理高维视觉表征需要较高内存带宽，远超 Jetson 等边缘计算平台的能力。即使采用模型量化等技术，在需要高精度操作的任务中仍面临显著的性能损失。计算成本方面，Transformer 架构的计算复杂度随序列长度与输入维度二次方增长，导致其在训练和推理阶段均需庞大的计算资源(Black 等，2024)。多数 VLA 系统需要高额显存，严重制约了其在自主无人机、移动机器人等平台上的部署可行性。最新的工作 UD-VLA(Chen 等，2026)将联合离散去噪扩散过程引入 VLA 框架，不同于以往将视觉预测与动作生成分离的架构，UD-VLA 在同一去噪轨迹中同步优化图像生成与动作预测，实现了理解、生成与执行的深度耦合。通过离散扩散机制大幅压缩了采样步数，使得推理速度达到传统自回归方法的 4 倍以上，为 VLA 在实时控制场景中的部署提供了可行路径。

表 7 展示了不同 VLA 模型的推理效率。表中数据综合整理自多篇原始研究论文中公布的结果，仅对其进行归纳呈现，并保持原文的实验设置不变。

表 7 不同 VLA 模型的推理效率

Table 7 Inference efficiency of different VLA models

模型	执行效率/Hz	时延/s
OpenVLA(Kim 等，2024)	4.2	0.239 6
OpenVLA-OFT(Kim 等，2025b)	71.4	0.112 0
VLA-Adapter(Wang 等，2026)	219.2	0.036 5

5.2 泛化能力与跨场景迁移障碍

泛化能力是决定 VLA 能否真正走向大规模落地的核心因素之一。当前 VLA 在跨场景(跨任务、

跨环境）、跨平台（跨机器人本体）和跨模态等方面的泛化方面仍面临严峻挑战，其根本原因可归纳为数据、表征与模型架构、本体异构性3个层面。

在数据层面，VLA的性能高度依赖于大规模、高质量的多模态演示数据，但现有数据生态在规模、多样性与标注质量等方面均存在瓶颈（Black等，2024）。一方面，网络爬取的视觉—语言数据虽然规模客观，但缺失动作标注；另一方面，通过遥操作收集的机器人演示数据虽然具有较高标注质量，但采集成本高昂且难以持续规模化。现有开源数据集（O'Neill等，2024）在任务类型和环境多样性方面均相对有限，使得模型训练所依赖的数据分布与开放世界中的长尾场景存在显著偏差，从而严重制约了VLA在复杂开放环境中的迁移与泛化能力。

在表征与模型架构层面，动作编码方式直接影响跨平台泛化性能。离散词元化方法在执行精细操作时往往难以达到足够的分辨率，且词表难以通用；连续回归方法（Cheng等，2024）虽然具备更高的表达精度，却容易出现模态坍塌，且对分布外情形的泛化能力较弱；基于扩散的策略能够刻画更丰富的动作分布，但其计算开销通常远高于传统解码器（Ding等，2025b）。在模型层面，尽管基于Transformer的架构展现了强大的容量，但其主要依托训练数据分布内的相关性进行学习，难以显式学习物理层面的因果规律，决策过程更易依赖数据中的统计关联而非深层物理原理。因此，当环境中出现动态内容、目标部分遮挡、复杂背景干扰或指令语义歧义等情况时，模型性能往往出现显著退化。

在跨平台迁移层面，机器人形态的多样性（如机械臂、人形机器人、轮式平台）导致其观测空间（关节状态、相机视角等）和动作空间（自由度、运动范围、动力学特性等）存在本质差异。即使在相同任务和环境条件下，更换机器人本体、传感器配置或控制接口之后，常需要重新收集数据并进行大规模微调，说明当前VLA对机器人形态、关节维度、控制频率和执行器特性仍具有较强依赖。不同机器人在关节配置、动力学特性、传感器类型与布局上的差异，使得为特定平台训练的策略难以直接迁移到于另一平台（Black等，2024），这要求VLA学习一种更为抽象和普适的动作表征。

此外，与LLM/VLM领域中已被反复验证的规模法则相比（Zitkovich等，2023），VLA目前的性能提升主要体现为随数据规模线性增长而非显著的性能涌现。除数据规模限制外，另一个重要原因在于LLM离散符号空间与VLA连续物理空间的本质差异：1）离散符号空间可通过有限词表构成清晰的组合结构，为模型提供了明确的语义边界和较高的容错率，易于通过规模化捕捉统计规律。相比之下，VLA面临高维连续动作空间，缺乏清晰的语义边界，从高维语义意图到低维物理执行的映射对数值精度极度敏感，微小的连续值漂移即可能导致物理交互失败。2）真实世界的物理动作分布往往具有多模态性（Multimodality），即同一指令可对应多条不同的轨迹。常用的行为克隆回归目标（如均方误差）倾向于输出均值动作，难以像离散分类那样形成清晰的决策边界。3）文本数据具有高语义密度，而机器人轨迹数据包含大量冗余的时序帧，且受限于本体异构性，其通用物理知识的提取难度远高于语言知识，模型难以直接通过规模化数据学习到深层的物理因果规律。尽管如此，现有研究已给出规模法则的初步证据，Open X-Embodiment的实验表明，在异构的跨本体数据上小容量模型（如RT-1-X）易出现欠拟合，而更大容量的RT-2-X能够更充分吸收多源数据，通过汇聚多种不同机器人的跨本体数据进行联合训练，在特定任务上的成功率相比单本体训练有大幅提升（O'Neill等，2024）。这说明虽然VLA尚未达到LLM级别的通用推理涌现，但在统一动作表征的前提下，跨本体数据的规模化确实能显著改善泛化边界。

针对上述挑战，提升VLA泛化能力需要在数据与算法两个层面协同推进。一方面，应构建覆盖多平台、多任务和多场景的大规模具身数据集，并在统一动作语义空间下对动作进行语义化、规范化与重参数化，使不同机器人在共享高层策略的同时保留必要的低层控制差异；同时结合域随机化、仿真—真实混合训练等策略拓展训练数据分布，从而支撑在统一框架下的跨平台协同训练与标准化评测。另一方面，可以利用世界模型进行模拟推演，增强模型对物理规律与因果结构的理解。世界模型对VLA泛化能力的贡献存在两条路径：时空先验注入与隐空间想象力辅助。第1条路径通过大规模视频预训练，使模型学习到物体运动的物理规律、场景演化的时序一致性等时空先验知识，从而在面对新场景时能够基于有限的观测进行合理的外推。GR-1（Wu

等，2024a）的消融实验表明，在大规模Ego4D视频数据上进行未来帧预测预训练后，模型在未见场景中的零样本任务成功率大幅提升，这直接证明了时空先验对泛化能力的增益（Grauman等，2024）。第2条路径则利用世界模型在隐空间中构建环境动力学模型，使智能体能够通过多步滚动预测在动作执行前推演不同候选动作的潜在后果，从而选择最优策略。WorldVLA（Cen等，2025）的实验数据显示，具备隐空间推演能力的模型在抓取任务中的成功率较基线方法有明显提升，且在面对未知物体时表现出更强的适应性。世界模型的引入为破解VLA领域的规模法则难题提供了新思路，不同于LLM依赖海量数据直接训练，世界模型通过动力学对齐的隐空间使模型能够从少量真实交互中推断环境的物理参数，从而在数据稀缺条件下实现更好的泛化。此外，隐空间想象力允许模型在虚拟环境中进行探索和学习，避免了对真实世界数据的无限需求。DALI（dynamics-aligned latent imagination）（Röder等，2025）框架的理论分析和实验证明，动力学对齐的隐空间编码器能够从有限的历史交互中恢复出环境的潜在上下文变量，这一能力是实现高效泛化的关键；DeepVerse（Chen等，2025c）进一步扩展到4D空间，通过显式建模几何结构显著减少了长时间预测中的误差累积，使世界模型能够生成更长的、物理一致的未来序列。这些进展表明，世界模型的核心价值在于通过结构化的隐空间建模，以更少的数据实现更强的泛化。

此外，结合元学习、终身学习以及测试时自适应等机制，使模型在严格的安全约束下进行在线适应与持续更新，在遇到新场景、新平台和新任务时，能够通过有限交互实现快速调整，从而在真实复杂环境中维持稳定、可解释的跨场景泛化与跨平台迁移性能。目前，这些方案仍处于探索阶段，尚未形成成熟通用解决方案，可为后续VLA泛化能力的提升指明研究路径。

5.3 环境适应与安全保障难题

VLA在动态、非结构化环境中的适应性与鲁棒性仍显不足。其视觉感知模块在低光照、阴影等复杂场景中准确率显著下降（Wang等，2025d）；语言理解模块在面对模糊或歧义指令时表现不佳；动作执行模块在杂乱环境中对部分遮挡物体的位姿估计常常失效（Zhang等，2025a）。这些缺陷源于模型学习到的多为静态的表征关联，而非环境动态变化的因果关系（Wen等，2025a），导致其难以在未见过的场景中进行有效的推理与决策。如何定义一套既能确保安全又不过度限制任务执行能力的通用规则，并将其有效融入模型决策过程，仍是VLA在可靠部署时的关键问题（Zhang等，2025d；郭园方 等，2025）。当前VLA系统的安全实践大体可划分为被动安全与主动安全两个层级，整体上正处在从前者向后者演进的过渡期。

1）基于阈值与规则的被动安全响应机制。当前大多数VLA系统或其部署框架的安全保障仍主要依赖于被动安全策略，包括预设的力/力矩阈值、关节运动限位以及基于简单规则的碰撞检测。此类方法本质上是一种事后反应机制，其优势在于实现简单、行为可预测且便于形式化验证。然而，其局限性也十分明显，难以及时预测并规避即将发生的危险，在应对意外碰撞、环境突变等复杂情形时适应性不足。同时，现有碰撞预测模型在动态空间中的可靠性仍有限（Zhen等，2024），在仓库物流、家庭服务等安全性要求较高的场景中存在潜在风险。

2）基于预见、评估与规避的主动安全响应机制。主动安全被普遍视为下一代安全VLA的重要发展方向，其核心目标是赋予智能体风险预判、风险评估及自主规避的能力，要求VLA能够对自身及他者的短期未来动作进行推演，并在此基础上做出将安全优先于任务效率的决策。实现这一目标，一方面需要构建具备风险预见能力的世界模型，例如将动态障碍物轨迹预测模块集成至VLA的视觉编码或决策闭环中，使系统能够提前数秒预演潜在碰撞风险；另一方面，在动作生成层面需将安全约束纳入优化目标，引入代表风险的代价函数，结合安全约束强化学习，鼓励模型生成低风险的动作序列。与传统被动安全的急停策略不同，理想的主动安全机制应支持平滑的中断与恢复：当预测到高风险时，系统可自主切换至保守的安全运行模式，并在风险解除后自动恢复任务，而非完全依赖人工干预（蹇木伟 等，2026）。已有研究尝试将主动安全理念融入VLA，例如SafeVLA（Zhang等，2025a）提出集成式安全方案，主动收集多种不安全行为样本，在此基础上通过安全约束的强化学习对VLA策略进行训练，并配合针对性的安全评测进行验证。实验结果表明，该类方法可以在基本不牺牲任务成功率的前提下，显著降

低长程移动与操作任务中的安全违规次数。

具身智能的安全性不仅取决于控制算法的稳定性，还面临对抗鲁棒性的严峻挑战，当前的威胁主要体现在推理阶段的对抗攻击与训练阶段的后门植入两个维度。1）物理一致性对抗攻击：与传统计算机视觉中的像素级扰动不同，针对VLA的攻击需考虑机器人的物理动力学与时序依赖性。最新的研究（Wang等，2025d）设计了针对VLA的物理对抗补丁（adversarial patch），仅需在相机视野内放置一个小型的彩色补丁，即可诱导机器人动作失控。为突破VLA的时序防御，攻击者提出了动作差异目标以最大化轨迹偏离，以及几何感知目标以操纵机器人在三维空间中的运动方向。实验表明，此类攻击在模拟与真实环境中均导致任务成功率大幅下降，揭示了VLA对空间几何与时序依赖的脆弱性；2）“训练即服务（training-as-a-service）”下的后门攻击（backdoor attacks）：随着“训练即服务”模式的兴起，外包训练引入了隐蔽的后门风险。BadVLA（Zhou等，2025a）提出了基于目标解耦优化的攻击范式，首先通过特征空间分离隔离触发器表征，然后冻结感知模块仅微调动作头以植入条件控制偏差。这种攻击方式能在保持正常任务性能的同时，实现近100%的后门攻击成功率，且对输入扰动与模型微调具有极强的鲁棒性，证明了在长程复杂任务中植入持久性后门的可行性。

5.4 评估体系与伦理问题

现有评估体系的局限性导致了对VLA能力的片面认知。当前主流基准大多集中于实验室或高度结构化的仿真环境，且任务类型单一。这种理想化的评估环境无法有效衡量模型在开放世界中的泛化性、鲁棒性与安全性，使得在基准测试中表现优异的模型，一旦部署于真实工业、户外或家庭环境中，性能往往会大幅衰减。

除评估体系本身的局限性外，VLA的快速发展还引发了深刻的伦理关切，由于模型缺乏真正的因果推理能力（Wen等，2025a），其在开放环境中的行为可能超出设计者的预期，引发关于责任归属、信任和可解释性的难题。此外，数据偏见问题尤为突出。由于VLA的训练数据（包括互联网图文数据、人类操作视频以及遥操作演示等）内嵌了人类社会已有的刻板印象，这些偏见可能被模型继承并在物理交互中放大，导致不公平甚至有害的决策行为。例如，当指令涉及职业相关任务时，模型可能表现出性别刻板印象，对于“让助手打扫厨房”的指令，VLA倾向于关联为女性形象，而对于“搬运重物”的指令，则倾向于关联为男性形象。这种偏见可能影响机器人在人机协作中对人类意图的准确理解。针对上述问题，1）在数据层面，需要构建公平性感知的数据集，通过再平衡采样、反事实数据增强等技术，确保不同群体在训练数据中的均衡代表性，发展去偏见的数据标注规范，在数据收集阶段即避免强化刻板印象；2）在算法层面，引入公平性约束学习，在训练目标中加入群体间性能差异的正则化项，通过生成反事实样本（如将指令中的性别代词互换），检验模型决策是否依赖于敏感属性，并在训练中强化决策的一致性；3）在系统层面，建立偏见检测与审计工具，在模型部署前进行系统性偏见测试，生成偏见报告；4）在评估层面，发展公平性基准测试，用于评估VLA在不同人口统计群体上的性能差异、指令理解的一致性以及任务分配的无偏性等指标，应覆盖多维度公平性度量，并建立统一的评分体系。长远来看，如何确保VLA系统的可控性、透明度与价值对齐，是决定其能否被广泛接纳的核心议题。

6 应用场景与未来展望

6.1 应用场景

随着VLA技术的逐步成熟，其应用范围正从实验室环境逐步拓展至多样化的真实场景，并在多个领域展现出广阔的应用前景（Kawaharazuka等，2025）。

1）人形机器人平台。近期研究表明（Chen等，2025a），VLA已能驱动人形机器人进行高频全身控制，实时协调手臂、手部、躯干及精细的手指动作。采用双系统架构，慢系统负责指令与环境的语义解析及任务分解；快系统以数百赫兹频率输出密集控制指令，从而在未知任务和动态环境下实现具有一定泛化能力的操作，而无需针对每个具体任务单独训练。例如，在家庭等非结构化场景中，人形机器人可完成抓取、放置、倒水和折叠等多种日常家务操作（Sapkota等，2025）。

2）自动驾驶系统。VLA为突破传统模块化自动驾驶系统提供了新的技术路径。传统方案通常将感知、预测与规划分模块建模，而基于VLA的端到端

决策系统则尝试在统一的自回归模型中融合多视角视觉数据、车载传感器信息与自然语言指令，直接输出如转向角、加速度等控制量。以 OpenDriveVLA（Zhou 等，2026）为例，通过构建大规模实车采集数据集，将驾驶视频、激光雷达、里程计、自然语言描述与高精度轨迹进行对齐，用于训练具备视觉问答与轨迹规划能力的驾驶 VLA 模型。该类系统能够在开放道路场景中对交通标志、行人行为及语音导航指令进行综合理解，在复杂城市路况下生成安全且具备可解释性的驾驶策略。

3）工业制造机器人。工业场景中的 VLA 系统通过联合编码视觉输入（如工件布局、输送带状态）、自然语言指令与机器人自身状态，使机器人能够依据场景语义自主选择操作策略。近期提出的架构（Li 等，2024）采用动作扩散解码器，在保证动作生成精度与鲁棒性的同时，显著提升对未知工具及未知工件布局的泛化能力。与传统基于演示或刚性编程的工业自动化方案相比，基于 VLA 的系统能够理解非结构化的自然语言指令，并具备极强的少样本学习能力。这使得此类系统仅通过语言提示和少量动作演示，即可实现任务的快速重配置，有望在物流分拣、非标零件装配以及人机协作等潜在方向进行验证。

4）医疗护理与手术机器人。在医疗场景中，VLA 有望提升手术机器人与护理机器人的自主性与协作能力（Li 等，2025c）。在手术机器人方向，VLA 系统通过融合腔镜视频、解剖结构信息与语言指令，将高层次语义任务映射为具有亚毫米级精度的多步运动序列，从而实现对器械位置、力反馈及关键组织结构的动态适应，减轻术者的操作负担。在护理与康复场景中，具备 VLA 能力的辅助机器人能够感知患者行为，理解语音或手势指令，并执行取药及安全监测等日常任务。

5）农业机器人。VLA 模型正逐渐成为构建智能农机与农用无人系统的重要技术手段。相较于依赖固定规则与预定义脚本的传统农业自动化系统，农业 VLA 通过统一建模多模态感知与动作生成，使地面机器人与无人机能够理解自然语言形式的农业指令，并据此生成具备情境感知能力的操作策略（Sautenkov 等，2025）。结合 RGB-D、多光谱等多种传感器，VLA 可完成作物长势监测、病虫害识别、精准施药及选择性采摘等任务，并通过合成数据与跨季节训练机制提升其泛化能力，以满足可持续农业与智慧农场的发展需求（Sapkota 等，2025）。

6.2 未来展望

本文系统性地梳理了基于 VLA 的具身智能研究进展。当前 VLA 技术已在语义理解、指令跟随与基础操控任务中取得显著进步，但本质上仍局限于一种被动的、基于静态数据关联的学习范式。为实现向通用具身智能的跨越，未来研究需在以下关键方向取得突破。

1）从被动感知走向主动推理。当前 VLA 多数仍停留在感知—动作直接映射层面，缺乏对物理世界底层因果机制的显式建模，模型在面对长尾分布、组合性任务或动态干扰时，往往表现为简单的模式匹配，泛化性能显著下降。从可行性来看，世界模型、模型预测控制、视频预测与状态空间模型以及因果学习等方向已积累了丰富的基础工作（Feng 等，2025），为在 VLA 框架内引入主动推理提供了现成的工具与理论支撑。

一种可行路径是构建分层推理框架，即在现有 VLA 的编码器—解码器结构中嵌入可学习的世界模型子系统，将环境状态与动力学编码为结构化的状态表征，使智能体在执行物理动作前，能够在隐空间中对候选动作序列进行多步滚动预测与评估，再由 VLA 生成高层子目标或约束，交由底层控制器完成细化执行（Cen 等，2025）。与此同时，需要考虑将因果发现与主动干预引入 VLA，从海量多模态交互数据中抽取结构化因果图，识别关键环境变量之间的因果关系，并通过设计探索性动作来验证假设、降低不确定性，提高学习效率。此外，可将基于模型的强化学习与先验知识相结合，利用世界模型在仿真环境或数字孪生中进行大规模规划与探索，再通过策略蒸馏等方式进行知识迁移，以缓解现实世界探索成本高、风险大的问题。

在这一方向上仍存在多项关键技术瓶颈亟需突破：其一，如何构建可扩展且计算高效的世界模型，既能处理高维视觉输入，又能在统一表示中兼顾语义、几何与动力学信息；其二，如何从非结构化、多模态交互数据中可靠学习可迁移的因果表征，避免噪声与偏置诱导伪因果关系；其三，如何在主动干预和探索过程中严格保障安全性与可控性，在实时性约束下实现稳定可靠的多步推理与决策。

2）实现持续学习与终身适应。现有 VLA 系统

在训练完成后通常冻结模型参数，对新任务、新场景的适应能力有限（Achiam等，2023）。赋予VLA持续在线学习与终身自适应能力，使其在长期运行中不断吸收关于新物体、新任务和新环境的知识，同时避免灾难性遗忘，是其走向开放世界的关键一步。这一目标要求模型结构、参数组织、记忆机制与学习算法之间进行协同设计。从可行性看，终身学习、增量学习以及领域自适应等方向已经提出了多种缓解遗忘的机制，为VLA的持续学习提供了技术参考。

在实现路径上，一方面可以发展参数高效与模块化的持续学习机制：通过为不同任务或场景分配独立或部分共享的适配器模块、稀疏激活子网络或可扩展子结构，在不显著增加整体模型体积的前提下容纳增量知识。层次化VLA架构在这一点上具有优势，如只对技能库、规划头或特定动作解码模块进行局部更新。另一方面，应为VLA设计情景记忆与经验回放机制，将关键示范轨迹、失败案例和代表性环境状态存储在外部或内部记忆中，通过周期性回放与对比学习巩固旧知识。另外，还可以建立基于人类反馈与自监督信号的在线学习机制，通过自然语言纠正、偏好排序或任务成败信号进行轻量级在线微调；在缺乏人工标注时，则借助下一帧预测、动作一致性约束等自监督信号推动领域自适应。

持续学习方向同样面临若干重要难题：如何在有限的边缘计算与存储资源下实现高效、稳定的在线更新；如何设计通用且可扩展的记忆表征与检索机制，兼顾时间跨度和查询效率；以及如何在实际部署中严格约束在线学习的安全边界，防止模型因错误反馈、对抗样本或异常事件而发生不可逆的性能退化。

3）深化应用场景与系统集成。VLA的最终价值体现在解决复杂现实问题的能力，其发展路径必然要从实验室中单一形态、单一任务的机器人系统，走向多模态、多平台和多智能体协同的复杂系统工程。这一方向在智能驾驶、高端制造、医疗与健康以及精准农业等领域已初步探索（Sapkota等，2025），但在系统集成与工程落地层面仍存在明显差距。从可行性看，现有控制理论、同步定位建图等系统已经相对成熟，为VLA的集成嵌入提供了重要基础。

在具体路径上，可以优先推进跨平台通用技能的移植与复用，开发具备强语义理解与规划能力的VLA大脑，再通过轻量级控制器或本体适配层映射到不同机器人体型与执行器上，借助大规模仿真和跨本体数据预训练通用技能表征实现多平台部署。其次，需要发展支持人机混合智能与多智能体协作的VLA，使其不仅能被动执行人类指令，还能理解人类意图、预测人类行动，与人类及其他机器人进行任务协商、角色分配和动态调整。此外，在自动驾驶、高端制造等场景中，VLA须与经典感知、建图、任务调度和安全控制模块深度集成，重点在于设计稳定可靠的系统接口和故障安全边界，使学习型组件与确定性组件在可验证的框架下协同工作。

这一方向的技术瓶颈主要包括：如何在异构硬件平台之间实现低延迟、高保真的技能迁移与策略共享；如何构建高效的人机互信与沟通机制，使VLA在协作中既不过度自主又不过度依赖人为干预；以及如何对集成了VLA的复杂系统进行形式化验证与安全保障，在确保整体系统满足工业级安全和可靠性要求的前提下，逐步提升系统的智能化水平。

4）构建可信赖的负责任VLA生态系统。随着VLA在关键领域的应用不断推进，伴随出现可解释性不足、安全风险难以完全预见以及伦理边界模糊等问题。VLA能否被社会广泛接纳，不仅取决于技术指标，更取决于能否建立贯穿研发、评估与部署全流程的可信赖生态系统。从可行性看，形式化验证、安全强化学习、价值对齐以及可解释机器学习等方向已有丰富成果，为构建多层次的安全与问责机制提供了可用工具。

在实现路径上，一方面需要提升可解释性与决策溯源能力，发展能够为关键行为生成自然语言解释和多模态可视化归因的技术，并建立标准化的决策日志系统，实现从输入、内部表征到输出动作的全链路记录；另一方面，需要系统研究VLA面对分布外输入、对抗样本和恶意指令时的鲁棒性与安全对齐问题，研发内嵌安全护栏，将强化学习中的安全奖励设计与控制理论中的安全约束、形式化规约有机结合。同时，应推动建立标准化评估体系与开源生态，构建覆盖从基础技能到长尾复杂任务、从仿真到真实世界的分层基准。

在人机协作场景下，VLA系统的引入使得责任归属问题愈发复杂，这涉及算法开发者（如模型训练者）、系统集成商、部署机构（如医院、工厂）和最终用户之间的复杂链条。由于VLA系统具有高度的端到端黑盒特性，当发生安全事故时，难以通过简单的

规则回溯判定是数据偏见、模型推理失误还是本体机械故障。从技术角度，应建立基于可解释路径存证的体系，通过记录模型内部的注意力热力图与预测的中间状态（如预见未来帧），为事故回溯提供证据链；从伦理法律角度，需要建立分级责任框架，明确不同参与方的注意义务和免责条件。例如，若事故源于训练数据中的偏见或未预见的场景，开发者可能需要承担部分责任；若因用户未按规范操作导致，则责任可能归于用户。此外，还需考虑人机信任的动态演化。在长期协作中，人类对机器人的信任会基于其行为表现而调整。设计能够主动沟通自身意图和置信度的 VLA 系统，有助于维持健康的信任关系。例如，VLA 可在执行高风险动作前以自然语言向人类确认，或在感知不确定时请求帮助，从而实现责任共担。

在上述过程中仍然存在多重难点：如何在不显著牺牲性能的前提下提升模型的内在可解释性；如何在开放、动态场景下形式化定义并验证安全性与合规性指标；以及如何在全球范围内建立跨学科、跨行业的伦理与标准共识，协调不同领域之间在安全要求和监管力度上的差异。这些问题很大程度上决定了 VLA 从实验室原型走向大规模社会部署的边界与进程，也需要技术社区、产业界与政策制定者的长期协同与迭代推进。

参考文献（References）

Abdin M, Aneja J, Behl H, Bubeck S, Eldan R, Gunasekar S, et al. 2024. Phi-4 Technical Report [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2412.08905.pdf

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F L, et al. 2023. GPT-4 technical report [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2303.08774.pdf

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]. [2025-11-17]. 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

Asif U, Tang J B and Harrer S. 2018. GraspNet: an efficient convolutional neural network for real-time grasp detection for low-powered devices//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: ijcai.org: 4875-4882 [DOI: 10.24963/ijcai.2018/677]

Bai J Z, Bai S, Chu Y F, Cui Z Y, Dang K, Deng X D, et al. 2023. Qwen technical report [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2309.16609.pdf

Belkhale S, Ding T L, Xiao T, Sermanet P, Vuong Q, Tompson J, et al. 2024. RT-H: action hierarchies using language//Robotics: Science and Systems XX. Delft, the Netherlands: MIT Press [DOI: 10.15607/RSS.2024.XX.049]

Beyer L, Steiner A, Pinto A S, Kolesnikov A, Wang X, Salz D, et al. 2024. PaliGemma: a versatile 3B VLM for transfer [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2407.07726.pdf

Bi J X, Ma K Y, Hao C, Shou M Z and Soh H. 2025. VLA-touch: enhancing vision-language-action models with dual-level tactile feedback [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2507.17294.pdf

Bjorck J, Castañeda F, Cherniadev N, Da X Y, Ding R Y, Fan L X, et al. 2025. GR00T N1: an open foundation model for generalist humanoid robots [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2503.14734.pdf

Black K, Brown N, Darpinian J, Dhabalia K, Driess D, Esmail A, et al. 2025. π0.5: a vision-language-action model with open-world generalization [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2504.16054.pdf

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]. [2025-11-17]. https://arxiv.org/pdf/2410.24164.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//Proceedings of Robotics: Science and Systems XIX. Daegu, Korea (South): MIT Press [DOI: 10.15607/RSS.2023.XIX.025]

Bruce J, Dennis M D, Edwards A, Parker-Holder J, Shi Y G, Hughes E, et al. 2024. Genie: generative interactive environments//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR: 4603-4623

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]. [2025-11-17]. https://arxiv.org/pdf/2503.06669.pdf

Bu Q W, Li H Y, Chen L, Cai J S, Zeng J, Cui H M, et al. 2024. Towards synergistic, generalized, and efficient dual-system for robotic manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2410.08001.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]. [2025-11-17]. https://arxiv.org/pdf/2505.06111.podf

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]. [2025-11-17]. https://arxiv.org/pdf/2506.21539.pdf

Chameleon Team. 2025. Chameleon: mixed-modal early-fusion foundation models[EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2405.09818.pdf

Cheang C, Chen S J, Cui Z R, Hu Y D, Huang L Q, Kong T, et al. 2025. GR-3 technical report [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2507.15493.pdf

Cheang C L, Chen G Z, Jing Y, Kong T, Li H, Li Y F, et al. 2024. GR-2: a generative video-language-action model with web-scale knowledge for robot manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2410.06158.pdf

Chen H, Liu J M, Gu C Y, Liu Z Y, Zhang R R, Li X Q, et al. 2025a. Fast-in-slow: a dual-system foundation model unifying fast manipulation within slow reasoning [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2506.01953.pdf

Chen H P, Li S F, Fan J M, Duan A Q, Yang C G, Navarro-Alarcon D, et al. 2025b. Human-in-the-loop robot learning for smart manufacturing: a human-centric perspective. IEEE Transactions on Automation Science and Engineering, 22: 11062-11086 [DOI: 10.1109/TASE.2025.3528051]

Chen J Y, Song W X, Ding P X, Zhou Z Y, Zhao H, Tang F L, et al. 2026. Unified diffusion VLA: vision-language-action model via joint discrete denoising diffusion process//Proceedings of the 14th International Conference on Learning Representations. Rio de Janeiro, Brazil: OpenReview.net

Chen J Y, Zhu H Y, He X L, Wang Y F, Zhou J J, Chang W Z, et al. 2025c. DeepVerse: 4D autoregressive video generation as a world model [EB/OL]. [2025-06-11]. https://doi.org/10.48550/arXiv.2506.01103

Chen X Y, Wei H X, Zhang P S, Zhang C H, Wang K X, Guo Y J, et al. 2025d. villa-X: enhancing latent action modeling in vision-language-action models [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2507.23682.pdf

Chen Y, Ge Y Y, Li Y Z, Ge Y X, Ding M Y, Shan Y, et al. 2024. Moto: latent motion token as the bridging language for robot manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2412.04445.pdf

Chen Y H, Tian S, Liu S G, Zhou Y T, Li H R and Zhao D B. 2025e. ConRFT: a reinforced fine-tuning method for VLA models via consistency policy//Proceedings of Robotics: Science and Systems. Los Angeles, USA: RSS Foundation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/XXI.019.pdf

Chen Z J, Niu R L, Kong H and Wang Q. 2025f. TGRPO: fine-tuning vision-language-action model via trajectory-wise group relative policy optimization [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2506.08440.pdf

Cheng A C, Ji Y D, Yang Z J, Gongye Z T, Zou X Y, Kautz J, et al. 2024. NaVILA: legged robot vision-language-action model for navigation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2412.04453.pdf

Cherepanov E, Kachaev N, Kovalev A K and Panov A I. 2025. Memory, benchmark and robots: a benchmark for solving complex tasks with reinforcement learning [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2502.10550.pdf

Cherti M, Beaumont R, Wightman R, Wortsman M, Ilharco G, Gordon C, et al. 2023. Reproducible scaling laws for contrastive language-image learning//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 2818-2829 [DOI: 10.1109/CVPR52729.2023.00276]

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]

Choi J W, Yoon Y, Ong H, Kim J and Jang M. 2024. LoTa-Bench: benchmarking language-oriented task planners for embodied agents//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

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]. [2025-11-17]. https://arxiv.org/pdf/2505.03912.pdf

Damen D, Doughty H, Farinella G M, Furnari A, Kazakos E, Ma J, et al. 2022. Rescaling egocentric vision: collection, pipeline and challenges for EPIC-KITCHENS-100. International Journal of Computer Vision, 130(1): 33-55 [DOI: 10.1007/s11263-021-01531-2]

Dang R H, Yuan Y Q, Zhang W Q, Xin Y F, Zhang B Q, Li L, et al. 2025. ECBench: can multi-modal foundation models understand the egocentric world? A holistic embodied cognition benchmark//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 24593-24602 [DOI: 10.1109/CVPR52734.2025.02290]

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

Davisson L D. 1972. Rate-distortion theory and application. Proceedings of the IEEE, 60(7): 800-808 [DOI: 10.1109/proc.1972.8779]

Deng S L, Yan M, Wei S L, Ma H X, Yang Y X, Chen J Y, et al. 2025. GraspVLA: a grasping foundation model pre-trained on billion-scale synthetic action data [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2505.03233.pdf

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, Volume 1 (Long and Short Papers). Minneapolis, USA: ACL: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Dey S, Zaech J N, Nikolov N, Van Gool L and Paudel D P. 2025. ReVLA: reverting visual domain limitation of robotic foundation models//Proceedings of 2025 IEEE International Conference on Robotics and Automation (ICRA). Atlanta, USA: IEEE: 8679-8686 [DOI: 10.1109/ICRA55743.2025.11128823]

Ding P X, Ma J F, Tong X Y, Zou B H, Luo X X, Fan Y G, et al. 2025a. Humanoid-VLA: towards universal humanoid control with visual integration [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2502.14795.pdf

Ding P X, Zhao H, Zhang W J, Song W X, Zhang M, Huang S T, et al. 2025b. QUAR-VLA: vision-language-action model for quadruped robots//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 352-367 [DOI: 10.1007/978-3-031-72652-1_21]

Doshi R, Walke H R, Mees O, Dasari S and Levine S. 2024. Scaling cross-embodied learning: one policy for manipulation, navigation, locomotion and aviation//Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR: 496-512

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. [s.l.]: OpenReview.net.

Driess D, Xia F, Sajjadi M S M, 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: PMLR: 8469-8488

Ebert F, Yang Y L, Schmeckpeper K, Bucher B, Georgakis G, Daniilidis K, et al. 2022. Bridge data: boosting generalization of robotic skills with cross-domain datasets//Proceedings of Robotics: Science and Systems XVIII. New York, USA: MIT Press [DOI: 10.15607/RSS.2022.XVIII.063]

Ehsani K, Gupta T, Hendrix R, Salvador J, Weihs L, Zeng K H, et al. 2024. SPOC: imitating shortest paths in simulation enables effective navigation and manipulation in the real world//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 16238-16250 [DOI: 10.1109/CVPR52733.2024.01537]

Esser P, Rombach R and Ommer B. 2021. Taming transformers for high-resolution image synthesis//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12868-12878 [DOI: 10.1109/CVPR46437.2021.01268]

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

Feng M T, Shen J H, Wu Z J, Peng W X, Zhong H, Guo Y L, et al. 2025. Advancements in 3D vision understanding using multimodal large language models. Journal of Image and Graphics, 30(6): 1744-1791 (冯明涛, 沈军豪, 武子杰, 彭伟星, 钟杭, 郭裕兰, 等. 2025. 多模态大模型驱动的三维视觉理解技术前沿进展. 中国图象图形学报, 30(6): 1744-1791) [DOI: 10.11834/jig.240588]

Feng T, Wang W G and Yang Y. 2025. A survey of world models for autonomous driving [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2501.11260.pdf

Firoozi R, Tucker J, Tian S, Majumdar A, Sun J K, Liu W Y, et al. 2025. Foundation models in robotics: applications, challenges, and the future. The International Journal of Robotics Research, 44(5): 701-739 [DOI: 10.1177/02783649241281508]

Gao C K, Liu Z X, Chi Z H, Huang J S, Fei X, Hou Y W, et al. 2025. VLA-OS: structuring and dissecting planning representations and paradigms in vision-language-action models [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2506.17561.pdf

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

Gong J Y, Lou Y J, Liu F Q, Zhang Z W, Chen H M, Zhang Z Z, et al. 2023. Scene point cloud understanding and reconstruction technologies in 3D space. Journal of Image and Graphics, 28(6): 1741-1766 (龚靖渝, 楼雨京, 柳奉奇, 张志伟, 陈豪明, 张志忠, 等. 2023. 三维场景点云理解与重建技术. 中国图象图形学报, 28(6): 1741-1766) [DOI: 10.11834/jig.230004]

Goyal R, Kahou S E, 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, Cartillier V, Chavis Z, Furnari A, et al. 2025. Ego4D: around the world in 3 600 hours of egocentric video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(11): 9468-9509 [DOI: 10.1109/TPAMI.2024.3381075]

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 (CVPR). Seattle, USA: IEEE: 19383-19400 [DOI: 10.1109/CVPR52733.2024.01834]

Guo Y F, Yu Z T, Liu A S, Zhou W B, Qiao T, Li B, et al. 2025. Recent progress of the security research for multimodal large models. Journal of Image and Graphics, 30(6): 2051-2081 (郭园方, 余梓彤, 刘艾杉, 周文柏, 乔通, 李斌, 等. 2025. 多模态大模型安全研究进展. 中国图象图形学报, 30(6): 2051-2081) [DOI: 10.11834/jig.250067]

Guo Y J, Zhang J K, Chen X Y, Ji X, Wang Y J, Hu Y C, et al. 2025. Improving vision-language-action model with online reinforcement

learning [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2501.16664.pdf

Han B O, Kim J and Jang J. 2024. A dual process VLA: efficient robotic manipulation leveraging VLM [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2410.15549.pdf

Hao P, Zhang C F, Li D Z, Cao X G, Hao X S, Cui S W, et al. 2025. TLA: tactile-language-action model for contact-rich manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2503.08548.pdf

He K M, Chen X L, Xie S N, Li Y H, Dollár P and Girshick R. 2022. Masked autoencoders are scalable vision learners//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 15979-15988 [DOI: 10.1109/CVPR52688.2022.01553]

He K M, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2980-2988 [DOI: 10.1109/ICCV.2017.322]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hirose N, Shah D, Sridhar A and Levine S. 2024. SACSoN: scalable autonomous control for social navigation. IEEE Robotics and Automation Letters, 9(1): 49-56 [DOI: 10.1109/LRA.2023.3329626]

Hong Y N, Zhen H Y, Chen P H, Zheng S H, Du Y L, Chen Z F, et al. 2023. 3D-LLM: injecting the 3D world into large language models//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #900

Hu E J, Shen Y L, Wallis P, Allen-Zhu Z, Li Y Z, Wang S A, et al. 2022. LoRA: low-rank adaptation of large language models//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: OpenReview.net

Hu Y D, Lin F Q, Zhang T, Yi L and Gao Y. 2023. Look before you leap: unveiling the power of GPT-4V in robotic vision-language planning [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2311.17842.pdf

Huang C P, Wu Y H, Chen M H, Wang Y C F and Yang F E. 2025a. ThinkAct: vision-language-action reasoning via reinforced visual latent planning [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2507.16815.pdf

Huang J L, Wang S, Lin F Q, Hu Y H, Wen C and Gao Y. 2025b. Tactile-VLA: unlocking vision-language-action model′s physical knowledge for tactile generalization [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2507.09160.pdf

Huang J Y, Yong S L, Ma X J, Linghu X K, Li P H, Wang Y, et al. 2024a. An embodied generalist agent in 3D world//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR: 20413-20451

Huang S Y, Jiang Z K, Dong H, Qiao Y, Gao P and Li H S. 2023b. Instruct2Act: mapping multi-modality instructions to robotic actions with large language model [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2305.11176.pdf

Huang W L, Wang C, Li Y Z, Zhang R H and Li F F. 2024b. ReKep: spatio-temporal reasoning of relational keypoint constraints for robotic manipulation//Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR: 4573-4602

Huang W L, Wang C, Zhang R H, Li Y Z, Wu J J and Li F F. 2023c. VoxPoser: composable 3D value maps for robotic manipulation with language models//Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR: 540-562

Huang W L, Xia F, Xiao T, Chan H, Liang J, Florence P, et al. 2022. Inner monologue: embodied reasoning through planning with language models//Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR: 1769-1782

Hung C Y, Sun Q, Hong P F, Zadeh A, Li C, Tan U X, et al. 2025. NORA: a small open-sourced generalist vision language action model for embodied tasks [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2504.19854.pdf

Hurst A, Lerer A, Goucher A P, Perelman A, Ramesh A, Clark A, et al. 2024. GPT-4o system card [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2410.21276.pdf

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. 2021. BC-Z: zero-shot task generalization with robotic imitation learning//Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR: 991-1002

Jatavallabhula K M, Iyer G and Paull L. 2019. gradSLAM: dense SLAM meets automatic differentiation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/1910.10672.pdf

Ji Y H, Tan H J, Shi J Y, Hao X S, Zhang Y, Zhang H Y, et al. 2025. RoboBrain: a unified brain model for robotic manipulation from abstract to concrete//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 1724-1734 [DOI: 10.1109/CVPR52734.2025.00168]

Jia X G, Wang Q, Donat A, Xing B W, Li G, Zhou H Y, et al. 2024. MaIL: improving imitation learning with selective state space models//Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR: 3888-3907

Jian M W, Ling Y K, Zhang H R, Zhang L S and Ma J J. 2026. Embod-

ied intelligence-driven distracted driving detection: a framework and research prospects[J/OL]. Journal of Image and Graphics, 1-16. https://doi.org/10.11834/jig.250514 (蹇木伟, 凌钰坤, 张昊然, 张琳松, 马嘉骏. 2026. 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, 1-16). [2026-5-29] https://doi.org/10.11834/jig.250514

Jiang J P, Xiao W Y, Lin Z Y, Zhang H Z, Ren T X, Gao Y, et al. 2025a. SOLAMI: social vision-language-action modeling for immersive interaction with 3D autonomous characters//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 26887-26898 [DOI: 10.1109/CVPR52734.2025.02504]

Jiang S, Fang X L, Roy N, Lozano-Pérez T, Kaelbling L P and Ancha S. 2025b. Streaming flow policy: simplifying diffusion/flow-matching policies by treating action trajectories as flow trajectories//Proceedings of the 9th Conference on Robot Learning. Seoul, Korea (South): PMLR: 238-257

Jiang Y F, Gupta A, Zhang Z C, Wang G Z, Dou Y Q, Chen Y J, et al. 2022. VIMA: general robot manipulation with multimodal prompts [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2210.03094.pdf

Jiang Z Y, Xie Y Q, Lin K, Xu Z J, Wan W K, Mandlekar A, et al. 2025c. 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. Atlanta, USA: IEEE: 5961-5968 [DOI: 10.1109/ICRA55743.2025.11127987]

Jülg T, Burgard W and Walter F. 2025. Refined policy distillation: from VLA generalists to RL experts [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2503.05833.pdf

Kawaharazuka K, Oh J, Yamada J, Posner I and Zhu Y K. 2025. Vision-language-action models for robotics: a review towards real-world applications. IEEE Access, 13: 162467-162504 [DOI: 10.1109/ACCESS.2025.3609980]

Kerbl B, Kopanas G, Leimkuehler T and Drettakis G. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4): #139 [DOI: 10.1145/3592433]

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 XX. Delft, the Netherlands: MIT Press [DOI: 10.15607/RSS.2024.XX.120]

Kim H, Kang J, Kang H, Cho M, Kim S J and Lee Y. 2025a. UniSkill: imitating human videos via cross-embodiment skill representations [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2505.08787.pdf

Kim M J, Finn C and Liang P. 2025b. Fine-tuning vision-language-action models: optimizing speed and success [EB/OL]. [2025-02-27]. 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

Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, et al. 2023. Segment anything//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 3992-4003 [DOI: 10.1109/ICCV51070.2023.00371]

Lee J, Duan J F, Fang H Q, Deng Y Q, Liu S, Li B Y, et al. 2025. MolmoAct: action reasoning models that can reason in space [EB/OL]. [2025-10-21]. https://arxiv.org/pdf/2508.07917.pdf

Li C M, Wen J J, Peng Y X, Peng Y and Zhu Y C. 2026. PointVLA: injecting the 3D world into vision-language-action models. IEEE Robotics and Automation Letters, 11(3): 2506-2513 [DOI: 10.1109/LRA.2026.3653303]

Li C S, Zhang R H, Wong J, Gokmen C, Srivastava S, Martín-Martín R, et al. 2022a. BEHAVIOR-1K: a human-centered, embodied AI benchmark with 1 000 everyday activities and realistic simulation // Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR: 80-93

Li H R, Chen Y H, Cui W B, Liu W H, Liu K, Zhou M C, et al. 2026. Survey of vision-language-action models for embodied manipulation. Acta Automatica Sinica, 52(1): 18-51 (李浩然, 陈宇辉, 崔文博, 刘卫恒, 刘锴, 周明才, 等. 2026. 面向具身操作的视觉-语言-动作模型综述. 自动化学报, 52(1): 18-51) [DOI: 10.16383/j.aas.c250394]

Li J N, Li D X, Savarese S and Hoi S C H. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR: 19730-19742

Li Q X, Liang Y B, Wang Z Y, Luo L, Chen X, Liao M Z, et al. 2024a. CogACT: a foundational vision-language-action model for synergizing cognition and action in robotic manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2411.19650.pdf

Li S, Puig X, Paxton C, Du Y L, Wang C, Fan L X, et al. 2022b. Pre-trained language models for interactive decision-making//Advances in Neural Information Processing Systems 35. New Orleans, USA: Neural Information Processing Systems Foundation: 31199-31212

Li S L, Gao L S, Cao J W and Hu Y B. 2025a. Graph-fused vision-language-action for policy reasoning in multi-arm robotic manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2509.07957.pdf

Li S L, Wang J, Dai R, Ma W Y, Ng W Y, Hu Y B, et al. 2025c. RoboNurse-VLA: robotic scrub nurse system based on vision-language-action model//Proceedings of 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hang-

zhou, China: IEEE: 3986-3993 [DOI: 10.1109/IROS60139.2025.11246030]

Li X H, Liu M H, Zhang H B, Yu C J, Xu J, Wu H T, et al. 2024b. Vision-language foundation models as effective robot imitators//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Li X L, Hsu K, Gu J Y, Mees O, Pertsch K, Walke H R, et al. 2024c. Evaluating real-world robot manipulation policies in simulation//Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR: 3705-3728

Li Y, Deng Y Q, Zhang J, Jang J, Memmel M, Garrett C R, et al. 2025d. HAMSTER: hierarchical action models for open-world robot manipulation//Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net

Li Y B, Gong Z Y, Li H Y, Huang X Q, Kang H L, Bai G P, et al. 2025e. Robotic visual instruction//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12155-12165 [DOI: 10.1109/CVPR52734.2025.01135]

Li Y L, Yan G, Macaluso A, Ji M Z Y, Zou X Y and Wang X L. 2025f. Integrating LMM planners and 3D skill policies for generalizable manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2501.18733.pdf

Li W, Zhang R S, Shao R, He J and Nie L Q. 2025g. CogVLA: cognition-aligned vision-language-action model via instruction-driven routing & sparsification//Advances in Neural Information Processing Systems 38. DiegoSan, USA: Neural Information Processing Systems Foundation [EB/OL]. [2025-11-17]. https://openreview.net/forum?id=Fg9HufTI0K

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. London, UK: IEEE: 9493-9500 [DOI: 10.1109/ICRA48891.2023.10160591]

Liang Z X, Mu Y, Ma H B, Tomizuka M, Ding M Y and Luo P. 2024. SkillDiffuser: interpretable hierarchical planning via skill abstractions in diffusion-based task execution//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 16467-16476 [DOI: 10.1109/CVPR52733.2024.01558]

Lin F Q, Nai R Q, Hu Y D, You J C, Zhao J M and Gao Y. 2025a. OneTwoVLA: a unified vision-language-action model with adaptive reasoning [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2505.11917.pdf

Lin J Y, Taherin A, Akbari A, Akbari A, Lu L, Chen G Y, et al. 2025b. VOTE: vision-language-action optimization with trajectory ensemble voting [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2507.05116.pdf

Lin T, Li G, Zhong Y L, Zou Y W, Du Y X, Liu J T, et al. 2025c. Evo-0: vision-language-action model with implicit spatial understanding [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2507.00416.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 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, Chen H, An P J, Liu Z Y, Zhang R R, Gu C Y, et al. 2025a. HybridVLA: collaborative diffusion and autoregression in a unified vision-language-action model [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2503.10631.pdf

Liu J M, Liu M Z, Wang Z Y, An P J, Li X Q, Zhou K C, et al. 2024a. 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 S L, Zeng Z Y, Ren T H, Li F, Zhang H, Yang J, et al. 2025b. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 38-55 [DOI: 10.1007/978-3-031-72970-6_3]

Liu S M, Wu L X, Li B G, Tan H K, Chen H Y, Wang Z Y, et al. 2025c. RDT-1B: a diffusion foundation model for bimanual manipulation//Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net

Liu S Y, Du J W, Xiang S C, Wang Z B and Luo D S. 2024b. ReLEP: a novel framework for real-world long-horizon embodied planning [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2409.15658.pdf

Liu Z, Mao H Z, Wu C Y, Feichtenhofer C, Darrell T and Xie S N. 2022. A ConvNet for the 2020s//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11966-11976 [DOI: 10.1109/CVPR52688.2022.01167]

Liu Z Y, Gu Y C, Zheng S X, Fu Y W, Xue X Y and Jiang Y G. 2025d. TriVLA: a triple-system-based unified vision-language-action model with episodic world modeling for general robot control [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2507.01424.pdf

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]. [2025-11-17]. https://arxiv.org/pdf/2505.18719.pdf

Lu J S, Clark C, Lee S, Zhang Z C, Khosla S, Marten R, et al. 2024. Unified-IO 2: scaling autoregressive multimodal models with vision, language, audio, and action//Proceedings of 2024 IEEE/

CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 26429-26445 [DOI: 10.1109/CVPR52733.2024.02497]

Luo H, Feng Y C, Zhang W P, Zheng S P, Wang Y, Yuan H Q, et al. 2025a. Being-H0: vision-language-action pretraining from large-scale human videos [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2507.15597.pdf

Lynch C, Wahid A, Tompson J, Ding T L, Betker J, Baruch R, et al. 2022. Interactive language: talking to robots in real time [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2210.06407.pdf

Ma Y E, Song Z X, Zhuang Y Z, Hao J Y and King I. 2024. A survey on vision-language-action models for embodied AI [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2405.14093.pdf

Ma Y J, Sodhani S, Jayaraman D, Bastani O, Kumar V and Zhang A. 2023. VIP: towards universal visual reward and representation via value-implicit pre-training//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net

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.: 1-10

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, Xu D F, Wong J, Nasiriany S, Wang C, Kulkarni R, et al. 2021. What matters in learning from offline human demonstrations for robot manipulation//Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR: 1678-1690

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. Zürich, Switzerland: PMLR: 879-893

Mao W X, Zhong W H, Jiang Z, Fang D, Zhang Z Y, Lan Z H, et al. 2024. RoboMatrix: a skill-centric hierarchical framework for scalable robot task planning and execution in open-world [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2412.00171.pdf

Mees O, Hermann L, Rosete-Beas E and Burgard W. 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]

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2022. NeRF: representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1): 99-106 [DOI: 10.1145/3503250]

Monache S D, Moleman R and Özcan E. 2024. Soundstorm, a collaborative ideation game for sound-driven design//Proceedings of the 19th International Audio Mostly Conference: Explorations in Sonic Cultures. Milan, Italy: ACM: 479-486 [DOI: 10.1145/3678299.3678348]

Mu Y, Zhang Q L, Hu M K, Wang W H, Ding M Y, Jin J, et al. 2023. EmbodiedGPT: vision-language pre-training via embodied chain of thought//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1090

Nair S, Rajeswaran A, Kumar V, Finn C and Gupta A. 2022. R3M: a universal visual representation for robot manipulation//Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR: 892-909

Nasiriany S, Maddukuri A, Zhang L C, Parikh A, Lo A, Joshi A, et al. 2024. RoboCasa: large-scale simulation of everyday tasks for generalist robots [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2406.02523.pdf

Nematollahi I, DeMoss B, Chandra A L, Hawes N, Burgard W and Posner I. 2025. LUMOS: language-conditioned imitation learning with world models//Proceedings of 2025 IEEE International Conference on Robotics and Automation. Atlanta, USA: IEEE: 8219-8225 [DOI: 10.1109/ICRA55743.2025.11127988]

Niu D T, Sharma Y, Biamby G, Quenum J, Bai Y T, Shi B F, et al. 2024. LLARVA: vision-action instruction tuning enhances robot learning//Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR: 3333-3355

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

Oquab M, Darcet T, Moutakanni T, Vo H V, Szafraniec M, Khalidov V, et al. 2024. DINOv2: learning robust visual features without supervision. Transactions on Machine Learning Research, 2024: 1-17

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L, Mishkin P, et al. 2022. Training language models to follow instructions with human feedback//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #2011

Patratskiy M A, Kovalev A K and Panov A I. 2025. Spatial traces: enhancing VLA models with spatial-temporal understanding [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2508.09032.pdf

Pertsch K, Stachowicz K, Ichter B, Driess D, Nair S, Vuong Q, et al. 2025. FAST: efficient action tokenization for vision-language-action models [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2501.09747.pdf

Pumacay W, Singh I, Duan J, Krishna R, Thomason J and Fox D. 2024. The COLOSSEUM: a benchmark for evaluating generalization for robotic manipulation//Proceedings of Robotics: Science and

Systems XX. Delft, the Netherlands: MIT Press [DOI: 10.15607/RSS.2024.XX.133]

Qi C R, Su H, Mo K C and Guibas L J. 2017a. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 77-85 [DOI: 10.1109/CVPR.2017.16]

Qi C R, Yi L, Su H and Guibas L J. 2017b. PointNet++: deep hierarchical feature learning on point sets in a metric space//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 5105-5114

Qi Z K, Dong R P, Zhang S C, Geng H R, Han C R, Ge Z, et al. 2025a. ShapeLLM: universal 3D object understanding for embodied interaction//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 214-238 [DOI: 10.1007/978-3-031-72775-7_13]

Qi Z K, Zhang W Y, Ding Y F, Dong R P, Yu X Q, Li J W, et al. 2025b. SOFAR: language-grounded orientation bridges spatial reasoning and object manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2502.13143.pdf

Qu D L, Song H M, Chen Q Z, Yao Y Q, Ye X Y, Ding Y, et al. 2025. SpatialVLA: exploring spatial representations for visual-language-action model [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2501.15830v1.pdf

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

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1): #140

Ranasinghe K, Li X, Mata C, Park J and Ryoo M S. 2025. Pixel motion as universal representation for robot control [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2505.07817.pdf

Rashid A, Sharma S, Kim C M, Kerr J, Chen L Y, Kanazawa A, et al. 2023. Language embedded radiance fields for zero-shot task-oriented grasping//Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR: 178-200

Reed S E, Zolna K, Parisotto E, Gómez Colmenarejo S, Novikov A, Barth-Maron G, et al. 2022. A generalist agent. Transactions on Machine Learning Research, 2022

Reimers N and Gurevych I. 2019. Sentence-BERT: sentence embeddings using siamese BERT-networks//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: ACL: 3982-3992 [DOI: 10.18653/v1/D19-1410]

Reuss M, Yagmurlu Ö E, Wenzel F and Lioutikov R. 2024. Multimodal diffusion transformer: learning versatile behavior from multimodal goals//Proceedings of Robotics: Science and Systems XX. Delft, the Netherlands: MIT Press [DOI: 10.15607/RSS.2024.XX.121]

Röder F, Benad J, Eppe M and Banerjee P K. 2025. Dynamics-aligned latent imagination in contextual world models for zero-shot generalization//Proceedings of the 39th Annual Conference on Neural Information Processing Systems. San Diego, USA: Curran Associates, Inc.

Ryoo M S, Piergiovanni A J, Arnab A, Dehghani M and Angelova A. 2021. TokenLearner: adaptive space-time tokenization for videos//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.: #979

Sameni S, Kafle K, Tan H and Jenni S. 2024. Building vision-language models on solid foundations with masked distillation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14216-14226 [DOI: 10.1109/CVPR52733.2024.01348]

Sanh V, Debut L, Chaumond J and Wolf T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/1910.01108v1.pdf

Sapkota R, Cao Y, Roumeliotis K I and Karkee M. 2025. Vision-language-action models: concepts, progress, applications and challenges [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2505.04769v1.pdf

Sautenkov O, Yaqoot Y, Lykov A, Mustafa M A, Tadevosyan G, Akhmetkazy A, et al. 2025. UAV-VLA: vision-language-action system for large scale aerial mission generation//Proceedings of the 20th ACM/IEEE International Conference on Human-Robot Interaction. Melbourne, Australia: IEEE: 1588-1592 [DOI: 10.1109/HRI61500.2025.10974117]

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]

Schulman J, Wolski F, Dhariwal P, Radford A and Klimov O. 2017. Proximal policy optimization algorithms [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/1707.06347.pdf

Shafiullah N M M, Rai A, Etukuru H, Liu Y Q, Misra I, Chintala S, et al. 2023. On bringing robots home [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2311.16098.pdf

Shao R, Li W, Zhang L S, Zhang R S, Liu Z Y, Chen R, et al. 2025. Large VLM-based vision-language-action models for robotic manipulation: a survey [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2508.13073.pdf

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 Annual Conference on Robot Learning. Zürich, Switzerland: PMLR: 906-915

Sharshar A, Khan L U, Ullah W and Guizani M. 2025. Vision-language

models for edge networks: a comprehensive survey. IEEE Internet of Things Journal, 12(16): 32701-32724 [DOI: 10.1109/JIOT.2025.3579032]

Shentu Y D, Wu P, Rajeswaran A and Abbeel P. 2024. From LLMs to actions: latent codes as bridges in hierarchical robot control//Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems. Abu Dhabi, United Arab Emirates: IEEE: 8539-8546 [DOI: 10.1109/IROS58592.2024.10801683]

Shi L X, Ichter B, Equi M R, Ke L Y M, Pertsch K, Vuong Q, et al. 2025. Hi robot: open-ended instruction following with hierarchical vision-language-action models//Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR: 54919-54933

Shorinwa O, Tucker J, Smith A, Swann A, Chen T, Firoozi R, et al. 2024. Splat-MOVER: multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting//Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR: 4748-4770

Shridhar M, Manuelli L and Fox D. 2021. CLIPort: what and where pathways for robotic manipulation//Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR: 894-906

Shridhar M, Thomason J, Gordon D, Bisk Y, Han W, Mottaghi R, et al. 2020. ALFRED: a benchmark for interpreting grounded instructions for everyday tasks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10737-10746 [DOI: 10.1109/CVPR42600.2020.01075]

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]. [2025-11-17]. https://arxiv.org/pdf/2506.01844.pdf

Singh H, Das R J, Han M, Nakov P and Laptev I. 2024. MALMM: multi-agent large language models for zero-shot robotics manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2411.17636.pdf

Singh I, Blukis V, Mousavian A, Goyal A, Xu D F, Tremblay J, et al. 2023. ProgPrompt: generating situated robot task plans using large language models//Proceedings of 2023 IEEE International Conference on Robotics and Automation. London, UK: IEEE: 11523-11530 [DOI: 10.1109/ICRA48891.2023.10161317]

Song C H, Sadler B M, Wu J M, Chao W L, Washington C and Su Y. 2023. LLM-planner: few-shot grounded planning for embodied agents with large language models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 2986-2997 [DOI: 10.1109/ICCV51070.2023.00280]

Song J M, Meng C L and Ermon S. 2021. Denoising diffusion implicit models//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Song W X, Chen J Y, Ding P X, Zhao H, Zhao W, Zhong Z D, et al. 2025a. Accelerating vision-language-action model integrated with action chunking via parallel decoding [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2503.02310v1.pdf

Song W X, Chen J Y, Li W X, He X, Zhao H, Cui C, et al. 2025b. RationalVLA: a rational vision-language-action model with dual system [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2506.10826.pdf

Song Z R, Ouyang G X, Li M Z, Ji Y H, Wang C X, Xu Z X, et al. 2026. ManipLVM-R1: reinforcement learning for reasoning in embodied manipulation with large vision-language models//Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore, Singapore: AAAI Press: 18558-18566 [DOI: 10.1609/aaai.v40i22.38922]

Sun Q, Fang Y X, Wu L, Wang X L and Cao Y. 2023. EVA-CLIP: improved training techniques for CLIP at scale [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2303.15389.pdf

Tan M X and Le Q V. 2019. EfficientNet: rethinking model scaling for convolutional neural networks//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 6105-6114

Tan S H, Dou K R, Zhao Y and Krähenbühl P. 2025a. Interactive post-training for vision-language-action models [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2505.17016.pdf

Tan X D, Yang Y X, Ye P, Zheng J L, Bai B Z, Wang X Y, et al. 2025b. Think twice, act once: token-aware compression and action reuse for efficient inference in vision-language-action models [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2505.21200.pdf

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, et al. 2023. LLaMA: open and efficient foundation language models [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2302.13971.pdf

van den Oord A, Vinyals O and Kavukcuoglu K. 2017. Neural discrete representation learning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6309-6318

Van Vo T, Nguyen T Q, Nguyen K M, Nguyen D H M and Vu M N. 2025. ReFineVLA: reasoning-aware teacher-guided transfer fine-tuning [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2505.19080.pdf

Wang C, Xia F, Yu W H, Zhang T N, Zhang R H, Liu C K, et al. 2025a. Chain-of-modality: learning manipulation programs from multimodal human videos with vision-language-models//Proceedings of 2025 IEEE International Conference on Robotics and Automation (ICRA). Atlanta, USA: IEEE: 6527-6535 [DOI: 10.1109/ICRA55743.2025.11128270]

Wang H Y, Xiong C Y, Wang R P and Chen X L. 2025b. BitVLA: 1-bit vision-language-action models for robotics manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2506.07530.pdf

Wang S S, Yu R C, Yuan Z H, Yu C, Gao F, Wang Y, et al. 2025c. Spec-VLA: speculative decoding for vision-language-action models with relaxed acceptance//Proceedings of 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou,

China: ACL: 26928-26940 [DOI: 10.18653/v1/2025.emnlp-main.1367]

Wang T W, Han C, Liang J C, Yang W H, Liu D F, Zhang L X, et al. 2025d. Exploring the adversarial vulnerabilities of vision-language-action models in robotics [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2411.13587.pdf

Wang W G, Yang Y and Pan Y H. 2025e. Visual knowledge in the big model era: retrospect and prospect. Frontiers of Information Technology and Electronic Engineering, 26(1): 1-19 [DOI: 10.1631/FITEE.2400250]

Wang Y H, Ding P X, Li L X, Cui C, Ge Z R, Tong X Y, et al. 2026. VLA-adapter: an effective paradigm for tiny-scale vision-language-action model//Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore, Singapore: AAAI Press: 18638-18646 [DOI: 10.1609/aaai.v40i22.38931]

Wang Y Q, Li X H, Wang W X, Zhang J B, Li Y Y, Chen Y T, et al. 2025f. Unified vision-language-action model [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2506.19850.pdf

Wang Y T, Zhu H Y, Liu M Y, Yang J G, Fang H S and He T. 2025g. VQ-VLA: improving vision-language-action models via scaling vector-quantized action tokenizers//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 11089-11099

Wei J L, Yuan S S, Li P F, Hu Q D, Gan Z X and Ding W C. 2024. OccLLaMA: an occupancy-language-action generative world model for autonomous driving [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2409.03272.pdf

Wen C, Lin X Y, So J, Chen K, Dou Q, Gao Y, et al. 2024a. Any-point trajectory modeling for policy learning//Proceedings of Robotics: Science and Systems XX. Delft, the Netherlands: MIT Press [DOI: 10.15607/RSS.2024.XX.092]

Wen J J, Zhu M J, Zhu Y C, Tang Z B, Li J M, Zhou Z Y, et al. 2025a. Diffusion-VLA: generalizable and interpretable robot foundation model via self-generated reasoning [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2412.03293.pdf

Wen J J, Zhu Y C, Li J M, Tang Z B, Shen C M and Feng F F. 2025b. DexVLA: vision-language model with plug-in diffusion expert for general robot control [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2502.05855.pdf

Wen J J, Zhu Y C, Li J M, Zhu M J, Wu K, Xu Z Y, et al. 2024b. TinyVLA: towards fast, data-efficient vision-language-action models for robotic manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2409.12514.pdf

Wu D, Fan J X, Zang J Z, Wang G B, Yin W, Li W H, et al. 2025a. Reinforced reasoning for embodied planning [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2505.22050.pdf

Wu H J, Zhang H, Zhang P S, Wang J and Wang C. 2025b. HiBerNAC: hierarchical brain-emulated robotic neural agent collective for disentangling complex manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2506.08296.pdf

Wu H T, Jing Y, Cheang C, Chen G Z, Xu J F, Li X H, et al. 2024a. Unleashing large-scale video generative pre-training for visual robot manipulation//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Wu K, Hou C K, Liu J M, Che Z P, Ju X Z, Yang Z Q, et al. 2024b. RoboMIND: benchmark on multi-embodiment intelligence normative data for robot manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2412.13877.pdf

Wu Z Y, Zhou Y H, Xu X W, Wang Z W and Yan H B. 2025c. MoManipVLA: transferring vision-language-action models for general mobile manipulation//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1714-1723 [DOI: 10.1109/CVPR52734.2025.00167]

Xia F, Shen W B, Li C S, Kasimbeg P, Tchapmi M E, Toshev A, et al. 2020. Interactive gibson benchmark: a benchmark for interactive navigation in cluttered environments. IEEE Robotics and Automation Letters, 5(2): 713-720 [DOI: 10.1109/LRA.2020.2965078]

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 C, Li Q Y, Luo J L and Levine S. 2024a. RLDG: robotic generalist policy distillation via reinforcement learning [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2412.09858.pdf

Xu R T, Zhang J, Guo M H, Wen Y P, Yang H T, Lin M, et al. 2025. A0: an affordance-aware hierarchical model for general robotic manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2504.12636.pdf

Xu Z Y, Wu K, Wen J J, Li J M, Liu N, Che Z P, et al. 2024b. A survey on robotics with foundation models: toward embodied AI [EB/OL]. [2024-02-05]. https://arxiv.org/pdf/2402.02385.pdf

Yan F, Liu F F, Zheng L M, Zhong Y F, Huang Y Y, Guan Z C, et al. 2024. RoboMM: all-in-one multimodal large model for robotic manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2412.07215v1.pdf

Yang J, Glossop C, Bhorkar A, Shah D, Vuong Q, Finn C, et al. 2024a. Pushing the limits of cross-embodiment learning for manipulation and navigation//Proceedings of Robotics: Science and Systems XX. Delft, the Netherlands: MIT Press [DOI: 10.15607/rss.2024.xx.093]

Yang L H, Kang B Y, Huang Z L, Xu X G, Feng J S and Zhao H S. 2024b. Depth anything: unleashing the power of large-scale unlabeled data//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10371-10381 [DOI: 10.1109/CVPR52733.2024.00987]

Yang R H, Yu Q X, Wu Y C, Yan R, Li B R, Cheng A C, et al. 2025a. EgoVLA: learning vision-language-action models from ego-

centric human videos［EB/OL］.［2025-11-17］. https://arxiv.org/pdf/2507.12440.pdf

Yang Y, Sun J X, Kou S Q, Wang Y H and Deng Z J. 2025b. LoHoVLA: a unified vision-language-action model for long-horizon embodied tasks［EB/OL］.［2025-11-17］. https://arxiv.org/pdf/2506.00411.pdf

Yang Y, Zhuang Y T and Pan Y H. 2022. The review of visual knowledge: a new pivot for cross-media intelligence evolution. Journal of Image and Graphics, 27(9): 2574-2588 (杨易, 庄越挺, 潘云鹤. 2022. 视觉知识: 跨媒体智能进化的新支点. 中国图象图形学报, 27(9): 2574-2588)［DOI: 10.11834/jig.211264］

Yang Z, Chen Y, Wang J K, Manivasagam S, Ma W C, Yang A J, et al. 2023. UniSim: a neural closed-loop sensor simulator//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1389-1399［DOI: 10.1109/CVPR52729.2023.00140］

Ye S, Jang J, Jeon B, Joo S J, Yang J W, Peng B L, et al. 2025. Latent action pretraining from videos//Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net

Yu F, Chen H F, Wang X, Xian W Q, Chen Y Y, Liu F C, et al. 2020. BDD100K: a diverse driving dataset for heterogeneous multitask learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2633-2642［DOI: 10.1109/CVPR42600.2020.00271］

Yu J W, Liu H R, Yu Q J, Ren J J, Hao C, Ding H T, et al. 2025. ForceVLA: enhancing VLA models with a force-aware MoE for contact-rich manipulation［EB/OL］.［2025-11-17］. https://arxiv.org/pdf/2505.22159.pdf

Yu S, Sohn K, Kim S and Shin J. 2023. Video probabilistic diffusion models in projected latent space//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 18456-18466［DOI: 10.1109/CVPR52729.2023.01770］

Yu T H, Quillen D, He Z P, Julian R, Hausman K, Finn C, et al. 2019. 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

Yuan C B, Wen C, Zhang T and Gao Y. 2024a. General flow as foundation affordance for scalable robot learning//Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR: 1541-1566

Yuan W T, Duan J F, Blukis V, Pumacay W, Krishna R, Murali A, et al. 2024b. RoboPoint: a vision-language model for spatial affordance prediction in robotics//Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR: 4005-4020

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 language 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. 2024. Robotic control via embodied chain-of-thought reasoning//Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR: 3157-3181

Ze Y J, Zhang G, Zhang K N, Hu C Y, Wang M H and Xu H Z. 2024. 3D diffusion policy: generalizable visuomotor policy learning via simple 3D representations//Proceedings of Robotics: Science and Systems XX. Delft, the Netherlands: MIT Press［DOI: 10.15607/RSS.2024.XX.067］

Zhai X H, Mustafa B, Kolesnikov A and Beyer L. 2023. Sigmoid loss for language image pre-training//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 11941-11952［DOI: 10.1109/ICCV51070.2023.01100］

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］.［2025-11-17］. https://arxiv.org/pdf/2503.03480v2.pdf

Zhang C F, Hao P, Cao X G, Hao X S, Cui S W and Wang S. 2025b. VTLA: vision-tactile-language-action model with preference learning for insertion manipulation［EB/OL］.［2025-11-17］. https://arxiv.org/pdf/2505.09577.pdf

Zhang D P, Sun J, Hu C H, Wu X Y, Yuan Z L, Zhou R, et al. 2025c. Pure vision language action (VLA) models: a comprehensive survey［EB/OL］.［2025-11-17］. https://arxiv.org/pdf/2509.19012.pdf

Zhang F and Gienger M. 2024. Affordance-based Robot Manipulation with Flow Matching［EB/OL］.［2025-10-29］. https://arxiv.org/pdf/2505.21851.pdf

Zhang H C, Yu H N, Zhao L, Choi A, Bai Q X, Yang B, et al. 2025d. SLIM: sim-to-real legged instructive manipulation via long-horizon visuomotor learning［EB/OL］.［2025-11-17］. https://arxiv.org/pdf/2501.09905.pdf

Zhang H Y, Zhuang Z F, Zhao H, Ding P X, Lu H C and Wang D L. 2025e. ReinboT: amplifying robot visual-language manipulation with reinforcement learning//Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR: 77254-77271

Zhang J H, Chen Y R, Xu Y M, Huang Z, Zhou Y P, Yuan Y J, et al. 2025f. 4D-VLA: spatiotemporal vision-language-action pretraining with cross-scene calibration［EB/OL］.［2025-11-17］. https://arxiv.org/pdf/2506.22242.pdf

Zhang J H, Luo Y S, Anwar A, Sontakke S A, Lim J J, Thomason J, et al. 2025g. ReWiND: language-guided rewards teach robot policies without new demonstrations［EB/OL］.［2025-11-17］. https://arxiv.org/pdf/2505.10911.pdf

Zhang J K, Guo Y J, Chen X Y, Wang Y J, Hu Y C, Shi C M, et al. 2024a. HiRT: enhancing robotic control with hierarchical robot transformers//Proceedings of the 8th Conference on Robot Learning.

Munich, Germany: PMLR: 933-946

Zhang J K, Guo Y J, Hu Y C, Chen X Y, Zhu X and Chen J Y. 2025h. UP-VLA: a unified understanding and prediction model for embodied agent//Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR: 74911-74922

Zhang J Y, Xu W Q, Yu Z J, Xie P F, Tang T T and Lu C W. 2025i. DexTOG: learning task-oriented dexterous grasp with language condition. IEEE Robotics and Automation Letters, 10(2): 995-1002 [DOI: 10.1109/LRA.2024.3518116]

Zhang J Z, Wang K Y, Wang S A, Li M H, Liu H R, Wei S L, et al. 2024b. Uni-NaVid: a video-based vision-language-action model for unifying embodied navigation tasks [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2412.06224v1.pdf

Zhang K D, Xu R T, Ren P Z, Lin J F, Wu H F, Lin L, et al. 2025j. RoBridge: a hierarchical architecture bridging cognition and execution for general robotic manipulation [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2505.01709.pdf

Zhang R Y, Dong M H, Zhang Y, Heng L, Chi X W, Dai G L, et al. 2026. MoLe-VLA: dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation//Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore, Singapore: AAAI Press: 18764-18772 [DOI: 10.1609/aaai.v40i22.38945]

Zhang S D, Xu Z, Liu P J, Yu X P, Li Y, Gao Q H, et al. 2024c. VLABench: a large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2412.18194.pdf

Zhang W Q, Wang M N, Liu G G, Xu H X, Jiang Y W, Shen Y L, et al. 2025k. Embodied-reasoner: synergizing visual search, reasoning, and action for embodied interactive tasks [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2503.21696.pdf

Zhang X, Zhang D, Li S M, Zhou Y Q and Qiu X P. 2024d. SpeechTokenizer: unified speech tokenizer for speech language models//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Zhang Y J, Zhang R Q, Zhou H J, Qi J, Yu Z F and Huang T J. 2025. Research status and development trends of vision foundation models. Journal of Image and Graphics, 30(1): 1-24 (张燚钧, 张润清, 周华健, 齐骥, 余肇飞, 黄铁军. 2025. 视觉基础模型研究现状与发展趋势. 中国图象图形学报, 30(1): 1-24) [DOI: 10.11834/jig.230911]

Zhao Q Q, Lu Y, Kim M J, Fu Z P, Zhang Z Y, Wu Y C, et al. 2025a. CoT-VLA: visual chain-of-thought reasoning for vision-language-action models//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1702-1713 [DOI: 10.1109/CVPR52734.2025.00166]

Zhao T Z, Kumar V, Levine S and Finn C. 2023. Learning fine-grained bimanual manipulation with low-cost hardware//Proceedings of Robotics: Science and Systems XIX. Daegu, Korea (South): MIT Press [DOI: 10.15607/RSS.2023.XIX.016]

Zhao W, Ding P X, Zhang M, Gong Z F, Bai S H, Zhao H, et al. 2025b. VLAS: vision-language-action model with speech instructions for customized robot manipulation//Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net

Zhao W, Li G S, Gong Z F, Ding P X, Zhao H and Wang D L. 2025c. Unveiling the potential of vision-language-action models with open-ended multimodal instructions [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2505.11214.pdf

Zhen H Y, Qiu X W, Chen P H, Yang J C, Yan X, Du Y L, et al. 2024. 3D-VLA: a 3D vision-language-action generative world model//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR: 61229-61245

Zheng J L, Li J X, Liu D X, Zheng Y N, Wang Z H, Ou Z H, et al. 2025a. Universal actions for enhanced embodied foundation models//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 22508-22519 [DOI: 10.1109/CVPR52734.2025.02096]

Zheng R J, Liang Y Y, Huang S Y, Gao J F, Daumé III H, Kolobov A, et al. 2025b. TraceVLA: visual trace prompting enhances spatial-temporal awareness for generalist robotic policies//Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net

Zhi P Y, Zhang Z Y, Zhao Y, Han M Z, Zhang Z Y, Li Z T, et al. 2025. Closed-loop open-vocabulary mobile manipulation with GPT-4V//Proceedings of 2025 IEEE International Conference on Robotics and Automation. Atlanta, USA: IEEE: 4761-4767 [DOI: 10.1109/ICRA55743.2025.11127975]

Zhong Y F, Bai F S, Cai S F, Huang X C, Chen Z, Zhang X W, et al. 2025a. A survey on vision-language-action models: an action tokenization perspective [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2507.01925.pdf

Zhong Z D, Yan H D, Li J F, Liu X C, Gong X, Song W X, et al. 2025b. FlowVLA: thinking in motion with a visual chain of thought [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2508.18269.pdf

Zhou J S, Wang J S, Ma B R, Liu Y S, Huang T J and Wang X L. 2024. Uni3D: exploring unified 3D representation at scale//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Zhou X C, Han X Y, Yang F, Ma Y P, Tresp V and Knoll A. 2026. OpenDriveVLA: towards end-to-end autonomous driving with large vision language action model//Proceedings of 2026 AAAI Conference on Artificial Intelligence. Singapore, Singapore: AAAI Press: 13782-13790 [DOI: 10.1609/aaai.v40i16.38386]

Zhou X Y, Tie G Y, Zhang G W, Wang H C, Zhou P and Sun L C. 2025a. BadVLA: towards backdoor attacks on vision-language-action models via objective-decoupled optimization//Proceedings of the 39th Annual Conference on Neural Information Processing Sys-

tems. San Diego, USA: Curran Associates, Inc.

Zhou Z Y, Zhu Y C, Wen J J, Shen C M and Xu Y. 2025b. Vision-language-action model with open-world embodied reasoning from pretrained knowledge [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2505.21906v1.pdf

Zhou Z Y, Zhu Y C, Zhu M J, Wen J J, Liu N, Xu Z Y, et al. 2025c. 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 and Martín-Martín R. 2020. Robosuite: a modular simulation framework and benchmark for robot learning [EB/OL]. [2025-11-17]. https://arxiv.org/pdf/2009.12293v1.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

刘锐，男，博士研究生，主要研究方向为具身智能和多模态大模型。E-mail: rui.liu@zju.edu.cn

杨易，通信作者，男，教授，主要研究方向为人工智能和跨媒体研究。E-mail: yangyics@zju.edu.cn

王文冠，男，研究员，主要研究方向为具身智能和人工智能赋能科学研究。E-mail: wenguanwang.ai@gmail.com

王俊，男，博士，主要研究方向为隐私计算。
E-mail: junwang.lu@gmail.com