平行智能范式视角下的视觉-语言-动作模型发展现状与展望

李柏[1]，郝金第[2]，孙跃硕[2]，孟雨晴[1]，黄峻[3]，田永林[4]，贺正冰[5]

（1. 华东师范大学通信与电子工程学院，上海 200241；
2. 湖南大学机械与运载工程学院，湖南 长沙 410082；
3. 澳门科技大学创新工程学院，澳门 999078；
4. 中国科学院自动化研究所多模态人工智能系统全国重点实验室，北京 100190；
5. 麻省理工学院信息与决策系统实验室，剑桥 02139-4307）

摘　要：视觉-语言-动作模型是一类面向具身智能的综合性建模方法，它将视觉感知、自然语言理解和动作执行在统一框架下进行表征与学习，旨在实现从环境感知到任务规划再到动作控制的连续闭环。视觉-语言-动作模型的运行逻辑与21世纪初提出的平行智能范式存在呼应。平行智能通过“人工系统、计算实验和平行执行”的三元架构，强调虚拟建模、可复现推演以及虚实交互的闭环机制，这些理念与视觉-语言-动作模型的发展路径在不同阶段形成了对应关系：早期的多模态探索可视为人工系统中的原型实践，随后的大规模模型与跨域训练扩展了计算实验的能力，近年来的分层控制和虚实闭环则体现了平行执行强调的反馈修正与规范指导。在这一框架下，视觉-语言-动作模型呈现出语义与动作深度耦合、虚拟与现实双向循环和可验证性增强等特征，但也存在泛化不足、语义对齐不稳、安全与解释机制薄弱以及部署效率受限等问题。未来研究可聚焦任务语义的契约化表达、长时序规划的可修复性、世界模型的工程化使用、多层次反馈与安全治理，以及跨平台迁移和人机协作等方向。以平行智能为参照重新审视视觉-语言-动作模型，不仅有助于厘清发展脉络，也为其在真实场景中的可信应用提供了方法论支持。

关键词：平行智能；视觉-语言-动作模型；具身智能；多模态融合；虚实交互

中图分类号：TP391.4

文献标志码：A

doi: 10.11959/j.issn.2096-6652.202536

Vision-language-action models under parallel intelligence paradigm: the state of the art and future perspectives

LI Bai[1], HAO Jindi[2], SUN Yueshuo[2], MENG Yuqing[1], HUANG Jun[3], TIAN Yonglin[4], HE Zhengbing[5]

1. School of Communication and Electronic Engineering, East China Normal University, Shanghai 200241, China
2. College of Mechanical and Vehicle Engineering, Hunan University, Changsha 410082, China
3. Faculty of Innovation Engineering, Macau University of Science and Technology, Macao 999078, China
4. State Key Laboratory for Multi-modal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
5. Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge 02139-4307, USA

Abstract: Vision-language-action (VLA) models are a comprehensive modeling approach for embodied intelligence that integrates visual perception, natural language understanding, and action execution within a unified framework, aiming to establish a continuous loop from environmental perception to task planning and action control. Their operational logic cor-

收稿日期：2025-08-02；**修回日期：**2025-09-10

通信作者：李柏，libai@zju.edu.cn

基金项目：澳门特别行政区科学技术发展基金项目（No.0157/2024/RIA2，No.0093/2023/RIA2，No.0145/2023/RIA3）；国家自然科学基金项目（No.62103139）；湖南省芙蓉计划湖湘青年英才项目（No.2023RC3115）

Foundation Items: Science and Technology Development Fund, Macao Special Administrative Region (No. 0157/2024/RIA2, No. 0093/2023/RIA2, No. 0145/2023/RIA3), National Natural Science Foundation of China (No. 62103139), Hibiscus Mutabilis Youth Talent Program of Hunan Province (No.2023RC3115)

responds closely to the paradigm of parallel intelligence articulated in the early 21st century. That paradigm comprises artificial systems, computational experiments, and parallel execution, emphasizing virtual modeling, reproducible inference, and closed-loop interaction between the virtual and the real. The initial stage of VLA development, driven by multimodal deep learning, can be regarded as prototypical work within artificial systems, the subsequent stage characterized by large-scale models and cross-domain training expanded the scope of computational experiments, and the more recent focus on hierarchical control and virtual-real closed loops reflects the feedback correction and normative guidance emphasized in parallel execution. VLA models exhibit deep coupling between semantics and action, iterative cycles linking simulation with reality, and steadily improving verifiability. Nonetheless, challenges remain in generalization, semantic alignment, safety and interpretability, and deployment efficiency. Addressing these issues calls for contract-based task semantics, repairable long-horizon hierarchical planning, engineering-oriented use of world models, multi-level feedback and safety governance, and cross-platform transfer with human-machine collaboration. Examining VLA through the lens of parallel intelligence clarifies its developmental logic and provides methodological support for advancing toward trustworthy real-world applications.

Key words: parallel intelligence, vision-language-action model, embodied intelligence, multimodal fusion, virtual-real interaction

0　引言

视觉-语言-动作（vision-language-action，VLA）模型已成为近年来具身智能研究的前沿方向。这类模型以大规模视觉-语言模型（vision-language model，VLM）的跨模态表征能力为基础，将视觉感知、自然语言理解与动作生成统一到同一体系中，从而驱动机器人等智能体在真实物理环境中完成复杂任务[1]。相较于仅能实现理解与表达的VLM，VLA模型进一步实现了“能看、能说、能做”的闭环联动，标志着人工智能由虚拟空间向物理世界延伸的关键突破。这一跃迁不仅在技术路径上拓展了多模态模型的边界，也在研究范式上推动了具身智能的体系化发展。随着Robotics Transformer（RT）系列（RT-1、RT-2、RT-X）、多模态提示驱动的通用机器人操作以及OpenVLA等代表性工作相继出现，VLA模型的研究与应用正呈现出加速升温的态势，显示出其在机器人操作、交互式任务执行乃至复杂开放场景中的巨大潜力[2]。

如果从更长远的学术脉络来看，二十余年前诞生的平行智能理论已经从宏观视角探讨过类似于VLA模型的虚实交互复杂系统设计方案[3-4]。平行智能理论构建了“人工系统-计算实验-平行执行”三位一体架构，以人工系统支撑现实行动，通过虚拟实验反馈修正现实决策，为复杂系统的智能控制奠定了基础[5]。这一理念与当前VLA模型的发展方式在本质上具有呼应。VLA通常借助仿真环境或数字孪生进行预训练，通过多模态输入（视觉+语言）驱动动作决策，使智能体能够在真实物理环境中执行复杂操作。这种“虚拟先行+现实反馈”的机制正与平行智能强调的仿真-决策-执行闭环高度契合[6]。平行智能为理解VLA模型的快速发展提供了理论参照，并在方法论层面与其若干机制形成了呼应关系。

本文从平行智能范式的视角出发，系统回顾VLA模型的发展历程，梳理其技术演进与典型成果，并在对比中揭示二者之间的内在联系。通过这种跨越时间的关联，本文意在说明VLA模型作为新兴研究方向，并非凭空而生，而是延续了平行智能长期以来的思路与方法体系。更重要的是，从平行智能的视角切入，有助于重新认识VLA模型的发展逻辑与未来趋势，呼吁学术界在推动模型技术进步的同时，注重其在虚实互动、复杂系统管控与可解释性等方面的深度融合。

1　平行智能的范式基础

1.1　平行智能的提出与发展

平行智能的提出源于对开放环境中不确定性与复杂性治理路径的系统思考。随着信息技术与人工智能不断渗透，单一数学模型或经验规则在复杂社会-工程系统中的适用性逐步受限，需要一种能够在模型-数据-决策-运行全流程形成反馈闭环的新范式。基于此，21世纪初在“人工系统（artificial systems）-计算实验（computational experiments）-平行执行（parallel execution）”（以下简称ACP）方法的框架上，逐步形成了以虚拟与现实并行协同

为特征的平行智能体系，并在随后二十余年得到完善与推广[5]。

在平行智能体系中，人工系统是真实物理系统在数字空间的化身。人工系统是多智能体环境、虚拟社会或数字孪生等软件定义的虚拟系统，其功能不止于静态映射，更强调对现实行为机制与社会因素的可计算刻画，从而突破现实试验条件与物理边界带来的约束。在人工系统的基础上，计算实验通过可重复、可控、可扩展的虚拟试验来探索运行轨迹与策略空间，遵循重复、随机化、分组等实验设计原则，既降低现实试错成本，也在更大范围、更高维度揭示潜在规律。进一步地，平行执行强调虚实互动的双向反馈：虚拟世界产出的规范性指导反哺现实系统的在线控制与管理，现实系统的最新状态又用于校准与迭代人工系统，从而形成“描述-预测-规约”一体化的动态闭环。

学术界常见的一个误区是将平行智能与近年来广受关注的数字孪生相混淆，但必须指出的是，平行智能不是数字孪生。数字孪生作为一种以虚拟体映射现实体为特征的技术，侧重高保真同步与状态感知，强调对现实系统的描述性刻画与运行监测。与之不同，平行智能并非停留在描述的层面，而是在同一机制中整合3类智能，即描述智能、预测智能与处方智能，面向复杂系统的治理与优化输出可执行的规范性指导，这也是其能够实现“虚拟先行-现实反馈-持续校准”的重要原因。在具体实施路线上，平行智能通过管理与控制、实验与评估、学习与训练这3类运行模式协同推进，依托虚实并行与闭环迭代，实现对复杂系统的可靠管理与持续进化，如图1所示。

随着平行智能技术的发展，其理论与方法从早期聚焦如何借助虚拟实验改进现实中的交通调度系统[4,7]，逐步外延至更广泛的社会与产业场景，包括城市建设规划[8]、能源与矿业[9]、公共安全[10]、医疗[11]、饮食[12]、教育[13]、文旅[14]以及军事[15]等方向，形成面向多行业的通用方法论布局。平行智能系统共性的本质机理可概括为“在人工系统中先学、先算、先试，再将规范性指导回流现实系统”的闭环，即以人工系统承载可计算的机制刻画，在计算实验中反复评估与优化策略，随后通过平行执行不断校准现实运行状态，体现出“虚拟先行+现实受益”的特征。这一范式在多类行业实践中得到印证，表明平行智能正由单域示范扩展为多域协同的可复用框架。

1.2 平行智能的三元交互模式及其与VLA的关联

平行智能以ACP为方法论底座，其根本动机在于以可计算、可验证的虚实闭环来对冲复杂系统中的不确定性与异质性。平行智能强调以生物人、数字人、机器人三元协同来实现认知-计算-执行的贯通[16]：生物人提供价值目标、情境知识与评判准则；数字人依托人工系统在虚拟世界中承载模型、数据与规则，承担知识的计算与推演；机器人作为具身执行体在物理世界直接作用于环境并返回可观测反馈。设计这种三元循环共生的模式可将虚拟与现实的互动演化落到计算机可实现的程序化流程之中。

值得注意的是，平行智能的三元交互模式近年来已得到国家层面的呼应与支持。国务院在2025年8月发布的《国务院关于深入实施“人工智能+”行动的意见》（国发〔2025〕11号）[17]中明确提出，要推动构建面向自然人、数字人和智能机器人多元一体的发展格局。这表明，平行智能已经在相关政

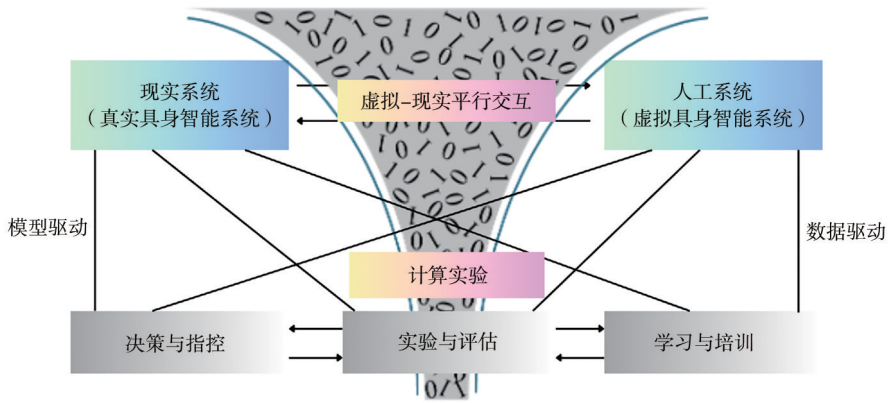

图1 平行智能系统基本架构示意图

策中获得了进一步的肯定和延展，显示出其前瞻性与一定的战略价值。

值得指出的是，VLA 模型与上述三元交互模式其实具有天然映射关系：VLA 的视觉侧重对物理环境的感知与状态获取，可对应机器人的具身观测与交互；VLA 的语言承担意图表达、知识注入与目标约束，可对应生物人提供的规则、语义与评价；VLA 的动作将推理与计划转化为可执行的控制序列，闭环回到物理世界完成控制动作的实施；在 VLA 模型内部，利用大规模多模态表征与动作令牌（token）的推演/生成过程，恰是数字人在人工系统中进行语义计算、策略演化与风险外推的实例化体现。将 VLA 模型与平行智能的三元交互模式联系起来，不仅揭示了当前具身智能领域快速发展的深层逻辑，也为未来复杂系统的智能管控和虚实融合技术发展提供了范式依据。

1.3　平行智能的复杂系统范式与 VLA 运行逻辑的映射

1.2 节强调了平行智能的三元交互模式及其与 VLA 在角色结构上的对应关系。平行智能不仅提供了智能构成要素的框架，还在复杂系统研究中提出了一种运行机制，即“虚拟先行+现实反馈”的范式[18]。该机制与 VLA 的训练和部署过程契合。

复杂系统往往规模庞大、要素多样且演化动态显著，传统依赖解析建模或单纯现实试错的方法常常受限于高昂的成本与风险。平行智能通过人工系统先行构建虚拟环境，在计算实验中对策略和方案进行反复推演与比较，从而在虚拟空间中筛选和优化可行路径。随后，平行执行将虚拟实验的成果迁移到现实系统中落地实施，而现实运行生成的数据又不断回灌人工系统，使虚实之间形成动态循环。这种范式使复杂系统的研究与治理摆脱了单一依赖现实试错的局限，能够在更安全、可控的条件下持续积累经验并提升表现。

这一运行逻辑与 VLA 的发展路径在多环节上相互对应。VLA 在训练阶段往往依赖大规模仿真环境和多模态数据[19]，其过程相当于在人工系统中进行的计算实验；当模型部署到具身智能体之后，现实环境中的交互反馈则不断修正和强化模型性能，对应于平行执行环节。随着虚拟训练与现实反馈的迭代循环，模型逐步获得更强的泛化与适应能力，形成了与平行智能“描述-预测-处方”相似的演进机制。

2　VLA 模型的发展脉络：平行智能视角下的解读

VLA 模型是近年来具身智能领域的重要进展，其发展大体经历了早期深度学习驱动的多模态探索、大模型与 Transformer 引领的拓展阶段，以及 RT 系列与 OpenVLA 带动的快速发展阶段，如图 2 所示[20]。

从平行智能的视角来看，VLA 模型并非凭空出现的。早期研究中已经可以看到自然人、数字人和机器人的互动雏形，与视觉、语言和动作三要素的结合存在内在呼应。本节在回顾 VLA 技术演进的同时，结合平行智能的发展历程，揭示二者在思路上的一致性与在方法上的延展关系。

2.1　萌芽阶段：深度学习与人工系统雏形

VLA 模型的萌芽阶段离不开深度学习的发展与崛起。2006 年之后，深度神经网络逐渐展现潜力，特别是在 2012 年 ImageNet 挑战赛上，卷积神经网络（convolutional neural network，CNN）取得突破性成果[21]，标志着深度学习时代的到来。此后，深度学习迅速扩展到图像识别、视频理解和自然语言处理[22]，为多模态学习奠定了基础。AlphaGo 在 2016 年击败人类顶尖围棋选手[23]，将深度学习成果推向社会大众，成为人工智能进入大范围推广应用时期的标志。

在机器人领域，这一时期的研究多以 CNN 和循环神经网络（recurrent neural network，RNN）为核心架构：前者用于提取视觉观测中的空间特征，后者则在处理语言指令与动作序列等时序信息方面展现出显著优势。Tanwani 等[24]的研究展示了如何利用序列建模方法（包括 RNN）进行机器人模仿学习，体现了深度神经网络在处理多任务时序数据上的潜力。在此基础上，CLIPort 框架提出将大规模视觉语义表征与空间推理融合，从而提升机器人在抓取与放置任务中的泛化能力和操作精度[25]。尽管这些方法在数据规模和任务复杂性上仍受限制，但它们首次在实践中给出了从视觉与语言到动作的完整通路，为后续 VLA 模型的发展奠定了方法论与实验基础。

从平行智能的角度来看，这些探索可以视为人工系统的雏形。当时的研究者已经在虚拟环境中训练模型，并在其中开展试错实验，再将经验迁移到现实场景。这种虚拟先行的方式，与平行智能提出

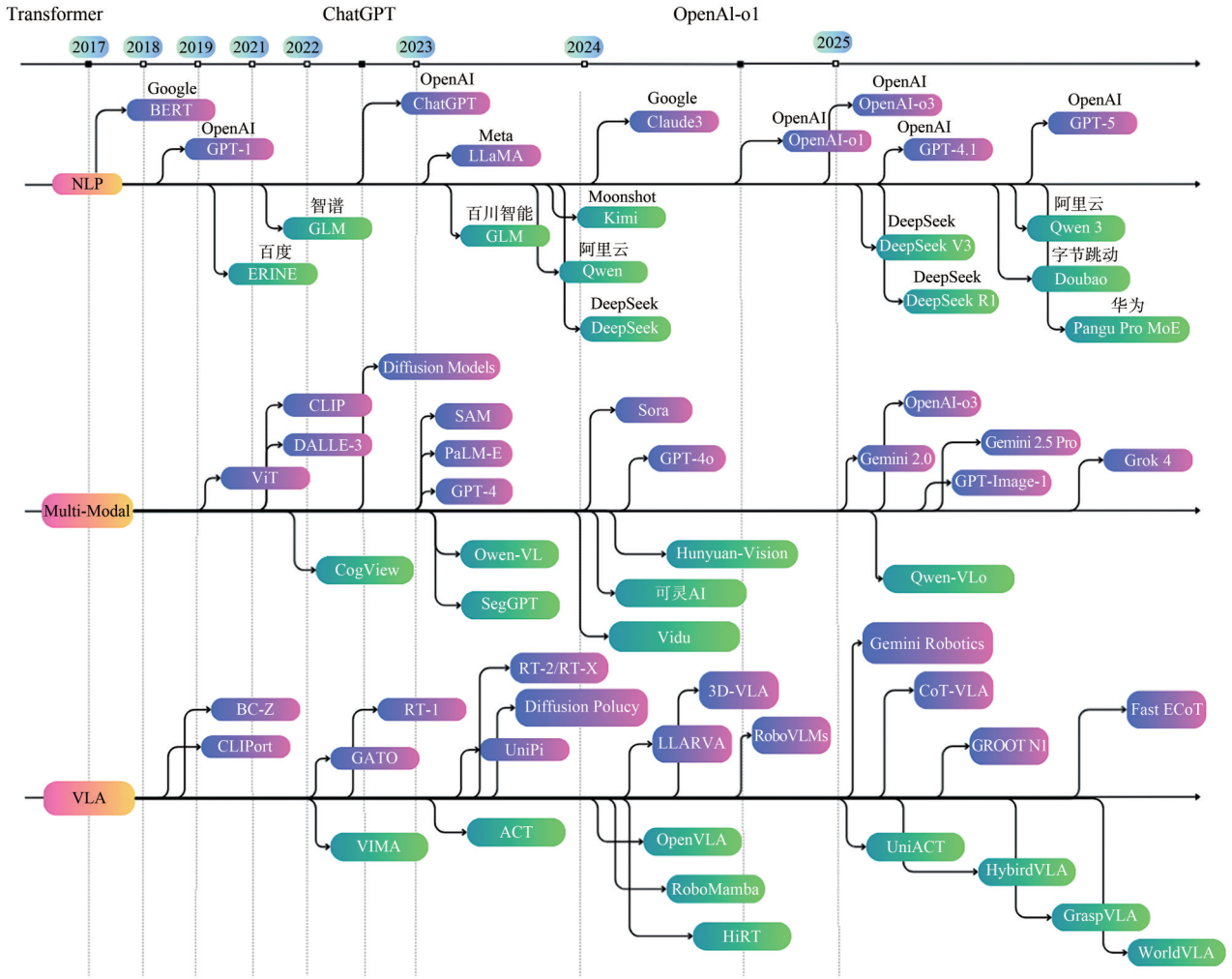

图2　VLA 模型技术发展的时间线

的 ACP 架构在理念上相通。尽管形式与内容尚不完善，但它为后续大模型驱动的探索阶段奠定了基础。

2.2　探索阶段：大模型范式与计算实验拓展

在萌芽阶段之后，VLA 模型进入了以大模型为核心的探索阶段。Transformer 架构逐渐确立为跨模态任务的主干架构，使视觉、语言与动作序列能够在同一表示空间内进行处理[26]。这一时期出现了一系列代表性成果，例如，RT-1 提出了大规模真实机器人数据上的 RT 框架，实现了跨任务的可迁移性与泛化[27]；RT-2 将机器人动作离散化为文本 token，并与互联网规模的视觉-语言模型共同训练，从而把 Web 语义迁移到控制层[19]；OpenVLA 基于开源 7B 模型，结合 DINOv2+SigLIP 视觉编码与 Llama-2 语言大模型，在近百万条真实演示数据上训练，展示了在多机型、多任务上的鲁棒性[28]。这些探索首次在开放场景下展现了 VLA 稳定执行任务的能力。

在动作表示的研究中，学术界逐渐意识到单一的状态或局部动作特征难以支撑复杂任务的泛化需求，因此开始强调多样化的表征方式。除离散原子动作与目标状态外，轨迹表征逐步成为一种能够覆盖长时程行为序列的方案，从而在不同任务与环境间展现出更强的可迁移性[29]；CaP（code as policies）通过大语言模型直接生成可调用应用程序接口（API）的策略代码，使语言到动作的映射具备更强的可解释性[30]；VoxPoser 利用语言模型推理可供性约束，组合 3D 值函数进行零样本的闭环轨迹规划[31]；此外，LAPA（latent action pretraining）在大规模互联网视频上学习潜动作，并以少量真实数据对齐物理控制[32]；ECoT（embodied chain-of-thought）框架在 OpenVLA 基础上进一步训练具身化的推理链，提升了复杂任务下的泛化表现[33]。动作表示的多样化探索，标志着 VLA 在“如何做”

这一核心问题上迈出关键的一步。

在数据与训练机制方面，探索阶段呈现出跨域联合与规模化实验的趋势。OXE（open x-embodiment）数据集整合了 22 种机器人、百万级回合的操作数据，显著推动了跨机型的策略迁移[34]；DROID 数据集覆盖 564 个真实场景和 84 个任务，提供了大规模自然演示[35]；RH20T 提供了 11 万条接触丰富的多模态序列[36]。另外，CALVIN 基准支持语言条件的长时序操作学习[37]，RLBench 提供了多任务的仿真环境[38]，Franka Kitchen 场景常用于评估策略在复杂操作链条下的表现[39]。通过在这些虚拟与现实资源上的联合训练，探索阶段的 VLA 模型能够在虚拟空间获得通用能力，再通过少量真实数据进行校准与对齐。

从平行智能的视角看，VLA 模型的发展探索阶段在计算实验环节实现了突破：人工系统由萌芽期的原型迅速演化为支撑大规模实验的知识-环境平台；依托大模型与大数据，研究者在虚拟环境中系统化检索与评估多种动作表示（离散原子、轨迹、目标状态、代码、潜动作等），以多轮仿真对比确定兼具可迁移性与可解释性的策略；同时，通过跨域联合训练将互联网多模态语料与多源机器人轨迹耦合，在统一表示空间内实现跨任务、跨平台对齐，并以少量实际数据完成校准。由此，VLA 逐步形成可计算、可复现的试验机制；尽管虚实闭环的平行执行尚未成熟，但是其雏形已现，并为随后的快速发展阶段奠定了直接基础。

2.3　快速发展阶段：分层模型与虚实反馈闭环

这一阶段的“快速”首先体现在能力基座与共同底层的成形：RT-1 在多机型、多任务的真实演示上建立了可迁移的控制框架，证明了大规模轨迹预训练的有效性[27]；RT-2 在此基础上将互联网视觉-语言知识与机器人轨迹协同微调，使开放语义下沉至执行层[19]；OXE 聚合百万级跨机构演示并训练出跨平台通用策略 RT-X，确立了数据与能力共享的范式[34]；围绕这些基座，开源 VLA（如 OpenVLA）降低了使用门槛，并在近百万个真实演示上报告稳健的多机型、多任务表现[28]，π0 在“感知-控制”一体建模上以预训练 VLM 为语义底座，并采用 flow-matching 的连续动作生成范式，在多类抓取/放置等操控任务上，其连续控制性能与泛化能力优于若干自回归/行为克隆基线[40]。在系统设计上，VLA 模型在其快速发展阶段的标志性特征是高层语义/任务分解+低层实时控制的分层结构：高层聚焦任务编排与语义约束，低层负责闭环执行与时延控制，从而在长时序复杂任务与实时稳定性之间取得折中，VLA 研究范式由单一“感知-动作”映射转向层次化智能[20]。

与之配套的，是数据与训练机制的体系化：联合互联网多模态语料与多源机器人演示成为常态，使模型既具备开放领域语义理解，又具备具体控制技能（RT-2 直接验证了开放语义→执行层的迁移可行性）；EgoVLA 利用大规模第一人称视频-语言数据学习通用表征，并结合人-机动作映射（如逆运动学/模仿对齐）将人类演示转移到机器人控制，从而在跨场景与新任务上获得更强的迁移与泛化[41]；同时，引入世界模型/视频生成以搭建可计算的虚拟试验支架，在模拟中开展长序列预测与策略演练，再迁移到现实执行（如 ReBot 通过“现实-模拟-现实”的数据循环，在模拟中对真实策略进行大规模增广与回放，再与真实数据联合训练以扩展任务多样性并提升跨域鲁棒性，从而改进 VLA 在新环境/新分布下的总体性能[42]；ECoT 在动作生成前显式生成具身化思维链，将任务/子任务分解与视觉-运动要素对齐联动，从而在未见场景与指令下提升泛化并增强可解释性[33]）。在动作建模层面，扩散策略为连续高维动作提供了稳定而多峰的生成表征，缓解了行为克隆一对多映射的模糊性，并在多个机器人操作基准上取得提升，成为动作表示的重要手段[43]。

从平行智能的视角看，这一阶段的发展体现出平行执行的理念。不同于探索阶段主要依赖虚拟实验，这一时期的模型在虚拟与现实之间形成了双向流动：虚拟环境生成大规模数据和策略，现实执行提供反馈与修正，二者相互作用，逐渐逼近知行合一的状态。这与平行智能提出的“虚拟先行+现实反馈”机制一致，如图 3 所示。可以说，VLA 模型在这一阶段让平行执行的设想更好地落到了实际工程之中。

3　平行智能视角下的 VLA 模型详细解析

本节以平行智能的 ACP 范式为纲，从机理层而非方法层阐释 VLA 模型的工作原理，强调“表征、接口、约束”的统一组织与在环闭合。

3.1　人工系统：虚拟表示与环境构造

在平行智能范式中，人工系统是现实世界的虚

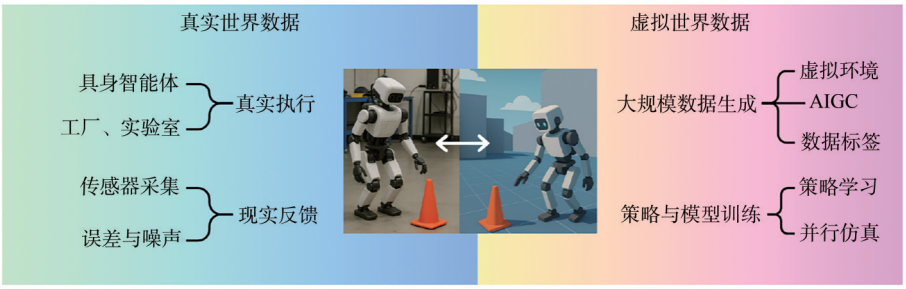

图3 平行智能体系对于虚实数据的生成与使用方式

拟映射与知识载体，而VLA模型的语义到动作推理与执行需在这一虚拟环境中编译、注入约束并接受验证。基于此，本节将依次说明数据如何规约为可计算环境、跨模态表征如何在该环境中对齐与保持、任务意图如何通过接口组织映射为行动等。

3.1.1 数据到人工环境构建

人工系统的首要职责，是将多源感知与知识材料转化为可计算的虚拟环境，使语义与动作在其中形成稳定映射。如果仅将数据视为被动样本堆积，系统便难以支撑闭环推理与执行。因此，可将这一过程理解为面向任务的知识转化，其核心在于表征规约、约束外显与接口协约的协同落地。

在表征层面，关键不在模态多少，而在能否提炼出对任务最小且充分的表示，并在同一任务本体下实现几何-语义-拓扑的层级对齐。单纯的点云或网格可满足定位与避障，但复杂操作往往要求将场景模型提升为“度量-语义-拓扑”多层结构，以便承载对象实例、场所与关系等高层概念，由此为上层推理与规划提供可操作的空间理解[44]。据此，从数据到环境的映射不能止于几何重建，而必须在跨层级表征上完成规约与对齐。

在约束层面，单纯数据并不能刻画行动边界。只有把几何规律、物理定律与基于常识的语义约束显式嵌入，虚拟环境才具备可验证性与可复现性。针对家庭操作的研究以语义约束为桥，提出将物理接触约束与常识性语义约束并置建模，使擦拭、倒液等动作在遵守物理合理性的同时，满足“不断贴附”“不外溢”等语义条件，从而避免表面上可行但语义上失当的行为[45]。这说明，数据到环境的转化必须把隐含边界条件外显化，才能保证生成策略的可执行与可审计。

在接口层面，人工环境不仅是知识存储，更是多模态交互的中介。为保证语言、视觉与动作在同一任务本体下协同，需要以契约化接口明确输入输出语义、尺度与不确定性表达。以知识图谱为骨架的任务-动作-技能分层抽象已证明，可以为接口组织提供统一的结构化语义，并支持跨模态查询、推理与模板迁移，从而把上层意图稳定地下传为可执行目标，再把低层观测一致地上返为语义证据[46]。

从样本数据到人工环境的构建，本质上是一条结构化的知识编译链：感知材料规约为最小充分表征，隐含规则转化为显式约束，多模态输入经由契约化接口形成可验证的语义闭环。面向VLA模型的发展，只有在多模态融合、任务规划与交互中建立这种语义闭环与可验证性，人工环境才能真正成为先虚拟后现实的枢纽[47]。

3.1.2 多模态感知与任务表示

在人工系统中，多模态感知与任务表示的要务在于构造可计算、可验证、可迁移的中介，使视觉、语言与动作围绕同一任务本体协同工作。单纯叠加模态或拼接特征难以承载从指令理解到动作生成的因果链，必须以清晰的结构来组织信息与约束。一个行之有效的思路是以“分解-对齐-兑现”为框架，先将任务拆解为可操作的要素，再将不同模态在这些要素上对齐，最终将对齐后的意图稳定地落到可执行的控制量。

所谓分解，是将任务内容还原为相互支撑的4类信息：语义目标界定“做什么”，几何可供性与接触关系界定“在何处、以何方式可做”，时序因果规定“先后与条件”，约束边界明确“在何范围内安全有效”。在这一基础上进行对齐，意味着在上述4类信息上实现语言、视觉与动作的同位投影，使一条指令、一帧视觉观测信号与一段动作在判据上保持一致[25]；换言之，形成跨模态不变量，减少模态间的语义漂移与尺度误配。

为使理解能够落到执行，实践中常用一种轻量的执行接口，可视为token的抽象。上行方向，观测与语言生成token的目标、优先级与置信区间；

下行方向，规划器或策略网络将token解读为关键帧、参考轨迹或动作分布；其间配以在线的安全过滤与回退机制，将物理与语义约束转化为可监控的约束满足问题。token化带来的收益在于，把高维感知压缩为离散、可检验、可组合的决策单位，从而抑制跨模态误差的级联，并为复用与解释提供清晰的颗粒度。

表示形态包含两类：其一为任务表示的符号规格，以时序逻辑等形式化方法把达成、避免、顺序、时限等语义落为可验证约束，从而收紧搜索空间并明确先后关系[48]；其二为策略表示的数据驱动分布式，以生成式策略在观测条件下直接学习多峰动作分布，因而能覆盖不确定性并对感知噪声保持较强适应力[43]。工程上宜将二者并行，用符号规格给出边界与时序架构，用分布式策略在架构内表达多样性与置信度信息，并以坐标与单位一致、置信传递与失效回退等接口契约维持对齐；评测应从离线准确率转向跨模态保持性、可执行充要性与可迁移性下界3项面向执行的指标。

3.1.3　虚拟环境与场景生成

在人工系统的语境下，虚拟环境旨在为推理、学习与评测提供可治理的实验空间，所有关键变量均可被声明、干预并复现。虚拟环境的品质可基于可控性、保真度与覆盖度这3个维度加以衡量：可控性强调以显式参数重现实验条件，并支持因果干预与严格对照；保真度要求几何、接触与动力学在数值上稳定且在物理上可信，能够支撑策略跨域迁移；覆盖度关注场景、任务与长尾扰动的分布丰富性，同时避免引入系统性偏差。三者相互制约又相互支撑，需要在统一的语义、几何与动力学坐标系内共同优化与验证。

围绕上述3个维度，现有的虚拟环境构建方案可概括为两类机制：其一是程序化仿真与脚本化示教，通过情景语法将语义、几何与动力学进行可编排的耦合，批量生成可审计的轨迹与对照实验，优势在于因果可控与规模化生产，短板在于逼真度与稀有事件覆盖不足[49]；其二是学习型世界模型，通过潜在动力学或生成式表征在观测条件下外推未来，再经逆动力学或策略映射落到动作，优势在于多样性、可组合性与对长时序的外推能力，短板在于物理一致性、误差可见性与稳定训练[50]。两类机制本质上是在可控性、保真度与覆盖度之间选择不同的折中点，前者偏向因果与可再现性，后者偏向分布覆盖与组合泛化。

为弥补两类机制的差异，可以采用“规范驱动×模型补全”的融合闭环。以规范化任务描述与情景语法给出可控且可审计的上界，在该上界内由模型扩展反事实与多样性以提升覆盖度，并以物理在环与一致性度量对生成结果进行筛校以维持保真度，度量应覆盖接触合规、时序一致与迁移下界等关键证据[51]。工程上将虚拟场景建设为可追责的数据生产线，包含明确变量与契约、可信动力学实现、可回放种子与完整元数据以及多粒度评测指标，同时引入现实反馈对仿真参数与随机化策略进行自适应校准。由此在统一坐标系内形成从规范到数据、从数据到策略、从策略到诊断的闭合链条，使3个维度的指标获得可验证的协同优化。

3.1.4　表征融合与人工系统的知识功能

在平行智能的视角下，VLA模型的表征融合旨在把视觉、语言与动作统一到同一语义-几何-时序坐标系，上承任务语义与场景证据，下接可操作的决策接口与约束检验。围绕这一目标，现有的表征融合方法可提炼为4类：其一，任务本体对齐，以任务/对象本体或知识图谱为载体，将语言意图与感知证据对齐到对象-关系-动作原语的结构化空间，便于跨任务迁移与可解释检索；其二，几何一致中间态，在统一的三维语义场或场景图中表达实体、接触与安全走廊，让语义定位与空间约束在可操作的中层表征合流；其三，时序预测，以世界模型在观测条件下外推未来，提供可达性与因果链的内在证据，并显式管理模型偏差与不确定性；其四，约束嵌入，把任务目标、先后关系与安全边界以形式化知识注入表征与决策接口，使学习与规划在同一表示层上接受可验证的约束与回退策略的监督。

上述4类思路顺序衔接、功能互补：本体对齐确保语义一致，几何中间态提供空间与接触合规，时序预测支撑长时程计划与反事实检查，约束嵌入保证可验证与可回退。由此形成从语义到行动的闭环，使VLA的表征既能理解任务，又能落到执行，还能经受验证。

3.2　计算实验：虚拟推演与训练机制

计算实验为VLA模型的发展提供了可控且高效的验证环境，通过虚拟推演可以在无风险条件下检验推理与决策机制的可靠性。与此同时，仿真训练能够反复积累多模态交互数据，为模型迭代提供了规模化的经验支撑。

3.2.1 仿真环境中的可重复实验

为了将表征融合知识底座转化为可靠能力，必须将其置于可重复的计算实验中，经由对照、反事实与度量反复求证。因此，仿真环境中的一项重要任务是把表征层的承诺转译为可复现的实验单元与可比对的证据：明确变量，固定流程，公开判据，使不同的研究者在相同配置与扰动设定下得到一致结论，并在受控变化中检验稳健性与迁移性。

围绕这一目标，可重复实验应当将 3 类要素组织起来：其一是实验单元与变量治理，以任务脚本与场景参数声明任务语义、对象集合、初始状态与成功判据，将随机种子、视角与扰动分布纳入元数据，使相同实验设置的试验可回放、可追责；其二是对照与反事实设计，在保持语义约束不变的前提下，通过参数化扰动与受控域随机化生成最小差异样本[52]，配合自动重置与批量回放[49]，既保障横向可比，又在低成本下放大知识差异，用于检验跨模态一致性、可执行充要性与抗干扰能力；其三是证据与指标体系，除任务达成类指标（成功率、路径或轨迹一致性、时间与能耗）外，应同时纳入过程与迁移指标（时序一致、接触合规、域外泛化表现），并以标准化日志与完整元数据记录试验配置与随机性来源，支撑回放、误差归因与长期回归分析。由此形成可治理的科学实验：在统一栈内定义任务与度量，在受控扰动下放大差异，在标准指标上进行统计检验，并形成“配置-运行-复盘”的闭环。对 VLA 而言，这使语义一致、空间合规、时序可达与约束可验得到系统化的验证与诊断，也使跨模型、跨团队的比较从演示式展示转为可重复的证据生产，为后续从仿真到现实的迁移提供可审计的实验基础。

3.2.2 实验设计学：从数据扩展/迁移到可验证的可靠性

如果说可重复实验为计算实验建立了最低限度的可比性，那么实验设计学的任务就在于把这种“可重复”进一步提升为“可验证”，即在虚拟推演中形成能够支撑跨域与跨体现可靠性检验的制度化机制。

实验设计强调对变量的精细治理、对失败的可见追踪、对判据的统一规范以及对迁移的预测性校准，从而把原本松散的数据资源与表征接口转译为具有审计意义的证据链。首先，实验不应停留在规模扩张的堆砌，而要在域空间中明确区分语义、扰动与体现 3 类变量，在保持语义约束恒定的条件下，通过单因素扰动构造最小差异对照，使实验能够在低成本下放大知识差异并形成因果可解释的比较[53]。第二，失败并非偶然噪声，而是系统性证据的一部分，因而必须完整记录和分类：状态快照、随机种子、控制日志应构成可回溯的闭环，将错误锁定到语义解读、空间几何、接触约束或时序逻辑等环节，使实验不仅能回答“成败如何”，更能回答“为何如此”[54]。第三，机器人操作与操控泛化基准已不满足单一成功率输出，且在多个扰动维度（如背景、光照、对象外观 / 形态、视角变化等）下报告性能下降，以及引入演示一致性等过程指标，以评估跨场景与跨体现扰动下的稳健性[55]。最后，迁移的可预测性应建立在仿真与现实之间的校准闭环之上：在虚拟环境中通过受控扰动识别最具区分性的位置、视觉、几何、背景等维度，再在现实环境中进行抽样验证，并将现实反馈用于校正模拟环境配置，从而使仿真性能排序越来越能预测真实性能排序[56]。

3.2.3 推理增强与策略生成

在计算实验中，推理增强与策略生成承担着把抽象表征转化为可执行行动的任务，其核心在于让策略生成具备结构化与可验证的能力：实验设计中通过引入显式中间结构（如子任务/阶段分解与状态比对）实现任务拆分与行为阶段监控，使策略生成过程可观察、可检验，并在发现偏差或异常时插入校正步骤，从而增强策略的稳健性与可执行性[57]。借助这种分解，计算实验能够在虚拟环境中逐步检验每个子目标的合理性，并在发现偏差时进行调整，从而保持整体策略的逻辑一致性与可执行性。

与此同时，策略生成以对未来状态的可计算预测为基础：在虚拟环境或世界模型中对候选轨迹进行模拟推演，并结合任务与安全约束进行一致性与可行性检验；剔除不满足物理规律或环境约束的方案，必要时触发轨迹重规划，从而形成“生成-验证-修正”的闭环流程[58]。这一机制不仅提升了策略在仿真环境中的稳健性，也强化了其在跨场景与长时序任务中的可迁移性。更重要的是，推理增强与策略生成实验把可解释性纳入显式考量：实验不仅关注策略是否完成任务，还要求能追溯其推理链路、检验其中间结构与结果的一致性。由此，策略不再只是“能做”，而是“能说明为什么这样做”，其透明性和可验证性使其更契合平行智能强调的证据化和可审计性。

3.3　平行执行与虚实交互

平行执行的核心在于虚与实的双向耦合：现实执行不断将状态和偏差反馈至虚拟系统以修正环境与模型，虚拟系统则通过推演与约束生成规范性指导反哺现实。两者形成持续迭代的循环，使虚拟与现实逐步收敛并保持一致性。本节分别从“现实到虚拟的反馈修正”与“虚拟到现实的规范性指导”两个方面展开论述。

3.3.1　现实到虚拟：反馈修正与环境校准

在平行执行的框架下，现实执行并不是虚拟实验的单向验证，而是虚拟与现实之间不断互证的环节。现实世界的感知数据、执行偏差与失败样例，为虚拟系统提供了不可替代的校准信息，使人工环境能够在几何建模、动力学约束和语义接口等层面持续修正。正是这种“反馈-修正”的循环，使虚拟环境不再是静态的理想化副本，而是随现实运行而演化的动态系统。

具体而言，反馈修正从多个层面展开。首先是动力学模型参数的更新：现实机器人执行往往揭示出虚拟环境中动力学模型精度的不足，例如摩擦系数偏差导致的抓取力不稳、质量分布误差引起的运动轨迹漂移；通过对比真实执行轨迹与虚拟预测轨迹，系统能够迭代修正这些底层参数，使虚拟环境的响应曲线在力学、运动学和接触动力学层面逐渐贴近真实世界，从而避免“眼高手低”的情况[59]。其次是环境配置与感知模型的刷新：现实场景中光照的动态变化、背景遮挡、传感器噪声甚至微小的相机标定误差，都会与虚拟环境的理想化假设产生偏差。如果不修正这些偏差，虚拟推演就会在关键细节上出现系统性错误；通过将现实观测转译为统一的语义层次信息，虚拟环境能够不断重建物体位置、材质属性、背景纹理乃至干扰物分布，并结合多模态传感器的统计特性更新感知模型，从而在仿真与现实之间维持长期的一致性[60]。最后是策略接口与约束条件的校正：在现实执行出现失败或异常时，系统不仅回溯并分类失败类型（如顺序偏差、对象操作漏掉、视觉标定误差等），而且会调整任务约束或目标界面，例如修正子目标顺序、收紧任务边界、重新估计相机与操作工具之间的空间关系，以防同类失误再次发生[61]。

需要说明的是，反馈修正过程并不是单次性的补丁，而是伴随执行持续进行的动态校准，使虚拟环境和现实环境在长期迭代中逐渐趋同，为后续推理增强与策略生成提供稳固的支撑。

3.3.2　虚拟到现实：规范性指导与策略约束

在平行执行的另一侧，虚拟系统将推演与预测结果转化为对现实执行的规范性指导。通过在虚拟环境中进行轨迹生成、约束评估与未来状态预测，系统能够在现实操作发生之前就识别潜在的风险与冗余，从而形成对现实的前瞻性支持。这种从虚拟到现实的映射并非简单地下发任务指令，而是通过策略生成和安全约束的形式，为现实系统提供可解释、可检验且带有边界条件的执行方案。

具体而言，虚拟系统对现实的规范性指导至少包括3个方面。其一，轨迹与策略的筛选：虚拟推演能够在候选动作或路径之间进行对比，剔除不符合物理规律或约束条件的方案，从而保证现实执行的稳健性[62]。其二，安全边界与约束的注入：虚拟环境可提炼出操作空间中的禁止区、安全阈值或容许范围，这些约束一旦传递给现实控制层，就能有效避免硬件损伤或危险操作[63]。其三，解释与诊断信息的生成：虚拟推演过程中形成的中间状态与因果链条，可以作为对现实执行的注解，帮助识别任务中潜在的失败点或优化点，并在必要时为人机协作提供参考[64]。

4　VLA 模型技术发展的未来展望

在前3节的分析中，本文已经从人工系统的构建、计算实验的设计到平行执行的机制，梳理了VLA模型在平行智能范式下的发展现状。这一系列研究奠定了模型具备多模态表征、可复现实验和虚实互证的基础，但要真正实现可信赖的应用，还需要在方法论和系统工程上进一步突破。未来的研究重点不应仅仅停留在“能重复”的实验演示，而要迈向“可验证”的证据链条，即如何让语义理解、过程执行与跨域迁移都能够被系统化度量、审计和解释，从而使虚拟与现实的结合更具透明性和可追责性。

在实现可验证性的前提下，另一个亟须突破的方向是任务语义的契约化表达。现实应用中的指令往往以自然语言呈现，但其模糊性和上下文依赖性使虚拟到现实的映射容易出现偏差。未来需要一种“契约化编译”机制，将任务语义系统地转写为形式化的约束与接口：任务契约明确目标、前置条件与时限，环境契约声明几何与动力学边界，安全契约设定禁止域、容许范围及回退策略。这样的契约

不仅能够被虚拟系统解析并在计算实验中进行推演与对照，也能直接下传到现实控制层，为策略生成提供透明且可检验的边界条件。

在任务语义得到契约化表达之后，新的挑战在于长时序任务的分层规划如何具备可修复性。现实环境不可避免地伴随突发扰动与执行偏差，如果规划仍然停留在“一次性生成”的模式，便难以应对动态变化。未来研究需要将“生成-验证-修正”机制融入分层体系之中：高层侧重任务分解与可达性校核，中层侧重关键帧与安全走廊的动态重构，低层依靠滚动时域控制吸收外部扰动。这样形成的分层修复链条，不仅能够在语义目标变化或环境条件突变时保持规划与执行的一致，也能在硬件漂移与感知误差下维持任务的连续性，从而实现长时序任务的稳健执行。

世界模型的工程化使用也是未来值得重点推进的方向。当前研究多集中于其生成未来状态或视频序列的能力，但真正的价值在于如何把这些推演结果转化为对现实执行有约束力的证据，而不是单纯的可视化预测。未来需要更系统的方法来度量仿真结果与真实表现之间的一致性，并建立相应的指标体系，以检验虚拟预测的可靠性。同时，应当形成“识别-抽样-回灌-再评估”的循环机制：先识别最具区分力的扰动维度，再在现实中进行抽样验证，将结果反馈到虚拟环境，从而逐步提升预测效度。进一步而言，世界模型的输出不仅要包含状态与轨迹，还要具备不确定性、置信水平以及约束违背的结构化标注，使其能够为平行执行提供规范性指导。通过这种工程化方式，世界模型将从一种预测工具转变为支撑决策与约束的证据平台，更好地发挥虚拟与现实之间的桥梁作用。

在平行执行中，现实反馈也不应仅被看作误差补偿，而需要被分解为多层次的修正路径。未来研究可从3个方面展开：参数回写，持续标定和辨识动力学模型与传感链路；场景回写，利用现实观测不断更新物体、材质、光照、遮挡与干扰布局；接口回写，在出现失败模式时调整搜索空间、安全阈值与回退策略。3类回写若能形成版本化与脉络化的记录，将使修正的动因、对象与程度都具备可解释性与可审计性，为模型的长期演化提供清晰轨迹。

在实际运行中，安全问题是VLA模型能否实现可信落地的关键。未来研究应建立从静态规则到在线监护的多层安全治理框架，不仅要在任务开始前明确禁止区、容许域和接触阈值等硬性边界，更要在运行过程中实时监控执行状态，形成可动态介入的安全机制。具体而言，可以通过设定安全预算来量化风险暴露范围，在关键边界附近布设监护器实现提前干预，并配套制定处置单以规范触发、降级与恢复的流程。

对于VLA模型而言，平行执行中的安全治理尤其重要。未来研究需要在虚拟端预设禁止区、容许域和安全阈值，并在现实执行中通过实时监控及时干预，防止危险操作。同时，还应建立触发、降级和恢复的标准流程，使安全约束成为系统运行全周期内的主动机制，从而在确保任务完成的同时保障硬件与环境的安全性。

在VLA模型的应用中，不同机器人的形体、自由度与末端机构存在显著差异，这导致同一策略难以直接迁移[65]。未来研究应探索中性动作语的设计，即将姿态、关键帧与约束原语抽象为统一表示，用于承载不同机器人间的动作与观察对齐。配合感知-坐标-控制的对齐协议和平台差异补偿机制，可以使策略在新的机器人形体上实现冷启动，并在少量监督下迅速调整。这种不同机器人形态之间的策略对齐与语义插槽，将是推动VLA模型通用化部署的重要方向。

未来的研究还需要回应效率、数据和人机协作3个方面的挑战。其一是在资源与时延受限的条件下实现高效部署：如何在云、边、端之间合理分配推理与控制任务，使VLA模型既能保持高层语义的复杂推理能力，又能满足低层执行的实时性要求。其二是在数据生产与基准治理上形成系统化流程：不仅要追求样本规模，更要关注覆盖度、保真度与可控性，通过标准化的配置、运行和复盘机制，为实验与对比提供长期有效的支撑。其三是在虚拟到现实的下传中强调解释与规范性指导：模型不仅要输出“做什么”，更要明确“为什么这样做”“何时需要修正”以及“如何进行回退”，以增强人机协作中的透明性与可托付性[66]。

5 结论

本文以平行智能范式为参照，对VLA模型的技术发展进行了系统梳理。可以看到，VLA模型已成为多模态智能迈向现实世界的关键桥梁，其演进逻辑贯穿了人工系统的虚拟建模、计算实验的推理验证和平行执行的虚实交互。与传统的单向试错

相比，这一范式凸显了“虚拟先行、现实反馈”的迭代机制，为复杂任务中的可靠性与可解释性提供了新的路径。未来的发展不只是模型能力的扩展，更关乎方法论的成熟与系统化：如何让语义契约化、任务长时序化、反馈结构化与安全治理一体化，决定了VLA模型能否实现从实验室演示到现实应用的跨越。作为平行智能思想的具体化身，VLA模型的持续演进将推动人工智能由认知层面的理解，走向可验证、可托付的真实世界实践。

参考文献：

[1] SHAO R, LI W, ZHANG L, et al. Large VLM-based vision-language-action models for robotic manipulation: a survey[J]. arXiv preprint, 2025, arXiv: 2508.13073.

[2] SAPKOTA R, CAO Y, ROUMELIOTIS K I, et al. Vision-language-action models: concepts, progress, applications and challenges[J]. arXiv preprint, 2025, arXiv: 2505.04769.

[3] 王飞跃. 平行系统方法与复杂系统的管理和控制[J]. 控制与决策, 2004, 19(5): 485-489, 514.
WANG F Y. Parallel system methods for management and control of complex systems[J]. Control and Decision, 2004, 19(5): 485-489,514.

[4] WANG F Y, TANG S M. Artificial societies for integrated and sustainable development of metropolitan systems[J]. IEEE Intelligent Systems, 2004, 19(4): 82-87.

[5] WANG X X, YANG J, LIU Y H, et al. Parallel intelligence in three decades: a historical review and future perspective on ACP and cyber-physical-social systems[J]. Artificial Intelligence Review, 2024, 57(9): 255.

[6] HU X M, LI S, HUANG T Y, et al. How simulation helps autonomous driving: a survey of sim2real, digital twins, and parallel intelligence[J]. IEEE Transactions on Intelligent Vehicles, 2023, 9(1): 593-612.

[7] WANG F Y, TANG S M. A framework for artificial transportation systems: from computer simulations to computational experiments[C]//Proceedings of 2005 IEEE Intelligent Transportation Systems. Piscataway: IEEE Press, 2005: 1130-1134.

[8] 吕宜生, 王飞跃, 张宇, 等. 虚实互动的平行城市: 基本框架、方法与应用[J]. 智能科学与技术学报, 2019, 1(3): 311-317.
LYU Y S, WANG F Y, ZHANG Y, et al. Parallel cities: framework, methodology, and application[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(3): 311-317.

[9] 陈龙, 王晓, 杨健健, 等. 平行矿山: 从数字孪生到矿山智能[J]. 自动化学报, 2021, 47(7): 1633-1645.
CHEN L, WANG X, YANG J J, et al. Parallel mining operating systems: from digital twins to mining intelligence[J]. Acta Automatica Sinica, 2021, 47(7): 1633-1645.

[10] 周芳, 王瑞. 基于平行系统的网络舆情试验方法[J]. 指挥信息系统与技术, 2013, 4(3): 1-7.
ZHOU F, WANG R. Test method for network opinion based on parallel system[J]. Command Information System and Technology, 2013, 4(3): 1-7.

[11] 王惠珍, 张捷, 俞怡, 等. 平行手术室:围术期护理流程与智慧手术平台管理的新模式[J]. 模式识别与人工智能, 2023, 36(10): 867-876.
WANG H Z, ZHANG J, YU Y, et al. Parallel operating rooms: a new model of perioperative nursing process and smart surgical platform management[J]. Pattern Recognition and Artificial Intelligence, 2023, 36(10): 867-876.

[12] 李柏, 宋秭函, 李鑫源, 等. 基于代理智能的平行厨师:从AI Agents到智慧数字机器人饮食系统[J]. 模式识别与人工智能, 2025, 38(3): 252-267.
LI B, SONG Z H, LI X Y, et al. Parallel chefs via agentic intelligence: from AI agents to smart digital robotic cuisine systems[J]. Pattern Recognition and Artificial Intelligence, 2025, 38(3): 252-267.

[13] 王飞跃. 数字教师与平行教育: 关于ChatGPT之后教学变革的探讨[J]. 智能科学与技术学报, 2023, 5(4): 446-455.
WANG F Y. Digital teachers and parallel education: a paradigm shift in teaching and learning after ChatGPT[J]. Chinese Journal of Intelligent Science and Technology, 2023, 5(4): 446-455.

[14] 张腾超, 田永林, 林飞, 等. 平行旅游:基础智能驱动的智慧出游服务[J]. 智能科学与技术学报, 2024, 6(2): 164-178.
ZHANG T C, TIAN Y L, LIN F, et al. Parallel tourism: foundation intelligence driven smart trip services[J]. Chinese Journal of Intelligent Science and Technology, 2024, 6(2): 164-178.

[15] 王飞跃. 指控5.0:平行时代的智能指挥与控制体系[J]. 指挥与控制学报, 2015, 1(1): 107-120.
WANG F Y. CC5.0: intelligent command and control systems in the parallel age[J]. Journal of Command and Control, 2015, 1(1): 107-120.

[16] 王飞跃, 王雨桐. 数字科学家与平行科学:AI4S和S4AI的本源与目标[J]. 中国科学院院刊, 2024, 39(1): 27-33.
WANG F Y, WANG Y T. Digital scientists and parallel sciences: the origin and goal of AI for science and science for AI[J]. Bulletin of Chinese Academy of Sciences, 2024, 39(1): 27-33.

[17] 国务院. 国发〔2025〕11号. 国务院关于深入实施“人工智能+”行动的意见[S]. 2025.
STATE COUNCIL. Guo Fa [2025] No. 11. Opinions of the state council on further implementing the “AI+” initiative[S]. 2025.

[18] 王飞跃. 平行控制: 数据驱动的计算控制方法[J]. 自动化学报, 2013, 39(4): 293-302.
WANG F Y. Parallel control: a method for data-driven and computational control[J]. Acta Automatica Sinica, 2013, 39(4): 293-302.

[19] ZITKOVICH B, YU T H, XU S C, et al. Rt-2: vision-language-action models transfer web knowledge to robotic control[C]//Proceedings of the 7th Conference on Robot Learning. Cambridge: PMLR, 2023: 2165-2183.

[20] 李浩然, 陈宇辉, 崔文博, 等. 面向具身操作的视觉-语言-动作模型综述[J]. arXiv preprint, 2025, arXiv: 2508.15201.
LI H R, CHEN Y H, CUI W B, et al. Survey of vision-language-action models for embodied manipulation[J]. arXiv preprint, 2025, arXiv: 2508.15201.

[21] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[22] DONG S, WANG P, ABBAS K. A survey on deep learning and its applications[J]. Computer Science Review, 2021, 40: 100379.

[23] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529 (7587): 484-489.

[24] TANWANI A K, YAN A, LEE J, et al. Sequential robot imitation learning from observations[J]. The International Journal of Robotics Research, 2021, 40(10/11): 1306-1325.

[25] SHRIDHAR M, MANUELLI L, FOX D. CLIPort: what and where pathways for robotic manipulation[C]//Proceedings of the 5th Conference on Robot Learning. Cambridge: PMLR, 2021: 894-906.

[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017: 5998-6008.

[27] BROHAN A, BROWN N, CARBAJAL J, et al. Rt-1: robotics transformer for real-world control at scale[J]. arXiv preprint, 2022, arXiv: 2212.06817.

[28] KIM M J, PERTSCH K, KARAMCHETI S, et al. OpenVLA: an open-source vision-language-action model[J]. arXiv preprint, 2024, arXiv: 2406.09246.

[29] GE Z, CHEN C, SINHA A, et al. On learning informative trajectory embeddings for imitation, classification and regression[J]. arXiv preprint, 2025, arXiv: 2501.09327.

[30] LIANG J, HUANG W L, XIA F, et al. Code as policies: language model programs for embodied control[J]. arXiv preprint, 2022, arXiv: 2209.07753.

[31] HUANG W L, WANG C, ZHANG R H, et al. VoxPoser: composable 3D value maps for robotic manipulation with language models[J]. arXiv preprint, 2023, arXiv: 2307.05973.

[32] YE S, JANG J, JEON B, et al. Latent action pretraining from videos[J]. arXiv preprint, 2024, arXiv: 2410.11758.

[33] ZAWALSKI M, CHEN W, PERTSCH K, et al. Robotic control via embodied chain-of-thought reasoning[J]. arXiv preprint, 2024, arXiv: 2407.08693.

[34] O'NEILL A, REHMAN A, MADDUKURI A, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration[C]//Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2024: 6892-6903.

[35] KHAZATSKY A, PERTSCH K, NAIR S, et al. DROID: a large-scale in-the-wild robot manipulation dataset[J]. arXiv preprint, 2024, arXiv: 2403.12945.

[36] FANG H S, FANG H J, TANG Z Y, et al. RH20T: a comprehensive robotic dataset for learning diverse skills in one-shot[C]//Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2024: 653-660.

[37] MEES O, HERMANN L, ROSETE-BEAS E, et al. CALVIN: a benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 7327-7334.

[38] JAMES S, MA Z C, ARROJO D R, et al. RLBench: the robot learning benchmark & learning environment[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 3019-3026.

[39] GUPTA A, KUMAR V, LYNCH C, et al. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning[J]. arXiv preprint, 2019, arXiv: 1910.11956.

[40] BLACK K, BROWN N, DRIESS D, et al. π0: a vision-language-action flow model for general robot control[J]. arXiv preprint, 2024, arXiv: 2410.24164.

[41] YANG R H, YU Q X, WU Y C, et al. EgoVLA: learning vision-language-action models from egocentric human videos[J]. arXiv preprint, 2025, arXiv: 2507.12440.

[42] FANG Y, YANG Y, ZHU X H, et al. ReBot: scaling robot learning with real-to-sim-to-real robotic video synthesis[J]. arXiv preprint, 2025, arXiv: 2503.14526.

[43] CHI C, XU Z J, FENG S Y, et al. Diffusion policy: Visuomotor policy learning *via* action diffusion[J]. The International Journal of Robotics Research, 2025, 44(10/11): 1684-1704.

[44] MASCARO R, CHLI M. Scene representations for robotic spatial perception[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2025, 8: 351-377

[45] IKEUCHI K, WAKE N, SASABUCHI K, et al. Semantic constraints to represent common sense required in household actions for multimodal learning-from-observation robot[J]. The International Journal of Robotics Research, 2024, 43(2): 134-170.

[46] MIAO R Q, JIA Q X, SUN F C, et al. Semantic representation of robot manipulation with knowledge graph[J]. Entropy, 2023, 25(4): 657.

[47] GARG S, SÜNDERHAUF N, DAYOUB F, et al. Semantics for robotic mapping, perception and interaction: a survey[J]. Foundations and Trends in Robotics, 2020, 8(1-2): 1-224.

[48] ZHAO Z G, CHENG S, DING Y, et al. A survey of optimization-based task and motion planning: From classical to learning approaches[J]. IEEE/ASME Transactions on Mechatronics, 2025, 30(4): 2799-2825.

[49] SZOT A, CLEGG A, UNDERSANDER E, et al. Habitat 2.0: training home assistants to rearrange their habitat[J]. Advances in Neural Information Processing Systems, 2021, 34: 251-266.

[50] HU Y C, GUO Y J, WANG P C, et al. Video prediction policy: a generalist robot policy with predictive visual representations[J]. arXiv preprint, 2024, arXiv: 2412.14803.

[51] VRABIČ R, ŠKULJ G, MALUS A, et al. An architecture for sim-to-real and real-to-sim experimentation in robotic systems[J]. Procedia CIRP, 2021, 104: 336-341.

[52] DEITKE M, VANDERBILT E, HERRASTI A, et al. Procthor: large-scale embodied AI using procedural generation[J]. Advances in Neural Information Processing Systems, 2022, 35: 5982-5994.

[53] AHMED O, TRÄUBLE F, GOYAL A, et al. CausalWorld: a robotic manipulation benchmark for causal structure and transfer learning[J]. arXiv preprint, 2020, arXiv: 2010.04296.

[54] ALT B, PICKLUM M, ARION S, et al. Open, reproducible and trustworthy robot-based experiments with virtual labs and digital-twin-based execution tracing[J]. arXiv preprint, 2025, arXiv: 2508.11406.

[55] LUO J L, XU C, LIU F C, et al. FMB: a functional manipulation benchmark for generalizable robotic learning[J]. International Journal of Robotics Research, 2025, 44(4): 592-606.

[56] LI X L, HSU K, GU J Y, et al. Evaluating real-world robot manipulation policies in simulation[J]. arXiv preprint, 2024, arXiv: 2405.05941.

[57] WILLIBALD C, LEE D. Hierarchical task decomposition for execution monitoring and error recovery: Understanding the rationale behind task demonstrations[J]. arXiv preprint, 2025, arXiv: 2505.04565.

[58] LI Y X, ZHU Y C, WEN J J, et al. WorldEval: world model as real-world robot policies evaluator[J]. arXiv preprint, 2025, arXiv: 2505.19017.

[59] SHI L, XU Y X, WANG S Y, et al. An real-sim-real (RSR) loop framework for generalizable robotic policy transfer with differentiable simulation[J]. arXiv preprint, 2025, arXiv: 2503.10118.

[60] YU Y W, LIU L T. Neural fidelity calibration for informative sim-to-real adaptation[J]. arXiv preprint, 2025, arXiv: 2504.08604.

[61] THODUKA S, HOUBEN S, GALL J, et al. Enhancing video-based robot failure detection using task knowledge[J]. arXiv preprint, 2025, arXiv: 2508.18705.

[62] BAR A, ZHOU G Y, TRAN D, et al. Navigation world models[C]//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2025: 15791-15801.

[63] CUI Z J, BAENA F R Y. Forbidden region dynamic active constraints in robot-assisted minimally invasive surgery[J]. IEEE Robotics and Automation Letters, 2025, 10(3): 2950-2957.

[64] KUBE D, HADWIGER S, MEISEN T. Beyond performance: explaining generalisation failures of robotic foundation models in industrial simulation[J]. Biomimetic Intelligence and Robotics, 2025: doi.org/10.1016/j.birob.2025.100249.

[65] SEO M, PARK H A, YUAN S L, et al. LEGATO: cross-embodiment imitation using a grasping tool[J]. IEEE Robotics and Automation Let-

ters, 2025, 10(3): 2854-2861.

[66] 张慧, 梁姝彤, 李明轩, 等. 视觉-语言-动作模型综述: 从前史到前沿[J]. 自动化学报, 2025, 51(9): 1922-1950.
ZHANG H, LIANG S T, LI M X, et al. Vision-language-action models: from the early foundations to the state-of-the-art[J]. Acta Automatica Sinica, 2025, 51(9): 1922-1950.

[作者简介]

李柏（1989– ），男，博士，华东师范大学通信与电子工程学院教授，主要研究方向为自动驾驶轨迹规划、计算最优控制理论、多模态大语言模型与具身智能系统。

郝金第（2002– ），男，湖南大学机械与运载工程学院硕士生，主要研究方向为大语言模型、计算最优控制方法在智慧医疗领域的应用。

孙跃硕（2002– ），男，湖南大学机械与运载工程学院硕士生，主要研究方向为自主无人系统协同运动规划、计算最优控制以及具身智能机器人。

孟雨晴（2001– ），女，华东师范大学通信与电子工程学院硕士生，主要研究方向为多模态大语言模型与具身智能在自动驾驶领域的应用。

黄峻（1998– ），男，澳门科技大学创新工程学院博士生，主要研究方向为平行智能、自动驾驶轨迹预测规划、提示工程以及大语言模型。

田永林（1994– ），男，博士，中国科学院自动化研究所多模态人工智能系统全国重点实验室助理研究员，主要研究方向为平行系统、自动驾驶以及场景工程。

贺正冰（1981– ），男，博士，麻省理工学院信息与决策系统实验室高级研究员，主要研究方向为城市交通、系统工程与人工智能的交叉领域。