/Author (Homer Simpson) /Title (Robots: Our new overlords) /CreationDate (D:20101201120000) /Subject (Robots) /Keywords (Robots;Overlords)

NeSyPack:
一种用于双手物流包装的神经符号框架

Bowei Li *,1 , Peiqi Yu *,1 , Zhenran Tang 1 , Han Zhou 1 , Yifan Sun 1 , Ruixuan Liu 1 and Changliu Liu 1
* Equal Contribution, 1 Carnegie Mellon University

Abstract—**本文介绍了 NeSyPack，一种用于双手物流包装的神经符号框架。我们的 NeSyPack 结合数据驱动模型和符号推理，构建了一个可解释的分层框架，该框架具有广泛的适应性、数据高效性和可靠性。它通过分层推理将任务分解为子任务，进一步通过符号技能图管理的原子技能进行分解。该图选择技能参数、机器人配置和任务特定的控制策略进行执行。这种模块化设计使得系统具有稳健性、适应性和高效复用性，优于需要大规模重新训练的端到端模型。使用 NeSyPack，我们团队在 2025 年 IEEE 国际机器人与自动化会议（ICRA）上赢得了双手能做什么（WBCD）竞赛的第一名。**

I. 介绍

物流包装是仓库行业中的一项关键任务，需要工作人员选择适当的物品并将其装入运输箱。现有的研究假设物体具有光滑的表面，并使用吸嘴抓手 [1, 14, 18] 。然而，具有不平整或漏气表面的物体无法通过吸嘴抓手操控。最新的工作 [6] 集成了视觉和触觉反馈，设计了一种创新的双指抓手，使机器人能够拾取物品并将其存放到货架上。然而，它考虑的是单臂操作，执行复杂任务如封装箱子仍然困难。仿人机器人最近的进步使人们关注到双手系统，这提供了更高的效率、更大的工作空间和更灵活的操作能力 [5] 。然而，如何让机器人迅速学习不同物品的包装技能仍在探索中。一种广泛研究的方法是视觉-语言-动作模型（VLA）[3, 9, 7] ，使机器人能够执行各种不同的任务，如折叠衣服、悬挂杯子、擦拭桌子等。[9] 使用感知模型 [13, 17] 提取视觉特征，并使用大型语言模型（LLM）[15] 将提取的特征编码为隐藏状态。然后，他们训练一个动作策略，即动作解码器，从隐藏状态生成机器人动作指令。[3, 7] 将 VLA 管道分解为一个较大的视觉-语言模型（VLM）主体 [2] 和一个较小的动作专家模型，用于从 VLM 输出生成机器人指令。然而，基于 VLA 的方法无法解释，并且需要大量高质量数据以实现良好的性能和通用性。而在工业环境中获取大量高质量数据是困难的，甚至是不可能的。[14] 采用了一种基于模型的方法，他们训练了一个神经网络模型来提取物体的几何特征，并手动设计了一种抓取策略，即使用吸盘抓手从可见平面中心抓取。然而，由于专用的抓手和单一的统一动作策略，难以推广到更复杂的任务中。

本文提出了 NeSyPack ，一种神经符号框架 [16] ，用于双手物流打包，使机器人能够有效地学习和执行涉及各种对象的打包任务。我们的方法在多个层面上结合了数据驱动模型和符号推理。首先，感知特征被符号化表示——例如，跟踪可变形对象的边缘——同时学习将原始视觉和深度数据映射到这些特征的感知模型。其次，通过人类演示获取笛卡尔空间中的操作技能，例如基于感知输入确定手的轨迹。这些操作技能从简单的帧相关偏移（例如对刚性对象的拾放）到基于神经网络的策略（例如对可变形对象的操控）。第三，虽然笛卡尔空间中的操作技能是学习得来的，但支持机器人手臂的动作是符号计算的，考虑了在跟踪笛卡尔空间运动时所有潜在的目标和约束。最后，所有学习的模型、控制策略和符号计算模块被组织成一个统一的神经符号结构——技能图。在线符号规划器利用该框架选择适当的表征、感知模型和运动策略，以实现高级任务规范。通过将符号推理注入神经模型，NeSyPack 是可解释的，并表现出强大的广泛性、高数据效率和可靠性。我们在 2025 年 IEEE 国际机器人与自动化大会 (ICRA) 的“What Bimanuals Can Do” (WBCD) 比赛中部署 NeSyPack 。由于强大的广泛性和数据效率，我们是唯一一个能够将我们的系统完全部署到全新比赛设置的小组。成功转移后，我们的系统表现出高度可靠性，最终赢得了一等奖。

II. NeSyPack ： 用于双手物流打包的神经符号框架

本文解决了物流打包问题，其中系统需要从存储箱中挑选指定的物品，将其放入运输箱中，并正确封闭运输箱。一个完整的打包任务可以表示为 $T = [(O_1, N_{O_1}), (O_2, N_{O_2}), \ldots, (O_N, N_{O_N}), S]$ ，它由 N 种不同类型的物品打包任务以及一个封箱任务 S 组成。N_{O_i} 表示物品 O_i 所需的数量。Fig. 1 展示了 NeSyPack 的系统概述，包含两个主要组件：1）分层任务推理（HTR）模块，以及 2）机器人技能图。给定打包任务，HTR 将长时间跨度任务 T 分解为子任务，并监督整体任务进度。基于监控的任务状态，技能图提取符号表示，然后推理出最佳适配的行动策略，并指令机器人执行。

如 Fig. 2 所示，对于给定的打包任务，HTR 模块将长期任务 T 分解为子任务 $o_1^1, o_1^2, \ldots, o_1^{N_{O_1}}, o_2^1, \ldots, o_N^{N_{O_N}}, S$ 。例如，对于任务“打包一个碗、两个网球并封箱”，如 Fig. 2 所示，分解后的子任务为：1）o_1^1 ：打包一个碗，2）o_2^1 ：打包一个网球，3）o_2^2 ：打包一个网球，4）S ：封箱。对于每个子任务，HTR 进一步将其分解为更小的原子任务。例如，对于子任务 o_1^1 （在 Fig. 2 中打包一个碗的 i.e., ），分解的原子任务计划可以是：1）检测碗的 3D 姿态，2）拾起碗，3）检测成功拾取，4）将碗转移到打包箱，5）将碗放入打包箱。本质上，HTR 将长期任务分解为一个逐步的执行计划 [12] ，其中每个单独的原子任务都在机器人能力范围内。由于范围减少，每个原子任务都可以被轻松参数化或学习。为了实际执行原子任务，HTR 查询如 Fig. 1 中所示的技能图。

此外，HTR 监控整体任务进度，并在必要时调整名义计划 [4] 。如 Fig. 2 所示，某些原子任务——e.g., “搅拌箱” ——并不是每个子任务 o_i^j 都需要。因此，它们被省略在默认计划之外，但可以在感知失败时作为故障恢复策略动态插入，如虚线箭头所示。

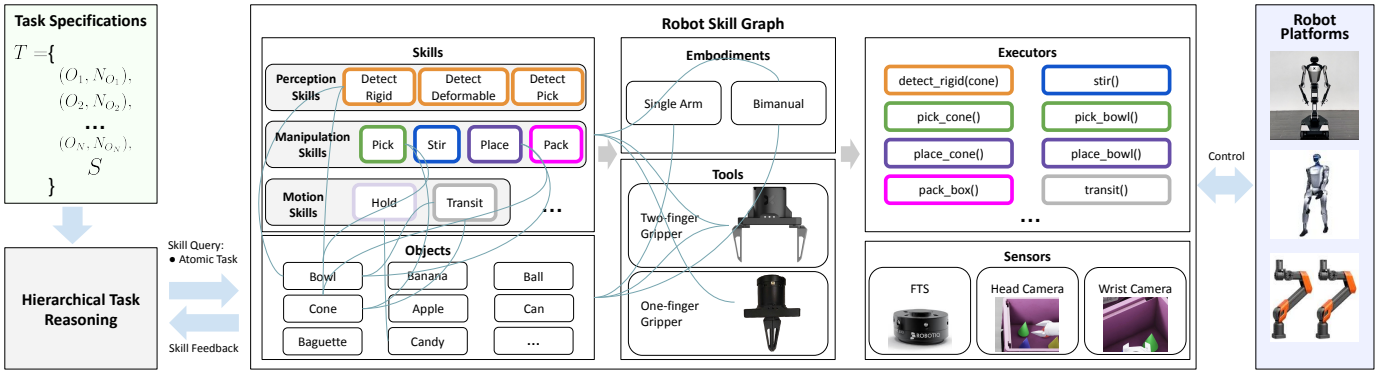


Fig. 1: NeSyPack 的概述。对于给定的包装任务 T ，HTR 将长时间的任务分解为较小的原子任务，并查询机器人技能图，以使用适当的技能执行各个原子任务。

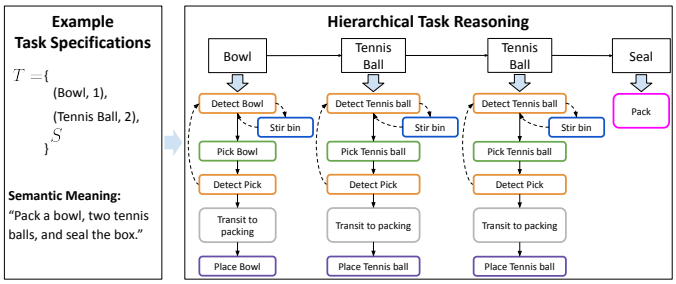


Fig. 2: HTR 在一个示例任务中的说明。实线箭头：确定的任务计划。虚线箭头：可能会更改的暂定任务计划。

A. 技能图谱

如 Fig. 1 所示，技能图以 HTR 的一个原子任务为输入，然后提取其符号表示并推理出最佳适配的行动策略。具体地，它识别出原子任务的目标物体和所需技能，确定合适的实现形式和工具，并输出接地执行功能作为机器人的控制指令。该技能图结合了学习模型和符号组件，贯穿于六个相互依赖的元素中，促进结构化和自适应的任务执行：

1) 技能：表示高级语义机器人操作，以促进直观的任务规划，包括感知技能，如“检测对象”，操作技能，如“抓取”、“放置”、“搅拌”和“包装”，以及运动技能，如“保持”和“移动”等。
2) 对象：要包装的物品，每个物品都由其物理属性（如大小、形状、重量和刚性）以及指导操作的感知特征来描述。例如，球状物体仅用位置表示，而柱状物体（如罐头）则额外包括一个从姿态估计中得出的直立标志。这些特征直接影响操作策略的选择：具有直立标志的罐头从顶部抓取，而倾斜的罐头则垂直于其纵轴抓取。
3) 实施例：定义支持的机器人配置，包括单臂和双臂。
4) 工具：指定要使用的现有末端执行器（EOAT），e.g.，“二指夹”和“一指夹”。
5) 执行器：将每个技能-对象集链接到可执行的函数，生成执行策略（预训练），同时整合体现、工具和反馈传感器。
6) 传感器：指定提供反馈的传感器。

技能图中的各个组件高度相互依赖。例如，如 Fig. 1 所示，给定一个由 HTR 指派的原子任务“拿起碗”，技能图提取所需技能为“拿起”，对象为“碗”，并确定为“双手”体现，使用单指夹持器，整合头部和手腕摄像头进行反馈。基于这些选择，技能图执行“pick_bowl()”以生成机器人动作并完成任务。这种结构化的方法有几个优点：1）其模块化设计允许机器人重用技能，从而无需在大量数据上进行端到端的重新训练即可有效地适应多样化的任务；2）它存储技能与硬件资源之间的关系，从而使机器人技能的落实高效且稳健；3）它支持技能集的高效扩展而不影响现有技能。

在技能图中，我们将机器人技能定义为一种参数化的、目标导向的能力，它能够改变世界状态或改变机器人对世界的认识，以推进特定任务的目标。根据这一定义，我们将机器人技能分为三类：感知技能、操作技能和运动技能。

感知技能 感知技能是模块化的、不依赖特定任务的组件，它们更新机器人的内部世界模型并提供符号感知特征。它们的设计确保可以灵活地集成到各种对象类型和操控环境中。

1) 检测刚体：对于已知几何形状的刚体物体（例如，罐子和碗），该技能使用 YOLOV11 [8] 对 RGB-D 输入进行实例分割，然后将掩码反投影到 3D 中以形成密集的点云。语义先验（例如，从盒子点云中移除物体点）有助于减少遮挡伪影。使用 ICP 对每个物体和盒子点云与其 CAD 模型进行配准，并使用多个初始猜测。得到的 6D 姿态被转换到机器人的基座坐标系中，以便于下游操作。
2) 检测可变形物体：对于没有 CAD 模型的形变或未建模的物体（例如衣物），系统绕过模型登记，直接分析分割后的点云。它利用遮罩和深度数据检测边界并计算显著的边缘。通过对干净、可接近的边缘进行排序，生成候选抓取点和方向，以指导对软性或不规则物体的操控。
3) 检测抓取：这个技能监测抓取是否成功。通过比较手爪姿态和从点云数据中感知到的物体姿态来判断是否成功。使用空间对齐来验证物体是否已被牢固抓取。

模块化感知技能使我们的系统能够根据对象类型和场景上下文调整策略。该设计确保了对新对象或传感器的可扩展性，并为长时间任务提供了可靠和具备上下文感知的输出。

操作技能 操作技能代表结构化的、可重用的动作原语，通过协调的机器人运动和夹持器控制与物体进行物理交互。这些技能包括任务级别的操作，如“拾取”、“放置”、“搅拌”和“打包”。每个技能都是参数化和模块化的，能够在不同的任务和平台上泛化，而无需重新训练。为了促进技能的泛化和稳健性，我们首先在笛卡尔空间中进行运动规划。具体来说，把硬约束（e.g., 运动学可达性、碰撞避免）

和软约束（e.g., 首选抓取方向、轨迹平滑度）都结合到运动规划中，作为路径点 [10, 11] 的标签。然后，我们将这些标记的笛卡尔路径点映射到配置空间。这种混合约束公式可以在杂乱或动态的环境中实现稳健且可解释的控制。

具体来说，我们按如下方式定制每个单独的操作技能。“拾取”技能通常包括移动到一个预抓取姿势，接近并执行抓取，然后提升物体。特别是，我们通过参数模板按物体类型定制“拾取”技能：

- 小型刚性物体：通过多个方向选择和评估以中心为基础的抓取姿势，以避免碰撞。
- 圆柱形物体：抓取对齐适应物体姿态——水平物体垂直于其主轴抓取，直立物体使用基于中心的策略。
- 大型或重型物体：使用两只手臂进行对角角点抓取可以改善负载分布并增强稳定性。
- 可变形物体：通过分析分割点云中的物体边界和边缘生成抓取候选项，而无需依赖 CAD 模型。

“放置”技能涉及选择目标位置，移动到该位置，并打开夹爪以放置物体。其他操作任务如“搅拌”和“打包”可能遵循类似的结构顺序，但由于其复杂性或多变性，通常通过远程操作或通过模仿学习作为一种基于箱位条件的神经策略来执行。

运动技能 不同于操作技能，运动技能不与环境进行物理交互。技能图包括两个基本的运动技能：“保持”和“移动”。“移动”技能负责关键姿势之间的无碰撞运动。在我们的系统中，我们使用 RRT-Connect [10] 生成无碰撞的双臂运动。“保持”技能使机器人在指定时间内暂停运动。它用于在关键决策、感知或协调阶段保持机器人静止，从而确保准确的后续动作和整体系统的稳定性。

III. 结果

我们考虑了十个类别中的各种对象，每个类别都提出了不同的操作挑战，如 ?? 所示。这些对象包括圆柱体（e.g., 锡罐）、立方体（e.g., 木块）、球体（e.g., 网球）和软体物品（e.g., T 恤），以及需要双手操作的更复杂的物品，如长方体和圆锥体；需要用两只手臂分开的紧密堆叠或连接的物品；需要小心包装的大型平面物品，e.g., 长棍面包；以及需要精确操作的小物品，e.g., 糖果棒棒糖。为了确保系统的通用性，开发过程中只提供了黄格中的项目。绿格中的物品是在比赛前三天才揭晓的。此外，比赛前未提供机器人硬件，因此我们在自己的平台上开发了 NeSyPack ，并在比赛前将其转移到官方平台上。

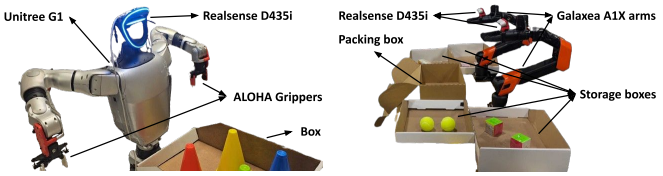
A. Unitree G1 的开发

如图 Fig. 3(a) 所示，我们在 G1 人形机器人上开发了提出的 NeSyPack 。机器人在每只手臂上配备了 ALOHA 夹持器，用于双手操作。机器人头部安装了一台 RealSense D435i 深度相机，为感知技能提供实时视觉反馈。要包装的物品放置在机器人前面的一个开口储物箱中。由于 G1 手臂的可达性有限，机器人前只有一个箱子，我们手动在储物箱和包装箱之间切换。

我们使用在准备过程中的可用物体来评估该系统，如 ?? 所示。每个物体进行了 10 次试验，如 Table. I 所示，系统在所有物体上都取得了高成功率，所有类别都达到了 9/10 或 10/10。这表明所提出的 NeSyPack 在处理不同物体的双手物流包装方面是可靠的。

TABLE I: G1 上的封装成功率

Object	Success Count	Total Attempts	Success Rate
Cube	9	10	90 %
Sphere	9	10	90 %
Stacked	10	10	100 %
Can	9	10	90 %
Cone	9	10	90 %
Small	10	10	100 %



(a) Unitree G1 (b) Dual Galaxea A1X

Fig. 3: 机器人硬件设置。左：准备设置。右：部署（WBCD 比赛）设置。

B. 2025 年 ICRA 的 WBCD 竞赛

我们参与了将在 ICRA 2025 举办的“Bimanuals 可以做什么（WBCD）”竞赛。Fig. 3(b) 展示了比赛中的硬件设置，这与我们准备阶段的设置有显著不同。比赛设置包括两个 Galaxea A1X 机械臂，每个机械臂的手腕上都有一个 RealSense D435i。由于更大的可达范围，双手系统需要处理五个盒子，包括四个储物盒和一个包装盒。机器人必须从每个储物箱中取出正确数量的物品，并将它们运输到中央的包装盒中。我们的系统在四次正式试验中进行了评估。各种物体类型的包装成功率总结在 Table. II 中。我们的试验视频可在 http://icontrol.ri.cmu.edu/news/icra2025-wbcd.html 观看。

泛化能力 我们的系统支持快速适应新的机器人平台，所需的工程工作量极少。对于不同的机器人，所有涉及机器人特定参数的计算都是符号计算的。摄像头的确切位置并不重要，因为感知特征最终会被转换到全局坐标系中以进行推理。同样，机械手和手臂的运动学参数也仅用于配置空间规划，这也是通过符号方式处理的。因此，该系统可以轻松地迁移到新平台，而无需重新训练模型或重写控制逻辑。

此外，我们的方法表现出很强的任务泛化能力。提取的感知特征和操作策略主要基于物体的几何特性。这使得系统能够很好地泛化到具有相似形状的物体上——例如相似的尺寸、边缘结构或轮廓——特别是在避免碰撞和实现稳定抓取方面。然而，由于缺乏力反馈，系统在表面材料变化上的泛化能力有限。例如，光滑的物体更容易掉落。为了解决这个问题，我们在抓手指尖增加了具有更大摩擦力的橡胶垫，这显著提高了抓持的稳固性和适应性。

数据效率 感知模型在不到一小时的时间内在 RTX 4060

TABLE II: WBCD 竞赛中的打包成功率

Object	Success Count	Total Attempts	Success Rate
Cube	2	2	100 %
Sphere	5	6	83.3 %
Large	2	2	100 %
Cuboid	4	4	100 %
Stacked	5	6	83.3 %

笔记本电脑上经过适配，只使用了每个物体类别少量的标记图像；操作技能则通过一次性关键帧演示获得，演示中指定了关键姿势和夹持器动作。

可靠性 我们原有的基于 G1 的管道在受控实验室评估中实现了超过 90 % 的任务成功率。当部署在双 Galaxea A1X 平台上时，如 Table. II 所示，这一性能水平得以保持，并在 WBCD 竞赛中获得第一名。

可扩展性 我们的神经符号框架允许无缝集成新的感知和操作技能，以应对未见过的对象类型。这些技能可以作为独立模块添加或通过增强已有模块来实现，使用有限的标注数据和演示。由于技能图的模块化设计，新增的技能不会干扰现有技能——这与整体型基础模型不同，后者在更新时经常会退化。

如 Table I and II 所示，我们的系统在开发和比赛过程中遇到了几次失败。开发过程中的一个主要挑战是 G1 机械臂的可达性有限——一些物体被放置在机器人可达工作空间之外，导致抓取失败。这部分是因为我们的技能图是在笛卡尔空间中构建的，其中路径点用语义标签（例如，必需还是可选）标注，然后再映射回配置空间。然而，这种映射不保证是完整或最佳的；在某些情况下，对于给定的笛卡尔方案可能不存在可行的配置空间运动。虽然该方法在实际中表现良好，但可以通过使整个身体控制扩展可达工作空间来减轻这些限制。在比赛期间，失败的主要原因转向了感知：腕部相机视野（FOV）有限常常导致单次观察不完整，从而产生感知错误并降低抓取成功率。结合更广视野的相机或多视角融合技术可能会缓解这一问题。未来的工作将研究如何在配置空间约束下评估笛卡尔运动计划的可行性，以及如何增强感知的鲁棒性以实现更可靠的执行。

References

[1] Amazon. Amazon introduces sparrow-a state-of-the-art robot that handles millions of diverse products, 2022. https://tinyurl.com/2p8h4w7v [Accessed: May, 2025].

[2] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726, 2024.

[3] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164, 2024.

[4] Hongyi Chen, Yunchao Yao, Ruixuan Liu, Changliu Liu, and Jeffrey Ichnowski. Automating robot failure recovery using vision-language models with optimized prompts. arXiv:2409.03966, 2024.

[5] Philip Huang, Ruixuan Liu, Shobhit Aggarwal, Changliu Liu, and Jiaoyang Li. Apex-mr: Multi-robot asynchronous planning and execution for cooperative assembly. In Robotics: Science and Systems, 2025. URL https://arxiv.org/abs/2503.15836.

[6] Nicolas Hudson, Josh Hooks, Rahul Warrier, Curt Salisbury, Ross Hartley, Kislay Kumar, Bhavana Chandrashekhar, Paul Birkmeyer, Bosch Tang, Matt Frost, Shantanu Thakar, Tony Piaskowy, Petter Nilsson, Josh Petersen, Neel Doshi, Alan Slatter, Ankit Bhatia, Cassie Meeker, Yuechuan Xue, Dylan Cox, Alex Kyriazis, Bai Lou, Nadeem Hasan, Asif Rana, Nikhil Chacko, Ruinian Xu, Siamak Faal, Esi Seraj, Mudit Agrawal, Kevin Jamieson, Alessio Bisagni, Valerie Samzun, Christine Fuller, Alex Keklak, Alex Frenkel, Lillian Ratliff, and Aaron Parness. Stow: Robotic packing of items into fabric pods. arXiv preprint arXiv:2505.04572, 2025.

[7] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054, 2025.

[8] Glenn Jocher, Jiacong Qiu, and Ayush Chaurasia. Ultralytics yolov11 (version 11.0.0) [computer software]. https://github.com/ultralytics/ultralytics, 2023. Open-source codebase, GitHub repository.

[9] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246, 2024.

[10] James J. Kuffner and Steven M. LaValle. RRT-Connect: An efficient approach to single-query path planning. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), volume 2, pages 995–1001, 2000.

[11] Changliu Liu, Chung-Yen Lin, and Masayoshi Tomizuka. The convex feasible set algorithm for real time optimization in motion planning. SIAM Journal on Control and Optimization, 56(4):2712–2733, 2018.

[12] Xusheng Luo, Shaojun Xu, Ruixuan Liu, and Changliu Liu. Decomposition-based hierarchical task allocation and planning for multi-robots under hierarchical temporal logic specifications. IEEE Robotics and Automation Letters, 9(8):7182–7189, 2024.

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023.

[14] Siemens. Siemens presents the future of intralogistics: Simatic robot pick ai pro enables machine builders to develop their own adaptive picking robots, 2025. https://tinyurl.com/4wv3n2te [Accessed: May, 2025].

[15] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-

tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[16] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-symbolic vqa: disentangling reasoning from vision and language understanding. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, page 1039–1050, Red Hook, NY, USA, 2018. Curran Associates Inc.

[17] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF international conference on computer vision, pages 11975–11986, 2023.

[18] Tianqi Zhang, Zheng Wu, Yuxin Chen, Yixiao Wang, Boyuan Liang, Scott Moura, Masayoshi Tomizuka, Mingyu Ding, and Wei Zhan. Physics-aware robotic palletization with online masking inference. arXiv preprint arXiv:2502.13443, 2025.