人形机器人与 AI 大模型之 Robot +AI 的 Transformer 之旅

2024 年 11 月 07 日

➢ **从 Transformer 到多模态大模型的演进与应用。**Transformer 不仅在语言处理上广泛应用，还扩展至图像、视频、音频等多模态任务。诸如 Stable Diffusion、VideoPoet 和 MusicLM 等模型展现了其强大的生成能力，推动了多模态大模型（MLLM）的发展。

➢ **机器人现实世界至数据化的突破：RT-2、RoboCat 与 MimicGen。**RT-2 通过大规模的视觉-语言预训练，将视觉识别与低级机器人控制结合，实现了机器人在复杂任务和未见环境中的强大泛化能力。RoboCat 则基于 Gato 模型，展示了多任务和多具身平台上的自我迭代学习能力，能够快速适应新任务并生成跨任务策略。英伟达的 MimicGen 自动生成大量模仿学习数据，有效减少了人工干预，提升了机器人学习的效率。

➢ **特斯拉 FSD，端到端算法成为主流，数据为关键。**2020 年 FSD 引入 Transformer 模型，走向了数据驱动的模型范式，2024 年初 FSD V12 完全采用神经网络进行车辆控制，从机器视觉到驱动决策都将由神经网络进行控制。FSD V12 能够模拟人类驾驶决策，成为自动驾驶领域全新发展路径。

➢ **英伟达 Robocasa：具体智能关键节点，首次论证 real-sim-real。**通过升级模拟平台并构建模拟框架，基于厨房场景和原子任务、复合任务、真实世界三个场景收集行为数据集并进行结果评估。说明模拟器的丰富多样性以及视觉和物理真实性显著改善了模拟效果，实验结果首次论证了 real-sim-real 可行。

➢ **后续演绎：在机器人 real-sim-real 可行，证明存在 scaling law 的基础上，持续推荐可执行任务的泛化能力，迈向真正的 AGI 智能化：1)李飞飞 Rekep：一种针对机器人操作任务的新型空间和时间约束表示方法，提供了一种三任务闭环的解决方案。**通过关键点约束解构机器人行为，将操作行为分为多阶段，并构建子目标约束和路径约束，基于此提出一种三任务闭环的解决方案。同时，融入大型视觉模型和视觉-语言模型，利用 VLM 和 GPT-4o 生成 Rekep 约束，避免了手动指定 Rekep 的需要。**2）1x 世界模型：首证扩展定律，能通过大量学习理解周围环境。**通过大量的真实数据学习和模拟，机器人能够预测复杂的物体互动，理解周围环境，并灵活应对日常任务。1x 的进展首次在机器人上证明了扩展法则。**3) GR-2 的高效动作预测与泛化能力。**由字节跳动研究团队开发的第二代机器人大模型，凭借大规模视频预训练和多模态学习技术，展示了卓越的泛化能力与多任务通用性。**4）数字表亲：机器人训练法优化，以更低的成本获取更好的泛化能力。**在保留数字孪生优势的基础上，数字表亲表现出了更强的适应能力和鲁棒性，成功实现了从模拟到现实的零样本迁移，为机器人学习在复杂、多变的真实环境中的应用开辟了新的可能性。

➢ **投资建议：**1）关注算法训练中，需要使用的传感器公司，如视觉方案奥比中光，力学方案安培龙；2）关注同步受益的机器人本体公司，如总成方案三花智控、拓普集团；丝杆公司北特科技、五洲新春、贝斯特、双林股份、震裕科技等；3）关注其他产业链可延伸公司。

➢ **风险提示：**机器人算法迭代进步速度不及预期，人形机器人落地场景实际需求不及预期

推荐

维持评级

分析师 李哲
执业证书：S0100521110006
邮箱：lizhe_yj@mszq.com
分析师 罗松
执业证书：S0100521110010
邮箱：luosong@mszq.com

相关研究

1.一周解一惑系列：轮胎模具需求稳步增长，出海空间广阔-2024/10/20
2.一周解一惑系列：苹果海外供应链受阻，国内有望新设产线-2024/10/13
3.一周解一惑：AI 驱动+产业转移，PCB 曝光设备受益-2024/09/30
4.一周解一惑系列：复盘日本拖拉机历史，大拖占比提升大势所趋-2024/09/21
5.扬帆系列：船舶需求分析（二）：干散货船需求与测算-2024/09/18

目录

1 Transformer 模型的演进：从语言翻译到多模态智能的前沿探索

1.1 开篇：Robot +AI 的核心时间线与关键节点结论

下图是机器人和 transformer 模型结合的重点时间线及关键节点突破。

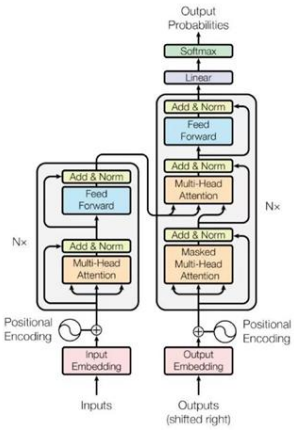
图1：Robot +AI 的核心时间线与关键节点

资料来源：Anthony Brohan《RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control》，Ajay Mandlekar《MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations》，Konstantinos Bousmalis《RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation》，tesla，Tianyuan Dai《ACDC: Automated Creation of Digital Cousins for Robust Policy Learning》，Jack Monas《1x world model》，Chi-Lam Cheang《GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation》，Soroush Nasiriany《RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots》，Wenlong Huang, Li Fei-Fei《ReKep: Spatio-Temporal Reasoning of Relational Keypoint Constraints for Robotic Manipulation》，民生证券研究院

1.2 Transformer 网络架构的提出

2017 年，Google 的 Brain 团队发布了一篇文章“Attention Is All You Need”，这篇文章中提出了 Transformer 网络结构。其一开始的提出是为了解决翻译问题，仅仅依赖于注意力机制就可处理序列数据，从而摈弃了 RNN 或 CNN。这个新的网络结构，刷爆了各大翻译任务，同时创造了多项新的记录（英-德的翻译任务，相比之前的最好记录提高了 2 个 BLEU 值）。而且，该模型的训练耗时短，并且对大数据或者有限数据集均有良好表现。

图2：Transformer 核心架构

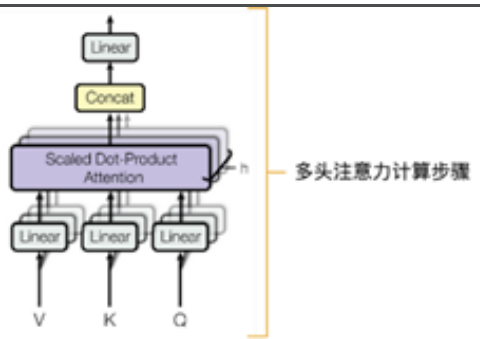


资料来源：Ashish Vaswani, Noam Shazeer《Attention Is All You Need》，民生证券研究院

模型的核心架构的示意图如上。Transformer 模型的架构就是一个 seq2seq 架构，由多个 Encoder Decoder 堆叠而成。在此示意图中，Encoder 和 Decoder 都包含 6 个 block。Transformer 将所有的单词向量化，通过矩阵编译的方法开始翻译以及预测，在翻译上一个词的同时对后续的单词进行预测，达到语句通顺的效果。其实际上是一个编码器-解码器结构，其中编码器将原始语言的句子作为输入并生成基于注意力的表征，而解码器关注编码信息并以回归方式生成翻译的句子，和之前的 RNN 相同。不同的是，Transformer 模型引入了注意力机制和残差链接，也就是所谓“Attention Is All You Need”，最终输出结果。

Transformer 的意义体现在它的长距离依赖关系处理和并行计算，而这两点都离不开其提出的自注意力机制。首先，Transformer 引入的自注意力机制能够有效捕捉序列信息中长距离依赖关系，相比于以往的 RNNs，它在处理长序列时的表现更好。而自注意力机制的另一个特点时允许模型并行计算，无需 RNN 一样 t 步骤的计算必须依赖 t-1 步骤的结果，因此 Transformer 结构让模型的计算效率更高，加速训练和推理速度。

图3：自注意力机制示意图



资料来源：Ashish Vaswani, Noam Shazeer《Attention Is All You Need》，民生证券研究院

1.3 语言、图片、视频大模型的出现

语言，图片，视频大模型以大语言模型为基础，将强大的大语言模型作为大脑来执行多模态任务。但 LLM 只能理解离散文本，在处理多模态信息时不具有通用性。另一方面，大型视觉基础模型在感知方面进展迅速，但推理方面发展缓慢。这两者的优缺点形成了巧妙的互补。

由于上述不同点中的互补性，单模态 LLM 和视觉模型同时朝着彼此运行，结合上部分的图像、视频和音频等等模态，最终带来了 MLLM 的新领域。形式上，它指的是基于 LLM 的模型，该模型能够接收多模态信息并对其进行推理。从发展人工通用智能的角度来看，MLLM 可能比 LLM 向前迈出一步。MLLM 更加符合人类感知世界的方式，提供了更用户友好的界面（可以多模态输入），是一个更全面的任务解决者，不仅仅局限于 NLP 任务。

图4：MLLM 的模型结构

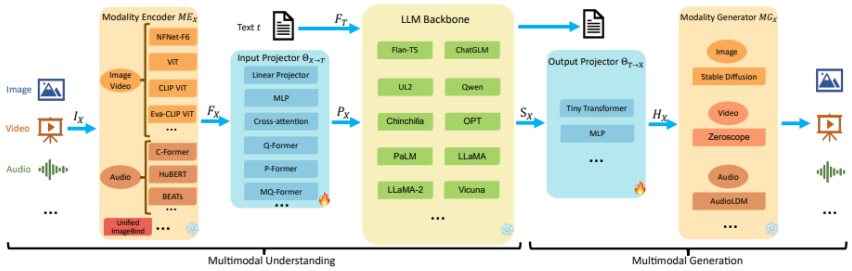


Figure 2: The general model architecture of MM-LLMs and the implementation choices for each component.

资料来源：Duzhen Zhang《MM-LLMs: Recent Advances in MultiModal Large Language Models》，民生证券研究院

上图包含了通用多模态模型结构的五个组件部分，以及每个组件部分的常用选择。

Modality Encoder：负责将不同模态的输入数据编码为模型可理解的表示，目前技术可以实现输入图片、视频、音频文件，对于图像而言，可能涉及到将像素数据转换成一个特征向量，该向量捕捉了图像中的重要信息；

Input Projector：将不同模态的输入数据映射到共享的语义空间，这意味着无论输入数据的形式如何，它们都会被转换成一个统一的格式，以便模型可以在一个统一的框架中处理它们；

LLMS：大型语言模型，用于处理文本数据，可以将不同模态的信息对齐到一个共同的语义空间中，整合由前面两个部分转换后输入的信息，融合后再生成一个统一的、丰富的语义表示，可能是相应的指导性文本或脚本，与专门的生成模型协同工作，实现高质量的图片和音频生成；

Output Projector：将模型生成的输出映射回原始模态的空间，如果模型的输出是文本，那么输出投影器将确保生成的文本与输入数据的语义空间相匹配；

Modality Generator：根据输入数据生成对应的输出数据，将模型的内部表

示转换成最终的输出形式，如生成图像、文本或音频。

多模态理解主要是前三个部分。（模态对齐）训练期间，encoder，LLM Backbone 和 generator 一般保持冻结。主要优化输出和输出的 projector。由于 Projector 是轻量级的模块，MM-LLMs 中可以训练的参数比例和总参数相比非常小（2%左右），模型的总体参数规模取决于 LLM 部分。由此，Transformer 模型随着 LLM 的广泛应用而成为了目前多模态大模型的核心思想和目前较为先进的网络架构。

截至 2024 年 10 月，中国移动在多模态大模型领域取得了显著进展，其九天善智多模态基座大模型表现尤为突出。该模型可以处理长文本的智能化解析，全双工语音交互，拥有高质量的视频与图像处理能力，可以对结构化数据做深度洞察。

1.4 多模态、跨模态大模型的难点

其一是异质化数据的处理与整合存在困难： 多模态大模型中，由于输入输出的数据具有多样性，面临的主要问题包括数据的异质性导致的表示难题、不同模态间的数据转换挑战、确定模态间元素联系的对齐问题、多模态信息的有效融合难点，以及如何在不同模态间进行知识迁移的协同学习挑战。需要综合应用多元化多样化的模型对其进行处理，将各个异质性的数据再整合规划，才能真正读懂要求，输出数据.

其二是训练过程挑战重重：获取跨多个模态的充足数据可能非常困难和昂贵，且数据可能会偏向于某些模态，导致模型产生偏见，从而导致模型偏向于数据量更多或特征更强的模态，导致模型产生偏见；同时由于特定于模态的编码器通常分别训练，他们声称的表示是存在差异的，对投影/对齐模块的有效学习过于依赖。

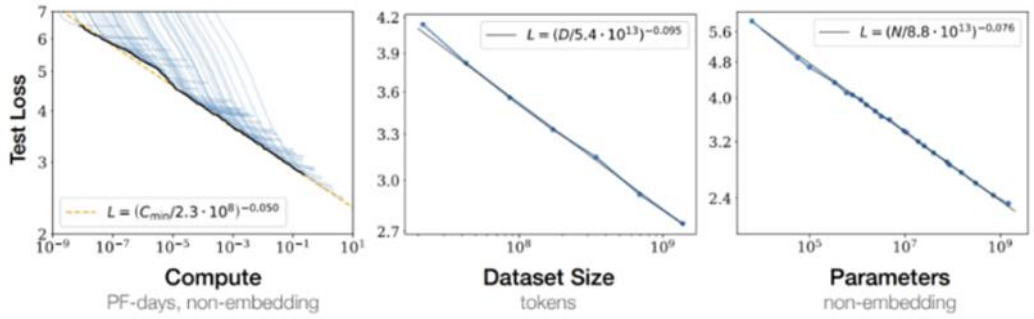
1.5 Scaling Law 的存在

在这其中，值得关注的是语言模型的 scaling law。大模型的 Scaling Law 是 OpenAI 在 2020 年提出的概念，可以概括为“预测即压缩、压缩即泛化、泛化即智能”将大语言模型用在别的领域之后，从计算理论的角度，联合压缩多模态数据理应获得比单模态更好的理论最优压缩器。

对于所有的模态来说，他们都必须要服从的 scaling law 是，随着数据规模的提升，模型的表现也会随之提升，如果法则正确，那么要想使得模型更好，只需要搭建好算法和框架，不断收集数据就可以了。一旦证明 scaling law 的存在和有效性，就可以预测模型性能与规模的关系，投入恰当规模的数据集，使得计算资源可以更高效的应用。多模态模型会变得更加可预测和可计算，其不确定性就极大的降

低了。

图5：Scaling Law 的效果图示



资料来源：Jared Kaplan《Scaling Laws for Neural Language Models》，民生证券研究院

在此基础上，本文想要按时间线和核心 milestone 贡献，来帮助大家拆解最近 1 年时间，robot 的 transformer 结合之旅是怎么演进的，从而去探讨真正前沿的，以 transformer 为基础的模型到底会去往何处。

2 机器人现实世界至数据化的突破：RT-2、RoboCat 与 MimicGen

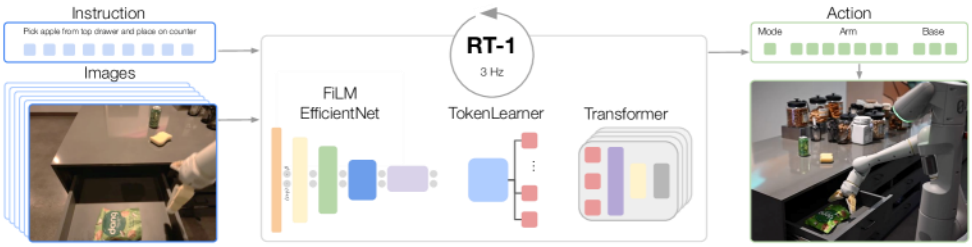
2.1 谷歌 RT-2：具身智能学习

2.1.1 提出的问题与研究意义

大语言模型可以实现流畅的文本生成、问题解决、创意写作以及代码生成，视觉-语言模型（VLM）则能够实现开放词汇的视觉识别。以上能力对于现实环境中的通用型机器人非常有用，然而它们如何获得这些能力还是未知。**如何将大型预训练的视觉-语言模型直接集成到低级机器人控制中，以促进泛化并实现紧急语义推理，成为了机器人下一步发展的方向。**

Google 提出的 RobotTransformer(RT)系列使用了更大规模的语言模型和更多的具身智能任务数据，在大量具身智能任务中获得较好效果。其中 RT-1 算法使用预训练的 EfficientNet-B3 网络初始化，以机器人状态和历史图片作为输入，通过 EfficientNet 特征提取后直接输出动作。

图6：RT-1 结构概览



资料来源：Anthony Brohan《RT-1: ROBOTICS TRANSFORMER FOR REAL-WORLD CONTROL AT SCALE》，民生证券研究院

RT-1 将机器人动作的每个维度进行均匀离散化，并将动作词元化，然后使用监督学习的损失进行训练。为了使视觉 - 语言模型能够控制机器人，还差对动作控制这一步。该研究采用了非常简单的方法：他们将机器人动作表示为另一种语言，即文本 token，并与 Web 规模的视觉-语言数据集一起进行训练。

图7：机器人动作数字 token 化

对机器人的动作编码基于 Brohan 等人为 RT-1 模型提出的离散化方法。

如下图所示，该研究将机器人动作表示为文本字符串，这种字符串可以是机器人动作 token 编号的序列，例如「1 128 91 241 5 101 127 217」。

资料来源：Anthony Brohan《RT-1: ROBOTICS TRANSFORMER FOR REAL-WORLD CONTROL AT SCALE》，民生证券研究院

RT-2 在机器人任务上展示了更强的泛化能力，以及对超出其接触的机器人数据之外的语义和视觉的理解。RT-2 在 RoboticTransformer1(RT-1)的基础上进行，直接训练视觉-语言模型以实现开放词汇视觉问答和视觉对话，输出低级机器人动作，同时解决其他互联网规模的视觉语言任务。相较于 RT-1，RT-2 模型在机器人理解新任务，并通过执行基本推理来响应用户命令，例如推理物体类别或高级描述等方面具有更大的优势。

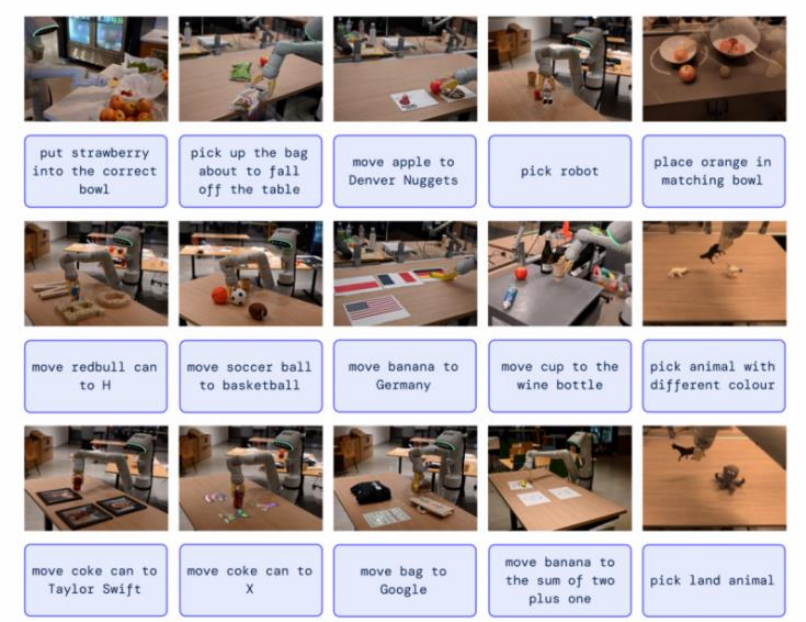
2.1.2 核心方法与进步：以预训练为基础升级泛化能力

与 RT-1 模型的泛化能力相比，RT-2 的目标是训练机器人从观测到动作的端到端模型，并且从大规模视觉-语言模型预训练模型中学习泛化知识。最终，Google 提出一个在机器人轨迹数据和互联网级别的视觉语言任务联合微调视觉-语言模型的学习方式。这类学习方法产生的模型被称为视觉-语言-动作(VLA)模型，具有泛化到新对象的能力、解释命令的能力以及根据用户指令思维推理的能力。

RT-2 算法整体使用大规模预训练的视觉-语言模型结构，模型参数可以达到 55B 的参数量，远超 RT-1 的参数规模，同时利用大规模预训练视觉-语言模型模型中编码的丰富视觉问答知识来帮助具身模型的训练。RT-2 将输出的动作进行和 RT-1 相同的离散化操作后将词元加入视觉-语言模型原先的词表中，可以把动作词元视为另外一种语言进行处理，无需改变原有视觉-语言模型结构设计。由于 RT-2 已经在海量的视觉问答任务中进行预训练，在对图片和任务指令的理解上有更加丰富的经验，在任务集合上具有更强的泛化能力。

RT-2 能够运用其大规模预训练的视觉问答经验进行泛化，在现实世界的任务中进行推广，实现推理、理解和识别。例如在下图的拾取、移动、放置等具体任务中，智能体能够精准识别任务需求并且以过往训练经验为基础准确地完成。

图8：RT-2 能够推广到各种需要推理、符号理解和人类识别的现实世界情况



资料来源：Anthony Brohan《RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control》，民生证券研究院

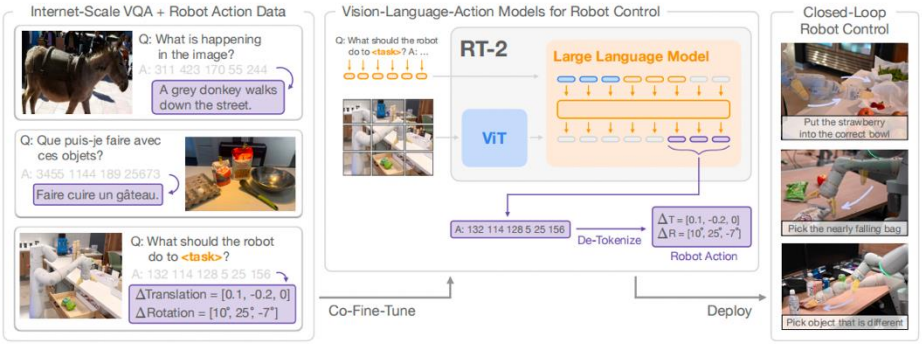
RT-2 的核心方法是采用视觉-语言-动作模型（VLA）与联合微调。具体步骤如下：

步骤一：RT-2 通过 Vision Transformer（ViT）提取图像特征，把动作 tokens 转化为语言 tokens，将相应动作转化为动作字符串（例如“1 128 91 241 5 101”）。在此过程中，机器人动作被离散化为多个参数（如位移和旋转），每个参数映射为预定义的 token。这些 token 被嵌入到模型的语言字典中，与自然语言 token 共用同一表示空间。

步骤二：RT-2 将任务指令和图像信息结合，通过 de-tokenize 转化为具体的机器人动作序列。此过程使用大语言模型（LLM）解析任务，像自然语言处理那样，动作模块使用 tokenizer 来处理这串 token 转成对应的机器人动作，将视觉信息和任务指令解码为具体的机器人动作序列（如平移和旋转参数），进而分析这串字符串对应的开始符、命令、停止符。

步骤三：在执行任务的过程中，模型同步实时进行**联合微调（Co-Fine-Tuning）：**机器人根据传感器和摄像头反馈的最新图像信息，判断任务执行的状态和完成情况。如果任务执行过程中出现误差或环境发生变化，模型会利用新的视觉数据重新规划动作，直至任务完成。总而言之，语言模型负责持续理解任务场景和需求，而动作模块根据视觉反馈实时调整操作，确保任务顺利完成。完成训练与微调后，RT-2 被部署到机器人系统中，并具备了在复杂环境下执行多任务的能力，实现**高效的闭环控制。**

图9：RT-2 全流程概览



资料来源：Anthony Brohan《RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control》，民生证券研究院

2.1.3 核心结论+未来进展

RT-2 展示了视觉-语言模型（VLMs）可以转变为强大的视觉-语言-动作（VLA）模型，通过结合 VLM 预训练和机器人数据，直接控制机器人。RT-2 基于 PaLM-E 和 PaLI-X 的两种 VLA 模型，提高了机器人策略的完成率，并且继承了视觉语言数据预训练的优势，具有更好的泛化能力和涌现能力。这不仅是对现有视觉-语言模型的有效改进，也展示了通用型机器人的发展前景。未来的机器人能够进行推理、解决问题，并进行高级规划和低级指令控制，在现实世界中执行大量多样化的任务。

RT-2 也具有局限性。该模型对于泛化能力的强化并没有提高机器人执行新动作的能力，智能体知识学会了以新的方式部署学习到的技能。同时，由于高频控制的设置应用场景，实时推断可能成为主要瓶颈。**未来工作的方向主要集中于如何通过新的数据收集范式（如人类视频）获得新技能，同时开发出更多的开元模型以支撑高速率和低成本的运作。**

2.2 英伟达 MimicGen：自动化数据生成系统

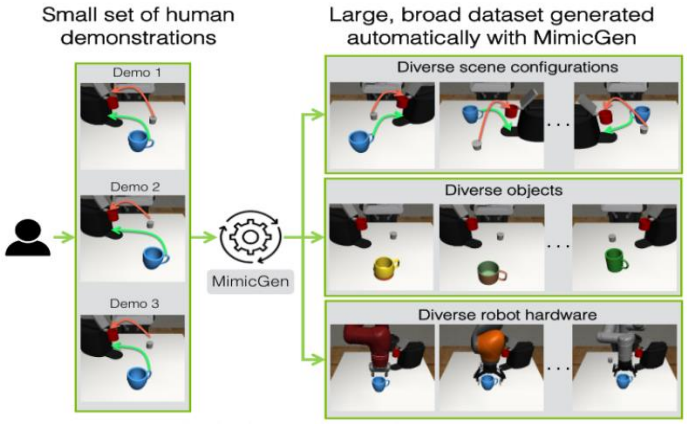
2.2.1 MimicGen：用于大规模机器人学习的数据生成系统

MimicGen 是一个用于大规模机器人学习的数据生成系统，目的是解决机器人学习过程中人工数据收集成本高、时间耗费大的问题。当前基于模仿学习的机器人研究依赖大量的人工演示数据来训练模型，但这些数据的收集非常昂贵。MimicGen 提出了从少量人类演示数据中自动生成大规模、多样化的演示数据集的系统。该系统通过将人类演示数据适应于新场景，生成多达 50,000 条演示数据，覆盖 18 项任务，从而显著降低了人工数据收集的需求。

这一方法能够加速机器人学习的进展，使得机器人能够在复杂场景中表现出

更强的泛化能力，尤其是在长时间任务和高精度任务（如多部件装配、咖啡准备）中表现出色。研究结果表明，利用 MimicGen 生成的数据进行模仿学习能够取得与传统人工数据收集相媲美的效果。

图10：MimicGen 从原始人类演示数据到生成的广泛数据集的过程



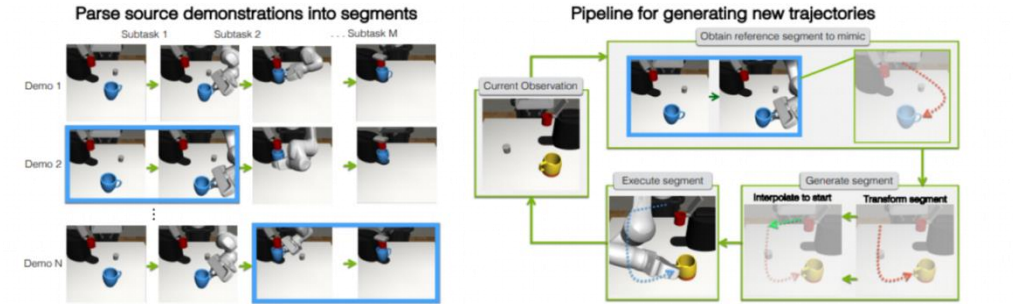
资料来源：Ajay Mandlekar《MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations》，民生证券研究院

2.2.2 核心方法与进步：数据分割与重组

MimicGen 的设计来源于模仿学习与数据增强两个技术背景。模仿学习是一种通过观察人类示范来训练机器人的方法。MimicGen 利用这一理念，通过生成多样化的示范来扩展模仿学习的应用范围。数据增强技术被广泛应用于提高模型的泛化能力。通过对现有数据进行变换或修改来生成新训练样本的技术，旨在提高模型的泛化能力和鲁棒性。常见的数据增强方法包括旋转、缩放、平移等，这些变换可以在不改变数据标签的情况下生成新的样本。

MimicGen 的核心方法是数据分割与重组。将少量人类演示数据分割成以物体为中心的子任务，然后在新的场景中通过空间变换和轨迹生成，自动生成新的演示数据。传统方法中，数据生成通常基于静态场景的回放，或通过复杂的模拟器进行大量数据收集。而 MimicGen 的创新点在于，它提出了一种简单但有效的策略，通过“对象中心片段”的变换和拼接，将少量的人类演示数据转化为大规模的多样化数据。这种方法可以直接融入现有的模仿学习管道中，适用于各种长时间、高精度的任务，并且能够生成比单纯回放方法更加多样和有效的数据。

图11：MimicGen 数据分割与重组示意图

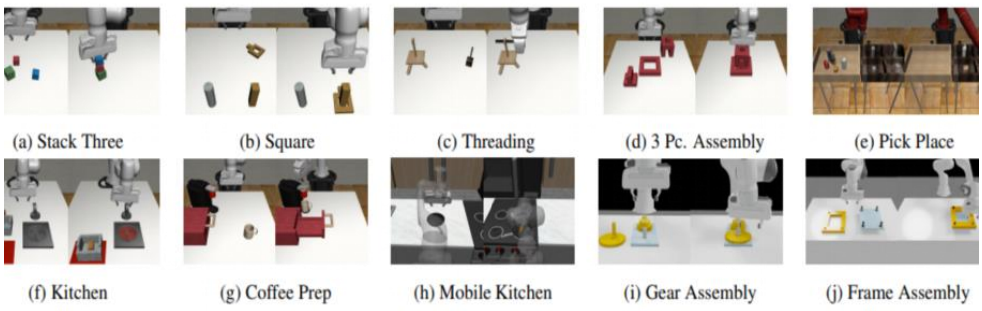


资料来源：Ajay Mandlekar《MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations》，民生证券研究院

2.2.3 核心结论：主要测试任务成功率大幅提升

通过对比使用 MimicGen 生成的数据集与传统人类示范数据集的结果可以得出，机器人在使用 MimicGen 生成的数据集后成功率显著上升。研究团队通过对 MimicGen 的实验，评估了其在不同任务中的表现，具体测验任务主要包括 Stack Three（堆叠三个物体）、Square（方形物体插入和对齐）、Threading（机器人在穿线或穿孔时的精细操作能力）、Kitchen（长时间多步骤任务）等十项。

图12：MimicGen 主要测试任务



资料来源：Ajay Mandlekar《MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations》，民生证券研究院

结果显示使用 MimicGen 后机器人成功率显著提升，例如“Square”任务的成功率从11.3%提升至90.7%，“Threading”任务的成功率从19.3%提升至98.0%。

图13：MimicGen 主要测试任务结果

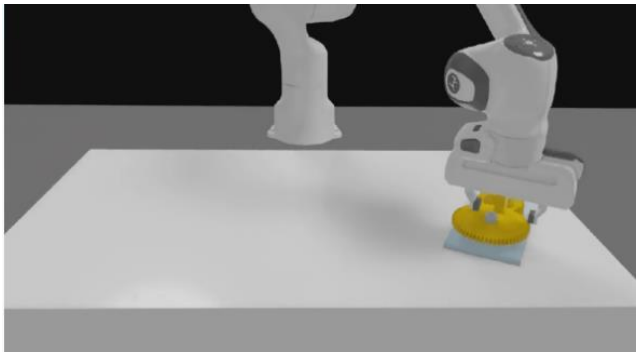
Task	Source	D_0	D_1	D_2
Stack	26.0 ± 1.6	100.0 ± 0.0	99.3 ± 0.9	-
Stack Three	0.7 ± 0.9	92.7 ± 1.9	86.7 ± 3.4	-
Square	11.3 ± 0.9	90.7 ± 1.9	73.3 ± 3.4	49.3 ± 2.5
Threading	19.3 ± 3.4	98.0 ± 1.6	60.7 ± 2.5	38.0 ± 3.3
Coffee	74.0 ± 4.3	100.0 ± 0.0	90.7 ± 2.5	77.3 ± 0.9
Three Pc. Assembly	1.3 ± 0.9	82.0 ± 1.6	62.7 ± 2.5	13.3 ± 3.8
Hammer Cleanup	59.3 ± 5.7	100.0 ± 0.0	62.7 ± 4.7	-
Mug Cleanup	12.7 ± 2.5	80.0 ± 4.9	64.0 ± 3.3	-
Kitchen	54.7 ± 8.4	100.0 ± 0.0	76.0 ± 4.3	-
Nut Assembly	0.0 ± 0.0	53.3 ± 1.9	-	-
Pick Place	0.0 ± 0.0	50.7 ± 6.6	-	-
Coffee Preparation	12.7 ± 3.4	97.3 ± 0.9	42.0 ± 0.0	-
Mobile Kitchen	2.0 ± 0.0	46.7 ± 18.4	-	-
Nut-and-Bolt Assembly	8.7 ± 2.5	92.7 ± 2.5	81.3 ± 8.2	72.7 ± 4.1
Gear Assembly	14.7 ± 5.2	98.7 ± 1.9	74.0 ± 2.8	56.7 ± 1.9
Frame Assembly	10.7 ± 6.8	82.0 ± 4.3	68.7 ± 3.4	36.7 ± 2.5

资料来源：Ajay Mandlekar《MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations》，民生证券研究院

2.2.4 MimicGen 未来潜力：生成训练数据，减少人工干预

MimicGen 在机器人系统（尤其是机械臂）中的应用潜力巨大。通过利用少量人类演示 (少于 200 个)，MimicGen 可自动生成超过 50,000 个覆盖 18 种任务的高质量数据，有效减少人工干预，提升生产效率。其灵活性使其能够适应不同机器人硬件和复杂操作环境，为工业自动化、医疗和服务机器人等领域提供广泛的应用前景。MimicGen 的核心优势包括：显著提升任务表现、良好的广泛适应性、跨物体和硬件适用性、适用于复杂移动操作任务、模拟器无关，精度表现卓越、支持非专家演示。

图14：MimicGen 操作机械臂完成毫米级精度接触任务示意图

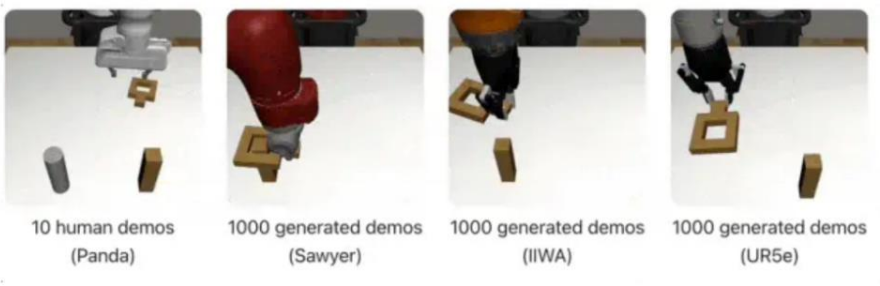


资料来源：Ajay Mandlekar《MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations》，民生证券研究院

MimicGen 依赖于任务开始时已知的对象位姿和操作步骤，这在完全未知或动态环境中存在局限性。此外，仅通过任务成功与否来筛选生成数据，可能导致数据集存在偏差，影响模型泛化能力。其应用场景主要限于准静态任务，并假设新对

象与已有对象同类，限制了其在动态环境和异构对象上的推广能力。未来研究应进一步提升系统对复杂场景的理解和分割能力，减少对人类参与的依赖。扩展 MimicGen 在更多物体类别、机器人硬件和任务类型中的应用能力。

图15：MimicGen 能够适应不同的机械臂



资料来源：Ajay Mandlekar《MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations》，民生证券研究院

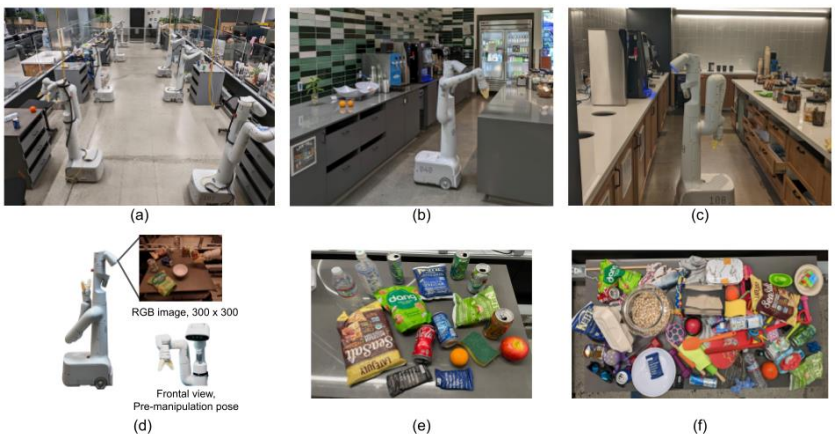
2.3 谷歌 RoboCat：多任务具身智能

2.3.1 RoboCat：多任务、多具身通才智能体

在机器人领域，如何大规模利用异构机器人数据仍然是机器人领域的难题，大多数现实中的机器人学习研究集中于一次开发一个任务的智体。在机器人技术领域，近期研究专注于通过训练有语言条件的 Transformer 策略来解决具有相同观测和动作空间的多个简单、视觉多样化的任务，从而弥合大型预训练语言模型和视觉基础操作之间的差距。

Google 曾经提出 RobotTransformer，采集了移动机器人完成日常任务的轨迹片段，构成了真实移动机器人的专家数据集，包含了 700 多个任务，如移动物体、拉开抽屉、开罐子等，学习到的策略在新的任务指令上有一定的泛化能力。

图16：RT 数据收集和评估场景

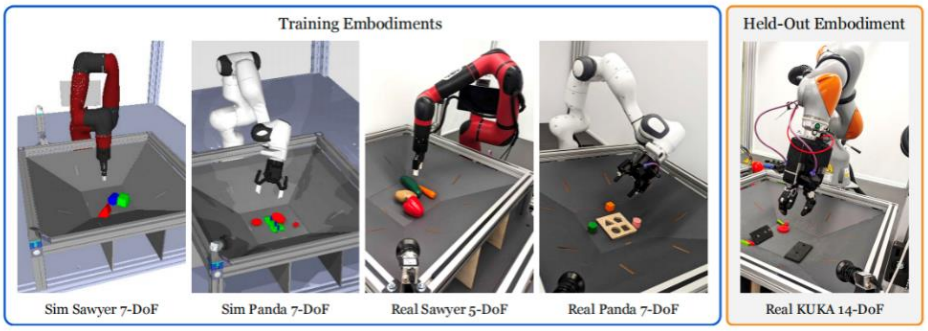


资料来源：Anthony Brohan《RT1：ROBOTICS TRANSFORMER FOR REAL-WORLD CONTROL AT SCALE》，民生证券研究院

RoboCat 在 Gato 模型的基础上进行了改进，是一项受视觉和语言基础模型最新进展启发而提出的自我改进型多任务、多具身通才智能体。RoboCat 使用了跨实体、跨任务的具身模仿学习框架，在 VQ-GAN 对视觉输入词元化之后，使用标准的 DT 回归损失根据历史的状态、观测、目标信息对未来的智能体动作和观测进行预测。同时，RoboCat 不断提升智能体的能力。在新任务上，RoboCat 仅需 100~1000 个示教样本就能完成快速策略泛化。

通过 RoboCat，Google 能成功展示其在新任务和不同机器人平台上的泛化能力，以及通过后续迭代利用大模型辅助具身智能数据生成，从而为构建一个自主改进循环提供基本的构建板块。随着训练数据的增长和多样化，RoboCat 不仅表现出了跨任务迁移的迹象，也能更有效地适应新任务。

图17：RoboCat 支持多种机器人具身和控制模式

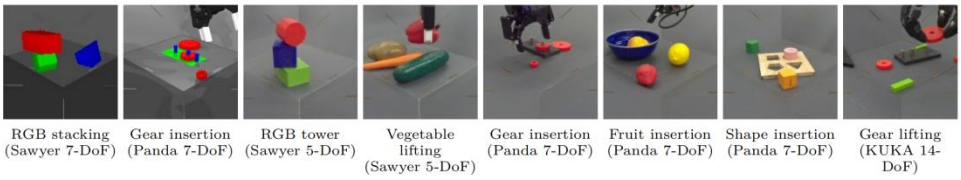


资料来源：Konstantinos Bousmalis《RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation》，民生证券研究院

2.3.2 RoboCat：以数据集为基础实现任务的快速适应

RoboCat 的最终目标是创建一个能够通过大量机器人情景经验进行训练的基础智能体，使其能够通过微调快速适应广泛的新下游任务。为了实现这一目标，RoboCat 拥有一个非常丰富的多样化操控行为数据集并在此基础上进行训练。RoboCat 基于 Gato 架构，使用在广泛图像集上预训练过的 VQ-GAN 编码器（Esser,2021），在涵盖多个领域和具身的广泛数据集上进行训练，通过视觉目标条件来指定任务。这种编码器的选择使得训练和迭代更加快速，这种训练方式也具有理想的自动事后目标生成属性，即轨迹中的任何图像都可以被标记为所有导致它的所有时间步骤的有效“后见目标”（Andrychowicz,2017）。这意味着现有数据中的后见目标可以在没有额外人为监督的情况下提取。此外，视觉目标提供了一个直观的界面，用于指示机器人应该执行什么任务。

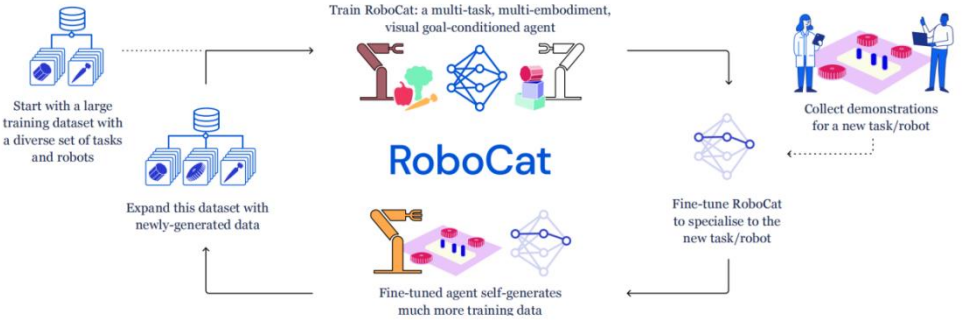
图18：目标图像示例：图 1、2 为虚拟环境，图 3-8 为现实世界



资料来源：Konstantinos Bousmalis《RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation》，民生证券研究院

RoboCat 能进行自我微调和迭代。首先智能体将在初始使用多样化的训练集进行训练，可以通过 100-1000 次演示微调以适应新任务，然后部署在真实机器人上，生成更多数据。其次，将生成轨迹添加进入下一次迭代的训练数据集中，从而提高跨任务的性能。RoboCat 的自我改进过程如图所示：主要以架构和预训练、微调和自我改进、真实世界部署作为全流程。

图19：RoboCat 自我改进进程



资料来源：Konstantinos Bousmalis《RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation》，民生证券研究院

2.3.3 机器人未来发展展望

未来机器人的研究工作将着眼于更灵活的多模态任务规划。首先是将现有的公开可获取的数据集与注释语言相结合，以语言为媒介的任务规划和视觉目标相辅相成，得以实现对不同任务的更精准定位。此外，尽管当前研究主要关注视觉目标条件反射以及基于视觉-前馈模型（VFM）的基线研究，但仍在图像推理方面表现出色；同时，语言条件反射和 LLM/VLM 基线研究可能提供更好的时间推理能力。

3 特斯拉 FSD：端到端算法成为研究主流，数据集成为关键

3.1 FSD V12：全新的端到端自动驾驶

FSD 全称 Full Self-Driving（完全自动驾驶），是特斯拉研发的自动化辅助驾驶系统，目标是实现 L5 级别的自动驾驶。

图20：FSD V12（Supervised）虚拟界面显示

资料来源：Tesla，民生证券研究院

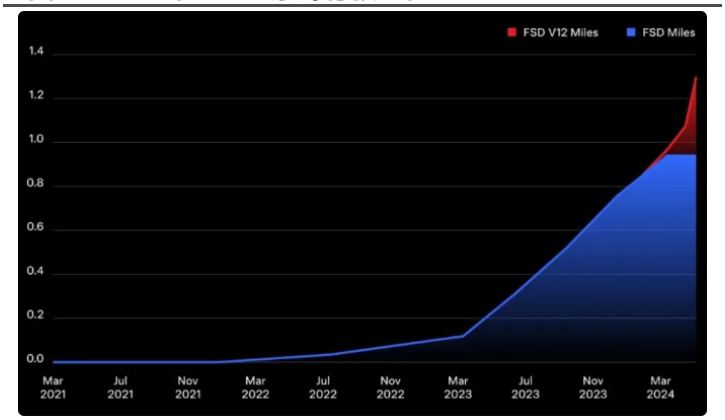
图21：自动驾驶的六个等级

资料来源：九章智驾，民生证券研究院

FSD V12（Supervised）是全新的“端到端自动驾驶”，模型架构发生了重大变化。据特斯拉 CEO 埃隆·马斯克表示，特斯拉 FSD V12（Supervised）需要人工干预的频率只有 FSD V11 的百分之一。FSD V12（Supervised）完全采用神经网络进行车辆控制，从机器视觉到驱动决策都将由神经网络进行控制。该神经网络由数百万个视频片段训练而成，取代了超过 30 万行的 C++代码。FSD V12（Supervised）减少了车机系统对代码的依赖，使其更加接近人类司机的决策过程。

根据特斯拉发布 2024Q2 的自动驾驶报告，自动驾驶大幅减少事故率：开启 Autopilot 的情况下，平均每行驶 1107.2 万公里（688 万英里）会发生一起事故，而未开启平均每行驶 233.3 万公里（145 万英里）会发生一起事故。

图22：FSD 和 V12 累计行驶里程



资料来源：Tesla，民生证券研究院

图23：每发生一次事故行驶的英里数

资料来源：特斯拉官网，民生证券研究院

3.2 FSD 的前世今生

早期特斯拉自动驾驶采用外部合作方式，合作厂商包括 Mobileye 和英伟达等。在 2019 年特斯拉步入自研时代，首次推出自研自动驾驶芯片 HW3.0。HW3.0 采用特斯拉全栈自研的 FSD 芯片。2020 年 10 月，特斯拉小范围推送 FSD Beta，对 Autopilot 基础架构进行了重大重写。2021 年 7 月，特斯拉开始推送 FSD Beta V9，该版本采用纯视觉自动驾驶方案，摒弃了传统的毫米波雷达和超声波雷达，是特斯拉在自动驾驶技术的重要发展节点。

图24：特斯拉自动驾驶主要发展历程

时间	软硬件版本	主要发展
2014年10月	Autopilot 1.0 Hardware 1.0	基于Mobileye的EyeQ3平台打造
2016年10月	Hardware2.0 Autopilot 2.0	配置Nvidia Drive PX2计算系统
2019年4月	Hardware3.0	搭载自研自动驾驶芯片FSD
2021年7月	FSD Beta(测试版) 9.0	确定纯视觉方案，取消毫米波雷达及超声波雷达
2024年1月	FSD 12.0（Supervised）	端到端神经网络
2024年2月	Hardware4.0	搭载Model Y算力提升5倍

资料来源：汽车财经，IT 之家，易车网，中国新闻周刊，新浪网，民生证券研究院

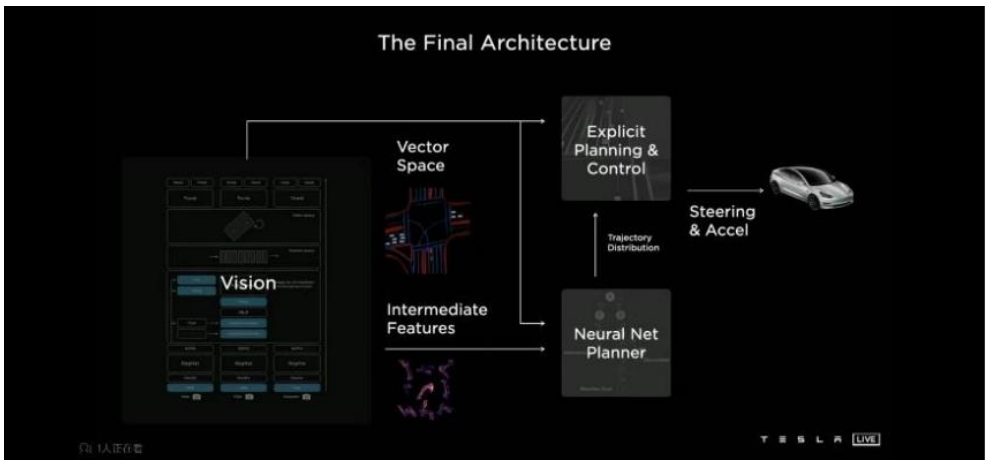
2024 年 1 月，特斯拉 FSD V12 正式向用户推送，将城市街道驾驶堆栈升级为端到端神经网络。2024 年 2 月，特斯拉 Model Y 迎来 HW4.0 自动辅助驾驶硬件升级，与 HW3.0 相比，HW4.0 算力提升 5 倍，在硬件设计上实现并行处理能力增强、内存管理优化和专用加速器集成等多项创新。从最初的辅助驾驶系统，到全栈自研自动驾驶技术，特斯拉持续引领智能驾驶技术发展浪潮。

3.3 FSD 架构变革：Transformer 模型的引入

复盘 FSD 历史，最重大的架构变革莫过于 2020 年引入 Transformer 模型（基于深度学习的神经网络），算法得以从重人工、规则驱动，转向重 AI，数据驱动。FSD 主要分为感知和规划模块，在两个模块中都运用到了 Transformer 模型，神经网络的介入使得端到端模型逐步实现。

2022 年特斯拉 FSD 感知模块即形成了 BEV +Transformer+Occupancy 神经网络架构。通过摄像头的图片输入，端到端输出汽车周围环境向量空间数据，为规划模块决策提供支持。特斯拉 FSD 规划模块在 2021 年引入基于神经网络的规划模块和蒙特卡洛树搜索，最终 FSD 规划模块由基于显性规则的规划模块和基于神经网络的规划模块构成。

图25：FSD 感知规划控制总体架构

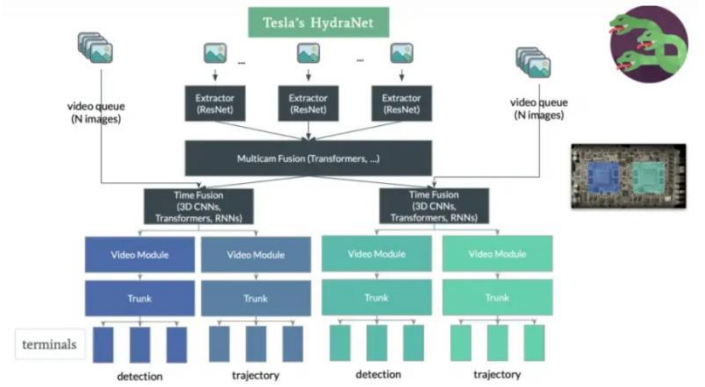


资料来源：特斯拉 2021 AI Day，民生证券研究院

HydraNets 是特斯拉开发的一种深度学习网络架构。这个网络的特点在于它能够将多个任务集成到一个网络中，例如车道线检测、行人检测与追踪、交通信号灯检测等，这些任务对于自动驾驶汽车来说至关重要。HydraNets 的核心在于其共享的主干网络，该主干网络通过分支成多个“头”，可以同时输出多个预测张量，每个“头”负责不同的任务或对象类别。

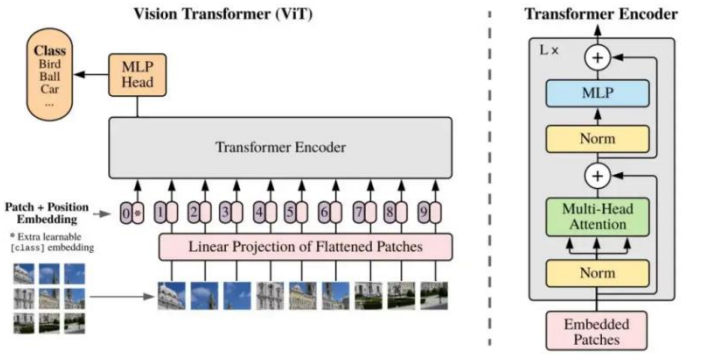
此外，这种架构的优势在于其能够有效地利用可用的计算资源，并且通过端到端的训练和推断，提高了处理不同视觉信息的效率。HydraNets 能够将来自多个摄像头的视觉内容转换为向量空间和道路特征，这对于构建车辆周围的综合视图至关重要。

图26：HydraNets 网络架构



资料来源：特斯拉 2022 CVPR，民生证券研究院

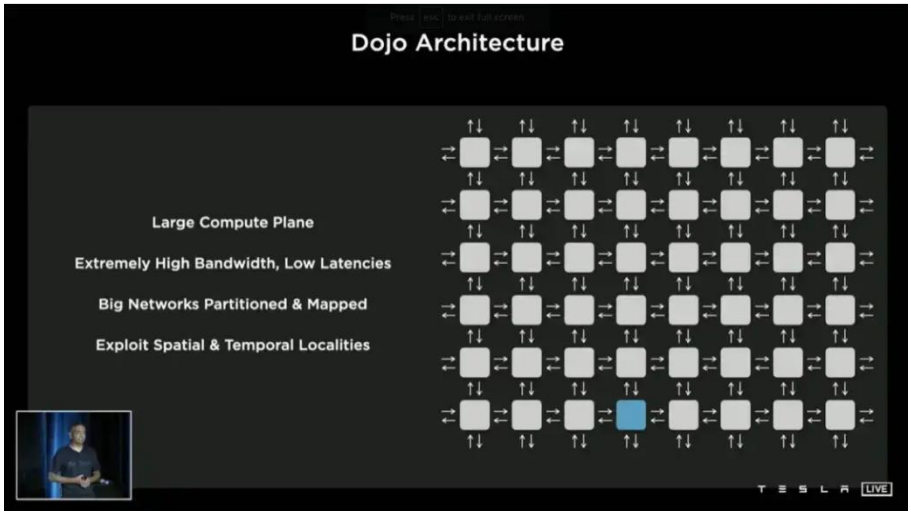
图27：视觉 Transformer 模型架构



资料来源：Alexey Dosovitskiy：《An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale》，民生证券研究院

Dojo 是特斯拉公司开发的一套高性能计算系统，用于处理和训练自动驾驶系统产生的海量数据。Project DOJO 的负责人 Ganesh Venkataramanan 表示，DOJO 是一种通过网络连接的分布式计算机架构，它具有高带宽、低延时等特点，将会使人工智能拥有更高速的学习能力，从而使 Autopilot 更加强大。

图28：Dojo 内核示例



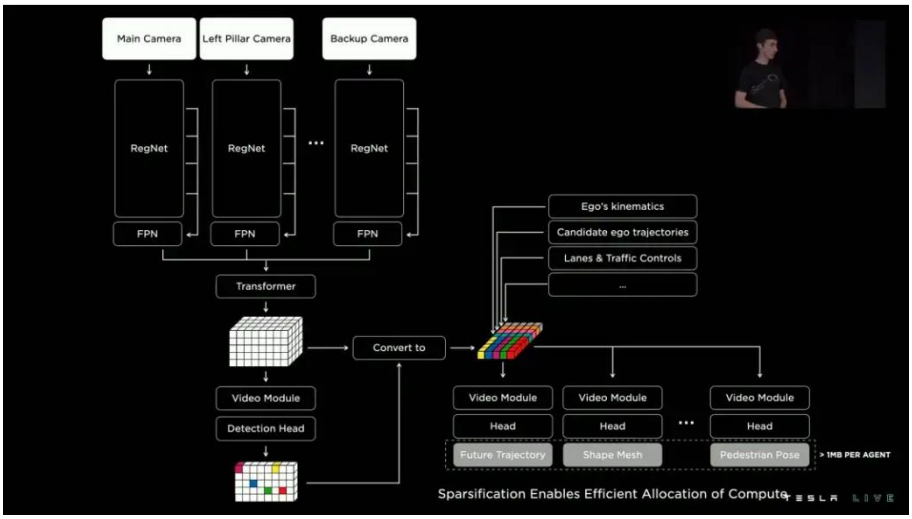
资料来源：特斯拉 2021 AI Day，民生证券研究院

3.4 FSD 端到端：感知决策一体化

FSD V12 为首个端到端自动驾驶系统，实现感知决策一体化。特斯拉 FSD v12 采用端到端大模型，消除了自动驾驶系统的感知和定位、决策和规划、控制和执行之间的断面，将三大模块合在一起，形成了一个大的神经网络，直接从原始传感器数据到车辆操控指令，简化了信息传递过程，因而减少了延迟和误差，提高了系统的敏捷性和准确性。FSD V12 能够模拟人类驾驶决策，成为自动驾驶领域全新发

展路径。FSD V12 也被称为“Baby AGI（婴儿版通用人工智能）”，旨在感知和理解现实世界的复杂性。

图29：Baby AGI 架构



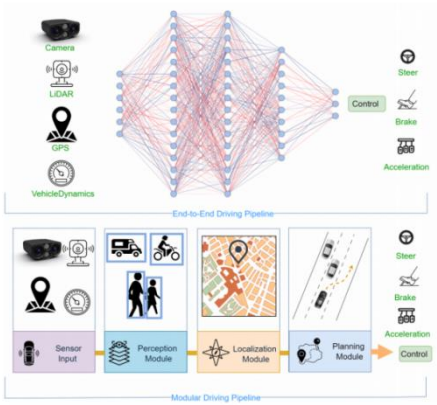
资料来源：特斯拉 2021 AI Day，民生证券研究院

4 端到端算法成为研究主流，数据集成为关键

4.1 端到端算法：直接连接数据输入与控制指令输出

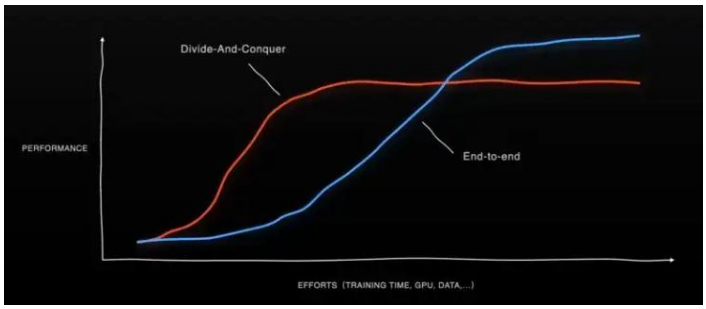
模块化自动驾驶分为传感器数据输入、感知模块、定位模块、规划模块和控制指令输出五部分。而端到端算法则通过单一神经网络直接连接传感器数据输入与控制指令输出。与传统的模块化自动驾驶相比，端到端自动驾驶神经网络逐渐接管了系统的各个部分，其架构设计简单，减少中间数据降维的成本，同时减小误差以达到全局最优。端到端的优势在数据量达到一定程度后性能显著提高，但是缺点是数据量较小时候性能上升缓慢，远低于解耦的传统基于专家模型的策略。

图30：端到端算法与模块化系统框架对比



资料来源：Pranav Singh Chib《Recent Advancements in End-to-End Autonomous Driving using Deep Learning: A Survey》，民生证券研究院

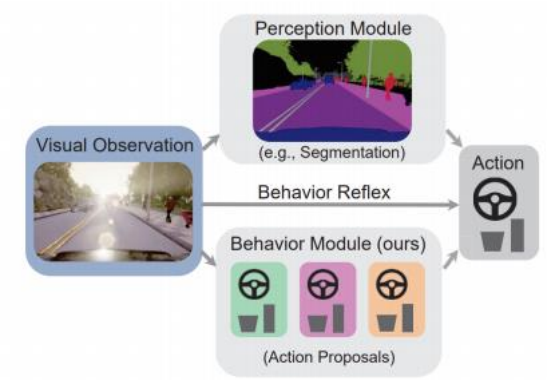
图31：端到端模型与基于规则模型表现曲线对比



资料来源：2023 年 CVPR，民生证券研究院

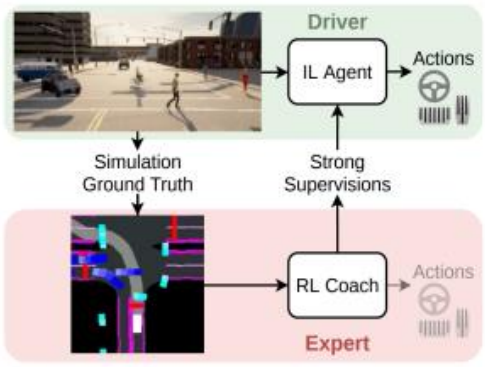
端到端算法实现自动驾驶有两种主要方法：通过强化学习探索和改进驾驶模型、使用模仿学习以监督的方式训练它模仿人类驾驶行为。强化学习的工作原理是通过与环境的相互作用，随着时间的推移最大化累积奖励，网络根据自己的行为做出驱动决策，以获得奖励或惩罚。它在利用数据方面的效率较低。而模仿学习是在专家演示中学习驾驶风格，因此需要大量的实际驾驶场景来作为模型的训练样例，数据集的规模与多样性成为关键问题。

图32：模仿学习框架示例



资料来源：Pranav Singh Chib《Recent Advancements in End-to-End Autonomous Driving using Deep Learning: A Survey》，民生证券研究院

图33：强化学习框架示例



资料来源：Pranav Singh Chib《Recent Advancements in End-to-End Autonomous Driving using Deep Learning: A Survey》，民生证券研究院

4.2 端到端算法相比传统的技术架构的优势

4.2.1 更容易解决 corner case

在传统的决策规划框架中，研发人员会根据不同的 ODD 定义好规则，面对特定场景时找到对应的规则，然后调用相应的规划器生成控制轨迹。这种架构需要事先写好大量的规则，故称为“重决策方案”。重决策方案较易实现，在简单场景下也堪称高效，但在需要拓展 ODD、或把不同的 ODD 连接起来时，就需要大量的手写规则来查缺补漏，从而实现更连续的智驾体验。当遇到未学习过的场景，即 corner case 时，系统会表现得不够智能甚或无法应对。

端到端是通过对场景的理解进行判断，比如环境车辆动态、车道线、交通灯、转向灯灯，通过多维度的元素，甚至是人类没有意识到的要素进行综合分析，判断意图，所以其理解的天花板更高。

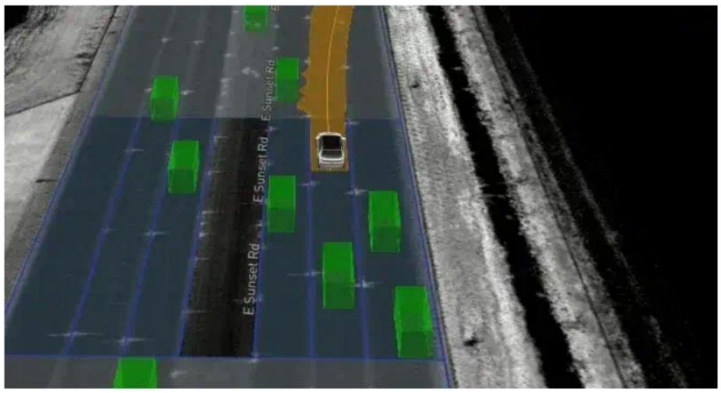
图34：城市中加塞场景，基于规则模型很难处理

资料来源：长城汽车测试城市 NOA 自动驾驶加塞场景，民生证券研究院

4.2.2 拟人化自动驾驶

传统智驾通过横向策略和纵向策略进行车辆的行为控制，基于确定的规则和精确的控制参数，导致车辆动作机械化，要做到拟人驾驶需要开展大量工作，定义控车曲线和匹配场景。端到端的本质是学习，所以其可以模仿人类驾驶汽车的行为，直接输出包括方向盘转角、方向盘转速、油门踏板开度、制动踏板开度等，从而实现接近人类驾驶的习惯进行任务的过程控制。

图35：端到端感知-决策模型示例



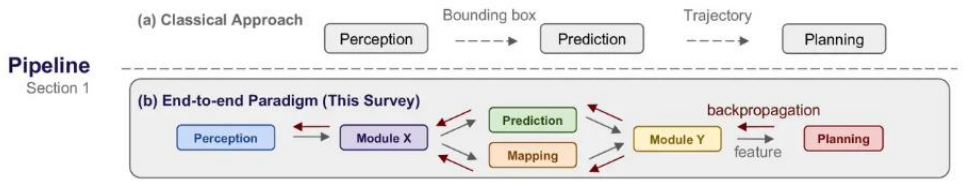
资料来源：Tesla，民生证券研究院

4.2.3 全局最优，成本低且泛用性强

传统“分而治之”的模块化架构，可能囿于局部最优解而难以达到全局最优。由于每个任务相对独立，人工标注使数据的针对性强，监督学习使模型训练的信号强，因此 AI 模型能迅速提升性能，有利于快速实现一个完整的产品。但在到达“局部最优解”之后，这些模型难以进一步提升，且串在一起之后形成累积误差，不利于追求全局最优解。

与传统的模块化自动驾驶系统相比，端到端自动驾驶系统设计难度低，硬件成本较小，并且通过多样性的数据，能够获得在不同场景下的泛用性。所以从算法架构设计的角度，其具有高度的整合度和一体化，省去了多个模块的独立架构设计和算法开发，降低代码量和运行所调度的模块数量。另一方面，由于模型直接从原始数据中学习，而不需要依赖于人工设计的特征或规则，所以删去了枯燥的标注工作。最重要的还有一点就是省去了后期无穷尽的规则补充和场景补充，从而减少了人工维护和升级的成本。

图36：误差依次反向传播给所有模块达到全局最优



资料来源：Li Chen《End-to-end Autonomous Driving: Challenges and Frontiers》，民生证券研究院

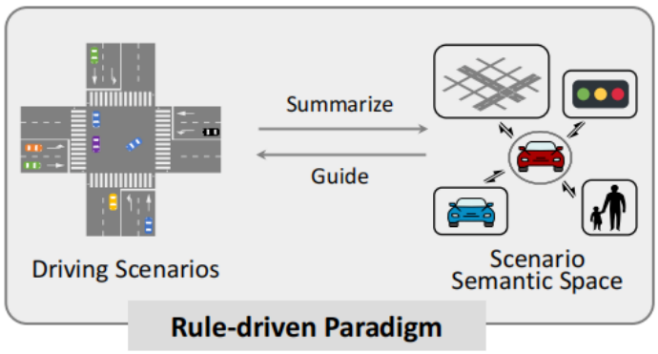
4.3 自动驾驶端到端算法迁移至人形机器人的优势

4.3.1 端到端算法迁移优势一：数据驱动的技术范式

自动驾驶端到端算法代表了一种数据驱动的学习范式，这种范式同样适用于机器人领域。通过大量的数据训练，模型能够学习到复杂的驾驶或操作行为，从而实现高度的智能化。

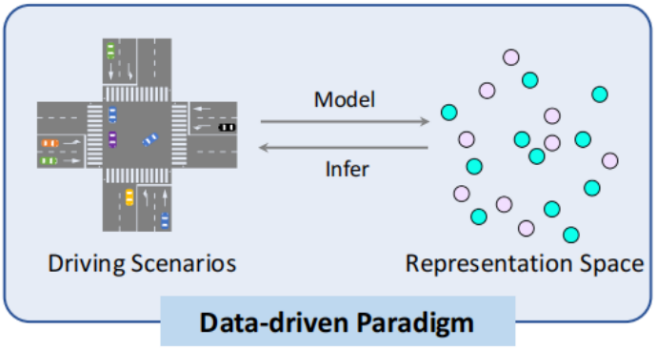
自动驾驶系统在道路上收集的数据，以及通过仿真和合成数据技术获取的数据，都可以为人形机器人的训练提供有力支持。

图37：基于规则驱动



资料来源：csdn，民生证券研究院

图38：基于数据驱动



资料来源：csdn，民生证券研究院

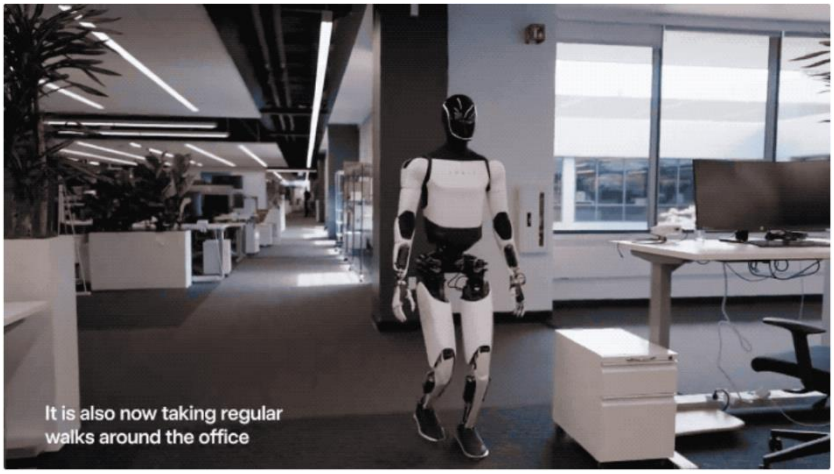
4.3.2 端到端算法迁移优势二：算法架构的通用性

完全端到端算法采用“Bev（鸟瞰视角）+Transformer（预训练）+Teacher-student（知识蒸馏）”方式实现力位的双控，典型代表是特斯拉的 Optimus 人形机器人，根据上文所述，特斯拉人形机器人采用了相同的算法架构。

端到端算法从汽车自动驾驶迁移至人形机器人几乎不需要做太多额外工作，车本身就是一种机器人。早期的特斯拉 Optimus 机器人使用了与汽车完全相同的计算机和摄像头，通过让汽车的神经网络在机器人上运行，它在办公室里走动时仍试图识别“可驾驶空间”，而实际上它应该识别的是“可行走空间”。这种通用化能力表明了很多技术是可以迁移的，虽然需要一些微调，但大部分系统和工具都是

通用的。

图39：特斯拉 optimus 机器人避障行走



资料来源：tesla，民生证券研究院

4.3.3 端到端算法迁移优势三：拟人化行为的实现

端到端算法是自动驾驶拟人化行为实现的关键。它采用整体化的神经网络，将感知、预测和规划等任务整合到一个模型中。通过输入感知信息（如摄像头、雷达等传感器数据），模型能够直接输出轨迹或控制信号，实现类似人类的驾驶行为。自动驾驶端到端算法能够学习到人类驾驶的拟人化行为，如平滑的转向、加速和减速等。

这种拟人化行为在人形机器人上同样重要，可以提升机器人的交互能力和用户体验。通过迁移自动驾驶的拟人化算法范式，人形机器人可以更加自然地与人类进行交互，如理解人类手势、面部表情等。

4.4 机器人端到端算法的关键问题

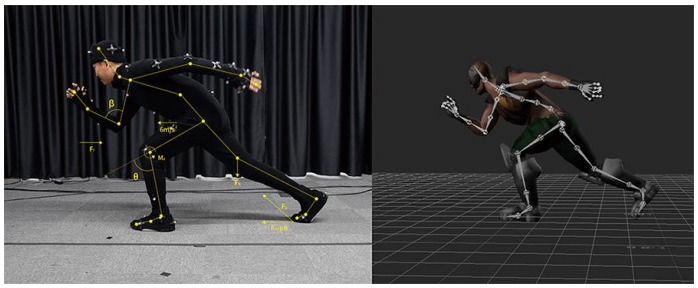
4.4.1 关键问题一：真实数据收集与标注

端到端算法需要大量连续时序的驾驶行为视频进行标注，这种数据收集、标注及闭环验证的过程在人形机器人上同样困难。人形机器人需要面对更加复杂的环境和任务，因此数据收集的难度和成本都更高。同时，由于人形机器人的操作具有更高的风险性，因此数据标注的准确性也要求更高。人形机器人需要大量实际人类真实的数据集给机器人进行训练。

动作捕捉技术和 VR 远程操作是实现人形机器人拟人化动作数据采集的有效途径。动作捕捉技术通过在人体关键部位贴上反光标记点或使用惯性传感器等方式，捕捉人体的运动姿态和动作数据。VR 远程操控技术是人类戴着 VR 眼镜和手

套，通过远程操作的方式来采集机器人数据。这些数据可以被用于训练人形机器人的动作模型，使其能够模拟出类似人类的动作和行为。

图40：动作捕捉技术采集数据



资料来源：武汉零智妙境科技 VR，民生证券研究院

图41：VR 远程操控采集数据

资料来源：特斯拉，民生证券研究院

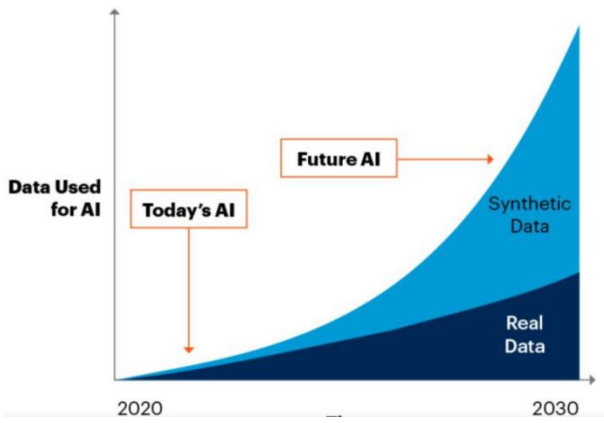
4.4.2 关键问题二：合成数据的生成和使用

由于扩展法则（Scaling Law）的存在，机器人的数据集大小决定了其性能的好坏，真实数据的采集消耗较大的人力物力成本，合成数据仅依赖 AI 算法实现数据生成，数据采集快并且成本低廉。

同时人形机器人面临着场景复杂性与模型泛化能力的问题，合成数据构建的世界模型就起到了很大的作用。自动驾驶场景相对结构化，主要操作在可预测和规范化的环境中。而人形机器人需要应用于多样的场景，如工厂、家庭、办公室等，对泛化能力的要求远高于自动驾驶汽车。基于世界模型生成高质量的动作视频和规划策略，在仿真环境中模拟各种复杂场景，就能够提升系统的鲁棒性。

合成数据生成的关键问题是保持数据集的熵和多样性，避免生成的数据与真实数据差距过大或者样式单一。

图42：未来合成数据的使用



资料来源：Gartner，民生证券研究院

4.4.3 关键问题三：模型的可解释性，展现模型思维链条

现有感知决策一体化模型缺乏可解释性，这一问题在人形机器人上同样存在。由于人形机器人需要与人类进行交互，因此模型的可解释性对于提升用户的信任度和接受度至关重要。

曾是特斯拉自动驾驶项目负责人的 Andrej Karpathy 指出，互联网数据确实是曾经用来训练模型的主要来源，但它并不是最理想的数据。现在真正需要的是大脑内部的思维轨迹、解决问题时的思维过程，如果能有数十亿条这样的数据，那么 AGI 就基本实现了。然而，目前还没有这样的数据。因此，当前的活动很多都集中在如何将数据集重构为这些内部思维轨迹的形式，同时大量依赖合成数据生成来填补这一空白。

4.5 特斯拉 grok 模型：模拟思维链思考过程

2024 年 3 月 28 日 xAI 发布了 Grok-1.5 模型。Grok-1.5 的核心在于使用“思维链”语言。这种语言帮助汽车分解复杂的场景，利用规则和反事实进行推理，并解释其决定。这种创新性的方法将自动驾驶的“像素到行动”映射提升到“像素到语言到行动”的新模式。

通过特斯拉自有的数据管道大规模标注高质量的“人工解释痕迹”，Grok-1.5 可以超越现有的语言模型，在复杂场景下进行更加细致入微的多模态推理。这不仅有助于解决自动驾驶的“边缘情况”，还可以使系统的决策更加透明和可信。

图43：Grok1.5 模型参数对比

Benchmark	Grok-1	Grok-1.5	Mistral Large	Claude 2	Claude 3 Sonnet	Gemini Pro 1.5	GPT-4	Claude 3 Opus
MMLU	73% 5-shot	81.3% 5-shot	81.2% 5-shot	75% 5-shot	79% 5-shot	83.7% 5-shot	86.4% 5-shot	86.8 5-shot
MATH	23.9% 4-shot	50.6% 4-shot	—	—	40.5% 4-shot	58.5% 4-shot	52.9% 4-shot	61% 4-shot
GSM8K	62.9 8-shot	90% 8-shot	81% 5-shot	88% 0-shot CoT	92.3% 0-shot CoT	91.7% 11-shot	92% 5-shot	95% 0-shot CoT
HumanEval	63.2% 0-shot	74.1% 0-shot	45.1% 0-shot	70% 0-shot	73% 0-shot	71.9% 0-shot	67% 0-shot	84.9% 0-shot

资料来源：特斯拉官网，民生证券研究院

模拟思维链思考过程包括三步：场景分解、规则和反事实推理、决策解释。

场景分解：当特斯拉车辆搭载 Grok-1.5V 模型时，模型会首先通过摄像头等传感器收集周围环境的信息，并将这些信息转化为数字信号。然后，模型会使用思维链语言对复杂的驾驶场景进行分解，将其拆分成多个简单的子场景或任务。

规则和反事实推理：在分解场景后，Grok-1.5V 会利用预先学习的规则和反事

实进行推理。这些规则可能包括交通规则、道路标志的含义、车辆动力学原理等。反事实推理则是指模型会考虑如果采取某种行动，可能会发生什么结果，并据此做出决策。

决策解释：与传统的自动驾驶系统不同，Grok-1.5V 不仅能够做出决策，还能够解释其决策过程。模型会将思维链语言中的推理步骤转化为人类可理解的语言或图像，以便驾驶员或相关人员了解系统的决策依据。

5 英伟达 Robocasa：具体智能关键节点，首次论证 real-sim-real

5.1 英伟达 Robocasa：基于厨房场景的模拟数据收集

5.1.1 提出的问题与研究意义

随着人工智能（AI）的快速发展，机器人领域因缺乏大规模机器人数据集而受到限制。之前的一些研究尝试创建大规模，多样化的数据集来训练通用机器人模型，但这些数据集在泛化能力上仍存在差距，此外，现有的模拟框架在场景、任务和资产多样性方面存在不足，且大多数框架没有结合生成式 AI 工具。英伟达提出了 RoboCasa，这是一个用于训练通用机器人的大型模拟框架，专注于现实生活环境，尤其是厨房环境，Robocasa 数据集提供了超过 150 个对象类别的数千个 3D 资产以及数十种可交互的家具和电器，它通过现实物理模拟来扩展环境、任务和数据集，以促进机器人学习方法的扩展。**目的是为了解决如何通过模拟环境来扩展机器人学习方法的规模，特别是针对通用机器人在日常环境中的训练的问题。**实验结果表明：在使用生成的机器人数据进行大规模模仿学习方面有着显著的效果提升，在现实世界任务中利用模拟数据来提升实际效果方面显示出巨大的前景。

Robocasa 有以下特点：1）多样化资产：在生成性 AI 工具的帮助下创建 120 个厨房场景和 2500 多个 3D 对象，比如从文本到三维模型的对象资产，以及从文本到图像模型的环境纹理；2）跨化身支持：支持移动机械手和仿人机器人；3）多样化的任务：在大型语言模型（LLM)的指导下创建任务；4）大规模训练数据集：有超过 100,000 条轨迹。

5.1.2 核心方法与进步

Robocasa 的模拟框架中包含 5 个方面内容：

1）模拟平台：Robocasa 构建在 RoboSuite 之上，并通过提供了大量的场景、对象和硬件平台，继承了几个核心组件，包括环境模型格式和机器人控制器，延续了 RoboSuite 框架模块化、快速、方便的特性，为了支持空间尺度环境，团队还扩展了 RoboSuite 以适应移动操纵器，包括安装在轮式基座上的机器人、人形机器人和带臂的四足机器人。

2）厨房场景：团队根据标准尺寸和空间规格对世界各种风格的厨房进行建模，并将其与一个大型的可交互的家具和应用程序、橱柜、炉子、微波炉、咖啡壶等仓库相匹配，构建模拟使用的厨房场景，并使用高质量的 AI 生成纹理来增加视觉多样性，这些纹理可以用作现实领域随机化的一种形式，以显著增加训练数据集的视觉多样性。

图44：Robocasa 模型使用的厨房场景

资料来源：Soroush Nasiriany《RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots》，民生证券研究院

3）资产库：Robocasa 创建了一个包含 2509 个高质量资产的库，涵盖 153 个不同的类别。这些资产包括家具、电器和其他厨房用品，大部分由 luma.ai 生成，

4） 任务集：该模拟包含 100 个系统临时评估的任务，前 25 个是基础原子任务（如抓取和放置、开关门等），另外 75 个是在大型语言模型（LLMs），尤其是 GPT-4o 的指导下生成的复合任务。如图 44 所示，英伟达研究团队使用 LLM 来概括不同的任务。首先，提示 GPT-4 提供不同的高级厨房活动，例如煮咖啡或洗碗等,团队共编制了 20 个任务清单；随后，对于每个活动，提示 GPT-4（或 Gemini1.5）提出一组不同的表征任务,包括：任务、目标、对象、家具、技能等。例如烹饪或清洁。

5) **数据集：为了增加数据集，团队扩展了 MimicGen，为原子任务生成 100K 额外的轨迹。使用数据生成工具来扩展数据量、利用自动轨迹生成方法来收集大规模演示数据集。**一个由四名人类操作员组成的团队使用 3D 为每个原子任务收集了 50 个高质量的演示集， 每个任务演示都是在一个随机的厨房场景中收集的（随机的厨房平面图 、随机的厨房风格和随机的 ai 生成纹理）。这就通过人工远程操作（1250 个演示）产生了大型和多样化的模拟数据集。然而，即使是这个规模的人类数据也不足以解决大多数任务。MimicGen 先于 Robocasa 出现，团队选择使用数据生成工具 MimicGen 来扩展数据量，MimicGen 可以从人类演示的种子集中自动合成丰富的数据集。**核心一代首先将每个人类演示分解为一个以对象为中心的操作片段。然后， 对于一个新场景，它根据相关任务的当前姿态转换为每一个以对象为中心的片段，并将片段缝合在一起，让机器人按照新的轨迹收集新的任务演示。**MimicGen 需要一些关于模拟的基本假设：任务具有以对象为中心的子任务序列的一致性 。而由八种核心技能组成的原子任务，所有与某一技能对应的任务都具有相同或相似的以对象为中心的子任务序列，其主要区别来自于引用对象的身份。因此，指定子任务序列较容易实现。 此外，提供给 MimicGen 的每个

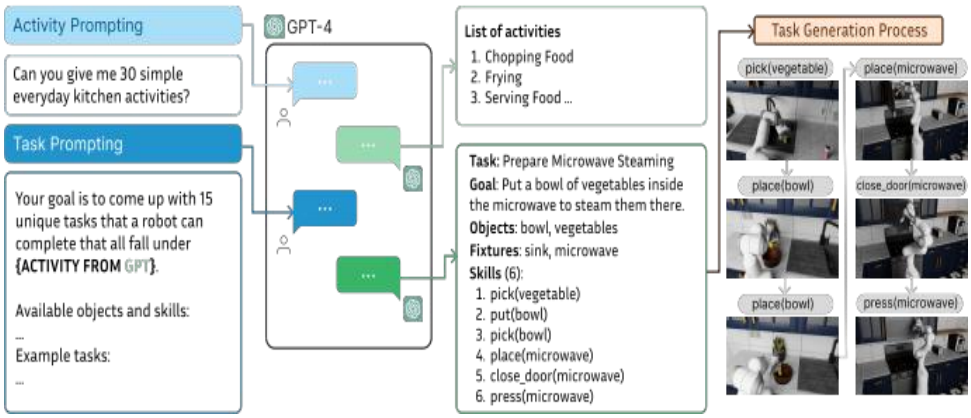
人类演示还必须用与每个以对象为中心的子任务对应的分段进行注释。这可以通过检测每个子任务结束的自动化度量来实现。

Robocasa 与其他流行的模拟框架相比，进步如下：1)Robocasa 支持移动操作，而非仅限于桌面操作；2）具有逼真的渲染、大量的任务、房间比例和对象；3）Robocasa 支持端口室缩放，其他模型仅支持在房间中较小部分进行移动操作等。**Robocasa 是唯一一个支持大量任务、房间规模的场景和物体的框架，同时结合了人工智能生成的任务和资产任务确保场景和任务可能无限多样性。此外，Robocasa 提供了大规模的任务演示数据集以及 MimicGen 系统，并提供了在大型任务集合中通过模仿学习训练的代理的全面分析。各种场景、任务和资产与 RoboCasa 提供的广泛数据集相结合，将满足机器人学习社区中任何其他模拟都没有解决的关键要求。**

在实验中，团队主要探讨了以下问题：1)在学习多任务策略时，机器生成的轨迹有多有效？2)随着训练数据集规模的增加，模拟学习策略规模的泛化性能将如何提高？3)大规模模拟数据集促进知识转移到下游任务，并促进现实世界任务的政策学习？**Robocasa 共涉及了原子任务、复合任务和真实世界实验三个场景。**

在对原子能任务的模拟学习中，团队设计了 25 个原子任务，涵盖八种基础技能（如抓取和放置，开关门等）,通过人类操作和 MinicGen 生成数据集，分别训练多任务策略，并评估其在不同数据集上的表现。在人类数据上，整体成功率为 20.8%，在使用全部生成的数据集上，成功率显著提升至 47,6%，从使用机器生成数据中观察到调整趋势：随着生成数据数量的增加，模型性能稳步提高，说明未来数据生成工具使模型能够以相对较低的成本学习更多的性能代理。

图45：Robocasa 使用 GPT-4 生成不同任务的模型流程

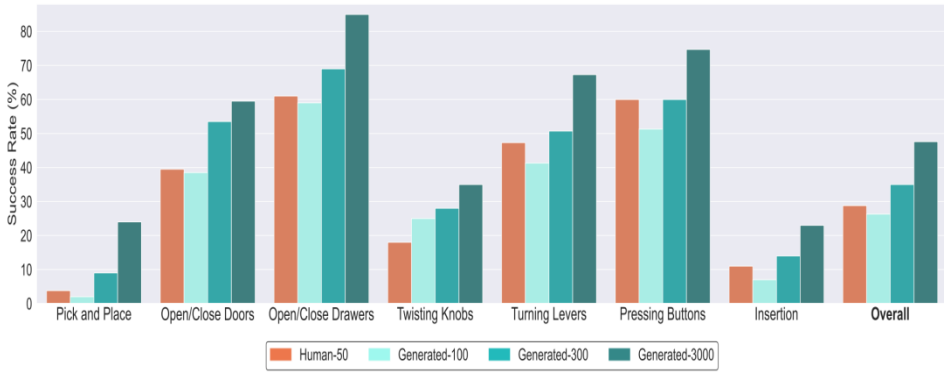


资料来源： Soroush Nasiriany《RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots》，民生证券研究院

在对复合任务的模拟学习中发现，选了五个代表性的复合任务（如放蔬菜，解冻食物等）,分别采用从头开始学习和微调预训练策略进行实验，每个任务收集 50

个人类演示，并使用 MimcGen 生成的数据进行微调，微调后的模型在质量上表现更好，策略架构、学习算法和微调策略的进步可能在性能方面发挥关键作用。

图46：人工演示和机器生成的数据集之间的比较结果



资料来源： Soroush Nasiriany《RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots》，民生证券研究院

将模拟转移到真实世界的环境中，比较了仅在真实数据（Real only）和模拟数据（Real +Sim）上共同训练的策略表现，并根据相应的感觉运动技能对任务结果进行分组，包括三个任务(如从柜台到水槽的抓取和放置)，对于每个任务，Robocasa 团队收集了 50 个演示，每个演示都超过 5 个不同的对象类别。团队为每个任务训练一个策略，并比较 Real only 和 Real +Sim 两种设置。同时，团队研究了 3 种任务的平均政策成功率（平均值和标准数据偏差，百分比），并评估了 5 个可见的对象类别和 3 个不可见的对象类别（在现实世界的演示中看不到）。结果表明，在某些对象上，在真实数据上训练的策略在已知对象上的平均成功率为 13.6%,而在真实数据和模拟数据上共同训练的平均成功率为 24.4%，最高提高了 79%，说明模拟器的丰富多样性以及视觉和物理真实性显著改善了模拟效果。

图47：Real only 和 Real+Sim 下不同对象训练成功率评估

Setting	Task	Real only	Real + Sim (Ours)
Seen Obj	Counter to sink	12.7 ± 2.5	**22.0 ± 2.8**
	Sink to counter	20.0 ± 5.9	**29.3 ± 4.1**
	Counter to cabinet	8.0 ± 1.6	**22.0 ± 5.8**
	Task average	13.6	**24.4**
Unseen Obj	Counter to sink	3.3 ± 4.7	**8.9 ± 7.9**
	Sink to counter	1.1 ± 1.6	**7.8 ± 4.2**
	Counter to cabinet	3.3 ± 4.7	**11.1 ± 11.0**
	Task average	2.6	**9.3**

Fig. 10: **Real Robot Evaluations.** In a real-world kitchen domain with only a handful of demonstrations, we explore co-training policies with our simulation data. Compared to training policies exclusively on in-domain real-world demonstrations, co-training substantially improves policy performance.

资料来源： Soroush Nasiriany《RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots》，民生证券研究院

5.1.3 核心结论+未来发展

英伟达提出了 Robocasa，一个用于训练通用机器人的大规模模拟框架，Robocasa 结合了生或式 AI 工具，创建了多样化，真实的厨房场景和任务，并通过大规模数据集提高了机器人在真实世界任务中的表现，实验结果表明，合成数据在模拟环境中学习机器人的策略是有效的，并且可以显著促进知识迁移到下游任务和真实世界任务中。

但实验表明，复合任务的微调产生了低性能，未来可以研究更强大的策略架构和学习算法，并提高机器基因比率和数据集的质量；使用 LLM 创建任务的过程仍然需要人工指导来编写相关注释，未来随着 LLM 成为模型生成体，使用 LLM 提出数千个新的场景和任务并编写代码，以最小的语言来实现这些场景和任务将成为可能。此外，目前的模拟仅限于厨房环境中，未来可以拓展到该环境和任务之外。

6 机器人 real-sim-real 可行，迈向真正的 AGI 智能化

6.1 李飞飞团队 Rekep：一种针对机器人操作任务的新型空间和时间约束表示方法，提供了三任务闭环的解决方案

6.1.1 提出的问题与研究意义

如何制将机器人操控任务表示为关联机器人和环境的约束条件，使它们既适用于多样化任务，又无需手动标记，还能被现成的求解器实时优化以产生机器人动作，是一个亟待解决的问题。

李飞飞团队 Rekep 项目提出了关系关键点约束（ReKep），这是一种针对机器人操控约束的视觉基础表示方法。ReKep 用 Python 函数表示，将一组 3D 关键点映射到数值成本上。Rekep 展示了通过将操控任务表示为一系列关系关键点约束，可以采用层次化优化过程来求解机器人动作(由一系列末端执行器姿态 SE(3)表示），并实现实时频率的感知-动作循环。此外，为了避免为每项新任务手动指定 ReKep，团队设计了一个自动化流程，利用大型视觉模型和视觉-语言模型从自由形式的语言指令和 RGB-D 观测中产生 ReKep（Relational Keypoint Constraints）。

机器人操控涉及与环境中的物体进行复杂的交互，这些交互通常可以表示为空间和时间域中的约束。例如，将茶倒入杯中的任务，机器人必须在手柄处抓握，在运输过程中保持杯子直立，对准壶嘴与目标容器，然后倾斜杯子以正确角度倒茶。这些约束不仅编码了中间子目标（例如，对准壶嘴），还编码了过渡行为（例如，在运输过程中保持杯子直立），共同决定了机器人动作在与环境的关系中的空间、时机和其他组合要求。然而，有效地为现实世界的大量任务制定这些约束条件将面临重大的挑战。虽然使用直接和广泛使用的方法来表示相对姿态之间的约束，但刚体变换不能描述几何细节，需要先验获得对象模型，并且不能在变形对象上工作。另一方面，数据驱动的方法可以直接在视觉空间中实现学习约束。虽然很灵活，但随着对象和任务的约束数量组合增加，如何有效地收集训练数据仍不清楚。为解决无操作的约束，李飞飞团队提出了关系关键点约束（ReKep），**该方法就是将任务表示成一个关系关键点序列。并且，这套框架还能很好地与 GPT-4o 等多模态大模型很好地整合。**

6.1.2 核心方法与进步

1）关系关键点约束理论的核心思想原理

核心实现方式是：对于每个阶段 i，该优化问题的目标是：基于给定的 ReKep

约束集和辅助成本，找到一个末端执行器姿势作为下一个子目标（及其相关时间），以及实现该子目标的姿势序列，该公式可被视为轨迹优化中的 direct shooting。

例如，下图的杯子任务可分为三个步骤：①步骤一：机器人抓住手柄并在搬运杯子时保持直立，避免茶水洒出。该过程中，子目标约束是将末端执行器伸向茶壶把手。此时 Rekep 限制茶壶手把的抓取位置（蓝色），②步骤二：将茶壶口与杯子口对齐，该过程中子目标约束是让茶壶口位于杯口上方，路径约束是保持茶壶直立，避免茶水洒出，ReKep 将茶壶喷口（红色）拉到杯开口的顶部（绿色）。③步骤三：使茶壶到达倾斜的角度，并将茶壶中的水倒出。该过程目标约束是到达指定的倒茶角度。ReKep 通过关联手柄（蓝色）和喷口形成的矢量（红色）来限制茶壶的方向。该过程中约束编码了中间子目标（对齐嘴），也编码了转换行为（在运输中保持杯子直立），这些共同决定了机器人动作与环境相关的空间、时间和其他组合要求。这就将多过程的任务分解为多个目标和约束条件，通过优化求解输出并实现机器人的行为。

图48：关系关键点约束 (Rekep)将不同的操作行为指定为在语义关键点上操作的约束功能的时空约束序列

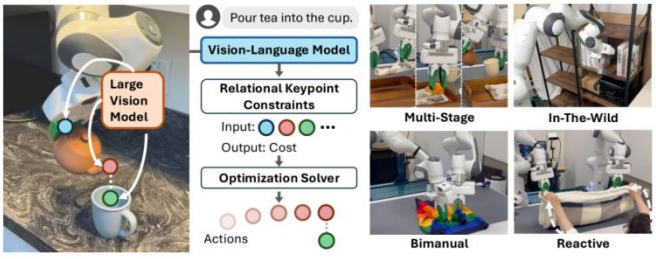


Figure 1: Relational Keypoint Constraints (ReKep) specify diverse manipulation behaviors as an optimizable spatio-temporal series of constraint functions operating on semantic keypoints. In the pouring task, one ReKep first constrains the grasping location at the handle of the teapot (*blue*). A subsequent ReKep pulls the teapot spout (*red*) towards the top of the cup opening (*green*) while another ReKep constrains the desired rotation of the teapot by associating the vector formed by the handle (*blue*) and the spout (*red*).

资料来源：Wenlong Huang, Li Fei-Fei《ReKep: Spatio-Temporal Reasoning of Relational Keypoint Constraints for Robotic Manipulation》，民生证券研究院

图49：Rekep 构建一组子目标约束和一组路径约束

$$\mathcal{C}^{(i)}_{\text{sub-goal}} = \{f^{(i)}_{\text{sub-goal},1}(\boldsymbol{k}), \ldots, f^{(i)}_{\text{sub-goal},n}(\boldsymbol{k})\}$$

- 一组子目标约束

$$\mathcal{C}^{(i)}_{\text{path}} = \{f^{(i)}_{\text{path},1}(\boldsymbol{k}), \ldots, f^{(i)}_{\text{path},m}(\boldsymbol{k})\}, \text{ where } f^{(i)}_{\text{sub-goal}}$$

- 一组路径约束

资料来源：Wenlong Huang, Li Fei-Fei《ReKep: Spatio-Temporal Reasoning of Relational Keypoint Constraints for Robotic Manipulation》，民生证券研究院

使用 ReKep，可将机器人操作任务转换成一个涉及子目标和路径的约束优化问题。一个操作任务通常涉及多个空间关系，并且可能具有多个与时间有关的阶段，其中每个阶段都需要不同的空间关系，Rekep 将一个任务分解成 N 个阶段并使用 ReKep 为每个阶段 i ∈ {1, ..., N} 指定两类约束：子目标约束和路径约束。其中子目标约束编码了阶段 i 结束时要实现的一个关键点关系，而路径约束编码了阶段 i 内每个状态要满足的一个关键点关系。

现实环境复杂多变，有时候在任务进行过程中，上一阶段的子目标约束可能不再成立（比如倒茶时茶杯被拿走了），这时候需要重新规划。该团队的做法是检查路径是否出现问题。如果发现问题，就迭代式地回溯到前一阶段 Rekep 的关键特点如下：

（1）多模态输入处理：ReKep 能够处理 RGB-D 图像和自由形式的语言指令，利用大型视觉模型（如 DINOv2）和视觉-语言模型（如 GPT-4o）来识别场景中

的关键点，并生成 ReKep 约束。

（2）层次化优化：通过将操作任务分解为多个阶段，并为每个阶段指定子目标约束和路径约束，ReKep 采用层次化优化方法来实时求解机器人动作。

（3）实时性能：ReKep 能够在大约 10Hz 的频率下实时解决优化问题，适用于需要快速反应的机器人操作任务。

（4）自动化关键点提议和约束生成：ReKep 通过自动化流程，减少了手动指定任务特定数据的需求，提高了任务的可扩展性和适用性。

（5）系统实现：ReKep 在单臂和双臂机器人平台上进行了系统实现，展示了其在多种操作任务中的应用潜力。

（6）代码和视频资源：ReKep 的研究团队提供了相关代码和演示视频，以便研究社区进一步探索和应用这一方法。

2）融入视觉-语言模型后的指定关系关键点约束

为了让该系统能在实际情况下自由地执行各种任务，**该团队还使用了大模型。具体来说，他们使用大型视觉模型和视觉 - 语言模型设计了一套管道流程来实现关键点提议和 ReKep 生成。**

图50：Rekep 实现方式概览

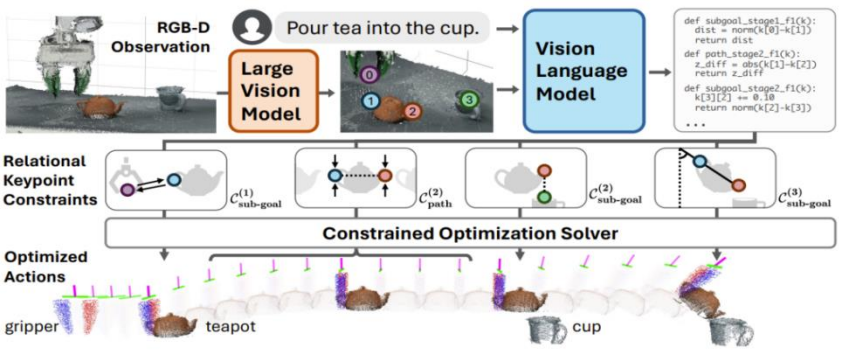


Figure 2: Overview of ReKep. Given RGB-D observation and free-form language instruction, DINOv2 [5] is used to propose keypoint candidates on fine-grained meaningful regions in the scene. The image overlaid with keypoints and the instruction are fed into GPT-4o [6] to generate a series of ReKep constraints as python programs that specify desired relations between keypoints at different stages of the task ($\mathcal{C}^{(i)}_{\text{sub-goal}}$) and any requirement on the transitioning behaviors ($\mathcal{C}^{i}_{\text{path}}$). Finally, a constrained optimization solver is used to obtain a dense sequence of end-effector actions in $SE(3)$, subject to the generated constraints.

资料来源：Wenlong Huang, Li Fei-Fei《ReKep: Spatio-Temporal Reasoning of Relational Keypoint Constraints for Robotic Manipulation》，民生证券研究院

虽然约束通常是针对每个任务手动定义的，李飞飞团队展示了 ReKep 的具体形式具有独特的优势，即它们可以通过预训练的大型视觉模型（LVM）和视觉语言模型（VLM）实现自动化，从而能够根据 RGB-D 观测和自由形式的语言指令在野外环境中指定 ReKep。**基本原理是：利用 LVM 在场景中提出细粒度且具有语义意义的关键点，并利用 VLM 将叠加了关键点的图像和指令输入 GPT-4o，以生成一**

系列 ReKep 约束，这些约束以 Python 程序的形式指定了在任务不同阶段（子目标）关键点之间所需的关系以及过渡行为（路径）的任何要求。即利用大型视觉模型和视觉-语言模型自动化地从自由形式的语言指令和 RGB-D 观测中生成 ReKep，避免了手动指定 ReKep 的需要。有了生成的约束，就可以使用现成的求解器通过跟踪关键点重新评估约束来生成机器人动作。李飞飞团队采用分层优化程序，首先求解一组作为子目标（表示为 SE(3)末端执行器姿态）的路点，然后求解滚动时域控制问题以获得实现每个子目标的密集动作序列。

团队通过一系列任务检查了该系统的多阶段、野外/实用场景、双手和反应行为。这些任务包括倒茶、摆放书本、回收罐子、给盒子贴胶带、叠衣服、装鞋子和协作折叠。结果显示，就算没有提供特定于任务的数据或环境模型，新提出的系统也能够构建出正确的约束并在非结构化环境中执行它们。值得注意的是，ReKep 可以有效地处理每个任务的核心难题。

此外，该团队还基于叠衣服任务探索了新策略的泛化性能。尝试了不同的机器人叠衣服场景，结果显示，该系统为不同衣服采用了不同的策略，其中一些叠衣服方法与人类常用的方法一样。

综上所述，ReKep 通过关键点来指定机器人臂部、物体（部件）和其他代理之间的期望空间关系。这些关键点是任务特定且语义上有意义的 3D 点，并允许使用现成的求解器通过重新评估基于跟踪关键点的约束来产生机器人动作。为了减少手动指定新任务生成 ReKep 的工作量，团队设计了一个自动化流程，利用大型视觉模型和视觉-语言模型从自由形式的语言指令和 RGB-D 观测中自动产生 ReKep。**Rekep 是实现空间智能的关键技术之一，它通过提供一种结构化的方式来理解和操作三维空间，从而增强了机器人的空间感知和操作能力。**

图51：使用 ReKep 模拟折叠不同类别的服装及成功率

	sweater	shirt	hoodie	vest	dress	pants	shorts	scarf	**Total**
Strategy Success	6/10	4/10	4/10	6/10	3/10	7/10	7/10	5/10	52.5%
Execution Success	6/10	5/10	6/10	9/10	7/10	8/10	9/10	9/10	73.8%

Figure 4: Novel *bimanual* strategies of ReKep for folding different categories of garments and their success rates. Since ReKep in this task always associates two points at a time, two keypoints are connected by an arrow if they need to be aligned. The coloring of the keypoints denotes the order. In the sweater task, two sleeves are first folded simultaneously with two arms, and then the two arms grasp the crew neck to align to the bottom.

资料来源：Wenlong Huang, Li Fei-Fei《ReKep: Spatio-Temporal Reasoning of Relational Keypoint Constraints for Robotic Manipulation》，民生证券研究院

6.1.3 核心结论+未来进展

关系关键点约束（ReKep）是一种结构任务表示，使用对语义关键点进行操作的约束，以指定机器人臂、对象（部件）和环境中的其他代理之间的期望关系。结合点跟踪器，李飞飞团队认为 **ReKep 约束可以在层次优化框架中重复和有效地解决，从而作为以实时频率运行的辅助循环策略**，ReKep 的独特优势是它是由大型视觉模型和视觉语言模型合成的细胞。

李飞飞团队 Rekep 主要贡献是：1）将操作任务表述为具有关系关键点约束的分层优化问题；2）设计了一条使用大型视觉模型和视觉语言模型自动指定关键点和约束的管道；3）在两个真实机器人平台上展示了系统实现，这些系统以语言指令和 RGB-D 观测为输入，为各种操作任务产生多阶段、野外、双手和反应式行为，而无需特定任务的数据或环境模型。

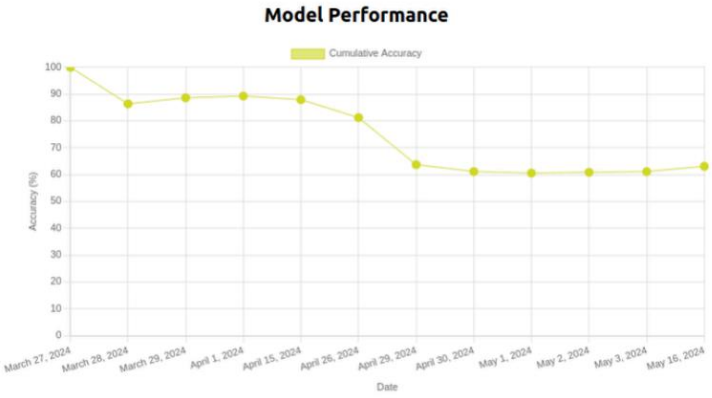
但仍有几个限制。首先，优化框架依赖于基于刚性假设的关键点前向模型，尽管是一个高频反馈回路，放宽了模型的精度要求。其次，ReKep 依赖于精确的点跟踪来正确地优化闭环动作，这本身就是一项具有挑战性的三维视觉任务，由于严重的间歇性遮挡。最后，当前的公式假设每个任务都有一个固定的阶段序列（即骨架）。使用不同的骨架重新规划需要在高频上运行关键点提议和 VLM，这就带来了不利的计算挑战。

6.2 1x 世界模型：首证扩展定律，能通过大量学习理解周围环境

6.2.1 核心问题：在复杂多变的真实环境中进行自我决策和适应

由于真实环境的复杂多变性，即使是同一场景，也会经历光照细微的变化，机器人在模型权重不变的情况下，会在几天内经历性能的快速下降。

图52：机器人性能随时间变化曲线



资料来源：Jack Monas《1x world model》，民生证券研究院

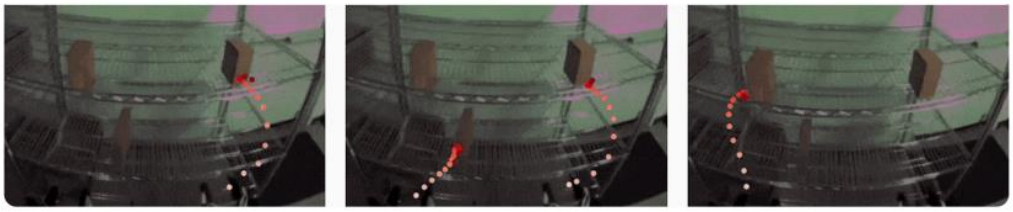
1X 世界模型解决的核心问题是如何使机器人在复杂多变的真实环境中进行自我决策和适应。传统的物理模拟方法往往难以适应大环境变化带来的挑战，且手动创建资产的复杂性高。而 1X 世界模型通过从原始传感器数据中学习，直接构建模拟器，能够在数百万种场景中评估机器人的行为，从而大大提高了机器人的适应性和智能性。

6.2.2 核心突破：从原始传感器数据中直接学习构建模拟器

1X 世界模型的核心突破在于其能够从原始传感器数据中直接学习，构建出能够预测世界如何响应机器人动作的模拟器。这一技术突破了传统物理模拟方法的局限性，使得机器人能够在更广泛、更真实的场景中进行学习和适应。

在过去的一年（2023 年）里，1X 收集的 1X 旗舰产品 EVE 机器人的数据高达数千小时，这些数据包括在家中和办公室中执行各种移动操作任务以及与人互动的任务。研究人员将这些视频和动作数据结合起来，训练了一个世界模型，该模型可以根据观察和动作预测未来的视频。

图53：世界模型轨迹预测

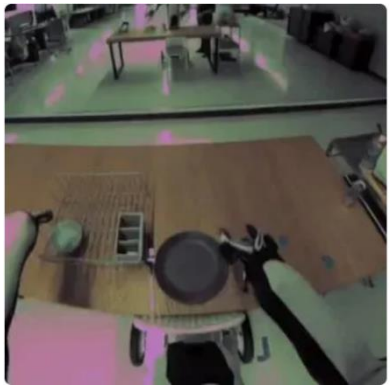


资料来源：Jack Monas《1x world model》，民生证券研究院

机器人通过观看数千小时的视频和来自机器人执行任务的感应器数据，模型能够观察当前的世界状况，并预测机器人在特定动作下会发生什么。

EVE 人形机器人在家庭和办公室环境中执行的各种任务为这一模型提供了宝贵的“生活素材”。通过不断与人类互动并收集这些真实数据，模型学会了如何更贴近真实世界进行模拟。

图54：EVE 人形机器人家庭环境训练



资料来源：Jack Monas《1x world model》，民生证券研究院

图55：EVE 人形机器人办公室环境训练

资料来源：Jack Monas《1x world model》，民生证券研究院

6.2.3 核心结论：首证扩展定律，能通过大量学习理解周围环境

1X 世界模型的核心结论是，通过大量的真实数据学习和模拟，机器人能够预测复杂的物体互动，理解周围环境，并灵活应对日常任务。这一模型使得机器人能够在神经网络空间内进行有效的规划和模拟操作，从而提高了其在复杂环境中的任务执行能力和智能水平。1X 的进展首次在机器人上证明了扩展法则：随着数据、计算和模型规模的增加，机器人在认知和行为上的能力也必将显著提升。

然而，尽管取得了显著进展，1X 世界模型在物体交互中仍可能出现物体失真或逻辑错误的现象，且目前还缺乏真正的自我认知。

图56：执行长视野任务

资料来源：Jack Monas《1x world model》，民生证券研究院

图57：工程师称没有出现自我认知

资料来源：澎湃新闻，民生证券研究院

6.2.4 未来发展方向：利用传感器信息实现完全端到端算法

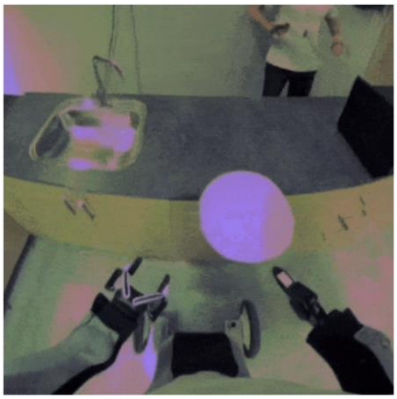
传感器数据融合：1X 世界模型可以进一步融合来自不同传感器的数据，如摄像头、激光雷达、惯性测量单元等，以构建更全面、更准确的世界模型。通过数据融合，可以实现对环境的更精细感知和更深入理解，为机器人的决策和规划提供更丰富的信息支持。

增强环境理解能力：利用传感器信息，1X 世界模型可以进一步提升对环境的理解能力，包括识别物体的形状、颜色、纹理等特征，以及理解物体之间的空间关系和运动规律。这有助于机器人在复杂环境中进行更准确的定位、导航和避障。

图58：对象连贯性问题

资料来源：Jack Monas《1x world model》，民生证券研究院

图59：物理定律理解丢失问题



资料来源：Jack Monas《1x world model》，民生证券研究院

6.3 字节 GR-2：高效动作预测与泛化能力

6.3.1GR-2:高效动作预测和视频生成

现阶段的大语言模型可以实现流畅的文本生成、问题解决、创意写作以及代码生成，视觉-语言模型则能够实现开放词汇的视觉识别，但是具体实践中如何获取这些能力仍需要进一步探索。字节跳动 ByteDance Research 致力于**让机器人模仿学习人类成长过程，将多模态素材的学习与预测直接集成到机器人控制中，以促进泛化并实现高效动作预测和视频生成，开辟智能决策和自主操作新可能性，成为了机器人下一步发展的方向。**

泛化能力与多任务通用性是机器人大模型目前最重要的突破方向，近日 ByteDance Research 的第二代机器人大模型 GR-2 发布了视频和技术报告，展示出卓越的泛化能力和多任务通用性，例如图中所示，GR-2 模型可以在接收指令后生成机器人完成倒咖啡指令并生成视频，这一进步预示着机器人大模型技术将释放出巨大潜力和无限可能。

GR-2 通过在网络规模的视频数据集上进行预训练，显著超越了传统的机器人数据源模型。例如，许多早期模型（如 RoboCasa）主要依赖于有限的机器人数据集进行训练，而这些数据集通常只涵盖少数场景或任务，导致模型在新场景中表现不佳。相较之下，GR-2 的视频数据源更加广泛和多样化，涵盖了从厨房到户外等各种场景，并且结合了多个公开的机器人数据集（如 RT-1 和 Bridge）。这使得 GR-2 能够更好地在新的、未见过的环境中执行任务，因为它已经从大量不同的场景中学习到如何操作和适应不同类型的任务。

图60：GR-2 模型接收倒咖啡指令并生成视频示意图

资料来源：CVer，民生证券研究院

图61：GR-2 视频-语言模型与视频-语音-动作模型示例

资料来源：Chi-Lam Cheang《GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation》，民生证券研究院

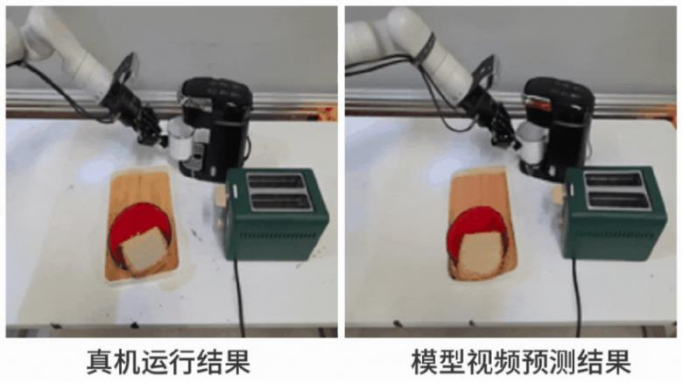
6.3.2GR-2 核心方法：预训练与微调

和许多大模型一样，GR-2 的训练包括预训练和微调两个过程。GR-2 在 3800 万个互联网视频片段上进行生成式训练，也因此得名 GR-2（Generative Robot 2.0）。这些视频来自学术公开数据集，涵盖了人类在不同场景下（家庭、户外、办公室等）的各种日常活动，以期迅速学会人类日常生活中的各种动态和行为模式。这种预训练方式使 GR-2 具备了学习多种操作任务和在多种环境中泛化的潜能。庞大的知识储备，让 GR-2 拥有了对世界的深刻理解。

在微调阶段，GR-2 通过几项关键改进提升了其在实际任务中的表现。首先，GR-2 引入数据增强技术，通过改变训练数据中的背景和物体，使其在未见环境下更具泛化能力。此外，模型通过多视角训练，利用不同角度的视觉数据，增强了其在复杂场景中的操作灵活性和准确性。为了保证动作的流畅性，GR-2 使用了条件变分自编码器（cVAE），生成连续、平滑的动作序列，确保任务执行时的动作更加高效和精准。

在经历大规模预训练后，通过在机器人轨迹数据上进行微调，GR-2 能够预测动作轨迹并生成视频。GR-2 的视频生成能力，让它在动作预测方面有着天然的优势，显著提高了准确率。它能够通过输入一帧图片和一句语言指令，预测未来的视频，进而生成相应的动作轨迹。如下图所示，只需要输入一句语言指令：“pick up the fork from the left of the white plate”，就可以让 GR-2 生成动作和视频。可以看到，机械臂从白盘子旁边抓起了叉子。右图中预测的视频和真机的实际运行也相差无几。

图62：真机预测结果与模拟视频预测结果对比

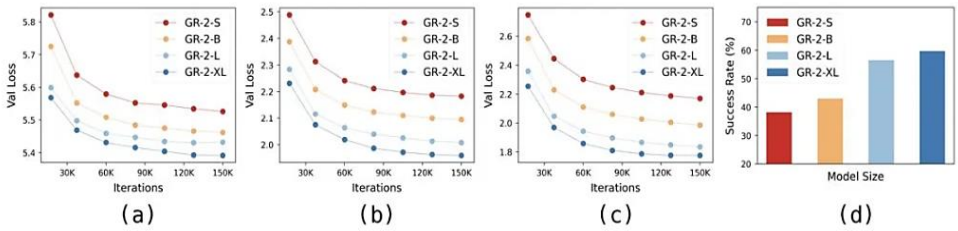


资料来源：CVer，民生证券研究院

6.3.3GR-2 核心结论：仍然符合 Scaling Law

经过多次大模型预训练与微调后，研究团队发现 GR-2 的视频生成与动作预测模型符合 Scaling Law，并且对于 GR-2 这样的机器人模型来说，这一法则尤为关键。随着模型规模的增加，GR-2 的性能呈现出显著的提升。在 7 亿参数规模的验证中，GR-2 团队发现，更大的模型不仅能够处理更多复杂的任务，而且在泛化到未见过的任务和场景时也表现得更加出色。如图所示，在预训练过程中，**视频预测的验证损失随着模型大小的增加而减小，**以图（a）测试为例，在重复 150k 预测后，最小的模型 GR-2-S 产生的视频验证损失最大，为 5.54 单位，而稍大的模型 GR-2-B 与 GR-2-L 产生的视频验证损失均在 5.4-5.5 单位区间内，最大的模型 GR-2-XL 产生的视频验证损失最小，为 5.38 单位。

图63：四种从小到大模型的视频预测验证损失



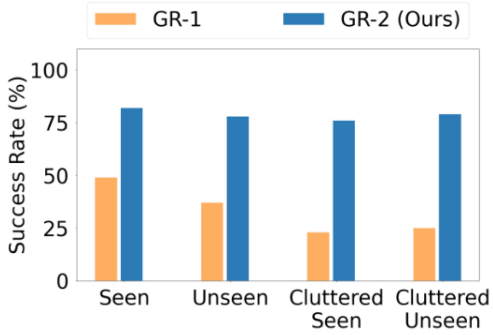
资料来源：Chi-Lam Cheang《GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation》，民生证券研究院

6.3.4GR-2 核心突破：性能较 GR-1 与其余视频语言模型提升显著

核心突破 1：在各场景端到端测试中，GR-2 的成功率相较 GR-1 提升迅速。

以未训练(Unseen)端到端拣选测试为例，GR-2 在未见（Unseen）场景中的成功率显著提升，主要是因为它依赖于大规模的视频数据进行预训练。与 GR-1 仅依赖于有限的机器人数据不同，GR-2 融合了来自网络的视频数据和多任务学习，这些数据覆盖了更多样化的场景和物体操作。通过预训练阶段，GR-2 学会了从视频中推测物体的动态和操作语义，提升了其在新环境中的泛化能力。此外，GR-2 在微调阶段引入了数据增强技术（如新的物体和背景变换），进一步提高了它处理未见场景的适应性。因此，GR-2 能够在未见测试中将成功率从 33% 提升到 79%。

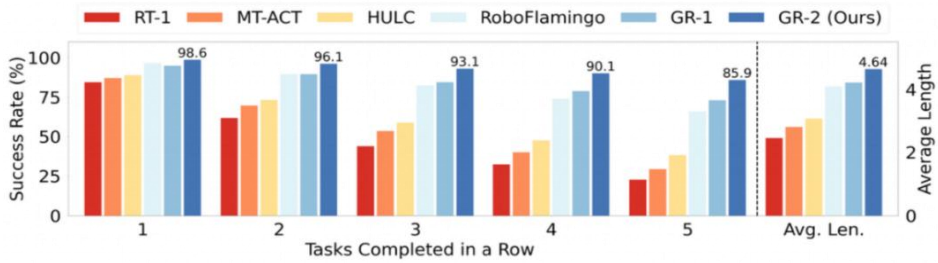
图64：各场景端到端测试中，GR-2 的性能相较 GR-1 提升迅速



资料来源：Chi-Lam Cheang《GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation》，民生证券研究院

核心突破 2：在 CALVIN 机器人操作仿真基准测试中，GR-2 大幅超越五种最先进的基线方法 RT-1 、 MT-ACT 、 HULC 、RoboFlamingo 和 GR-1 。如图显示，横轴代表机器人在连续 5 个任务序列中能够完成的平均任务数，纵轴代表成功率。GR-2 建立了一种新的技术水平，在成功率和平均长度方面优于所有比较基线方法。

图65：CALVIN 机器人操作仿真测试，GR-2 大幅超越五种最先进的基线方法



资料来源：Chi-Lam Cheang《GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation》，民生证券研究院

6.3.5 GR-2 未来应用：强大泛化能力实现多场景任务（如端到端拣选）

GR-2 的强大之处不仅在于它能够处理已知任务，更在于其面对未知场景和

物体时的泛化能力。无论是全新的环境、物体还是任务，GR-2 都能够迅速适应并找到解决问题的方法。在多任务学习测试中，GR-2 能够完成 105 项不同的桌面任务，平均成功率高达 97.7%。此外，GR-2 还能够与大语言模型相结合，完成复杂的长任务，并与人类进行互动，并可以鲁棒地处理环境中的干扰，并通过适应变化的环境成功完成任务。。

在实际应用中，GR-2 相比前一代的一个重大突破在于能够端到端地完成两个货箱之间的物体拣选。无论是透明物体、反光物体、柔软物体还是其他具有挑战性的物体，GR-2 均能准确抓取。这展现了其在工业领域和真实仓储场景的巨大潜力。除了能够处理多达 100 余种不同的物体，如螺丝刀、橡胶玩具、羽毛球，乃至一串葡萄和一根辣椒，GR-2 在未曾见过的场景和物体上也有着出色的表现。

图66：GR-2 完成流畅端到端物体拣选示意图

资料来源：CVer，民生证券研究院

图67：GR-2 在实验中顺利完成 122 项物体拣选，其中过半物体 GR-2 未曾见过

资料来源：Chi-Lam Cheang《GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation》，民生证券研究院

6.4 数字表亲：机器人训练法优化，以更低的成本获取更好的泛化能力

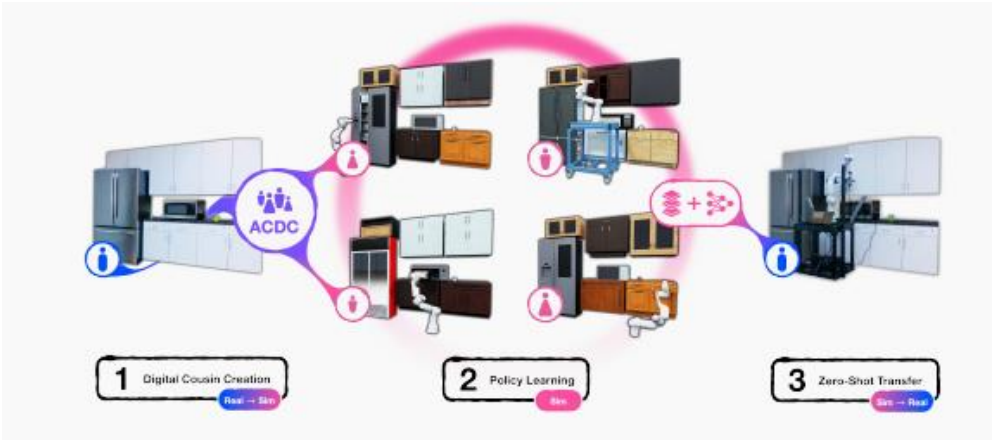
6.4.1 问题聚焦：在机器人训练中兼顾降低成本和补足泛化能力

模拟是一种廉价且潜力极强的训练数据来源，但模拟环境和真实世界环境之间存在语义和物理差异。这些差异可以通过“数字孪生”的训练最小化，将数字双胞胎作为真实场景的虚拟副本。然而，这一办法的生成成本极为昂贵，且无法提供良好的跨域泛化能力。

作为机器人训练法的最新成果，数字表亲力求在保留数字孪生训练优势的基础上，降低从真实到模拟环境的生成成本并提高机器人学习的泛化能力。这一点与很多机器人大模型的突破方向并无二致。与数字孪生不同，它没有直接模拟现实世界的特定对应物，但仍然能表现出类似的几何形状和语义功能。因此，数字表亲降

低了生成类似虚拟环境的成本，同时还通过提供一系列相似但不完全相同的训练场景，提高从模拟到真实环境的迁移鲁棒性。另外，数字表亲还能实现将单幅图像转换为完全交互式的虚拟场景的目标，同时全自动处理过程，无需人工注释，并且训练出的机器人策略可以直接在原始场景中进行零样本部署。

图68：数字孪生与数字表亲生成的模拟环境



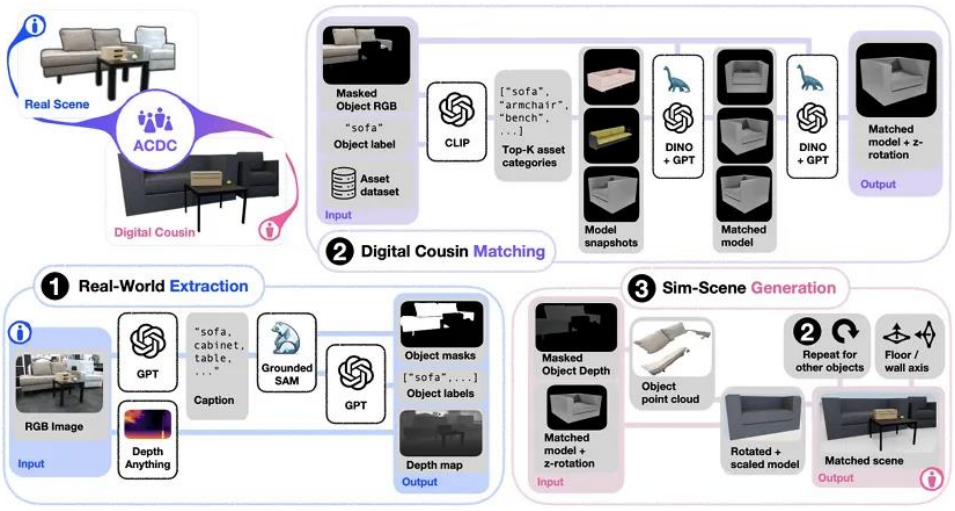
资料来源：Tianyuan Dai《ACDC: Automated Creation of Digital Cousins for Robust Policy Learning》，民生证券研究院

6.4.2 核心算法：自动创建数字表亲（ACDC）

为了实现数字表亲的自动生成，研究团队提出了名为 ACDC 的算法。**ACDC 是一个完全自动化的端到端流程，从单个 RGB 图像生成完全交互式的模拟场景，由信息提取、数字表亲匹配、场景生成三个关键步骤组成。**算法首先从输入的单张 RGB 图像中提取每个物体的关键信息，包括位置、大小、朝向等，再利用这些信息结合预先准备的 3D 模型资产库，为检测到的每个物体匹配最合适的数字表亲模型，最后对选定的数字表亲模型进行后处理和组合，生成一个物理上合理且完全可交互的虚拟场景。

通过以上步骤，ACDC 能够自动创建与输入图像语义相似但不完全相同的虚拟场景，为机器人策略训练提供多样化的环境。从而在这些环境中进一步训练机器人策略。

图69：ACDC 算法的关键步骤



资料来源：Tianyuan Dai《ACDC: Automated Creation of Digital Cousins for Robust Policy Learning》，民生证券研究院

6.4.3 核心结论与展望

研究团队设计了一系列实验，以全面评估"数字表亲"方法的有效性。首先，研究者们在 sim-to-sim 场景中对 ACDC 场景重建进行了定量和定性评估。结果显示，**ACDC 能够快速、自动地生成与单张真实世界 RGB 图像对应的交互式数字表亲场景。**这些虚拟复制品不仅在物体识别上表现出色，还能准确还原其在场景中的位置和尺寸。

图70：场景重建质量评估结果

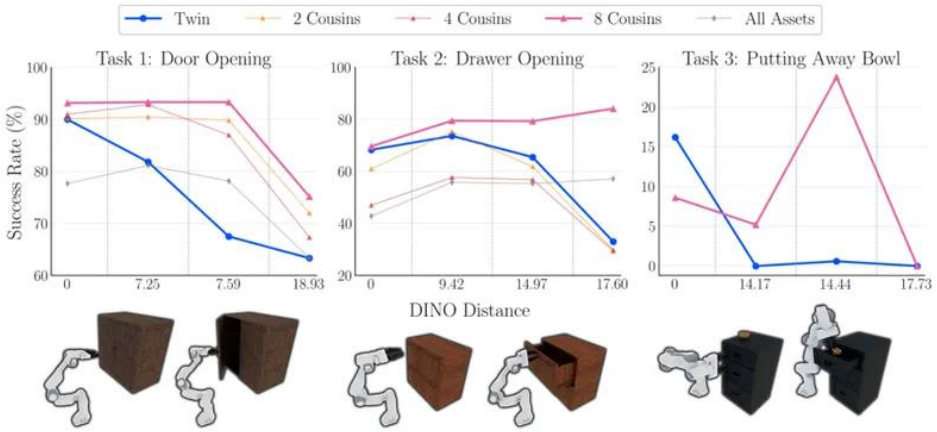
Input Scene	ACDC Output	Scale (m)	Cat.	Mod.	$\mathcal{L}_2$ Dist. (cm) ↓	Ori. Diff. (rad) ↓	Bbox IoU ↑	Cen. IoU ↑
		3.42	6/6	6/6	4.15 ± 2.04	0.10 ± 0.14	0.64 ± 0.23	0.73 ± 0.22
		4.17	8/8	8/8	7.65 ± 5.62	0.05 ± 0.00	0.66 ± 0.21	0.74 ± 0.16
		6.89	10/10	10/10	4.77 ± 3.38	0.03 ± 0.01	0.74 ± 0.20	0.77 ± 0.19
		10.23	15/15	15/15	15.67 ± 8.86	0.12 ± 0.11	0.59 ± 0.14	0.72 ± 0.14

资料来源：Tianyuan Dai《ACDC: Automated Creation of Digital Cousins for Robust Policy Learning》，民生证券研究院

数字表亲在保持分布内性能的同时，还能提供更好的分布外泛化能力。在“开门、打开抽屉和收起碗”三个典型任务中，数字表亲训练的策略通常可以匹配，甚至优于数字孪生的表现；然而，针对所有 All Assets 进行训练的策略要比数字孪生差得多，这表明朴素的领域随机化并不总是有用的。此外，随着随着测试环境与

训练环境差异的增大，数字孪生的策略性能通常会出现成比例的显著下降，但数字表亲策略的整体表现更为稳定，这表明，**数字表亲训练的策略展现出更强的鲁棒性，特别是在分布外场景中。**

图71：sim2sim 策略学习效果



资料来源：Tianyuan Dai《ACDC: Automated Creation of Digital Cousins for Robust Policy Learning》，民生证券研究院

数字表亲在真实世界中的应用表现同样出色。经过在数字表亲环境的专门模拟训练后，机器人在完全真实的厨房环境中成功完成了开启厨房橱柜的任务，从模拟到现实的迁移成功率高达 90%，有力证明了 ACDC 方法在真实场景中的适用性和有效性。结合以上实验，数字表亲方法的优势不言而喻：在原始分布上，其性能与基于数字孪生训练的策略相当；在面对分布外场景时，数字表亲表现出了更强的适应能力和鲁棒性；最为关键的是，这些策略成功实现了从模拟到现实的零样本迁移，无需额外调整就能在真实环境中有效运作，为机器人学习在复杂、多变的真实环境中的应用开辟了新的可能性。

图72：real2sim2real 全流程验证结果

Policy	Sim Success	Real Success
Twin	100%	25%
Twin + ↑ DR	70%	55%
Twin + Cousin	92%	95%
Cousin	94%	90%

资料来源：Tianyuan Dai《ACDC: Automated Creation of Digital Cousins for Robust Policy Learning》，民生证券研究院

7 投资建议

1) 关注算法训练中，需要使用的传感器公司，如视觉方案**奥比中光**，力学方案**安培龙**；

2）关注同步受益的机器人本体公司，如总成方案**三花智控、拓普集团**；丝杆公司**北特科技、五洲新春、贝斯特、双林股份、震裕科技；**

3）关注其他产业链可延伸公司。

8 风险提示

1） 机器人算法迭代进步速度不及预期：机器人的算法进步速度可能并非线性，在某些数据缺失的情况下，算法训练的进步速度可能下降。

2） 人形机器人落地场景实际需求不及预期：机器人的实际应用场景还需要结合 B 端/C 端客户的实际付费购买点，可能会与仿真环境中模拟的使用场景有差异

插图目录

分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并登记为注册分析师，基于认真审慎的工作态度、专业严谨的研究方法与分析逻辑得出研究结论，独立、客观地出具本报告，并对本报告的内容和观点负责。本报告清晰准确地反映了研究人员的研究观点，结论不受任何第三方的授意、影响，研究人员不曾因、不因、也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

评级说明

投资建议评级标准	评级		说明
以报告发布日后的 12 个月内公司股价（或行业指数）相对同期基准指数的涨跌幅为基准。其中：A 股以沪深 300 指数为基准；新三板以三板成指或三板做市指数为基准；港股以恒生指数为基准；美股以纳斯达克综合指数或标普 500 指数为基准。	公司评级	推荐	相对基准指数涨幅 15%以上
		谨慎推荐	相对基准指数涨幅 5%～15%之间
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上
	行业评级	推荐	相对基准指数涨幅 5%以上
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上

免责声明

民生证券股份有限公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。

本报告仅供本公司境内客户使用。本公司不会因接收人收到本报告而视其为客户。本报告仅为参考之用，并不构成对客户的投资建议，不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，客户应当充分考虑自身特定状况，不应单纯依靠本报告所载的内容而取代个人的独立判断。在任何情况下，本公司不对任何人因使用本报告中的任何内容而导致的任何可能的损失负任何责任。

本报告是基于已公开信息撰写，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、意见及预测仅反映本公司于发布本报告当日的判断，且预测方法及结果存在一定程度局限性。在不同时期，本公司可发出与本报告所刊载的意见、预测不一致的报告，但本公司没有义务和责任及时更新本报告所涉及的内容并通知客户。

在法律允许的情况下，本公司及其附属机构可能持有报告中提及的公司所发行证券的头寸并进行交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问、咨询服务等相关服务，本公司的员工可能担任本报告所提及的公司的董事。客户应充分考虑可能存在的利益冲突，勿将本报告作为投资决策的唯一参考依据。

若本公司以外的金融机构发送本报告，则由该金融机构独自为此发送行为负责。该机构的客户应联系该机构以交易本报告提及的证券或要求获悉更详细的信息。本报告不构成本公司向发送本报告金融机构之客户提供的投资建议。本公司不会因任何机构或个人从其他机构获得本报告而将其视为本公司客户。

民生证券研究院：

上海：上海市浦东新区浦明路 8 号财富金融广场 1 幢 5F； 200120

北京：北京市东城区建国门内大街 28 号民生金融中心 A 座 18 层； 100005

深圳：广东省深圳市福田区益田路 6001 号太平金融大厦 32 层 05 单元； 518026