Innata-0：原生具身世界系统与模型和智能体协同设计

Innata团队

北京人工智能研究院

https://baai-eai.github.io/innata-0

arXiv:2611.30254v1 [cs.RO] 23 Nov 2026

摘要

物理世界开放、动态且部分可观测。尽管基础模型推动了具身智能的发展，但独立模型难以胜任长时程执行，而简单地将模型与智能体集成，可能导致规划与物理控制之间不一致。为支持物理世界中可靠且适应性强的行为，我们推出了**Innata-0**，一个原生具身世界系统。Innata-0 对具身世界—动作模型与智能体进行联合设计，以结合二者互补的优势。模型**Innata-0-WAM**采用三塔混合变换器架构，整合理解、预测与动作生成。智能体**Innata-0-Agent**通过记忆、验证与恢复来协调长时程执行。我们还提出了一种协同设计协议，使模型训练与智能体执行相协调，并利用经验证的经验支持后续适应和模型改进。大量评估表明，该系统在理解、世界预测和机器人操作方面具备强大的具身能力。Innata-0 是我们迈向物理智能的第一步，我们认为这是一段漫长旅程的开端。

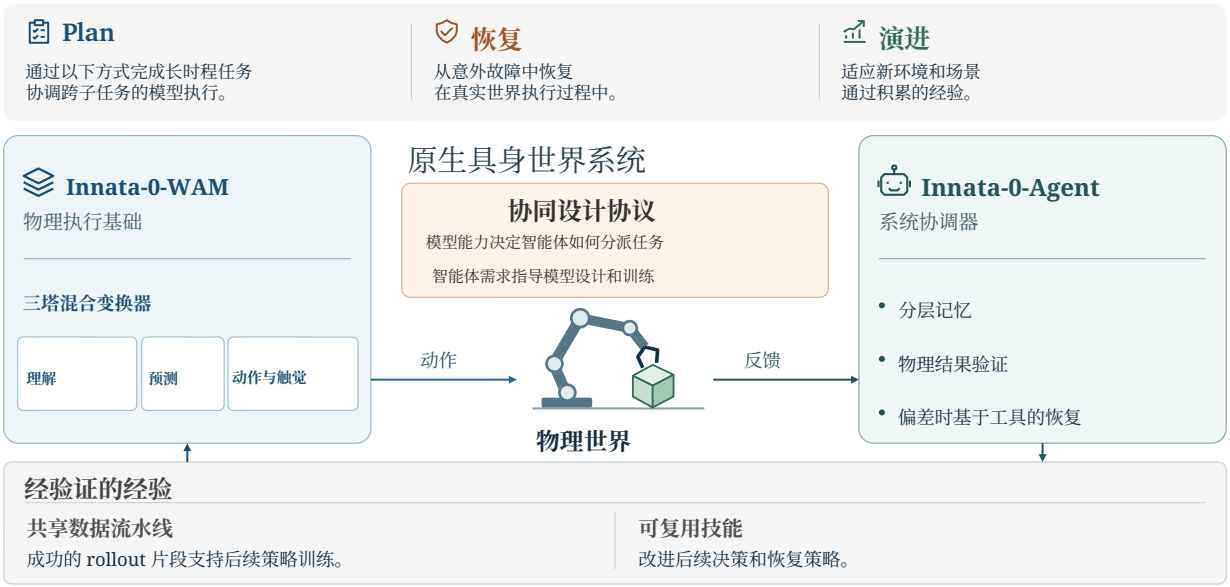


图1：Innata-0 中的原生模型—智能体协同设计。具身模型与智能体在实现规划、恢复和演进方面发挥互补作用。统一的交互协议使模型能力与智能体需求保持一致。成功的 rollout 片段支持后续策略训练，而经验证的交互和经验则被整合为可复用技能。

目录

1 NativeModel-Agent 协同设计 3

2 具身数据构建 **4**

2.1 概述 .4

2.2 统一信息提取与标注 .52.3 具身视觉-语言理解数据 .82.4 视频生成数据 .92.5 用于动作生成的数据 .92.6 用于触觉生成的数据.13

3 世界动作模型 **15**

3.1 世界模型架构 .153.2 基础设施 .19

3.3 训练 .24

4 具身 智能体系统 28

4.1 任务结构与执行 .294.2 验证与自适应恢复 .304.3 经验驱动的演化 .314.4 智能体基础设施 .33

5 评估 **35**

5.1 理解能力评估 .355.2 视频生成 .355.3 策略评估 .375.4 触觉反馈策略评估 .40

6 局限性与未来方向 **43**

7 贡献与致谢 43

参考文献 44

A 统一注释记录 64B 理解能力训练细节 67C 灵巧手 69D 触觉感知 76E 基础设施细节 78F Innata-0-Agent细节 86

1. 原生模型—智能体协同设计

物理世界开放、部分可观测且动态演变，动作的后果难以确定，失败还可能造成不可逆的代价。尽管基础模型极大推动了具身智能的发展，但独立的基础模型仍难以在长时间跨度内维持可靠行为，尤其是在执行需要自我纠正和适应时。要拓展其能力边界，就需要将原生模型能力与持续的智能体推理 [1, 2],记忆 [3, 4], 和交互 [5–7]。然而，单纯整合模型与智能体，未必能充分发挥二者的综合潜力。规划与控制之间的不一致可能导致物理执行不可靠，也限制了从积累的经验中进行适应。这促使我们从 独立具身模型 转向 原生具身世界系统。

为此，我们将原生特性概括为三项核心能力：规划通过协调模型在各子任务中的执行，为系统完成长时程任务奠定基础；恢复进一步增强系统，使其能够应对不确定性，并在真实世界执行过程中从意外故障中恢复；以及演进通过积累的经验，使系统能够持续适应新环境和新场景。为共同实现这些能力，我们推出了**Innata-0**，这是一个模型—智能体协同设计系统，推动了原生具身智能的发展。如图1所示，我们的协同设计为具身模型和智能体确立了互补的角色，使模型能够充分发挥其原生能力，同时由智能体扩展其有效运行范围。

在模型方面，**Innata-0-WAM** 作为系统进行物理执行的基础。我们采用三塔混合变换器架构 [8],，该架构具备理解、预测和动作生成能力。模型以完整指令、当前子任务和经验证的任务历史作为条件输入。通过在训练过程中融入当前计划标注和历史计划标注，Innata-0-WAM 使其学习到的条件接口与智能体对任务进度不断演进的表征保持一致。

在智能体侧，**Innata-0-Agent** 担任系统协调器，支持长时程规划、失败恢复和持续演化。它在维护分层记忆的同时，协调模型在各子任务中的执行。智能体以观测和执行反馈为依据，对照预期的物理结果进行验证，并在出现偏差时调用基于工具的恢复机制。经验证的交互和经验会被整合为可复用技能，用于改进后续决策和恢复策略。Innata-0-Agent 使其执行逻辑与模型能力相匹配，从而实现协同执行、结果验证和经验驱动的适应。

实现规划、恢复和演进，需要模型与智能体协同工作、各展所长。因此，我们建立了一套协同设计协议，使二者能够协同运作，发挥各自优势，并指导后续设计。关键在于，这套协议是双向的：模型的执行能力决定了智能体如何组织和委派任务，而智能体对规划、恢复和适应的需求则指导模型设计与训练。我们的协议使模型训练数据与智能体的任务指令、子任务计划和恢复命令保持一致，从而通过协同执行完成复杂任务。反过来，智能体验证执行结果并整合经验，以改进其任务命令和可复用的恢复策略。它还会将成功的 rollout 片段，包括纠正动作和恢复后续执行，送入共享数据流水线，用于后续策略训练。这些原则指导我们的数据构建、模型训练和智能体设计，为原生具身智能奠定基础。

我们评估 Innata-0 的基础物理能力及其在闭环任务执行中的应用。在参与比较的模型中，理解塔在九项基准测试中排名第一，在十四项中位列前二。在 WorldArena 2.0 Track 1 [9], 上，Innata-0 的 EWM 分数为 67.63，在开源模型中最高。RoboCasa365 [10] 和RoboDojo-Sim [11] 上的策略评估考察了操作与任务完成情况。在 RoboCasa365 上，Innata-0 的总体成功率为 60.2%，并在 Atomic-已见任务中排名第一，成功率达 88.4%。这些评估刻画了模型的定位、预测，

以及控制能力，并评估其在智能体引导的执行循环中的应用。

总而言之，我们提出了**Innata-0**，这是一种原生具身世界系统，将物理智能拓展到单个任务执行之外。我们对具身模型和智能体进行协同设计，以发挥二者的互补优势。共享交互协议使模型能力与智能体需求保持一致，并指导数据构建、模型训练和智能体设计。这些设计共同实现了长时程规划、失败恢复和持续演化。我们的系统协同设计为研究以下问题奠定了基础：对齐的任务上下文、物理结果验证以及经验证的经验如何促进可靠的开放世界交互。

2. 具身数据构建

2.1. 概述

原始的视觉、语言、动作和触觉数据包含关于物体、交互和物理过程的丰富知识，但其现有标注仅明确表达了其中一部分知识。遵循协同设计协议，我们明确任务目标、子任务关系和完成证据，同时保留纠正行为与后续任务执行之间的联系。因此，我们开发了面向任务的知识精炼，这是一套统一的数据构建流程，从异构数据源中提取共享标注，并针对每个学习目标和训练阶段准备相应的监督信息。在具备必要观测和标注的情况下，每个数据源都可以支持多个训练任务。由此生成的任务标注使模型的训练上下文与智能体用于规划的子任务目标保持一致。纠正轨迹和任务延续为恢复提供监督，而经验证的 rollout 片段则进入同一流程，用于后续策略训练，并通过经验复用支持演进 。

我们的初始来源池主要由开源数据集和可通过网络访问的语料库构成，包括图文对及其他视觉语言语料库、通用视频、第一人称人类视频、真实机器人录制数据和机器人仿真数据。视频提供了物理经验的主要来源，并由图文监督、原生状态与动作通道以及触觉观测加以补充。这些来源共同涵盖了多样的场景、物体、具身形态和操作任务。

来源中可获取的知识，可能多于其原始标注明确提供的监督信息。视频可能记录了物体、空间关系、中间目标和状态变化，而其描述文本可能只提及总体活动。相关证据也可能仅出现在较长录制内容中的一小段时间内。不同的学习目标需要以不同方式呈现这些信息。例如，同一段将杯子放入橱柜的视频，可以支持关于杯子的指称定位问题、对其放置位置进行指令条件预测；如果有相应的录制或重建轨迹，还可作为动作学习示例。如果有同步触觉观测，同一交互过程还可以用于预测抓取或放置时的接触与载荷。构建这些视图需要定位事件、识别其参与对象和任务结构，并以适当的粒度对齐上下文与目标。空间问题可能只需一帧，而视觉预测需要连贯的状态转移，动作预测还需要对齐的运动监督。**因此，我们会按照各个目标所需的形式明确呈现相关知识**，同时保留理解这些知识所需的上下文。这样，一段经验就能服务于多种学习任务，每个示例都聚焦于特定能力。

图 2总结了面向任务的知识精炼的三个阶段：源数据获取、统一标注和任务专用数据准备。统一标注阶段对每个样本执行一次语义解释，以创建包含层级任务标签、事件边界以及场景和动作描述的共享记录。它保留原始观测以及可用的状态、动作和触觉流，连同空间、接触和质量信息，且不筛选样本。

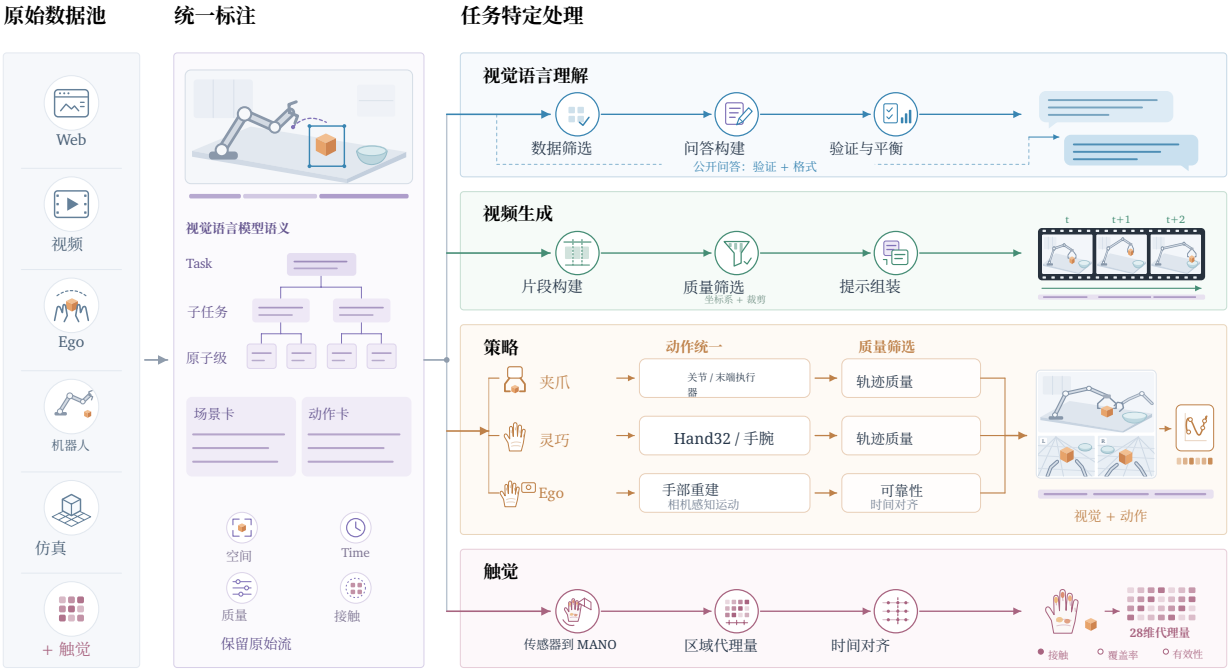


图 2：面向任务的知识精炼。异构数据源在不进行样本筛选的情况下接受统一标注。任务专用流水线复用共享标注，为视觉语言理解、视频生成、策略训练和触觉生成准备监督，并应用各自的质量标准。一个数据源可根据其可用的观测与标注支持多项任务。

视觉语言流水线利用共享的空间与任务标注来选择观测，并通过任务专用改写构建指令-目标对，随后进行验证和分布平衡。现有公开问答经过验证和重新格式化，而不重新生成。视频生成流水线利用共享的时间标注进行片段构建，针对每个训练阶段应用帧级和片段级质量标准，并从共享的场景与动作描述中组装描述性提示或指令式提示。策略流水线将原生或重建动作与共享指令和任务区间配对，统一动作语义，并利用任务、结果和质量标注进行轨迹筛选和多视图组装。对于符合条件的人类视频，这包括手部重建和相机感知动作对齐。触觉流水线将传感器观测数据与共享的接触、覆盖率和有效性记录相结合，随后进行传感器到 MANO 的对齐、区域代理构建，以及与视觉流和动作流的时间对齐。因此，同一来源可根据其可用观测和监督生成多种训练形式。

2.2. 统一信息提取与标注

我们为每个数据源附加一条共享标注记录，并在可用时保留其原始时间戳、相机视角、源元数据、原生状态与动作通道以及触觉观测。该记录将数据源级上下文与帧级及区间级标注相结合。我们提示 Qwen3.8-Flash-Next [12]对每个样本进行一次调用，在单次结构化响应中生成适用的语义、关系、时间与感知质量字段。专用感知模型向该记录添加稠密掩码，并一并纳入数据源提供的深度信息与相机标定。因此，一次视觉语言模型处理即可提供对体验的广泛描述，而专用视觉模块恢复空间测量结果，可用的触觉观测则为下游任务提供接触证据。

空间与几何标注。 我们将专用感知模型与来源提供的深度和相机标定相结合，以获取物体定位和空间信息。

• **检测、分割与跟踪**。 我们使用 Grounding DINO [13] 进行物体检测，并使用 SAM 3。1 [14, 15] 进行实例分割和视频跟踪。对于需要进一步标注的数据，视觉语言模型可提供额外的物体检测和定位。我们保留框、掩码、可用的预测得分，以及关联至共享实体记录的跨坐标系物体身份。

• **深度与三维结构**。 我们使用数据源提供的深度信息和相机测量值，并结合可用的相机内参和外参，以获取空间输入和监督信号。

语义与时间标注。 共享的视觉语言模型响应通过以下分组描述场景、交互和任务结构。

• **场景与实体上下文**。 我们标注场景类别、活动设置、工作空间类型和视角标签，包括自我中心视角、外部视角和机器人搭载视角。实体记录用于识别参与的人类或机器人、相关物体和工具、可观察到的物体属性及其初始状态。场景标签从多个层级描述环境，例如室内环境、厨房和台面工作空间。我们同时保留类别标签和简洁的自然语言描述。

• **物体关系与交互状态**。 共享的视觉语言模型响应会明确标注物体与物体、参与者与物体之间的关系，并通过以语言为依据的空间描述 [16] 补充局部实体信息。这些关系包括左/右、上方/下方、前方/后方等方向关系；内部、位于……上以及由……支撑等拓扑关系；以及接近、持有、触碰或释放等可见交互状态。每条关系都会指出参与对象，以及关系成立的坐标系或时间区间。我们还记录这些关系的变化，例如物体从支撑面移入容器，并保留每条空间描述所依据的视角或坐标上下文。

• **层级化任务与可验证的子目标**。 我们按照每段录制内容所支持的粒度，对活动类别、总体任务、子任务和原子级目标进行标注。每个子任务都明确执行实体、相关物体和预期的状态变化，并附上简洁的指令，以及用于判断任务是否完成的可观测条件。与父任务的关联和时间顺序则保留了该子任务在整体活动中的作用。这些标注提供了明确的、以执行主体为中心的单元，可用于评估执行情况和进度。

• **时间边界与执行阶段**。 我们标注任务、子任务和原子级时间间隔的起止时间，以及它们的顺序和执行阶段。对于选定的机器人示范和自我中心视角的录制内容，共享标注流程会将服务于同一操作目标的运动归为一组，并在目标发生变化时标记新的时间间隔。对于能够观察到的阶段，我们使用阶段标签描述准备、接近、交互和完成或终止。我们还识别暂停、重复尝试和目标之间的状态转移，并将观察到的状态变化与相应的时间间隔关联起来。

• **动作类型与运动特征**。 我们为伸手、抓取、抬升、传输、放置、推、拉、旋转、插入、打开、关闭、倾倒和释放等动作分配动作类型标签。任务标签描述预期目标；动作标签则描述为实现目标而采用的运动或交互方式。对于每个时间间隔，我们标注正在活动的手部或末端执行器、单臂或双臂执行方式、运动方向和表观运动幅度。幅度标签用于描述定性幅度，例如操纵器和被操作物体各自的运动幅度是小、中等还是大。我们将这些视觉评估与原生轨迹中可用的指标位移或关节运动测量区分开来，并单独描述相机或视角运动。

• **任务结果、失败与恢复**。我们根据支持任务和子任务结果的可见状态变化进行标注；当影像资料不完整时，则记录为部分完成或未确定结果。对于恢复事件，我们将先前的失败尝试与其纠正区间关联起来，并描述纠正行为，例如重新抓取滑脱物体、重新对齐插入动作，或重试打开动作。我们记录纠正是否恢复了继续执行原始任务所需的状态，并在可见的情况下标识随后进行的子任务。初始尝试、恢复和后续执行的结果彼此保持区分。原生成功标志与

基于视频的评估。

静态和动态描述卡。 我们将同一结构化响应中的场景描述和动作描述整理为 静态 和 动态卡片。当一段录制内容始终处于同一场景时，其中的片段共用一张静态卡，描述工作空间、光照以及保持不动的物体；当场景有所变化时，则为每个片段保留一段场景描述。每个片段都有一张动态卡，其中包含简短的动作摘要，以及对参与对象、运动和状态变化的详细描述。对于连续动作，描述会跟随运动展开；对于离散变化，则记录初始状态、动作和可见结果。在第一人称视角影像中，我们会区分视角变化与手部和物体运动。如果能看到操作，动态卡还会包含相应指令。附录中的图 A.1展示了同一段录制内容中的三个片段如何共用一张静态卡，同时保留各自不同的动作描述。

视觉质量、可见性和时间完整性。 我们将技术属性和感知评估分别归入不同的组。

- **录制属性**。 我们从源元数据中提取分辨率、宽高比、帧率、时长、编解码器和比特率。我们保留派生片段与原始帧时间戳之间的对应关系。
- **感知质量和时间连续性**。 共享的视觉语言模型响应会评估失焦和运动模糊、曝光、光照、压缩伪影、相机抖动和抖动现象。它还会识别镜头切换、明显的画面冻结或中断、视角突然变化，以及可见动作中的不连续。我们将局部质量问题与其受影响区间关联起来。

- **具身主体和物体可见性**。 我们会明确标注机器人或人体、手臂、手或夹爪、被操作物体以及交互区域是否始终可见。记录会区分部分裁切、完全离开画面以及被其他实体遮挡的情况，并在可观察时记录进入或离开区间。我们会分别跟踪身体可见性、活动操作器的可见性和目标物体的可见性：机器人底座可能位于画面之外，但操作过程仍然可以完整可见。这些字段用于记录理解动作或重建其运动所需的证据是否存在。

触觉观测和接触证据。 触觉观测提供局部接触和力的信息，有助于理解物理交互。不过，这些观测的可靠性取决于信号质量和感测覆盖范围，而这两者会因感测技术和传感器配置而异 [17, 18]。因此，我们会进行源特定基线校正和去噪，以抑制由传感器噪声和基线漂移导致的伪响应。对于覆盖不完整的复杂传感器布局，我们会将覆盖率和测量有效性与接触标签分开记录，避免将缺失或不可靠的读数当作未发生接触的证据。共享记录会保留这些观测和标注，并附上其源时间戳以及与视觉流和动作流的对应关系。

我们根据源模态填充字段：图像提供场景和空间标注，视频还提供时间和运动证据，而机器人记录则可提供原生状态与动作通道、触觉观测、标定和结果信息。每个准备分支都会对这些属性应用各自的质量要求。

从共享标注到聚焦监督。生成的标注支持非排他路由至四个准备分支。VLM 分支根据实体位置、空间关系、任务层级、执行阶段和结果，构建特定任务问题及目标。视频生成分支利用共享场景上下文、按时间间隔划分的动作描述，以及依阶段而定的质量筛选，准备带字幕序列。动作分支将相关时间间隔和指令与原生轨迹或重建动作结合，并按需执行对齐、重定向、轨迹筛选和多视图组装。触觉分支将触觉观测与接触标注及有效性信息结合，在通用表示中构建对齐的触觉目标。同一来源可以为多个分支提供数据，每个分支会选择适合其学习目标的观测、标注和监督

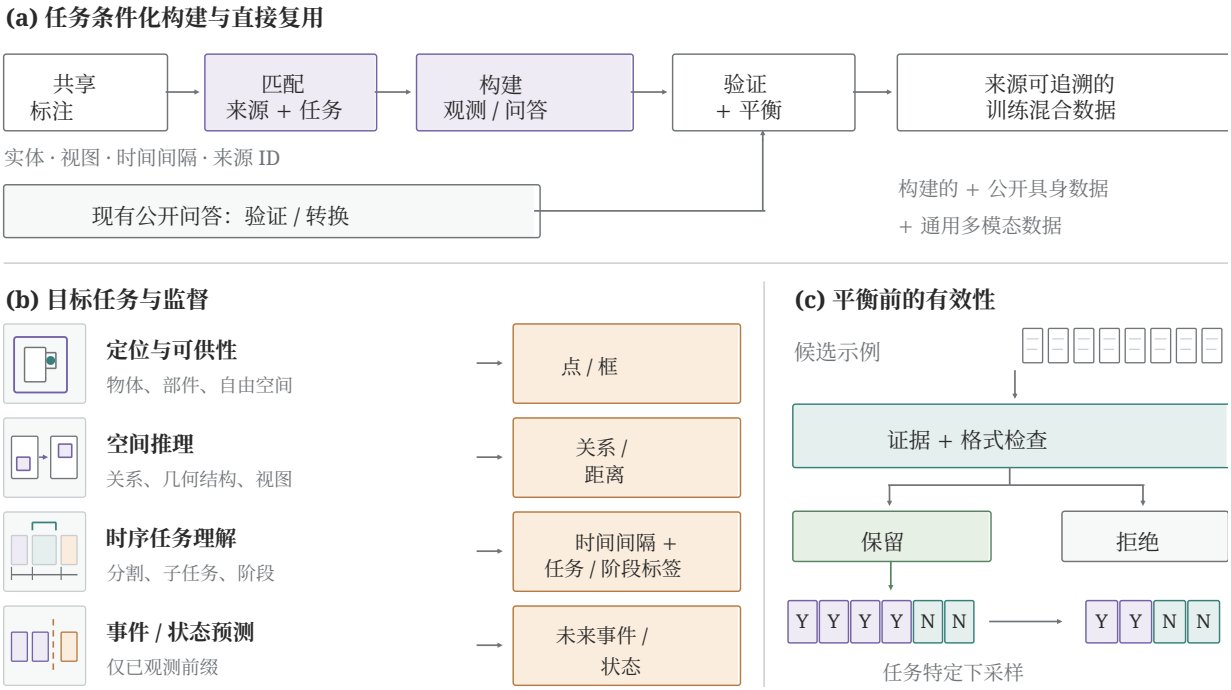


图 3：具身理解数据流水线。 共享的公开来源和源级标注支持任务条件筛选、观测—目标构建和质量控制。现有公开问答遵循独立的验证与格式转换路径。最终形成的语料库结合了构建的任务导向监督、公开具身数据和通用多模态数据。

内容。

2.3. 具身视觉—语言理解数据

VLM 分支将共享的源标注转换为用于具身感知和推理的任务导向监督。如图3所示，该分支选择与任务兼容的观测，基于可用证据构建问题和目标，并将所得样例与公开多模态训练语料库相结合。

任务覆盖范围与规范。 我们围绕物体定位、空间推理和时序任务理解组织监督，其中包括子任务识别和执行阶段识别。在证据充分时，还会纳入时间分割以及未来事件或状态预测。每项任务都定义其观测要求、相关实体或事件以及目标格式（例如文本、离散选项、空间坐标或时间区间），以指导来源选择和标注。

任务条件源选择。 我们的数据来源包括通用视频、第一人称人类视频、真实机器人录制数据、仿真数据和公开视觉—语言语料库。共享的空间、语义、时间和质量标注支持任务条件筛选，以确定可观测实体、交互和充足的上下文。筛选标准因任务而异：不适合精确定位的裁剪仍可能支持语义识别，而单一来源也可能服务于多项任务。

观测与监督构建。 对于每个选定的源—任务对，我们会整理图像、多视角图像集或视频片段，并构建指令—目标对。监督信息直接来自可靠标注，或通过任务专用改写和模型辅助标注获得。我们保留观测—标注对应关系，并统一目标格式和坐标约定。对于未来预测，目标可以基于后续观测生成，但模型输入不包含待预测的未来信息。

质量保证与分布控制。 我们会验证媒体可用性、指令与目标一致性、标注支持和任务专用格式，同时保留来源溯源信息，以便去重并开展基准重叠审计。为减少分布捷径，我们会检查任务覆盖范围和答案分布，对过度代表的二元或方向性答案进行下采样，并在保留语义和定位的前提下，重新平衡符合条件的多项选择选项。这些措施可补充标注有效性检查。

与公共多模态语料库的整合。 构建的示例与公开具身数据集和通用多模态语料库相结合。现有问答数据通过验证和格式转换纳入，无需重新生成。公共语料库提供广泛的视觉和语义覆盖，而面向任务的示例则强化了具身监督。

数据来源与覆盖范围。 我们汇集了一个异构语料库，涵盖通用多模态理解、空间感知、多视图与视频推理、语言条件定位、可供性理解以及以机器人为中心的问答。来自 Cambrian 和 LLaVA-OneVision 的通用数据提供广泛的视觉和指令遵循监督。具身数据则为空间关系、度量估计、视角变化、物体定位以及与交互相关的区域提供互补监督。表B.1 附录 B 总结了公共数据集，以及我们构建的面向任务的数据集；后者被纳入理解塔训练混合数据中。

2.4. 视频生成数据

视频生成分支基于通用视频、第一人称人类视频和机器人视频，构建 WAM 的视觉动态基础。该分支准备时间连贯的训练序列，生成与其内容和时间跨度相匹配的描述文本，并针对每个训练阶段应用相应的质量标准。通用视频提供广泛的视觉和运动覆盖，而人类与机器人交互视频则使模型的学习聚焦于具身事件。

片段构建。 我们在解码前筛选录制内容，根据适用于各数据来源的标准检查时长、帧尺寸和帧率。我们仅从指定的摄像机视角中提取每个选定区间一次，并以通用视频编码格式存储裁剪片段。随后，质量评分、描述文本生成和潜在编码都读取同一个裁剪片段文件，从而减少对特定格式进行预处理的需求。

帧级与片段级评分。 对于采样帧，我们使用中央处理器上的经典图像处理方法测量亮度、饱和度和锐度。亮度和饱和度有助于标记曝光不佳或色彩缺失的情况。Gamma [19] 评估感知质量，LAION审美预测器 [20] 评估视觉吸引力，PaddleOCR [21, 22] 定位文本。我们对每个裁剪片段采样帧的 LAION 得分取平均值。对于 OCR，我们比较各帧中的文本位置，以标记固定的角落叠加层。持久存在本身并不意味着要剔除裁剪片段：产品标签可能会在整个动作过程中持续可见。

整体评估裁剪片段的模型包括用于评估技术和美学质量的 DOVER [23] ，以及用于评估视频训练适用性的 VTSS [24]。我们将每项得分与对应的裁剪片段一同保留，并采用更严格的质量标准，从持续预训练池中筛选出质量调优子集。

提示组装。 在持续预训练的每个时期，我们为每个裁剪采样一个与其标注字段兼容的提示模板。描述性提示将场景与观察到的动作配对；指令可以与场景一同出现，也可以不包含场景。因此，同一个裁剪可以在一个时期与描述配对，在另一个时期与指令配对，如图 A.2所示。

2.5. 用于动作生成的数据

策略数据流水线从公开的真实机器人和仿真轨迹，以及从符合条件的

人类视频重建并重定向的运动中，为 Innata-0-WAM 准备配对的视觉与运动监督。共享的任务与时间标注将每个轨迹片段与其执行目标关联起来。对于表示任务进度的字幕变体，我们将完整指令与当前子任务及此前已完成的任务配对，以匹配智能体根据协同设计协议提供的任务上下文。

准备流程统一动作语义、提升轨迹质量，并汇集同步相机视角。系统执行轨迹展开期间收集的轨迹遵循相同的标注和准备流程。成功的执行片段，包括纠正动作和恢复后续执行，都会予以保留，用于后续策略训练，使经验证的系统经验能够通过共享流程提供监督。

2.5.1. 夹爪操作数据

夹爪演示数据涵盖多种机器人具身形态、工具和控制接口，因此相同的向量形状可能表示不同的物理量。位姿可能描述腕部法兰或抓取中心，而动作字段可能包含控制器指令或后续测量状态。我们围绕监督信号的物理含义进行统一，使不同具身形态的演示数据能够提供兼容的学习目标，同时保留其运动学差异。因此，准备流程分两个阶段进行：**先进行语义统一，再进行轨迹质量筛选**。第一阶段明确每个通道所代表的内容及其适用时机；第二阶段则评估所得轨迹的可靠性。

阶段 1：统一动作语义。对于每种具身形态，我们首先确定源数据中各字段的含义，然后将每条观测与其动作目标相匹配。这些配对构成了两种机器人运动表示的基础：一种基于预定义的统一末端执行器（EEF）空间，另一种基于关节空间。

规范化末端执行器几何结构与运动。我们在闭合参考构型下，将工具中心点（TCP）固定定义于有效指垫抓取区域的中心。其 +Z 轴从手掌指向指尖， +X 轴沿着朝向指定手指的张开方向， +Y 轴则补全为右手坐标系。匹配的工具与安装变换将源位姿映射到该 TCP；有测量关节数据时，可通过正向运动学获得位姿，而直接使用位姿并不合适。未经过历史标定的规范化几何结构会定义一个明确标识的虚拟 TCP。

EEF 状态记录当前 TCP 在指定参考坐标系中的位姿，例如相机坐标系或机器人基座／机身坐标系 [25]。双臂机器人形态中的两条手臂共用一个参考坐标系。**我们根据连续测量状态推导 EEF 动作的位姿分量，并在当前 TCP 坐标系中表示当前至下一时刻的 TCP 变换**。当前 TCP 定义了该增量的原点和坐标轴。规范化位姿与局部增量如下：

$$
\begin{aligned}
T^{B}_{\mathrm{TCP},t} &= T^{B}_{F,t} H^{F}_{\mathrm{TCP}}, \\
\Delta T_t &= \left(T^{B}_{\mathrm{TCP},t}\right)^{-1} T^{B}_{\mathrm{TCP},t+1}.
\end{aligned}
\tag{1}
$$

其中， F 表示记录的末端执行器坐标系， H^{F}_{TCP} 是相对于 F的规范 TCP 坐标系的固定位姿，其中包括工具安装变换； B 表示状态转移过程中的公共参考坐标系。我们先计算相对旋转，再对其进行编码，而不是直接相减旋转编码 [26]。两个位姿在整个状态转移过程中必须共享同一公共参考坐标系。如果有必要的信息，我们会补偿相机或机身运动，并记录任何固定参考系假设。

每条机械臂占用十个通道。状态包含三个 TCP 位置分量、六个绝对旋转分量以及测得的夹爪开口。动作包含三个局部平移分量、六个相对旋转分量以及一个绝对夹爪目标。6D旋转编码将相应旋转矩阵的前两列拼接起来 [27]。

关节状态和目标。我们在每种具身形态中按名称排列关节，旋转关节位置使用弧度，移动关节位置使用米。对于普通关节位置通道，状态包含当前测得的位置，动作包含下一时刻测得的绝对位置，不受原生控制器接口影响。可用的头部、躯干和其他关节仍会予以表示，而移动基座通道则保留其声明的位姿或位移语义。

夹爪开口和控制信号。 我们根据 URDF 或制造商文档中规定的限位，将观测到的夹爪开口归一化到 0（完全闭合）至 1（完全张开）的尺度。某些来源提供单独的观测信号和控制信号，命令有时会编码为离散值，例如 {−1，0，1}。预训练期间，我们优先使用观测到的夹爪状态；随后在监督微调期间使用控制信号，以匹配目标机器人的控制接口。

布局与存储。 Joint 使用固定的128通道布局，涵盖手臂、夹爪、手部、头部、躯干、移动底座、腿部和扩展部件。EEF 使用单独的20通道布局，每条手臂占十个通道。单臂数据仅使用前十个通道。不可用通道以零填充并进行掩码处理，同时使用独立的状态和动作有效性掩码。LeRobot v3 回合 [28] 保留源时间戳、行映射、各视图的视频偏移量以及任务引用。显式的时间戳映射可确立观测—目标对齐；归一化与语义统一保持分离。原始值、变换元数据、代理值和假设均保持可追溯。

阶段2：轨迹质量筛选。 我们会检查标准化轨迹中是否存在非有限活动值、缺失观测、无效掩码或旋转，以及回合或媒体边界不一致等问题。在时间基准和参考坐标系适用的情况下，运动诊断会检查位移、角度变化、速度、加速度、加加速度、局部趋势残差和振荡。阈值取决于具身形态、表示方式、字段和采样机制。命令—状态检查仅使用彼此独立且语义可比的命令；由状态派生的目标不能独立验证其来源状态。

我们会结合这些诊断结果以及共享任务、子任务、结果和视觉质量标注，拒绝有缺陷的轨迹或时间间隔。轨迹有效性与任务结果分开评估：即使先前尝试失败，成功的纠正片段仍符合条件，同时保留其与失败尝试及后续任务执行的关联。当源结果标签可靠时，其优先级高于模型评估。仅凭动作幅度小或加加速度高，不足以判定数据损坏：任务上下文有助于区分重复空闲与具有信息价值的保持、接触、恢复和等待。

移除一个时间区间会将一条轨迹拆分为连续回合。我们会在相机视角之间以及配对的关节/末端执行器表示之间同步切分，保留包含完整子任务描述且已调整时间戳的数据，并在边界处重新计算或使派生目标失效。目标绝不会跨越重置或移除的间隔；即使没有有效的下一步目标，也可以保留终止观测。配对表示计为一次物理演示。

选择性平滑仅适用于经过验证的噪声候选项。该过程作用于连续的位置或关节序列，并在旋转空间中处理旋转，同时保护接触事件、夹爪状态转换和任务边界。我们会重新计算由观测派生的目标，保留原生命令，并且仅在一致性检查和事件保留检查均通过时保留纠正结果。

2.5.2. 灵巧操作数据

灵巧操作示范捕捉协调的手指关节运动，用于抓取适应和手内操作，从而扩展夹爪数据所提供的开合监督。手部机构、关节编码和末端执行器坐标系的差异，使这种监督难以在不同具身形态之间组合。我们沿用第2.5.1节中针对夹爪数据所述的两阶段流程，复用其视觉处理、语义标注和通用轨迹检查。此处详述对齐关节式手部及其运动所需的适应。表C.6 （附录C.2.3 ）列出了动作学习数据源，包括互补的仪器化示范和机器人示范。

统一手部表示。 机器人手在关节数量、轴和连杆布局上各不相同，因此其原生坐标并不能提供共同的关节运动标准。尽管存在这些结构差异，手部仍借助共同的物体可供性完成相似目标，例如抓取把手或转动盖子。这些共享交互促使我们按其功能角色对齐手部运动，从而使不同具身形态的示范能够关联到共同的任务语义与接触证据。我们因此设计 Hand32 时围绕人手解剖结构，并遵循 FAAS [29]的功能对齐原则。每个坐标标识一根手指、一个关节层级和一种运动功能：屈曲、外展或扭转。拇指贡献七个角坐标，其余每根手指各贡献六个，共计 31 个共享坐标，并保留一个坐标用于已声明的具身特定扩展。机器人示范与人类示范因此共享关节运动接口，同时保留每只手所支持的运动。

多样手部机构映射。人手参考模型并不意味着关节一一对应：机械关节可能位于不同位置，多个关节轴可以实现同一种手指功能，单个电机也可以驱动多个关节。因此，映射既要考虑预期运动，也要考虑产生该运动的机构。我们首先根据来源运动学链确定相对应的手指功能，并统一关节轴、中性位置和单位。若直接对应关系不足以满足需求，则通过逆运动学匹配对齐后的手指端点，以支持多个来源自由度与参考自由度之间的映射。三指或四指手部采用明确的功能性手指对应关系，并对缺失的手指和运动进行掩码处理。已知的模仿关系仍作为耦合约束处理，而整指驱动值则使用声明的等效角度转换。实测关节运动与推导出的关节运动仍予以区分，并单独记录独立控制的可用性、观测有效性和目标有效性。

关节与末端执行器对齐。与夹爪数据一样，我们将灵巧运动组织为关节和 EEF 表示。关节数据描述机械臂构型及其 Hand32 关节运动；EEF 数据描述手部的位置和方向，并在可用时单独记录手指关节运动。表 C.4 列出了手部局部 Hand32索引，表 C.5 则汇总了设备映射。记录的末端执行器位姿可能以法兰、手掌或跟踪设备为参考，它们各自具有不同的原点和方向。利用匹配的安装变换和运动学模型，我们将这些位姿转换到规范手腕坐标系，其原点位于参考手的手腕关节处。在手掌朝下、手指朝前的中性构型下， $+X$、$+Y$和 $+Z$ 轴对双手均分别指向右、下和前方。该坐标系会在整个轨迹中随手腕刚性运动。轨迹审查还会考虑手指关节运动和任务上下文：手腕静止时也可能伴随抓取调整，而手部保持稳定也可能表示一次有意义的保持动作。

仪器化手套数据。 仪器化手套数据（包括 PaXini 的 OmniSharing 录制内容） [30],通过关节传感器和已知的机械连杆提供关节运动，无需进行基于图像的手部姿态重建。我们将这些录制内容作为一种特殊的灵巧操作具身形态进行处理，按照相同的功能规则，将手套运动学映射到 Hand32 。对于 PaXini 的非拇指手指，将两个手套远端铰链的数值相加，得到相应的远端屈曲坐标。在有可用的手根位姿时，还会通过源到手腕的标定，将其置于规范手腕坐标系中，以进行末端执行器监督。因此，手指关节运动和手部位置通过彼此独立且匹配的变换进入共享表征，同时保留它们与视觉观测及可用触觉观测之间的对应关系。

2.5.3. 自我中心人类数据

为了从第一人称人类视频中学习，我们将第2.5.1节中的两阶段流程调整为适用于重建的手部运动。我们保留 Hand32 以及为灵巧操作数据引入的规范腕部约定。阶段 1 中，我们分别恢复并表示手部运动和相机运动，以适应视角变化。阶段 2 中，我们根据重建可靠性筛选轨迹，并调整其时间尺度，使其与机器人轨迹的运动节奏相匹配。

手部重建与功能对齐。我们使用 `Wan2.2-Fun-5B-Control` [31, 32]的特征和 ACE-Ego-Hand [33]的解码器架构，在 13 个数据集上训练手部重建模型，并采用互补的关节与 MANO 监督。我们将该模型应用于第一人称视频，以估计手部骨架、MANO 参数 [34], 以及手部存在和可见性分数。对于较长的录制内容，我们融合重叠预测，为每只手在每段连续序列中估计共享形状，并在固定形状的情况下优化手部位置。我们将重建的关节运动转换为 `Hand32` 功能角度，并应用固定的源根节点到手腕的变换，使手部与规范手腕对齐。在共享形状下进行正向运动学计算，可确保关节运动与重建几何形状保持一致。由此得到的规范手腕姿态和 `Hand32` 关节运动作为阶段 1 中相机感知运动表示的输入。附录 C.1.4 提供了更多细节。

阶段 1：相机感知运动表示。我们统一不同来源数据中的单位、左右手系、MANO 规范和光学相机坐标。当头部追踪器姿态可用时，我们应用固定的追踪器到相机的变换，以获得相应的相机位姿。

令 $G_t = {}^{W}T_{C_t}$ 表示公共度量世界坐标系 W 中的主摄像头位姿，$P_t^h = {}^{W}T_{E_t^h}$ 表示手部 $h \in \{L, R\}$ 的规范手腕姿态。变换 ${}^{A}T_B$ 将列向量坐标从 B 映射到 A，光学相机轴分别指向右、下和前方。我们以每个连续追踪片段中的首个有效相机位姿 $A = {}^{W}T_{C_0}$ 为锚点，并在从该片段提取的所有回合中重复使用此锚点。

我们将相机位姿与其源帧同步，并在世界空间中使用置信度加权样条平滑手腕位置，使用符号连续的四元数样条平滑手腕旋转 [35], ，且在每个有效连续区间内分别处理。手腕动作 D_t^h 遵循方程 1中的局部增量定义，根据连续的世界空间规范手腕姿态计算得出。此外，我们还定义相机相对手腕状态、锚定相机状态和局部相机运动动作，如下所示：

$$\begin{aligned} X_t^h &= G_t^{-1} P_t^h, \\ S_t^C &= A^{-1} G_t, \quad D_t^C = G_t^{-1} G_{t+1}. \end{aligned} \tag{2}$$

这里，X_t^h 是相机坐标系手腕状态，S_t^C 是锚定相机状态，D_t^C 是局部相机运动动作。结合世界坐标系手腕动作，这些量能够区分手部运动与相机运动，同时保留相机相对手腕状态。

最终得到的 128 通道表示 `Ego128` （表C.7，见附录C.2.3），包含双手腕、双手的 `Hand32` 关节运动以及主摄像头。平移以米为单位，旋转采用旋转矩阵的前两列构成的 6D 表示，如第 2.5.1 节所述。手部关节运动将 q_t^h 用作其状态，并将下一步的绝对关节运动 q_{t+1}^h 用作其动作。

阶段 2：重建可靠性与时间对齐。 我们通过手部存在、可见性、源置信度、有效时长和时间覆盖率来评估重建可靠性，并采用源特定阈值以及关节级或手部级可靠性掩码。我们还应用用于机器人数据的图像质量检查。我们对平滑后的世界空间手腕轨迹进行筛选和分段，去除突发的运动不连续以及双手均未被跟踪的时间间隔，以保留可用的交互片段 [36]。第 2.2 节中的任务与时间标注进一步指导完整操作子任务的选择。

最后，我们使用 2 的全局减速因子和 30Hz 的训练时间基准 [37–39]，将人类轨迹与机器人训练数据对齐。我们对手腕和相机位姿以及手部关节运动进行联合重采样：对平移和有效的关节角轨迹使用连续插值，对旋转使用球面插值。

2.6. 触觉生成数据

触觉观测可直接提供接触和载荷的证据，而仅凭视觉观测可能无法充分辨识这些信息。不过，触觉观测的形式因传感原理而异：电阻式、磁性和光学传感器可以

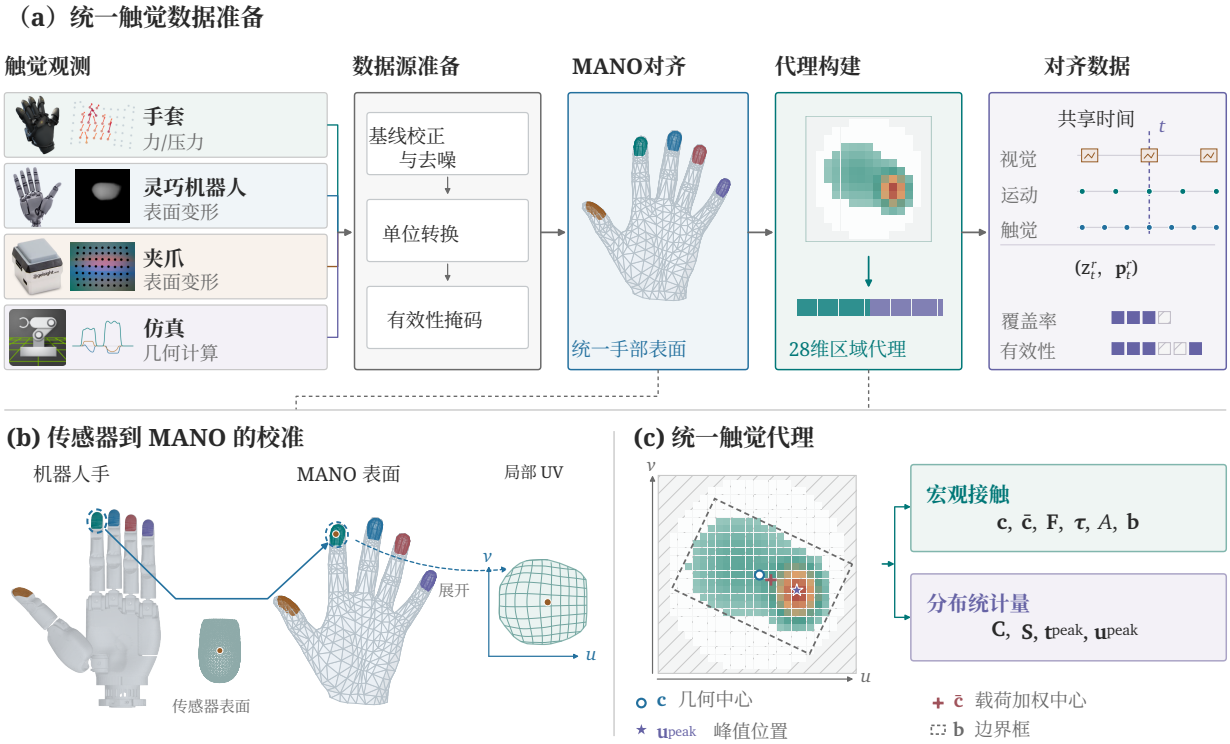


图 4：统一触觉数据准备。异构触觉观测经过数据源准备、对齐到统一的 MANO 手部表面，以及区域代理构建。由此得到的存在性–代理对会与视觉和运动观测在时间上对齐，同时分别记录传感器覆盖范围和特征有效性。

输出压力读数、方向力阵列或表面变形图像，其物理可观测性各不相同。 [18]不同具身形态中的传感器密度、覆盖率和放置位置也各不相同，从独立的指尖垫到覆盖手指和手掌的阵列皆有。完成数据源准备后，我们通过统一的几何参考系以及用于表示接触几何和载荷的统一触觉代理，对这些异构观测进行对齐。图 4概述了准备流程。

传感器到 MANO 的对齐。我们使用 MANO [34] 作为统一的手部表面。为适应手部结构和传感器安装方式的差异，我们开发了一种传感器到 MANO 的校准工具。在 FAAS [29], 的功能对齐原则指导下，该工具将每个传感器的覆盖区域及局部方向与相应的 MANO 区域对齐，例如，将单个夹爪垫对齐到食指指尖。校准会为每种传感器布局建立固定对应关系，并在不同录制内容间重复使用，以便将传感器测量值一致地映射到手部表面上。固定的区域 UV 图表提供二维坐标，用于描述手部表面上的接触位置和形状。映射后的传感器覆盖区域标识出该表面上可观测的部分，因此未感测区域仍保持掩码状态。

统一触觉代理。 为适配不同的信号维度和空间密度，我们将对齐后的接触场汇总为区域触觉代理，用于描述宏观接触和分布统计量。对于传感器标定支持绝对力的情况，测量值会转换为牛顿；压力则在其物理支撑面积上积分，方向力则以通用局部坐标系表示。对牵引（单位面积上的力）进行表面积分，可得合力及其相对于载荷加权接触中心的力矩，从而构成六维力螺旋。未经标定的源则保留其相对响应语义和分量有效性掩码。

我们根据三维表面几何结构计算接触面积和力臂，并用上文定义的区域 UV 坐标描述接触中心和覆盖区域形状。存储的表面对应关系可将这些 UV 特征映射回 MANO 手部。区域代理为

$$\mathbf{p}_t^r = \big[\mathbf{c}_t^r;\ \bar{\mathbf{c}}_t^r;\ \mathbf{F}_t^r;\ \boldsymbol{\tau}_t^r;\ A_t^r;\ \mathbf{b}_t^r;\\ \operatorname{symvec}(\mathbf{C}_t^r);\ \operatorname{symvec}(\mathbf{S}_t^r);\ \mathbf{t}_{\text{peak},t}^r;\ \mathbf{u}_{\text{peak},t}^r\big] \in \mathbb{R}^{28}. \tag{3}$$

第一组描述宏观接触：几何和载荷加权 UV 中心 $\mathbf{c}$ 和 $\bar{\mathbf{c}}$、合力$\mathbf{F}$、力矩 $\boldsymbol{\tau}$、接触面积 A，以及边界框尺度、各向异性和方向$\mathbf{b}$。其余字段保留更精细的分布信息：牵引协方差$\mathbf{C}$的六个独立元素、载荷加权 UV 协方差$\mathbf{S}$的三个元素、三维峰值牵引向量$\mathbf{t}^{\text{peak}}$，以及其二维位置$\mathbf{u}^{\text{peak}}$。这里，symvec 将对称矩阵的独立元素打包为向量。这些统计量描述载荷变化、空间集中度和局部峰值，同时避免将表示形式绑定到固定的触觉像素数量。

每个代理项都与同一 MANO 区域和时间戳处的接触存在标签配对，形成同步配对（z_t^r，$\mathbf{p}_t^r$）。传感器覆盖范围和特征有效性与该标签保持独立，以区分缺失观测与测得的无接触状态。特定来源归一化保留了耦合方向分量的共享尺度。这些配对在共享时间轴上与视觉和运动观测保持对齐，并适用于不同采样率，从而提供触觉输入和预测目标。

3. 世界动作模型

我们开发了 Innata-0-WAM，将其打造为统一的世界动作模型（WAM），以连接视觉—语言理解、预测与具身控制。根据协同设计协议，WAM 基于观测以及智能体维护的任务上下文生成动作块。反过来，其执行能力也会指导智能体如何组织和委派任务。基于这一共享任务上下文进行训练，为规划提供执行基础；纠正轨迹则教会模型应对偏差并继续执行任务，以实现恢复。经验证的经验通过智能体对任务上下文的优化，以及在后续策略训练中重用精选轨迹，为演进 提供助力。

视觉观测为在场景中对语言进行定位、理解物体及其关系提供语义证据 。当提供动作时，视觉观测也能作为预测上下文 ，通过视频预训练学习时间动态，包括可观测到的动作后果。对于策略学习，观测历史提供控制上下文 ，用于解读当前情境并生成可执行的动作序列。

我们通过混合变换器（MoT）架构整合这些作用，该架构对动作、视觉动态和触觉反馈进行联合建模。我们借助分布式训练与推理框架支持这一架构，以处理可变长度的多模态序列。在使理解塔适配具身任务后，我们通过视频生成和动作条件预测训练世界预测塔。随后，策略预训练引入动作生成，接着进行机器人状态条件化和动作–触觉联合学习。

3.1. 世界模型架构

3.1.1. 世界理解塔

理解塔采用以 Qwen3-VL- 8B 为初始权重的仅解码器视觉语言架构 [16]。它由视觉编码器、视觉—语言合并器和自回归语言模型组成。视觉编码器从图像和视频帧中提取视觉表征，再通过合并器将其投影到语言模型的嵌入空间中。给定视觉观测和语言指令，模型会生成自然语言响应和结构化空间预测，例如点和边界框。

3.1.2. 世界预测塔

世界预测塔采用 Cosmos 3 Nano 的 8B 主干网络 [40], , 包含 36 个 Transformer 层，隐藏宽度为 4096。视频变分自编码器（VAE）将观测编码为时空潜变量，再将其切分为补丁并投影为视频令牌。在语言和可用视觉上下文的条件下，该塔预测目标视频潜变量；同时提供与时间对齐的动作轨迹时，它还会预测相应的视觉变化。视频生成是用于学习时间动态的预训练目标，重点关注与操作相关的空间位移和物体交互。

我们在视频潜空间中使用条件流匹配训练该塔。将高斯噪声与干净的目标潜变量进行插值，该塔则根据噪声水平预测沿此路径的速度。损失作用于目标视频潜变量；观测到的条件帧保持干净，不计入损失。这种形式既支持根据视觉和语言上下文进行预测，也支持动作条件世界建模。

3.1.3. 动作与触觉生成塔

动作与触觉生成塔是一个独立配置参数、约有 10 亿参数的 Transformer。它保留了两个 8B 塔的 36 层深度，同时将隐藏宽度缩减至 1280，前馈宽度缩减至 3840。每层使用 32 个查询头、8 个键/值头，头部维度为 128，从而与更大规模塔的注意力接口保持兼容。这些兼容的注意力维度使不同隐藏宽度的塔也能交换信息。

跨塔交互。 视频和动作生成令牌，以及存在时的触觉令牌，在每个样本内通过全注意力进行交互，而理解塔则向生成流提供条件特征。生成令牌之间的注意力是双向的，从而能够跨模态进行联合预测。观测帧以及可用的机器人状态和触觉观测，仍作为干净条件输入；未来视频、动作和触觉令牌则是预测目标。每个生成流都以自身的噪声水平为条件。这种共享计算将动作生成与伴随的视觉变化和接触反馈预测联系起来。

域感知投影。 我们在动作塔的输入端和输出端使用域感知仿射映射，将异构控制空间连接到共享的 Transformer。对于投影域 d，每个映射的形式如下：

$$f_d(x) = xW_d + b_d, \tag{4}$$

每个域分别使用一个矩阵和偏置。输入映射将归一化并填充后的 128 维状态或动作表示投影到 1280 维隐藏空间；输出映射则生成 128 维动作空间流速度。EEF 状态和动作目标使用独立投影，因为绝对观测与相对运动目标具有不同的语义。另有一组共享的投影对，用于包含 EEF 的灵巧手和自我中心表示，包括腕部运动和手部关节运动。关节空间表示则根据其控制语义和关节布局，采用具身形态专属或共享的投影族。表 1 总结了这些投影族及其实现标识符。

动作监督。 我们将条件流匹配直接应用于连续动作块。令 a 表示干净归一化动作块，$\epsilon_a \sim \mathrm{N}(0, I)$ 表示独立高斯噪声，$\sigma_a \in [0, 1]$ 表示其噪声水平，数值越大表示输入噪声越大。我们定义

$$a_{\sigma_a} = (1 - \sigma_a)a + \sigma_a\epsilon_a, \qquad v_a^\star = \epsilon_a - a. \tag{5}$$

让 c 收集干净条件输入，包括视觉历史、语言条件化，以及可用的机器人状态和触觉观测。语言输入包含任务指令，以及对应的

表 1：领域感知投影接口。该实现提供 64 个投影槽位。兼容的源可以共享一个接口。投影 ID 用于选择输入/输出映射，并且与用于选择归一化统计量的源标识符相互独立。

投影族	状态 ID	动作 ID
标准 EEF 表示	32	0
共享的含末端执行器灵巧手与自我中心表示	62	63
关节空间表示	按具身形态或已注册族划分	

字幕变体、当前子任务和已完成子任务历史（第3.1.4节）。令 $\tilde{z}_v$ 和 $\tilde{z}_\tau$ 分别表示相应噪声级别下的视频流和触觉流（如有）。动作目标函数为

$$\mathcal{L}_{\text{action}} = \mathbb{E}\left[\frac{\left\|M_a \odot \left(\hat{v}_\theta^a(a_{\sigma_a}, \sigma_a \mid c, \tilde{z}_v, \tilde{z}_\tau) - v_a^\star\right)\right\|_2^2}{\sum M_a}\right], \tag{6}$$

其中，M_a 用于选取有效目标动作条目，并排除条件状态和填充维度。在推理时，从噪声端点向干净端点积分预测速度场，即可生成一个动作块。

在智能体循环中，连续生成的动作块利用更新后的观测来完成当前子任务。一个子任务可能跨越多个动作块；智能体会在启动下一个子任务前验证其物理结果。确认恢复成功后，WAM 会在更新后的任务上下文中，从修复后的状态继续执行。

触觉生成与监督。 操作动作通过接触对物体产生影响，接触是否发生及其载荷决定了机器人运动如何传递至环境。触觉观测为这些交互提供直接反馈。因此，将触觉观测与动作一同预测，可使动作塔同时获得关于运动及其接触后果的监督。接触具有稀疏性，且可能突然出现或消失；因此，我们将每个区域代理 $\mathbf{p}$ （来自 第 2.6 节 ）与一个存在状态 z配对，同时对接触发生情况和活动接触的属性进行建模。对于每个未来时刻的配对，两部分都遵循动作流路径，$x_\sigma = (1-\sigma)x + \sigma\epsilon_x$，并加入独立高斯噪声 ϵ_x。独立的输入投影将$\mathbf{p}_\sigma$ 和 z_σ 组合为每个区域和时间戳对应的一个词元，共享动作塔会将其与动作令牌共同处理。两个输出投影分别预测其流速度。这些速度更新配对时，随之变化的存在状态会在下一个去噪步骤再次编码，使代理预测能够利用模型对接触发生情况的当前估计。为在同一连续生成过程中训练该配对，我们对每个分支应用速度均方误差损失。存在状态还会在突变处确定离散接触决策，因此我们在其重建的干净估计上加入一个较小的二元交叉熵项。预测目标为

$$\begin{aligned}\mathcal{L}_{\text{tactile}}^{\text{pred}} = &\frac{\text{MSE}_{M_p}(\hat{v}_p, v_p^\star) + \lambda_z \,\text{MSE}_{M_z}(\hat{v}_z, v_z^\star)}{1+\lambda_z} \\ &+ \lambda_{\text{BCE}}\,\text{BCE}_{M_z}(\hat{q}_\sigma, y),\end{aligned} \tag{7}$$

其中，$v_x^\star = \epsilon_x - x$, y 是二元接触标签，而$\hat{q}_\sigma$ = Sigmoid函数(2[$z_\sigma - \sigma\hat{v}_z$])则在采样噪声水平下重建干净接触概率。每一项均对其有效目标条目取平均。由于没有接触时接触位置和形状未定义，M_p 会在接触目标处选择有效代理特征；M_z还会纳入有效的无接触标签，在排除缺失测量值的同时保留对无接触状态的监督。在推理时，对预测速度进行积分，从高斯噪声中得到未来的配对结果 ($\hat{\mathbf{p}}$, $\hat{z}$)。对于触觉读出，我们计算 $\hat{q}_{\text{输出}}$ = Sigmoid函数(2$\hat{z}$)，并返回 $\tilde{\mathbf{p}} = \mathbb{1}[\hat{q}_{\text{输出}} \geq 0.5]\hat{\mathbf{p}}$，仅在模型预测存在接触的位置保留接触属性。

3.1.4. 多模态条件化与对齐

我们以相机观测和结构化任务指令作为模型的条件输入，并在可用时加入近期状态和触觉历史记录。视频、动作和触觉标记共用一条物理时间轴，而空间坐标则用于区分相机视角和触觉区域。

单视图与多视图观测。 该策略接受单相机或同步多相机观测。对于多相机输入，我们在统一的时间网格上解码帧，并在 VAE 编码前将其调整大小，合并到一个复合画布中。两个视图垂直堆叠。对于三个视图，主视图占据上方区域，其余视图分别占据左下方和右下方区域。细黑色分隔线区分相邻的相机区域。在每个采样裁剪中，相机分配始终保持一致。在同一时间戳捕获的视图共享同一时间位置，并占据视频标记网格中不同的空间区域。

结构化任务条件化。 语言条件化提供了共享任务接口，协同设计协议通过该接口将智能体决策连接到模型执行。表示任务进度的字幕变体将完整指令与当前子任务及已完成子任务的历史记录配对。完整指令保留总体目标，当前子任务指定即时执行目标，历史记录则记录已完成的阶段。历史记录则记录已完成的阶段。训练标注填充这些字段；在执行过程中，智能体维护相同的任务表示，并根据经核实的物理结果更新子任务进度。

该接口使智能体能够根据模型的执行能力，将长时程任务组织为相应的目标，同时相应的训练监督教会 WAM 在上下文中围绕这些目标采取动作。经验证的经验还可以进一步优化后续执行中通过同一接口提供的任务命令，而无需立即更改模型参数。

结构化字幕还会保留任务标题、具身身份、片段时长、视频帧/秒、动作频率、画布分辨率和宽高比。对于多视图观测，它们会指定每个相机的语义角色、画布位置和区域尺寸。这些字段使采样率和相机布局在条件文本中明确呈现；下方的位置编码则以数值方式表示它们在时间和空间上的组织结构。

帧率对齐旋转位置。 我们使用具有时间、高度和宽度坐标的三轴多模态 RoPE。文本令牌沿三个轴同时推进，而视频令牌遵循空间块网格，并采用由物理时间决定的时间坐标。给定参考帧率 $f_{\text{ref}} = 24$ 以及参考 VAE 时间压缩率 $c_{\text{ref}} = 4$，时间为 τ 秒的令牌所获得的时间坐标为

$$p_t(\tau) = b + \frac{f_{\text{ref}}}{c_{\text{ref}}}\tau = b + 6\tau, \tag{8}$$

其中， b 是关联的视频、动作和触觉流共享的时间原点。在规则的视频潜在网格上，索引 k 对应于 $\tau = kc_v/f_v$，其中视频帧率为 f_v ，变分自编码器的时间压缩率为c_v。动作令牌在时间上未压缩：其坐标遵循动作采样率和序列对齐偏移，或在提供显式时间戳时遵循这些时间戳。128 维状态/动作通路使用显式时间戳，因此，当当前状态和第一个动作对应同一时间时，两者可以共享同一时间坐标。浮点位置 ID 保留小数偏移量。因此，对于采样率不同的流，相同的经过时间会对应相同的时间位移。这些物理时间坐标与流匹配所用的噪声水平相互独立。

拼接视图的空间位置。 空间坐标在每个视频片段中重置。对于多视图画布，我们以相机上、下区域之间的接缝为基准表示垂直位置：

$$p_h = s - h, \tag{9}$$

其中，h 是补丁行坐标，s 是以补丁为单位的接缝位置，根据实际的相机布局几何结构计算得出。接缝位置可以是小数。宽度坐标保留其在画布上的位置，从而区分三相机布局中的两个下方视图。单视角观测保留常规空间网格。此转换不改变时间坐标，并保留各模态之间的物理时间对齐。

触觉对齐。 无接触读数既可能出现在首次触碰之前，也可能出现在释放之后，因此仅凭单次观测无法判断交互阶段。近期的触觉与状态历史有助于消除这种歧义，因为它们记录了先前的接触以及导致当前状态的运动。我们保留这些历史作为干净的条件输入，用于联合生成未来的触觉、动作和视频令牌，且损失仅应用于未来目标。为保留接触变化相对于运动发生的时刻，触觉 mRoPE 在 (8) 的共享时间轴上使用实际时间戳。触觉样本在时间维度上保持未压缩，同一时刻的所有区域共享同一时间坐标，从而在不同采样率下保留它们相对于动作和视频的计时。两个空间轴分别编码规范功能区域和手侧，因此每个区域在不同时间和具身形态下都保留相同的解剖坐标；区域内的接触位置仍属于预测代理的一部分。由于这些坐标标识固定的解剖槽位，观测到的状态嵌入还会提供相应的姿态上下文：历史词元使用对应的观测状态，而未来查询使用当前观测状态。

3.2. 基础设施

我们扩展了英伟达的 `cosmos-framework` 和 Cosmos 3 执行栈 [40, 41]，以训练 第 3.1 节所述的世界模型。变长视频、动作和触觉序列会限制设备内存、计算图复用和输入供给。我们的训练框架通过混合分片、固定容量执行和共享数据准备来应对这些限制，随后导出权重和预处理状态以供推理。以下测量采用不含触觉的历史工作负载（$T = 0$）；附录 E.1 介绍了其配置和范围。

图 5 展示了执行契约：理解词元 (U) 使用世界理解塔，视频令牌 (V) 使用世界预测塔，动作令牌 (A) 后接触觉令牌 (T)，二者使用共享的动作与触觉生成塔。塔专用投影和前馈神经网络位于联合注意力模块周围；掩码将打包样本彼此隔离，并在监督中排除填充内容。干净的条件输入和预测目标使用独立的有效性掩码和损失掩码。

3.2.1. 训练框架

对可变长度的多模态样例进行训练时，需要将模型状态和词元批次装入 GPU 内存，同时确保设备持续获得数据。我们通过分布式分片、固定容量的词元缓冲区，以及将批次传送与计算重叠的数据加载器来满足这些需求。

分布式执行。 全分片数据并行（FSDP2）对参数、梯度和优化器状态进行分片；混合分片（HSDP）在副本之间同步相应的分片 [42, 43]。在 128 个 GPU 上，分片8 × 副本16 将参数集合通信限制在八 GPU 节点内。在我们的目标生产配置中，分片8 无法容纳 96K 的视频词元预算。分片16 × 副本8 将每个设备上的持久状态从$M/8$ 降至 $M/16$，对应总计未分片状态 M，代价是跨节点参数通信。采用两个梯度累积步数（GA2）、BF16、完全重计算和静态编译后，所得的文本转视频（T2V）配置支持每个微批次 U3,072/V98,304 个词元。历史上的 64K 到 96K 离线转换报告显示，样本吞吐量提高了 **8.9%**；这是配置级比较。 附录 E.1将该配置与后续的分片8/96K 内核基准测试区分开来。

计算优化。 固定的 U、V 和共享的 A+T 词元缓冲区将物理容量与逻辑样本长度分离。位置和掩码保留变长语义；动态准备仍在

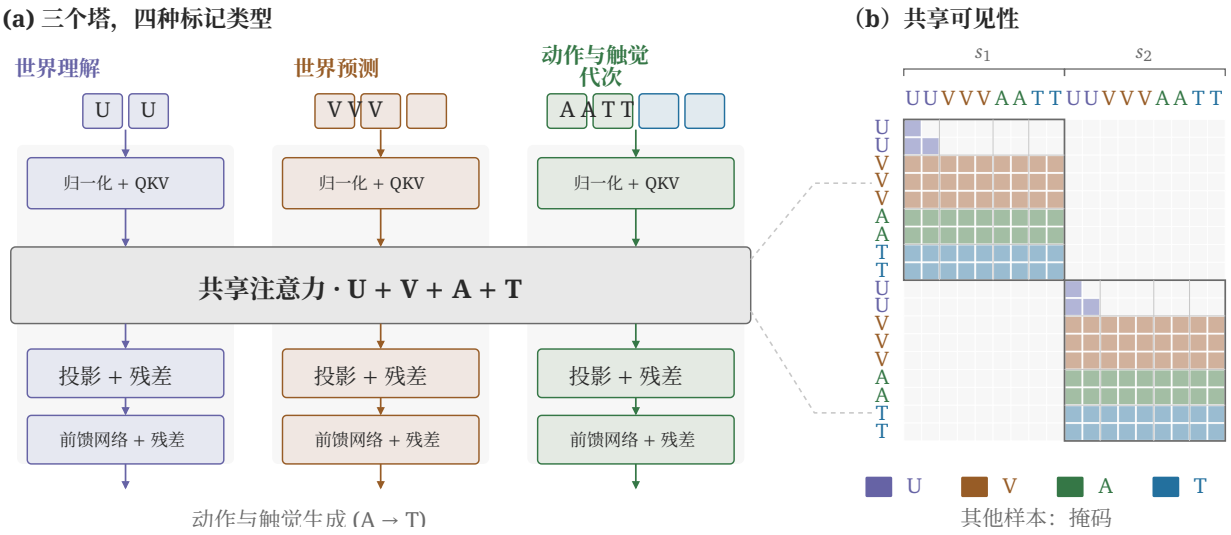


图 5：四种令牌类型，三个塔。 紫色 U、赭色 V、绿色 A 和蓝色 T 分别表示理解、视频、动作和触觉令牌。世界预测塔处理视频令牌。每个样本示意两个 U、三个 V、两个 A 和两个 T 令牌；在共享的动作与触觉生成塔中，A 位于 T 之前。U 的注意力是因果的；V/A/T 查询会关注同一样本中所有有效令牌。打包样本彼此隔离；令牌数量仅为示意。

编译后的解码器块，从而支持计算图复用 [44]。在 图 6中的 DROID Action 动作消融实验中，打包将吞吐量从 9.64 提升至 22.39 个样本 / 秒。固定形状和编译使吞吐量达到 36.76 个样本 / 秒，**比仅打包高 64.2%**；增加工作进程后，吞吐量达到 39.09 个样本/秒，即 **4.06× 初始短窗口基线**。全局批次大小固定为 32。启动和编译时间不计入其中。

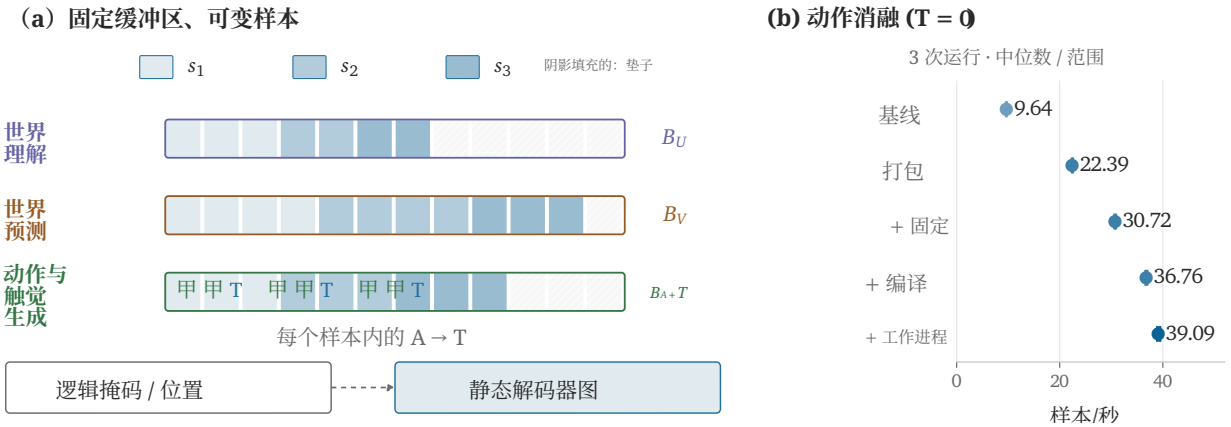


图 6：静态令牌缓冲区与训练吞吐量。 (a) U 和 V 填充世界理解缓冲区和世界预测缓冲区。A 和 T 在动作与触觉生成塔中共享一个容量为 B_{A+T} 的缓冲区。填充部分不参与注意力计算和监督。(b) 不含触觉的历史 DROID Action 消融实验（$T = 0$）：`width_only_1b`，不使用状态条件化，冻结世界理解塔，使用 8 个 GPU、BF16、在线 VAE、完全重计算，全局批次大小为 32。每种配置运行三次，更新步数为 11–40；点表示运行均值时间中位数的倒数，范围表示观测极值倒数。 表 E.2 列出了各项配置。

对于文本转视频，离线潜变量省去了重复解码、预处理和冻结 VAE 编码，同时保留帧选择和时间上下文。噪声和目标仍在线构建。四 GPU 性能分析记录到，每次捕获的 rank-0 更新步包含 1.069 秒的在线 VAE 工作，而离线跟踪中没有 VAE 工作；由于使用了不同的目录，无法估算匹配条件下的端到端加速比。预处理和潜变量存储是单独的成本。

我们扩展 FlashAttention-3 [45]，使其能够按分段选择注意力掩码。单次调度分发即可在每个样本内，对理解查询应用因果注意力，对视频、动作和触觉查询应用全注意力

。在 T2Vtrace 中，分组使主内核启动次数减半，而更新时间几乎没有变化。算子检查通过；全更新数值验收仍未解决。 附录 E.2 介绍了内核设计、通信与调度实验，以及计算与内存权衡。

数据加载。我们将来源身份、采样顺序和可恢复游标与物理工作进程分离（图 7）。共享池在全秩信用额度预算下准备批次 K：交付会返还信用额度，而消费会提交可恢复进度。 K 限制的是已提交但未交付的任务，而非字节数；活动批次和媒体缓存各有独立预算，且有序交付仍可能停滞。

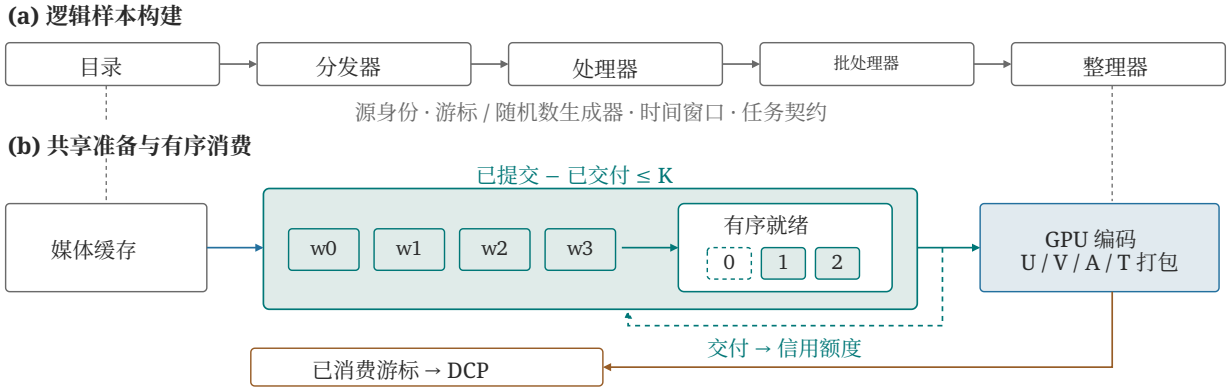


图 7：逻辑构造与物理准备。打包会保持 U/V/A/T 的顺序，其中 A/T 会路由至动作与触觉生成塔。每秩预算 K 限制跨工作进程、传输、可选固定和有序就绪的已提交但未交付任务。交付会返还信用额度；消费会提交可恢复进度。活动批次和媒体缓存各有独立预算。

逐源加载器基线使用每个 rank 23 个工作进程，预取因子为 2。共享池使用每个 rank 4 个工作进程，全秩信用额度为 $K = 16$。在编译后比较中，节点主机内存从 **695.13 降至 334.57 GiB**（−51.9%），而测得吞吐量相近（图 8(a)）。独立重复实验节省约 52% 的内存，吞吐量变化为 +2.73% 和 −0.21%。这些比较涵盖完整的加载器/资源配置；简短的编译后重放覆盖 23 个源中的 9 个。

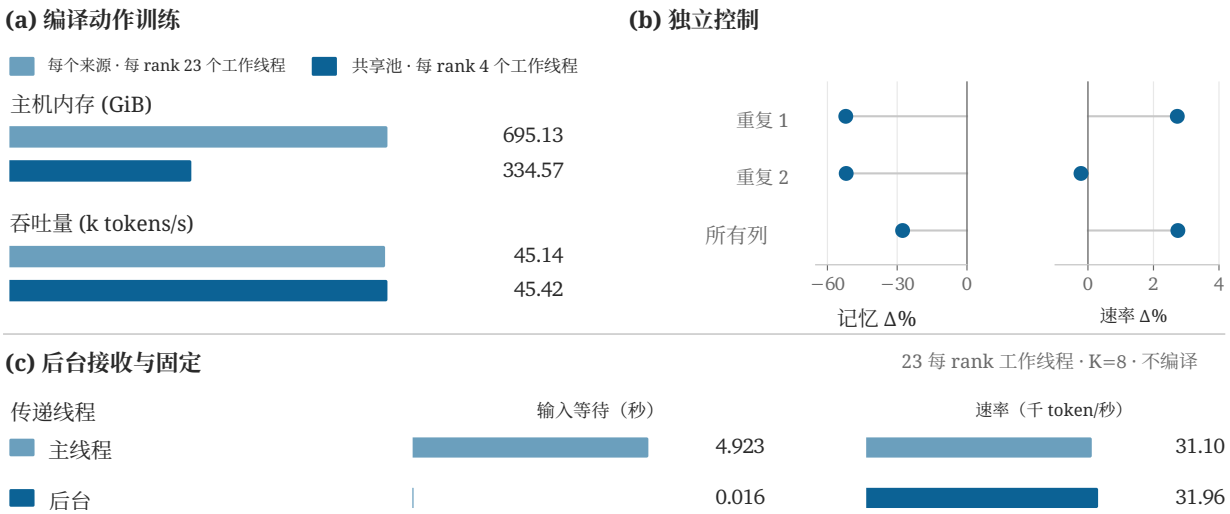


图 8：8 个 GPU，23 个源。(a) 编译后：4 次预热/8 次计时更新，每个机械臂运行 1 次。(b) 两次独立重复实验：每次各进行 4/12 次更新；全列对照：2/6 次，读取完整的回合表。每种设置均使用自身的逐源加载器基线。(c) 每个 rank 固定 23 个工作进程和 $K = 8$，进行 2/6 次更新，每个机械臂运行 1 次。输入等待时间为 rank 0 的 6 次更新总和。在 (b、c) 中，解码器和提前 VAE 编译均已禁用。内存是整个运行期间节点匿名内存+共享内存的最大值，每两秒采样一次，不包括文件缓存。

背景　　d IPC 接收、反序列化和可选固定使交付与计算重叠。在　　23

每个 rank 配置工作线程、 $K = 8$，并在两侧均固定记忆后，rank-0 的输入等待在六次更新期间从**4.923 降至 0.016 秒**。吞吐量从每秒 31.10 k 个 token 提升至 31.96 k 个 token，主机内存从 244.92 GiB 增至 265.91 GiB。

指纹索引元数据还将 58 个来源的数据初始化时间从 643.39 秒缩短至 14.59 秒，并实现了缓存命中。该比较保留操作系统文件缓存，并排除模型加载和编译。附录 E.3 报告了控制实验和更长时间的输入等待测量。

其他优化。 选择性重计算以激活内存换取更少的重复计算 [46]；精度和变分自编码器编译研究见 表 E.5。按需通信 [47] 消除了逐层会合点，而成本感知重新分配降低了预测的注意力不平衡 [48–50]，但在所测试的工作负载中，两者都未能证明端到端收益（附录 E.2）。

为支持恢复，分布式检查点 [51]会保存模型/优化器状态，以及逻辑数据分片标识、已消费游标和随机数生成器状态。离线所有权重新映射支持 160 → 144**图形处理器续训**：全部 11,136 个逻辑状态均予以保留，且每个秩都通过了前十项微批次序列检查（表 E.10）。全局批次组成可能发生变化；已保存状态恢复并不能证明数值延续等价性（附录 E.5）。

训练效率。 表 2 列出了 80 GB 图形处理器上的历史配置。模型浮点运算利用率（模型浮点运算利用率）**不包括激活重计算和填充** [46, 52]。128 个图形处理器的在线文本转视频审计得到 **31.04%**，其中包括通过冻结的变分自编码器所需的前向传播；或者，对于 **25.28%**，仅针对 Transformer。这些是基于 677 次完整更新的秩 0 分析估算，而非全秩平均值。完整的更新间隔时间保留了输入和同步停滞；附录 E.4 提供了浮点运算次数分子和密集 BF16 峰值约定。

表 2：历史诊断配置。 这些配置不包括触觉令牌（$T = 0$），并且不同于第 3.3 节中的课程安排。令牌容量按每个图形处理器微批次计算。只有第一行重建了有效工作模型浮点运算利用率。

工作负载	并行度	物理令牌；有效工作模型浮点运算利用率
文本转视频，在线编码（已审计）	128 个 GPU，梯度累积 4 分片8 × 副本 16	U 5,958；V 59,578；**31.04%**，包括变分自编码器前向传播
三塔动作，在线编码	128 个 GPU，梯度累积 2 分片16 × 副本 8	U 11,776；V 62,592；A 2,944；模型浮点运算利用率未经审计
文本转视频，离线编码	128 个 GPU，GA2 分片16 × 副本8	U 3,072；V 98,304；模型浮点运算利用率未经审计

3.2.2. 推理

导出权重、预处理资源和归一化状态用于初始化常驻推理。策略和服务结果来自对历史三塔模型的单独测试，这些模型不包含状态或触觉输入（附录 E.1）。

- **视频生成内核。** 在 50 步文本转视频推理中，前向注意力占秩 0 上累计计算内核时间的 34.0%。为解决这一瓶颈，我们使用 FlagGems 注意力算子优化了注意力计算， [53] 该算子来自 FlagOS，41× 在从实际推理跟踪中提取的 BF16 张量形状上实现了 1.39–1.表 E.7中的内核级加速比。

- **条件复用。** 为每个请求和无分类器引导（CFG）分支缓存世界理解塔中的时间步不变状态和键/值张量，可在单个 GPU 上节省 72–73 毫秒，在双 GPU CFG 并行时节省 45–47 毫秒。在已测试的案例中，输出保持一致。CFG 并行基线已启用此缓存。

- **双 GPU CFG 并行**。 在两块 GPU 上并发运行条件/无条件分支 [54] 可将仅动作延迟降低 2.004 → 1.054 **秒**（1.90×），测试条件为 960×640、41 帧/40 个动作以及五个采样步。已分配的 GPU·秒/请求从 2.004 增至 2.108。
- **世界预测塔 FFN 复用**。 相邻步骤的输出相似，因此可以缓存选定的 FFN 输出 [55, 56]。跳过 20% 的调用可使准备执行延迟降低约**10%**，同时仅需 2.42 GiB 的缓存。旋转 6D 均方根误差平均上升 3.42%；单个夹爪开合平均绝对误差最差上升 21.03%。默认关闭此近似方法。

- **有界队列**。 每个副本服务一个请求，按先进先出顺序将两个请求加入队列，拒绝溢出请求，并跳过已过期请求。在八秒截止期限下，当速率为每图形处理器每秒 0.5 个请求时，及时有效吞吐量提高**22.20%**；在 0.75 时下降**57.13%**。

离线手部重建结合了混合精度、按输入形状编译和重叠窗口处理（附录 C.1.4）。

3.2.3. 评估基础设施

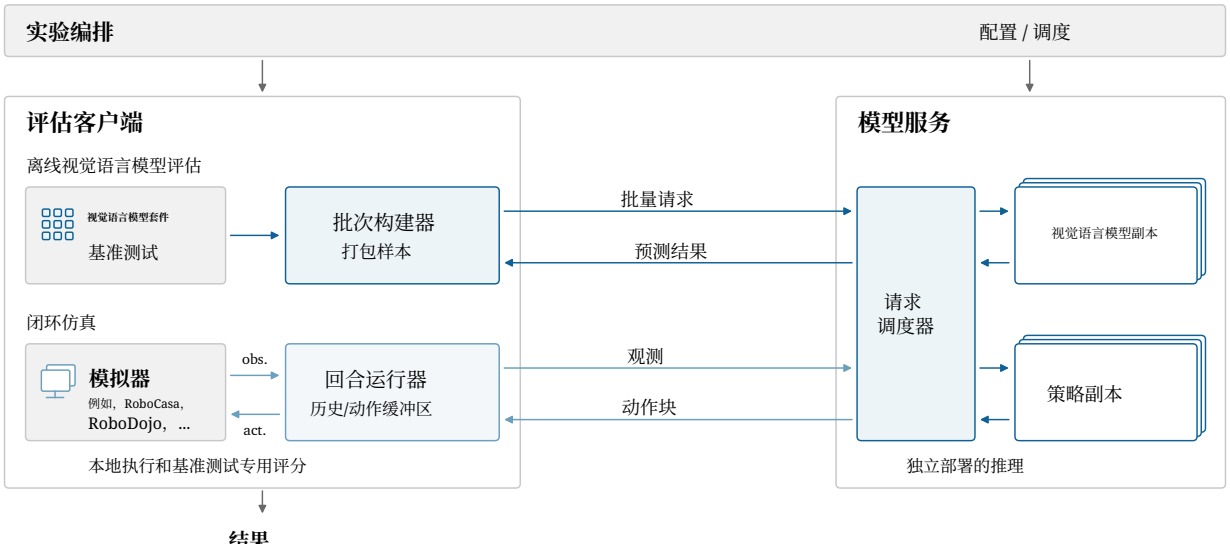


图 9：基于服务的评估基础设施。视觉语言基准测试提交批次请求，而仿真客户端则在每个回合推进时查询策略服务。每个基准测试均保留其自身的评分流程。

我们的评估系统支持跨多个环境开展视觉语言基准测试和闭环仿真，其中包括若干仿真基准（图 9）。评估客户端执行基准测试，并通过请求-响应接口与独立部署的模型服务通信，使数据准备、仿真和推理能够在不同资源上运行。

对于 VLM 评估，基准适配器会准备任务专用提示、媒体输入和生成设置。批次执行流水线将请求分组并分发至模型服务，随后将返回的预测结果与相应示例关联起来。该框架支持原生基准测试，并可与 VLMEvalKit 和 lmms-eval 集成，同时保留任务专用的提示构建和评分流程。

仿真客户端维护环境状态、观测历史记录和动作缓冲区，并在需要新的动作块时请求策略推理。返回的动作在本地执行，后续请求则会纳入更新后的观测。请求调度器将并发环境发出的推理请求调度分发至可用的模型副本。每个环境分别维护策略状态，并且

响应会被路由至对应的回合。这种分离方式支持并行仿真，无需在每个环境进程中嵌入模型推理。

评分遵循各基准测试的协议，可在执行期间进行，也可在单独的后处理阶段进行。对于需要进行数据集级评估的视觉语言模型基准测试，会将收集到的预测结果用于评分。仿真评估会记录任务完成情况，并在有定义时记录部分任务进度。报告会保留各基准测试的汇总指标和单个评估案例。

视觉语言模型后端会保存问题、预测结果、得分和运行清单，以及模型和基准测试配置、执行设置和源代码版本信息。这些记录可用于检查、恢复中断的运行以及延后评分。仿真框架会保留推理请求与响应、回合轨迹和轨迹展开视频，以便结合相应的观测和动作分析策略行为。对于系统轨迹展开数据收集，同一套执行接口会保留准备策略训练轨迹所需的观测、指令、动作和结果。

3.3. 训练

训练分为两个主要阶段：视频生成预训练，通过视频生成和动作条件世界建模（ACWM）发展视觉预测能力；随后进行策略模型预训练，引入动作生成，继而实现联合动作与触觉生成。在此之前，会调整理解塔以适应具身理解。

原始数据池在筛选前包含约 100,000 小时的数据；各阶段的训练混合数据均从该数据池中构建。当末端执行器（EEF）和关节空间表示均可用时，我们会同时使用同一经验的两种表示：末端执行器目标描述末端执行器运动，而关节空间目标则保留具身特定关节运动。这样无需增加底层记录时长，就能扩充可用的监督数据。

3.3.1. 世界理解训练

在 MoT 架构中，理解塔旨在发展与机器人操作相关的空间表征，重点关注三维空间定位、物体定位和空间关系。我们并非只追求通用视觉理解，而是优先发展可能有助于动作生成的能力，为下游机器人决策提供感知和空间基础 [57, 58]。我们对理解塔进行微调，以增强这些能力；训练数据和微调流程的详情见附录B。

3.3.2. 视频生成预训练

视频预训练教会世界预测塔根据观测到的状态预测操作场景如何演进。我们将通用视频生成器调整为适用于具身交互的生成器，然后加入将给定动作轨迹与生成视频相联系的示例。

从通用视频先验到具身动力学。要将预训练视频生成器调整为适用于机器人操作场景的生成器，需要在领域和条件输入两方面实现对齐：从 Cosmos 3 Nano 继承的以文本转视频为中心的先验涵盖广泛的视觉内容，而下游操作任务则需要根据特定的观测状态对具身交互进行建模。

因此，我们继续在由通用视频、真实机器人演示以及第一人称人类交互组成的千万片段混合数据上预训练模型，将重点转向具身动力学和文本与图像转视频（TI2V）预测。每个裁剪都配有动作描述或指令，训练以单视角为主，建模场景如何从一个视角演变。我们保留原有的纯文本转视频（T2V）监督，同时加入文本与图像转视频任务，从而保留文本驱动的通用生成能力

表 3：视频生成预训练设置。适配后的理解塔被冻结，而代次参数及其输入/输出投影则会更新。令牌容量是可变长度样例的打包上限。

设置	视频生成预训练
优化器	AdamW
视频塔学习率	10^{-4}
学习率调度	预热 50 步，然后保持恒定
AdamW 系数	$(\beta_1,\ \beta_2) = (0.9,\ 0.95)$
AdamW ϵ	10^{-6}
权重衰减	0
梯度累积	4
梯度范数裁剪	0.1
每设备理解 token 容量	2，176
每设备视频令牌容量	65，536
视频噪声级别采样器	Waver

同时开发以观测到的场景为基础的预测能力。所有设置均使用条件流匹配目标函数。在这一阶段，理解塔始终保持冻结，作为条件骨干网络。

在本阶段后期，我们根据第 2.4 节中的质量评分，增加对视觉可靠样本的采样权重，同时保持各来源的覆盖率。

将视频生成与动作条件预测结合。 视频预训练的最后阶段将视频生成与 ACWM 结合。一个 ACWM 样本包含初始观测、语言指令以及时间对齐的动作轨迹；预测目标是由此产生的未来视频。提供的动作仍作为干净条件输入。没有未来动作标注的视频生成样本则继续提供视觉预测监督，将广泛的视觉覆盖与明确的动作—结果学习相结合。

训练配置。 我们使用 AdamW 更新视频生成参数及其输入/输出投影 [59] ，学习率为 10^{-4}。经过 50 步预热后，学习率保持恒定。视频噪声级别使用 Waver 进行采样。训练采用 BF16 精度、分片数据并行、完整激活检查点，并为理解和视频令牌设置单独的打包预算。表 3 列出了优化器和打包设置。

3.3.3. 策略模型预训练

在视频预训练的基础上，我们训练策略模型，使其根据观测和任务指令生成动作。该课程首先利用广泛的交互数据学习日常操作行为，然后重点学习机器人示范中可获得的状态和接触信息。

三阶段课程。 从预训练世界模型开始，策略学习依次经历广泛动作学习、状态条件动作学习以及动作–触觉联合学习。动作监督在整个过程中持续有效，触觉监督则在最后一个阶段加入。图 10展示了课程如何从自我中心视角体验逐步转向机器人示范，其中抓取与放置及家务任务占最终混合数据的大部分。

第一阶段：广泛动作学习。我们最初训练动作生成，不使用明确的机器人状态条件或触觉预测。自我中心示范广泛覆盖日常交互，而机器人任务则遵循源分布。此阶段着重于语言条件运动，以及在多样化场景中与物体交互和任务进度相关的视觉线索。

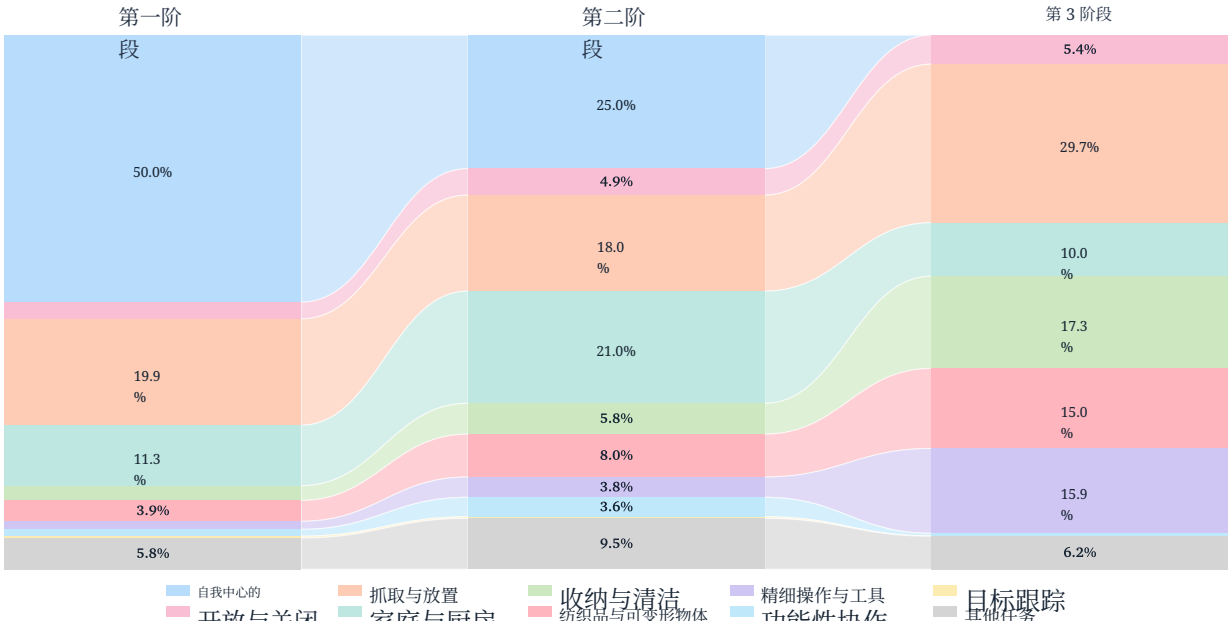


图 10：三阶段策略训练混合数据示意图。 色带高度表示根据来源权重和任务标注估算的占比。自我中心体验构成一个组。机器人示范（包括使用灵巧手完成的示范）按任务分组。

第二阶段：状态条件动作学习。我们引入观测到的机器人状态作为干净的条件输入，并减少自我中心示范的占比。状态条件化提供明确的机器人状态信息，这些信息仅凭图像可能难以判断。我们还重新平衡剩余来源，以提高任务类别的覆盖率，例如精细操作与工具使用、存储与清洁、纺织品与可变形物体，以及功能性协作。凡是可用，末端执行器和关节空间目标都会保留在混合数据中，并采用相应的领域感知投影。

第 3 阶段：联合动作与触觉学习。 最后阶段保留动作监督，并为具有对齐触觉观测的示范加入触觉预测。动作流与触觉流共享动作生成塔，并采用架构部分所述的联合监督。没有触觉测量的示例仍会用于动作监督；触觉损失仅适用于有效目标。本阶段移除自我中心数据，训练混合数据侧重于机器人示范，尤其是抓取与放置任务以及家庭和厨房任务。触觉标签可用性与任务类别分别记录，因此触觉监督可应用于不同类别。

恢复监督与轨迹复用。 经过筛选的系统执行轨迹用于策略预训练和选定的监督微调设置。这些轨迹通常包含恢复过程，以系统执行期间到达的状态所产生的纠正行为补充示范。第 2.2 节中的共享标注将每个恢复片段关联到之前的失败尝试，并在可用时关联到后续的任务继续执行。我们为包含恢复行为且经过验证的子任务轨迹分配更高的采样权重，以监督纠正动作以及后续执行；其动作目标与其他示范所用的相同。

重复使用前，轨迹展开轨迹会经过共享的标注与策略数据准备流水线。随后，成功的任务分段和恢复片段会被纳入训练。这种重复使用通过将经验证的交互经验纳入模型训练，补充了智能体对任务上下文的完善。

从交互覆盖率到具身对齐。这一课程逐步将重点从丰富多样的交互经验转向与机器人控制相对齐的监督。自我中心视角数据源广泛覆盖物体、场景和运动，而机器人示范则在机器人专属控制空间中提供动作和状态监督

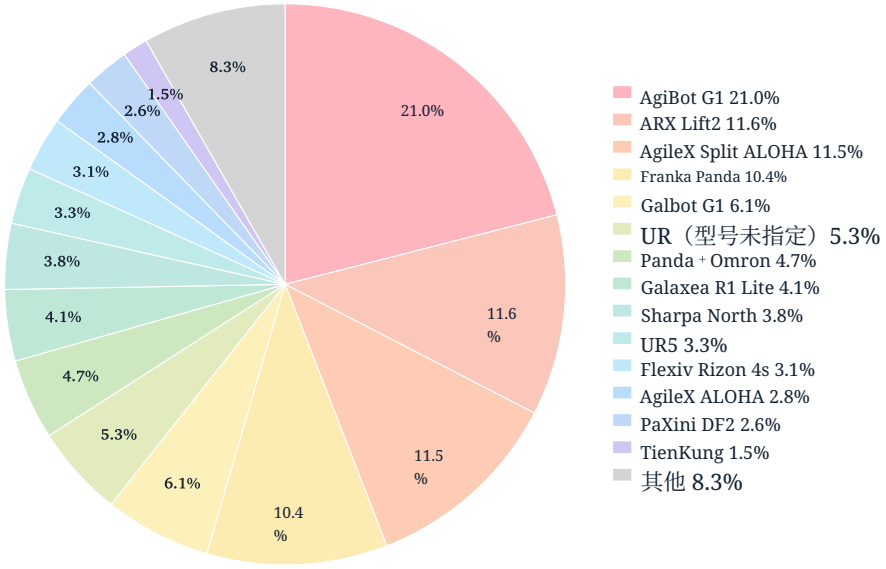


图 11：关节空间来源池的具身组成。各占比按筛选前的源时长计算。单独展示占比最大的 14 个组，其余部分合并为“其他”。对应的真实变体和仿真变体，以及末端执行器变体，均根据源图例进行分组。末端执行器和人类第一视角表示不计入其中。这些占比描述的是源池覆盖率，而不是图10中的训练混合占比。

表 4：状态条件策略配置的设置。视频和动作学习率分别设定。令牌容量描述序列打包限制，因此每个批次的示例数量取决于序列长度。

设置	状态条件策略训练
视频塔学习率	10^{-5}
动作塔学习率	10^{-4}
学习率预热	100 个优化器步数
梯度累积	3
每设备理解 token 容量	11，776
每设备生成 token 容量	62,592
预留动作 token 容量	2,944，在生成容量范围内
视频／动作损失权重	1.0 / 1.0
视频／动作噪声级别采样器	偏移均匀分布，偏移量为 3；独立采样

，并在条件允许时提供接触观测。这一渐进过程受到相关方法的启发：这些方法将广泛的操作经验与后续控制学习相结合：VPP2 先进行操作视频预训练，再进行动作学习 [60]，；GR00T 2 从人类与手持设备演示迁移到机器人具身形态 [3]，；Xiaomi-Robotics-1 则在对手持操作轨迹进行广泛预训练后，进行本体形态与指令对齐 [61]。随着引入机器人专属状态和接触监督，我们的课程逐步降低自我中心视角贡献。

具身形态覆盖率。 图 11 展示了关节空间来源池，与训练混合视图相辅相成。AgiBot G1（21.0%）、ARX Lift2（11.6%）、AgileX Split ALOHA（11.5%）和 Franka Panda（10.4%）是最大的本体形态组，此外还有移动操作机器人和灵巧操作平台。域感知投影可在共享的动作 Transformer 中适配它们各自不同的控制表征。

噪声级别采样。 噪声实现和噪声级别在视频之间独立采样，并在存在动作目标时，也在动作目标之间独立采样。对于策略训练中的视频和动作，我们分别从 $u_m \sim \mathrm{U}(0,\ 1)$ 中抽取

并应用 3 的偏移:

$$\sigma_m = \frac{3u_m}{1 + 2u_m}, \qquad m \in \{v, a\}. \tag{10}$$

这种平移均匀分布会更侧重较高噪声级别，同时保留对低噪声的覆盖率。对高噪声动作输入的侧重，源于 π_0 [62]中的时间步采样分析，以及 [63]在 FLUX 3 Action 中研究的高噪声分布。在第 3 阶段，触觉目标与动作流共享 RF 时间步，并为每种模态独立采样高斯噪声。

超参数。在策略预训练期间，视频塔和动作塔分别采用固定学习率 10^{-5}和 10^{-4}。训练采用 BF16 精度、完整激活检查点和打包的多模态序列。状态条件配置采用 100 步预热。其他设置见表 4。令牌预算会限制打包序列的长度，因此每个批次中的样本数会随序列长度而变化。

4. 具身智能体系统

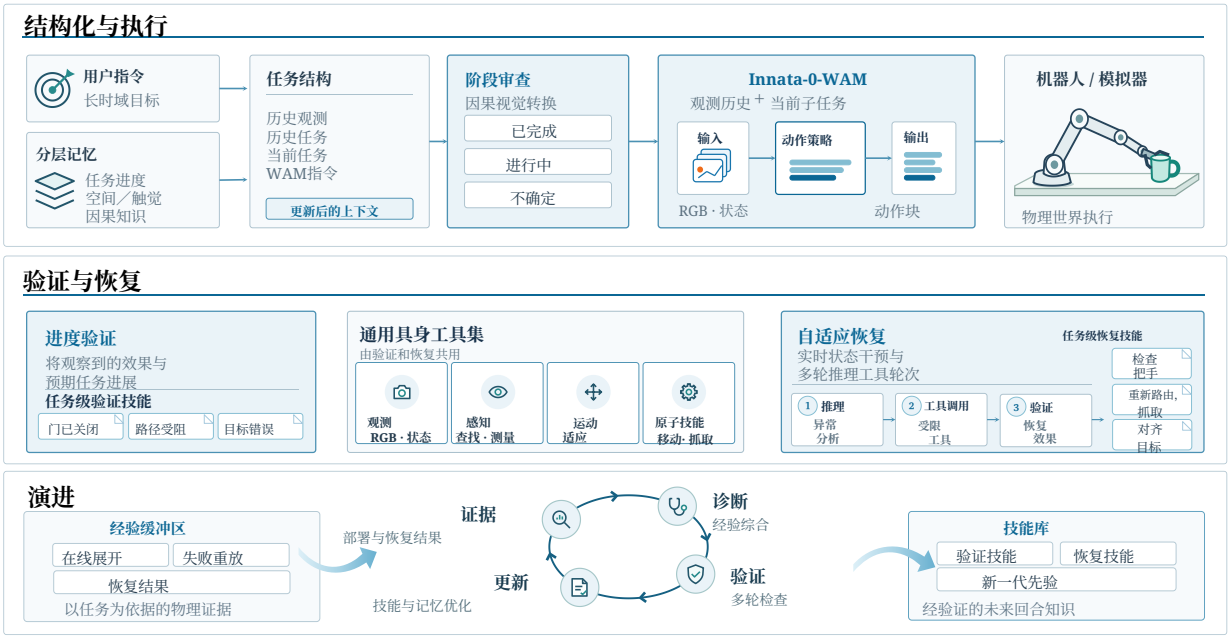


图12：Innata-0 智能体运行时。智能体将长时域目标落实到分层记忆中，并维护持续更新的任务结构。定期进行的阶段审查会评估因果视觉转变并更新任务进度，而 WAM 则预测动作块，以便在机器人或模拟器中执行。借助任务级验证技能，验证器使用与恢复共用的通用具身工具集，将观测到的物理效应与预期任务进展进行比较。确认发生偏差后，自适应恢复会进行迭代推理、受限工具调用和结果验证。随后，轨迹展开和恢复证据会经过综合与验证，以改进任务级技能，并为未来回合提供可复用先验。

Innata-0-WAM 提供理解、预测和动作生成所需的物理能力。要将这些能力扩展到长时程任务，需要超越单个模型执行的协调机制。因此，我们遵循协同设计协议，开发了 **Innata-0-Agent** ，以支持规划、恢复和演进。

对于规划，智能体根据模型的能力将复杂指令分解为任务结构，并协调其执行。它会维护完整指令、当前子任务和经验证的任务历史记录，提供与模型训练相一致的结构化任务上下文。对于恢复，智能体会验证执行结果，并识别与预期任务进度之间的偏差。当发生失败时，它会调用恢复工具，以恢复有效的执行条件，并协调后续执行以完成

剩余任务。对于演进，智能体会积累经验证的交互经验，并将其整合为可复用技能，以改进未来的决策和恢复策略。成功的 rollout 片段，包括纠正动作和恢复后续执行，会通过共享数据流水线进一步收集，用于后续的策略训练。Innata-0-Agent 运行时的概览如图12所示。

为在统一的执行框架内支持这些能力，我们开发了智能体基础设施，用于连接任务结构、模型执行与工具使用。该基础设施通过可组合的任务图组织执行，提供对模型和工具的统一访问，并集成物理反馈与安全机制。任务隔离和资源复用进一步保障了系统运行的可靠性与效率。

4.1. 任务结构与执行

为支持规划，Innata-0-Agent 会协调模型在相互依赖的子任务间执行，以完成长时程任务。给定用户指令后，智能体会根据模型的执行能力组织任务，并检索相关的分层记忆，以构建当前任务上下文和 WAM 指令。随后，WAM 通过动作块在机器人或模拟器中执行指定子任务。在执行过程中，智能体会定期检查任务进度，并根据观测到的结果更新当前子任务和 WAM 指令。由此形成任务规划、物理执行与进度驱动的适应闭环。

4.1.1. 层级记忆架构

长时域交互要求系统维护不断演变的任务状态，而不是反复依据原始指令 [64] 进行推理。因此，我们引入一种层级记忆架构，逐步将物理交互证据组织成与任务相关的状态和可复用知识。这些记忆共同构成共享状态基础，为任务分解、规划、验证、恢复和模型执行提供支持。

任务记忆。 执行过程中，智能体会持续将交互历史压缩为以任务为中心的状态，其中概述当前进度、已完成子任务、未解决的依赖关系、相关物体状态、检测到的异常以及恢复结果。这份不断演变的摘要为长时域推理提供主要上下文，避免每次都从完整执行历史中重建任务状态。智能体会按固定时间间隔，依据最新观测和执行证据审视当前任务阶段，并相应更新任务记忆。

物理状态记忆。 物理状态记忆通过互补的空间组件和触觉组件，持续记录不断演变的物理世界。空间组件会持续记录物体位置、场景结构和空间关系，即使物体超出当前视野也能保留相关信息。它为任务状态跟踪提供稳定的定位，使规划和恢复能够推理暂时被遮挡或不再直接观测到的物体或区域 [65, 66]。触觉组件则提供稀疏但富含交互信息的局部接触证据，尤其适用于交互状态在视觉上难以判定的情况 [67, 68]。智能体会结合视觉观测和本体感觉观测来解读这些反馈，并保留经验证的交互状态，例如物体是否仍被持握或已被释放。这两个组件共同依据场景的空间布局和正在发生的物理交互，为任务状态提供定位。

因果交互记忆。在最高抽象层级，经验证的交互经验被整合为关于动作如何改变物理状态的可复用知识。因果记忆表示以动作为中心、作用于结构化对象状态和关系的前置条件与效果 [69]:

$$\text{Precondition} \rightarrow \text{Action} \rightarrow \text{Effect}. \tag{11}$$

这种抽象使得从以往交互中获得的经验能够扎根于当前状态，并在无需立即更新模型的情况下，重新用于动作选择、前向预测、反向推理和恢复

因此，记忆层级支持双向信息流。在执行过程中，空间和触觉证据会更新由智能体概括的任务状态，而经过验证的状态转移则逐步整合为因果知识。在决策过程中，相关的因果经验会重新扎根于当前任务和物理状态，以指导规划、执行和恢复。

4.1.2. 扎根式感知与执行

在分层记忆的基础上，智能体将一个长时域目标分解为具有因果顺序的子任务，并将其调度分发给 WAM，以进行扎根式执行。我们进一步引入周期性的阶段审查机制，自动决定何时应维持或推进当前子任务。在每次审查时，智能体会采样较早的关键帧和最近的观测，以分析机器人或模拟器执行 WAM 生成的动作块后所呈现的状态转移。该时间证据会与回合内任务记忆、可用的空间记忆和触觉记忆，以及从长期分层记忆中检索到的子任务依赖关系和时序知识结合进行解读。

随后，审查器判断当前子任务预期产生的物理效应是否得到视觉支持，并将其状态分类为 已完成、进行中 或 不确定。只有在有充分证据支持完成时，才会接受子任务状态转移；随后将已完成阶段添加到任务历史中，并激活下一个子任务及相应的 WAM指令。否则，WAM 会继续执行当前指令，同时后续审查会积累更多证据。该过程将高层任务推进与扎根式执行结合起来，而无需 WAM 自身推断全局任务结构或决定何时切换子任务。

每次审查都会被记录，包括环境步数、审查前后的阶段索引、状态转移是否被接受，以及模型返回的完整结构化决策。这些记录使任务结构的演变可供审计，并为后续分析和系统改进提供可复用的执行证据。因此，整体执行过程会在 WAM 动作生成与保守且以证据为依据的阶段审查之间交替，使长时程任务得以推进，而无须依赖模拟器的特权状态，也不会让模型基于缺乏依据的假设推进规划。

4.2. 验证与自适应恢复

为支持 恢复，Innata-0-Agent 通过物理结果验证和自适应恢复，处理真实世界执行过程中的不确定性和意外故障。由于部分可观测性、接触不确定性或环境变化，物理动作可能产生非预期结果。为应对这些挑战，智能体会通过环境反馈持续监控执行进度，并验证是否已实现预期的物理结果 [70, 71]。发生偏差时，智能体会分析其原因并选择适当的恢复策略 [71, 72]。随后，它会调用互补的具身工具和技能，以恢复有效的执行条件。根据恢复结果，后续执行将通过 WAM 或由智能体控制的工具继续，以完成剩余任务。这一闭环流程将物理结果验证与自适应恢复相结合，使系统能够在不确定性下维持任务进度。

4.2.1. 执行监控与干预 on

在正常执行过程中，WAM 会反复预测并执行动作块，而智能体则在预定的验证检查点评估任务进度。每次验证都会根据采样得到的历史关键帧和近期观测构建一个紧凑的时间证据包，使智能体能够将观测到的状态转移与当前任务上下文下预期的物理效应进行比较。最终的决策会说明是否应触发异常、相应的故障类型、支持

证据，以及观测到的物理变化与预期变化之间的差异。单凭进度缓慢或视觉歧义不会被视为失败；只有在时间证据明确表明执行偏离预期任务推进时，才会启动干预。

除了定期验证外，系统还会在可能存在高风险的交互附近自适应地提高检查频率，例如抓取或物体搬运期间夹爪状态发生变化时。这种事件驱动的调度将推理集中在对接触敏感的阶段，而无需持续查询推理模型。时间冷却机制会抑制由同一持续视觉症状引发的重复触发，避免冗余干预。异常一经验证，便会暂停WAM执行，并将控制权移交给智能体。交接直接在当前实时环境中进行。智能体继承相同的场景实例、物体配置、机器人关节与夹爪状态、环境步数以及剩余执行预算。

4.2.2. 基于工具的验证与恢复

为支持在不同任务和异常模式下进行恢复，我们收集了一套通用且可扩展的工具。恢复代理不会直接生成底层控制代码，而是选择工具、设定其有界参数，并将其组合成简短的闭环干预序列。

工具集分为四个功能层。观测工具可提供任务信息、机器人状态、同步的视觉观测、相机标定以及此前发现的场景实体，且不会改变环境。感知工具会将这些观测转化为可用于行动的语义与几何表示，包括目标分割、三维测量，以及图像与世界坐标之间的转换。运动适应 工具提供参数化的几何调整，使智能体能够重新定位末端执行器、执行接触感知弧线运动，或根据当前场景重新定位移动底座。最后，原子技能封装了有界且可重复执行的物理操作，例如驱动夹爪、使机械臂返回安全构型、保持位姿，或执行基于几何条件的边缘抓取。

这些层支持闭环工作流的不同阶段。在验证过程中，智能体主要使用观测和感知工具检查不断变化的场景，对与任务相关的实体进行定位，并判断预期的物理效应是否已经发生。借助这些工具，智能体可以收集更多证据，降低语义或几何不确定性。其输出可补充时间证据，用于区分正常进展、模糊观测和已验证的异常。

异常得到验证后，同一观测与感知接口便可提供恢复所需的、有依据的诊断结果和几何目标。智能体可以随后引入运动适应工具和原子技能，以改变当前的物理状态。在通过原生机器人或模拟器接口执行有界物理基元之前，可以先检查候选运动是否可行。因此，验证与恢复共享一致的物体、几何结构和机器人状态表征，避免在实施干预时另行设置感知流程。

工具使用遵循迭代式的观察—推理—行动—验证流程。智能体获取证据，根据当前任务或已诊断的失败选择工具，执行该工具，并将其结果纳入下一步决策。进行物理干预后，智能体会收集新的观测，以判断预期的局部变化是否已经发生。未成功的动作、几何约束以及工具级别的失败反馈都会保留在交互历史中，使智能体能够调整参数、选择其他工具，或终止无效策略，而不是反复发出相同的干预指令。

4.3. 经验驱动的演化

演进 会保留并利用早期运行积累的经验，使智能体能够识别类似的失败，并在后续回合中作出应对。执行记录有助于诊断失败并指导纠正

动作 [72, 73], , 而技能库则让智能体能够保留成功行为，以便日后使用 [74]。在 Innata-0 中，WAM 会为当前子任务生成动作。根据由此产生的观测，智能体学习识别哪些视觉转变表明确实发生了失败。当智能体进行干预时，工具调用产生的物理效应会揭示适用于该任务的有效工具组合。验证技能会捕捉失败的可观察迹象，而恢复技能则描述哪些工具组合有效，以及何时使用它们。智能体会为每项任务将这两项技能作为版本化工件加以维护。在经验积累期间，智能体会诊断失败，并可重放失败过程以测试候选干预。它利用新的证据来完善现有技能对，并在推广每个修订版本之前进行验证。在部署期间，它会使用最新的已验证技能对进行在线验证和实时恢复，并在整个运行过程中保持该版本固定。最后，选定的轨迹展开片段也会为后续的 WAM 训练提供监督。

经验积累与初始技能合成。经验积累在留出的任务实例和场景配置上进行。每次轨迹展开都会记录同步的观测、WAM 动作、验证决策、工具调用、物理干预和最终结果。第零代开始时不具备特定任务的验证或恢复经验。系统使用原生 WAM 执行任务，生成初始的成功与失败轨迹集合。随后对失败回合进行回顾性分析，以识别最早出现的因果可观测偏差、导致失败的物理机制，以及仍可实施干预的状态。尽管可以在事后诊断中使用回合的最终结果，但由此得出的失败特征必须能够仅依据拟议验证点上可获得的观测来表述。这确保了回顾性获得的知识仍适用于因果在线应用。

在经验积累过程中，可以重放失败轨迹，以重建已诊断出的状态并测试候选恢复策略。诊断智能体可以调整工具选择、几何目标、运动参数和干预时机，同时观察这些操作产生的物理效应。只有在执行有效物理动作并产生预期的符合任务要求的结果时，才会保留候选策略。失败的恢复尝试也会予以保留，因为它们揭示了不合适的工具、不可达的几何构型、不安全的干预点，以及不应采用某项策略的条件。

由此产生的证据被整合为初始的任务级技能对。重复出现的失败观测被概括为紧凑的验证技能，而经验证的干预轨迹及其适用条件则构成相应的恢复技能。特定于种子的坐标和物体实例被语义引用、相对几何关系以及可观测前置条件所替代，使所得技能能够迁移到新的场景配置中。

代际技能演化。 初始技能综合完成后，演化按代次推进。在每一代中，前一代的验证技能和恢复技能会被整合，并用于新的任务实例。验证技能指导在线异常识别，而恢复技能则在确认异常后，为选择和组合工具提供特定于任务的参考。因此，系统评估的不仅是现有技能能否复现先前成功的干预，还评估其能否泛化到新的物体配置、机器人状态和失败变化。

新的运行轨迹会产生额外证据。成功执行能够细化正常状态转移的范围并减少误干预，而失败的验证与恢复尝试则会暴露缺失的签名、不正确的适用条件或不足的恢复策略。在这一积累阶段，新出现的失败回合还可以再次进行诊断和重放，以形成候选修订。在每一代结束时，现有的任务级验证技能和恢复技能会与两类新近经验证的经验相整合：一类是闭环在线执行期间收集的证据，另一类是通过失败重放获得的诊断或恢复经验。所得修订构成下一代的候选技能集，但仍需在全新的任务配置上进行验证。

整合并非只是追加完整轨迹，而是提取运行时智能体所需的最少可复用知识，合并等效的失败签名和工具流程，记录正面与负面的适用性证据，并保留指向支持性回合的链接以供审计。每次技能修订

都会进行版本管理，并在全新的留出配置上评估，通过后才予以晋升。一旦某个评估实例被用于技能修订，它便成为开发经验，不再作为下一代的留出证据重复使用。

经过多代演进，验证技能能够更准确地判断何时应针对表面上的偏差进行干预；恢复技能则积累了更丰富的策略，并明确了在这些策略之间进行选择的条件。因此，技能演进能够提升任务层面的能力，而无须将每种意外情况都纳入 WAM 策略参数。

利用积累的技能进行部署。部署遵循比经验积累更严格的协议。对于当前任务，智能体仅调用最新经过验证的验证技能和恢复技能，并在整个执行过程中保持它们不变。验证技能以当前任务状态和因果观测历史为条件，用于判断物理执行是否偏离预期。当有充分证据表明存在异常时，会将相应的恢复技能提供给恢复代理，作为任务特定的指导，帮助其组合使用通用具身工具。

部署期间的恢复直接作用于当前实时环境状态。失败回合不会被重置或重放，也无法通过重建先前状态来测试替代恢复策略。这一约束确保了真正的在线干预，而非反复进行离线搜索。部署轨迹仍会被记录，并可在之后纳入新一轮经验积累，但在当前评估代次中不会修改正在使用的技能。这种分离既能避免未经验证的单回合特定改动影响运行时行为，也为后续技能演进保留了清晰的路径。

4.4. 智能体基础设施

随着机器人的动作改变环境，智能体必须根据观察到的效果来决定下一步行动。我们将这一反馈循环组织为任务图（图 13），并遵循 ReAct [1] 和 Inner Monologue [70]中推理与动作交错处理的方式。该图规定了规划、动作生成、执行、观测和验证之间的依赖关系。对于规划，智能体会选择当前子任务，并向 WAM 发送相应指令。物理接口将 WAM 生成的动作传递给机器人或模拟器，并将观测结果返回给智能体。智能体会验证物理结果；如果确认存在异常，就暂停 WAM，并使用恢复工具进行干预（恢复）。该图根据验证结果选择下一步，而运行时会保留完整的执行记录，以支持后续的技能修订（演进）。我们的设计将任务逻辑保留在图中，因此同一工作流可以通过兼容接口，搭配不同模型、技能和机器人平台运行。

任务图。 我们借鉴 LangGraph 通过共享状态、节点和条件转换来组织工作流的方式，同时采用自有运行时来执行具身任务。节点代表规划、动作生成和验证等操作。节点之间的连接指定必须具备哪些结果，以及不同结果将导向何处。编译器解析图定义，调度器则根据这些依赖关系运行节点。任务上下文保留总体目标、当前进度和中间结果，以便后续决策能够基于先前的执行情况展开。复杂步骤可以展开为子图，例如在验证期间将观测分析与目标检查分开。每个子图都有自己的输入，可以访问总体任务上下文，并将其内部结果单独保存。父级使用子图定义的输出，而非其内部节点结果。因此，验证过程可以增加额外检查，而无需改变父级调用它或使用其判断结果的方式。

工具与记忆。 图决定何时需要某项能力，以及其结果如何影响任务；工具运行时则提供对该能力的访问。规划和验证可以使用不同的输入和标准调用同一个推理模型，而动作节点则可以使用策略或现有技能。只要替换后的策略保留预期的输入和输出，替换策略就不会改变该节点在工作流中的作用。记忆通过相同的工具接口参与其中。任务上下文保存当前执行期间所需的信息

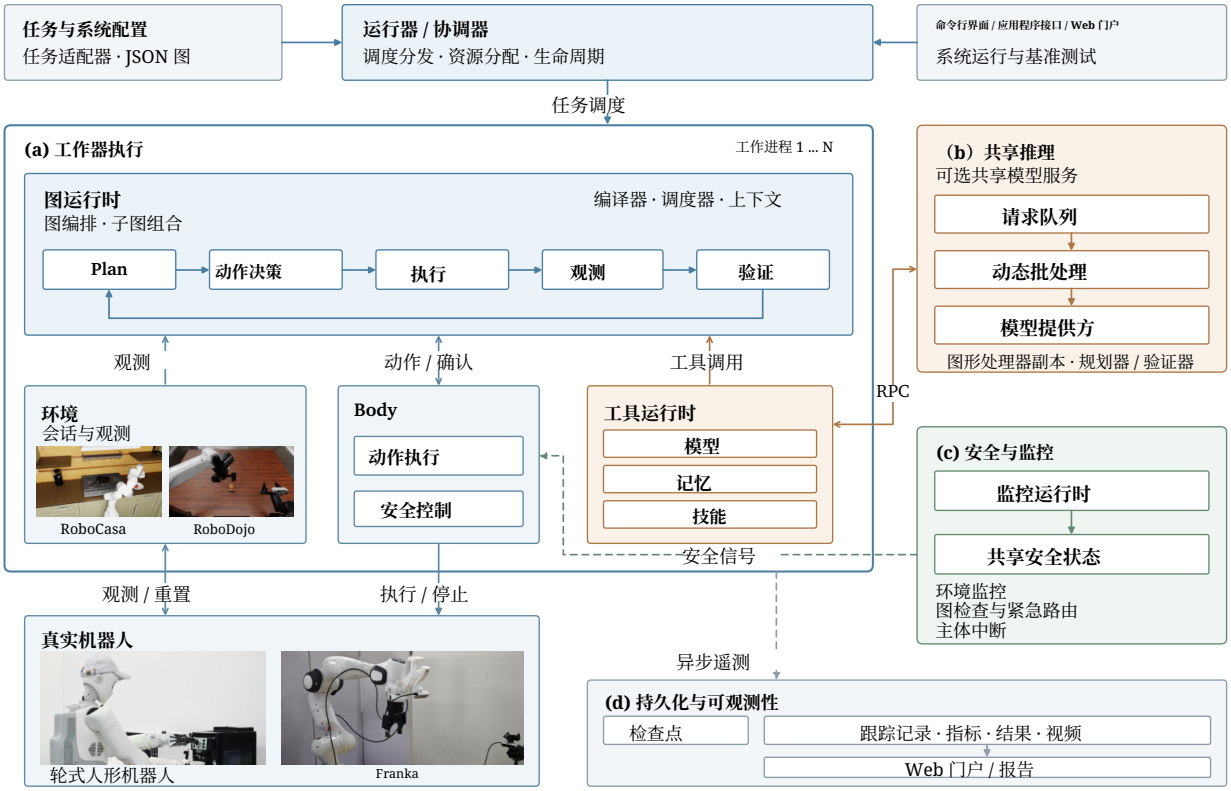


图13：智能体基础设施与执行流程。协调器将任务调度给工作器。在工作器内部，示例图将规划、动作生成、执行、观测和验证连接起来；反馈决定后续决策。节点通过工具运行时调用模型、记忆和技能。环境提供观测，而身体通过物理集成执行动作。共享推理服务使工作器能够复用模型资源。安全监控会影响图的推进和动作执行，执行记录则会汇入 Web 门户和报告。实线箭头表示所示的执行、数据或服务交互；虚线路径传递安全信号和遥测。平台照片展示了模拟器和机器人类别。

，而持久记忆则保留跨任务的信息。沿用记忆增强智能体中检索经验的做法， [76], 节点可以检索与当前决策相关的信息。存储、选择和摘要仍属于记忆实现的一部分。任务图规定何时需要这些信息以及如何使用它们，而不会将记忆流程嵌入每个调用节点。

物理反馈与安全。 动作生成后，身体会执行该动作并报告其状态，而环境则提供结果场景的观测。验证会结合这两个来源，因为命令完成与任务成功是不同的事件。夹爪命令可能已经执行完毕，但物体未必已被抓取；需要通过下一次观测来判断是继续还是尝试恢复。验证会将这一判断反馈给图，由图选择下一步。本体与环境接口以一致的方式呈现这些交互，适用于各种模拟器和机器人；平台适配器则负责处理坐标系、传感器格式和动作语义。安全监控通过共享安全状态，同时作用于图执行和本体控制。本体控制可以停止正在进行的动作，图也可以针对同一状况改变后续路径，从而在任务推进中考虑此次中断。

运行时与共享资源。 协调器将配置好的任务调度分发给工作器。每个工作器都维护自己的任务上下文，以及与环境或机器人的交互会话；而模型、技能和记忆服务则可以常驻，以供后续任务使用。这样就将任务的生命周期与其所用资源的生命周期分离开来：一个任务结束时，不必卸载下一个任务仍需使用的模型。并发

工作器也可以使用共享模型服务，由这些服务对兼容的推理请求进行排队和批处理 [77, 78]。基础设施维护请求与任务之间的关联，而模型服务负责推理调度。结果会返回给发出请求的工作器，每个工作器都根据自身的观测和反馈继续执行。检查点、跟踪记录、指标、结果和视频记录了这一执行过程。可视化网页界面可用于查看这些记录，以检查任务进度，并将决策与其执行结果联系起来。

5. 评估

我们评估了 Innata-0 的具身能力及其任务执行表现。理解能力基准测试评估语义与空间定位，而 WorldArena 2.0 [9] 评估动作条件视频预测。在 RoboCasa365 [10] 和 RoboDojo-Sim [11], 上，我们评估了 AGENT+WAM 配置在闭环任务中的表现，考察任务上下文、物理结果验证和经验证的经验如何引导执行与恢复。UniVTAC [79] 和 Bench2Dex [80]则进一步评估触觉条件控制在夹爪操作和多阶段灵巧操作任务中的表现。

5.1. 理解能力评估

我们报告了涵盖通用视觉理解、具身推理和空间定位的 18 项基准测试结果。表 5 将理解塔与七个参考模型进行了比较。在报告的 18 项基准测试中，我们的模型在所列模型中有九项排名第一，十四项跻身前两名。它在 SPARBench 上达到 69.08，在 EmbSpatial-Bench 上达到 84.42，在 ERQA 上达到 48.50。相较于 Qwen3-VL-8B [16] 初始化模型，它在 18 项基准测试中的 16 项上表现有所提升，其中 SPARBench 提高了 29.79 点，RoboAfford 提高了 22.30 点，VSI-Bench 提高了 6.14 点。通用视觉任务的表现也有所提升，包括 CountBenchQA、RealWorldQA 和 V-Star Bench。

5.2. 视频生成

我们在 WorldArena 2.0 Track 1 [9] 上评估视频预测。公开测试集包含 1,000 个取自 RoboTwin 2.0 [102] 轨迹的机器人操作片段。每个片段提供一个初始 RGB 帧、一条语言指令和一条机器人动作轨迹。任务是预测初始帧之后的视频。

评估指标。 我们报告六个类别中的 15 项归一化指标：视觉质量、运动质量、内容一致性、物理规律遵循度、三维准确度和可控性。得分范围为 0 到 100，越高越好。对于评估的模型，在汇总各回合得分前，我们将动态程度、光流得分和运动平滑度分别封顶为每个回合对应的真实值。总体 EWM 分数按 15 项分数不加权取平均计算。

基线模型与推理协议。 我们使用相同的官方 Track 1 评估代码评估我们的模型，以及已发布的 ABot-PhysWorld RoboTwin 2.0 后训练检查点 [103], FlowWAM [104], 和 BWM [105] 。对于这四个模型，我们使用 Real-ESRGAN [106] 将初始帧放大至 640×480。我们的模型以动作轨迹为条件；每个基线模型则使用其已发布检查点所支持的输入。对于评估所用检查点尚未公开或未发布的模型，表 6列出了原始论文中的得分，未重新计算。

结果。 我们的模型取得了 67.63 的 EWM 分数，在本次比较中纳入的开源模型及公开报告结果的模型中排名第一（表 6）。相比 BWM、FlowWAM 和 ABot-PhysWorld，分别高出 1.48、1.77 和 6.44 点。相较于各项指标上表现最佳的本地基线模型，最大提升体现在图像质量（67.49, +13。18）和美学质量（51.02, +9。17）上。我们的

表 5：18 项基准测试的理解性能。我们将模型与七个参考模型进行比较。粗体和下划线分别表示每行中的最高分和第二高分。所有得分均以 0–100 的尺度报告，分数越高越好。

基准测试	Gemma 4 8B [81]	Qwen3-VL 8B [16]	Embodied-R1.5 8B [82]	MiMo-Embodied 7B [83]	RoboBrain 2.5 8B [84]	RynnBrain 8B [85]	Cosmos 3 Nano 8B [40]	**我们的方法 8B**
通用视觉理解								
BlinkDepth [86]	71.77	83.87	86.29	78.23	83.06	91.13	**93.55**	89.52
BlinkSpatial [86]	77.62	83.22	75.52	76.92	**86.01**	78.32	81.12	79.02
CausalVQA [87]	43.00	65.00	56.00	54.00	48.00	46.00	63.00	**69.00**
CountBenchQA [88]	55.65	90.35	77.21	79.47	89.12	90.55	85.01	**92.81**
RealWorldQA [89]	49.02	71.11	69.02	63.27	73.20	70.33	72.55	**73.86**
RefCOCO [90]	72.09	**87.79**	75.01	78.91	78.23	83.34	84.02	86.14
V-Star Bench [91]	23.56	82.72	81.68	48.17	57.59	85.34	86.91	**89.53**
CV-Bench [92]	52.91	82.34	86.95	77.12	86.06	88.15	87.96	**89.18**
具身理解与定位								
3DSRBench [93]	49.01	56.60	56.25	54.36	49.48	56.14	55.05	**58.38**
EmbSpatial-Bench [94]	62.75	77.94	75.16	76.07	77.72	79.20	81.59	**84.42**
ERQA [7]	38.25	43.00	42.00	38.25	44.75	27.75	48.00	**48.50**
MindCube [95]	38.36	29.70	34.48	44.80	27.74	**57.63**	40.45	50.92
MMSI-Bench [96]	27.70	31.40	29.80	25.90	28.90	**39.30**	36.10	34.70
MMSI-Video-Bench [97]	22.33	29.93	28.12	29.66	**30.56**	30.02	26.40	30.47
RoboAfford [98]	30.78	51.67	72.57	62.02	**77.17**	76.04	70.67	73.97
SPARBench [99]	26.44	39.29	40.33	16.27	42.34	49.47	55.02	**69.08**
SWITCH [100]	24.87	23.27	21.93	**26.53**	26.33	21.93	20.80	26.47
VSI-Bench [101]	31.65	55.46	56.15	46.70	43.77	**62.98**	55.00	61.60

模型在该组的光度一致性（44.73）和透视性（95.68）指标上也均居首位。其轨迹准确度为 62.81，比 BWM 高 3.54 点，比 FlowWAM 低 1.25 点。在背景一致性和指令遵循方面，表现仍低于这些指标上最强的基线模型。

表 6：WorldArena 2.0 Track 1 视频预测。我们评估了自己的模型和已发布的 RoboTwin 2.0 后训练检查点（†）；标记 ‡ 的得分来自原始论文。在我们评估的四个模型中，粗体和下划线分别标出 EWM 的最佳值和次佳值。

模型	EWM	视觉质量			运动质量			内容一致性			物理学遵循性		3D 准确性		可控性	
		IQ	AQ	JS	DD	FS	MS	SC	BC	PC	IN	TA	DA	PV	IF	SA
在本文中评估																
ABot-PhysWorld† [103]	61.19	54.31	40.66	89.95	22.44	12.02	67.10	83.28	89.48	18.65	70.20	29.35	81.70	93.66	75.92	89.20
FlowWAM† [104]	65.86	48.95	40.84	97.31	21.89	11.77	66.77	79.51	86.85	29.60	73.70	64.06	98.98	94.74	83.06	89.92
BWM† [105]	66.15	52.08	41.85	97.77	22.28	11.52	67.30	76.47	82.88	40.55	74.70	59.27	97.82	94.78	83.12	89.91
Ours	**67.63**	67.49	51.02	95.37	22.01	11.48	65.18	76.35	81.14	44.73	72.28	62.81	98.29	95.68	80.80	89.88
原论文中报告																
DreamX-Phi 1‡ [107]	60.65	63.25	43.38	92.93	22.90	5.81		63.26	71.93	84.52	14.29	57.36	57.15	98.55	82.24	61.62	90.53
CtrlWAM‡ [108]	61.75	39.40	37.10	96.30	21.20	10.90	66.00	73.70	78.50	47.40	66.80	41.20	95.50	88.70	73.80	89.90
Ropedia W0‡ [109]	66.92	53.57	41.04	96.00	25.48	15.75	67.55	79.51	84.17	27.98	77.42	63.78	98.56	97.72	85.50	89.73

指标缩写。 智商：图像质量；逆商：美学质量；JS：JEPA 相似度；DD：动态程度；FS：光流得分；MS：运动平滑度；SC：主体一致性；BC：背景一致性；PC：光度一致性；IN：交互质量；TA：轨迹准确度；DA：深度准确度；PV：透视性；IF：指令遵循；SA：语义对齐。

5.3. 策略评估

我们通过针对基准测试的监督微调（SFT），将预训练模型适配到 RoboCasa365 [10] 和 RoboDojo-Sim [11]，并在闭环仿真中评估所得策略。

SFT 设置。 对于 RoboCasa365，我们使用 Human300 演示数据进行微调，并根据每项任务可用的训练窗口数量重新平衡任务采样：窗口较多的任务降低采样权重，窗口较少的任务则提高采样权重。这提高了原子任务的相对曝光度，与预训练期间采用的子任务分解和条件输入相辅相成。为同时支持独立执行，以及与使用同一策略的规划系统集成，我们将 70% 的纯指令示例与 30% 的增强示例混合，后者由 `current_plan` (string) 和 `history_plan` (list) 补充，并基于视觉语言模型子任务分割和标注生成。我们训练 26k 步，全局批次大小为 1,848，并以当前多视角 RGB 观测和机器人状态（H1）作为条件输入。

对于 RoboDojo，我们从预训练的 WAM 检查点初始化，并在 RoboDojo-Sim 演示数据上进行微调。为减少任务不平衡，我们首先从 35 个任务组中均匀采样，然后在每个组内采样轨迹。每个样本都包含原始任务指令以及 `current_plan`和`history_plan`。策略以 H5 个最近的多视角 RGB 观测以及当前机器人状态为条件。对于记忆任务，最早的观测会替换为回合初始观测。模型以 25Hz 预测未来 16 个视频帧及相应的 14D 关节控制目标。我们训练 24k 步，全局批次大小为 1,680。表7总结了训练设置。

表 7：策略监督微调设置。 RoboCasa365 仅使用 Human300 演示数据，不包括 MG60、Target Human50 以及 RoboCasa365 的其他演示数据集。JPEG 质量指存储的训练图像的质量。

设置	RoboCasa365	RoboDojo
监督微调数据	Human300	RoboDojo-Sim 轨迹
初始化	预训练模型	预训练模型
训练步数 / 全局批次 size	26k / 1,848	24k / 1,680
条件输入	语言、多视角 RGB、机器人状态	语言、多视角 RGB、机器人状态
图像存储	JPEG，质量 95	JPEG，质量 95
多视角处理	视图合成为一个 RGB画布	视图合成为一个 RGB画布
观测历史	H1：当前图像和机器人状态	H5：最近五个时间点的多视角 RGB 点和当前机器人状态；最早观测替换为回合初始记忆任务的观测
预测时域	16 个未来动作及对应的视频	16 个未来动作及对应视频
动作接口	12维控制指令；掩码 128维内部表示	14维关节目标；掩码 128维内部表示
条件损失掩码	不包含观测图像和机器人状态	不包含观测图像和机器人状态
视频/动作学习率	10^{-5} / 10^{-4}	10^{-5} / 10^{-4}
学习率衰减	余弦衰减	在 1 万步内保持不变，随后进行线性衰减
EMA	完整 MoT 网络；动量 0。995	完整 MoT 网络；动量 0。995
视频时间步采样	Waver；分辨率相关偏移	均匀偏移分布；偏移 5
动作时间步采样	均匀基分布；偏移 3，独立于视频	与视频时间步共享
训练目标	视频与动作流匹配	视频与动作流匹配
文本丢弃	0.1	0.0

评估设置。 在 RoboCasa365 上，我们使用经过 26k 步训练的 EMA 检查点，在 50 个任务上进行评估；每个任务采用 50 个固定初始条件，共计 2,500 个回合：900 个 Atomic-Seen、800 个 Composite-Seen 和 800 个 Composite-Unseen。策略接收使用初始观测扩展后的详细任务指令；既不提供 `current_plan` nor `history_plan`，也没有在线规划器在执行期间更新引导信息。每张相机图像都会先经过 JPEG 质量-95 编码和解码，然后进行视图合成。推理采用 30 个去噪步骤、分类器自由引导 2 和时间步偏移 5，并使用相互独立的视频噪声与动作噪声，同时在每次决策时更新采样种子。策略以 20Hz 执行 16 个动作，然后接收下一次观测。每个控制步都会检查成功与否，并采用特定于任务的回合时长上限。

在 RoboDojo-Sim 上，我们评估了一个视觉语言模型 + WAM 系统，该系统涵盖 42 项官方任务，分为 54 个评估条目。视觉语言模型首先根据初始观测和任务指令生成有序的子任务计划序列。执行期间，第二个视觉语言模型会评估当前子任务是否已完成，使系统每次检查最多推进一个计划；它不会生成、改写或重新排序计划。WAM 接收原始指令、当前计划、已完成计划的历史记录、H5 多视角 RGB 观测以及当前机器人状态。图像在视图合成前经过 JPEG 质量为 95 的编码和解码。推理会预测 16 个动作，并以 25 Hz 执行其中 8 个，采用 30 个 UniPC [110] 去噪步骤、引导尺度 1、时间步偏移 5，并为每个动作块使用新的采样种子。我们报告得分和成功率，并对五个能力维度的各项指标进行等权平均。

对于完整系统评估（AGENT+WAM），我们在 RoboCasa365 和 RoboDojo-Sim 上，通过四代开发时交互收集技能。对于每项任务，我们在每一代中随机采样五个场景初始化，并通过种子或（种子、布局）进行标识。布局、物体放置及其他随机化场景属性，均根据这些标识对应的基准测试进行实例化。请注意，这些场景与正式部署中的场景互不相同。每一代的所有回合终止后，我们会将其中的成功执行、可观测的失败模式、事后诊断和恢复结果汇总为一对新的任务级验证与恢复技能。部署使用最终一代经过验证的技能对。遵循 RoboDojo 基准测试，我们因仿真不稳定或场景初始化失败，将这些回合从得分和成功率中同时排除。

表 8：RoboCasa365 成功率（%）。参考得分遵循其发布的协议，并取自公开排行榜 （2026年11月18日），GR00T 2 的结果除外，该结果取自其报告中的表19 [3]。阴影行（†）在 50 项任务中的 23 项中使用 WAM 作为默认执行器，在其余 27 项中使用 ASPIRE 生成的程序；两种情况下均可使用代码即策略恢复。

模型	Type	总体原子已见		综合已见	综合未见
GR00T 2–监督微调 [3]	WAM	**67.2**	87.4	**73.6**	38.2
Awomo-0.5V [111]	WAM	62.5	85.0	60.4	**39.4**
公理-0 [112]	VLA	61.6	86.5	59.4	35.8
RPent [113]	智能体+视觉-语言-动作	59.2	87.8	43.8	42.5
WorldDreamer v5 [114]	WAM	57.5	82.0	56.0	31.3
小米-机器人学-1 [61]	VLA	57.4	80.2	57.1	32.1
美团-机器人学-0 [115]	VLA	55.0	77.3	57.8	27.1
ABot-M0.5 [116]	WAM	40.3	75.6	37.7	3.3
$\pi_{0.5}$ [117]	VLA	16.9	39.6	7.1	1.2
GR00T 2† [3]	智能体+WAM	81.8	88.2	86.9	69.5
Ours	WAM	60.2	**88.4**	56.1	32.4
Ours	智能体+WAM	65.4	**90.6**	58.9	**43.6**

在 WAM执行期间，每经过 80 个模拟器步数就会触发一次阶段审查。智能体在正常运行期间每经过 256 个模拟器步数、在关键窗口内每经过 128 个步数，以及在明确的高风险事件发生后立即验证执行情况。对于在线恢复，智能体接管当前的实时模拟器状态

表 9：在 RoboDojo-Sim 上的评估。 每个条目以“得分 / 成功率（%）”的形式报告，数据来自 公开排行榜，截至 2026年11月18日。报告中的平均值是泛化、精度、长时域、记忆和开放五项得分的等权平均值。Cosmos3-Nano† 结果是对 Cosmos3-Nano 进行独立复现和微调后的评估结果， [40]由 FluxVLA 项目报告 [118]。最佳结果以粗体显示。

方法	Type	平均值	泛化	精度	长时域	记忆	Open
模型赛道							
InternW0-Δ [119]	WAM	**30.77** / 23.91	30.09 / 22.78	31.98 / 23.25	**46.29** / 29.33	34.67 / 34.00	10.84 / 10.17
Liber-0 预览版 [120]	视觉语言模型 + WAM	30.74 / **25.52**	24.99 / 18.61	**38.28** / **33.17**	45.98 / **33.17**	37.77 / 37.33	6.68 / 5.33
美团-机器人学-0 [115]	VLA	21.68 / 15.73	16.31 / 10.50	21.24 / 11.17	27.07 / 13.92	**39.52** / **39.22**	4.28 / 3.83
ME-Brain-1.0 [121]	视觉-语言-动作 + WAM	21.67 / 15.99	21.81 / 15.34	24.23 / 15.92	32.96 / 20.58	23.53 / 22.78	5.83 / 5.33
GalaxeaVLA（G0.5） [122]	VLA	20.23 / 14.88	18.46 / 12.83	28.25 / 20.42	44.12 / 32.25	8.61 / 7.33	1.73 / 1.58
小米-机器人-1 [61]	VLA	20.07 / 13.93	23.54 / 17.00	26.69 / 18.83	38.39 / 23.67	7.81 / 6.56	3.94 / 3.58
MachEmbodied-U0 [123]	WAM	17.78 / 11.51	16.07 / 8.78	24.20 / 16.67	36.66 / 22.25	8.16 / 6.67	3.78 / 3.17
Hy-Embodied-0.5-VLA [124]	VLA	13.07 / 8.80	11.78 / 8.39	13.81 / 8.00	25.74 / 14.92	13.37 / 12.11	0.65 / 0.58
Cosmos3-Nano [40]†	WAM	3.03 / 0.93	2.58 / 1.00	2.29 / 0.00	8.74 / 2.75	1.23 / 0.67	0.30 / 0.25
Ours	视觉语言模型 + WAM	23.04 / 15.98	30.88 / 23.67	21.58 / 11.04	32.31 / 18.56	19.90 / 17.00	10.53 / 9.63
Ours	智能体 + WAM	30.02 / 24.17	**39.01**/ **33.20**	27.07 / 19.20	32.99 / 20.00	23.58 / 21.78	**27.46** / **26.66**

无需重放之前的轨迹。最多可进行 20 轮推理—工具交互。调用工具时，纯感知操作不会推进模拟器，而执行运动操作则会消耗控制步数。根据我们的协同设计协议，在这两种设置中，WAM 均作为默认执行器。智能体会验证执行情况，并在需要时介入恢复。我们不会合成任务级代码控制来充当默认执行器。

结果分析。 在 RoboCasa365（表 8）上，Innata-0-WAM 的总体成功率达到 60.2%（1,504/2,500）。原子级已见任务的成功率达到 88.4%，而综合已见任务和综合未见任务的成功率分别为 56.1% 和 32.4%。原子级任务与综合任务之间的差距凸显了将单项操作技能转化为可靠多阶段执行的难度，对于未见过的任务组合尤其如此。

使用 Innata-0-Agent 后，总体成功率提升至 65.4%，增加了 5.2 个百分点。Composite-Unseen 上的提升最大，成功率从 32.4% 升至 43.6%，增加了 11.2 个百分点。两种配置使用相同的 WAM 检查点和完整任务指令。智能体会验证物理结果并调用基于工具的恢复，使用从经验证的交互经验中不断完善的技能。这些结果表明，结果验证与基于经验的恢复相结合，能够共同提升任务完成率，并通过智能体—模型协作扩大模型的有效运行范围。GR00T 2 的 AGENT+WAM 在 50 项任务中的 27 项上直接使用 ASPIRE 生成的程序作为默认执行器，同时也提供 Code-as-Policy 用于恢复 [3]。而在我们的系统中，WAM 始终是各项任务的名义动作策略，智能体则负责任务结构化、因果验证以及通过工具组合实现实时恢复。因此，我们将此结果标记为不同执行设置下的参考值。

在 RoboDojo-Sim（表9）上，我们的视觉语言模型 +WAM 系统取得了 23.04 的总体得分和 15.98% 的成功率。其最强的结果出现在要求泛化的任务上。在泛化维度中，其得分/成功率达到 30.88 / 23.67%，分别超过 InternW0-Δ 0.79 个得分点和 0.89 个百分点。在开放任务上，其得分/成功率也达到 10.53 / 9.63%。这些结果表明，WAM 提供了强大且可迁移的物理执行基础，尤其是在任务要求将先前学习到的操作行为适应于陌生物体、布局或指令时。通过完整的系统参与，Innata-0-Agent 可以进一步扩展 Innata-0-WAM 的能力。加入 Innata-0-Agent 后，五个维度中的得分和成功率均有所提升。平均得分从 23.04 上升至 30.02，平均成功率则从 15.98% 提高至 24.17%（+8提高了 8.19 个百分点）。除开放任务上取得显著提升外，Innata-0-Agent 还将泛化和精度维度的成功率分别提高至 33.20% 和 19.20%。

视觉分析。 我们通过分析智能体干预如何改变 RoboCasa365 和 RoboDojo 中各回合的结果，评估系统的恢复能力。如图 14所示，在 RoboCasa365 上，

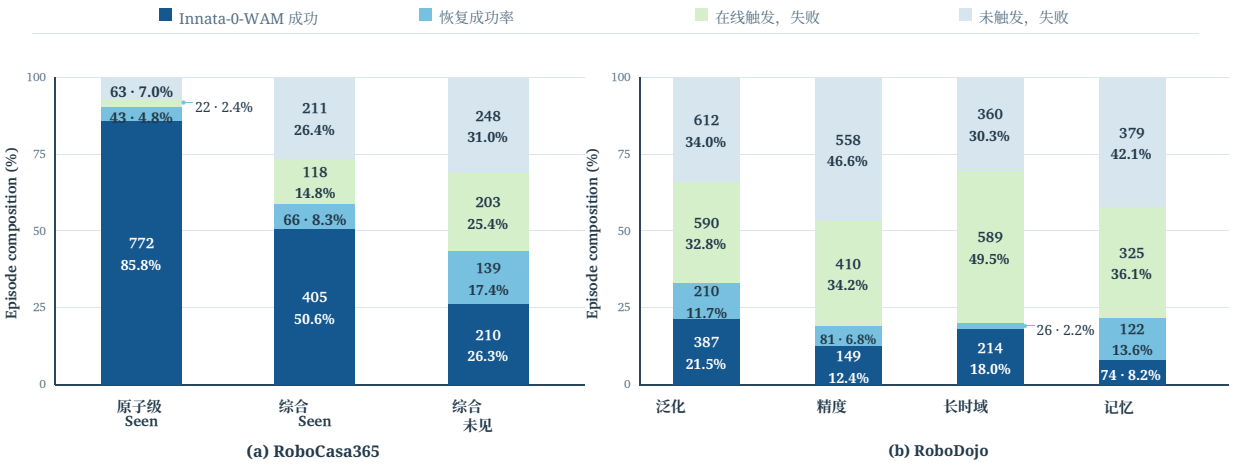


图14：RoboCasa365 和 RoboDojo 上的恢复能力分析。每个柱形将回合划分为四种互斥的最终结果：Innata-0-WAM 成功，即由 Innata-0-WAM 完成且未进行恢复；恢复成功，即在恢复干预后完成；在线触发但失败，即在线验证触发了恢复，但回合仍未成功；未触发且失败，即未触发恢复。

随着任务难度增加，恢复参与度也随之提高。Atomic-Seen 主要依赖 Innata-0-WAM 执行，其成功率为 85.8%，恢复参与度最低。Composite-Seen 的恢复参与度有所提高，Composite-Unseen 则最高；与此同时，Innata-0-WAM 的成功率分别从 85.8% 降至 50.6% 和 26.3%。这些结果表明，系统能够自适应地权衡速度更快但能力较弱的基础模型与速度较慢但能力更强的智能体：简单的回合由 Innata-0-WAM 完成，而随着任务复杂度增加，智能体参与得更加频繁。

在 RoboDojo 上，恢复在不同任务维度上的作用各不相同。泛化带来的恢复成功回合数最多，而记忆相较于 Innata-0-WAM 执行的相对提升最大。记忆的 Innata-0-WAM 成功率最低（8.2%），但恢复增加了 122 个成功回合，使最终成功率升至 21.8%。该智能体的一项优势是，它能够将当前场景与回合早些时候设定的任务要求进行核对，例如，检查正在拿起的物体是否与此前展示的目标一致。即便抓取本身看起来成功，这项检查也能发现抓错物体的情况。

图 15 所示的两个案例展示了在不同基准测试、任务时程和失败结构下的在线恢复。在 RoboCasa365 煎制任务中，智能体确认物体与固定装置的关系有误：平底锅没有放在工作中的炉头上。它将目标转向平底锅手柄，重新建立抓取，抬升平底锅，并将其移至正确的热源。在 RoboDojo 记忆任务中，智能体确认两个 T 形积木与所需的姿态排列不匹配。随后，它抓取两个积木，交换它们的位置，并以正确的朝向将其释放。

由此产生的干预针对具体失败而设计：一种通过重新抓取和传输来纠正物体与固定装置的关系，另一种则执行双手协同重排，以恢复记忆中的空间配置。这些示例展示了恢复模块如何补充 WAM：它将已确认的视觉差异转化为有针对性的状态依赖型动作组合，使原本未完成的执行达到正式成功状态。

5.4. 触觉反馈策略评估

在本次评估中，我们关注触觉反馈如何帮助 WAM 处理仅凭视觉观测可能无法解决的接触交互，涵盖从原子操作技能到具身系统内的多阶段执行。为测试统一触觉表征的通用性，我们选择了两种使用不同类型触觉传感器和机械臂的基准测试。UniVTAC [79] 提供了八项基于夹爪的任务



图 15：RoboCasa365 和 RoboDojo 中成功的在线恢复案例。在 WAM执行期间验证任务状态不匹配后，智能体直接介入实时环境。在 RoboCasa365 中，智能体重新抓取错位的平底锅，并将其转移到工作中的炉头上（上图）。在 RoboDojo 记忆任务中，智能体识别出两个 T 形积木的错误排列，抓取两个物体，交换它们的位置，并按照要求的朝向释放它们（下图）。两次恢复最终都满足相应的正式成功条件。

用于评估接触控制和触觉感知。Bench2Dex [80]将触觉反馈与双手灵巧操作和长时程任务相结合，使我们能够评估系统多阶段执行中的接触处理和手指协调。

监督微调设置。 对于两个基准测试，我们都从联合动作与触觉预训练后的最终 WAM 检查点初始化任务专属策略。在 UniVTAC 上训练 20k 步，在 Bench2Dex 上训练 10k 步，全局批次大小为 128。每个策略以 H9 个同步的 RGB、机器人状态和触觉观测作为条件输入，其中包括当前观测和八个过去的观测。模型联合预测 16 帧未来视频，以及相应的动作和触觉输出。目标函数结合了视频与动作流匹配、触觉存在信号和代理监督，定义见（7）。视频生成的学习率为 10^{-5}，动作与触觉模块的学习率为 10^{-4}。

对于 UniVTAC，我们使用每项任务 50 个官方发布的示范（总计 400 个），为八项任务分别微调一个策略，并以任务指令作为条件输入，不使用规划。对于 Bench2Dex，我们在每个选定的任务—具身形态设置下使用 50 个专家示范，并在任务指令中加入`current_plan` 和 `history_plan`，这些内容源自分段后的专家示范。

评估设置。 对于两个基准测试，我们会在请求下一个动作块之前执行所有预测的动作。

在 UniVTAC 上，WAM 作为独立的闭环策略运行，在每个任务的 100 条官方测试运行轨迹中使用触觉反馈，不进行种子筛选或替换。仿真以 120Hz 运行，每个动作保持两个物理步（60Hz 控制）。推理使用非指数移动平均检查点进行 10 步欧拉采样，流时间偏移和引导尺度均设为 1.0。因果 H9 历史记录通过重复第一个观测进行初始化；触觉输入包括代理信号和存在掩码。

对于 Bench2Dex，我们排除记录的轨迹中不包含任何触觉反馈的任务，并从保留的任务中选出最具挑战性的十项任务。每项任务评估 200 次运行轨迹，四种官方设置各评估 50 次。我们在具身系统中评估 WAM。智能体调用

基于视觉语言模型的进度检查，利用近期的视觉证据、触觉事件摘要和任务上下文。确认任务已完成后，它会更新提供给 WAM 的当前计划和已完成子任务历史。我们报告综合得分、0.45 成功率 + 0.45 锁定阶段完成率 + 10 效率.，综合了稳定成功率、锁定阶段完成率和成功条件效率。对于十项任务的汇总结果，成功率和锁定阶段完成率取所有选定任务的平均值，而效率仅对有成功运行轨迹的任务取平均值。

表 10：UniVTAC 任务成功率（%）。平均值为全部八项任务的未加权平均值。ACT^{Tac} 使用 UniVTAC编码器；GW-Policy 和 Xiaomi-R0 分别是 GigaWorld-Policy 和 Xiaomi-Robotics-0 的简称。粗体和下划线分别表示显示数值中的最佳值和次佳值。

方法	插入 HDMI	插入 Hole	插入 Tube	抓取分类	Lift Can	Lift 瓶子	拉出 Key	将瓶子放入架子	Avg.
GW-Policy [125]	0.0	12.0	9.0	20.0	0.0	38.0	32.0	21.0	16.5
LingBot-VA [126]	38.0	42.0	96.0	17.0	0.0	0.0	58.0	0.0	31.4
$\pi_{0.5}$ [117]	8.0	25.0	74.0	49.0	6.0	**100.0**	35.0	34.0	41.4
ACT^{Tac} [79, 127]	28.0	24.0	56.0	99.0	29.0	71.0	46.0	31.0	48.0
Fast-WAM [128]	19.0	66.0	**98.0**	72.0	0.0	21.0	73.0	35.0	48.0
小米-R0 [129]	**69.0**	96.0	**98.0**	45.0	13.0	21.0	80.0	12.0	54.3
StarVLA-α [130]	24.0	52.0	69.0	68.0	65.0	32.0	51.0	88.0	56.1
FTP-1 [18]	4.0	64.0	79.0	99.0	65.0	97.0	48.0	47.0	62.9
InternVLA-A1 [131]	12.0	93.0	95.0	86.0	90.0	58.0	66.0	37.0	67.1
T-Rex [132, 133]	8.0	**100.0**	95.0	95.0	75.0	99.0	**99.0**	79.0	81.3
TacDyn-WAM [134]	12.0	91.0	97.0	99.0	87.0	97.0	97.0	72.0	81.5
DECO.p [135]	23.0	59.0	88.0	99.0	**99.0**	**100.0**	96.0	93.0	82.1
VTAM [136]	45.0	67.0	97.0	**100.0**	75.0	91.0	92.0	**100.0**	83.4
N_0-TWAM [137]	68.0	99.0	**98.0**	94.0	93.0	58.0	79.0	87.0	84.5
Ours	50.0	99.0	**98.0**	95.0	82.0	99.0	96.0	70.0	**86.1**

表 11：使用综合得分评估 Bench2Dex 十项任务的策略（越高越好）。T 表示官方任务标识符；总体得分根据这十项任务重新计算。粗体和下划线分别表示最高和第二高得分；— 表示每任务效率未定义。Tac 表示触觉输入。基线结果反映截至 2026年11月18日 的排行榜。

方法	任务得分 ↑										总体
	T21	T24	T26	T32	T34	T43	T60	T64	T73	T79	
ACT [127]	28.33	20.46	29.51	17.16	30.06	12.09	21.35	9.31	—	8.67	19.18
DP [138]	10.87	—	—	12.81	—	10.29	10.33	8.98	—	9.42	11.63
$\pi_{0.5}$ [117]	30.86	—	32.13	40.06	31.02	25.00	22.24	32.49	—	19.40	25.67
GR00T N1.5 [139]	23.35	—	27.79	33.22	33.55	17.21	13.90	19.79	25.46	31.91	23.58
GR00T N1.5^{Tac} [139]	22.15	—	30.30	34.54	31.85	19.17	14.74	21.14	23.43	28.93	23.60
GR00T N1.7 [139]	37.59	—	26.47	36.54	32.21	19.15	19.99	22.47	15.08	27.77	24.71
GR00T N1。7^{Tac} [139]	41.37	24.67	30.20	30.70	42.29	20.10	29.72	22.72	21.88	33.90	29.76
Mora V0 [140]	52.55	28.28	49.24	49.31	39.79	24.66	37.85	33.13	25.31	24.36	36.45
MXI [141]	56.25	14.87	36.36	36.91	44.65	22.41	37.21	41.18	—	21.57	32.47
Phoenix-MWA [142]	65.83	44.26	38.83	48.64	76.35	**45.03**	71.20	37.43	—	13.09	46.18
PyRUA-Dex [143]	37.35	31.14	**65.35**	51.00	—	30.56	53.73	55.51	—	**86.45**	46.35
GW-Policy-0。6 [125]	**84.12**	**63.24**	57.92	46.44	**77.59**	42.15	71.65	43.16	**31.51**	17.40	53.52
Ours	77.86	20.89	61.51	**51.76**	76.12	39.56	**72.79**	**58.45**	26.10	53.17	**53.82**

结果分析。 在 UniVTAC（表10）上，WAM 的平均成功率为 86.1%。在插入 HDMI 任务中达到 50.0%，在提起瓶子任务中达到 99.0%，表明其能够处理精确的插入操作并稳定维持抓取。这些结果表明，所学习的触觉表征保留了接触

对不同形式的操作有用的信息。

在 Bench2Dex（表 11）上，同一预训练 WAM 在智能体系统中取得了 53.82 的总体得分，在选定的十项任务子集中优于所有对比方法。在烘焙托盘准备（T32）、面包篮快餐装载（T60）和运动球杯分类（T64）任务中均排名第一，得分分别为 51.76、72.79 和 58.45。这两项基准测试中的表现表明，共享区域表征为跨不同传感器和具身形态学习高接触控制提供了有益基础。通过将异构触觉数据统一为通用格式，触觉中期训练使任务特定微调能够基于从多个来源收集的交互经验进行。

长时程任务进一步体现了模型与智能体协同设计的优势。冰箱葡萄酒双手交接倾倒（T34）涉及复杂的双手交接，现有编码智能体难以执行；而双手弹钢琴（T79）则对独立的 WAM/VLA 策略提出了严格的顺序编排和计时要求。协同设计的系统在这两项任务上均取得了有竞争力的结果。智能体负责维持任务结构并更新当前子目标，而 WAM 则生成完成目标所需的灵巧动作。这些结果支持我们对任务级编排与学习式操作进行协同设计，结合两者优势，以应对任何单一组件都难以独立完成的任务。

6. 局限性与未来方向

尽管 Innata-0 在具身理解、世界预测和机器人操作方面取得了令人瞩目的进展，但迈向原生具身智能仍面临挑战。跨环境、任务和机器人具身形态的泛化能力仍有待进一步探索。我们需要开展更广泛的真实世界验证，以评估系统在不同机器人具身形态、多样化环境和日益复杂的任务中的泛化能力，以及其在物理交互过程中应对动态变化和意外故障的鲁棒性。

展望未来，我们认为模型和智能体能力都有进一步提升的空间。在模型方面，我们旨在提升跨任务和机器人具身形态的泛化能力，增强物理理解与预测能力，并支持更复杂的灵巧操作。这需要扩大具身数据的多样性、质量和覆盖率，包括第一人称视频、真实机器人轨迹、触觉交互和灵巧操作数据。我们还计划利用真实世界执行和恢复经验，推动迭代式模型改进与智能体持续演进，使原生具身世界系统更具泛化能力和可靠性。在智能体方面，我们旨在借助更丰富的视觉、空间和触觉信息，提升多模态推理能力。我们还计划探索上下文学习和经验检索，以增强对陌生情境的适应能力。此外，降低词元消耗并优化推理效率，对于实现可扩展的真实世界部署也至关重要。

7. 贡献与致谢

所有贡献者均按姓氏字母顺序排列。

核心贡献者： Yu Bai, Ziyi Bai, Yuxuan Cai, Chenrui He, Wenhao Li, Zhe Lin, Heng Wang, Hai Yuan, Zhiwei Yu, Yuanhang Zhang

贡献者： Luyou Chen, Tiezhu Gao, ZhengGeng, Jinbang Guo, Dongxu Han, Diqi He, Börje F. Karlsson, Jintao Li, Zhiyu Li, Xiongkun Linghu, Guang Liu, You Liu, Zhen Luo, Mengsi Lyu, Yufeng Lyu, Zheng Lyu, Quanyue Ma, Chunlei Men, Ziqiang Ni, Tengfei Pan, Yijun Shen, Liang Sun, Shaoxuan Xie, Xiang Yang, Guocai Yao, Hongjun Zhang, Yingli Zhao

项目负责人： Yonghua Lin, Zhongyuan Wang✉, Huan Yang✉

致谢： 我们衷心感谢 Lightwheel 对自我中心数据的贡献，以及 PaXini 对触觉数据的贡献。

✉ 通讯作者：Zhongyuan Wang、Huan Yang（hyang@fastmail.com）

参考文献

[1] Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik Narasimhan 和 Yuan Cao。《ReAct：协同语言模型中的推理与行动》。发表于 国际学习表征会议，2023 年。网址 https://openreview.net/forum?id=WE_vluYUL-X。

[2] Timo Schick、Jane Dwivedi-Yu、Roberto Dessi、Roberta Raileanu、Maria Lomeli、Eric Hambro、Luke Zettlemoyer、Nicola Cancedda 和 Thomas Scialom。《Toolformer：语言模型可以自学使用工具》。载于神经信息处理系统进展，第36卷，第 68539–68551 页，2023。doi: 10.52202/075280-2997。URL https://proceedings.neurips.cc/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html。

[3] 英伟达。《GR00T 2：扩展机器人基础模型，以实现泛化、灵巧性和适应》。英伟达白皮书，2026年。URL https://research.nvidia.com/labs/gear/gr00t-2/。

[4] Runyu Lu、Yubo Wu、Ethan Kou、Haotian Lin、Letian Fu、Wenli Xiao、Ajay Mandlekar、Yinzhen Xu、Guanya Shi、Ken Goldberg、Ang Chen、Mosharaf Chowdhury、Jan Kautz、Yuke Zhu、Linxi "Jim" Fan 和 Guanzhi Wang。《ASPIRE：用于机器人学的智能体/技能发现》。*arXiv* 预印本 *arXiv:2607.00272*，2026年。

[5] 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、Anthony Brohan、Montserrat Gonzalez Arenas 和 Kehang Han。《RT-2：视觉-语言-动作模型将网络知识迁移至机器人控制》。载于 Jie Tan、Marc Toussaint 和 Kourosh Darvish 编，《第七届机器人学习会议论文集》，《机器学习研究论文集》第229卷，第2165–2183页。PMLR，2023。网址 https://proceedings.mlr.press/v229/zitkovich23a.html。

[6] Moo Jin Kim、Karl Pertsch、Siddharth Karamcheti、Ted Xiao、Ashwin Balakrishna、Suraj Nair、Rafael Rafailov、Ethan P Foster、Pannag R Sanketi、Quan Vuong、Thomas Kollar、Benjamin Burchfiel、Russ Tedrake、Dorsa Sadigh、Sergey Levine、Percy Liang 和 Chelsea Finn。《OpenVLA：一种开放源代码视觉-语言-动作模型》。载于 Pulkit Agrawal、Oliver Kroemer 和 Wolfram Burgard（编），第八届机器人学习会议论文集，机器学习研究论文集第 270 卷，第 2679–2713 页。PMLR，2025 年。网址： https://proceedings.mlr.press/v270/kim25c.html。

[7] Gemini Robotics Team、Saminda Abeyruwan、Joshua Ainslie、Jean-Baptiste Alayrac、Montserrat Gon- zalez Arenas、Travis Armstrong、Ashwin Balakrishna、Robert Baruch、Maria Bauza、Michiel Blokzijl、Steven Bohez、Konstantinos Bousmalis、Anthony Brohan、Thomas Buschmann、Arunkumar Byravan、Serkan Cabi、Ken Caluwaerts、Federico Casarini、Oscar Chang、Jose Enrique Chen、Xi Chen、Hao- Tien Lewis Chiang、Krzysztof Choromanski、David D'Ambrosio、Sudeep Dasari、Todor Davchev、Coline Devin、Norman Di Palo、Tianli Ding、Adil Dostmohamed、Danny Driess、Yilun Du、Debidatta Dwibedi、Michael Elabd、Claudio Fantacci、Cody Fong、Erik Frey、Chuyuan Fu、Marissa Giustina、Keerthana Gopalakrishnan、Laura Graesser、Leonard Hasenclever、Nicolas Heess、Brandon Hernaez、Alexander Herzog、R. Alex Hofer、Jan Humplik、Atil Iscen、Mithun George Jacob、DeepaliJain、Ryan Julian、Dmitry Kalashnikov、M. Emre Karagozler、Stefani Karp、Chase Kew、Jerad Kirkland、Sean Kirmani、Yuheng Kuang、Thomas Lampe、Antoine Laurens、Isabel Leal、Alex X. Lee、Tsang-Wei Edward Lee、Jacky Liang、Yixin Lin、Sharath Maddineni、Anirudha Majumdar、Assaf Hurwitz Michaely、Robert Moreno、Michael Neunert、Francesco Nori、Carolina Parada、Emilio Parisotto、Peter Pastor、Acorn Pooley、Kanishka Rao、Krista Reymann、Dorsa Sadigh、Stefano Saliceti、Pannag Sanketi、Pierre Sermanet、Dhruv Shah、Mohit Sharma、Kathryn Shea、Charles Shu、Vikas Sindhwani、Sumeet Singh、Radu Soricut、Jost Tobias Springenberg、Rachel Sterneck、Razvan Surdulescu、Jie Tan、Jonathan Tompson、Vincent Vanhoucke、Jake Varley、Grace Vesom、Giulia Vezzani、Oriol Vinyals、Ayzaan Wahid、Stefan Welker、Paul Wohlhart、Fei Xia、Ted Xiao、Annie Xie、Jinyu Xie、Peng Xu、Sichun Xu、Ying Xu、Zhuo Xu、Yuxiang Yang、Rui Yao、Sergey Yaroshenko、Wenhao Yu、Wentao Yuan、Jingwei Zhang、Tingnan Zhang、Allan Zhou 和 Yuxiang Zhou。《Gemini Robotics：将人工智能带入物理世界》。*arXiv* 预印本 *arXiv:2503.20020*，2025 年。

[8] Weixin Liang、Lili Yu、Liang Luo、Srinivasan Iyer、Ning Dong、Chunting Zhou、Gargi Ghosh、Mike Lewis、Wen tau Yih、Luke Zettlemoyer 和 Xi Victoria Lin。《混合变换器：一种面向多模态基础模型的稀疏且可扩展架构》。*arXiv* 预印本 *arXiv: 2411.04996*，2024年。

[9] Zhuohang Li、Yiding Ma、Yu Shang、Lei Jin、Shengyuan Wang、Jianjie Fang、Zhaolu Wang、Xvyuan Liu、Yinzhou Tang、Xin Jin、Ziyou Wang、Weikang Su、Xin Zhang、Haisheng Su、Weizhen He、Wei Wu、Haoyi Duan、Gordon Wetzstein、Xihui Liu、Dhruv Shah、Zhaoxiang Zhang、Zhibo Chen、Jun Zhu、Yonghong Tian、Tat-Seng Chua、Xin Wang、Wenwu Zhu、Chen Gao 和 Yong Li。《WorldArena 2.0：拓展具身世界模型在模态、功能性和平台方面的基准测试。 *arXiv* 预印本 *arXiv:2605.17912*，2026年。网址 https://arxiv.org/abs/2605.17912。

[10] Soroush Nasiriany、Sep Nasiriany、Abhiram Maddukuri 和 Yuke Zhu。RoboCasa365：用于训练和基准测试通用机器人的大规模仿真框架。载于国际学习表征会议，第 2026 卷，第 98643–98667 页，2026 年。URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a05003fdb1e9562ab0c0a9719ea4de10-Abstract-Conference.html。

[11] Tianxing Chen、Yue Chen、Zixuan Li、Junyuan Tang、Kailun Su、Haoran Lu、Weijie Wan、Baijun Chen、Songling Liu、Haowen Yan、Honghao Su、Zhiyang Dou、Kaixuan Wang、Dandan Zhang、Yunze Liu、Yan Qin、Qiwei Liang、Qiwei Wu、Zijian Lin、Wenwei Lin、Yuran Wang、Minghua He、Tianshu Wu、Ruihai Wu、Jingquan Zhou、Kai-Chong Lei、Haibao Yu、Yuanfeng Ji、Weiyang Jin、Guanyu Lin、Xiaofan Li、Qi Xiong、Renjing Xu、Zhongyu Li、Wenhao Chai、Enze Xie、Ziwei Wang、Yao Mu、Hao Dong、Wojciech Matusik、Mingyu Ding、Wenbo Ding、Ping Luo 和 Masayoshi To mizuka。RoboDojo：用于全面评估通用机器人操作策略的统一仿真与真实基准测试。*arXiv* 预印本 *arXiv:2607.04434*，2026年。URL https://arxiv.org/abs/2607.04434。

[12] Qwen团队。Qwen3.8-Flash-Next：迈向终极成本效益的新架构，2026年8月。网址：https://qwen.ai/blog?id=qwen3.8-flash-next。

[13] Shilong Liu、Zhaoyang Zeng、Tianhe Ren、Feng Li、Hao Zhang、Jie Yang、Qing Jiang、Chunyuan Li、Jianwei Yang、Hang Su、Jun Zhu 和 Lei Zhang。Grounding DINO：将 DINO 与基于定位的预训练相结合，用于开放集目标检测。载于 欧洲计算机视觉会议（*ECCV*），2024年。网址：https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/6319_ECCV_2024_paper.php。

[14] Nicolas Carion、Laura Gustafson、Yuan-Ting Hu、Shoubhik Debnath、Ronghang Hu、Didac Suris Coll-Vinent、Chaitanya Ryali、Kalyan Vasudev Alwala、Haitham Khedr、Andrew Huang、Jie Lei、Tengyu Ma、Baishan Guo、Arpit Kalla、Markus Marks、Joseph Greer、Meng Wang、Peize Sun、Roman Rädle、Triantafyllos Afouras、Effrosyni Mavroudi、Katherine Xu、Tsung-Han Wu、Yu Zhou、Liliane Momeni、RISHI HAZRA、Shuangrui Ding、Sagar Vaze、Francois Porcher、Feng Li、Siyuan Li、Aishwarya Kamath、Ho Kei Cheng、Piotr Dollar、Nikhila Ravi、Kate Saenko、Pengchuan Zhang 和 Christoph Feichtenhofer。《SAM 3：用概念分割万物》。发表于国际学习表征会议，第 2026 卷，第 138846–138923 页，2026年。网址：https://proceedings.iclr.cc/paper_files/paper/2026/hash/e0982cbc81401df3430ee1ff780dc7a2-Abstract-Conference.html。

[15] Meta。SAM 3.1 发行说明。官方软件发行文档，2026年3月。URL https://github.com/facebookresearch/sam3/blob/main/RELEASE_SAM3p1.md。于2026年3月27日发布。

[16] Shuai Bai、Yuxuan Cai、Ruizhe Chen、Keqin Chen、Xionghui Chen、Zesen Cheng、Lianghao Deng、Wei Ding、Chang Gao、Chunjiang Ge、Wenbin Ge、Zhifang Guo、Qidong Huang、Jie Huang、Fei Huang、Binyuan Hui、Shutong Jiang、Zhaohai Li、Mingsheng Li、Mei Li、Kaixin Li、Zicheng Lin、Junyang Lin、Xuejing Liu、Jiawei Liu、Chenglong Liu、Yang Liu、Dayiheng Liu、Shixuan Liu、Dunjie Lu、Ruilin Luo、Chenxu Lv、Rui Men、Lingchen Meng、Xuancheng Ren、Xingzhang Ren、Sibo Song、Yuchong Sun、Jun Tang、Jianhong Tu、Jianqiang Wan、Peng Wang、Pengfei Wang、Qiuyue Wang、Yuxuan Wang、Tianbao Xie、Yiheng Xu、Haiyang Xu、Jin Xu、Zhibo Yang、Mingkun Yang、Jianxin Yang、An Yang、Bowen Yu、Fei Zhang、Hang Zhang、Xi Zhang、Bo Zheng、Humen Zhong、Jingren Zhou、Fan Zhou、Jing Zhou、Yuanzhi Zhu 和 Ke Zhu。Qwen3-VL 技术报告。*arXiv* 预印本 *arXiv: 2511.21631*，2025年。

[17] Carolina Higuera、Akash Sharma、Chaithanya Krishna Bodduluri、Taosha Fan、Patrick Lancaster、Mrinal Kalakrishnan、Michael Kaess、Byron Boots、Mike Lambeta、Tingfan Wu 和 Mustafa Mukadam。Sparsh：面向基于视觉的触觉感知的自监督触觉表征。载于 Pulkit Agrawal、Oliver Kroemer 和 Wolfram Burgard 编，《第八届机器人学习会议论文集》，第 270 卷（机器学习研究论文集），第 885–915 页。PMLR，2025年。网址：https://proceedings.mlr.press/v270/higuera25a.html。

[18] Chengbo Yuan、Zicheng Zhang、Mingjie Zhou、Wendi Chen、Yi Wang、Zhuoyang Liu、Dantong Niu、Shuo Wang、Hui Zhang、Wenkang Zhang、Yingdong Hu、Yuanqing Gong、Wanli Xing、Chuan Wen、Cewu Lu、Kaifeng Zhang 和 Yang Gao。FTP-1：面向接触丰富型操作、适用于多种触觉传感器的通用基础触觉策略。*arXiv* 预印本 *arXiv:2606.13102*，2026年。网址：https://arxiv.org/abs/2606.13102。

[19] Hantao Zhou、Rui Yang、Longxiang Tang、Guanyi Qin、Runze Hu 和 Xiu Li。Gamma：借助评估专家混合模型实现通用图像评估。载于第 *33* 届 *ACM*国际多媒体会议论文集，第 8815–8824 页，2025年。doi：10.1145/3746027.3755699。URL https://doi.org/10.1145/3746027.3755699。

[20] Christoph Schuhmann。CLIP+MLP 美学评分预测器。https://github.com/christophschuhmann/improved-aesthetic-predictor，2022年。

[21] Yuning Du、Chenxia Li、Ruoyu Guo、Xiaoting Yin、Weiwei Liu、Jun Zhou、Yifan Bai、Zilin Yu、Yehua Yang、Qingqing Dang 和 Haoshuang Wang。PP-OCR：一种实用的超轻量级 OCR 系统。arXiv 预印本 arXiv:2009.09941，2020年。doi：10.48550/arXiv.2009.09941。URL https://arxiv.org/abs/2009.09941。[22] PaddleOCR 开发者。PP-OCRv4。PaddleOCR 版本 2.7.0。 https://github.com/PaddlePaddle/PaddleOCR/releases/tag/v2.7.0，2023。[23] Haoning Wu、Erli Zhang、Liang Liao、Chaofeng Chen、Jingwen Hou、Annan Wang、Wenxiu Sun、Qiong Yan 和 Weisi Lin。从美学和技术视角探索用户生成内容的视频质量评估。载于 IEEE/CVF 国际计算机视觉会议论文集，第 20144–20154 页，2023。URL https://openaccess.thecvf.com/content/ICCV2023/html/Wu_Exploring_Video_Quality_Assessment_on_User_Generated_Contents_from_Aesthetic_ICCV_2023_paper.html。[24] Qiuheng Wang、Yukai Shi、Jiarong Ou、Rui Chen、Ke Lin、Jiahao Wang、Boyuan Jiang、Haotian Yang、Mingwu Zheng、Xin Tao、Fei Yang、Pengfei Wan 和 Di Zhang。Koala-36M：一个大规模视频数据集，旨在提升细粒度条件与视频内容之间的一致性。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 8428–8437 页，2025年。URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Koala-36M_A_Large-scale_Video_Dataset_Improving_Consistency_between_Fine-grained_Conditions_CVPR_2025_paper.html。[25] Wim Meeussen。REP 105：移动平台的坐标系，2010年。URL https://reps.openrobotics.org/rep-0105/。[26] Kevin M. Lynch 和 Frank C. Park。《现代机器人学：力学、规划与控制》。剑桥大学出版社，2017年。doi：10.1017/9781316661239。URL https://modernrobotics.org/。[27] Yi Zhou、Connelly Barnes、Jingwan Lu、Jimei Yang 和 Hao Li。神经网络中旋转表示的连续性。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 5745–5753 页，2019年。URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zhou_On_the_Continuity_of_Rotation_Representations_in_Neural_Networks_CVPR_2019_paper.html。[28] Hugging Face。LeRobotDataset v3.0，2026年。URL https://huggingface.co/docs/lerobot/lerobot-dataset-v3。访问日期：2026年11月14日。[29] Gu Zhang、Qicheng Xu、Haozhe Zhang、Jianhan Ma、Long He、Yiming Bao、Zeyu Ping、Zhecheng Yuan、Chenhao Lu、Chengbo Yuan、Tianhai Liang、Xiaoyu Tian、Maanping Shao、Feihong Zhang、Mingyu Ding、Yang Gao、Hao Zhao、Hang Zhao 和 Huazhe Xu。UniDex：一套机器人基础工具，用于通过第一人称人类视频实现通用灵巧手控制。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 1841–1852 页，2026年。URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_UniDex_A_Robot_Foundation_Suite_for_Universal_Dexterous_Hand_Control_CVPR_2026_paper.html。[30] Jiacheng Fan、Zhiyue Zhao、Yiqian Zhang、Chao Chen、Peide Wang、Hengdi Zhang 和 Zhengxue Cheng。RoboPaint：从人类示范到任意机器人和任意视角。arXiv 预印本 arXiv:2602.05325，2026年。URL https://arxiv.org/abs/2602.05325。[31] aigc apps。VideoX-Fun：用于扩散变换器的视频生成流水线，2026年。URL https://github.com/aigc-apps/VideoX-Fun。

[32] Team Wan、Ang Wang、Baole Ai、Bin Wen、Chaojie Mao、Chen-Wei Xie、Di Chen、Feiwu Yu、Haiming Zhao、Jianxiao Yang、Jianyuan Zeng、Jiayu Wang、Jingfeng Zhang、Jingren Zhou、Jinkai Wang、Jixuan Chen、Kai Zhu、Kang Zhao、Keyu Yan、Lianghua Huang、Mengyang Feng、Ningyi Zhang、Pandeng Li、Pingyu Wu、Ruihang Chu、Ruili Feng、Shiwei Zhang、Siyang Sun、Tao Fang、Tianxing Wang、Tianyi Gui、Tingyu Weng、Tong Shen、Wei Lin、Wei Wang、Wei Wang、Wenmeng Zhou、Wente Wang、Wenting Shen、Wenyuan Yu、Xianzhong Shi、Xiaoming Huang、Xin Xu、Yan Kou、Yangyu Lv、Yifei Li、Yijing Liu、Yiming Wang、Yingya Zhang、Yitong Huang、Yong Li、You Wu、Yu Liu、Yulin Pan、Yun Zheng、Yuntao Hong、Yupeng Shi、Yutong Feng、Zeyinzi Jiang、Zhen Han、Zhi-Fan Wu 和 Ziyu Liu。《Wan：开放且先进的大规模视频生成模型》。*arXiv* 预印本 *arXiv:2503.20314*，2025年。

[33] Yufei Liu、Xixi Wang、Hao Li、Ganlong Zhao、Kaitong Cai、Chengkai Jin、Chunxiao Liu、Jianbo Liu、Siyuan Huang、Xingang Pan 和 Hongsheng Li。《ACE-Ego-Hand：重新利用视频扩散模型，实现遮挡鲁棒的自我中心三维手部运动恢复》。*arXiv* 预印本 *arXiv:2608.20308*，2026年。

[34] Javier Romero、Dimitrios Tzionas 和 Michael J. Black。《具身手部：手与身体的联合建模和捕捉》。*ACM* 图形学汇刊，36(6)：1–17，2017年。doi：10.1145/3130800. 3130883。URL https://doi.org/10.1145/3130800.3130883。

[35] Qixiu Li、Yu Deng、Yaobo Liang、Lin Luo、Lei Zhou、Chengtang Yao、Lingqi Zeng、Zhiyuan Feng、Huizhi Liang、Sicheng Xu、Yizhong Zhang、Xi Chen、Hao Chen、Lily Sun、Dong Chen、Jiaolong Yang 和 Baining Guo。《利用真实人类活动视频进行机器人操作的可扩展视觉-语言-动作模型预训练》。发表于*2026*年 *IEEE* 国际机器人与自动化会议（*ICRA*）。IEEE，2026年。URL https://microsoft.github.io/VITRA/。

[36] Hongzhe Bi、Zihao Zhou、Yihang Tang、Jingrui Pang、Shuhe Huang、Haitian Liu、Runqing Wang、Shuai Huang、Yichen Wang、Yiming Cheng、Ruowen Zhao、Zhenghua Li、Hengkai Tan、Xiaolong Liu、Jinhui Wan、Jiabao Liu、Min Zhao、Fan Bao 和 Jun Zhu。Motus2：用于灵巧操作的自进化通用世界模型。*arXiv* 预印本 *arXiv:2608.30237*，2026年。

[37] Chengbo Yuan、Rui Zhou、Mengzhen Liu、Yingdong Hu、Shengjie Wang、Li Yi、Chuan Wen、Shanghang Zhang 和 Yang Gao。MotionTrans：人类虚拟现实数据助力机器人操作策略的运动层级学习。载于*2026*年 *IEEE* 国际机器人与自动化会议（*ICRA*），第 8731–8738 页，2026年。doi：10.1109/icra57385.2026.11697199。URL https://doi.org/10.1109/icra57385.2026.11697199。

[38] Simar Kareer、Dhruv Patel、Ryan Punamiya、Pranay Mathur、Shuo Cheng、Chen Wang、Judy Hoffman 和 Danfei Xu。《EgoMimic：通过第一人称视角视频扩展模仿学习》。载于*2025*年 *IEEE* 国际机器人与自动化会议 *(ICRA)*，第 13226–13233 页。IEEE，2025年。URL https://ieeexplore.ieee.org/document/11127989.[39] Ri-Zhao Qiu、Shiqi Yang、Xuxin Cheng、Chaitanya Chawla、Jialong Li、Tairan He、Ge Yan、David J. Yoon、Ryan Hoque、Lars Paulsen、Ge Yang、Jian Zhang、Sha Yi、Guanya Shi 和 Xiaolong Wang。《人形机器人策略与人类策略》。载于 Joseph Lim、Shuran Song 和 Hae-Won Park 编，《style id='7'>第九届机器人学习会议论文集》，第 305 卷，机器学习研究论文集，第 2888–2906 页。PMLR，2025年。URL https://proceedings.mlr.press/v305/qiu25a.html.[40] 英伟达。Cosmos 3：面向物理人工智能的全模态世界模型。*arXiv* 预印本 *arXiv:2606.02800*，2026年。URL https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf.[41] 英伟达。Cosmos-Framework（GitHub：NVIDIA/cosmos-framework），2026年。URL https://github.com/NVIDIA/cosmos-framework。

[42] Samyam Rajbhandari、Jeff Rasley、Olatunji Ruwase 和 Yuxiong He。《ZeRO：面向万亿参数模型训练的内存优化》。载于 SC20：国际高性能计算、网络、存储与分析会议，第 1–16 页，2020 年。网址： https://doi.org/10.1109/sc41405.2020.00024。[43] PyTorch 贡献者。《开始使用全分片数据并行（FSDP2）》，2026 年。网址 https://docs.pytorch.org/tutorials/intermediate/FSDP_tutorial.html。文档，访问日期：2026年11月11日。[44] Jason Ansel、Edward Yang、Horace He、Natalia Gimelshein、Animesh Jain、Michael Voznesensky、Bin Bao、Peter Bell、David Berard、Evgeni Burovski、Geeta Chauhan、Anjali Chourdia、Will Constable、Alban Desmaison、Zachary DeVito、Elias Ellison、Will Feng、Jiong Gong、Michael Gschwind、Brian Hirsh、Sherlock Huang、Kshiteej Kalambarkar、Laurent Kirsch、Michael Lazos、Mario Lezcano、Yanbo Liang、Jason Liang、Yinghai Lu、C. K. Luk、Bert Maher、Yunjie Pan、Christian Puhrsch、Matthias Reso、Mark Saroufim、Marcos Yukio Siraichi、Helen Suk、Shunting Zhang、Michael Suo、Phil Tillet、Xu Zhao、Eikan Wang、Keren Zhou、Richard Zou、Xiaodong Wang、Ajit Mathews、William Wen、Gregory Chanan、Peng Wu 和 Soumith Chintala。《PyTorch 2：通过动态 Python 字节码转换和图编译加速机器学习》。载于第 29 届 ACM 编程语言与操作系统架构支持国际会议论文集，第 2 卷，第 929–947 页，2024年。网址 https://doi.org/10.1145/3620665.3640366。[45] Jay Shah、Ganesh Bikshandi、Ying Zhang、Vijay Thakkar、Pradeep Ramani 和 Tri Dao。《FlashAttention- 3：采用异步和低精度技术实现快速且准确的注意力》。载于神经信息处理系统进展，第 37 卷，第 68658–68685 页，2024年。doi：10.52202/079017-2193。网址 https://proceedings.neurips.cc/paper/2024/hash/7ede97c3e082c6df10a8d6103a2eebd2-Abstract-Conference.html。[46] Vijay Anand Korthikanti、Jared Casper、Sangkug Lym、Lawrence McAfee、Michael Andersch、Mo- hammad Shoeybi 和 Bryan Catanzaro。《减少大型 Transformer 模型中的激活重计算》。载于机器学习与系统会议论文集，第 5 卷，第 341–353 页，2023 年。网址 https://proceedings.mlsys.org/paper_files/paper/2023/hash/80083951326cf5b35e5100260d64ed81-Abstract-mlsys2023.html。[47] Xinyi Wan、Penghui Qi、Guangxing Huang、Chaoyi Ruan、Min Lin 和 Jialin Li。《重新审视大语言模型后训练中的参数服务器》。载于国际学习表征会议，2026年。网址https://openreview.net/forum?id=iIEEgI6WsF。[48] Haoyang Li、Fangcheng Fu、Sheng Lin、Hao Ge、Xuanyu Wang、Jiawen Niu、Jinbao Xue、Yangyu Tao、Di Wang、Jie Jiang 和 Bin Cui。《Hydraulis：通过共同设计并行策略和数据分配来平衡大型 Transformer 模型训练》。ACM 数据管理会刊，3(6)：1–30，2025年。网址 https://doi.org/10.1145/3769802。[49] Hao Ge、Junda Feng、Qi Huang、Fangcheng Fu、Xiaonan Nie、Lei Zuo、Haibin Lin、Bin Cui 和 Xin Liu。《ByteScale：在超过 12,000 个 GPU 上以 2048K 上下文长度高效扩展大语言模型训练》。arXiv:2502.21231，2025年。[50] Zili Zhang、Yinmin Zhong、Yimin Jiang、Hanpeng Hu、Jianjian Sun、Zheng Ge、Yibo Zhu、Daxin Jiang 和 Xin Jin。《DistTrain：通过面向多模态大语言模型的解耦式训练应对模型和数据异构性》。载于 2025年 ACM SIGCOMM 会议，2025年。网址https://doi.org/10.1145/3718958.3750472。[51] PyTorch 贡献者。《使用分布式检查点（DCP）进行异步保存》，2026年。网址 https://docs.pytorch.org/tutorials/recipes/distributed_async_checkpoint_recipe.html。文档，访问日期：2026年11月11日。

[52] Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann、Parker Schuh、Kensen Shi、Sasha Tsvyashchenko、Joshua Maynez、Abhishek Rao、Parker Barnes、Yi Tay、Noam Shazeer、Vinodkumar Prabhakaran、Emily Reif、Nan Du、Ben Hutchinson、Reiner Pope、James Bradbury、Jacob Austin、Michael Isard、Guy Gur-Ari、Pengcheng Yin、Toju Duke、Anselm Levskaya、Sanjay Ghemawat、Sunipa Dev、Henryk Michalewski、Xavier Garcia、Vedant Misra、Kevin Robinson、Liam Fedus、Denny Zhou、Daphne Ippolito、David Luan、Hyeontaek Lim、Barret Zoph、Alexander Spiridonov、Ryan Sepassi、David Dohan、Shivani Agrawal、Mark Omernick、Andrew M. Dai、Thanumalayan Sankaranarayana Pillai、Marie Pellat、Aitor Lewkowycz、Erica Moreira、Rewon Child、Oleksandr Polozov、Katherine Lee、Zongwei Zhou、Xuezhi Wang、Brennan Saeta、Mark Diaz、Orhan Firat、Michele Catasta、Jason Wei、Kathy Meier-Hellstern、Douglas Eck、Jeff Dean、Slav Petrov 和 Noah Fiedel。《PaLM：使用路径扩展语言建模》。 机器学习研究杂志，24(240):1–113，2023。URL http://jmlr.org/papers/v24/22-1144.html。

[53] FlagOS 贡献者。FlagOS/FlagGems：使用 Triton 语言实现的大语言模型算子库，2024年。网址 https://github.com/flagos-ai/FlagGems。FlagOS 开源社区 (https://flagos.io/)。

[54] Jonathan Ho 和 Tim Salimans。无分类器扩散引导。arXiv:2207.12598，2022年。

[55] Xuanlei Zhao、Xiaolong Jin、Kai Wang 和 Yang You。通过金字塔注意力广播实现实时视频生成。载于国际学习表征会议，第 2025 卷，第 3296–3319 页，2025年。网址 https://proceedings.iclr.cc/paper_files/paper/2025/hash/092c2d45005ea2db40fc24c470663416-Abstract-Conference.html。

[56] Feng Liu、Shiwei Zhang、Xiaofeng Wang、Yujie Wei、Haonan Qiu、Yuzhong Zhao、Yingya Zhang、Qixiang Ye 和 Fang Wan。时间步嵌入揭示：视频扩散模型是时候进行缓存了。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 7353–7363 页，2025年。网址 https://openaccess.thecvf.com/content/CVPR2025/html/Liu_Timestep_Embedding_Tells_Its_Time_to_Cache_for_Video_Diffusion_CVPR_2025_paper.html。

[57] Jiarui Yuan、Yuxin Zuo、Xiaoyu Huang、Wenbo Zhang、Kwanyoung Park、Junlin Yang、Tianxiang Zhang、Ruoqu Chen、Yutao Ouyang、Bingxiang He、Xingtai Lv、Hongjin Wu、Jiajun Liu、Jiahao Zhao、Tianwei Luo、Haozhan Li、Cheng Qian、Jiaxin Ge、Huan ang Gao、Kaisen Yang、Jiaqi Deng、Qixin Xu、Siqiao Huang、Bowen Zhou、Mengdi Xu、Zhiyuan Liu、Pieter Abbeel 和 Ning Ding。《让机器人使用智能体普及化：基于开放模型的统一测试框架》。*arXiv* 预印本 *arXiv: 2611.01560*，2026年。

[58] Gemini Robotics Team。《Gemini Robotics 1.5：借助先进的具身推理、思考和运动迁移，推动通用机器人发展前沿》。技术报告，Google DeepMind，2025年。URL https://storage.googleapis.com/deepmind-media/gemini-robotics/Gemini-Robotics-1-5-Tech-Report.pdf。

[59] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。载于《国际学习表征会议》，2019年。URL https://openreview.net/forum?id=Bkg6RiCqY7。[60] Yanjiang Guo、Haodong Yan、Zhide Zhong、Zhongru Zhang、Qingyuan Yang、Qingzhou Lu、Xiaoyu Chen、Yen-Jen Wang、Shuying Deng、Chenghan Yang、Puzhen Yuan、Chenxin Liu、Tun Ban、Xiang Zhu、Yichen Liu、Kun Feng、Haoang Li 和 Jianyu Chen。视频预测策略 2：预测得更好，行动得更好。《arXiv 预印本 arXiv:2610.10270》，2026年。URL https://arxiv.org/abs/2610.10270。[61] Xiaomi Robotics Team。Xiaomi-Robotics-1：利用超过 100K 小时的真实世界轨迹扩展视觉-语言-动作模型。《arXiv 预印本 arXiv:2607.15330》，2026年。URL https://arxiv.org/abs/2607.15330。

[62] Kevin Black、Noah Brown、Danny Driess、Adnan Esmail、Michael Robert Equi、Chelsea Finn、Niccolo Fusai、Lachy Groom、Karol Hausman、Brian Ichter、Szymon Jakubczak、Tim Jones、Liyiming Ke、Sergey Levine、Adrian Li-Bell、Mohith Mothukuri、Suraj Nair、Karl Pertsch、Lucy Xiaoyang Shi、Laura Smith、James Tanner、Quan Vuong、Anna Walling、Haohuan Wang 和 Ury Zhilinsky。 π_0：用于通用机器人控制的视觉-语言-动作流模型。载于《机器人学：科学与系统》论文集，美国加利福尼亚州洛杉矶，2025年。doi：10.15607/RSS.2025.XXI.010。URL https://www.roboticsproceedings.org/rss21/p010.html。[63] Black Forest Labs。FLUX 3 Action：用于机器人控制的 7B 世界动作模型。技术博客，2026年。URL https://bfl.ai/models/flux-3-action。[64] Diqi He、Jinbang Guo、Yijun Shen、Zhiwei Yu、Huan Yang、Dingwen Zhang、Junwei Han 和 Zhongyuan Wang。MOST：面向长时域移动操作的物体状态与任务进度记忆。《arXiv 预印本 arXiv:2611.00734》，2026年。URL https://arxiv.org/abs/2611.00734。[65] Chloe Su、Jiahan Li、Zhenting Qi、Chaoran Cheng、Haoyu Chen、Sejoon Jun、Weirui Ye 和 Yilun Du。面向长时域具身智能体的多模态文件系统记忆。《arXiv 预印本 arXiv:2610.22686》，2026年。URL https://arxiv.org/abs/2610.22686。[66] Tianyidan Xie、Shenyi Wang、Qiang Tang、Mingjie Wang、Zhicheng Qiu、Xuanfu Li、Zhan Xu、Jian Yang、Lanjun Wang 和 Zili Yi。面向具身智能体高效长时域空间记忆的语言轨迹编码。《arXiv 预印本 arXiv:2609.04802》，2026年。URL https://arxiv.org/abs/2609.04802。已被 NeurIPS 2026 接收。[67] Jianxin Bi、Kevin Yuchen Ma、Ce Hao、Mike Zheng Shou 和 Harold Soh。VLA-Touch：通过双层触觉反馈增强视觉-语言-动作模型。《IEEE机器人与自动化快报》，11(7)：8487–8494，2026年。doi：10.1109/LRA.2026.3692345。URL https://doi.org/10.1109/LRA.2026.3692345。[68] Jianyi Zhou、Feiyang Hong、Yunhao Li、Yicheng Zhao、Yongjue Cen、Zirui Liu、Jiakang Huang、Zirui Chen、Ruiyang Zhang、Weizhuo Zhu、Xuhua Song 和 Shuo Yang。TouchWorld：用于灵巧操作的预测式与反应式触觉基础模型。《arXiv 预印本 arXiv:2607.07287》，2026年。[69] Yijun Shen、Diqi He、Jinbang Guo、Wanyue Zhang、Jiaming Mi、Huan Yang、Zhongyuan Wang、Zhiwei Yu 和 Kai Zhang。CausGround：用于自适应具身规划的因果交互定位。《arXiv 预印本 arXiv:2610.29702》，2026年。URL https://arxiv.org/abs/2610.29702。[70] Wenlong Huang、Fei Xia、Ted Xiao、Harris Chan、Jacky Liang、Pete Florence、Andy Zeng、Jonathan Tompson、Igor Mordatch、Yevgen Chebotar、Pierre Sermanet、Tomas Jackson、Noah Brown、Linda Luu、Sergey Levine、Karol Hausman 和 brian ichter。Inner Monologue：通过使用语言模型进行规划实现具身推理。载于 Karen Liu、Dana Kulic 和 Jeff Ichnowski 编，《第六届机器人学习会议论文集》，《机器学习研究论文集》第205卷，第1769–1782页。PMLR，2023。URL https://proceedings.mlr.press/v205/huang23c.html。[71] Christopher Agia、Rohan Sinha、Jingyun Yang、Ziang Cao、Rika Antonova、Marco Pavone 和 Jeannette Bohg。解构生成式策略的失效模式：一致性与进度的运行时监控。载于 Pulkit Agrawal、Oliver Kroemer 和 Wolfram Burgard 编，《第八届机器人学习会议论文集》，《机器学习研究论文集》第270卷，第689–723页。PMLR，2025年。URL https://proceedings.mlr.press/v270/agia25a.html。[72] Zeyi Liu、Arpit Bahety 和 Shuran Song。REFLECT：总结机器人经验，用于失效解释与纠正。载于 Jie Tan、Marc Toussaint 和 Kourosh Darvish 编，《第七届机器人学习会议论文集》，《机器学习研究论文集》第229卷，第3468–3484页。PMLR，2023。URL https://proceedings.mlr.press/v229/liu23g.html。

[73] Xin Ding、Liang Mi、Mingzhe Huang、Zixuan Wang、Chao Zhang、Zixu Hao、Fu Chen、Xiangyu Li、Yikai Zheng、Yaoyu Guo、Weijun Wang、Kun Li、Hao Wu、Yunxin Liu 和 Ting Cao。Zetta ζ：用于自进化物理智能的高效闭环具身智能框架。 *arXiv* 预印本 *arXiv:2608.16590*，2026年。[74] Guanzhi Wang、Yuqi Xie、Yunfan Jiang、Ajay Mandlekar、Chaowei Xiao、Yuke Zhu、Linxi Fan 和 Anima Anandkumar。Voyager：一种借助大语言模型的开放式具身智能体。《机器学习研究汇刊》，2024 年。网址 https://voyager.minedojo.org/assets/documents/voyager.pdf。[75] LangChain。LangGraph：图 API 概览。https://www.langchain.com/langgraph，2024年。访问日期：2026年11月16日。[76] Joon Sung Park、Joseph O'Brien、Carrie Jun Cai、Meredith Ringel Morris、Percy Liang 和 Michael S. Bernstein。生成式智能体：人类行为的交互式拟像。载于第 *36* 届 *ACM* 用户界面软件与技术年度研讨会论文集，第 1–22 页，2023。doi：10.1145/3586183.3606763。网址 https://doi.org/10.1145/3586183.3606763。[77] Gyeong-In Yu、Joo Seong Jeong、Geon-Woo Kim、Soojeong Kim 和 Byung-Gon Chun。Orca：面向基于 Transformer 的生成模型的分布式服务系统。载于第 *16* 届 *USENIX* 操作系统设计与实现研讨会（*OSDI 22*），第 521–538 页。USENIX 协会，2022年。网址 https://www.usenix.org/conference/osdi22/presentation/yu。[78] Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph Gonzalez、Hao Zhang 和 Ion Stoica。借助 PagedAttention 实现大语言模型服务的高效内存管理。载于第 *29* 届操作系统原理研讨会论文集，第 611–626 页，2023。doi：10.1145/3600006.3613165。网址 https://doi.org/10.1145/3600006.3613165。[79] Baijun Chen、Weijie Wan、Tianxing Chen、Xianda Guo、Congsheng Xu、Yuanyang Qi、Haojie Zhang、Longyan Wu、Tianling Xu、Zixuan Li、Yizhe Wu、Rui Li、Xiaokang Yang、Ping Luo、Wei Sui 和 Yao Mu。UniVTAC：用于视觉触觉操作数据生成、学习和基准测试的统一仿真平台。*arXiv* 预印本 *arXiv:2602.10093*，2026年。doi：10.48550/arXiv.2602.10093。网址https://arxiv.org/abs/2602.10093。已被机器人学习会议（CoRL 2026）接收。[80] Zhenjie Yang、Yideng Zhang、Dongjie Zhang、Chenyu Jiang、Xianshuai Liu、Yufeng Li、Zuhao Ge、Xingyu Jiao、Zheng Zhang、Kaiyu He、He Wang、Yuwen Zhong、Yi Deng、Muyun Jiang、Xianliang Huang、Haisheng Su、Donghang Zhang、Jian Zhang、Xue Yang、Hongyang Li、Zuxuan Wu、Yu-Gang Jiang、Xiaosong Jia 和 Junchi Yan。Bench2Dex：跨灵巧手的视觉触觉双手灵巧操作基准测试。*arXiv* 预印本 *arXiv:2609.15726*，2026年。doi：10.48550/arXiv.2609.15726。网址 https://arxiv.org/abs/2609.15726。[81] Gemma 团队、Sherif El Abd、Vaibhav Aggarwal、Robin Algayres 等。Gemma 4 技术报告。*arXiv* 预印本 *arXiv: 2607.02770*，2026年。[82] Yifu Yuan、Yaoting Huang、Xianze Yao、Yutong Li、Shuoheng Zhang、Linqi Han、Pengyi Li、Jiangeng Sun、Wenting Jia、Zhao Zhang、Yuhao Liu、Ruihao Liao、Yucheng Hu、Qiyu Wu、Yuxiao Li、Zibin Dong、Fei Ni、Yan Zheng、Shuyang Gu、Yi Ma、Hongyao Tang、Han Hu 和 Jianye Hao。Embodied-R1.5：通过具身基础模型演进物理智能。*arXiv* 预印本 *arXiv: 2606.11324*，2026年。

[83] Xiaoshuai Hao、Lei Zhou、Zhijian Huang、Zhiwen Hou、Yingbo Tang、Lingfeng Zhang、Guang Li、Zheng Lu、Shuhuai Ren、Xianhui Meng、Yuchen Zhang、Jing Wu、Jinghui Lu、Chenxu Dang、Jiayi Guan、Jianhua Wu、Zhiyi Hou、Hanbing Li、Shumeng Xia、Mingliang Zhou、Yinan Zheng、Zihao Yue、Shuhao Gu、Hao Tian、Yuannan Shen、Jianwei Cui、Wen Zhang、Shaoqing Xu、Bing Wang、Haiyang Sun、Zeyu Zhu、Yuncheng Jiang、Zibin Guo、Chuhong Gong、Chaofan Zhang、Wenbo Ding、Kun Ma、Guang Chen、Rui Cai、Diyun Xiang、Heng Qu、FuliLuo、Hangjun Ye 和 Long Chen。MiMo-Embodied：X-具身基础模型技术报告。*arXiv* 预印本 *arXiv: 2511.16518*，2025年。

[84] Huajie Tan、Enshen Zhou、Zhiyu Li、Yijie Xu、Yuheng Ji、Xiansheng Chen、Cheng Chi、Pengwei Wang、Huizhu Jia、Yulong Ao、Mingyu Cao、Sixiang Chen、Zhe Li、Mengzhen Liu、Zixiao Wang、Shanyu Rong、Yaoxu Lyu、Zhongxia Zhao、Peterson Co、Yibo Li、Yi Han、Shaoxuan Xie、Guocai Yao、Songjing Wang、Leiduo Zhang、Xi Yang、Yance Jiao、Donghai Shi、Kunchang Xie、Shaokai Nie、Chunlei Men、Yonghua Lin、Zhongyuan Wang、Tiejun Huang 和 Shanghang Zhang。《RoboBrain 2.5：洞察深度，心怀时间》。*arXiv* 预印本 *arXiv: 2601.14352*，2026年。

[85] Ronghao Dang、Jiayan Guo、Bohan Hou、Sicong Leng、Kehan Li、Xin Li、Jiangpin Liu、Yunxuan Mao、Zhikai Wang、Yuqian Yuan、Minghao Zhu、Xiao Lin、Yang Bai、Qian Jiang、Yaxi Zhao、Minghua Zeng、Junlong Gao、Yuming Jiang、Jun Cen、Siteng Huang、Liuyi Wang、Wenqiao Zhang、Chengju Liu、Jianfei Yang、Shijian Lu 和 Deli Zhao。《RynnBrain：开放式具身基础模型》。*arXiv* 预印本 *arXiv: 2602.14979*，2026年。

[86] Xingyu Fu、Yushi Hu、Bang-Zheng Li、Yu Feng、Haoyu Wang、Xudong Lin、Dan Roth、Noah A. Smith、Wei-Chiu Ma 和 Ranjay Krishna。《BLINK：多模态大语言模型能看见，却无法感知》。欧洲计算机视觉会议，2024年。doi：10.48550/arXiv.2404.12390。[87] Aaron Foss、Chloe Evans、Sasha Mitts、Koustuv Sinha、Ammar Rizvi 和 Justine T. Kao。《CausalVQA：面向视频模型的物理基础因果推理基准测试。*arXiv* 预印本 *arXiv: 2506.09943*，2025年。[88] Lucas Beyer、Andreas Steiner、André Susano Pinto、Alexander Kolesnikov、Xiao Wang、Daniel Salz、Maxim Neumann、Ibrahim Alabdulmohsin、Michael Tschannen、Emanuele Bugliarello、Thomas Un- terthiner、Daniel Keysers、Skanda Koppula、Fangyu Liu、Adam Grycner、Alexey Gritsenko、Neil Houlsby、Manoj Kumar、Keran Rong、Julian Eisenschlos、Rishabh Kabra、Matthias Bauer、Matko Bošnjak、Xi Chen、Matthias Minderer、Paul Voigtlaender、Ioana Bica、Ivana Balazevic、Joan Puigcerver、Pinelopi Papalam- pidi、Olivier Henaff、Xi Xiong、Radu Soricut、Jeremiah Harmsen 和 Xiaohua Zhai。《PaliGemma：用于迁移的多功能 3B 视觉语言模型。*arXiv* 预印本 *arXiv: 2407.07726*，2024年。[89] xAI。《Grok-1.5 Vision Preview》。 https://x.ai/news/grok-1.5v，2024年4月。介绍了 RealWorldQA 基准测试。[90] Licheng Yu、Patrick Poirson、Shan Yang、Alexander C. Berg 和 Tamara L. Berg。《指代表达中的上下文建模》。 *arXiv* 预印本 *arXiv: 1608.00272*，2016年。[91] Penghao Wu 和 Saining Xie。《V*：将引导式视觉搜索作为多模态大语言模型的核心机制。*arXiv* 预印本 *arXiv: 2312.14135*，2023。[92] Shengbang Tong、Ellis Brown、Penghao Wu、Sanghyun Woo、Manoj Middepogu、Sai C. Akula、Jihan Yang、Shusheng Yang、Adithya Iyer、Xichen Pan、Austin Wang、Rob Fergus、Yann LeCun 和 Saining Xie。《Cambrian-1：对多模态大语言模型进行完全开放、以视觉为中心的探索》。载于 神经信息处理系统进展，第37卷，第87310–87356页，2024年。doi：10.52202/079017-2771。网址 https://proceedings.neurips.cc/paper_files/paper/2024/hash/9ee3a664ccfeabc0da16ac6f1f1cfe59-Abstract-Conference.html。[93] Wufei Ma、Haoyu Chen、Guofeng Zhang、Yu-Cheng Chou、Jieneng Chen、Celso de Melo 和 Alan Yuille。《3DSRBench：综合三维空间推理基准测试》。载于 *IEEE/CVF* 国际计算机视觉会议（*ICCV*）论文集，第6924–6934页，2025年10月。

[94] Mengfei Du、Binhao Wu、Zejun Li、Xuanjing Huang 和 Zhongyu Wei。《EmbSpatial-bench：利用大型视觉语言模型对具身任务的空间理解进行基准测试》。Lun-Wei Ku、Andre Martins 和 Vivek Srikumar 编，载于 第*62*届计算语言学协会年会论文集（第*2*卷：短论文），第346–355页，泰国曼谷，2024年8月。计算语言学协会。doi：10.18653/v1/2024.acl-short.33。网址 https://aclanthology.org/2024.acl-short.33/。

[95] Qineng Wang、Baiqiao Yin、Pingyue Zhang、Jianshu Zhang、Kangrui Wang、Zihan Wang、Jieyu Zhang、Keshigeyan Chandrasegaran、Han Liu、Ranjay Krishna、Saining Xie、Jiajun Wu、Li Fei-Fei 和 Manling Li。基于有限视角的空间心理建模。发表于国际学习表征会议，第 2026 卷，第 82559–82622 页，2026年。URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/858de961874816f894285ec02e50c7fe-Abstract-Conference.html。

[96] Sihan Yang、Runsen Xu、Yiman Xie、Sizhe Yang、Mo Li、Jingli Lin、Chenming Zhu、Xiaochen Chen、Haodong Duan、XiangyuYue、Dahua Lin、TaiWang 和 Jiangmiao Pang。MMSI-Bench：多图像空间智能基准测试。*arXiv* 预印本 *arXiv: 2505.23764*，2025年。

[97] Jingli Lin, Runsen Xu, Shaohao Zhu, Sihan Yang, Peizhou Cao, Yunlong Ran, Miao Hu, Chenming Zhu, Yiman Xie, Yilin Long, Wenbo Hu, Dahua Lin, Tai Wang 和 Jiangmiao Pang。《MMSI-Video-Bench：基于视频的空间智能综合基准测试》。*arXiv* 预印本 *arXiv: 2512.10863*，2025年。

[98] Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang, Yanbiao Ma, Yunfeng Diao, Ziyu Jia, Wenbo Ding, Hangjun Ye 和 Long Chen。《RoboAfford++：一个用于机器人操作与导航中多模态可供性学习的生成式人工智能增强数据集》。*arXiv* 预印本 *arXiv: 2511.12436*，2025年。

[99] Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Chunhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu 和 Li Zhang。《从平面世界到空间：教视觉语言模型在三维中进行感知与推理》。载于神经信息处理系统进展，第38卷，2025年。doi：10.52202/085713-5180。URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e3a0db7c0a191854c176af1d20cdec80-Abstract-Datasets and Benchmarks Track.html。

[100] Jieru Lin、Zhiwei Yu 和 Börje F. Karlsson。《SWITCH：长时程具身场景中有形界面的建模与操作基准测试》。*arXiv* 预印本 *arXiv: 2511.17649*，2025年。

[101] Jihan Yang、Shusheng Yang、Anjali W. Gupta、Rilyn Han、Li Fei-Fei 和 Saining Xie。《空间思维：多模态大语言模型如何观察、记忆和回忆空间》。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第 10632–10643 页，2025年6月。

[102] Tianxing Chen、Zanxin Chen、Baijun Chen、Zijian Cai、Yibin Liu、Zixuan Li、Qiwei Liang、Xianliang Lin、Yiheng Ge、Zhenyu Gu、Weiliang Deng、Yubin Guo、Tian Nian、Xuanbing Xie、Qiangyu Chen、Kailun Su、Tianling Xu、Guodong Liu、Mengkang Hu、Huan-Ang Gao、Kaixuan Wang、Zhixuan Liang、Yusen Qin、Xiaokang Yang、Ping Luo 和 Yao Mu。《RoboTwin 2.0：一种可扩展的数据生成器和基准测试，借助强域随机化实现稳健的双臂机器人操作》。载于 Tong Zhang、Miroslav Dudik、Martin Jaggi、Alekh Agarwal、Sharon Li、Dale Schuurmans、Jerry Zhu、Felix Berkenkamp、Hanze Dong 和 Alberto Bietti（编），第*43*届国际机器学习会议论文集，第 306 卷，机器学习研究论文集，第 13673–13699 页。PMLR，2026年。网址： https://proceedings.mlr.press/v306/chen26j.html。

[103] Yuzhi Chen、Ronghan Chen、Dongjie Huo、Yandan Yang、Dekang Qi、Haoyun Liu、Tong Lin、Shuang Zeng、Junjin Xiao、Xinyuan Chang、Feng Xiong、Xing Wei、Zhiheng Ma 和 Mu Xu。ABot-PhysWorld：通过物理对齐实现机器人操作的交互式世界基础模型。*arXiv* 预印本 *arXiv:2603.23376*，2026年。网址 https://arxiv.org/abs/2603.23376。

[104] Yixiang Chen、Peiyan Li、Yuan Xu、Qisen Ma、Jiabing Yang、Kai Wang、Jianhua Yang、Dong An、He Guan、Gaoteng Liu、Jianlou Si、Jun Huang、Jing Liu、Nianfeng Liu、Yan Huang 和 Liang Wang。FlowWAM：将光流作为世界动作模型的统一动作表征。arXiv 预印本 arXiv:2607.13017，2026年。网址 https://arxiv.org/abs/2607.13017.[105] BWM 团队。BWM：一种低成本、高保真的机器人学习世界模拟器。arXiv 预印本 arXiv:2607.29302，2026年。网址 https://arxiv.org/abs/2607.29302.[106] Xintao Wang、Liangbin Xie、Chao Dong 和 Ying Shan。Real-ESRGAN：使用纯合成数据训练真实世界盲超分辨率。在 IEEE/CVF 国际计算机视觉会议论文集，第1905–1914页，2021年。网址 https://openaccess.thecvf.com/content/ICCV2021W/AIM/html/Wang_Real-ESRGAN_Training_Real-World_Blind_Super-Resolution_With_Pure_Synthetic_Data_ICCVW_2021_paper.html.[107] DreamX 团队、Rui Chen、Xiangxiang Chu、Geng Li、Jifan Li、Qingfeng Shi、Datao Tang、Jing Tang、Jun Wang 和 Pengfei Zhang。DreamX-Phi 1.0：面向机器人操作的动作条件视频世界模型。arXiv 预印本 arXiv:2608.13489，2026年。网址 https://arxiv.org/abs/2608.13489.[108] Chensheng Peng、Wenhao Ding、Ran Tian、Zewei Zhou、Jef Packer、Maximilian Igl、Peter Karkus、Yan Wang、Masayoshi Tomizuka、Boris Ivanovic、Marco Pavone 和 Yuxiao Chen。CtrlWAM：具有对齐意图和前瞻能力的可控世界动作模型。arXiv 预印本 arXiv:2610.00859，2026年。网址 https://arxiv.org/abs/2610.00859.[109] Tianqi Liu、Hanfei Zhao、Qifeng Liang、Hui En Pang、Jiarui Xu、Sayang Mu、Penghao Qiu、YingSheng Zhang、Jun Wang、Ziang Cao、Mingze Sun、Yukang Cao、Fangzhou Hong、Zhaoxi Chen 和 Ziwei Liu。Ropedia W0：利用人类经验扩展跨具身形态的移动操作世界模型。arXiv 预印本 arXiv:2610.26855，2026年。网址 https://arxiv.org/abs/2610.26855.[110] Wenliang Zhao、Lujia Bai、Yongming Rao、Jie Zhou 和 Jiwen Lu。UniPC：用于扩散模型快速采样的统一预测器-校正器框架。在《神经信息处理系统进展》第36卷，第49842–49869页，2023。doi：10.52202/075280-2170。网址 https://proceedings.neurips.cc/paper_files/paper/2023/hash/9c2aa1e456ea543997f6927295196381-Abstract-Conference.html.[111] Awomo-WAM 团队。Awomo-0.5 技术报告。技术报告，2026年。网址 https://github.com/Awomo-WestlakeDI/Awomo-0.5/blob/main/Awomo_0_5.pdf.[112] Renbin Shu、Jing Yang、Xiangyuan You、Guolin Qiu、Yufei Xia、Yiming Zhu、Nanjun Dai、Fei Zhang、Feng Wu 和 Keke Tang。Axiom-0：视觉-语言-动作策略的分层自我细化。arXiv 预印本 arXiv:2610.14155，2026年。网址 https://arxiv.org/abs/2610.14155.[113] Yixian Zhang、Huanming Zhang、Feng Gao、Xiao Li、Zhihao Liu、Chunyang Zhu、Jiaxing Qiu、Yuchen Yan、Jiyuan Liu、Wenhao Tang、Zhengru Fang、Yi Nie、Changxu Wei、Yu Wang、Wenbo Ding 和 Chao Yu。Harness VLA：通过记忆引导智能体引导冻结的视觉-语言-动作模型，实现可靠的操作原语。arXiv 预印本 arXiv:2607.08448，2026年。网址 https://arxiv.org/abs/2607.08448.[114] Guoning Wu、Zijian Cai、Yuhang Zhang、Ying Liu 和 Yongbin Zheng。World Dreamer v5-lite：由一个梦境生成器驱动的多机器人世界动作模型。arXiv 预印本 arXiv:2610.37518，2026年。网址 https://arxiv.org/abs/2610.37518.[115] 美团机器人团队。Meituan-Robotics-0：面向桌面操作的视觉-语言-动作基础模型，2026年。网址 https://arxiv.org/abs/2611.08577。

[116] Ronghan Chen、Yandan Yang、Zuojin Tang、Dongjie Huo、Tong Lin、Haoning Wu、Haoyun Liu、Yuzhi Chen、Lulu Zheng、Botai Yuan、Tianlun Li、Mingxin Wang、Dekang Qi、Bin Hu、Wei Mei、Yuze Xuan、Haolong Yang、Yanqing Zhu、Mu Xu、Zhiheng Ma 和 Xinyuan Chang。ABot-M0.5：统一移动与操作世界动作模型。*arXiv* 预印本 *arXiv:2607.00678*，2026年。网址 https://arxiv.org/abs/2607.00678。

[117]Kevin Black、Noah Brown、James Darpinian、Karan Dhabalia、Danny Driess、Adnan Esmail、Michael Robert Equi、Chelsea Finn、Niccolo Fusai、Manuel Y. Galliker、Dibya Ghosh、Lachy Groom、Karol Hausman、Brian Ichter、Szymon Jakubczak、Tim Jones、Liyiming Ke、Devin LeBlanc、Sergey Levine、Adrian Li-Bell、Mohith Mothukuri、Suraj Nair、Karl Pertsch、Allen Z. Ren、Lucy Xiaoyang Shi、Laura Smith、Jost Tobias Springenberg、Kyle Stachowicz、James Tanner、Quan Vuong、Homer Walke、Anna Walling、Haohuan Wang、Lili Yu 和 Ury Zhilinsky。$\pi_{0.5}$：一款具备开放世界泛化能力的视觉-语言-动作模型。载于 Joseph Lim、Shuran Song 和 Hae-Won Park（编），第九届机器人学习会议论文集，第305卷，机器学习研究论文集，第17–40页。PMLR，2025年。网址 https://proceedings.mlr.press/v305/black25a.html。

[118] Yinhao Li, Weixin Mao, Zihan Lan, Jikun Rong, Qirui Hu, Yiming Zhang, Weipeng Deng, Bo wen Shen, Minzhao Zhu, Yiming Mao, Yan Yang, Chenguang Cui, HongyuanChen, Xu Huang, Zheyi Zhao, Pinxi Shen, Bozhen He, Zhen Fu, Yifan Wang, Zexin Zhang, Ang Gao, Haoyu Chen, Chengqi Shi, Wenda Wu 和 Hua Chen。FluxVLA Engine：面向具身智能的一站式视觉-语言-动作工程平台，2026年。URL https://arxiv.org/abs/2609.17210。

[119] Xingyu Miao, Zizun Li, Baole Fang, Kaiwen Song, Tenghui Wang, Hanxue Zhang, Yating Wang, Xudong Li, Yuping He, Xueyuan Wei, Chao Gao, Xijie Yang, Yingxiang Xu, Kerui Ren, Wenqi Guo, Jianjun Zhou, Xinzhe Wang, Weiguang Zhao, Ni Yang, Zetao Cai, Yufei Xue, Hengjie Li, Zeyu He, Yuanzhen Zhou, Rong Fu, Jianyang Zhang, Siwei Cui, Fuxian Huang, Yunsong Zhou, Xing Gao, Yifei Yao, Qiaojun Yu, Kailin Li, Ming Zhou, Mu Huang, Xinyue Li, Wenze Cui, Bingqi Jiang, Xueyue Zhu, Junting Dong, Haoyu Guo, Tao Lu, Mulin Yu, Bowen Zhou, Bin Zhao, Tianfan Xue, Weinan Zhang 和 Chunhua Shen。InternW0-Δ：世界动作模型，通过 20K+小时开放数据，将预测动力学与动作连接起来。*arXiv*预印本 *arXiv:2609.31394*，2026年。URL https://arxiv.org/abs/2609.31394。

[120] LiberAI 团队。Liber-0：从 220 万小时的人类经验中涌现的多步操作零样本能力，2026年。网址 https://www.liberai.com/blog/liber0/。

[121] Wei He、Hengtao Li、Chenfeng Wang、Zhongrui Yu、Xuhan Zhu、Maokui He、Zide Liu、Xiyue Zhang、Xianwei Mao、Chunpeng Zhou、Jia Shi、Yanze Xin、Jingwen Li、Jingxie Zheng、Sijie Zeng、Fan Lu、Zeyu Zhang、Shuai Guo、Hengxuan Zhang、Pengfei Yu、Jia Shi、Yu Liu、Kun Zhan 和 Yan Xie。ME-Brain-1.0：面向不断演进的具身智能的记忆、认知与动作。*arXiv* 预印本 *arXiv: 2609.24271*，2026年。

[122] Yicheng Liu、Zibin Dong、Baijun Ye、Tianyuan Yuan、Tao Jiang 等。G0.5：用于机器人推理与动作的单一自回归流。*arXiv* 预印本 *arXiv:2608.11739*，2026年。网址 https://arxiv.org/abs/2608.11739。

[123] Haoran Wen、Wenfu Wang、Kunsong Shi、Jingke Wang、Wancheng Feng、Yiren Zhang、Yueran Zhao、Xuancheng Zhang、Nanfei Ye、Xingru Chen、Zhaohong Sun、Chengmin Yang、Zikang Yu、Penghao Bi、Jia Shi、Yu Liu、Kun Zhan 和 Yan Xie。MachEmbodied-U0：面向具身智能的统一理解与生成模型。*arXiv* 预印本 *arXiv:2609.25627*，2026年。URL https://arxiv.org/abs/2609.25627。

[124] He Zhang, Lingzhu Xiang, Haitao Lin, Zeyu Huang, Minghui Wang, Dingyan Zhong, Yubo Dong, Yihao Wu, Yongming Rao, Dongsheng Zhang, Wanjia He, Ling Chen, Kai Huang, Jiahao Chen, Sichang Su, Xumin Yu, Ziyi Wang, Chengwei Zhu, Xiao Teng, Yuchun Guo, Yufeng Zhang, Yuandong Liu, Rui Wang, Zisheng Lu, Han Hu, and Zhengyou Zhang. Hy-Embodied-0.5-VLA：从视觉-语言-动作模型到真实世界机器人学习技术栈。*arXiv* 预印本 *arXiv: 2606.14409*，2026年。

[125] Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, Min Cao, Peng Li, Qiuping Deng, Wenjun Mei, Xiaofeng Wang, Xinze Chen, Xinyu Zhou, Yang Wang, Yifan Chang, Yifan Li, Yukun Zhou, Yun Ye, Zhichao Liu, and Zheng Zhu. GigaWorld-Policy：一种高效的以动作为中心的世界-动作模型。*arXiv* 预印本 *arXiv:2603.17240*，2026年。doi：10.48550/arXiv.2603.17240。URL https://arxiv.org/abs/2603.17240。[126] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Luyao Zhang, Mingrui Yu, Zelin Gao, Nan Xue, Boyu Zhou, Xing Zhu, Mingyu Ding, Yujun Shen, and Yinghao Xu. 用于机器人控制的因果世界建模。发表于《机器人学：科学与系统》论文集，澳大利亚悉尼，2026年。doi：10.156 07/RSS.2026.XXII.016。URL https://www.roboticsproceedings.org/rss22/p016.html。[127] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. 使用低成本硬件学习精细双手操作。发表于《机器人学：科学与系统》论文集，大韩民国大邱，2023。doi：10.15607/RSS.2023.XIX.016。URL https://www.roboticsproceedings.org/rss19/p016.html。[128] Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM：世界动作模型是否需要测试时的未来想象？*arXiv* 预印本 *arXiv:2603.16666*，2026年。doi：10.48550/arXiv.2603.16666。URL https://arxiv.org/abs/2603.16666。

[129] Rui Cai、Jun Guo、Xinze He、Piaopiao Jin、Jie Li、Bingxuan Lin、Futeng Liu、Wei Liu、Fei Ma、Kun Ma、Feng Qiu、Heng Qu、Yifei Su、Qiao Sun、Dong Wang、Donghao Wang、Yunhong Wang、Rujie Wu、Diyun Xiang、Yu Yang、Hangjun Ye、Yuan Zhang 和 Quanyun Zhou。Xiaomi-Robotics-0：一种开源的视觉-语言-动作模型，可实时执行。*arXiv* 预印本 *arXiv:2602.12684*，2026年。URLhttps://arxiv.org/abs/2602.12684。

[130] Jinhui Ye、Ning Gao、Senqiao Yang、Jinliang Zheng、Zixuan Wang、Yuxin Chen、Pengguang Chen、Yilun Chen、Shu Liu 和 Jiaya Jia。StarVLA-α：降低视觉-语言-动作系统的复杂度。发表于欧洲计算机视觉会议，2026年。URL https://github.com/starVLA/starVLA-alpha.[131] Junhao Cai、Zetao Cai、Jiafei Cao、Yilun Chen、Zeyu He、Lei Jiang、Hang Li、Hengjie Li、Yang Li、Yufei Liu 等。InternVLA-A1：统一机器人操作中的理解、生成与动作。*arXiv* 预印本 *arXiv:2601.02456*，2026年。URL https://arxiv.org/abs/2601.02456.[132] Dantong Niu、Zhuoyang Liu、Zekai Wang、Boning Shao、Zhao-Heng Yin、Anirudh Pai、Yuvan Sharma、Stefano Saravalle、Ruijie Zheng、Jing Wang、Ryan Punamiya、Mengda Xu、Yuqi Xie、Yunfan Jiang、Letian Fu、Konstantinos Kallidromitis、Matteo Gioia、Junyi Zhang、Jiaxin Ge、Haiwen Feng、Fabio Galasso、Wei Zhan、David M. Chan、Yutong Bai、Roei Herzig、Jiahui Lei、Li Fei-Fei、Ken Goldberg、Jitendra Malik、Pieter Abbeel、Yuke Zhu、Danfei Xu、Linxi Fan 和 Trevor Darrell。T-Rex：触觉响应型灵巧操作。*arXiv* 预印本 *arXiv:2606.17055*，2026年。doi：10.48550/arXiv.2606. 17055。URL https://arxiv.org/abs/2606.17055.[133] Jiaming Liu、Xinghao Zhu、Yiming Zhang、Shanghang Zhang、Yinxi Wang、Yichu Yang、Zhuoyang Liu、Dantong Niu、Soha Pouya、Juekun Li 和 Yan Chang。物理词汇：用于构建机器人操作中潜在推理结构的词汇表。*arXiv* 预印本 *arXiv:2610.26854*，2026年。doi：10.48550/arXiv.2610.26854。URL https://arxiv.org/abs/2610.26854.[134] Enyi Wang、Mingxin Wang、Quan Shi、Hetian Guo、Hongyu Wang、Xi Wang、Bin Qian、Yupeng Zheng、Wenxuan Song、Houde Liu、Yong Xu、Cheng Chi、Wenchao Ding、Yilun Chen 和 Yan Wang。TacDyn-WAM：在异构视觉-触觉世界动作模型中学习隐式触觉动力学。*arXiv* 预印本 *arXiv:2610.00638*，2026年。doi：10.48550/arXiv.2610.00638。URLhttps://arxiv.org/abs/2610.00638。

[135] Xukun Li、Yu Sun、Lei Zhang、Bosheng Huang、Yibo Peng、Yuan Meng、Haojun Jiang、Shaoxuan Xie、Guocai Yao、Alois Knoll、Zhenshan Bing、Xinlong Wang 和 Zhenguo Sun。DECO：用于双手灵巧操作的解耦式多模态扩散变换器，配备插件式触觉适配器。arXiv 预印本 arXiv:2602.05513，2026年。doi：10.48550/arXiv.2602.05513。URL https://arxiv.org/abs/2602.05513.[136] Haoran Yuan、Weigang Yi、Zhenyu Zhang、Wendi Chen、Yuchen Mo、Jiashi Yin、Xinzhuo Li、Yue Luo、Jialin Li、Xiangyu Zeng、Chuan Wen、Cewu Lu、Katherine Driggs-Campbell 和 Ismini Lourentzou。VTAM：无需触觉预训练、面向接触丰富型操作的高效视觉触觉世界动作模型。arXiv 预印本 arXiv:2603.23481，2026年。doi：10.48550/arXiv.2603.23481。URLhttps://arxiv.org/abs/2603.23481.[137] NeoteAI 团队和复旦大学 TEAI 团队。 n_0-TWAM：扩展触觉原生世界动作模型，以实现接触丰富型操作，2026年。URL https://arxiv.org/abs/2607.23783.[138] Cheng Chi、Zhenjia Xu、Siyuan Feng、Eric Cousineau、Yilun Du、Benjamin Burchfiel、Russ Tedrake 和 Shuran Song。扩散策略：通过动作扩散进行视觉运动策略学习。《国际机器人研究杂志》，44(10-11)：1684–1704，2025年。doi：10.1177/02783649241273668。URLhttps://doi.org/10.1177/02783649241273668.[139] 英伟达、Johan Bjorck、Fernando Castañeda、Nikita Cherniadev、Xingye Da、Runyu Ding、Linxi "Jim" Fan、Yu Fang、Dieter Fox、Fengyuan Hu、Spencer Huang、Joel Jang、Zhenyu Jiang、Jan Kautz、Kaushil Kundalia、Lawrence Lao、Zhiqi Li、Zongyu Lin、Kevin Lin、Guilin Liu、Edith Llontop、Loic Magne、Ajay Mandlekar、Avnish Narayan、Soroush Nasiriany、Scott Reed、You Liang Tan、Guanzhi Wang、Zu Wang、Jing Wang、Qi Wang、Jiannan Xiang、Yuqi Xie、Yinzhen Xu、Zhenjia Xu、Seonghyeon Ye、Zhiding Yu、Ao Zhang、Hao Zhang、Yizhou Zhao、Ruijie Zheng 和 Yuke Zhu。GR00T N1：面向通用人形机器人的开放基础模型。arXiv 预印本 arXiv:2503.14734，2025年。doi：10.48550/arXiv.2503.14734。URL https://arxiv.org/abs/2503.14734.[140] Morphi Robot。Mora V0：Bench2Dex 评估结果。官方基准测试提交，2026年。URL https://github.com/Bench2Dex/Bench2Dex-Leaderboard/pull/4.[141] 1011.ai。MXI：Bench2Dex 评估结果。官方基准测试提交，2026年。URL https://github.com/Bench2Dex/Bench2Dex-Leaderboard/pull/6.[142] Anyverse Dynamics 和 OpenPhoenix 团队。Phoenix-MWA：Bench2Dex 评估结果。官方基准测试提交，2026年。URL https://github.com/Bench2Dex/Bench2Dex-Leaderboard/pull/5.[143] 北京大学 DAGroup。PyRUA-Dex：作为策略学习器的编码智能体，用于灵巧操作。项目页面与基准测试结果，2026年。URL https://github.com/DAGroup-PKU/PyRUA-Lean/tree/dex.[144] Bo Li、Yuanhan Zhang、Dong Guo、Renrui Zhang、Feng Li、Hao Zhang、Kaichen Zhang、Peiyuan Zhang、Yanwei Li、Ziwei Liu 和 Chunyuan Li。LLaVA-OneVision：轻松实现视觉任务迁移。《机器学习研究汇刊》，2025年。URL https://openreview.net/pdf/5f36c0c4436ea26159310510d02dd062b913402d.pdf.[145] Jianhui Liu、Haoze Sun、Wenbo Li、Yanbing Zhang、Rui Yang、Zhiliang Zhu、Yijun Yang、Shenghe Zheng、Nan Jiang、Jiaxiu Jiang、Haoyang Huang、Tien-Tsin Wong、Nan Duan 和 Xiaojuan Qi。OpenSpatial：用于赋能空间智能的系统化数据引擎。arXiv 预印本 arXiv: 2604.07296，2026年。

[146] Yuanyuan Gao、Hao Li、Yifei Liu、Xinhao Ji、Yuning Gong、Yuanjun Liao、Fangfu Liu、Manyuan Zhang、Yuchen Yang、Dan Xu、Xue Yang、Huaxi Huang、Hongjie Zhang、Ziwei Liu、Xiao Sun、Dingwen Zhang 和 Zhihang Zhong。Holi-Spatial：将视频流演进为全面的三维空间智能。Tong Zhang、Miroslav Dudik、Martin Jaggi、Alekh Agarwal、Sharon Li、Dale Schuurmans、Jerry Zhu、Felix Berkenkamp、Hanze Dong 和 Alberto Bietti 编，第 *43* 届国际机器学习会议论文集，第 306 卷，收录于机器学习研究论文集，第 33300–33316 页。PMLR，2026年。网址：https://proceedings.mlr.press/v306/gao26o.html。

[147] Rui Yang、Ziyu Zhu、Yanwei Li、Jingjia Huang、Shen Yan、SiyuanZhou、Zhe Liu、XiangtaiLi、Shuangye Li、Wenqian Wang、Yi Lin 和 Hengshuang Zhao。视觉空间调优。*arXiv* 预印本 *arXiv: 2511.05491*，2025年。

[148] Xiongkun Linghu、Jiangyong Huang、Baoxiong Jia 和 Siyuan Huang。3D-RFT：面向基于视频的3D场景理解的强化微调。载于 Tong Zhang、Miroslav Dudik、Martin Jaggi、Alekh Agarwal、Sharon Li、Dale Schuurmans、Jerry Zhu、Felix Berkenkamp、Hanze Dong 和 Alberto Bietti 编，第*43*届国际机器学习会议论文集，机器学习研究论文集第306卷，第75240–75261页。PMLR，2026年。URL https://proceedings.mlr.press/v306/linghu26a.html。

[149] Shusheng Yang、Jihan YANG、Pinzhi Huang、Ellis Brown、Zihao Yang、Yue Yu、Shengbang Tong、Zihan Zheng、Yifan Xu、Muhan Wang、Rob Fergus、Yann LeCun、Li Fei-Fei 和 Saining Xie。Cambrian-S：迈向视频中的空间超感知。载于国际学习表征会议，2026年，第78185–78225页，2026年。URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7e3dcf772fa2e8d9087b599c7c07d4cf-Abstract-Conference.html。

[150] Zhongang Cai、Ruisi Wang、Chenyang Gu、Fanyi Pu、Junxiang Xu、Yubo Wang、Wanqi Yin、Zhitao Yang、Chen Wei、Tongxi Zhou、Qingping Sun、Hui En Pang、Jiaqi Li、Oscar Qian、Zhiqian Lin、Xuanke Shi、Kewang Deng、Xiaoyang Han、Zukai Chen、Xiangyu Fan、Hanming Deng、Lewei Lu、Liang Pan、Bo Li、Ziwei Liu、Quan Wang、Dahua Lin 和 Lei Yang。利用多模态基础模型扩展空间智能。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 7879–7890 页，2026年。URL https://openaccess.thecvf.com/content/CVPR2026/html/Cai_Scaling_Spatial_Intelligence_with_Multimodal_Foundation_Models_CVPR_2026_paper.html。

[151] Zhiwen Fan、Jian Zhang、Renjie Li、Junge Zhang、Runjin Chen、Hezhen Hu、Kevin Wang、Peihao Wang、Huaizhi Qu、Shijie Zhou、Dilin Wang、Zhicheng Yan、Hongyu Xu、Justin Theiss、Tianlong Chen、Jiachen Li、Zhengzhong Tu、Zhangyang Wang 和 Rakesh Ranjan。VLM-3R：以指令对齐的三维重建增强视觉语言模型。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 31054–31065 页，2026年。URL https://openaccess.thecvf.com/content/CVPR2026/html/Fan_VLM-3R_Vision-Language_Models_Augmented_with_Instruction-Aligned_3D_Reconstruction_CVPR_2026_paper.html。

[152] Mengfei Du、Binhao Wu、Zejun Li、Xuanjing Huang 和 Zhongyu Wei。《EmbSpatial-Bench：使用大型视觉语言模型为具身任务评测空间理解》。载于 Lun-Wei Ku、Andre Martins 和 Vivek Srikumar 编，《计算语言学协会第 *62* 届年会论文集（第 *2* 卷：短论文）》，第 346–355 页，泰国曼谷，2024年。计算语言学协会。doi：10.18653/v1/2024.acl-short.33。URL https://aclanthology.org/2024.acl-short.33/。

[153] Suchae Jeong、Jaehwi Song、Haeone Lee、Hanna Kim、Jian Kim、Dongjun Lee、Dong Kyu Shin、Changyeon Kim、Dongyoon Hahm、Woogyeol Jin、Juheon Choi 和 Kimin Lee。从跨视角关系中学习多视图空间推理。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 2570–2581 页，2026年。URL https://openaccess.thecvf.com/content/CVPR2026/html/Jeong_Learning_Multi-View_Spatial_Reasoning_from_Cross-View_Relations_CVPR_2026_paper.html。

[154] Xiaoyu Zhan、Wenxuan Huang、Hao Sun、Xinyu Fu、Changfeng Ma、Shaosheng Cao、Bohan Jia、Shaohui Lin、Zhenfei Yin、LEI BAI、Wanli Ouyang、Yuanqi Li、Jie Guo 和 Yanwen Guo。《Actial：激发多模态大语言模型的空间推理能力》。载于神经信息处理系统进展，第 38 卷，第 83946–83971 页，2025年。doi: 10.52202/085713-2813。URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/78e36c70d5051e9e271b00289624d709-Abstract-Conference.html。

[155] Enshen Zhou、Jingkun An、Cheng Chi、Yi Han、Shanyu Rong、Chi Zhang、Pengwei Wang、Zhongyuan Wang、Tiejun Huang、Lu Sheng 和 Shanghang Zhang。《RoboRefer：面向机器人学视觉语言模型的空间指称与推理》。载于神经信息处理系统进展，第 38 卷，第 28404–28481 页，2025年。doi: 10.52202/085713-0957。URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/29416b66c2149872b9d1415a3fd2c5e0-Abstract-Conference.html。

[156] Yingbo Tang、Lingfeng Zhang、Shuyi Zhang、Yinuo Zhao 和 Xiaoshuai Hao。RoboAfford：用于提升机器人操作中物体与空间可供性学习的数据集和基准测试。载于第*33*届 *ACM*国际多媒体会议论文集，第 12706–12713 页，2025年。doi：10.1145/3746027.3758209。URL https://doi.org/10.1145/3746027.3758209。

[157] Wentao Yuan、Jiafei Duan、Valts Blukis、Wilbert Pumacay、Ranjay Krishna、Adithyavairavan Murali、Arsalan Mousavian 和 Dieter Fox。RoboPoint：用于机器人学中空间可供性预测的视觉-语言模型。载于 Pulkit Agrawal、Oliver Kroemer 和 Wolfram Burgard（编），第八届机器人学习会议论文集，第 270 卷，机器学习研究论文集，第 4005–4020 页。PMLR，2025年。URL https://proceedings.mlr.press/v270/yuan25c.html。

[158] Haozhan Shen、Peng Liu、Jingcheng Li、Chunxin Fang、Yibo Ma、Jiajia Liao、Qiaoli Shen、Zilun Zhang、Kangjia Zhao、Qianqian Zhang、Ruochen Xu 和 Tiancheng Zhao。《VLM-R1：一种稳定且具有泛化能力的 R1 风格大型视觉语言模型》，2025年。网址： https://arxiv.org/abs/2504.07615。

[159] Delin Qu、Haoming Song、Qizhi Chen、Zhaoqing Chen、Xianqiang Gao、Dong Wang、Xinyi Ye、Qi Lv、Modi Shi、Guanghui Ren、Cheng Ruan、Maoqing Yao、Haoran Yang、Jiacheng Bao、Bin Zhao 和 Xuelong Li。《EO-1：一种用于通用机器人控制的开放统一具身基础模型》，2026年。网址 https://arxiv.org/abs/2508.21112。

[160] Kaiyuan Eric Chen、Shuangyu Xie、Zehan Ma、Pannag Sanketi 和 Ken Goldberg。Robo2VLM：利用大规模机器人操作数据改进视觉问答。在神经信息处理系统进展第 38 卷，2025年。doi: 10.52202/085713-0733。URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1f467c3e37abf9f86c78f44c6a27ee7c-Abstract-Datasets and Benchmarks Track.html。

[161] Yuheng Ji、Huajie Tan、Jiayu Shi、Xiaoshuai Hao、Yuan Zhang、Hengyuan Zhang、Pengwei Wang、Mengdi Zhao、Yao Mu、Pengju An、Xinda Xue、Qinghang Su、Huaihai Lyu、Xiaolong Zheng、Jiaming Liu、Zhongyuan Wang 和 Shanghang Zhang。RoboBrain：从抽象到具体的机器人操作统一大脑模型。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 1724–1734 页，2025年。URL https://openaccess.thecvf.com/content/CVPR2025/html/Ji_RoboBrain_A_Unified_Brain_Model_for_Robotic_Manipulation_from_Abstract_CVPR_2025_paper.html。

[162] Yuan-Hong Liao、Sven Elflein、Liu He、Laura Leal-Taixé、Yejin Choi、Sanja Fidler 和 David Acuna。LongPerceptualThoughts：为系统一感知蒸馏系统二推理。发表于语言建模会议，2025年。URL https://openreview.net/forum?id=SrKdi4MsUW。

[163] Haochen Wang、Xiangtai Li、Zilong Huang、Anran Wang、Jiacong Wang、Tao Zhang、Jiani zheng、Sule Bai、Zijian Kang、Jiashi Feng、Zhuochen Wang 和 Zhaoxiang Zhang。可追溯证据增强的视觉接地推理：评估与方法。发表于国际学习表征会议，第 2026 卷，第 148769–148794 页，2026年。网址 https://proceedings.iclr.cc/paper_files/paper/2026/hash/f09ab24aa37ae2664b3d923cdb0319b5-Abstract-Conference.html。

[164] Yuhong Liu、Beichen Zhang、Yuhang Zang、Yuhang Cao、Long Xing、Xiaoyi Dong、Haodong Duan、Dahua Lin 和 Jiaqi Wang。Spatial-SSRL：通过自监督强化学习增强空间理解。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 9570–9581 页，2026年。网址 https://openaccess.thecvf.com/content/CVPR2026/html/Liu_Spatial-SSRL_Enhancing_Spatial_Understanding_via_Self-Supervised_Reinforcement_Learning_CVPR_2026_paper.html。

[165] Ryan Hoque、Peide Huang、David J. Yoon、Mouli Sivapurapu 和 Jian Zhang。EgoDex：从大规模第一人称视角视频中学习灵巧操作。载于国际学习表征会议，2026年。URL https://openreview.net/pdf/d4fea01944202c75aa13dc34757e154478eb8df7.pdf。

[166] Zicong Fan、Omid Taheri、Dimitrios Tzionas、Muhammed Kocabas、Manuel Kaufmann、Michael J. Black 和 Otmar Hilliges。ARCTIC：用于灵巧双手手物操作的数据集。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 12943–12954 页，2023。URL https://openaccess.thecvf.com/content/CVPR2023/html/Fan_ARCTIC_A_Dataset_for_Dexterous_Bimanual_Hand-Object_Manipulation_CVPR_2023_paper.html。

[167] Prithviraj Banerjee、Sindi Shkodrani、Pierre Moulon、Shreyas Hampali、Shangchen Han、Fan Zhang、Linguang Zhang、Jade Fountain、Edward Miller、Selen Basol、Richard Newcombe、Robert Wang、Jakob Julian Engel 和 Tomas Hodan。HOT3D：基于自我中心多视角视频的三维手部与物体跟踪。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 7061–7071 页，2025年。URL https://openaccess.thecvf.com/content/CVPR2025/html/Banerjee_HOT3D_Hand_and_Object_Tracking_in_3D_from_Egocentric_Multi-View_CVPR_2025_paper.html。

[168] Taein Kwon、Bugra Tekin、Jan Stühmer、Federica Bogo 和 Marc Pollefeys。H2O：通过双手操纵物体实现第一人称交互识别。载于 *IEEE/CVF* 国际计算机视觉会议论文集，第 10138–10148 页，2021年。URL https://openaccess.thecvf.com/content/ICCV2021/html/Kwon_H2O_Two_Hands_Manipulating_Objects_for_First_Person_Interaction_Recognition_ICCV_2021_paper.html。

[169] Xinyu Zhan、Lixin Yang、Yifei Zhao、Kangrui Mao、Hanlin Xu、Zenan Lin、Kailin Li 和 Cewu Lu。《OAKINK2：用于复杂任务完成的双手与物体操控数据集》。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 445–456 页，2024年。网址 https://openaccess.thecvf.com/content/CVPR2024/html/Zhan_OAKINK2_A_Dataset_of_Bimanual_Hands-Object_Manipulation_in_Complex_Task_CVPR_2024_paper.html。

[170] Shangchen Han、Po-Chen Wu、Yubo Zhang、Beibei Liu、Linguang Zhang、Zheng Wang、Weiguang Si、Peizhao Zhang、Yujun Cai、Tomas Hodan、Randi Cabezas、Luan Tran、Muzaffer Akbay、Tsz-Ho Yu、Cem Keskin 和 Robert Wang。《UmeTrack：面向虚拟现实的统一多视图端到端手部跟踪》。发表于*SIGGRAPH Asia 2022* 会议论文集，第 1–9 页，2022年。doi：10.1145/3550469.3555378。网址https://doi.org/10.1145/3550469.3555378。

[171] Yihang Li、Xuelong Wei、Jingzhou Luo、Yingjing Xiao、Yibo Bai、Guangyuan Zhou、Teng Zou、Chen- guang Gui、Jiajun Wen、He Zhang、Kangliang Chen、Xing Pan、Shuaiyan Liu、Daming Wang、Tao An、Jiayi Li、Shibo Jin、Wanwan Zhang、Tianyu Wang、Boren Wei、Zhixuan Huang、Fangsheng Liu、Ruodai Li、Hui Zhang、Anson Li、Yicheng Gong、Peng Cao、Jiaming Liang 和 Liang Lin。EgoLive：来自真实世界人类任务的大规模自我中心数据集。*arXiv* 预印本 *arXiv:2604.23570*，2026年。

[172] Patrick Rim、Kevin Harris、Braden Copple、Shangchen Han、Xu Xie、Ivan Shugurov、Sizhe An、He Wen、Alex Wong、Tomas Hodan 和 Kun He。SHOW3D：在自然环境中捕获三维手和物体的场景。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第7111页 – 7120，2026年。URL https://openaccess.thecvf.com/content/CVPR2026/html/Rim_SHOW3D_Capturing_Scenes_of_3D_Hands_and_Objects_in_the_CVPR_2026_paper.html。

[173] Lightwheel。EgoSuite-Open100K：一个完整标注的开放式第一人称视角人体数据集。 https://huggingface.co/collections/LightwheelAI/egosuite-open100k，2026年8月。数据集。

[174] Takehiko Ohkawa、Kun He、Fadime Sener、Tomas Hodan、Luan Tran 和 Cem Keskin。AssemblyHands：通过三维手部姿态估计实现第一人称视角活动理解。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第12999–13008页，2023年。URL https://openaccess.thecvf.com/content/CVPR2023/html/Ohkawa_AssemblyHands_Towards_Egocentric_Activity_Understanding_via_3D_Hand_Pose_Estimation_CVPR_2023_paper.html。

[175] Gyeongsik Moon、Shunsuke Saito、Weipeng Xu、Rohan Joshi、Julia Buffalini、Harley Bellan、Nicholas Rosen、Jesse Richardson、Mallorie Mize、Philippe De Bree、Tomas Simon、Bo Peng、Shubham Garg、Kevyn McPhail 和 Takaaki Shiratori。《重新照明的三维交互手部数据集》。载于神经信息处理系统进展，第36卷，第17689–17701页，2023年。doi: 10.52202/07528 0-0776。URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/396beafa6feba781a7114780e6837253-Abstract-Datasets and Benchmarks.html。

[176] Christian Zimmermann、Duygu Ceylan、Jimei Yang、Bryan Russell、Max Argus 和 Thomas Brox。FreiHAND：一个用于从单张RGB图像中进行手部姿态和形状无标记捕捉的数据集。载于*IEEE/CVF* 国际计算机视觉会议论文集，第813–822页，2019年。URL https://openaccess.thecvf.com/content_ICCV_2019/html/Zimmermann_FreiHAND_A_Dataset_for_Markerless_Capture_of_Hand_Pose_and_ICCV_2019_paper.html。

[177] Christian Zimmermann 和 Thomas Brox。学习从单张RGB图像估计三维手部姿态。载于*IEEE*国际计算机视觉会议论文集，第4903–4911页，2017年。URL https://openaccess.thecvf.com/content_iccv_2017/html/Zimmermann_Learning_to_Estimate_ICCV_2017_paper.html。

[178] Aiden Nibali、Zhen He、Stuart Morgan 和 Luke Prendergast。使用卷积神经网络进行数值坐标回归。*arXiv*预印本 *arXiv:1801.07372*，2018年。

[179] Xiao Sun、Bin Xiao、Fangyin Wei、Shuang Liang 和 Yichen Wei。积分人体姿态回归。载于欧洲计算机视觉会议（*ECCV*）论文集，第 529–545 页，2018 年。URL https://openaccess.thecvf.com/content_ECCV_2018/html/Xiao_Sun_Integral_Human_Pose_ECCV_2018_paper.html。

[180] Ryan Punamiya、Simar Kareer、Zeyi Liu、Josh Citron、Ri-Zhao Qiu、Xiongyi Cai、Alexey Gavryushin、Jiaqi Chen、Davide Liconti、Lawrence Y Zhu 等。EgoVerse：用于从全球范围学习机器人学习的自我中心人类数据集。arXiv 预印本 arXiv:2604.07607，2026年。[181] Yan Qin、Yue Chen、Wenwei Lin、Shujia Liu、Chuqiao Lyu、Kailun Su、Weiyang Jin、Chenze Yu、Ping Luo、Wenbo Ding 等。DexTouch-WM：从人类触摸中学习动作条件触觉世界模型，以实现灵巧机器人操作。arXiv 预印本 arXiv:2609.20649，2026年。[182] PX OmniSharing DB。PX OmniSharing DB。 https://huggingface.co/datasets/paxini/Omnisharing_DB_SampleData，2026年。[183] PsiBot。SynData：用于具身智能的大规模真实世界多模态数据集。 https://huggingface.co/datasets/PsiBotAI/SynData，2026年。[184] Sharpa。SharpaDex v1.0：大规模双手灵巧操作示范。 https://huggingface.co/datasets/Sharpa-Robotics/SharpaDex-v1.0，2026年。[185] RoCo Task Board Assembly。RoCo task board assembly LeRobot 数据集。 https://huggingface.co/datasets/SharpaIT/RoCo_TaskBoardAssembly，2026年。[186] Robotic Origami Challenge Fold Plane。Robotic origami challenge fold plane LeRobot 数据集。 https://huggingface.co/datasets/SharpaIT/Robotic_Origami_Challenge，2026年。[187] Robotic Origami Challenge。机器人折纸挑战赛。 https://robotic-origami-challenge.github.io/，2026年。[188] Jongbin Lim、Taeyun Ha、Mingi Choi、Jisoo Kim、Byungjun Kim、Subin Jeon 和 Hanbyul Joo。HRDexDB：用于跨具身灵巧抓取的配对人机数据集。arXiv 预印本 arXiv:2604.14944，2026年。[189] Dao-AILab。FlashAttention-3 变长注意力接口。 Dao-AILab/flash-attention，修订版 0dc2cb48，2026年。[190] 英伟达。NVSHMEM：使用编程接口，2026年。URL https://docs.nvidia.com/nvshmem/api/latest/using.html。文档，访问日期：2026-11-12。

A. 统一标注记录

这些示例展示了第 2.2 节中的共享标注层。一个可重复使用的记录结合了场景描述、任务目标、时间区间、执行结果和质量属性；字幕是其下游用途之一。该记录保留了指向原始任务文本、原生轨迹通道和传感器时间戳的链接。只有在来源支持时才填充字段：视频可以支持语义事件边界，但未必能提供关节度量运动或触觉测量。

A.1. 共享上下文与区间级描述

一个场景，三个动作。 图 A.1 中的三个裁剪片段共享场景上下文和任务级标注。动态卡片描述了分开扫帚和簸箕、将纸张扫入簸箕，以及将簸箕里的东西倒入垃圾桶的过程。扫掠描述依据扫帚反复挥动的动作；倒空描述则捕捉了一次状态变化：左臂将装有揉皱的纸和蓝色包装纸的簸箕倾倒在垃圾桶上方，然后放下已空的簸箕。对左、右夹爪、工具、纸张和垃圾桶的统一指称，将描述、任务区间和可见关系联系起来，例如纸张在被倒入垃圾桶前放在簸箕里。若空间模块提供了轨迹、掩码或深度信息，也可以将其与相同实体及其语言描述关联起来。

A.2. 语义子任务、执行结果与证据

目标定义了子任务边界。 对于机器人轨迹，共享的语义处理阶段会生成一段简短的`task_title`、一个有序的 `task_description`，以及带时间戳的子任务区间。每个子任务将朝向同一操作目标的运动归为一组：伸手、抓取并抬起一块面包构成一个拾取子任务；将面包放到盘子上并释放，则构成下一个子任务。若针对同一目标的重复尝试最终成功，这些尝试仍归为一个子任务，因此，两次抓取失败和一次成功抓取不会被拆分成三个互不相关的任务。父级区间内可记录可见的原子动作或执行阶段，而不将其拆分为独立的训练回合。

结构化示例。 表A.1 展示了标注协议；它并非额外测量得到的轨迹。输入目标是“将面包放到盘子上”，时长为 11.4 s，序列以 1 帧/秒的速率采样，并带有显式时间戳。详细描述为：“两次尝试失败后，用左夹爪拿起面包，将其放到盘子上并释放，然后将机械臂移回初始位置，为下一个任务做好准备。”这描述的是观察到的流程，而不是将通用目标与物体名称组合而成的描述。

表 A.1：一条共享记录中的子任务区间示例。时间以回合内秒数表示。每个区间都有一个行为标签和一个单独的空闲标记。终止边界由给定时长决定，而不是由采样图像的数量决定。

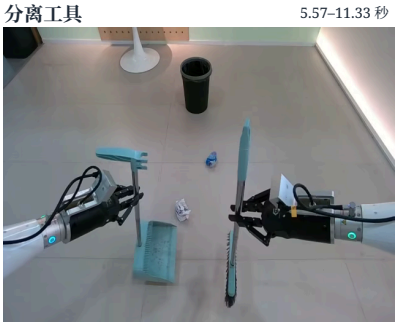
区间（秒）	子任务描述	行为	Idle
0–5	两次尝试未成功后，用左夹爪拿起面包。	重试后成功	false
5–10	用左夹爪将面包放在盘子上，然后释放。	成功	false
10–11.4	结束：将机械臂返回初始位置，为下一个任务做好准备。	end	false

区分意图、执行和可见性。 该记录保留原始目标，以及修正后的标题和详细描述。 `task_alignment` 检查文本是否指出了可见物体、操作和目的地； `outcome` 评估适用的目标是否已实现。正确的指令可能伴随失败，而不正确的物体描述可能伴随对视觉支持任务的成功执行。原生成功标志与基于视频的评估保持分离。在该示例中， 对齐为 `consistent` ， 总体结果为 `success`，尽管最初出现了抓取失败。

每个机器人 inte rval 从 {`normal`, `success`, `failure`, `retry then success` 中接收一个行为标签，

共享任务和场景具身形态 Galbot G1 · 双臂机器人**任务类别** 家务**任务标签** 双手操作 · 抓取与放置 · 工具使用 · 清洁**任务** 清扫垃圾并倒空簸箕**描述** 用左夹爪 ಹೊಡೆ住簸箕，用右夹爪拿开扫帚，将垃圾扫入簸箕，然后把收集到的垃圾倒入垃圾桶。**场景** 一间整洁的铺瓷砖的房间，墙壁为薄荷绿色，配有一张白色底座桌和一个黑色垃圾桶。**物体** 浅蓝色扫帚和簸箕 · 揉皱的纸 · 蓝色包装纸 · 黑色垃圾桶**任务结果** 成功 — 垃圾已收集并倒入垃圾桶。**视图质量** 工作空间、工具和活动夹爪均清晰可见。

分离工具 5.57–11.33 秒



动态卡片

摘要
A dual 机械臂机器人使用左夹爪握住浅蓝色簸箕，并用右侧夹爪抓取相配的扫帚，然后将这两件工具分开。

详细动作
... The 夹爪绕住扫帚柄闭合，随后右臂将扫帚向上并远离簸箕拉开，使这两件工具分离。 ...

子任务
用左夹爪拿住簸箕，用右侧夹爪将扫帚抬出。

执行
空闲：否
结果：成功

状态变化 离散变化

初始
左夹爪握着簸箕的手柄
扫帚放在簸箕内。

动作
右侧夹爪抓取扫帚柄
并将其从簸箕中抬起。

最终
扫帚由右侧夹爪单独握持
夹爪，离开左侧夹爪持有的簸箕
夹爪。

扫掠纸张 13.33–18.23 秒



动态卡片

摘要
A dual 配备机械臂的机器人用右侧夹爪中的扫帚，将揉皱的纸扫入左夹爪拿着的簸箕中。

详细动作
... The 机械臂用短促、重复的动作移动扫帚，将白色纸团推向由左臂稳稳拿着的簸箕。 ...

子任务
用右侧夹爪将纸张扫入簸箕，同时用左夹爪将其稳稳按住。

执行空闲：否
结果：成功

状态变化 连续过程扫帚反复扫动，将纸张扫入簸箕。

倒空簸箕 26.60–31.57 秒



动态卡片

摘要
A dual 机械臂机器人将浅蓝色簸箕倾斜到黑色垃圾桶上方，把揉皱的纸和蓝色包装纸倒入垃圾桶。

详细动作
... The 机械臂抬起簸箕，并逐渐将其倾斜到垃圾桶敞开的桶口上方，同时旋转平底锅，使其中的内容物滑出并落入垃圾桶。 .....

子任务
用左夹爪将簸箕倾斜到垃圾桶上方，以倒空垃圾。

执行
空闲：否
结果：成功

状态变化 离散变化**初始**左臂拿着浅蓝色簸箕，里面装着揉皱的纸和蓝色包装纸，位于地板上方、黑色垃圾桶附近。**动作**左臂将簸箕抬起并倾斜到垃圾桶上方，把里面的东西倒进去。**最终**簸箕已空，并被放低至地板附近，垃圾现在已倒入垃圾桶内。

图 A.1：任务级和子任务级标注。 共享卡片整合了 Galbot G1 具身形态、任务类别和标签、完整任务描述、场景上下文、结果及可见性。三个定时片段展示了子任务、执行结果和状态变化。显示标签依据提供的帧和保留的描述生成。

end}, 与 `is_idle` 一起。有意复位属于结束阶段，但并非空闲状态。结束片段是可选的：录制可能在成功释放后立即停止，无需复位或暂停。总体结果也允许 `uncertain`，当缺少上下文或可见性不足以进行评估时，可将其与明显未完成的任务或注释处理失败区分开来。

带时间戳的输入与输出验证。 输入包括源任务文本、实际时长、采样率、相机标识符和采样图像时间戳。采样率不同于原始视频帧率。在稀疏采样情况下，边界反映证据的时间分辨率，而非精确到帧的接触事件。多个视图展示的是同一个回合，而不是重复尝试。输出验证会检查 JSON 模式、标签词汇表、区间顺序，以及所提供时长内的边界。可恢复的格式错误可以得到纠正并记录；无法解析的响应会被标记为标注失败，而不是任务失败。保留的源时间偏移量用于在较长的共享视频文件中定位片段内区间。

语义信息同时附有质量与数值证据。 R2 第 1 阶段会在语义响应旁保留感知视频质量得分及其依据。得分范围为 1（任务关键内容不可用）至 5（清晰、稳定且具有足够可见性），不用于衡量任务成功。独立的低层测量指标用于刻画模糊、曝光、帧变化和解码问题。对于可用的原生轨迹，数值检查会收集运动、导数、趋势残差和平稳性证据，并记录适用性和校准状态。稀疏的视觉语言模型观测无法排除短暂的画面冻结或高频抖动，而未经标定的数值规则也无法证明某条轨迹正常。这些记录支持筛选与审查，而无需重新生成语义描述。阶段 1 提取证据；下游处理决定是否保留、拆分、排除或平滑轨迹。

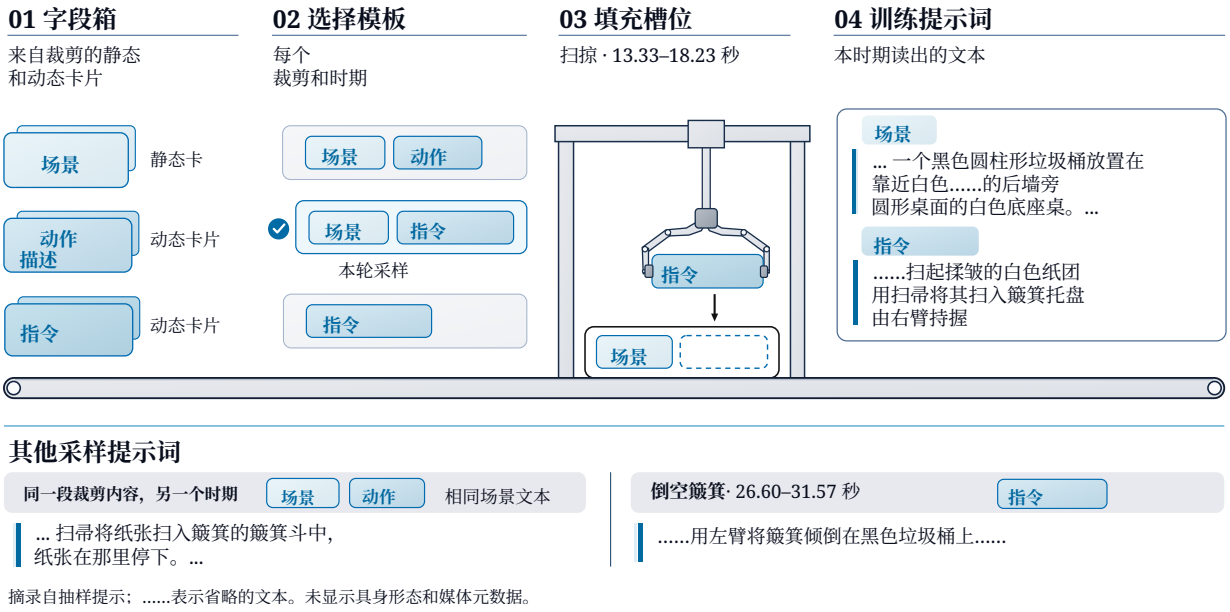
A.3. 在不同准备分支中重复使用同一条记录

同一片段，不同提示。 对于扫地片段（13.33–18.23 秒），来自不同时期的两个提示共享相同的场景描述，其中包括句子：“一个黑色圆柱形垃圾桶靠着后墙，位于一张圆形桌面的白色底座桌旁。”其中一个描述动作：“扫帚将纸张扫进簸箕的簸箕斗，纸张随后停在那里。”另一个则指示机器人用左臂将簸箕竖直拿稳，并用右侧夹爪中的扫帚将纸张扫进去。针对稍后清空片段（26.60–31.57 秒）的提示，则使用了不含场景文本的指令（图 A.2）。

图 A.2 展示了视频生成如何使用共享记录，而非单独的标注流程。其他分支会从同一证据中选取不同字段：

- **具身理解：** 仅根据问题所涉及的观测，构造有关活动夹爪、物体关系、子任务顺序或重试结果的问题。
- **视觉预测：** 将静态上下文与特定时间间隔的动态描述或指令相结合，如图所示。
- **动作预测：** 将选定的子任务指令和任务上下文与相应的状态/动作区间及相机流配对。语义边界用于指导选择，但不会建立数值同步，也不会创建控制目标。
- **触觉预测：** 将可用的接触观测和有效性掩码关联到相同的事件区间。仅凭可见的持握或触碰无法提供经过测量的触觉目标。

记录保留完整的任务描述、有序计划和局部区间。如果筛选过程将一条轨迹拆分，每个保留的分段都会保留其源时间对应关系和父计划引用。移除空闲区间既不会抹去原始任务上下文，也不会跨越由此产生的间隔连接动作。训练提示词会选择适当的上下文和局部指令；结果、质量、空闲和处理状态字段用于筛选与审计，而不是不加区分地附加到每条提示词中。



其他采样提示词

同一段裁剪内容，另一个时期 场景 动作 相同场景文本

... 扫帚将纸张扫入簸箕的簸箕斗中，
纸张在那里停下。...

倒空簸箕 · 26.60–31.57 秒 指令

……用左臂将簸箕倾倒在黑色垃圾桶上……

摘录自抽样提示；……表示省略的文本。未显示具身形态和媒体元数据。

图 A.2：基于可复用标注进行提示组装。替代模板将场景与动态卡片结合起来。下方的面板展示了提示组装与轨迹字段之间的联系：详细的任务文本和子任务指令提供条件输入，边界用于选择时间间隔，而结果、空闲状态和质量字段则支持下游筛选和审计。这些面板说明了字段的用途；它们并未报告所显示裁剪片段的其他预测结果。

B. 理解能力训练细节

我们通过分阶段的监督训练流程对理解塔进行调整，将广泛的空间与具身学习和以任务为重点的精炼相结合。从 Qwen3-VL-8B [16], 开始，训练逐步强化空间理解，同时保留通用视觉监督。下文将介绍经评估检查点的数据组成和训练轨迹，并区分可用的来源清单、准备好的训练混合数据，以及实际使用的监督样本曝光次数。

在中期训练阶段，我们的来源清单包含约 6,360 万条候选记录，其中 6,080 万条在经过针对各来源的预处理和验证后符合条件。符合条件的记录包括约 1,170 万条通用记录和 4,920 万条具身导向记录，其中包含内部汇编的集合。这些数字描述的是来源池中的记录，而非全局去重后的样本或实际训练曝光量。我们从这些记录中准备了一个包含 2,000 万个样本的中期训练混合数据，按来源类别计，其中约 13.3% 为通用数据，86.7% 为具身导向数据。详细数据集见表 B.1。

训练从广泛的中期训练逐步过渡到面向任务的后训练。中期训练侧重于大规模空间数据集和具身数据集，同时保留通用监督。后训练首先采用任务均衡多源混合数据，然后使用经过筛选的问答集合、紧凑的特定任务训练集和通用数据回放对模型进行精炼。最终精炼混合数据包含 57,565 个面向任务的样本、来自紧凑特定任务训练集的 42,737 个样本，以及 20,000 个通用样本。这些紧凑数据集包括 EmbSpatial-SFT、MindCube，以及 Embodied-R1.5 中与 ERQA 相关的训练子集。

我们使用 AdamW 对语言和视觉组件进行微调，参数为 $\beta_1 = 0.9$、$\beta_2 = 0.95$、$\epsilon = 10^{-8}$，权重衰减为 0.1。训练采用 bfloat16 计算、全分片数据并行、激活检查点，以及阈值为 1.0 的梯度范数裁剪。语言模型在整个训练期间的峰值学习率为 10^{-5}；视觉组件在中期训练期间使用 2×10^{-6}，在多源后训练和最终精炼期间均使用 10^{-6}。这三个阶段均采用带线性

表 B.1：理解塔训练混合数据中包含的数据来源。M 表示中期训练，P 表示后训练。纳入的内容是选定子集，而非整个公开发布版本。列出数据集发布系列是为了说明来源；不同系列可能共享上游媒体或标注，因此不一定互不重叠。内部汇编的数据集与公开发布版本分开列出。

数据集／发布系列	选定子集中的监督	阶段
通用多模态理解		
Cambrian [92]	通用视觉指令遵循	M
LLaVA-OneVision [144]	图像/视频理解和指令遵循	M
空间感知和推理		
SPAR [99]	空间关系、几何结构、多视图和视频推理	M, P
OpenSpatial [145]	基于视频的空间理解	M
HoliSpatial-QA [146]	空间问答	M
VST [147]	多视图与视频空间理解	M, P
3D-RFT [148]	空间推理与监督式推理轨迹	M, P
VSI-590K [149]	基于视频的空间与指标推理	M, P
SenseNova-SI [150]	空间理解	M, P
VLM3R-VSTI [151]	多视图空间推理	M, P
EmbSpatial-SFT [152]	具身空间关系	P
MindCube [95]	多视图空间推理	P
XVR [153]	跨视图空间关系	P
Viewpoint-100K [154]	视角依赖的空间推理	P
指称落地与可供性理解		
RefSpatial [155]	空间指称与点定位	M, P
RoboAfford [156]	功能性定位与可供性定位	M, P
RoboPoint [157]	指向、空间定位与放置	M, P
HoliSpatial-3D-Grounding [146]	结合几何上下文的三维定位	M
VLM-R1-RefCOCO [158]	指代表达定位	M, P
具身任务理解		
EO [159]	具身感知与任务理解	M, P
Robo2VLM [160]	以机器人为中心的视觉问答	M
ShareRobot [161]	可供性、轨迹与规划监督	M, P
Embodied-R1.5 [82]	与ERQA相关的具身问答	P
感知与基于语境的推理		
LongPerceptualThoughts [162]	视觉感知推理	P
TreeVGR [163]	基于语境的视觉推理	P
Spatial-SSRL [164]	空间推理监督	P
我们构建的数据		
Innata-Understanding（我们的方法）	用于物体定位、空间推理和具身任务理解	M, P

预热。初始学习率和最低学习率分别设为各自峰值的 1%。最终检查点源自一个中期训练模型，该模型经历了约 7.99M 次监督样本曝光，随后在多源后训练中经历了约 0.20M 次曝光，并在最终精炼阶段经历了 0.12M 次曝光。

C. 灵巧手

C.1. 自我中心视角手部重建模型

本小节介绍了用于构建第 2.5.3 节中自我中心视角运动监督的手部重建模型。其源训练混合数据包含 ~ 208小时视频和73。8k 张静态图像。我们基于 ACE-Ego-Hand [33], , 采用针孔相机输入、关节分布监督以及序列级形状与位置优化。该训练混合数据不同于表 C.6 中的动作学习语料库。

C.1.1. 架构与相机表示

我们沿用 ACE-Ego-Hand 的干净潜变量编码器和双向时空解码器 [33], , 使用 `Wan2.2-Fun-5B-Control` 特征 [31, 32]。为统一不同来源的相机模型，我们将所有输入校正到针孔域，并在裁剪和缩放后更新内参。随后，我们通过解析方式生成标准射线图 $\mathbf{r}_K(u,\ v) = K^{-1}[u, v, 1]^\top$，用于射线编码和几何解码，以此替代自由预测的射线场。有可用标定数据时，我们使用所提供的标定；否则，还会根据输入特征预测针孔相机参数，并通过相同的映射生成射线。表 C.1总结了模型和训练设置。

表 C.1：手部重建配置。 该配置描述了本工作中使用的模型。学习率用于区分完整训练的解码器和头部、LoRA 参数以及可训练的图块嵌入。

组件	配置
视觉初始化	`Wan2.2-Fun-5B-Control`，干净后验均值潜变量
特征提取	Transformer 块索引 15，LoRA 秩 / 尺度 64/64
时空解码器	4 层，宽度 384，8 个注意力头
视频输入	81 个 RGB 帧，对应 21 个时间潜变量样本
图像输入	5 个重复的 RGB帧，对应 2 个时间潜变量样本
训练时长 / 全局批次	80,000 个优化器步数 / 48 个输入样本
优化器	AdamW，权重衰减 10^{-2}
学习率	解码器 / 头部 2×10^{-4}、LoRA 10^{-4}、图块嵌入 2×10^{-5}
学习率调度	200 步线性预热，随后进行余弦衰减
热图目标	高斯，$\sigma = 0$。6 个网格间隔，JS系数 0。1
热图 / 相机损失递增	500 个优化器步数

C.1.2. 训练数据

表 C.2报告了训练数据及其采样概率。我们通过对任务和场景进行分层，从更大的数据池中抽取 EgoLive、Lightwheel 和 EgoStandard，并将未见场景和工作进程留作验证。对于 UmeTrack，我们使用“real”子集中的视频，其 `cam1`和 `cam2`中手部可见。我们有意将 EgoDex [165] 排除在训练混合数据之外，因为我们检查过的原生 ARKit 标注存在骨骼骨段过长以及三维到图像对齐不一致的问题。对于动作学习，我们改为使用模型从 EgoDex 视频中重新估计手部姿态，如第 2.5.3 节所述。图 C.1展示了这十三个数据源以及选取的 UmeTrack 视图。

预处理。我们将 HOT3D Aria、Re:InterHand 和 UmeTrack 的鱼眼输入重新映射到针孔画布，并对 Aria 录制内容应用固定的光度处理（去渐晕和颜色校正）。我们对关节、单位、左右手属性和 MANO 规范进行标准化，并使用有效性及存在性/可见性掩码表示各来源的标注覆盖率。

裁剪与缩放增强。为改变手部的表观大小和位置，我们采样覆盖感知裁剪，并在 VAE 编码前对每个序列统一应用一种空间变换。我们同步更新内参、二维目标、可见性和射线，同时保留相机坐标系三维几何结构和关节运动。

表 C.2：手部重建训练数据。FreiHAND 和 RHD 提供静态图像；其余各行均统计视频帧。时长不包括静态图像。画布尺寸为宽度 ×高度，采样表示来源选择概率。不包括验证数据。

来源	记录数	#帧数	小时	画布	采样
ARCTIC [166]	267	184,373	1.71	672 × 480	8%
HOT3D Aria [167]	126	444,649	4.12	480 × 480	10%
H2O [168]	114	68,929	0.64	832 × 480	5%
OakInk2 [169]	550	935,260	8.66	832 × 480	8%
UmeTrack Real [170]	1,992	876,018	8.11	480 × 480	5%
EgoLive [171]	6,823	5,390,474	50.28	480 × 480	11%
Lightwheel	6,452	5,309,132	49.99	832 × 480	11%
SHOW3D [172]	3,162	3,026,568	28.02	384 × 480	10%
EgoStandard [173]	6,493	5,309,250	49.99	832 × 480	11%
AssemblyHands [174]	800	697,054	6.46	640 × 480	6%
Re:InterHand [175]	43	21,315	0.59	480 × 480	5%
FreiHAND [176]	32,560	32,560	—	224 × 224	5%
渲染手部姿态数据集（RHD） [177]	41,258	41,258	—	320 × 320	5%
总计	**100,640**	**22,263,022 帧 +73,818 张图像**	**208.57**	—	**100%**

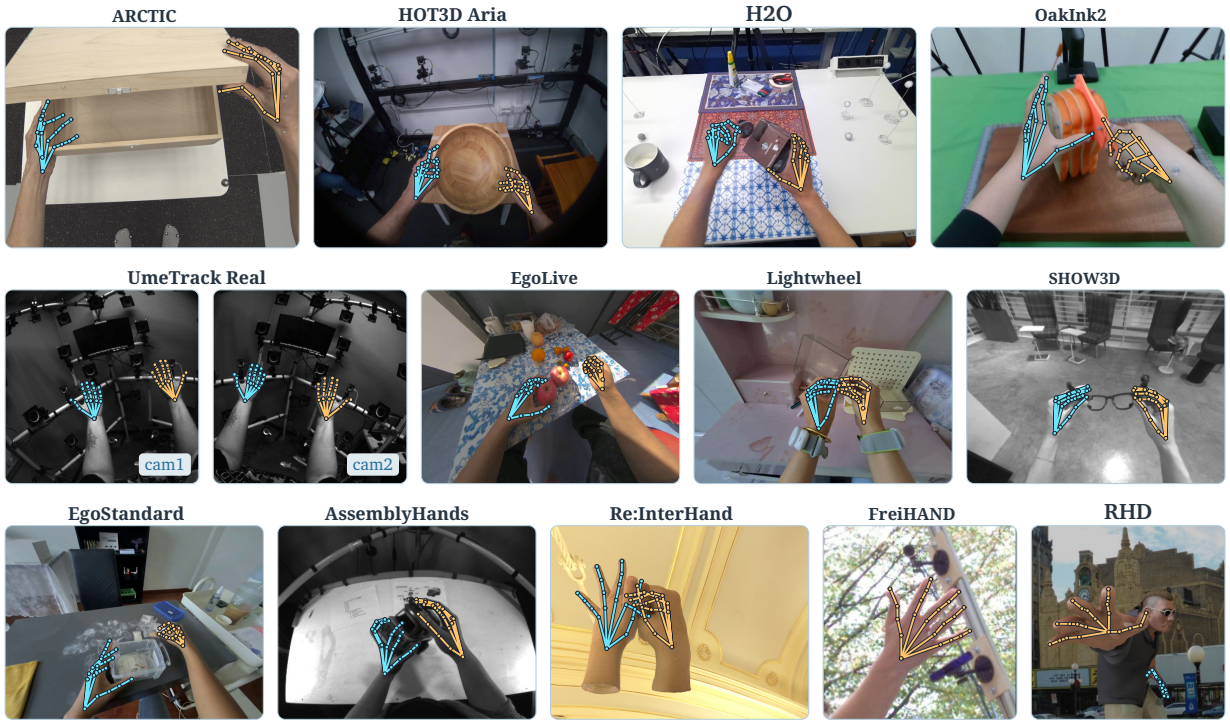


图 C.1：训练来源多样性。我们的训练集涵盖受控捕捉、日常操作和合成图像。叠加的手部骨架取自各数据集的原生标注。

我们使用基础画布进行验证。

C.1.3. 监督与优化

我们直接监督 MANO 几何结构、关节运动、二维关节定位与重投影、手部存在性与可见性，以及时间平滑性。具体而言，我们针对 UmeTrack Real、Assem- blyHands、Lightwheel 和 EgoStandard 拟合 MANO 参数，并保持回合内形状一致。我们根据目标可用性对每个目标函数应用掩码，按来源计算有效残差的平均值，并使用表 C.2 中的概率组合各来源。表 C.3 列出了各项系数。下文重点介绍额外的分布监督和相机监督。

关节分布监督。仅凭坐标均值，无法区分局部集中的关节预测与弥散或分离的空间支持。因此，我们使用以已标注关节 [178, 179] 为中心的高斯目标 Q，对基于注意力的关节分布 P 进行监督：

$$Q_i = \frac{\exp[-d_i^2/(2\sigma^2)]}{\sum_j \exp[-d_j^2/(2\sigma^2)]}, \qquad \mathcal{L}_{\text{dist}} = \text{JS}(P, Q). \tag{C.1}$$

其中，d_i 表示解码器端点网格上各网格间隔内与目标的距离，$\sigma = 0。6$，而 JS 表示詹森–香农散度。我们在原生解码器时间点，对具有有效图像和几何标注的关节应用此目标函数，其中包括已可靠标注的遮挡关节。我们也监督坐标均值，因为边界截断可能会使高斯分布的期望发生偏移。

几何感知相机监督。 相同的标定参数误差在不同相机上可能产生不同的射线误差。因此，我们根据视线方向灵敏度缩放对数焦距和归一化主点惩罚项。对于一个轴，使用归一化坐标和视线方向 $b = (u - c)/f$，我们采用

$$\begin{aligned} \mathcal{L}_{\text{camera}} &= r_n \left[3.8 \left\langle s_f \rho_{0.01}(\log \widehat{f} - \log f^\star) \right\rangle + 4.1 \left\langle s_c \rho_{0.005}(\widehat{c} - c^\star) \right\rangle \right], \\ s_f &= \left\langle \left| \frac{u - c^\star}{f^\star} \right| \right\rangle_{\text{ray cells}}, \qquad s_c = \frac{1}{f^\star}. \end{aligned} \tag{C.2}$$

我们根据已标定内参计算灵敏度，并在 SmoothL1 之外应用灵敏度，记作 ρ_δ，状态转移为 δ，然后对有效相机及两个轴取平均值。我们还使用余弦距离监督得到的单位射线。RHD 仅提供关节和图像监督。对于相机损失和分布损失，我们在从零开始的优化器步 n 使用渐增计划 $r_n = \min(1, (n + 1)/500)$。

表 C.3：目标函数系数。 我们针对每个来源，分别对各分量在其有效目标上取平均值。相机项和分布项采用 500 步渐增系数 r_n。

目标	残差 / 分支	系数
相对手腕关节	直接 / MANO 平均绝对误差	10/10
相机坐标系关节	直接 / MANO 平均绝对误差	2.5/2.5
相机坐标系腕部	直接 / MANO 平均绝对误差	1/1
MANO旋转	测地线 / 旋转矩阵误差	1/1
MANO形状 / 平移	参数平均绝对误差	0。1/1
图像关节 / 手腕 / 投影	归一化坐标绝对误差	1/0。5/1
存在性 / 可见性	二元交叉熵	0.5/0.25
时间关节	有效三元组二阶差分	0.5
单位射线方向	余弦距离	1
对数焦距 / 主点	雅可比缩放的SmoothL1	$3.8r_n/4.1r_n$
辅助平移	方位路径 XY 绝对误差	1
视野外朝向	全局测地线 + 矩阵误差	1/32
联合读出分布	高斯目标 JS 散度	0。$1r_n$

我们使用 AdamW 优化 LoRA、图块嵌入、解码器和头部，同时冻结变分自编码器和基础骨干网络。表 C.1 列出了优化和损失渐增计划。在推理时，为减少长录制内容中的重复计算，我们对每个视频仅编码一次，并使用常驻模型对重叠的重建窗口进行批次处理。我们对选定的骨干网络和变分自编码器操作使用经过校准的 FP8 内核，对变分自编码器的输入、输出和潜变量投影使用 FP16。

C.1.4. 序列重建与全局一致性

独立的窗口估计可能会随着图像支持度的变化而引入形状和尺度漂移。我们将 ACE 的窗口级形状先验扩展到完整的连续片段，在重新生成几何结构之前，为每只手估计一个形状，并估计一个共享相机。我们首先在各自的源时间戳处融合重叠预测，并在旋转空间中合并旋转。

为降低不可靠帧的影响，我们根据手部存在 e_{th}、可见性v_{th}以及用于置信度、图像支持度和投影一致性的门控项 g_{th}，为每个形状估计赋予权重。随后，我们通过 Huber 重新加权来优化共享形状：

$$w_{th} = e_{th}^2 v_{th}^2 g_{th}, \qquad \widetilde{w}_{th}^{(k)} = w_{th} \frac{1.5}{\max(1.5, d_{th}^{(k)})}, \\ \overline{\boldsymbol{\beta}}^{h,(k+1)} = \frac{\sum_t \widetilde{w}_{th}^{(k)} \boldsymbol{\beta}_{th}}{\sum_t \widetilde{w}_{th}^{(k)}}. \tag{C.3}$$

我们使用加权中位数初始化每个参数，并使用固定的逐参数尺度 $\max(1.4826\mathrm{MAD}, 0.05)$ 对残差进行标准化；该尺度根据初始估计计算加权中位数绝对偏差。距离 $d_{(k)\,th}$ 表示当前共享形状下的标准化残差均方根。如果门控未能筛选出可靠帧，我们会在同一稳健估计中使用相等的基础权重。

对于预测相机，我们在源画布上对焦距对数和归一化主点采用相同的处理流程，使用每帧左右手权重中的最大值，并分别设置 0.01 和 0.002 的尺度下限。对于标定序列，我们使用提供的内参。

几何结构再生成与平移优化。 更新形状和标定会改变手部的图像投影。我们根据共享形状重新生成 MANO 几何，并针对二维锚点 $\mathbf{a}_{thj}$优化平移，在图像对齐与围绕重新生成点原始放置位置的先验 $\mathbf{Q}_{thj}$之间取得平衡：

$$\min_{\mathbf{t}_{th}} \sum_j \omega_{thj} \rho\Big(\big\| \pi_{\overline{K}}(\mathbf{Q}_{thj} + \mathbf{t}_{th}) - \mathbf{a}_{thj} \big\|_2 \Big) + \tfrac{1}{2} \Big\| \mathbf{t}_{th} - \mathbf{t}_{th}^{\text{prior}} \Big\|_{\Sigma^{-1}}^2 . \tag{C.4}$$

此处，ρ为 Huber 损失，ω 结合了锚点支持度和手部可靠性，Σ 规定了横向和深度先验尺度。我们在固定关节运动和旋转的情况下优化平移，强制正深度，并限制手腕深度的变化。当锚点不足时，我们保留原始投影的手腕位置和深度。
我们导出时间戳对齐的关节、MANO 参数、内参和置信度分数，用于第 2.5.3 节中的动作数据准备。图 C.2 总结了这一优化过程。

C.1.5. 定性示例

图 C.3(a) 展示了 EgoStandard、EgoDex 和 H2O 中双手操作的模型重建。面板 (b) 和 (c) 将这些重建结果与发布的标注进行了比较。在 OakInk2 示例中，发布的关节投影到了桌面上，远离可见的手部。在 EgoStandard 示例中，发布的右手关节投影到了部分遮挡的手部下方。在这两种情况下，重建的关节都与可见的手部更加吻合。这些示例说明，在数据整理过程中，应用数据集专属阈值评估标注与图像的一致性，有助于标记并过滤噪声监督。

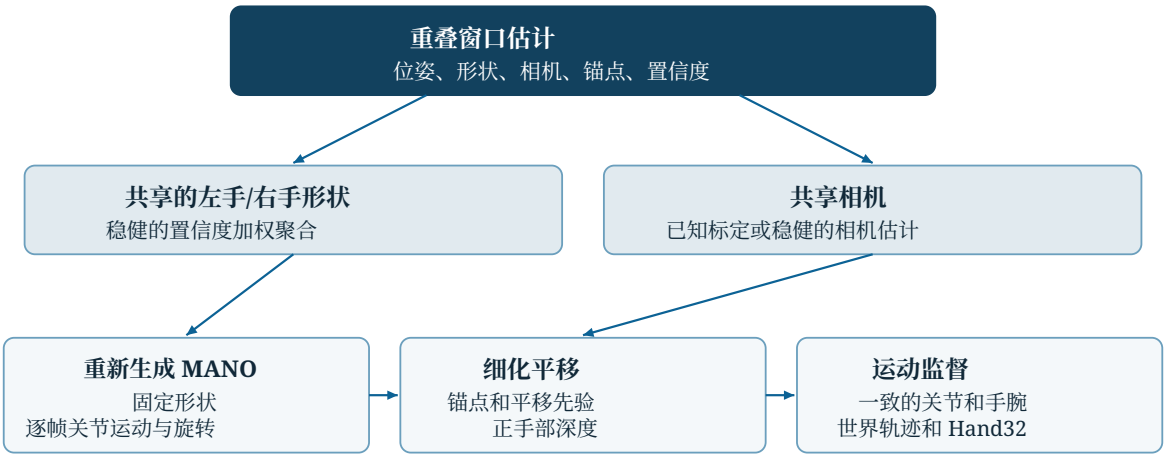


图 C.2　：序列级一致性。 我们将可靠的窗口估计汇总为共享的手部形状，并进行标定、重新生成几何结构，并根据图像锚点细化平移。

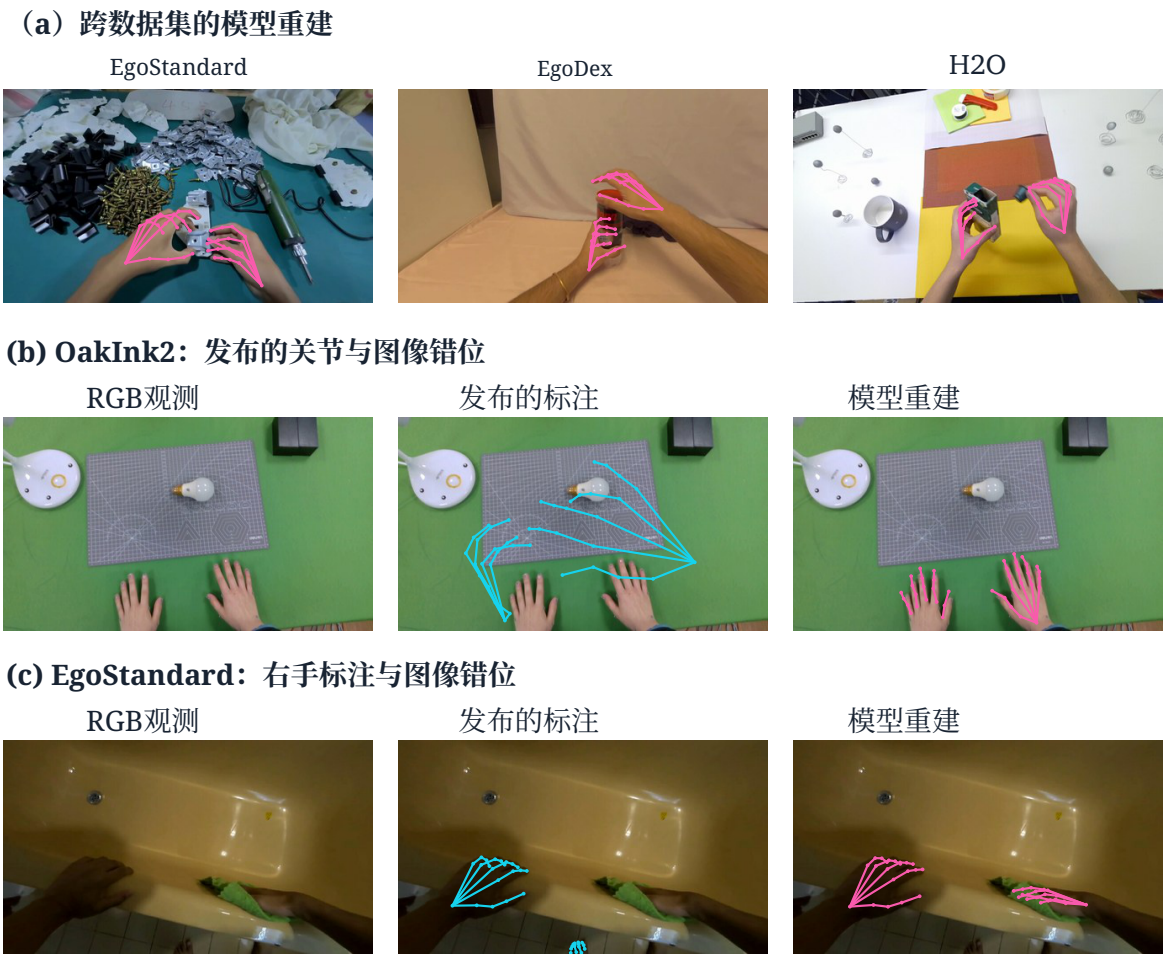


图 C.3：跨数据集的定性手部重建。（a）EgoStandard、EgoDex 和 H2O 上的模型重建。（b、c）从左至右分别展示 OakInk2 和 EgoStandard 的 RGB 观测、发布的关节标注（青色）以及模型重建（粉色）。关节叠加使用所提供的相机标定。

C.2. 灵巧状态与动作详情

C.2.1. Hand32定义

遵循 FAAS 的功能对齐原则 [29]， Hand32 按手指、关节层级和运动功能分配槽位。因此，不同的手部可以共用一种关节配置格式，而无需具有完全相同的机械结构：例如，即使机制不同，食指屈曲也可以占据同一槽位。表 C.4 列出了槽位定义。

表 C.4： Hand32 槽位定义。 条目为从零开始的槽位索引；角度以弧度为单位。**FE**表示屈曲/伸展，**AA** 表示外展/内收。**CMC**：腕掌关节；**MCP**：掌指关节；**IP**：指间关节；**PIP/DIP**：近端/远端指间关节。槽位 31 保留用于已声明的具身特定扩展。

手指	关节	FE	AA	扭转
拇指	CMC	1	2	26
	MCP	3	25	0
	IP	4	—	—
索引	MCP	7	6	5
	PIP	8	—	27
	DIP	9	—	—
中间	MCP	12	11	10
	PIP	13	—	28
	DIP	14	—	—
Ring	MCP	17	16	15
	PIP	18	—	29
	DIP	19	—	—
小型	MCP	22	21	20
	PIP	23	—	30
	DIP	24	—	—

C.2.2. 将灵巧手映射至 Hand32

我们将每只手的关节运动映射到 Hand32 ，依据的是动作功能，而非原生关节编号。多个机械关节可能共同对应一个参考坐标。有效性掩码用于标记不受支持的运动，特殊映射则保留每只手特有的运动。表C.5 列出了设备映射关系。

表 C.5： Hand32 代表性灵巧手的映射。 槽位列表给出状态覆盖率；影子动作另外排除槽位 25。特殊映射仅适用于所列出的手部。

Hand	已映射槽位	特殊映射
Sharpa Wave	1–4、6–9、11–14、16–19、21–25、31	31：小型腕掌关节
DexHand021	1–4 ，6–9，11–14 ，16–19 ，21–24	—
Allegro	1–4、6–9、11–14、16–19	四个功能性手指
Orca	1–4，6–8，11–13 ，16–18 ，21–23 ，31	31：内侧腕部
Revo2	2–4，7–8，12–13 ，17–18 ，22–23	—
RH56DFX	2–4，7–8，12–13 ，17–18 ，22–23 ，26	26：拇指额外屈曲
RH5DG2	2–4，6–9，11–14 ，17–19 ，22–24 ，26	26：拇指远端指间关节
Schunk	2–4、6–9、12–14、16–19、21–24、26、31	26：拇指铰链；31：耦合的掌侧运动
Shadow	1–4，6–9，11–14 ，16–19 ，21–25 ，30–31	30/31：内侧手腕
Wuji	1–4、6–9、11–14、16–19、21–24	—
能力	2–3、7–8、12–13、17–18、22–23	保留记录的模拟命令
LEAP	1–4、6–9、11–14、16–19	四个功能性手指

C.2.3. 动作学习数据源与 Ego128 布局

表C.6 列出了第2.5.2 节和第2.5.3节动作学习所使用的数据来源。配备仪器的人类和手套录制内容提供了额外的手部测量数据，以及在可获得时的触觉观测。

表 C.6：用于动作学习的人类与灵巧手演示数据源。 “筛选前”计数的是本地输入片段。“筛选后”计数的是指定发布版本中保留的训练片段，其中包括数据集划分选择的影响。小时数采用原始录制时间线计算，同步视角只计一次。在我们于2026年9月8日获取的快照中，EgoVerse仅限于短边严格大于480像素的人类录制内容。

	筛选前		筛选后	
来源	回合	小时 片段		小时
第一人称视角人类演示				
EgoDex [165]	314,839	721.93	303,391	700.80
EgoLive [171]	971,045	1,675.46	915,403	1,592.18
EgoPro [173]	175,284	1,035.85	170,476	1,011.70
EgoStandard [173]	1,185,239	5,687.81	1,160,797	5,589.46
EgoVerse [180]	89,690	101.74	80,406	94.47
Lightwheel（内部研发）	2,002,511	4,235.84	1,322,146	2,819.34
仪器化人类与手套示范				
Xspark-HumanTouch [181]	17,608	132.75	17,576	132.30
PaXini OmniSharing [182]	44,176	335.34	41,508	317.44
SynData（手套） [183]	36,780	145.46	35,303	120.39
灵巧机器人示范				
FTP-1 [18]	9,328	73.16	8,098	66.49
SharpaDex [184]	28,993	299.96	28,403	294.04
RoCo Task Board Assembly [185]	562	22.79	571	22.65
机器人折纸挑战赛 [186, 187]	1,845	101.39	1,821	100.12
T-Rex [132]	5,464	50.68	5,254	48.53
HRDexDB（机器人） [188]	1,630	9.20	1,592	8.15
Bench2Dex [80]	1,300	12.55	1,281	12.37
总计	4,886,294 14	,641.91 4	**,094,026 12**	**,930.43**

表 C.7：Ego128 状态与动作。索引从零开始，采用半开区间。手腕和相机位姿采用九维位置–旋转编码。手部关节目标是下一次连续观测时的绝对角度。

通道	**组**	**t 时的状态**	**来自 t 的动作**
$[0, 9)$	左手腕	pose9位姿编码(x_t^L)	pose9位姿编码(D_t^L)
$[10, 19)$	右手腕	pose9位姿编码(x_t^R)	pose9位姿编码(D_t^R)
$[20, 52)$	左 Hand32	q_t^L	q_{t+1}^L
$[52, 84)$	右侧 Hand32	q_t^R	q_{t+1}^R
$[84, 93)$	主摄像头	pose9位姿编码(s_t^c)	pose9位姿编码(D_t^c)
9, 19, $[93, 128)$	保留	零值、非激活掩码	零值、非激活掩码

Ego128 表C.7 中的状态和动作布局使用公式2定义的相机相对腕部和锚定相机量；腕部目标 D_t^h 遵循公式1 中规范手腕处的局部增量约定。每个位姿编码为 $\text{pose9}(T) = [\mathbf{p}; \mathbf{R}_{:,1}; \mathbf{R}_{:,2}]$，其中平移单位为米；手部

关节运动单位为弧度。状态记录当前关节运动，动作记录下一次连续观测时的绝对关节运动。

D. 触觉感知

D.1. 触觉数据处理详情

D.1.1. 表面校准与数据源准备

校准工具。 算子首先通过全手相似变换将数据源几何对齐到中性 MANO，然后在表面贴合前调整每个传感器模块的位置、方向和几何尺度。检查时需确认手指/指骨身份、掌侧/背侧放置，以及中性和关节姿态下的局部方向。PaXini 每只手使用 15 个模块和 1,005 个样本；HumanTouch 每只手使用 360 个已分配的图谱样本。源像素或触觉像素保留其原始标识符。

每个附着点都存储一个三角形 f_i、重心权重 $\boldsymbol{\beta}_i$ 和中性接触坐标系 Q_i^0。位姿下的位置为

$$x_i(t) = \sum_{k=1}^{3} \beta_{ik} V_{f_i,k}(t), \qquad \sum_k \beta_{ik} = 1, \quad \beta_{ik} \geq 0. \tag{D.1}$$

切向量以 $A_t A_0^{\dagger} q_x^0$ 的形式传输，使用中性/姿态三角形的边矩阵，然后针对姿态法向量进行正交化。生成的 Q_i 包含切向 X、切向 Y 和法向 Z 列；局部向量按 $Q_i g_i$ 映射。自动检查会验证唯一标识符、有效三角形、重心重建和 $Q_i \in SO(3)$。左右绑定会分别审核。几何结构、拓扑和校准哈希用于标识缓存。

数据源准备。表D.1 列出了特定数据源观测。仅法向响应 a_i 构成 $a_i n_i$ 沿其映射法向量的响应。Sharpa Step4f-v2 缓存会减去首帧基线，保留至少为 8 的正残差，要求活动区域达到 0.5%，并在连续四帧确认接触后进行离线起始时间回填；释放则立即生效。这些设置仅适用于该缓存版本。覆盖率、接触和特征有效性分别存储；只有已覆盖且无接触的区域才会提供负存在标签。

表 D.1：具有代表性的触觉数据源适配器和覆盖规则。信号解释、表面对应关系和覆盖率分别记录。

来源 / 布局	映射观测	覆盖率处理
PaXini 手套	局部方向阵列响应；已校准表面锚点	分离左右标识；保留手掌和手指绑定-关系
HumanTouch 图谱	相对法向响应	每只手分配 360 个样本；不含填-充
Sharpa FTP	光学指尖响应	拇指／食指／中指；无名指／小指无效
RoCo / Origami / T-Rex	来源指尖形变流	已声明的手部/手指瓦片及映射版本-sion
Bench2Dex	表面对齐的几何接触信号 [80]	手部专用感知表面和接触有效-性
单个触觉垫	有来源支持的触觉垫观测	指定的食指区域角色
UniVTact双颌布局	两个分别映射的感知面	右手拇指/食指角色；其他区域掩码

D.1.2. 代理构造与编码

对于每个区域，固定原点 o_r、右手基 B_r 和尺度 $l_r > 0$。位置、UV 坐标和载荷均为

$$y_i = B_r^{\top}(x_i - o_r), \qquad u_i = (y_{i,x}, y_{i,y})^{\top}/\ell_r, \qquad f_i = B_r^{\top} f_i^M. \tag{D.2}$$

UV 无量纲；面积和力臂采用三维几何结构。表 D.2 列出了完整布局。

表 D.2：R　区域代理项布局。 索引范围从零开始且采用半开区间。特征有效性单独记录 ly.

索引	宽度	符号	定义
[0, 2)	2	c	几何 UV 中心
[2, 4)	2	$\bar{c}$	载荷加权 UV 中心
[4, 7)	3	F	局部合力
[7, 10)	3	τ	关于载荷加权三维中心的力矩
[10, 11)	1	A	有效支撑面积
[11, 14)	3	b	稳定定向范围与各向异性编码
[14, 20)	6	对称向量化(C)	牵引协方差
[20, 23)	3	对称向量化(S)	载荷加权 UV 协方差
[23, 26)	3	$t_{\text{峰值}}$	峰值牵引向量
[26, 28)	2	$u_{\text{峰值}}$	峰值 UV 位置

对于有效接触样本 I、支撑面积 $a_i > 0$ 和 $w_i = \max(\|f_i\|_2,\ \epsilon)$，计算

$$\begin{aligned} c &= \frac{1}{|\mathcal{I}|}\sum_i u_i, \quad \bar{y} = \frac{\sum_i w_i y_i}{\sum_i w_i}, & \bar{c} &= (\bar{y}_x, \bar{y}_y)^\top / \ell_r, \\ F &= \sum_i f_i, & \tau &= \sum_i (y_i - \bar{y}) \times f_i, \quad A = \sum_i a_i. \end{aligned} \tag{D.3}$$

此处 f_i 是样本载荷；校准压力/牵引会先乘以物理支撑面积。顶点支撑可采用关联三角形面积的三分之一，或在可用时采用校准原生面积。未校准输入保留源响应单位。

主 UV 轴给出范围 $L \geq W$ 和角度 θ （模 π）。稳定形状描述符为

$$b = [L + W,\ \eta\cos(2\theta),\ \eta\sin(2\theta)], \qquad \eta = (L - W)/(L + W). \tag{D.4}$$

单点或零范围接触使用零形状描述符。对于牵引 $t_i = f_i/a_i$ 和 $\mu_t = \sum_i a_i t_i / A$，

$$\begin{aligned} C &= A^{-1}\sum_i a_i (t_i - \mu_t)(t_i - \mu_t)^\top, \\ S &= (\sum_i w_i)^{-1}\sum_i w_i (u_i - \bar{c})(u_i - \bar{c})^\top. \end{aligned} \tag{D.5}$$

打包顺序为（xx、yy、zz、xy、xz、yz），对于 C则为（uu、uv、vv）。峰值为 $i^* = \arg\max_i \|t_i\|_2$，存储为（$t_{i^*}$、$u_{i^*}$）。空接触使用零代理量存储，并禁用接触属性损失。

数值编码。 离线源归一化使用训练集划分统计量，并对耦合的 XYZ 分量使用共享尺度。代理编码还使用全区域样本数、空间尺度和支撑面积，对合力、力矩和几何字段进行缩放。归一化器标识和逆变换参数会被记录。在稠密场裁剪后计算的代理量不一定能够仿射逆变换为原始代理量，因此两种视图都会保留。训练读取器会验证已发布的值和时间戳，而不会重复进行标定或归一化拟合。

D.2. 触觉建模细节

D.2.1. 输入和时间对齐

输入结合了相互独立的代理量/存在性投影，以及来源和解剖元数据：

$$\begin{aligned} H^{(0)} = {} & E_p(p_\sigma \odot M_p^{\text{in}}) + g_z E_z(z_\sigma \odot M_z) \\ & + e_{\text{sensor}} + e_{\text{normalizer}} + e_{\text{part}} + e_{\text{role}} + E_s(s, M_s) + e_{\text{modality}}. \end{aligned} \tag{D.6}$$

可学习增益初始值为 $g_z = 0$。1；role 用于区分历史记录/目标。部件 ID 为 $r + 64h$，其中从拇指到小指的编号为 $r = 0, \ldots, 4$，$h = 0,1$ 表示左侧/右侧。投影路径和归一化器 ID 相互独立。
配置使用每种可用模态的九个实值观测，目标时域为 1.2 秒。目标行遵循半开区间；锚点行可因不同角色同时出现在上下文和目标中。缺失历史记录会被拒绝，填充目标无效。不同的原生速率会产生不同的历史时长。加载器使用带浮点时间戳和空间旋转坐标的公式 (8) (r, h)。历史姿态上下文使用相应的状态；目标上下文使用当前状态。

D.2.2. 损失实现

代理项和存在性分别按各自的有效元素计数进行归约：

$$\mathrm{MSE}_M(\hat{v}, v^*) = \frac{\sum_j M_j(\hat{v}_j - v_j^*)^2}{\max(1, \sum_j M_j)}. \tag{D.7}$$

无效条目在噪声插值前后均保持为零。辅助二元交叉熵读取使用带分离隐藏状态的存在性解码器：

$$\ell_\sigma = 2\big[z_\sigma - \sigma D_z(\mathrm{stopgrad}(H_\sigma))\big]. \tag{D.8}$$

因此，二元交叉熵会更新该解码器，而流损失则训练普通编码器/Transformer/解码器路径。在公式 (7) 中，$\lambda_z = 0.25$ 和 $\lambda_{\text{二元交叉熵}} = 0.02$。另有 $0.05 L_{\text{geo}}$ 用于惩罚重建的 UV 中心：

$$L_{\text{geo}} = \frac{\sum_{j \in \mathcal{U}} M_{p,j}[\max(|\hat{p}_{0,j}| - 1, 0)]^2}{\max(1, \sum_{j \in \mathcal{U}} M_{p,j})}, \quad \hat{p}_0 = p_\sigma - \sigma \hat{v}_p, \quad \mathcal{U} = \{0, 1, 2, 3\}. \tag{D.9}$$

该惩罚是软性的。存在性门控仅在最终读取时应用一次，同时保留原始输出用于诊断。

E. 基础设施细节

E.1. 实验条件

这些图展示了三个塔中的四种标记类型（U/V/A/T）。动作与触觉生成塔将 A 放在 T 之前。历史测量未使用触觉令牌（$T = 0$），且涵盖不同的工作负载和实现快照；这些增益不能合并为已发布模型的加速比。表 E.1 列出了训练对比，表 2 列出了利用率配置，表 E.3 列出了推理协议。吞吐量窗口不包括启动和编译；训练时间不包括离线预处理和存储。

96K 配置。生产配置在 128 个 GPU 上使用 shard16/replica8/GA2。历史 rank-0 汇总报告称，64K 离线配置约为 18.5 个样本/47 秒，96K 离线配置约为 15 个样本/35 秒：约为 0.394 → 0.429 个样本/秒/GPU，吞吐量提高了 8.9%。这些阶段汇总数据经过四舍五入；未保留精确的平均窗口。后续的 FA3 基准测试将 96K 配方调整为适用于八个 GPU、shard8/replica1、只读检查点器以及记录在案的内核构建。在相同的 U/V 容量下，峰值分配内存为 64.13–64.22 GiB；这并非匹配的 shard8/shard16 容量消融实验。另一个编译器修复将归约拆分大小设为 32：发生故障的分块需要 278,688 字节，超过了 232,448 字节的共享内存限制。

表 E.1：训练研究与实现范围。所有 GPU 均具有 80 GB 记忆。可选实验并不能证明这些路径训练出了已发布检查点。U/V/A/T 分别表示理解/视频/动作/触觉能力；在这些历史研究中，T 为零。

研究 / 角色	工作负载与实现	测量范围
静态执行 / 配置的训练路径	三塔式 DROID， `width_only_1b`，不使用状态条件化。冻结的世界理解塔；SDPA，路由投影；PyTorch 2.10/CUDA 13.0	八个 GPU，在线 VAE，BF16，完全重计算，全局批次大小为 32；每种配置运行三次，更新 11–40； 表 E.2
分组 FA3 / 实验	文本转视频，8 个 GPU，分片8/副本1，GA2，U3,072/V98,304；PyTorch 2.10/Triton 3.6。内核源代码 `6c9c171a`。测试修订版本 `f6447a5d`	同分支拆分/分组比较，每项运行一次，预热十次并进行三十次计时更新；rank-0 内核计数来自两次单独的性能分析更新
ODC / 实验	文本转视频、八个 GPU、分片8、GA2、在线 VAE、64K；快照 `016ac4e6`	每个机械臂运行一次 100次更新运行；计时范围为 51–94，跟踪记录覆盖 96–98，使用八个 GPU
重新分配 / 原型	文本转视频，八个 GPU，副本4/分片2，GA4，离线变分自编码器；U512/V32,768	每秩/窗口四个包，0.2 秒搜索预算；十七次队列匹配更新，排除三个
共享加载器 / 在测试配方中采用	动作、8 个 GPU、23 个源、相同的 i8800 权重/Wan2.2 VAE，种子42；74,368 个 token 的预算，shard8/GA2	完整更新；W&B 和检查点禁用写入。编译、重复、全列以及后台控制，如图 8
离线潜变量 / 文本转视频路径	四个图形处理器，分片 4，BF16；U19,456/V25,600；在线/离线目录	两次捕获的 rank-0 更新步；展示变分自编码器路径移除，而非匹配的端到端加速比

表 E.2：图 6 中的静态执行消融图 6。全局批次大小固定为 32。固定容量行使用 U/V/A/T 容量，分别为每包 512/2,688/128/0 个令牌。时间为三次运行平均完整更新耗时的中位数和范围。工作进程按秩计数。预取是每个工作进程的因子。变体名称未指定模型参数总量。

阶段	Pack	GA	容量	编译	工作线程 / 预取	s/更新 [范围]
基线	1	4	动态	Off	8/2	3.320 [3.305，3.323]
打包	4	1	动态	Off	8/2	1.429 [1.420, 1.429]
+ 固定	4	1	固定	Off	8/2	1.042 [1.037, 1.045]
+ 编译	4	1	固定	静态	8/2	0.871 [0.866, 0.873]
+ 工作进程	4	1	固定	静态	18/2	0.819 [0.811，0.822]

表 E.3：历史三塔推理协议。这些协议不包含状态和触觉输入，并且不同于当前的 Ray 后端。CFG 并行性研究采用等权重的输入/区块 p50 汇总以及插值基线；GPU 成本使用平均时长。其包含 20 次观测、每次重复两次的质量检查独立于计时。

研究	输入与执行	计时器/重复次数
CFG	80 GB 图形处理器； `width_only_1b`； 960×640，41 帧/40 个动作，6,600 个视频令牌；五个个采样步，分类器自由引导系数为 1.5；仅动作	从图像准备到在中央处理器上交付动作；三个 ABA 块、两个输入、三次预热和每种输入配置二十个计时请求；不包括启动/HTTP
前馈网络复用	三组计时输入；独立的六任务质量集；单个 GPU、双 CFG 分支、十次去噪器评估	准备完成后的同步执行，不包括准备/HTTP；两个会话，二十四个 AB/BA 配对/输入/会话
队列	四个单图形处理器副本；128 个请求/配置/速率；固定间隔到达，客户端并发数为八，八秒截止期限	每个分配的图形处理器在墙钟时间内的及时响应数从启动到最终排空；进行一次有限重放每种配置/速率

E.2. 执行机制与权衡

E.2.1. 分组注意力

上游 FA3 变长接口在每次调用时都需要单独提供查询和键的累积偏移量，以及一个因果设置 [189]。因此，因果注意力和全注意力需要分别调用。我们保留这些偏移量，并传入一个设备驻留的因果标志向量，每个分段对应一个标志。U 查询以因果方式关注 U，生成查询则关注同一样本中的所有有效上下文。描述符会打包 $G = V + A + T$，并在动作与触觉生成塔中保持 A 位于 T 之前。模式感知排序会改变瓦片调度顺序，同时保留分段标识（图 E.1）。

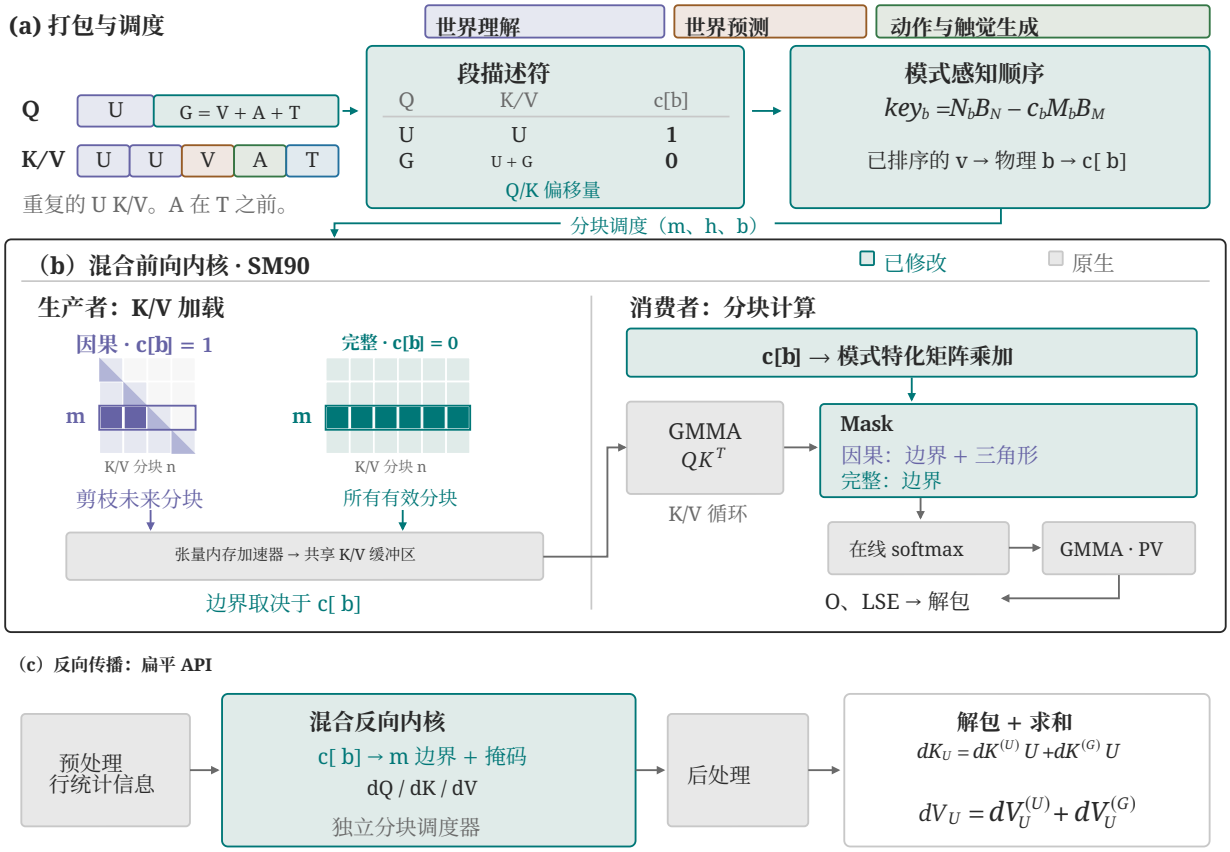


图 E.1：分组 FA3 内部结构。(a) U 和 G =V + A + T 构成因果分段和全注意力分段；A/T 共用动作与触觉生成塔的注意力投影，其中 A 在前。标志 c[b] 用于标识物理分段 b。模式感知排序期间会保留 Q/K 偏移量；M/N 统计查询/键瓦片数量，B_M/B_N 则给出瓦片尺寸。(b) 持久化调度器会调度查询瓦片 m、头部 h 和分段 b；一个前向主内核使用特定于模式的边界和掩码。(c) 全梯度扁平 API 使用混合反向主内核和独立的调度器。预处理/后处理仍然彼此独立；框架自动微分会对重复的 U 键和值的梯度求和。瓦片网格仅为示意图。

前向传播使用一个主内核，并采用特定于模式的分块边界和掩码。反向传播使用独立的混合调度器，并为重复的 U 键/值累积梯度；打包、准备和梯度转换仍然彼此独立。在历史 T2V 跟踪中，前向主内核启动次数减少576 → 288，反向传播启动次数 288 → 144。一项未进行性能分析的比较报告显示，中位数为 41。410 →41。311 秒/次更新：尽管调度次数更少，较长的生成注意力仍占据了大部分运行时间。

E.2.2. 按需通信

FSDP 参数全聚合要求各参与方进入同一集体通信。我们将按需通信（ODC） [47] 与 FSDP2 和同步梯度累积集成，使就绪的工作器无需等待每个所有者的模型工作进程，即可读取已发布分片。本地节点内的 CUDA 进程间通信迁移仍需要

复制完成。梯度发送方写入收件箱并向 CPU 辅助进程发出信号，后者安排累积，并在缓冲区可以重新使用时发送确认；NVSHMEM 提供控制信号 [190]。优化器仍会等待所有预期的贡献量、副本归约和梯度裁剪（图 E.2）。

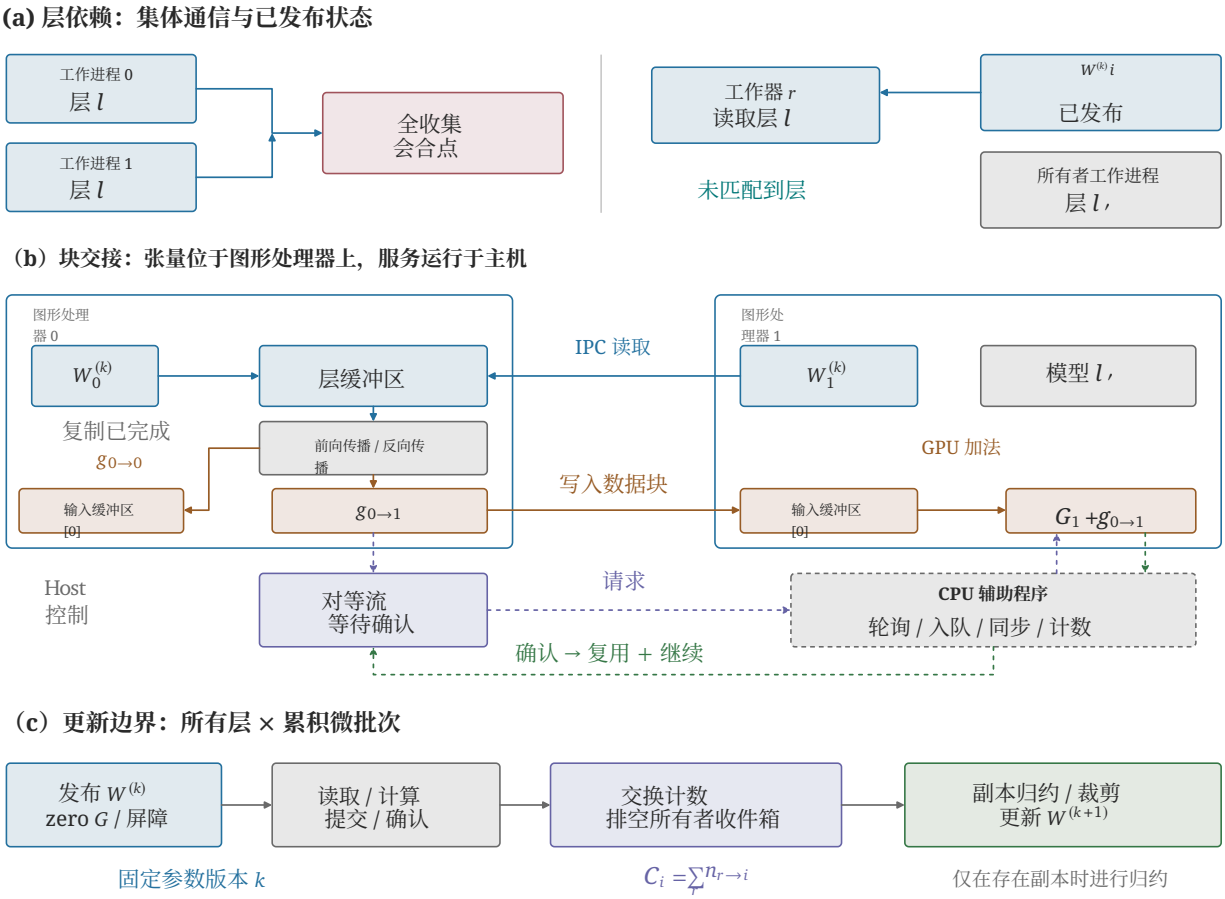


图 E.2：两种尺度下的 ODC 依赖关系。(a) 一次集体通信需要层 l 上的两个工作器；对等读取需要所有者已发布的分片，而不需要其在 l 层的工作器。(b) 同节点交接将 GPU 缓冲区与 CPU 辅助程序分离。$W_{(k)\,i}$ /G_i 是所有者 i 的参数/梯度分片；$g_{r\to i}$ 是发送方 r 的贡献量。本地贡献使用相同的收件箱协议（省略辅助程序 0）。请求/确认携带控制信息；实线箭头携带张量。(c) 更新 k 使 $W_{(k)}$ 在各层和 GA 微批次中保持固定；提交计数 $n_{r\to i}$ 必须在优化器运行前与已完成计数 C_i 匹配。几何结构为示意图。

表 E.4 总结了更新 96–98 期间的通信变化。在八个 GPU 上一个单独的未进行性能分析的窗口中，采用在线 VAE 的 shard8/GA2 每次更新耗时25。21 → 27。36 秒，覆盖更新 51–94（吞吐量–7。84%；每个机械臂运行一次）。收件箱处理、请求和完成操作会增加开销；无法从墙钟时间中减去重叠的跟踪时长来估算净节省。ODC 消除了逐层会合，但此工作负载运行速度更慢，且尚未通过数值验收（附录 E.5）。

表 E.4：ODC 通信变化。 更新 96–98 期间，八个图形处理器的跟踪事件总数。这些计数描述的是协议变化，而非节省的关键路径时间。

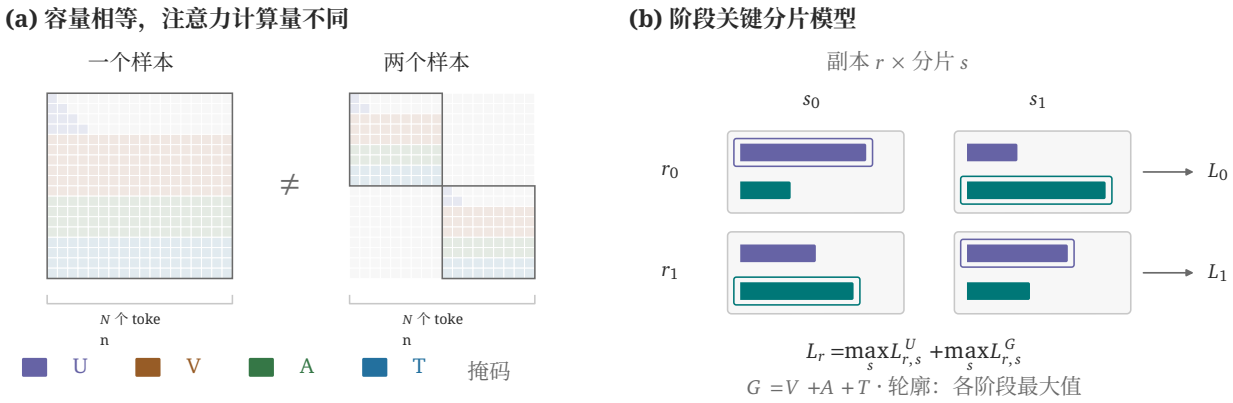
事件	原生 FSDP	ODC
参数全聚合	3,536	80
分片规约-分散	1,776	0
全规约	824	841
NVSHMEM 请求/等待	0 / 0	41,856 / 41,856

E.2.3. 工作负载重新分配

在物理容量固定的情况下，有效注意力计算量取决于样本布局（图 E.3）。对于样本 p，定义 $G_p = V_p + A_p + T_p$；当 U 为因果模式且生成查询完整时，可见查询-键对的总数为

$$\mathcal{A} = \sum_p \left[\frac{U_p(U_p+1)}{2} + G_p(U_p + G_p) \right].$$

在缓冲区大小固定时，稠密投影和前馈神经网络执行的计算量基本固定。我们的原型保持 HSDP 网格不变，并根据 $\sum U_p^2$、$\sum G_p^2$ 和 $\sum U_p G_p$ 拟合阶段开销，与 Hydraulis [48], ByteScale [49]，和 DistTrain [50] 等并行度和调度方法形成补充。



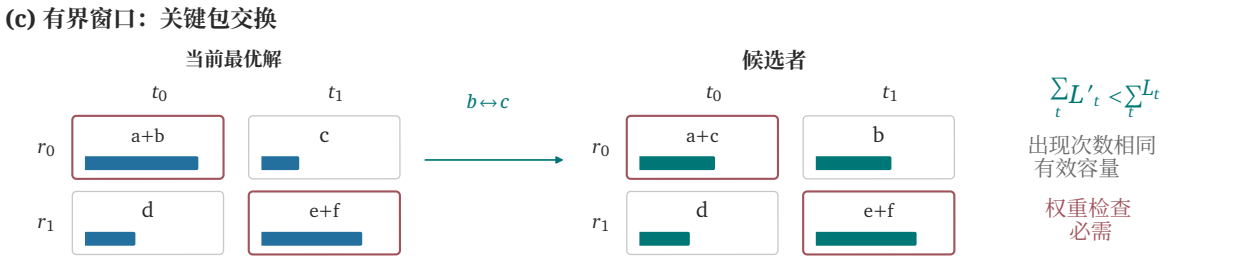


图 E.3：成本感知重新分配。(a) 容量相等的包具有不同的可见 U/V/A/T 注意力面积，顺序如图所示。(b) 阶段开销模型：每个副本 r 在每个成本组成部分中都受其最慢分片 s 的限制；$G = V + A + T$ 表示生成计算量，而非单独的注意力操作。(c) 两个副本各有一个分片：列为微批次 t，行为秩，a–f 为样本，条形表示预测成本。交换 b/c 可降低列最大值之和 L_t，同时保留出现次数和容量；红色轮廓标出最大值。损失权重不变性需要单独检查。所有几何结构均为示意图。

从顺序打包开始，各秩全收集完整的 Python 样本物体，并在时间限制内重复执行确定性交换搜索。对于微批次 t、副本 r 和分片 s，拟合得到的可变成本目标函数为

$$\sum_t \max_r \left[\max_s L^U_{t,r,s} + \max_s L^G_{t,r,s} \right].$$

打包延迟、汇总的副本负载和重排序距离用于打破平局。检查会验证容量和样本出现次数；各秩之间以及梯度累积中的等效样本/损失加权尚未验证。

文本转视频实验使用 $A = T = 0$、8 个 GPU、replica4/shard2、GA4、离线变分自编码器，并在每个窗口中为每个秩分配 4 个包。在 17 次队列匹配更新中，核心耗时减少 20.7519 → 20.4973 秒（吞吐量提升 +1.24%）。每个窗口的协调耗时为 0.2104 秒，完整更新吞吐量因此仅提升 +0.18%（$p = 0.468$；节省时间的 95% 置信区间为 [−0.0685,0.1426] 秒）。由于队列不同，有 3 次更新被排除；留出成本预测的 p 值为 $R^2 = -0.097$。协调开销和分配质量限制了该方法的采用。

E.2.4. 计算与内存策略

仅注意力重新计算可避免重复进行前馈网络计算，但会保留更多激活值 [46]。 表 E.5报告了这种权衡，同时介绍了世界预测塔多层感知机中的实验性 FP8 计算和变分自编码器编译。这些研究采用了不同的工作负载和计时窗口；它们的收益不能累加，而且 FP8 对模型质量的验证仍未完成。

表 E.5：在八个 GPU 上开展的独立研究。重新计算使用匹配的输入/RNG 指纹；每次运行包含十次预热和三十次计时更新。FP8 质量验证仍未完成。

策略 / 测量的工作负载	结果	成本与范围
仅注意力重计算 / 文本转视频	29.821 → 25.979 秒/次更新	峰值分配内存 47.30 → 64.50 GiB；每种模式运行三次
世界预测塔中的 FP8 MLP/文本转视频	45.944 → 42.964 秒/次更新	已分配内存 64.28 → 66.90 GiB；每种模式运行一次
提前 VAE 编译 / 动作	12.173 → 11.007 scompute	四次计时更新； 不包括输入、迁移和编译

E.3. 补充测量

输入供给。 编译加载器比较采用每源基线，每个 rank 配置 23 个工作进程和预取因子 2（每个 rank 有 46 个传统配额）。共享池每个 rank 配置 4 个工作进程，预算为 $K = 16$。吞吐量变化 +0.63%。独立重复实验将内存减少 52.13%/51.94%，吞吐量变化分别为 +2.73%/−0.21%。读取所有列可将内存减少 27.71%，吞吐量为 +2.74%。后台传输变体每个 rank 使用 23 个工作进程和 $K = 8$。两者都固定内存；较新的 58 来源方案则禁用固定内存，以避免锁定已准备好的窗口。列投影、有界读取、寻址前缀复用和选择性 RGB 转换进一步限制了实体化。

图 E.4(a) 扩展了简短的加载器比较。对于十二个源，按每个源的工作进程数和预取因子配置为 (1、1)、(2、1) 和 (1、2) 时，平均耗时分别为 21.7407、20.9629 和 20.9606 秒/次更新，输入等待 p95 分别为 3.0040、0.0182 和 0.0108 秒。尾部延迟大幅降低，但平均更新时间仅改善约 3.6%。这里是三次运行，而非 600 次独立重复实验。后台传输变体具有匹配的输入轨迹哈希；其记忆成本与主要结果一并报告。

FFN 策略与开销。 世界预测塔会在第 2/4/6 次去噪器评估时缓存第 1–18 层和第 31–36 层的 FFN 输出，并在第 3/5/7 次评估时重复使用（层编号为 1–36，评估编号为 0–9）。其他层以及第 0/1/8/9 次评估均照常运行。此举跳过了 720 次双 CFG FFN 调用中的 144 次（20%）。注意力和动作与触觉生成塔仍会运行。每个请求为每层和每个分支分别使用 BF16 缓冲区，缓存占用6，600 × 4，096 × 2 × 24 × 2 字节（2.42 GiB），不包括分配器开销和其他激活值。图 E.4（b–d）和表 E.6分别报告计时、开环误差和队列效应。这些加速比不能相加。

表 E.6：推 理资源与质量成本。质量区间在每个固定任务内对回合进行重采样，从而保留 对种子的依赖；每个任务仅有两个回合。开环误差并不能证明任务成功 ing s.
观测/

数量	基线 / 总体	替代方案 / 变化
CFG 延迟 / GPU 成本	单个 GPU 2.004 秒，分配的 GPU·秒/请求为 2.004	双 GPU CFG 并行 1.054 秒，2.108 GPU·秒/请求
FFN 专家错误变化	等任务均值；六任务 回合聚类自助法	平移 −0.51%；旋转 +3.42% [2.98%、3.87%]；夹爪开口 +0.31%
FFN个体尾部	72 对生成对/指标	夹爪平均绝对误差：7 个变差 >10%； 最差 +21。03%

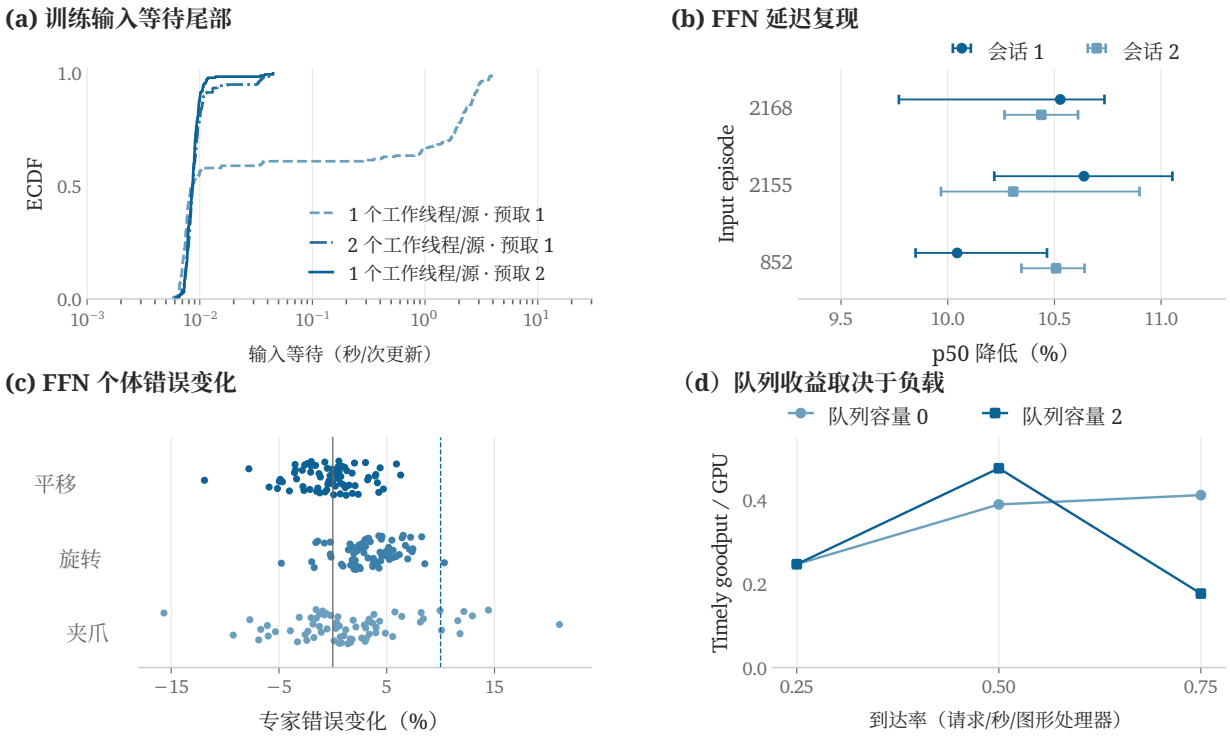


图 E.4：补充证据。(a) 十二来源、八张图形处理器的加载器研究：每种配置运行一次，进行二十次预热和 200 次计时更新。每个秩的等待时间为 GA2 微批次的总和，然后取所有秩中的最大值。图例标示每个来源的工作进程数，并给出每个工作进程的预取因子。(b) 会话 1 和 2 中的 FFN 延迟，每个会话使用三组相同输入，每组输入进行二十四次配对。时间间隔为 95% 配对循环区块 4 自助法区间。(c) 来自六项任务、十二个回合、三个随机种子的七十二组配对生成/指标； +10% 为任务级参考值。(d) 每个副本的队列容量设为零或两个待处理请求；每种配置/速率下使用四个副本和 128 个请求。

表 E.7：基于跟踪数据的形状上的 FlagGems 算子基准测试。 VLM 指视觉语言建模，T2V 指文本转视频生成。在世界动作模型（WAM）模式下，模型会生成动作块。启动次数和时间占比指预热后秩 0 上计算内核的情况。加速比是历史 FlagTree/PyTorch 2.8.0a0 技术栈上参考延迟与 FlagGems 延迟的比值；如有数据，则给出不同测试形状下的范围。T2V 在四个图形处理器上运行，每个图形处理器使用四条 WorldArena 提示（736 × 544，50 个 UniPC 步数，CFG 引导系数为 5 ）。其注意力调用次数包括 7,200 次初始化启动。具有代表性的查询和键值形状分别为 (55，131，32，128) 和 (66，503，8，128)。

工作负载	算子	启动次数	时间占比	加速比	跨形状
VLM	均值归约	4,640	3.443%	1.25×	1.23–1.27×
	前向注意力	1,296	0.244%	1.424×	—
	反向注意力	1,620	0.554%	1.397×	—
	索引写入（`index_put_`）	1,125	0.048%	2.170×	—
T2V	前向注意力	14,400	34.043%	1.40×	1.39–1.41×
	索引写入（`index_put_`）	800	0.112%	2.97×	2.95–3.00×
	求和归约	400	< 0。001%	1.44×	1.02–1.86×
	散射加法	100	0.040%	1.007×	—
WAM	前向注意力	1,728	26.022%	1.425×	1.38–1.47×
	融合式 RMSNorm（前向传播）	864	0.236%	1.03×	1.01–1.04×
	分块与拼接（`chunk_cat`）	222	0.052%	1.16×	1.06–1.26×

用于数据描述生成的结构化 JSON 生成。FlagOS 推理框架 [53]在推测解码期间对批次进行语法掩码计算，并使用经过优化的混合专家内核。在所评估的平台上，其每设备吞吐量达到相应基线模型的 1.4–2.2 倍。

E.4. FLOP 统计

经审计的 T2V 解码器有 36 层，宽度为 4,096，包含 32 个查询头/8 个 KV 头，头部宽度为 128，FFN 宽度为 12,288。对于经审计的在线 T2V 运行，有效 FLOP 包括对真实标记进行解码器前向/反向传播，以及所需的冻结 VAE 前向传播。不计入激活重计算和填充。使用秩 0 解析 FLOP F_i、完整更新耗时 t_i 和稠密 BF16 设备峰值 $P = 989 \times 10^{12}$FLOP/s，

$$\mathrm{MFU}_{\mathrm{useful}} = \frac{\sum_i F_{i,\mathrm{useful}}}{P \sum_i t_i}.$$

表 E.8 重建了更新 52–728 期间的结果。仅对最后一个秩 0 微批次进行采样，并按 GA4 缩放；这既非全秩平均值，也非对所有微批次的详尽统计。反向传播 FLOP 使用解析的 2× 前向传播近似值。分子包括测量更新期间所需的在线 VAE 前向传播；仅 Transformer 结果不包括该项。完整更新间隔时间包括数据、通信和优化器开销。

表 E.8：经审计的在线 T2V 配置的有效工作模型浮点运算利用率，见 表 2：677 次更新，35,350.814 秒。 两行均不包括重新计算和填充，且均为分析估算，而非硬件计数器测量值。

所需模型工作量	窗口内每 GPU FLOP	MFU
真实标记 Transformer 前向/反向传播	8.83882×10^{18}	25.28%
高于 + 在线冻结 VAE 前向传播	1.08533×10^{19}	**31.04%**

E.5. 正确性与恢复

表 E.9 报告内核和完整更新一致性。 表 E.10 报告已保存状态恢复和数值延续。即使原生重复运行结果有所变化，候选失败仍未解决。在这些研究中，分组 FA3 和 ODC 均未能证明其长期训练质量。

表 E.9：数值验证。 FA3 更新审计对所有八个 FSDP 分片中的第 0/17/35 层及输入/输出头进行抽样，并使用匹配的输入/随机数生成器和抽样的初始参数；所报告的归一化值并不涵盖完整模型。

比较	结果	解释
分组 FA3 算子	真实 QKV 重放：最大值 输出/源梯度相对 L2 0.002299， 低于 0.03	算子一致性通过了测试的 案例，包括 GQA、填充和 源梯度恢复
FA3 原生重复	裁剪梯度 / 更新相对L2范数： 0.8798 / 0.6825	基线更新变异性 仍未解决
FA3拆分为分组	裁剪梯度 / 更新相对L2范数： 1.301271 / 0.838888	未通过更新级数值验收
ODC / 原生	固定输入梯度相对L2范数 1.2007%； 原生重复 1.2847%；独立后端 比较 2.1583%	后端差异超过 2% 阈值； 严格的逐元素检查也会失败

使用较少的 GPU 恢复训练时，会恢复到保存的边界；检查点之后的工作可能会重放，全局批次或交错处理方式也可能发生变化。独立的状态条件输入测试可验证干净输入路由，以及输入掩码与损失掩码之间的区别；修正这一约定并不能验证此前训练的权重。

表 E.10：恢复验证。已保存游标跟踪的是已消耗的数据，而非已完成的预取操作。CP/TP/PP 分别表示上下文并行、张量并行和流水线并行。

边界	证据	范围
精确恢复	每个 10,149 个已保存的张量/元数据项 秩和九条输入/计划/随机数生成器记录 与其参考值匹配	已在保存状态边界处验证； 后续的数值等价性 尚未确立
160 → 144 GPU	58 个源 × 192 原始逻辑分片 = 11，保留了 136 个状态； 144 个秩各自都通过了前十项 微批次序列检查	离线所有权重新映射；CP/TP/PP=1, 指纹/处理过程均未变化。 逻辑分片保持独立 物理秩数量
数值延续	三次更新的采样 FP32 最大值 差异 1。24×10^{-5}–2。40×10^{-5}	10^{-5} 未达到阈值；内存回滚 也会失败

F. Innata-0-Agent 细节

F.1. 具身工具集的实现

我们实现了一套通用且可扩展的具身工具集，供验证和恢复共用。这些工具根据其在闭环交互中的作用，分为四个功能组：观测、感知、运动调整和原子技能。观测和感知工具不会推进环境，而运动调整工具和原子技能则通过原生机器人或模拟器接口执行操作，并消耗仿真步数。完整的工具集见表F.1。

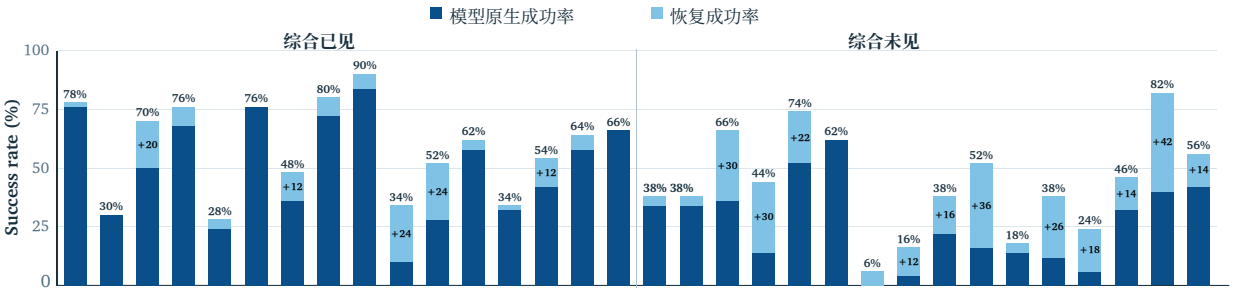


图 F.1：32 个 RoboCasa365 综合任务的逐任务恢复收益。每个堆叠柱状图报告了 Innata-0 的成功率。深蓝色表示由 Innata-0-WAM 原生完成的回合，浅蓝色表示通过在线恢复额外取得的成功。任务按 Composite-Seen 和 Composite-Unseen 划分分组。柱状图上方的百分比表示恢复后的最终成功率，浅蓝色分段内的数值表示恢复带来的绝对百分点提升。

F.2. 详细结果分析

逐任务恢复分析。图F.1显示，恢复带来的益处强烈依赖于任务，且主要集中在对原生 Innata-0-WAM 而言较具挑战性的任务上。在 Composite-Seen 划分中，恢复在*SearingMeat* 和*SetUpCuttingStation*等任务上带来了显著提升。在若干 Composite-Unseen 任务中，提升更为明显，包括*BreadSelection*、*CategorizeCondiments*、*PanTransfer*、*RecycleBottlesByType*和*WashFruitColander*。相比之下，对于原生表现已经较强的任务，例如*PreSoakPan*、*ScrubCuttingBoard*和*GarnishPancake*，几乎不需要恢复干预，甚至完全不需要。这些结果表明，恢复主要是在困难或陌生条件下弥补 WAM 失败，而非一概改变原本成功的执行过程。

表 F.1：具身工具集。 这些工具由模拟器共享。

类别	Tool	功能
观测	task_info	返回任务指令、剩余执行预算和当前任务上下文。
	robot_state	读取当前机械臂、末端执行器和夹爪状态。
	scene_objects	列出已被检测、定位或测量的场景实体。
	rgb_view	从指定相机获取原始RGB图像。
	depth_view	从指定相机获取深度观测。
	scene_view	向智能体呈现同步的多视角观测（可选）并附加视觉叠加层。
	camera_info	返回相机内参、外参和位姿信息。
感知	target_find	定位并分割由语义名称、图像点或边界框指定的目标。
	target_judge	估计目标的三维中心、空间范围、方向和测量不确定性。
	pixel_unproject	将图像像素反投影为世界坐标系中的三维点。
	point_project	将世界坐标系中的三维点投影到指定的相机图像平面上。
运动调整	pose_move	将末端执行器移动到参数化的目标位姿。
	arc_move	围绕指定的旋转中心执行参数化弧线运动和轴。
	arc_push	在沿着一条受限弧线。
	base_move	调整移动底座的位置和方向。
原子技能	gripper_set	打开或关闭夹爪，或将其设置为指定的开口宽度。
	arm_reset	使机器人手臂返回预定义的安全构型。
	motion_hold	在有限数量的控制步数内保持当前机器人配置。
	rim_grasp	利用估算出的容器容器。