



精灵构想者：用于机器人操作的统一世界基础平台

科学论文 廖悦周鹏飞* 黄思远* 杨东林 陈圣聪 蒋宇鑫 胡跃 蔡景彬 刘思 罗建兰
陈利亮[†] 严水成[○] 姚茂庆[○] 任广辉^{†○}

阿吉博特精灵团队 LV-NUS实验室 北航

科学论文<|startofcontentleak|><https://genie-envisioner.github.io>

摘要 学科文

我们推出了Genie Envisioner (GE)，这是一个用于机器人操作的统一世界基础平台，将策略学习、评估和仿真整合到一个单一的视频生成框架中。其核心组件GE-Base是一个大规模的指令条件视频扩散模型，能够在结构化的潜在空间中捕捉真实世界机器人交互的空间、时间和语义动态。在此基础上，GE-Act通过一个轻量级的流匹配解码器，将潜在表征映射为可执行的动作轨迹，从而在极少监督的情况下实现跨不同本体的精确且可推广的策略推理。为了支持可扩展的评估与训练，GE-Sim作为一个动作条件神经模拟器，能够生成高保真的回放数据，用于闭环策略开发。此外，该平台还配备了EWMBench，一套标准化的基准测试套件，用于衡量视觉保真度、物理一致性和指令与动作的匹配程度。这些组件共同使Genie Envisioner成为一种可扩展且实用的基础平台，适用于指令驱动的通用具身智能。所有代码、模型和基准测试都将公开发布。

1 引言 科学论文

"The best way to predict the future is to invent it."

— 艾伦·凯

能够在物理世界中感知、推理并采取行动的具身智能体，代表了人工智能系统的下一个前沿领域。其核心仍面临一项基础性的研究挑战：开发可扩展且稳健的机器人操作能力——即通过选择性接触来有目的地与物理环境互动并加以控制的能力 (Mason, 2001)。尽管这一领域已取得了显著进展，从解析方法 (Berenson等, 2009; Stilman, 2007)、基于模型的框架 (Ebert等, 2018; Janner等, 2019; Nagabandi等, 2020)，到利用大规模数据集学习操作策略的数据驱动方法 (Black等, 2024; Brohan等, 2023; Bu等, 2025b; Kim等, 2024)，现有的系统通常仍然依赖于独立的数据采集、训练和评估阶段。每个阶段都需要专门的基础设施、人工干预以及针对特定任务的调优；这种环节间的摩擦可能会减缓迭代速度，掩盖故障模式，并阻碍大规模下的可重复性。这些割裂的阶段凸显了当前缺乏一种能够以统一方式学习和评估操作策略的集成框架。

为此，我们推出了Genie Envisioner (GE)，这是一个统一的平台，将机器人感知、策略学习和评估整合到一个闭环的视频生成世界模型中，如图1所示。其核心是GE-Base，这是一种基于指令条件的多视角视频扩散模型，基于约3000小时的视频与语言配对数据进行训练，涵盖了来自AgiBot-World-Beta数据集 (Bu等人, 2025a) 中超过一百万个真实世界机器人操作场景。GE-Base以机器人的视觉观察为条件，通过自回归方式生成视频片段，捕捉在高层次指令指导下操作行为的时间演变过程。借助机器人

* Equal Contribution. † Project Leader. ○ Corresponding Author.

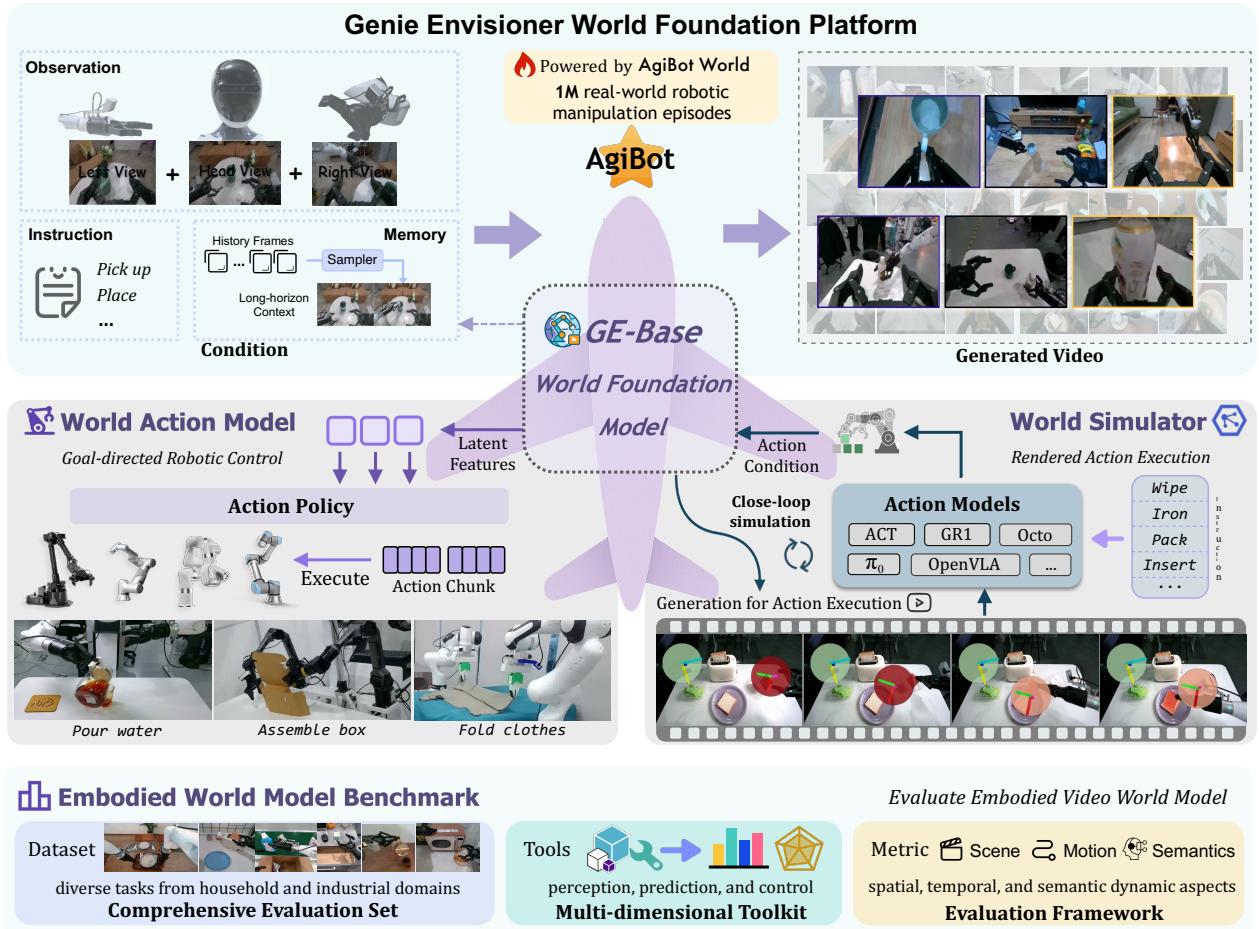
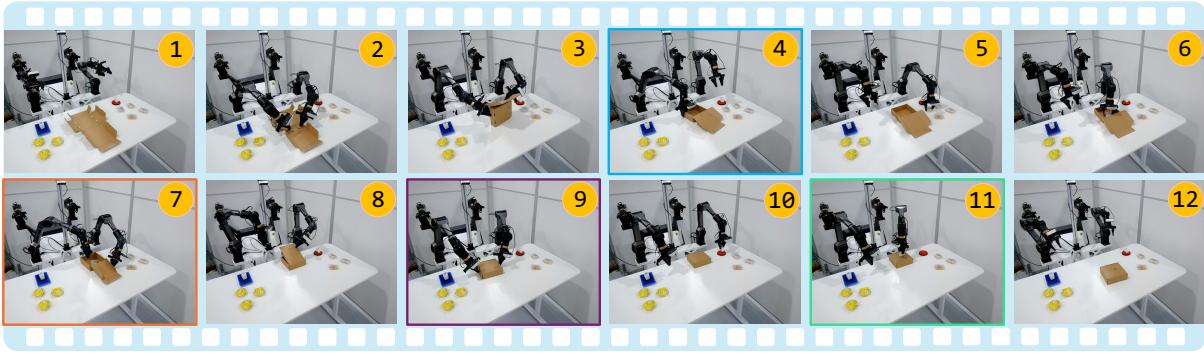


图1：Genie Envisioner世界基础平台概览。Genie Envisioner是一个统一的世界基础平台，在单一的视频生成框架中集成了操控策略的学习与评估。其核心是GE-Base，一个大规模世界模型，用于编码机器人交互的空间、时间和语义结构。围绕GE-Base构建了两个关键功能模块：GE-Act，一个基于指令条件推断策略的世界行动模型；以及GE-Sim，一个基于视频的世界模拟器，通过动作条件生成实现闭环执行。此外，该平台还配备了EWM Bench这一综合评估套件，用于评估视觉保真度、物理合理性以及指令与策略的一致性。因此，GE为通用智能的具身化提供了一个实用且可扩展的基础。

领域自适应预训练，GE-Base建立了一种从语言指令到具身视觉空间的映射，通过建模现实世界交互中的空间、时间和语义规律，捕捉机器人操作的本质。它通过推断潜在轨迹来实现这一目标，这些轨迹联合编码了机器人的感知输入以及在合理动作序列下场景的预期演变。为了弥合视觉表征与可执行机器人控制之间的差距，我们提出了GE-Act，一种轻量级的并行流匹配动作模型。GE-Act根据语言指令对视觉潜在特征进行条件化处理，将其转化为精细且低延迟的运动指令，从而实现从感知和指令到可执行物理动作的直接高效映射。除了策略学习之外，仿真在实现机器人系统的规模化训练、安全验证和快速迭代方面也发挥着至关重要的作用。为此，我们推出了GE-Sim，它利用了GE-Base的具身视频生成能力，并将其生成动力学重新应用于基于动作条件的世界模拟器。GE-Sim通过基于视频的仿真支持闭环策略评估，其速度远超真实世界的执行速度。在设计了核心基础模型之后，一个关键挑战依然存在：如何评估生成的视频是否忠实地模拟了机器人的行为。这要求我们超越通用的感知指标，转而评估合成行为是否同时



"Yellow candy requires a blue stamp, white candy requires a red stamp. Fold a box, place the appropriate candy inside, seal the box, and apply the correct stamp based on the candy type."



图2：GE-Act在一种全新机器人形态——Agilex协作机器人Magic上的实际演示，该机器人形态在预训练期间未曾见过。仅利用一小时针对具体机器人形态和任务的远程操作数据进行后训练后，GE-Act便成功执行了一项复杂的操控任务，涉及对可变形物体的精细控制以及基于记忆的决策。根据通用的包装规则，机器人需按要求完成每件物品的包装流程。在此，我们展示了第一个包装周期的详细执行过程。机器人首先堆叠一个可变形的盒子，根据指令将目标物体放入其中，并盖上盒盖，*rendering the object no longer visible*。随后，它仅依靠内部记忆，正确选择并使用与物体类型匹配的印章。这一演示充分展示了GE对新机器人形态的泛化能力、对可变形材料的精确操控能力，以及在各个步骤间保持任务相关记忆的能力。

基于物理且语义上与给定指令相一致。为此，我们提出了具身世界模型基准测试（EWMBench），这是一个原则性的评估套件，直接从视觉保真度、物理一致性和指令-动作对齐三个方面对视频生成式神经世界模拟器进行基准测试。因此，GE构建了一个统一的基于视频的机器人视觉空间，以促进在感知基础框架内对动作策略的学习、仿真和评估。与主流的视觉-语言-动作（VLA）方法（Black等人，2024；Kim等人，2024）不同，后者依赖于视觉-语言模型（VLMs）（Abouelenin等人，2025；Bai等人，2025；Chen等人，2024），将视觉输入映射到语义语言空间，并从这种以语言为中心的表征中学习动作策略，GE则通过生成式视频建模构建了一个以视觉为中心的空间。这一空间保留了精细的空间和时间线索，能够更真实地模拟机器人与环境之间的动态关系，并支持在单一、连贯的平台上实现端到端的策略学习与评估。

为了全面评估GE在具身视频生成、策略学习和仿真方面的性能，我们在一系列多样化的真实世界机器人操作任务上进行了大量实验。GE-Act通过生成54-step torque trajectories within 200 ms on a commodity GPU实现了低延迟的端到端控制，在域内AgiBot G1平台上能够精确执行任务，并展现出强大的跨具身泛化能力，即使面对Dual Franka和Agilex Cobot Magic等新型系统，仅使用1 hour of teleoperated demonstrations便能超越特定任务的基线方法（Bjorck等人，2025年；Black等人，2024年；Bu等人，2025b）。GE-Act在广泛的场景和任务中均表现出色，包括基于传送带的移动物体操作等工业应用，以及烹饪、餐桌清洁和倒液体等家庭任务。除了这些标准操作任务外，GE-Act的视觉世界建模能力还使其能够处理长时程、内存密集型序列任务，如图2所示。此外，GE-Sim通过分布式集群并行化，每小时可完成数千个剧集的策略回放评估，从而大幅加速了操作能力和策略训练的评估过程。EWMBench为基于视频的世界模型提供了一个全面的评估框架，系统性地将GE-Base与当前最先进的视频生成模型进行基准测试。结果表明，GE-Base在机器人世界3方面表现出卓越的性能。

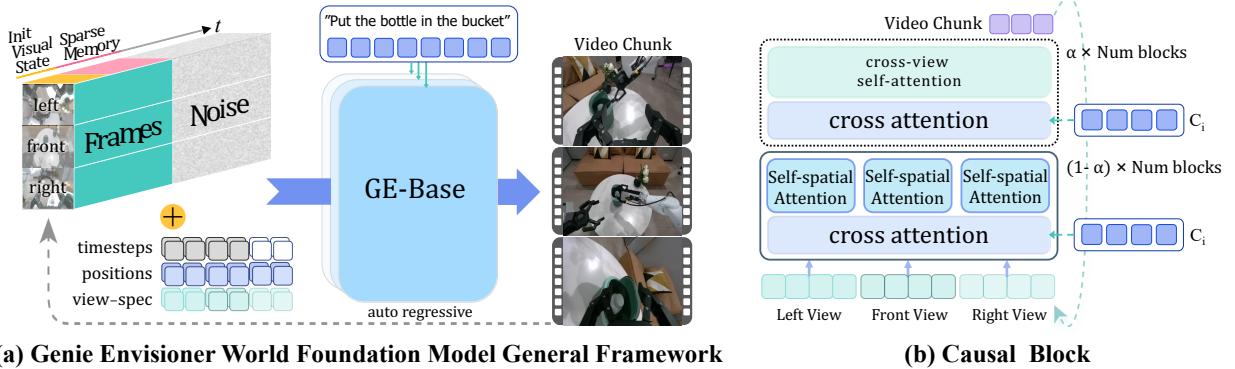


图3：GE-Base世界基础模型概览。（a）自回归视频生成过程的示意图。给定多视角视觉条件，包括初始观测和稀疏记忆，以及相应的噪声和位置嵌入，该模型根据语言指令生成下一个多视角视频片段。（b）一个专用的因果模块促进了不同视角之间的信息交换，确保在多视角视频片段生成过程中保持空间一致性。

建模与人类评估高度一致，凸显了其作为通用GE平台基础组件的重要作用。

这些贡献共同使Genie Envisioner成为一种实用且可扩展的现实世界操作基础，有助于推动下游研究。所有代码、预训练模型以及完整的EWM Bench工具集都将在论文发表时开源，以加速未来的研究进展。

2 GE-Base：世界基础模型

在本节中，我们介绍了GE-Base，这是Genie-Envisioner的核心组件。我们的目标是扩展通用视频生成模型的预测能力，构建一个*embodied predictive representation*——一种统一的生成式框架，能够根据任务指令并结合智能体的物理具身性，预测未来的机器人与环境交互过程。为此，我们将机器人视频世界建模问题定义为一个从文本和图像到视频的生成问题：给定一段语言指令和初始视觉观测，该模型能够预测未来视频片段，反映合理且连贯的机器人行为。GE-Base的一个关键设计特点是其稀疏记忆机制，该机制将当前的视觉输入与长期的历史背景相结合，通过统一的视觉条件实现更强的时间推理能力。基于这一框架，GE-Base采用了视频扩散Transformer架构，并引入了一种机器人自适应预训练策略，将通用视频数据集中的知识迁移到具身机器人领域中。我们在真实世界的机器人操作视频生成任务上展示了GE-Base的有效性。实验结果表明，GE-Base能够生成与指令一致、时间上连贯的视频序列，并且在多种操作任务和不同具身形态之间具有良好的泛化能力。

2.1 基本架构

科学论文

$$\mathbf{x}_{1:N}^{(t)} = \mathcal{W}(\hat{\mathbf{x}}_{0:t-1}, \mathbf{x}_0, q).$$

这种方案能够逐步生成在视觉和指令条件下均具时间一致性的视频片段。通过将长期稀疏记忆整合到视觉状态中，而非仅依赖于

在最近的帧中，该模型能够有效捕捉长时间跨度的依赖关系，同时在整个操作过程中保持语义一致性和视觉连贯性。

为了在机器人视频建模中兼顾效率与容量，我们采用了一种紧凑的视频生成模型作为核心架构。我们的GE-Base世界模型 \mathcal{W} 在设计时充分考虑了灵活性，能够无缝集成各种基于扩散变换器（DiT）的视频生成模型。具体而言，我们选择了LTX-Video 2B（HaCohen等人，2024年）和COSMOS2 2B（Agarwal等人，2025年）作为基础模型。LTX-Video提供了更快速、更轻量化的架构，支持高效的下游动作策略预测；而COSMOS2则能实现更高品质的视频合成，非常适合用于高保真度的仿真任务。考虑到双臂机器人系统感知的自我中心特性，我们将 \mathcal{W} 扩展为一个多视角、基于语言和图像条件的生成框架，该框架利用来自三个机载摄像头的时序同步输入：一个头戴式视角(v^h)，以及两个腕部安装的视角(v^l, v^r)。 $x_0, \hat{x}_{0:t-1}$ 和 x_t 中的每一帧都遵循这种三视角观测结构。

如图3所示，生成流程首先通过一个共享的视频编码器 \mathcal{E} 对初始视觉观测 x_0 和稀疏记忆 $\hat{x}_{0:t-1}$ 中的多视角观测进行编码。对于每个视角，我们得到潜在的视觉标记，分别用 $\mathcal{E}(v_0^{(i)})$ 和 $\mathcal{E}(v_{t-1}^{(i)})$ 表示 $i \in \{h, l, r\}$ 。每个视角的视觉标记序列由来自 x_0 和 $\hat{x}_{0:t-1}$ 的标记拼接而成。与每个视角相对应，还会初始化一个独特的噪声图 $z^{(i)}$ ，以指导生成过程。为了在区分不同视角信息的同时保持时空对齐，我们为每个标记和噪声输入同时添加了二维旋转位置嵌入 e_{pos} 以及视点特定的可学习嵌入 e_{view} 。所有视角中经过增强的标记和噪声图被拼接在一起，并进一步通过时间步编码 e_t 进行嵌入，随后输入到DiT主干网络中，以自回归方式生成下一个视频片段。

为了促进跨多个视图的连贯推理，我们将标准的空间自注意力机制从 (H, W) 扩展为跨视图自注意力机制 (N, H, W) ，其中 N 表示相机视角的数量。隐藏状态被重塑为 (B, N, T, H, W, C) ，以实现跨视图的联合推理。为确保计算上的可行性，跨视图注意力被稀疏地插入到选定的DiT模块中，而其余模块则通过将 N 维度折叠进批次维度，以独立处理各个视图，从而得到形状为 $(B \cdot N, T, H, W, C)$ 的张量。这种混合注意力机制在视图层面的一致性和效率之间实现了平衡。

为了融入语义层面的任务指导，指令 q 通过一个冻结的T5-XXL编码器（Raffel等，2020）进行处理，生成一组文本嵌入 $\mathcal{T}(q)$ 。这些嵌入通过DiT中的交叉注意力层整合到视觉标记流中，从而使模型能够将视频生成与指令语义对齐。

鉴于此设计，世界模型 \mathcal{W} 预测下一个视频片段 \hat{x}_t 为：

$$\hat{x}_t = \mathcal{W}\left(\{v_0^{(i)}, v_t^{(i)}, z^{(i)}\}_{i \in \{h, l, r\}}, \mathcal{T}(q)\right),$$

其中， $v_0^{(i)}$ 和 $v_t^{(i)}$ 表示来自视角 i 的编码后的初始和历史视觉标记， $z^{(i)}$ 代表相应的特定于视角的噪声图，而 $\mathcal{T}(q)$ 是编码后的语言指令。

这种统一的建模范式使 \mathcal{W} 能够同时捕捉空间布局、时间动态和语义意图，从而生成连贯且可控的具身机器人行为预测。

2.2 世界模型预训练

在构建用于机器人操作的基于视频的世界模型时，一个核心挑战在于如何将通用的视频生成能力适配到具身机器人领域的结构化动态和语义中。为此，我们开发了一种多阶段预训练框架，逐步使模型的时空表征与真实世界机器人行为的分布特征相一致。本节概述了我们的数据整理流程以及相应的领域适应训练策略。在训练过程中，我们会从先前的视频历史中随机采样稀疏的记忆帧，作为一种数据增强手段。这种设计增加了未来预测的难度，并增强了模型对时间变化的鲁棒性，最终提升了其在多样化操作场景中的泛化能力。

数据整理。我们采用AgiBot-World-Beta（Bu等人，2025a）数据集作为预训练的基础。该数据集包含约一百万个高质量的真实世界双臂机器人操作片段，总时长达到2,967小时，这些数据是通过人类远程操控收集的。该数据集涵盖了多种任务、物体类别以及

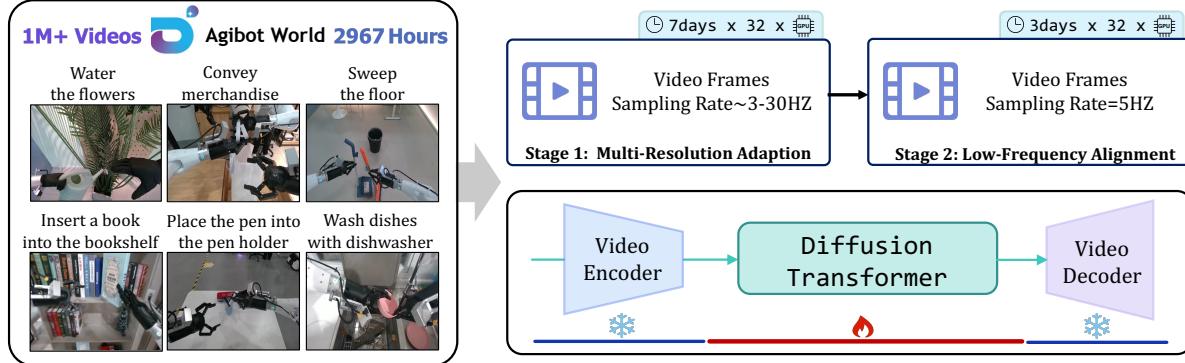


图4：GE-Base训练流程概述。GE-Base在AgiBot-World-Beta数据集上进行预训练，该数据集是一个大规模的真实世界双臂机器人操作数据集，包含100万个指令对齐的多视角视频序列。训练首先从领域适应阶段开始，通过高帧率序列和混合采样策略将通用视频生成能力迁移到机器人领域，以提升模型的鲁棒性。随后进入低帧率微调阶段，旨在使模型的时间分辨率与下游动作策略训练的需求相匹配。在整个过程中，视频编码器和视频解码器保持固定不变。

环境，其中每条轨迹都附有自然语言指令、多视角视觉观测以及结构化的动作策略。为了使该数据集适用于基于视频的建模，我们从三个校准后的摄像头视角中提取了时间同步的视频流，并确保每个视频片段与其配对指令之间具有语义一致性。这一预处理步骤生成了高质量的文本-视频对，能够反映连贯且可执行的操作行为。为了适应预训练阶段中不同的学习目标，我们采用了可变的帧采样策略，以在时间分辨率和训练稳定性之间取得平衡。

第一阶段：多分辨率时间适应（GE-Base-MR）。第一阶段旨在弥合通用视频表征学习与机器人特定运动动态之间的差距。我们使用以3 Hz至30 Hz之间随机帧率采样的57帧视频序列对模型进行预训练。每个训练样本包含四帧稀疏记忆帧，这些帧随机选取自之前的视频历史，以增强时间上的多样性。这些视频片段通过一个预训练的变分自编码器（VAE）被编码到一个8帧的潜在空间中，并在其中加入噪声，然后通过去噪目标对模型进行优化。

这种训练设置使被称为GE-Base-MR的模型接触到广泛的运动速度和时间模式，从而促使它学习对采样率不变的时空表征。该模型同时基于视觉观察和语言指令进行条件约束，能够将高层次的任务意图映射到低层次的视觉动态，并在部分观测条件下保持 robust 性。这种设计对于实际部署至关重要，因为在现实世界中，传感器延迟、帧丢失和异步数据等情况十分常见。经过这一阶段后，GE-Base-MR能够生成高质量的机器人操作视频，准确捕捉运动动态并保持视觉一致性。该模型使用32块NVIDIA A100 GPU，在AgiBot-World-Beta数据集上进行了端到端训练，耗时约七天。

阶段二：低频策略对齐（GE-Base-LF）。为了提高训练效率，并更好地与下游动作建模中使用的时序抽象相匹配，我们使用低帧率视频序列对GE-Base-MR进行微调。具体而言，我们以固定5 Hz的频率采样9帧的片段，并额外提供4帧稀疏的记忆帧作为时序上下文。这些序列通过一个预训练的视频编码器映射到由两帧潜在表示组成的紧凑潜在空间，该编码器的参数保持冻结。在此阶段，仅更新视频生成组件。最终得到的模型GE-Base-LF经过优化，能够在稀疏视觉采样的条件下捕捉语义上有意义的过渡。训练过程对于生成路径仍然是端到端的，并同时受任务指令和视觉条件的约束。这一过程有效地将视频DiT与时序控制中使用的抽象对齐，从而能够在离散动作步骤的粒度上实现可靠的视频反馈。GE-Base-LF为后续的动作模型预训练奠定了关键基础，使用32块NVIDIA A100 GPU训练约三天时间。

“Pick up the milk from the refrigerator”



“Place the held potato into the plastic bag in the shopping cart”



图5：由GE-Base在AgiBot G1上生成的多视角机器人操作视频。我们展示了GE-Base在涉及不同物体和环境的两项任务中生成的机器人操作序列。对于每个示例，分别呈现了来自三个视角的视频：i.e.为头戴式摄像头，以及左、右手臂上的摄像头。

2.3 基于GE-Base的机器人操作视频生成

我们使用GE-Base生成双臂机器人操作视频，基于LTX-Video 2B架构（目前还在探索其他基础架构）。这一过程采用自回归方法，每一步都根据初始观测、一系列记忆帧以及语言指令生成一个新的视频片段。生成过程以迭代方式进行，直至指令所指定的任务完全执行完毕，最终形成一段流畅的视频序列，精确地呈现整个操作过程。

在推理阶段，记忆帧会以固定间隔从先前的视频片段中均匀采样，从而确保稳定的时间动态和一致的预测效果。我们针对现实世界中的双臂机器人操作任务对这一流程进行了评估。如图5所示，GE-Base生成的多视角视频能够准确反映多样化的语言指令。结果表明，该模型能够在不同视角间保持空间一致性，保留背景和场景结构，并生成与指令语义相一致的稳定、分步执行过程。关于视频生成质量的进一步分析，请参阅基准测试部分（第6节）。

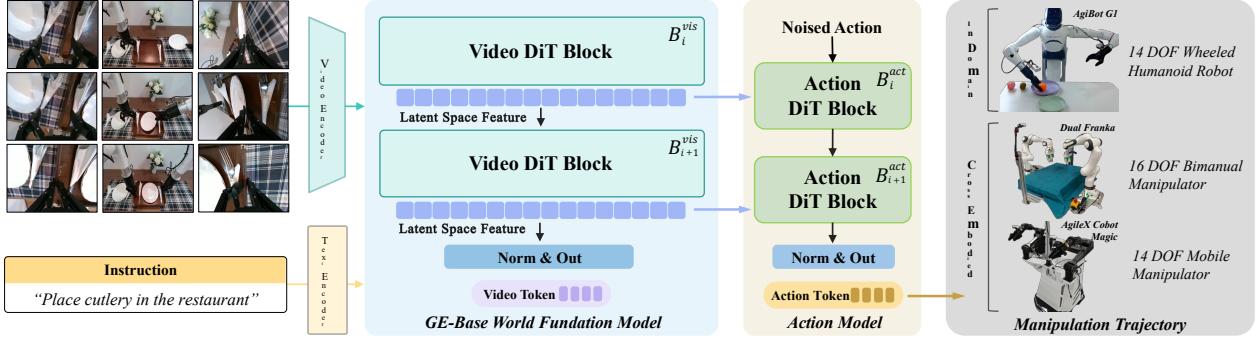


图6：GE-Act世界行动模型概览。GE-Act在GE-Base基础模型的基础上进行了扩展，引入了一个并行的动作分支，该分支可将视觉潜在表征转换为结构化的动作策略轨迹。它沿用了与GE-Base相同的模块设计和深度，但通过减少隐藏维度以提高效率。视觉潜在特征通过交叉注意力机制融入动作路径，确保动作的语义基础。最终的动作预测则采用基于扩散的去噪流匹配管道生成，将带有噪声的动作预测逐步细化为连贯的动作轨迹。

3 GE-行动：世界行动模型

在具身机器人中部署视觉-语言基础模型时，弥合高层次的世界建模与低层次的控制至关重要。我们提出了GE-Act，一个即插即用的世界行动模块，它通过一个轻量级的1.6亿参数自回归动作解码器，对基于LTX-Video的快速GE-Base基础模型进行增强。GE-Act能够将多模态潜在表征（基于多视角视觉观测和语言指令）转化为具有时间结构的动作策略，从而实现无需显式生成视频的指令遵循行为。这一架构将感知与控制紧密耦合，为跨不同环境的实时机器人操作提供了一种可扩展且高效的解决方案。

3.1 基本架构

GE-Act是一个即插即用的世界行动模块，它扩展了GE-Base基础模型，以实现基于指令的机器人控制。在架构上，它与GE-Base的视觉骨干并行运行，采用基于自回归DiT的设计，将潜在的视觉表征转化为具有时间结构的动作策略。这种集成实现了高层次感知理解与低层次运动执行之间的桥梁，支持从多视角视觉观测和语言指令中无缝生成策略。

如图6所示，GE-Act通过镜像其DiT模块的深度，同时采用较小的隐藏维度，以确保计算效率，从而保持与GE-Base的结构对齐。在每一步中，基础模型会处理由初始观测 \mathbf{x}_0 和稀疏采样的历史帧 $\hat{\mathbf{x}}_{t-1}$ 生成的视觉标记，并以指令嵌入 $\mathcal{T}(q)$ 为条件：

$$\mathbf{v}_i = \mathcal{B}_i^{\text{vis}}(\mathbf{v}_{\text{in}}, \mathcal{T}(q)),$$

其中 \mathbf{v}_{in} 表示输入的视觉标记，而 $\mathcal{B}_i^{\text{vis}}$ 则代表GE-Base中第*i*个视觉DiT块。

同时，在GE-Act中，动作路径通过一组对应的动作专用Transformer模块 $\mathcal{B}_i^{\text{act}}$ 处理由噪声初始化的动作标记 \mathbf{z}_{act} ，并通过交叉注意力机制融入相关上下文信息：

$$\mathbf{a}_i = \mathcal{B}_i^{\text{act}}(\mathbf{z}_{\text{act}}, \text{CrossAttn}(\mathbf{z}_{\text{act}}, \mathbf{v}_i)),$$

其中 \mathbf{a}_i 表示输出动作的表示。

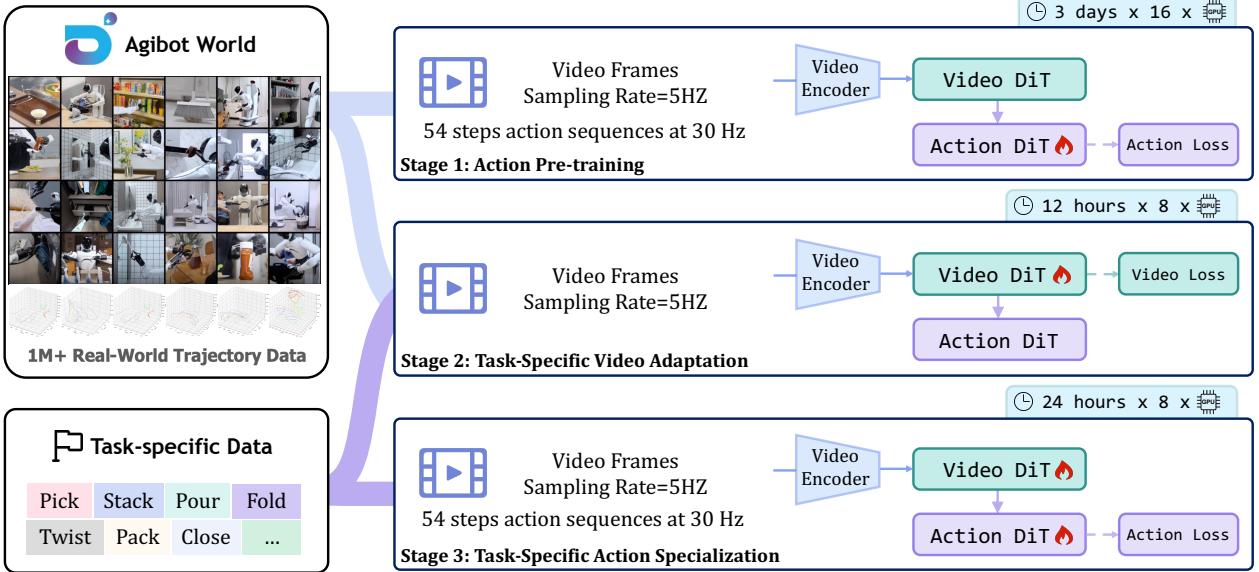


图7：GE-Act训练流程概览。GE-Act模型源自GE-Base基础模型，通过一个三阶段的训练过程获得，该过程使用了AgiBot-World-Beta数据集中的文本-视频-策略三元组。第一阶段进行动作空间预训练，优化视觉主干网络，以将视频序列映射到潜在的动作策略空间中。随后，进行两阶段的任务适配程序，使模型针对多样化的下游任务实现专业化。在此阶段，首先利用特定任务的视觉数据对视频编码器进行适配，然后使用相应的控制信号对动作头进行微调。

这种模块化架构使GE-Act能够完全在潜在特征空间中运行，从而在部署过程中无需显式生成视频即可实现控制推理。当集成到实际系统中时，该模型可以直接处理实时感知输入，并通过闭环形式保持策略的一致性。

3.2 训练流程

我们采用了一种受标准视觉-语言-行动（VLA）操控框架启发的两阶段训练范式，包括与任务无关的预训练，随后是针对特定任务的适应性调整。

预训练。在动作模型的预训练阶段，我们利用AgiBot-World-Beta数据集，将预训练的视觉-语言表征专门用于动作策略的学习。世界模型 \mathcal{W} 使用GE-Base-LF中的固定参数进行初始化，以保留其时空和语义先验，而仅更新动作解码模块的参数。为了降低计算开销，训练过程中禁用了视频生成功能。取而代之的是，采用低频率的视觉记忆序列作为条件输入，该序列由以5 Hz采样的四帧组成；同时，模型预测高频率的动作序列，即以30 Hz的频率生成包含54个步骤的动作序列。训练过程仅通过真实动作轨迹进行监督，使模型能够在预训练的潜在空间内完全学习与控制相关的动态特性。这一过程在由十六个NVIDIA A100 GPU组成的集群上大约需要三天时间即可完成。

特定任务的适配调优。为了使预训练模型适应下游机器人任务，我们采用了一个包含视频适配和动作专门化两个阶段的微调流程，旨在将通用的视觉-语言表征与特定任务的执行需求相匹配。在视频适配阶段，我们仅更新世界模型中负责视频生成的部分 \mathcal{W} ，其余参数保持冻结状态。微调过程基于一个复合数据集进行，该数据集由完整的AgiBot-World语料库和一个特定任务子集组成，其中后者被赋予了10倍的权重，以强化任务对齐效果，同时不牺牲模型的泛化能力。采样协议与GE-Base-LF所用的保持一致，以确保时间连贯性。此阶段使用8张NVIDIA A100 GPU，在约12小时内完成。在随后的动作专门化阶段，整个模型——包括GE-Base骨干网络和动作模块——仅在特定任务数据上进行微调，以捕捉精细的控制动态。这一过程与动作预训练的设置相类似，并沿用了相同的采样策略，以确保时间和控制层面的一致性。此阶段同样使用8张NVIDIA A100 GPU，训练时长约36小时。

3.3 异步推理

为了弥合视觉处理与运动控制之间的时间差距，我们提出了慢速-快速异步推理模式，该模式通过利用两个关键层面的不对称性——去噪复杂度和目标频率——来优化计算效率。

非对称去噪策略。我们的推理流程根据每种模态的不同需求分配计算资源。视频DiT在每次推理过程中仅执行一步流匹配去噪，以生成视觉潜在标记，这些标记随后被缓存并在动作生成阶段重复使用。而动作模型由于需要更高的时间分辨率以实现精确控制，因此会执行五步去噪，且所有步骤均基于相同的已缓存视觉表征。这种方法确保了在搭载于真实机器人上的NVIDIA RTX 4090 GPU上，54步的前向传播可在200毫秒内完成，从而保证了实时推理能力。

除了改进去噪过程之外，我们还利用了视觉感知与运动控制之间固有的频率不匹配。视频DiT的运行频率为5 Hz，而动作模型的运行频率则为30 Hz，从而形成了1比6的时间分辨率比。这种解耦设计使得稀疏的视频预测与密集的动作生成能够并行进行。通过仅表示选定的未来视频帧，我们显著降低了视频潜在空间的维度，从而无需处理高频的视觉序列。这一设计使视频DiT能够在紧凑的潜在空间中高效运行，同时动作模型仍保留了完整的时域分辨率，以实现精确且响应迅速的控制。

这种双层优化为训练和部署带来了显著优势。在训练过程中，我们通过使用随机高斯噪声初始化隐藏状态，消除了由视频加载和解码所导致的典型瓶颈，从而优化了大规模视频模型的训练流程。在部署阶段，单步视频去噪与降低潜在维度相结合，使得机器人硬件能够高效地实现实时运行，促进了视频生成与动作执行的无缝集成。

3.4 通过 GE-Act 在 AgiBot G1 上进行行动规划

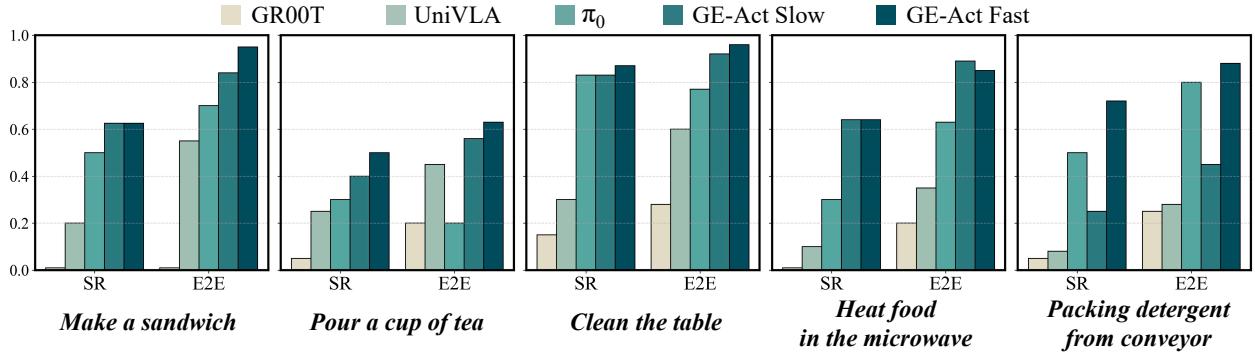


图8：AgiBot G1平台上特定任务真实世界机器人操作性能的对比。我们使用两种评估指标，将GE-Act与当前最先进的VLA基线方法在多个真实世界的双臂机器人任务中进行了比较。

为了严格评估我们方法在真实机器人操作中的有效性，我们在五个具有代表性的任务上进行了广泛的测试，每个任务都旨在检验控制精度、任务复杂度和泛化能力的不同方面。这些任务包括：(1) *Make a sandwich*：依次组装面包、培根、生菜和面包，该任务用于测试多物体协调、空间推理以及程序化任务的执行能力；(2) *Pour a cup of tea*：涉及抓取、精确倒液和重新定位茶壶，突出了流体操作中精细运动控制和灵巧性的重要性；(3) *Clean the table*：要求机器人抓取抹布并执行稳定的擦拭动作以清除表面污渍，评估轨迹稳定性和柔顺力的应用；(4) *Heat food in the microwave*：操作微波炉门、放入碗并与按钮交互，挑战系统对铰接式物体和多阶段界面操作的处理能力；(5) *Pack laundry detergent*：从传送带上抓取移动的洗涤剂袋并将其放入箱子中，旨在评估动态感知、运动跟踪以及工业规模的操作能力。这些任务涵盖了家庭和工业场景，为评估指令条件下的控制、时间定位以及闭环执行能力提供了全面的基准。

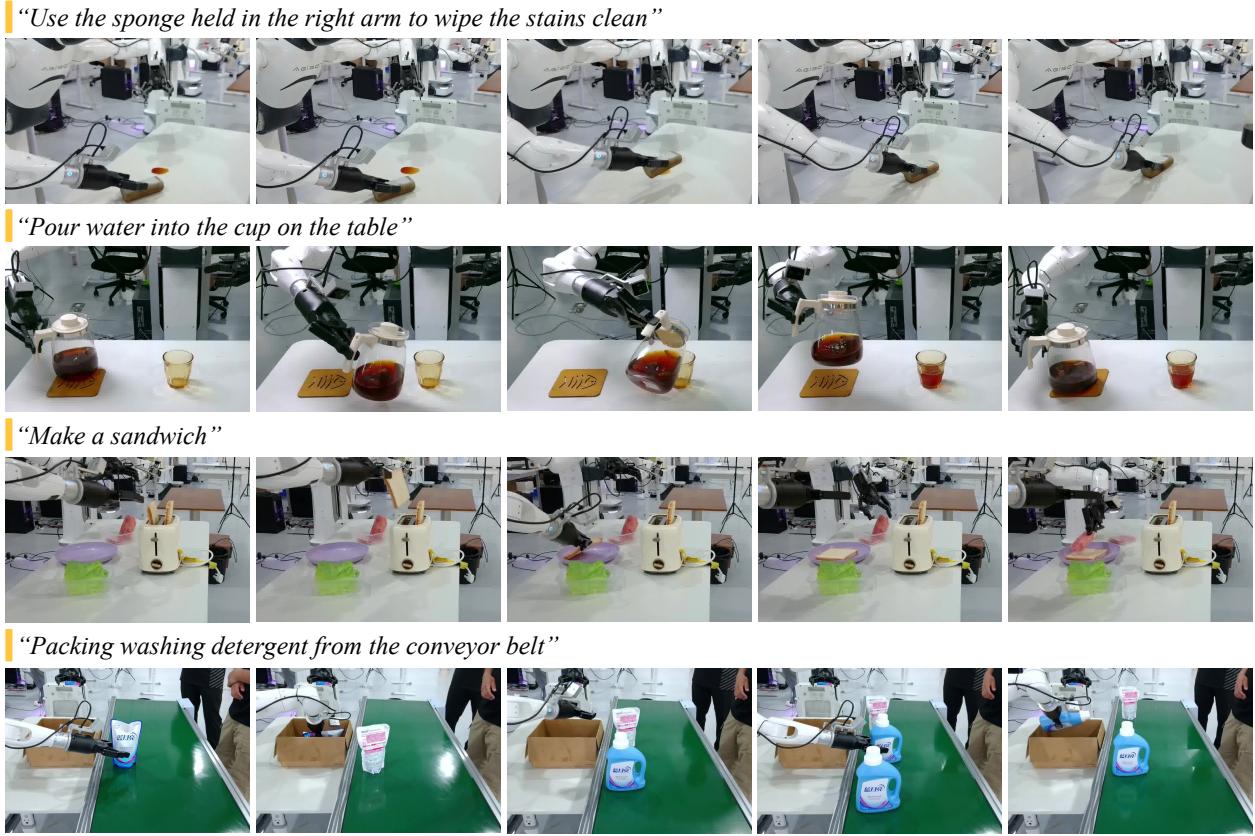


图9：通过GE-Act实现的AgiBot G1真实世界机器人操作可视化。在自然语言指令的条件下，GE-Act在AgiBot G1平台上生成并执行动作策略。视觉示例展示了该模型能够产生一致、可靠且符合上下文的操作行为，彰显了其在真实环境中的鲁棒性和有效性。

评估协议。我们采用了两种评估指标来衡量性能：分步成功率（SR）和端到端成功率（E2E）。SR指标独立评估每个子步骤，并将成功完成的子步骤数量与总子步骤数量之比作为整体成功率，从而提供对部分任务完成情况的细致洞察。相比之下，E2E指标仅评估整个任务的最终结果，在执行过程中允许对单个子步骤进行多次尝试，这更符合机器人在实际部署场景中能够从中间失败中恢复的情况。

AgiBot G1平台上的性能对比。我们以GE-Act为基准，将其与两种领先的基于VLA的机器人操作模型进行了比较：UniVLA（Bu等人，2025b），这是LIBERO基准测试中的最先进方法（Liu等人，2023）；以及GR00T N1（Bjorck等人，2025），一种大规模的VLA基础模型。所有模型均在AgiBot G1平台上进行评估，遵循相同的任务协议，并使用完全相同的任务特定遥操作演示数据进行微调。如图8所示，在一系列真实的日常操作任务中，GE-Act在SR和E2E指标上均持续优于基线模型。这种性能提升得益于预训练的GE-Base世界基础模型，该模型提供了强大的时空先验和精确的视觉-语言对齐能力，从而能够更高效、更稳健地适应各种下游操作场景。

我们通过两种运行模式进一步验证了这一设计：标准模式同步视觉与动作更新，而快速模式则利用时间抽象以提升效率。如图8所示，在各种操作任务中，快速模式均能达到相当或更优的性能，尤其在动态物体追踪和反应式抓取等对延迟敏感的场景中表现突出。值得注意的是，在“从传送带上包装洗涤剂”这类需要快速生成动作的短周期任务上，快速模型的表现显著优于其他模型。

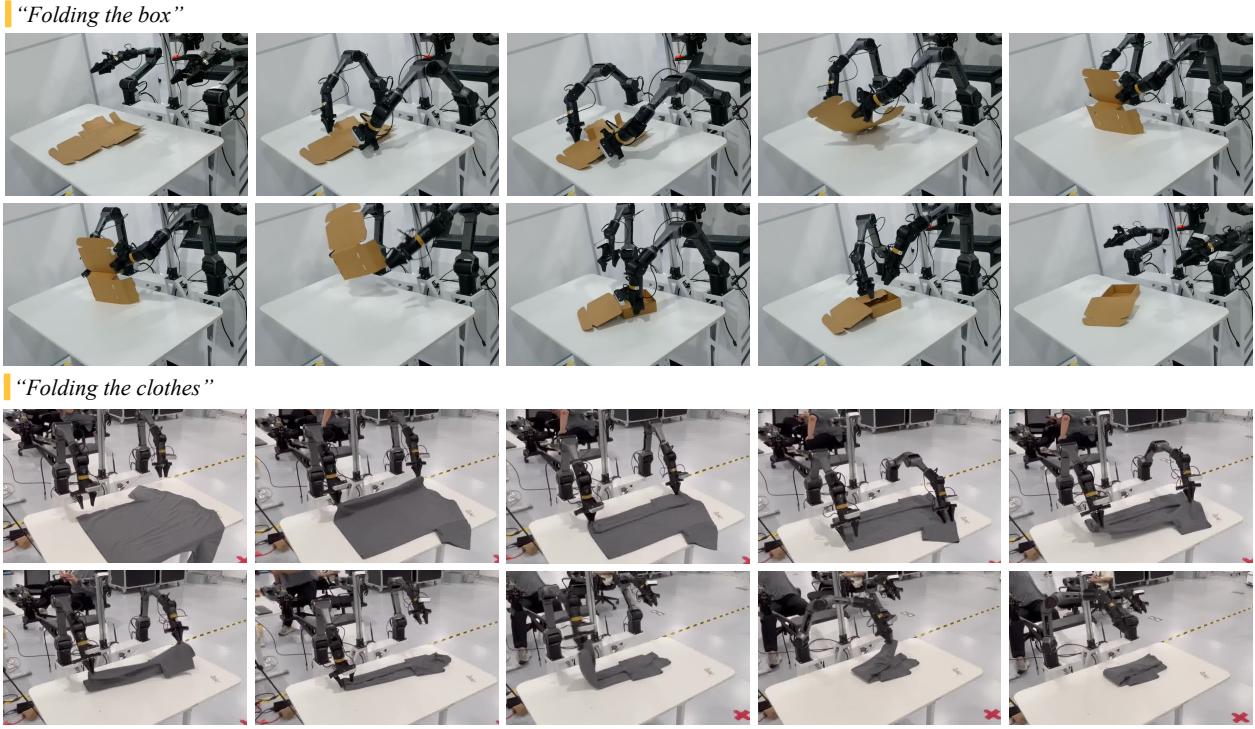


图12：GE-Act在Agilex Cobot Magic平台上的真实场景演示可视化。此图展示了GE-Act适配于一种新型Agilex Cobot Magic形态，执行包括叠布和叠盒在内的真实机器人操作任务。

在一种新型机器人平台上，以高精度和高可靠性完成任务，进一步增强了GE-Base有效迁移到新平台的能力。这一实验巩固了GE-Base作为可扩展、可适应的真实世界具身智能基础的潜力。

4.3 对双臂法兰克机器人本体的泛化

我们进一步在Dual Franka平台上评估了GE的跨本体泛化能力，通过使用250个远程操作片段（约一小时）针对布料折叠任务对GE-Act进行本体和任务特异性适配。由于缺乏专门的远程操作界面，我们在Dual Franka上的数据采集采用了一种更为简单的基于空间鼠标的操作控制系统。与Agilex Cobot的评估一致，我们选取GR0 OT N1 (Bjorck等, 2025)、 π_0 (Black等, 2024)以及UniVLA (Bu等, 2025b)作为基线，并在250个片段的适配数据集上对它们分别进行微调。图13展示了Dual Franka平台上市料折叠任务的示意图，包括由GE-Base预测的未来空间视频，以及由GE-Act执行的真实世界操作结果。结果表明，GE能够有效建模与任务相关的视觉动态，并泛化到新的本体以实现精确的操作。如图11所示，在Dual Franka平台上的真实世界执行中，GE-Act始终优于任务特异性的基线模型，这一趋势与Agilex Cobot Magic上的观察结果相吻合。值得注意的是，尽管 π_0 和GR0OT N1均接受了来自Franka本体的大规模数据的充分训练，但GE-Act仅用一小时的适配数据便实现了更优的性能。

4.4 对RoboTwin的泛化

我们进一步在双臂模拟器RoboTwin (Chen等人, 2025) 上评估了跨具身泛化能力。我们采用了一种一体化策略，使用200个演示数据（每项任务50个）对GE-Act进行四项任务的联合微调，并直接在所有任务上评估这一统一模型。相比之下，基线方法 (Black等人, 2024; Bu等人, 2025a) 则进行了特定任务的适应性调整。如图11所示，GE-Act的表现优于 π_0 和GO-1。

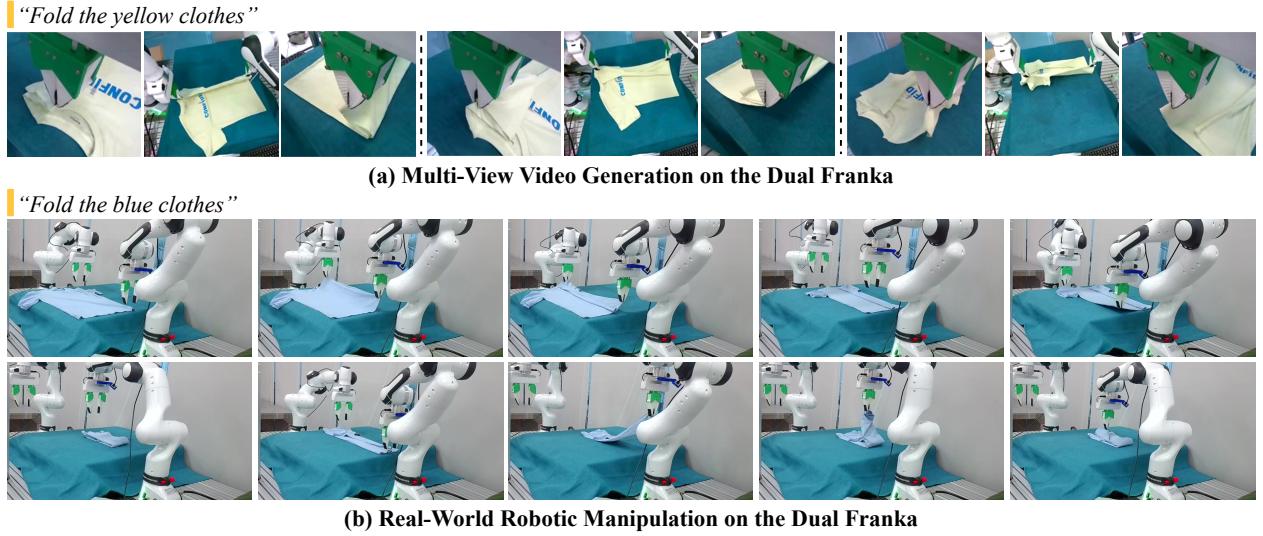


图13：通过GE实现的双法兰克机器人视频生成与现实世界操作的可视化。

在四项任务中有三项表现优异，尽管并未采用“单任务对应单模型”的设置，并且在提升壶盖任务上的表现仅略逊于VLA方法。这一细微差距可能归因于联合训练引入的任务干扰。

5 GE-Sim：世界模拟器

为了支持与现实世界相一致的评估和闭环控制，我们开发了一种基于视频的世界神经模拟器，该模拟器能够根据机器人动作生成时间上连贯的视觉预测。这一神经模拟器使具身策略模型能够与一致的视觉环境进行交互，且不受物理约束的影响，同时为策略学习和跨不同任务的泛化提供了一个统一的测试平台。

我们通过将GE-Base基础模型扩展为一个动作条件模拟器GE-Sim，实现了这一能力。在该框架中，动作轨迹作为主要控制信号，驱动视频随时间进行合成。为了实现GE-Sim，我们采用了两种GE-Base架构：一种是基于快速LTX-Video的变体，用于GE-Act；另一种是基于COSMOS2 2B的变体，用于高保真模拟和逼真视频生成。为了保持生成帧之间的视觉一致性，我们引入了一张由冻结的CLIP图像编码器编码的参考图像，作为轻量级的风格锚点。该参考图像通过交叉注意力机制注入到每个DiT模块中，与视觉观测提供的空间定位信息相辅相成。

这一转变面临的一个根本挑战是，如何调和低层控制指令与预训练世界模型所编码的高层潜在表征之间的语义差异。为了解决这一问题，如图14所示，我们提出了一种分层动作条件机制，将结构化的动作表征直接整合到GE-Base的标记空间中。该架构在保留模型预训练的时空语义的同时，能够与多种策略模型无缝对接，从而实现闭环、基于动作条件的神经仿真，并对各种机器人任务展现出强大的泛化能力。

5.1 分层动作条件机制

为了确保与多种动作策略模型的兼容性，我们采用了一种通用的机器人轨迹表示方法。对于单个机械臂，每个控制步骤被编码为一个7维向量 $[x, y, z, roll, pitch, yaw, o]$ ，其中 (x, y, z) 表示末端执行器的位置， $(roll, pitch, yaw)$ 表示其姿态（滚转、俯仰、偏航），而 o 则表示夹爪的开合程度。在我们的双臂配置中，每一步的控制信号由两个机械臂控制向量拼接而成，形成一个14维向量。在 K 步的时间范围内，完整的动作轨迹表示为 $\mathbf{A} \in \mathbb{R}^{K \times 14}$ 。为了将这种低层次的控制信号与GE-Base基础模型基于标记的输入接口相衔接，我们提出了一种分层的动作条件机制，该机制同时融合了空间和时间两个方面的信息。

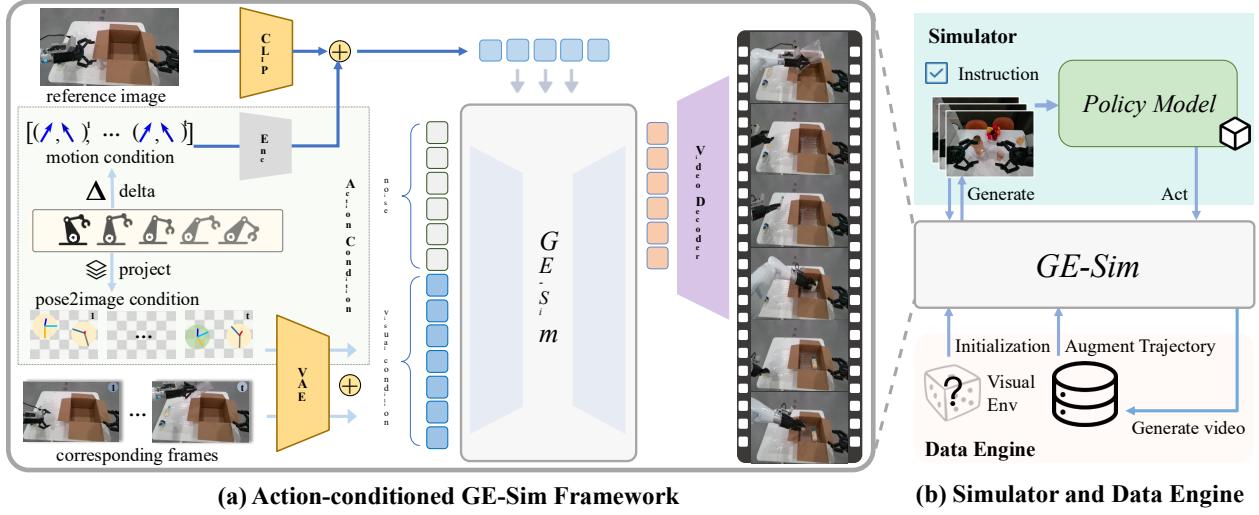


图14: GE-Sim世界模拟器概览。 (a) GE-Base被传输至一个动作条件的视频生成器中, 用于根据预测的动作模拟机器人行为。空间姿态条件被投影到图像空间, 并与历史视觉输入融合; 同时, 时间上的运动增量与参考图像拼接, 以保持风格一致性, 并通过交叉注意力注入生成模型中。 (b) GE-Sim通过生成动作条件的视频回放, 实现了闭环策略评估和可控的数据生成, 支持指令遵循以及在不同视觉情境下的一致轨迹重放。

姿态到图像的条件化。在每个时间步 i , 姿态向量 $a_i = [x_i, y_i, z_i, r_i, p_i, y_i, o_i]$ 编码了空间位置、方向以及夹爪状态。位置 (x_i, y_i, z_i) 通过标定后的相机内参和外参投影到像素坐标中。方向 (r_i, p_i, y_i) 被转换为一个旋转矩阵, 其正交轴同样被投影到图像平面, 以指示方向性。夹爪的张开程度 o_i 则在单位圆上进行渲染, 阴影的深浅反映了其状态——较浅表示打开, 较深表示关闭。不同的颜色编码用于区分左臂和右臂。这一过程生成的姿态图像 \mathbf{P}_i 在空间上与视觉场景对齐。

每个 \mathbf{P}_i 与其对应的采样历史帧 \mathbf{I}_i 配对。两者均使用共享的视频编码器 \mathcal{E} 进行编码, 并通过逐元素相加的方式融合其潜在特征:

$$\mathbf{v}_i = \mathcal{E}(\mathbf{I}_i) + \mathcal{E}(\mathbf{P}_i). \quad (1)$$

由此产生的融合标记 \mathbf{v}_i 同时捕捉了上下文视觉语义和显式的姿态信息, 并被插入到视觉标记流中以供下游处理。

运动矢量条件化。为了捕捉时间动态, 我们计算连续末端执行器位姿之间的运动增量。设 $\mathbf{a}_i = [\mathbf{p}_i, \mathbf{r}_i]$ 表示在时间步 i 处的 6 自由度位姿, 其中 $\mathbf{p}_i \in \mathbb{R}^3$ 为位置, $\mathbf{r}_i \in \mathbb{R}^3$ 为方向。增量的计算公式如下:

$$\Delta \mathbf{a}_i = \mathbf{a}_i - \mathbf{a}_{i-1} = [\Delta \mathbf{p}_i, \Delta \mathbf{r}_i], \quad (2)$$

它同时编码了位置和方向的变化。这些变化量通过一个可学习的编码器被转化为运动标记, 并与参考图像的风格标记拼接后, 经由交叉注意力注入到每个 DiT 模块中。这种具有时间感知的表征为 GE-Sim 中的动作条件视频生成提供了连贯的运动先验。

5.2 训练流程

为确保动作条件生成所需的高保真视频仿真, GE-Sim 从具有高时间分辨率的预训练模型 GE-Base-MR 中进行初始化, 该模型能够对机器人动力学进行精细建模。随后, 该模型基于完整的 AgiBot-World-Beta 数据集进行训练, 并以真实动作轨迹作为条件输入来生成视频。为了提高泛化能力和鲁棒性, 训练语料库还通过多样化的失败案例进行了扩充, 包括错误执行、行为不完整以及次优控制等情况。

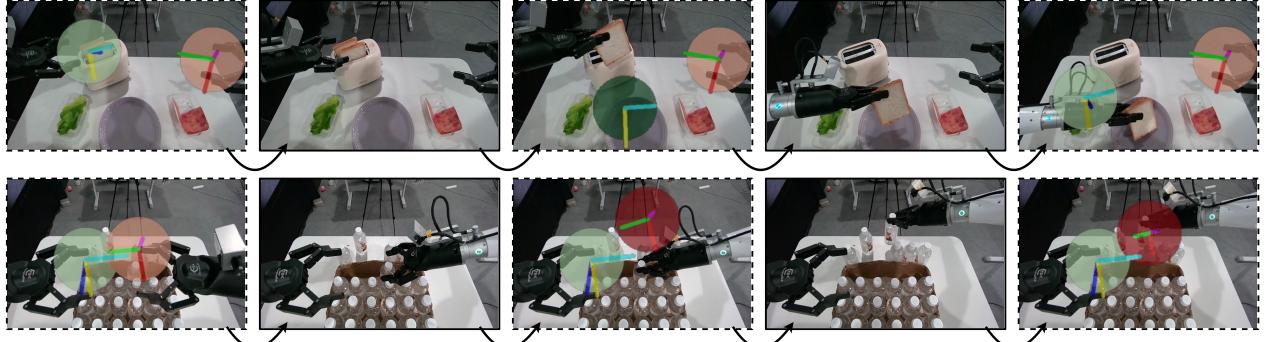


图15：GE-Sim生成的动作条件视频可视化。给定一个真实动作策略，我们使用GE-Sim生成相应的下一帧预测。对于每个样本，我们将投影的动作目标叠加到当前帧上，并与预测的下一帧并排显示，以展示模型在空间上与预期控制信号的对齐情况。

轨迹——分别来自人类遥操作和真实世界中的机器人部署。在此阶段，VAE和CLIP编码器保持冻结状态，以保留预训练的语义和空间先验，而其余参数则通过作用于预测视频表示上的流匹配损失进行优化。

5.3 行动条件下的视频生成

为了评估动作条件视频生成的精确性，我们基于真实控制序列，对GE-Sim的仿真输出进行了可视化展示。如图15所示，每个示例都展示了当前观测帧，并叠加了下一次动作的目标位置投影，同时附有模拟器合成的相应预测帧。在不同任务和视角下，生成的末端执行器运动始终与动作输入的空间意图保持一致，这表明GE-Sim能够准确地将低层级控制指令转化为连贯的视觉预测。此外，我们在图16中对比了基于两种基础架构构建的GE-Sim在相同动作条件下的表现。结果表明，基于COSMOS2的变体相比基于LTX-Video的模型具有更高的视觉保真度和更强的时间一致性，证实了其在生成高质量、与动作匹配的机器人仿真方面具有更优越的能力。

5.4 闭环仿真

为了支持对任意策略模型的闭环评估，GE-Sim作为一个基于视频的世界模拟器发挥作用。给定一段语言指令和初始视觉观察，策略模型首先将这些作为输入，并输出一条动作轨迹。随后，GE-Sim以初始观察和预测的动作策略为条件，生成一段模拟该动作结果的视频片段。这段生成的视频连同原始指令一起反馈回策略模型，用于生成下一步的动作。这一迭代过程持续进行，从而在一致且可控的视觉环境中实现策略模型的闭环仿真及与现实世界相契合的评估。

除了政策评估之外，GE-Sim还充当了一个多功能的数据引擎。通过在不同的初始视觉环境下执行相同动作轨迹，它可以生成反映多种情境的多样化操作序列。

这款基于真实世界数据的视频世界模拟器为传统物理模拟器提供了一种极具吸引力的替代方案，在实现高视觉保真度的同时，显著降低了部署成本。更重要的是，它无需手动构建环境模型，即可实现可扩展、灵活的仿真。因此，GE-Sim为一类新型的通用、逼真且低成本的世界模型奠定了基础，这些模型能够连接具身智能中的学习与评估。

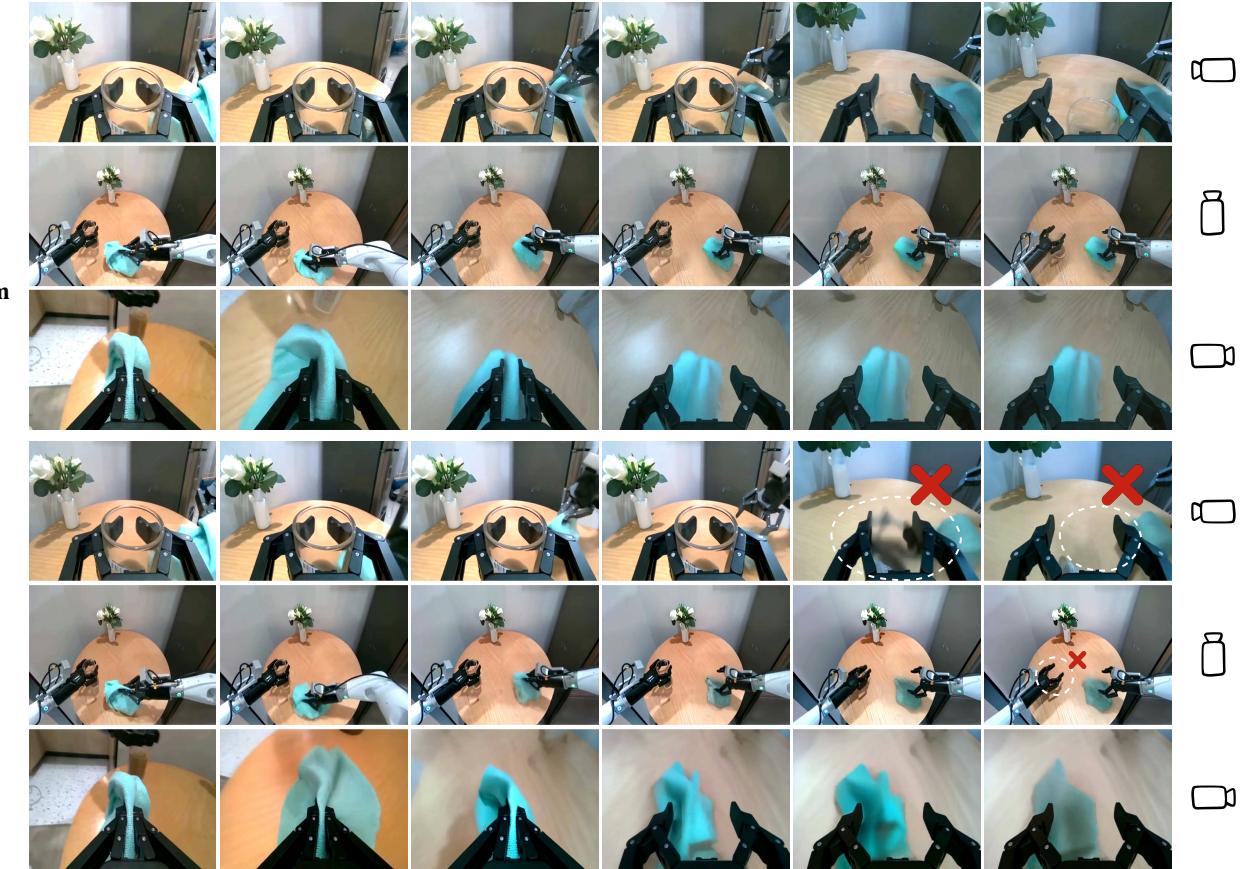


图16：两种GE-Sim变体在多视角动作条件下的生成结果对比。基于不同基础模型（COSMOS2和LTX-Video）的GE-Sim输出在相同动作条件下的可视化展示。

6 EWMBench：具身世界模型基准测试

科学论文 有效的评估框架犹如科学进步的导航工具，它确立了标准化的评判标准，并促进了不同方法论之间有意义的比较。在机器人世界建模的背景下，系统性地评估一个模型能否忠实地捕捉具身环境的结构、动态和语义特征，对于推动该领域的发展至关重要。为此，我们提出了具身世界模型基准测试——EWMBench，这是一套全面的评估工具，旨在衡量基于视频的世界模型在真实世界机器人操作中的表征保真度与实际应用价值。

除了传统视频生成领域中侧重于视觉保真度、语言对齐或人类偏好的基准测试之外，机器人操控视频还引入了更为严格的结构约束。在这一领域中，背景布局、物体配置以及本体结构（e.g.、机器人形态）应保持不变，而只有机器人的姿态和交互会根据指令发生变化。EWMBench在设计时充分考虑了这些特定领域的特性，提供了一系列面向任务的指标，用于评估视觉场景的一致性、动作的正确性以及语义对齐与多样性，从而能够更真实、更实用地评估以操控为核心场景中的世界模型。为此，EWMBench包含了一个高质量的真实世界基准数据集，以及一套开源的评估工具，为严格评估基于视频的世界模型在以操控为核心的任务中的能力建立了一个标准化框架。

6.1 基准数据集

基准数据集是从AgiBot-World-Beta测试集中精选出来的，共选取了10个具有代表性的任务，涵盖家庭和工业领域。这些任务的特点是具有明确的操作目标和强烈的序列依赖性，需要对可用性和动作顺序进行程序化推理。为确保评估的公平性，所有选定的任务均与100万规模预训练阶段所使用任务不重叠。每个任务被分解为4至10个原子级子动作，每个子动作都配有逐级说明，从而实现视频片段、动作标签与语言描述之间的细粒度对齐。对于每个任务，我们统一采样100个视频实例，以构建一个均衡且全面的评估数据集。

为了促进每项任务内部的多样性，我们实施了一种基于空间变化的轨迹选择策略。具体而言，首先提取双臂末端执行器的轨迹，并将其体素化为三维网格。然后，利用三维交并比（IoU）计算两两相似性矩阵，并采用贪心算法迭代地选择重叠程度最小的轨迹。这种方法能够确保运动模式的广泛覆盖，并最大限度地减少每项任务评估集内的冗余。

6.2 评估指标

我们建立了一个统一的评估框架，用于衡量基于视频的世界模型在多大程度上准确地捕捉了机器人操作中的空间、时间和语义动态。

场景一致性。为了评估生成视频的结构和视觉连贯性，我们引入了一种场景一致性指标，用于衡量视觉外观、环境布局和视角对齐在时间上的稳定性。具体而言，我们提出了一种基于连续帧与初始帧之间计算的补丁级特征相似度指标。首先，我们在机器人操作数据集上对强大的视觉编码器DINOv2 (Oquab等人，2023) 进行微调，使其表征空间与具身领域相一致。对于每一帧，我们使用该编码器提取补丁级别的嵌入特征。随后，我们计算各帧间对应补丁之间的余弦相似度，以量化时间一致性。较高的相似度分数表明在整个视频序列中场景结构和相机视角得到了更好的保持，从而反映出更强的时空保真度。

动作轨迹质量。为了评估根据指令执行的动作轨迹的质量，我们为每条指令手动标注一条参考轨迹作为真实标签（GT）。对于每段生成的视频，我们使用经过训练的EEF检测器来定位各帧中的夹爪位置，并重建轨迹。每条指令生成三段视频样本，并提取相应的轨迹。空间对齐（SA）通过对称豪斯多夫（symH）距离进行评估，该距离用于测量生成轨迹 P 与真实轨迹 G 之间的最大逐点偏差。为确保分数越高表示对齐效果越好，我们报告该值的倒数：

$$\text{SA}_{\text{score}} = \frac{1}{d_{\text{symH}}(G, P) + \epsilon}.$$

为了考虑生成差异，选择symH值最低的轨迹进行进一步评估。

然后，我们使用归一化动态时间规整（NDTW） (Ilharco等，2019) 对时间对齐（TA）进行评估，该方法能够捕捉生成轨迹与真实轨迹在序列和时间上的一致性。为了得到一个正相关的指标，我们报告NDTW距离的倒数：

$$\text{TA}_{\text{score}} = \frac{1}{d_{\text{NDTW}}(G, P) + \epsilon}$$

此外，我们引入了一种动态一致性（DYN）指标，用于通过比较预测轨迹与真实轨迹之间的速度和加速度曲线，评估运动动态的真实性。具体而言，我们计算了相应时间序列之间的Wasserstein距离 $W(\cdot)$ ，在无需严格时间对应的情况下捕捉分布的对齐情况。为了考虑运动幅度的变化，并防止低动态情况下的不稳定现象，我们利用幅度感知比率对每个分量进行归一化处理。最终得分定义如下：

$$\text{DYN}_{\text{score}} = \alpha \cdot \frac{\min(\Delta v^{\text{gt}}, \Delta v^{\text{pred}}) + \epsilon}{\max(\Delta v^{\text{gt}}, \Delta v^{\text{pred}}) + \epsilon} \cdot \frac{1}{W(v)} + \beta \cdot \frac{\min(\Delta a^{\text{gt}}, \Delta a^{\text{pred}}) + \epsilon}{\max(\Delta a^{\text{gt}}, \Delta a^{\text{pred}}) + \epsilon} \cdot \frac{1}{W(a)}$$

其中 $\Delta v = \max(v) - \min(v)$, $\Delta a = \max(a) - \min(a)$, $\epsilon = 10^{-8}$, 以及 $\alpha = 0.007$, $\beta = 0.003$ 。这种表述方式确保了该指标能够同时反映动态保真度和幅值鲁棒性。这种多层次的评估方法为空间、时间和动态保真度提供了全面的衡量标准。

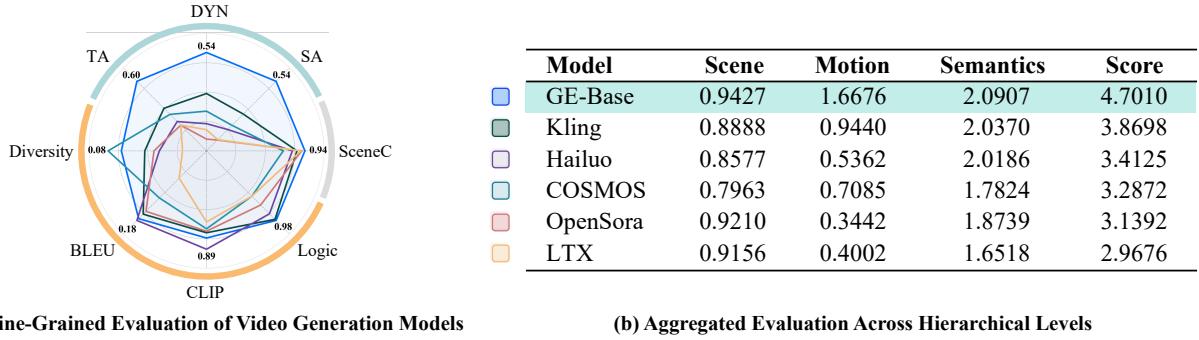


图17：机器人操控中视频世界模型的综合评估。我们利用EWM-Bench平台，系统性地评估了一系列源自最先进通用视频生成和具身世界建模方法的视频世界模型。所有模型均在统一的文本与图像到视频生成范式下进行评估。评估涵盖多个层面，包括场景、运动和语义，以捕捉视觉保真度、时间连贯性以及在多样化真实世界机器人操控任务中的语义基础。

运动语义指标。我们从语义一致性和行为多样性两个角度评估运动语义。语义一致性用于评估生成的操作行为是否与预期的任务指令相符，而多样性则衡量模型生成多样化且有效轨迹的能力。对于语义一致性，我们采用基于VLM、Qwen2.5-VL-7B-Instruct (Bai等人, 2025) 的多粒度评估框架：

- *Global-level alignment*: 一个VLM为每个生成的视频生成一个简洁的摘要标题，然后使用BLEU分数将该标题与原始的任务目标指令进行比较，以评估视频与预期任务语义之间的整体一致性。
- *Key-step consistency*: 为了评估关键子任务是否正确执行，VLM会为生成的操纵视频和真实操纵视频分别生成逐步描述。通过计算两种描述中对应步骤之间的基于CLIP的相似性来衡量一致性。
- *Logical correctness*: 为了识别违反物理或常识约束的情况，我们首先引导GPT定义机器人操作视频中常见逻辑错误的分类体系，例如幻觉动作、物体消失或物理上不可能的运动。随后，利用基于视频的视觉语言模型（VLM）来检测生成视频中是否存在这些预定义的错误类型。对于检测到的违规情况，我们会明确施加惩罚，以促使模型生成语义准确且物理上连贯的操作行为。

为了评估模型生成多样化输出的能力，我们使用基于CLIP的全局视频嵌入来衡量语义多样性。具体而言，我们计算在相同指令条件下生成的视频之间的两两CLIP相似度，并将多样性得分定义为1减去CLIP相似度。得分越高，表明语义变异性越大，反映了模型超越确定性执行进行泛化的潜力。

6.3 世界模型评估

为了全面评估基于视频的世界模型在机器人操作中的有效性，我们建立了一个综合评估框架，称为“评估竞技场”，该框架能够对多种模型架构进行直接的对比分析。在此框架下，我们对七种最先进的视频生成模型进行了基准测试，包括Open-Sora (Zheng等, 2024)、Kling (快手, 2025)、Hailuo (MiniMax, 2024)、LTX-Video (HaCohen等, 2024) 以及以场景为中心的COSMOS (Agarwal等, 2025)。所有模型均在标准化的文本与图像到视频生成范式下进行评估，在这一范式中，自然语言指令和俯视视角的视觉观测共同用于条件控制视频合成。值得注意的是，GE-Base模型基于LTX-Video架构构建，这使其能够专注于特定领域的任务，并充分利用经过微调的控制能力。

如图17所示，GE-Base在多个评估维度上始终优于基线模型，在时间对齐和动态一致性这两项核心指标上表现尤为突出，而这两项指标对于生成动作合理且时间稳定的机器人行为至关重要。尽管在运动语义方面的表现与通用视频生成相当，

整体质量。排名是基于标注者和样本的综合结果，并经过多轮评审以确保标注的一致性。如图18所示，实证结果表明，EWMBench的排名与人类判断具有高度一致性，能够有效捕捉时间对齐、语义忠实性和视觉连贯性等维度。相比之下，VBench则表现出一定的偏差，尤其是在需要具身一致性和目标条件推理的场景中。这些结果证实，EWMBench能够更真实、更贴近任务地评估机器人操作中基于视频的世界模型。

7 相关工作

用于机器人操作的世界模型。作为感知、规划和控制的内部预测性表征，世界模型的概念在机器人学中一直占据着核心地位（Chatila和Laumond，1985；Sutton和Barto，1981）。早期的方法依赖于解析建模和系统辨识（Murray等人，2017），需要针对特定任务进行工程设计，且泛化能力有限。神经网络世界模型的提出（Ha和Schmidhuber，2018）使得直接从感官输入中学习紧凑的动力学表征成为可能。此后，这些模型已发展到既能在像素空间中运行（Ebert等人，2018；Finn等人，2016），也能在学习的潜在空间中运行（Hafner等人，2019；Hu等人，2024；Wu等人，2023），并在控制和规划中得到了应用。然而，大多数先前的研究仍然局限于特定任务，受到交互数据有限的制约。最近的一些进展提出了基于视频的通用世界模型，这些模型通过大规模数据集进行训练（Agarwal等人，2025；Bruce等人，2024；Jang等人，2025；Russell等人，2025），但它们主要关注视觉合成，无法支持闭环机器人控制。相比之下，我们的工作开发了一个统一框架，将基于视频的世界建模与动作解码模块（GE-Act）及闭环仿真器（GE-Sim）相结合，从而可以直接应用于现实世界的机器人操作任务中。

用于机器人学习的视频生成模型。近年来，视频生成技术取得了显著进展，涌现出一批能够根据文本或图像提示合成高质量视频的强大模型（Blattmann等人，2023；Ho等人，2022；OpenAI，2024）。尽管这些模型在视觉质量上表现优异（Blattmann等人，2023；OpenAI，2024；Yang等人，2024），但其在机器人领域的应用仍受到限制，主要体现在缺乏动作条件、时间连贯性以及多视角推理能力方面。机器人操作需要能够基于动作指令预测未来状态、保持长期时间一致性，并对空间分布的观测信息进行推理的模型。目前，基于动作条件的视频模型（Bruce等人，2024）已展现出初步潜力，且不断有更复杂的系统被开发出来，包括自动驾驶（Russell等人，2025）和机器人模型（Agarwal等人，2025），同时通过引入相机可控性（Wang等人，2024）进一步提升了操控能力。然而，现有方法大多局限于单视角预测，且通常缺乏对任务的全面理解。GE-BASE通过多视角合成与带有记忆机制的自回归解码，有效解决了这些问题，从而提高了时空一致性和任务相关性。

视觉-语言-动作模型。视觉-语言-动作（VLA）模型已成为指令条件下的机器人技术中一种主要范式（Black等人，2024年；Brohan等人，2023年；Driess等人，2023年；Kim等人，2024年）。这些模型通常以大规模视觉-语言预训练为基础，并通过机器人演示数据进行微调，以预测动作序列。尽管这种方法在多种任务中表现出色，但仍存在一些固有的局限性。行为克隆将智能体限制于模仿，使其无法从错误中恢复或探索替代策略。此外，缺乏明确的世界模型也使得智能体无法进行内部模拟或对潜在结果进行推理。另外，高质量遥操作数据的收集仍然是一个主要瓶颈。其他方法则尝试将视觉-语言模型用作固定的编码器（Nair等人，2022年）或高层规划器（Ahn等人，2022年；Huang等人，2023年）。而我们的框架则采取了一种不同的方法，利用视觉-语言输入来指导生成式世界模型，从而实现通过内部模拟进行预测性推理和规划。

机器人学中的策略评估。高效的策略评估对于扩展机器人学习至关重要。传统的物理引擎，如MuJoCo（Todorov等人，2012）和Isaac Gym（Makoviychuk等人，2021），能够提供快速的仿真，但需要大量手动调优，并且在迁移到真实世界时仍存在差距。尽管真实世界的评估更为准确，但速度慢且资源消耗大（Zhou等人，2025）。最近的一些研究尝试将生成模型融入仿真器中（作者，2024；Nasiriany等人，2024），为高效且可扩展的评估提供了新的可能性。然而，这些方法大多局限于简化场景或受限的观测模态。GE-Sim通过将机器人模型嵌入到一个生成式循环中来应对这些挑战，该循环支持多视角下的长horizon操控，并包含成功和失败模式的轨迹，以提高系统的鲁棒性和可靠性。

科学论文 对具身世界模型的评估。评估具身世界模型的质量需要采用能够反映其在真实操作场景中表现的指标。传统的视频生成指标，如MSE或FVD，并不能很好地与现实任务的成功率相关联。近期的一些基准测试引入了结构化的评估协议（Huang等, 2024a,b），并涵盖了更广泛的指标，但其中许多仍然更强调视觉真实感，而非任务相关性。一些专门的框架，如PhyGenBench（Meng等, 2024）和T2V-CompBench（Sun等, 2024），分别用于评估物理理解和组合性，但缺乏与控制目标的一致性。我们的EWMBench填补了这一空白，提供了一套全面的评估工具，重点关注视觉保真度、运动一致性、语义对齐以及动作条件下的可控性（Yue等, 2025）。该评估工具专为在具身机器人背景下评估基于视频的世界模型能力而设计。

8 限制

在本工作中，我们系统性地研究了面向真实世界机器人操作的世界模型，重点解决了视觉运动表征、策略学习和具身评估中的核心挑战。尽管我们的Genie Envisioner框架为实现可扩展且通用的机器人智能奠定了基础，但仍存在一些局限性：

- *Data Coverage and Source Diversity.* 尽管我们进行了跨具身性的迁移实验，但我们的训练完全依赖于AgiBot-World-Beta数据集——这是一个大规模但单一平台的真实世界语料库。我们并未引入互联网规模或基于模拟的数据源，这限制了预训练过程中所遇到的具身类型、传感器模态和场景配置的多样性。虽然Genie Envisioner通过少量样本适应展现了良好的泛化潜力，但其在异构来源和低资源领域中的鲁棒性仍有待深入探索。未来若能结合大规模模拟或网络获取的演示数据，将对进一步拓展其迁移能力至关重要。
- *Embodiment Scope and Dexterity.* 本研究仅限于使用平行爪式夹持器进行上半身桌面操作。更复杂的实现方式，包括灵巧的手部协调和全身运动，并未在本研究中涉及。这些能力对于现实世界中的通用机器人技术至关重要，值得进一步整合到Genie Envisioner框架中，以支持精细的多接触交互和全身行为。
- *Evaluation Methodology.* 尽管我们的EWMBench为视觉保真度、动作一致性和语言关联提供了一种结构化的评估方法，但它仍然依赖于代理指标和部分人工验证。在面对多样化的失败模式和模糊语义时，实现任务成功的完全自动化且可靠的评估仍是一个尚未解决的挑战。构建与人类判断高度一致的可扩展评估协议，对于在真实场景中进行稳健的基准测试和安全部署至关重要。

尽管Genie Envisioner目前还不是一个完整的解决方案，但它标志着朝着Genie这一具身人工智能系统迈出了有意义的一步，这种系统具备实现AGI级别操控能力的潜力。

9 结论

在本工作中，我们推出了Genie Envisioner，这是一个统一且可扩展的双臂机器人操作平台，利用高保真视频生成技术。其核心组件GE-Base提供了一个坚实的底层基础，能够捕捉机器人交互中的时空和语义动态，从而实现与指令对齐的视频合成。通过集成GE-Act模块，系统能够实现高精度的任务执行，不仅在多种领域内任务中表现出色，还展现出卓越的跨机器人本体泛化能力。经过少量调整，GE-Act即可成功迁移到新型机器人平台上，并在诸如叠衣服和装箱等复杂任务中表现优异。此外，GE-Sim进一步增强了框架的功能，支持闭环仿真，从而实现策略的持续优化。EWMBench则提供了一套全面的评估工具，确保从视觉真实感、语义对齐到策略一致性的全方位稳健评估。广泛的现实世界测试验证了GE-Base、GE-Act和GE-Sim的优越性能，使Genie Envisioner成为构建通用型、指令驱动具身智能的强大基础。

致谢

我们衷心感谢先前工作的基础性贡献，包括EnerVerse（黄等人, 2025）、EnerVerse-AC（蒋等人, 2025）和EWMBENCH（岳等人, 2025），这些工作为我们提供了灵感和基础。

S.奈尔、A.拉杰斯瓦兰、V.库马尔、C.芬恩和A.古普塔。R3m：一种用于机器人操作的通用视觉表征。载于*Conference on Robot Learning (CoRL)*, 2022年。

S.纳西里亚尼、A.马杜库里、L.张、A.帕里克、A.洛、A.乔希、A.曼德尔卡和Y.朱。Robocasa：面向通用机器人的大规模日常任务仿真。*arXiv preprint arXiv:2406.02523*, 2024。

OpenAI. Sora, 2024. 网址 <https://openai.com/sora/>.

M.奥夸布、T.达尔塞、T.穆塔卡尼、H.沃、M.斯扎夫拉涅茨、V.哈利多夫、P.费尔南德斯、D.阿齐扎、F.马萨、A.埃尔-努比等。Dinov2：无监督学习鲁棒视觉特征。*arXiv preprint arXiv:2304.07193*, 2023。

A.拉德福德、J. W.金、C.哈拉西、A.拉梅什、G.戈赫、S.阿加瓦尔、G.萨斯特里、A.阿斯克尔、P.米什金、J.克拉克等。从自然语言监督中学习可迁移的视觉模型。载于*International conference on machine learning*, 第8748–8763页。PMLR, 2021年。

C.拉费尔、N.沙泽尔、A.罗伯茨、K.李、S.纳朗、M.马特纳、Y.周、W.李和P. J.刘。利用统一的文本到文本转换器探索迁移学习的极限。*Journal of machine learning research*, 21(140):1–67, 2020。

L.罗素、A.胡、L.贝托尼、G.费多塞耶夫、J.绍顿、E.阿拉尼和G.科拉多。Gaia-2：一种用于自动驾驶的可控多视角生成式世界模型。*arXiv preprint arXiv:2503.20523*, 2025年。

M.斯蒂尔曼. 机器人关节空间中的任务约束运动规划。载于*2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 第3074–3081页。IEEE, 2007年。

K.孙, K.黄, X.刘, Y.吴, Z.徐, Z.李和X.刘。T2v-compbench：一个用于组合式文本到视频生成的全面基准测试。*arXiv preprint arXiv:2407.14505*, 2024年。

R. S.萨顿和A. G.巴托。一种构建并使用其世界内部模型的自适应网络。*Cognition and Brain Theory*, 4(3):217–246, 1981。

E.托多罗夫、T.埃雷兹和Y.塔萨。Mujoco：一种用于基于模型控制的物理引擎。载于*2012 IEEE/RSJ international conference on intelligent robots and systems*, 第5026–5033页。IEEE, 2012年。

Z. Wang、Z. Yuan、X. Wang、Y. Li、T. Chen、M. Xia、P. Luo 和 Y. Shan. Motionctrl：一种用于视频生成的统一且灵活的运动控制器。载于*ACM SIGGRAPH*, 2024年。P. Wu、A. Escontrela、D. Hafner、P. Abbeel 和 K. Goldberg. Daydreamer：用于物理机器人学习的世界模型。载于*Conference on robot learning*, 第2226–2240页。PMLR, 2023年。Z. Yang、J. Te ng、W. Zheng、M. Ding、S. Huang、J. Xu、Y. Yang、W. Hong、X. Zhang、G. Feng 等. Cogvideox：具有专家级Transformer的文本到视频扩散模型。*arXiv preprint arXiv:2408.06072*, 2024年。H. Yue、S. Huang、Y. Liao、S. Chen、P. Zhou、L. Chen、M. Yao 和 G. Ren. Ewmbench：评估具身世界模型中的场景、运动和语义质量。*arXiv preprint arXiv:2505.09694*, 2025年。

Z.郑, X.彭, T.杨, C.申, S.李, H.刘, Y.周, T.李, 和 Y.尤。Open-sora：为所有人 democratize 高效视频制作, 2024年3月。网址：<https://github.com/hpcatech/Open-Sora>。

Z.周, P.阿特雷亚, Y. L. 谭, K.佩尔奇和S.列文。Autoeval：现实世界中通用型机器人操作策略的自主评估。*arXiv preprint arXiv:2503.24278*, 2025年。