Vision-Language-Action：从 OpenVLA 到 ECoT 看具身智能发展

耿浩轩

计算机科学与技术学院

浙江大学

August 31, 2025

摘要

Vision-Language-Action（VLA）模型代表了机器人学习的最新发展趋势，其核心目标是将视觉感知、语言理解与动作控制统一于单一框架。得益于大规模视觉-语言基础模型（如 CLIP、SigLIP、LLaMA）[1, 2, 3] 的预训练能力，VLA 能够在多任务和跨模态环境中展现出优越的泛化性能。近年来，闭源模型 RT [4, 5] 系列展示了将互联网规模知识迁移至机器人控制的可行性，而开源模型 OpenVLA [6] 则推动了研究的可复现性与普及化。更进一步，Embodied Chain-of-Thought (ECoT)[7] 将显式推理机制引入机器人策略，使模型在泛化性、可解释性和鲁棒性方面实现显著提升。

本文系统回顾了 VLA 的发展历程，重点介绍了 VLA 的通用架构与训练范式，分析了 OpenVLA 和 ECoT 的设计与贡献，并对比它们在开放性、可扩展性与推理能力方面的差异。在此基础上，本文总结了当前 VLA 研究仍面临的挑战，并对 VLA 的未来研究方向做出了展望。

VLA 的快速发展揭示了具身智能迈向通用性与可解释性的潜力，为构建能够理解自然语言、适应新环境并安全与人类协作的机器人奠定了基础。

1. 简介

机器人学习旨在赋予机器在非结构化现实环境中执行多样化任务的能力。传统方法，如模仿学习与强化学习，虽在特定狭窄领域表现出一定效果，但通常难以超越训练数据分布实现泛化。虽然现有针对单一技能或语言指令训练的策略能够在物体位置等新的初始条件下外推行为 [8, 9]，但他们在面对场景干扰物或新颖物体时缺乏鲁棒性 [10, 11], 并且难以执行未训练过的任务 [12, 9]

然而，近年来视觉-语言基础模型（如 CLIP、SigLIP 和 LLaMA）[1, 2, 3] 在跨任务和跨模态的任务中表现出出色的泛化能力。这类模型在大规模互联网多模态数据上预训练，能够学习丰富的语义表示，并以较少适应实现向新领域的迁移。受此启发，研究者将基础模型扩展至机器人领域，催生了 Vision-Language-Action（VLA）模型。

VLA 模型将感知、语言理解与动作生成统一于单一框架。模型在接收图像观测与自然语言指令后，可直接以 token 化形式预测机器人动作。该范式有望提升机器人的灵活性、可扩展性与易用性，并支持日常任务的自然语言交互。然而，数据稀缺、实时推理、安全性及可解释性仍为亟待解决的挑战。

本文综述了 VLA 模型的发展历程，重点介绍了 OpenVLA[6] 与 Embodied Chain-of-Thought（ECoT）[7] 的最新进展，并探讨了当前挑战及未来研究方向。

2. 发展历程

2.1. 早期探索

最初将视觉与语言与动作相连接的研究集中在 Vision-Language Navigation（VLN）以及模拟环境中的指令跟随任务。这些工作将自然语言作为强化学习或模仿策略的高级目标信号。在受限环境中，这些方法取得了一定成功，但在开放式、复杂环境中这些方法表现出泛化能力不足的问题。

2.2. 大规模机器人策略的出现

RT-1[4] 的引入成为一个重要转折点。RT-1 在超过 13 万条机器人实验数据上训练，采用 Transformer 架构将图像与语言直接映射为机器人动作。其后继模型 RT-2[5] 结合了互联网规模的视觉-语言数据与机器人示范，实现了对未见物体和任务的零样本泛化。然而，这些模型仍为封闭源，限制了可复现性和研究使用。

2.3. 面向开放与通用策略的发展

与此同时，诸如 Octo[13] 等工作汇聚了多机器人数据集 Open-X Embodiment，提出了能够控制多种机器人的通用策略。尽管取得一定进展，但 Octo 等策略在抗干扰能力方面仍不及封闭源的 VLA 模型，如 RT-2-X。这种差距进一步促使了首个大规模开源 VLA 模型，OpenVLA[6] 的开发，以及随后出现的 ECoT[7]，旨在解决推理能力的局限性。

3. 通用 VLA 架构与训练范式

Vision-Language-Action Models（VLA 模型）是一种结合 VLM（Vision-Language Model）与机器人控制的模型，旨在将预训练的 VLM 直接用于生成机器人动作 [14]。与以往 VLM 进行规划或从零构建控制策略的方法不同，VLA 模型无需重新设计新的网络架构，而是通过将机器人动作序列编码为 token，并微调预训练的 VLM，使模型能够直接生成动作指令。这种端到端的感知-决策-执行一体化方式，不仅保留了 VLM 在视觉理解和语言推理方面的能力，还显著降低了训练成本，并提升了模型在不同任务和场景下的迁移与泛化能力。

目前的 VLA 可以从以下几个方面进行区分：模型结构和大小（如 action head 的设计, tokenize 的方法如 FAST）预训练与微调策略和数据集，输入和输出（2D vs. 3D | TraceVLA 输入 visual trace），不同的应用场景等 [1]。

VLA 模型通常由视觉编码器、语言模型骨干和动作离散化三部分组成。其使用 CNNs、ViT 以及更先进的编码器（如 SigLIP[3]）对输入图像进行 patch-level 的嵌入提取，并利用 LLMs，如 LLaMA 或 T5，充当序列建模的核心骨干。LLM 接收融合后的视觉-语言 token，并以自回归的方式预测后续 token，此外，连续的机器人动作被离散化为若干区间，并嵌入到 LLM 的词汇表中。这种方法使 VLA 能够将动作生成建模转化为语言建模。

训练过程包括：预训练，在互联网规模的多模态数据上进行预训练，提供语义层面的基础表征；微调，在大规模机器人演示数据集 [15, 16] 上进行微调，使感知与语言更好地对齐运动控制；

这种简洁而又可扩展的训练范式，使 VLA 模型能够充分利用计算机视觉和自然语言处理基础模型的快速进展，并直接连接到具身动作生成。

4. OpenVLA：首个开源的通用 VLA 模型

由于闭源模型透明度有限和现有模型与消费级硬件的适配性差，研究者认为未来机器人研发需要开源的、通用性的 VLA，来支持高效的微调和适配类似于当前开源语言模型 LLAMA[17]、Mistral[18]。为此，研究者推出了 OpenVLA，其由一个预训练的视觉条件语言模型主干组成，能够在多个粒度层面提取视觉特征，并在 Open-X Embodiment[15] 数据集上进行了微调。

对于 OpenVLA 而言，其基于 Prismatic-7B[19]，融合 DINOv2 和 SigLIP 的视觉特征，并采用 LLaMA-2-7B 作为语言建模主干。一个轻量投影器将视觉特征映射到语言模型空间，从而实现统一的动作 token 预测。

在微调的过程中，为了使 VLM 的语言模型主干能够预测机器人动作，作者通过将连续的机器人动作映射到语言模型的分词器使用的离散 token，将动作表示在 LLM 的输出空间中。对于每个动作维度，作者设置设置区间宽度，使其在训练数据中动作的第 1 和第 99 分位数之间均匀划分，但由于分词器 LLAMA[17] 的限制，研究人员只能选择简化的边界方案，按照 Brohan 等人 [20] 的方法，简单地用动作 token 覆盖 Llama 分词器词表中使用频率最低的 256 个 token。

Figure 1

最终的 OpenVLA 模型在一个由 64 个 A100 GPU 组成的集群上训练了 14 天，总计 21,500 A100 小时，使用的批量大小为 2048。由于数据多样性的提升和新模型组件的引入，OpenVLA 在 WidowX 和 Google Robot 两种机器人形态的 29 项评测任务中，绝对成功率比之前的业界领先 VLA——拥有 550 亿参数的 RT-2-X 模型 [20]——高出 16.5%

OpenVLA 有效解决了以往 VLA 模型闭源限制与算力需求过高的两大瓶颈。作为一个易于获取且性能强大的基线，成为推动 VLA 研究的关键基础.

5. Embodied Chain-of-Thought:VLA 模型的推理机制

虽然 OpenVLA 在可用性与性能上取得突破，但仍然遵循反应式范式。他们学习从观察到动作的直接映射，直接将输入映射为动作，缺乏显式推理。为解决这一问题，研究人员提出了将思维链推理引入 VLAs 的想法。通过生成中间步骤，LLMs 能够更好地映射问题不同部分之间的关系，并提出更准确的解决方案。

ECoT 策略在预测下一个机器人动作 (见图 2 , 右) 之前执行多个步骤的文本推理，训练 VLA 执行关于计划、子任务、运动和视觉特征 (如目标边界框和末端执行器位置) 的多步推理。

此外，研究者还设计了一个可扩展的 pipeline，用于在大型机器人数据集上生成 ECoT 的合成训练数据。使用强大的预训练的 open-vocubulary 目标检测器和大型语言模型生成推理监督。

实验表明，通过训练最先进的 VLA 在动作预测之前执行多个步骤的推理，可以大幅提高它们执行具有挑战性的泛化任务的能力。在没有任何额外的机器人训练数据的情况下，ECoT 将 OpenVLA 的绝对成功率提高了 28 %[7]。涉及到对新物体、场景、视点和

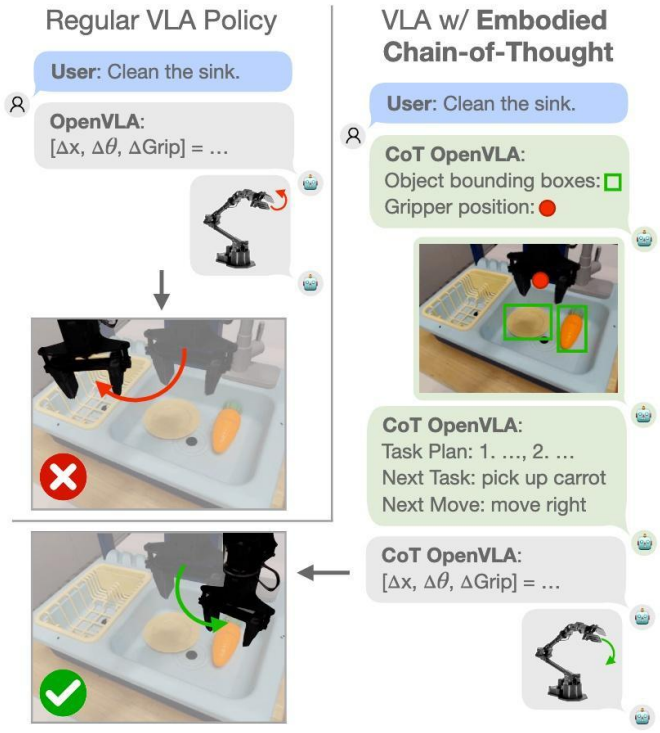

Figure 2

指令的泛化的一系列机器人操作任务。除了原始的性能改进外，实验表明，通过 ECoT，policy 失误更具有可解释性，允许人类通过自然语言反馈修改错误的推理链来轻松纠正 policy 行为。

6. 面临的挑战

OpenVLA 为社区提供了开放性与可用性，奠定了基线：ECoT 则在此基础上引入显式推理机制，使 VLA 在能执行任务的同时还能解释其行为。二者结合代表了一条发展路线：从开放、可扩展的 VLA 模型迈向可解释、具推理能力的具身智能体。

但是，在进展显著的背后，VLA 研究仍面临许多挑战：在训练过程中，多样化机器人演示数据采集昂贵，且数据间仍存在域间差异 [21]，限制了模型在真实环境中的表现；此外，ECoT 的推理过程增加了延迟，影响实时控制回路 [22]，提高了对计算的硬件要求；而且当前策略往往过拟合特定机器人形态，不同平台间的迁移能力有限 [23]，跨具身泛化能力仍亟待提升。

7. 未来展望

展望未来，VLA 研究可以从以下方面开展：

1. 与新兴开源 LLMs（如 LLaMA 3、Qwen-VL）结合，提升推理与知识迁移能力 [23]。
2. 利用仿真、合成环境以及自监督视频学习扩充稀缺的机器人数据。
3. 将推理扩展到视觉与语言之外，引入触觉、力反馈与本体感知，形成更全面的多模态推理链。
4. 通过自然语言交互让人类实时修正机器人推理链，提升学习效率与可控性。
5. 从实验室扩展到工业与家庭应用，确保模型在传感器噪声、光照变化、动态遮挡等复杂条件下仍具鲁棒性。

8. 总结

Vision-Language-Action 模型代表了机器人研究的范式转变，统一了多模态感知、自然语言理解与具身控制。OpenVLA 通过开放与可扩展性推动了研究，而 ECoT 在此基础上引入显式推理，显著提升了泛化能力与可解释性。二者共同勾勒出一条发展路径：从可复现、可扩展的 VLA 模型，到具备推理能力、可解释、并能够与人类交互的通用机器人智能体。尽管仍存在数据效率、实时推理、跨具身泛化等挑战，VLA 研究已展现出极大潜力。未来，随着开放模型、合成数据、多模态推理链和人机交互机制的逐步完善，具推理能力的通用机器人智能体将逐渐走向现实应用。

References

[1] 赵宇航. 神器 clip：连接文本和图像，打造可迁移的视觉模型. https://zhuanlan.zhihu.com/p/493489688, 2021. 知乎专栏, 访问日期: 2025-08-28.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Peter Welinder, Polina Kuznetsova, and et al. Learning transferable visual models from natural language supervision, 2021. Accessed: 2025-08-28.

[3] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023. Accessed: 2025-08-28.

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. Accessed: 2025-08-28.

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. Accessed: 2025-08-28.

[6] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. Accessed: 2025-08-28.

[7] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024. Accessed: 2025-08-28.

[8] Anonymous. Exploring the limits of vision-language-action manipulations in cross-task generalization. *arXiv preprint arXiv:2505.15660*, 2025. Accessed: 2025-08-28.

[9] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024. Accessed: 2025-08-28.

[10] Anonymous. Otter: A vision-language-action model. `https://ottervla.github.io`, 2024. Accessed: 2025-08-28.

[11] Anonymous. 3d cavla: Leveraging depth and 3d context to generalize vision language action models for unseen tasks. *arXiv preprint arXiv:2505.05800*, 2025. Accessed: 2025-08-28.

[12] Anonymous. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024. Accessed: 2025-08-28.

[13] Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Ria Doshi, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. Accessed: 2025-08-28.

[14] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, and Anthony Brohan. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 2023.

[15] Open x-embodiment: Robotic learning datasets and rt-x models. `https://robotics-transformer-x.github.io/`, 2023. Accessed: 2025-08-28.

[16] Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 1723–1736. PMLR, Nov 2023.

[17] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023. Accessed: 2025-08-29.

[18] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Timothée Lacroix, Thomas Wang, and William El Sayed. Mistral 7b: A 7b-parameter language model with state-of-the-art efficiency. arXiv preprint arXiv:2310.06825, 2023. Accessed: 2025-08-29.

[19] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 23123–23144. PMLR, Jul 2024.

[20] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Abhijit Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Carlos G. Arenas, Kuang-Huei Lee, Kanishka Gopalakrishnan, Kejun Han, Karol Hausman, Ariel Herzog, Jasmine Hsu, Brian Ichter, Alexander Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yao Lu, Igor Mordatch, Karl Pertsch, Kanishka Rao, Karl Reymann, Michael Ryoo, Sergio Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Henry Tran, Vincent Vanhoucke, Quan Vuong, Alexander Wahid, Stefan Welker, Peter Wohlhart, Jianyu Wu, Fei Xia, Ted Xiao, Peng Xu, Steve Xu, Tien-Ju Yu, Brian Zitkovich, Alejandro Escontrela, Kendra Byrne, and Erik Frey. Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818, 2023.

[21] Yi Zhao, Moo Jin Kim, Karl Pertsch, Dorsa Sadigh, and Sergey Levine. From grounding to manipulation: Case studies of foundation model integration in embodied robotic systems. *arXiv preprint arXiv:2505.15685*, 2025.

[22] Alan Burns, Robert I Davis, Sanjoy Baruah, and Giorgio C Buttazzo. The bottlenecks of ai: Challenges for embedded and real-time research in a data-centric age. *Real-Time Systems*, 2025.

[23] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Anybody: A benchmark suite for cross-embodiment manipulation. *arXiv preprint arXiv:2505.14986*, 2025.