arXiv:2508.05635v3 [cs.RO] 4 Nov 2025

Genie Envisioner: A Unified World Foundation Platform for Robotic Manipulation

Yue Liao* **Pengfei Zhou*** **Siyuan Huang*** **Donglin Yang** **Shengcong Chen**
Yuxin Jiang **Yue Hu** **Jingbin Cai** **Si Liu** **Jianlan Luo** **Liliang Chen†**
Shuicheng Yan◇ **Maoqing Yao◇** **Guanghui Ren†◇**

AgiBot Genie Team LV-NUS Lab BUAA

https://genie-envisioner.github.io

Abstract

We introduce Genie Envisioner (GE), a unified world foundation platform for robotic manipulation that integrates policy learning, evaluation, and simulation within a single video-generative framework. At its core, GE-Base is a large-scale, instruction-conditioned video diffusion model that captures the spatial, temporal, and semantic dynamics of real-world robotic interactions in a structured latent space. Built upon this foundation, GE-Act maps latent representations to executable action trajectories through a lightweight, flow-matching decoder, enabling precise and generalizable policy inference across diverse embodiments with minimal supervision. To support scalable evaluation and training, GE-Sim serves as an action-conditioned neural simulator, producing high-fidelity rollouts for closed-loop policy development. The platform is further equipped with EWMBench, a standardized benchmark suite measuring visual fidelity, physical consistency, and instruction-action alignment. Together, these components establish Genie Envisioner as a scalable and practical foundation for instruction-driven, general-purpose embodied intelligence. All code, models, and benchmarks will be released publicly.

1 Introduction

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

Embodied agents that sense, reason, and act in the physical world represent the next frontier of AI systems. At its core, a fundamental research challenge remains: developing scalable and robust robotic manipulation capabilities - the ability to purposefully interact with and control the physical environment through selective contact (Mason, 2001). Despite considerable progress has been made in this domain, ranging from analytic (Berenson et al., 2009; Stilman, 2007), model-based frameworks (Ebert et al., 2018; Janner et al., 2019; Nagabandi et al., 2020) to data-driven approaches that learn manipulation policies from large-scale datasets (Black et al., 2024; Brohan et al., 2023; Bu et al., 2025b; Kim et al., 2024), existing systems typically rely on a *patchwork* of separate data-collection, training, and evaluation stages. Each stage demands bespoke infrastructure, manual curation, and task-specific tuning; the resulting friction could potentially slow down iteration, obscure failure modes, and impede reproducibility at scale. These fragmented stages underscore the absence of an integrated framework capable of learning and evaluating manipulation policies in a unified manner.

To this end, we introduce Genie Envisioner (GE), a unified platform that collapses robot sensing, policy learning, and evaluation into a single closed-loop video generative world model, as illustrated in Figure 1. At its core lies **GE-Base**, an instruction-conditioned, multi-view video diffusion model trained on approximately 3,000 hours of video-language paired data spanning over one million real-world robotic manipulation episodes from the AgiBot-World-Beta dataset (Bu et al., 2025a). Conditioned on robot's visual observations, GE-Base autoregressively generates video chunks that capture the temporal evolution of manipulation behaviors following high-level instructions. Leveraging robotic

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

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

5202 v0N4]OR.sc[3v53650.8052:viXra

摘要　学稔文

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.

Figure 1: **Overview of the Genie Envisioner World Foundation Platform.** Genie Envisioner is a unified world foundation platform that integrates manipulation policy learning and evaluation within a single video-generative framework. At its core lies GE-Base, a large-scale world model that encodes the spatial, temporal, and semantic structure of robotic interactions. Built around it are two key functional modules: GE-Act, a world action model that infers instruction-conditioned policies, and GE-Sim, a video-based WORLD SIMULATOR that enables closed-loop execution through action-conditioned generation. The platform is complemented by EWMBench, an integrated evaluation suite that assesses visual fidelity, physical plausibility, and instruction-policy alignment. GE thus provides a practical and scalable foundation for general intelligence embodiment.

domain adaptation pretraining, GE-Base establishes a mapping from language instructions to an embodied visual space, capturing the essence of robotic manipulation by modeling the spatial, temporal, and semantic regularities of real-world interactions. It achieves this by inferring latent trajectories that jointly encode the robot's perceptual inputs and the anticipated evolution of the scene under plausible action sequences. To bridge the gap between visual representations and executable robotic control, we introduce **GE-Act**, a lightweight parallel flow-matching action model. GE-Act translates visual latent features, conditioned on language instructions, into fine-grained and low-latency motor commands, enabling direct and efficient mapping from perception and instruction to executable physical actions. Beyond policy learning, simulation plays a critical role in enabling scalable training, safety validation, and fast iteration for robotic systems. To this end, we introduce **GE-Sim**, which leverages the embodied video generation capabilities of GE-Base and repurposes its generative dynamics into an action-conditioned WORLD SIMULATOR. GE-Sim supports closed-loop policy evaluation through video-based simulation, achieving speeds significantly faster than real-world execution. After designing the core foundation model, a critical challenge remains: evaluating whether the generated videos faithfully simulate robotic behaviors. This requires moving beyond generic perceptual metrics to assess whether the synthesized behaviors are both

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EWMBench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

Figure 2: **Real-world demonstration of GE-Act on a novel robot embodiment, Agilex Cobot Magic, unseen during pretraining.** With only one hour of embodiment- and task-specific teleoperation data for post-training, GE-Act successfully executes a complex manipulation task involving fine-grained control of deformable objects and memory-based decision making. Given a general packaging rule, the robot is required to complete the packing process for each item accordingly. Here, we showcase the detailed execution of the first packing cycle. The robot first stacks a deformable box, places a target object inside based on instruction, and closes the lid, *rendering the object no longer visible*. It then correctly selects and applies the appropriate stamp, matching the object type, relying solely on internal memory. This showcases GE's generalization to new embodiments, its precise handling of deformable materials, and its ability to retain task-relevant memory across steps. .

physically grounded and semantically aligned with the given instructions. To address this, we propose the Embodied World Model Benchmark (**EWMBench**), a principled evaluation suite that directly benchmarks video generative neural world simulators in terms of visual fidelity, physical consistency, and instruction–action alignment. Therefore, GE constructs a unified video-based robotic vision space that facilitates the learning, simulation, and evaluation of action policies within a perceptually grounded framework. Different from mainstream vision-language-action (VLA) methods (Black et al., 2024; Kim et al., 2024) that rely on vision-language models (VLMs) (Abouelenin et al., 2025; Bai et al., 2025; Chen et al., 2024) to map visual inputs into a semantic linguistic space and learn action policies from this language-centric representation, GE constructs a vision-centric space through generative video modeling. This space preserves detailed spatial and temporal cues, enabling more faithful modeling of robot–environment dynamics and supporting end-to-end policy learning and evaluation within a single, coherent platform.

To comprehensively evaluate the capabilities of GE across embodied video generation, policy learning, and simulation, we conduct extensive experiments on a diverse set of real-world robotic manipulation tasks. GE-Act achieves low-latency end-to-end control by generating *54-step torque trajectories within 200 ms on a commodity GPU*. It delivers precise task execution on the in-domain AgiBot G1 platform and demonstrates strong cross-embodiment generalization to novel systems, including Dual Franka and Agilex Cobot Magic, with only *1 hour of teleoperated demonstrations*, outperforming task-specific baselines (Bjorck et al., 2025; Black et al., 2024; Bu et al., 2025b). GE-Act proves effective across a wide range of scenarios and tasks, including industrial applications such as conveyor-based moving object manipulation, and household tasks such as cooking, table cleaning, and pouring. Beyond these standard manipulation tasks, GE-Act's visual world modeling enables it to handle long-horizon, memory-intensive sequences, as shown in Figure 2. Furthermore, GE-Sim enables policy rollout evaluation at thousands of episodes per hour via distributed cluster parallelization, substantially accelerating the assessment of manipulation capabilities and policy training. EWMBench provides a comprehensive evaluation framework for video-based world models, systematically benchmarking GE-Base against state-of-the-art video generation models. The results reveal superior performance of GE-Base in robotic world

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。

(a) Genie Envisioner World Foundation Model General Framework **(b) Causal Block**

Figure 3: **Overview of the GE-Base World Foundation Model.** (a) An illustration of the autoregressive video generation process. Given multi-view visual conditions, including the initial observation and sparse memory, along with corresponding noise and positional embeddings, the model generates the next multi-view video chunk conditioned on a language instruction. (b) A dedicated causal block facilitates information exchange across different views, ensuring spatial consistency during multi-view video chunk generation.

modeling, with strong alignment to human assessments, underscoring its role as a foundational component of the unified GE platform.

Together, these contributions position Genie Envisioner as a practical, scalable foundation for real-world manipulation, facilitating downstream research. All code, pretrained models, and the complete EWMBench suite will be open-sourced upon publication to accelerate future research.

2 GE-Base: World Foundation Model

In this section, we present **GE-Base**, the core component of Genie-Envisioner. Our objective is to extend the predictive capacity of general video generation models toward constructing an *embodied predictive representation*—a unified generative formulation that anticipates future robot-environment interactions conditioned on task instructions and grounded in the agent's physical embodiment. To this end, we formulate robotic video world modeling as a text-and-image-to-video generation problem: given a language instruction and an initial visual observation, the model forecasts future video segments that reflect plausible and coherent robotic behaviors. A key design feature of GE-Base is its sparse memory mechanism, which augments the current visual input with long-term historical context, enabling stronger temporal reasoning through a unified visual condition. Built upon this formulation, GE-Base adopts a video diffusion transformer architecture and incorporates a robotic-adaptive pretraining strategy that transfers knowledge from generic video datasets into the embodied robotic domain. We demonstrate the effectiveness of GE-Base on real-world robotic manipulation video generation. Experimental results show that GE-Base produces instruction-aligned, temporally coherent video sequences that generalize well across diverse manipulation tasks and embodiments.

2.1 Basic Architecture

To align with the sequential nature of robotic manipulation data, we adopt an autoregressive video generation framework that segments the output into discrete video chunks, each containing N frames. At each step t, the world foundation model $\mathcal{W}$ generates the next chunk $\mathbf{x}_{1:N}^{(t)}$ conditioned on three components: the initial visual observation $\mathbf{x}_0$, the language instruction q, and a sparse memory $\hat{\mathbf{x}}_{0:t-1}$, constructed by sparsely sampling long-term historical frames from previous steps. The generation process is formally defined as:

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

This formulation enables the progressive generation of temporally coherent video segments, grounded in both visual and instruction conditions. By integrating long-term sparse memory into the visual state, rather than relying solely on

(a) Genie Envisioner World Foundation Model General Framework **(b) Causal Block**

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EWMBench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

recent frames, the model effectively captures extended temporal dependencies while maintaining semantic alignment and visual consistency throughout the manipulation process.

To balance efficiency and capacity in robotic video modeling, we adopt a compact video generation model as the core architecture. Our GE-Base world model $\mathcal{W}$ is designed with flexibility in mind, allowing seamless integration with various diffusion transformer (DiT)-based video generation models. Specifically, we select LTX-Video 2B (HaCohen et al., 2024) and COSMOS2 2B (Agarwal et al., 2025) as our base models. LTX-Video provides a faster and more lightweight architecture, supporting efficient downstream action policy prediction, whereas COSMOS2 offers higher-quality video synthesis, making it well-suited for high-fidelity simulation. Considering the egocentric nature of perception in dual-arm robotic systems, we extend $\mathcal{W}$ into a multi-view, language-and-image-conditioned generation framework that leverages temporally synchronized inputs from three onboard cameras: a head-mounted view (v^h) and two wrist-mounted views (v^l, v^r). Each frame in x_0, $\hat{\mathbf{x}}_{0:t-1}$, and x_t follows this tri-view observation structure.

As illustrated in Figure 3, the generation pipeline begins by encoding multi-view observations from the initial visual observation x_0 and the sparse memory $\hat{\mathbf{x}}_{0:t-1}$ using a shared video encoder $\mathcal{E}$. For each view, we obtain latent visual tokens, denoted as $\mathcal{E}(v_0^{(i)})$ and $\mathcal{E}(v_{t-1}^{(i)})$ for $i \in h, l, r$. The visual token sequence for each view is composed by concatenating tokens from x_0 and $\hat{\mathbf{x}}_{0:t-1}$. Corresponding to each view, a distinct noise map $z^{(i)}$ is initialized to guide generation. To preserve spatiotemporal alignment while distinguishing viewpoint-specific information, we augment each token and noise input with both a 2D rotary positional embedding e_{pos} and a view-specific learnable embedding e_{view}. These enriched tokens and noise maps from all views are concatenated and further embedded with a timestep encoding e_t, then fed into the DiT backbone for autoregressive generation of the next video chunk.

To facilitate coherent reasoning across multiple views, we extend standard spatial self-attention over (H, W) to cross-view self-attention over (N, H, W), where N denotes the number of camera perspectives. Hidden states are reshaped to (B, N, T, H, W, C) to enable joint cross-view reasoning. To ensure computational tractability, cross-view attention is sparsely inserted into selected DiT blocks, while the remaining blocks treat views independently by folding the N dimension into the batch dimension, yielding a shape of $(B \cdot N, T, H, W, C)$. This hybrid attention scheme balances view-level consistency and efficiency.

To incorporate semantic task-level guidance, the instruction q is processed using a frozen T5-XXL encoder (Raffel et al., 2020), producing a set of text embeddings $\mathcal{T}(q)$. These are integrated into the visual token stream through cross-attention layers within the DiT, allowing the model to align video generation with the instruction semantics.

Given this design, the world model $\mathcal{W}$ predicts the next video chunk $\hat{x}_t$ as:

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

where $v_0^{(i)}$ and $v_{\hat{t}}^{(i)}$ denote the encoded initial and historical visual tokens from view i, $z^{(i)}$ represents the corresponding view-specific noise map, and $\mathcal{T}(q)$ is the encoded language instruction.

This unified modeling paradigm enables $\mathcal{W}$ to jointly capture spatial layouts, temporal dynamics, and semantic intent—yielding coherent and controllable predictions of embodied robotic behavior.

2.2 World Model Pre-training

A core challenge in building video-based world models for robotic manipulation lies in adapting general video generation capabilities to the structured dynamics and semantics of the embodied robotic domain. To address this, we develop a multi-stage pre-training framework that progressively aligns the model's spatiotemporal representations with the distributional characteristics of real-world robot behavior. This section outlines our data curation pipeline and the corresponding training strategies for domain adaptation. During training, sparse memory frames are randomly sampled from preceding video history, serving as a form of data augmentation. This design increases the difficulty of future prediction and enhances the model's robustness to temporal variation, ultimately improving its generalization to diverse manipulation scenarios.

Data Curation. We adopt the AgiBot-World-Beta (Bu et al., 2025a) dataset as the foundation for pretraining. This dataset comprises approximately one million high-quality real-world dual-arm robotic manipulation episodes, totaling $2,967$ hours, collected via human teleoperation. The dataset spans a diverse range of tasks, object categories, and

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、$\hat{\mathbf{x}}_{0:t-1}$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$\hat{\mathbf{x}}_{0:t-1}$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$\hat{\mathbf{x}}_{0:t-1}$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：

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

其中，$v_0^{(i)}$ 和 $v_{\hat{t}}^{(i)}$ 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

Figure 4: **Overview of the GE-Base Training Process.** GE-Base is pre-trained on AgiBot-World-Beta, a large-scale real-world dual-arm robotic manipulation dataset containing 1 million instruction-aligned, multi-view video sequences. The training begins with a domain adaptation phase, transferring general video generation capabilities into the robotic domain using high-frame-rate sequences and mixed sampling strategies to enhance robustness. This is followed by a low-frame-rate fine-tuning stage designed to align the model with the temporal resolution required for downstream action policy training. Throughout the process, the video encoder and video decoder remain fixed.

environments, with each trajectory annotated with natural language instructions, multi-view visual observations, and structured action policies. To adapt the dataset for video-based modeling, we extract temporally synchronized video streams from three calibrated camera viewpoints and ensure semantic consistency between each video segment and its paired instruction. This preprocessing step results in high-quality text–video pairs that reflect coherent and executable manipulation behavior. To accommodate different learning objectives across pretraining stages, we employ variable frame sampling strategies that balance temporal resolution and training stability.

Stage I: Multi-Resolution Temporal Adaptation (GE-Base-MR). The first stage aims to bridge the gap between generic video representation learning and robotic-specific motion dynamics. We pretrain the model on 57-frame video sequences randomly sampled at frame rates between 3 Hz and 30 Hz. Each training sample includes four sparse memory frames, randomly drawn from prior video history to enhance temporal diversity. These clips are encoded into an 8-frame latent space via a pretrained VAE, where noise is added and the model is optimized through a denoising objective.

This training setup exposes the model, referred to as GE-Base-MR, to a wide spectrum of motion speeds and temporal patterns, encouraging it to learn spatiotemporal representations invariant to sampling rates. Conditioning on both visual observations and language instructions, the model learns to map high-level task intent to low-level visual dynamics while maintaining robustness to partial observations. This design is essential for real-world deployment, where sensor latency, frame drops, and asynchronous data are common. After this stage, GE-Base-MR is capable of generating high-quality robotic manipulation videos that accurately capture motion dynamics and visual consistency. The model is trained end-to-end on the AgiBot-World-Beta dataset using 32 NVIDIA A100 GPUs for approximately seven days.

Stage II: Low-Frequency Policy Alignment (GE-Base-LF). To improve training efficiency and better align with the temporal abstraction used in downstream action modeling, we fine-tune GE-Base-MR using low-frame-rate video sequences. Specifically, we sample 9-frame clips at a fixed rate of 5 Hz and provide 4 additional sparse memory frames as temporal context. These sequences are mapped into a compact latent space consisting of two latent frames via a pretrained video encoder, whose parameters remain frozen. Only the video generation components are updated during this phase. The resulting model, GE-Base-LF, is optimized to capture semantically meaningful transitions under sparse visual sampling. Training remains end-to-end for the generative pathway and is conditioned on both task instructions and visual conditions. This process effectively aligns the video DiT with the temporal abstraction used in control, enabling reliable video feedback at the granularity of discrete action steps. GE-Base-LF serves as a critical foundation for subsequent action model pretraining and is trained for approximately three days using 32 NVIDIA A100 GPUs.

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。

Figure 5: **Multi-View Robotic Manipulation Videos Generated on AgiBot G1 by GE-Base.** We visualize robotic manipulation sequences generated by GE-Base across two tasks involving varied objects and environments. For each example, videos from three views are presented, *i.e.*, the head-mounted, left-, and right-arm cameras, respectively.

2.3 Robotic Manipulation Video Generation via GE-Base

We generate dual-arm robotic manipulation videos using GE-Base, leveraging the LTX-Video 2B architecture (with additional base architectures currently being explored). This process follows an autoregressive approach, where each step generates a new video chunk conditioned on the initial observation, a series of memory frames, and the language instruction. The generation proceeds iteratively until the task specified by the instruction is fully executed, resulting in a seamless video sequence that precisely captures the complete manipulation procedure.

At inference time, memory frames are uniformly sampled from prior video chunks at fixed intervals, ensuring stable temporal dynamics and consistent prediction. We evaluate this pipeline on real-world dual-arm robotic manipulation tasks. As shown in Figure 5, GE-Base generates multi-view videos that accurately reflect diverse language instructions. The results highlight the model's ability to maintain spatial consistency across views, preserve background and scene structure, and produce stable, step-by-step execution aligned with the instruction semantics. Further analysis of video generation quality is provided in the benchmark section (Section 6).

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）。

Figure 6: **Overview of the GE-Act World Action Model.** GE-Act extends the GE-Base foundation model by incorporating a parallel action branch that converts visual latent representations into structured action policy trajectories. It follows the same block design and depth as GE-Base but reduces the hidden dimensions to improve efficiency. Visual latent features are integrated into the action pathway through a cross-attention mechanism, ensuring the semantic grounding of actions. Final action predictions are generated using a diffusion-based denoising flow-matching pipeline, refining noisy action predictions into coherent action trajectories.

3 GE-Act: World Action Model

Bridging high-level world modeling and low-level control is essential for deploying vision-language foundation models in embodied robotics. We present **GE-Act**, a plug-and-play world action module that augments the fast LTX-Video–based GE-Base foundation model with a lightweight 160M-parameter autoregressive action decoder. GE-Act translates multimodal latent representations—conditioned on multi-view visual observations and language instructions—into temporally structured action policies, enabling instruction-following behaviors without explicit video generation. This architecture tightly couples perception and control, providing a scalable and efficient solution for real-time robotic manipulation across diverse environments.

3.1 Basic Architecture

GE-Act is a plug-and-play world action module that extends the GE-Base foundation model to enable instruction-conditioned robotic control. Architecturally, it operates in parallel with the visual backbone of GE-Base, adopting an autoregressive DiT-based design that transforms latent visual representations into temporally structured action policies. This integration bridges high-level perceptual understanding with low-level motor execution, supporting seamless policy generation from multi-view visual observations and language instructions.

As shown in Figure 6, GE-Act preserves structural alignment with GE-Base by mirroring its DiT block depth while employing a reduced hidden dimension to ensure computational efficiency. At each step, the foundation model processes visual tokens derived from initial observations $\mathbf{x}_0$ and sparsely sampled historical frames $\hat{\mathbf{x}}_{t-1}$, conditioned on instruction embeddings $\mathcal{T}(q)$:

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

where $\mathbf{v}_{\text{in}}$ denotes the input visual tokens, and $\mathcal{B}_i^{\text{vis}}$ represents the i-th visual DiT block in GE-Base.

Simultaneously, the action pathway in GE-Act processes noise-initialized action tokens $\mathbf{z}_{\text{act}}$ via a corresponding set of action-specific transformer blocks $\mathcal{B}_i^{\text{act}}$, incorporating relevant contextual information through cross-attention:

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

where $\mathbf{a}_i$ denotes the output action representation.

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$\mathbf{v}_{\text{in}}$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$\mathbf{z}_{\text{act}}$，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。

Figure 7: **Overview of the GE-Act Training Pipeline.** GE-Act is derived from the GE-Base foundation model through a three-stage training process utilizing text–video–policy triplets from the AgiBot-World-Beta dataset. The first stage performs action-space pretraining, where the visual backbone is optimized to project video sequences into a latent action policy space. Subsequently, a two-stage task adaptation procedure is conducted to specialize the model for diverse downstream tasks. In this phase, the video encoder is first adapted using task-specific visual data, followed by fine-tuning of the action head using corresponding control signals.

This modular architecture allows GE-Act to operate entirely within the latent feature space, enabling control inference without requiring explicit video generation during deployment. When integrated into real-world systems, the model can directly consume live perceptual inputs, maintaining policy consistency through a closed-loop formulation.

3.2 Training Procedure

We adopt a two-stage training paradigm inspired by standard vision-language-action (VLA) manipulation frameworks, consisting of task-agnostic pretraining followed by task-specific adaptation.

Pre-training. During the pretraining phase of the action model, we leverage the AgiBot-World-Beta dataset to specialize the pretrained visual-linguistic representations for action policy learning. The world model $\mathcal{W}$ is initialized with fixed parameters from GE-Base-LF to retain its spatiotemporal and semantic priors, while only the parameters of the action decoding module are updated. To reduce computational overhead, video generation is disabled during training. Instead, low-frame-rate visual memory sequences, consisting of four frames sampled at 5 Hz, are used as conditioning inputs, while the model predicts high-frequency action sequences comprising 54 steps at 30 Hz. The training is supervised solely by ground-truth action trajectories, enabling the model to learn control-relevant dynamics entirely within the pretrained latent space. This process completes in approximately three days on a cluster of sixteen NVIDIA A100 GPUs.

Task-specific adaptation tuning. To adapt the pretrained model for downstream robotic tasks, we employ a two-stage fine-tuning pipeline comprising video adaptation and action specialization, aimed at aligning general visual-linguistic representations with task-specific execution requirements. During the video adaptation phase, we update only the video generation components of the world model $\mathcal{W}$, keeping the remaining parameters frozen. Fine-tuning is conducted on a composite dataset consisting of the full AgiBot-World corpus and a task-specific subset, with the latter upweighted by a factor of 10 to strengthen task alignment without sacrificing generalization. The sampling protocol is consistent with that used in GE-Base-LF to preserve temporal coherence. This phase is completed using 8 NVIDIA A100 GPUs over approximately 12 hours. In the subsequent action specialization phase, the full model—including both the GE-Base backbone and the action module—is fine-tuned exclusively on task-specific data to capture fine-grained control dynamics. This procedure mirrors the action pretraining setup and follows the same sampling strategy to ensure temporal and control-level consistency. This stage is trained using 8 NVIDIA A100 GPUs for approximately 36 hours.

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，采用低帧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 Asynchronous Inference

To bridge the temporal gap between visual processing and motor control, we introduce the Slow-Fast Asynchronous Inference mode, which optimizes computational efficiency by exploiting asymmetries at two critical levels: denoising complexity and target frequency.

Asymmetric Denoising Strategy. Our inference pipeline allocates computational resources based on the distinct requirements of each modality. The video DiT performs a single flow-matching denoising step per inference pass to generate visual latent tokens, which are then cached and reused throughout the action generation phase. The action model, requiring higher temporal resolution for precise control, executes five denoising steps, all conditioned on the same cached visual representations. This approach ensures that the forward pass for 54 steps is completed in 200ms on an onboard NVIDIA RTX 4090 GPU mounted on a real-world robot, ensuring real-time inference capability.

Beyond improving the denoising process, we exploit the inherent frequency mismatch between visual perception and motor control. The video DiT operates at 5 Hz, while the action model runs at 30 Hz, resulting in a temporal resolution ratio of 1 to 6. This decoupling enables sparse video prediction alongside dense action generation. By representing only selected future video frames, we significantly reduce the dimensionality of the video latent space, eliminating the need to process high-frequency visual sequences. This design allows the video DiT to operate efficiently in a compact latent space, while the action model retains the full temporal resolution necessary for accurate and responsive control.

This dual-level optimization delivers substantial advantages for both training and deployment. During training, we eliminate the typical bottleneck caused by video loading and decoding by initializing hidden states with random Gaussian noise, optimizing the training process for large-scale video models. In deployment, the combination of single-step video denoising and reduced latent dimensionality enables efficient real-time operation on robotic hardware, facilitating the seamless integration of video generation and action execution.

3.4 Action Planning via GE-Act on AgiBot G1

Figure 8: **Comparison of Task-Specific Real-World Robotic Manipulation Performance on the AgiBot G1 Platform.** We compare GE-Act with state-of-the-art VLA baselines across multiple real-world dual-arm robotic tasks, using two evaluation metrics to assess performance.

To rigorously assess the effectiveness of our approach in real-world robotic manipulation, we conduct extensive evaluations across five representative tasks, each designed to test distinct aspects of control precision, task complexity, and generalization. These include: (1) *Make a sandwich*: sequentially assembling bread, bacon, lettuce, and bread, which tests multi-object coordination, spatial reasoning, and procedural task execution; (2) *Pour a cup of tea*: involving grasping, precise pouring, and repositioning a teapot, highlighting the need for fine-grained motion control and dexterity in fluid manipulation; (3) *Clean the table*: requiring the robot to grasp a wiper and perform consistent wiping motions to remove surface stains, evaluating trajectory stability and compliant force application; (4) *Heat food in the microwave*: operating a microwave door, inserting a bowl, and interacting with buttons, challenging the system's ability to handle articulated objects and multi-stage interface operations; (5) *Pack laundry detergent*: grasping moving detergent bags from a conveyor belt and placing them into boxes, designed to assess dynamic perception, motion tracking, and industrial-scale manipulation. These tasks span both household and industrial settings, providing a comprehensive benchmark for evaluating instruction-conditioned control, temporal grounding, and closed-loop execution capabilities.

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。

Figure 9: **Visualization of Real-World Robotic Manipulation on AgiBot G1 via GE-Act.** Conditioned on natural language instructions, GE-Act generates and executes action policies on the AgiBot G1 platform. The visual samples demonstrate the model's capability to produce consistent, reliable, and contextually appropriate manipulation behaviors, showcasing its robustness and effectiveness in real-world environments.

Evaluation Protocols. We employed two evaluation metrics to assess performance: Step-wise Success Rate (SR) and End-to-End Success Rate (E2E). The SR metric evaluates each sub-step independently and computes the overall success as the ratio of successfully completed sub-steps to total sub-steps, providing fine-grained insight into partial task completion. In contrast, the E2E metric evaluates only the final outcome of the complete task, allowing multiple attempts for individual sub-steps during execution, which better reflects real-world deployment scenarios where robots can recover from intermediate failures.

Performance Comparison on the AgiBot G1 Platform. We benchmark GE-Act against two leading VLA-based robotic manipulation models: UniVLA (Bu et al., 2025b), the state-of-the-art method on the LIBERO benchmark (Liu et al., 2023), and GR00T N1 (Bjorck et al., 2025), a large-scale VLA foundation model. All models are evaluated on the AgiBot G1 platform following the same task protocols and using identical task-specific teleoperated demonstrations for fine-tuning. As shown in Figure 8, GE-Act consistently outperforms baseline models across both SR and E2E metrics on a range of real-world daily manipulation tasks. This performance gain is attributed to the pretrained GE-Base world foundation model, which supplies strong spatiotemporal priors and precise visual-language grounding, enabling more efficient and robust adaptation to diverse downstream manipulation scenarios.

We further validate this design through two operational modes: the standard mode, which synchronizes visual and action updates, and the fast mode, which leverages temporal abstractions for improved efficiency. As demonstrated in Figure 8, the fast mode achieves comparable or superior performance across various manipulation tasks, particularly excelling in latency-sensitive scenarios such as dynamic object tracking and reactive grasping. Notably, on short-horizon tasks like "Packing detergent from conveyor," which require rapid action generation, the fast model significantly outperforms

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。

Figure 10: **Multi-View Video Generation on the Agilex Cobot Magic Robotic Platform by GE-Base.** Visualization of instruction-conditioned video generated by GE-Base for two complex folding tasks on the cross-embodiment Agilex Cobot Magic robot. Each row displays temporally sampled frames from a multi-view sequence.

the standard model. Qualitative results are visualized in Figure 9, showcasing GE-Act's ability to execute complex manipulation tasks precisely and reliably in real-world settings, directly conditioned on natural language instructions.

Analysis. To systematically analyze our GE model, we conduct real-world robotic manipulation experiments on the AgiBot-G1 platform. We select a stable and controllable task, *"grasping a red cylinder from the table and placing it into a paper cup with fixed positions"*, using a dataset of 305 demonstrations. All models are trained for 40,000 steps under the same protocol. Our analysis focuses on the role of pretraining in action policy prediction, comparing general video pretraining with in-domain embodied pretraining (AgiBot-World-Beta). As shown in Table 1, training from scratch or adapting from a general video model such as LTX-Video yields near-zero success. In contrast, in-domain pretraining achieves 64 SR and 81 E2E, which further improve to 76% and 89% when combined with general video pretraining. We further validate the effectiveness of incorporating robot state as input, which yields additional performance gains. However, when applied directly to general video-pretrained models, the inclusion of state information reduces performance due to short-cut learning effects. These results demonstrate that the GE-Base pretrained world model offers strong representations and serves as a solid foundation for action policy prediction.

Table 1: Analysis of Pre-training. 'S' denotes inclusion of robot state; 'VidAW' indicates initialization from GE-Base, 'VidAda' indicates task-specific video adaptation.

VidAW	VidAda	E2E		SR	
		w/ $\mathcal{S}$	w/o $\mathcal{S}$	w/ $\mathcal{S}$	w/o $\mathcal{S}$
✗	✗	0.15	0.30	0.05	0.11
✗	✓	0	0.05	0	0
✓	✗	0.81	0.49	0.64	0.26
✓	✓	**0.89**	0.37	**0.76**	0.37

4 Cross-Embodiment Generalization with Genie Envisioner

Beyond validating GE on the in-domain AgiBot G1 platform, we assess GE's ability to generalize across different embodiments, an essential step toward developing a versatile robotic foundation model. Specifically, we evaluate GE on two widely used robotic platforms in manipulation research: the Franka arm and the Agilex Cobot Magic system, as well as a dual-arm simulator, RoboTwin. To maintain consistency with our dual-arm framework, all platforms are configured accordingly.. To maintain consistency with our dual-arm framework, both platforms are configured accordingly.

Due to differences in embodiment design and action space, directly deploying the pretrained GE models on new platforms is not feasible. To overcome this, we use a few-shot adaptation protocol, collecting a small number of high-quality teleoperated demonstrations for each task. These demonstrations are used to fine-tune both the GE and GE-Act models, enabling effective transfer and alignment with the new platforms. In addition to the standard task set on AgiBot G1, we evaluate GE on complex deformable object manipulation tasks such as *"cloth folding"* and *"box folding"*, selected for their real-world relevance and physical challenges. This framework provides a thorough evaluation of GE's transferability, robustness, and control precision across different robotic platforms

图10：由GE-Base在Agilex协作机器人Magic机器人平台上生成的多视角视频。展示了GE-Base为跨模态Agilex协作机器人Magic执行的两项复杂折叠任务所生成的指令条件视频的可视化效果。每一行显示了多视角序列中按时间采样的帧。

标准模型。定性结果如图9所示，展示了GE-Act在真实场景中直接根据自然语言指令，精确且可靠地执行复杂操作任务的能力。

分析。为了系统地分析我们的GE模型，我们在AgiBot-G1平台上开展了真实世界的机器人操作实验。我们选择了一项稳定且可控的任务*"grasping a red cylinder from the table and placing it into a paper cup with fixed positions"*，并使用包含305个示范数据集进行训练。所有模型均按照相同的协议训练了40,000步。我们的分析重点在于预训练在动作策略预测中的作用，比较了通用视频预训练与领域内具身预训练（AgiBot-World-Beta）。如表1所示，从零开始训练或基于如LTX-Video这样的通用视频模型进行迁移学习，其成功率几乎为零。

相比之下，领域内预训练分别达到了64%的SR和81%的E2E，而当与通用视频预训练相结合时，这一性能进一步提升至76%和89%。我们还进一步验证了将机器人状态作为输入的有效性，这带来了额外的性能提升。然而，当直接应用于经过通用视频预训练的模型时，加入状态信息反而会因捷径学习效应而导致性能下降。这些结果表明，GE-Base预训练的世界模型能够提供强大的表征能力，并为动作策略预测奠定了坚实的基础。

表1：预训练分析。"S"表示包含机器人状态；"VidAW"表示从GE-Base进行初始化，"VidAda"表示针对特定任务的视频适配。

VidAW	VidAda	E2E		SR	
		w/ $\mathcal{S}$	w/o $\mathcal{S}$	w/ $\mathcal{S}$	w/o $\mathcal{S}$
✗	✗	0.15	0.30	0.05	0.11
✗	✓	0	0.05	0	0
✓	✗	0.81	0.49	0.64	0.26
✓	✓	**0.89**	0.37	**0.76**	0.37

4 借助Genie Envisioner实现跨具身泛化

除了在域内AgiBot G1平台上验证GE之外，我们还评估了GE在不同机器人形态间的泛化能力，这是迈向开发多功能机器人基础模型的关键一步。具体而言，我们在操纵研究中广泛使用的两个机器人平台——Franka机械臂和Agilex协作机器人Magic系统，以及一个双臂仿真器RoboTwin上对GE进行了评估。为保持与我们双臂框架的一致性，所有平台均按此进行了相应配置。为保持与我们双臂框架的一致性，这两个平台也均按此进行了相应配置。

由于实体化设计和动作空间的差异，直接将预训练的GE模型部署到新平台上是不可行的。为此，我们采用了一种少样本适应协议，为每个任务收集少量高质量的遥操作示范数据。这些示范数据用于微调GE模型和GE-Act模型，从而实现与新平台的有效迁移和对齐。除了在AgiBot G1上标准的任务集之外，我们还针对复杂可变形物体操作任务对GE进行了评估，例如*"cloth folding"*和*"box folding"*，这些任务因其现实相关性和物理挑战性而被选中。这一框架全面评估了GE在不同机器人平台上的迁移能力、鲁棒性以及控制精度。

4.1 Few-shot Adaptation

To enable few-shot adaptation on a novel robotic embodiment, we adopt a two-stage task-specific fine-tuning strategy for GE-Act, as illustrated in the final two rows of Figure 7:

- In the first stage, we adapt the visual generative component to the new embodiment domain by fine-tuning the video DiT module using a small set of newly collected instruction-conditioned video demonstrations. During this process, the CLIP (Radford et al., 2021) and video encoders are kept frozen to preserve pretrained semantic and perceptual priors. This step enables the model to synthesize realistic, embodiment-consistent manipulation videos aligned with the new platform's visual characteristics.
- In the second stage, we train a new action DiT module from scratch using task-specific teleoperated trajectories. Owing to fundamental discrepancies in the embodiment structure and action space semantics, the pretrained action decoder is not reused. Instead, we retain the GE-Base visual backbone and learn a new action head tailored to the control dynamics and interface of the novel robotic platform.

This two-stage adaptation pipeline facilitates effective transfer of both perceptual and motor capabilities, enabling high-fidelity video generation and accurate, instruction-driven policy inference under minimal data supervision.

Figure 11: Comparison of Task-Specific Manipulation Performance on Various Embodiments.

4.2 Generalization to Agilex Cobot Magic Embodiment

We evaluate the generalization capability of GE on the Agilex Cobot Magic platform using two complex tasks: *"box folding"* and *"cloth folding"*. For each task, we collect 250 high-quality teleoperated demonstrations—approximately 1 hour of data—using the Aloha-based teleoperation system (Fu et al., 2024). These demonstrations serve as the adaptation dataset to fine-tune both GE-Base and GE-Act.

As shown in Figure 11, we compare GE-Act with three state-of-the-art VLA models: GR00T N1 (Bjorck et al., 2025), π_0 (Black et al., 2024), and UniVLA (Bu et al., 2025b). All models are fine-tuned on the same dataset for these tasks. The experimental results clearly show that GE-Act outperforms all three models. While UniVLA and GR00T N1 demonstrate strong capabilities in simpler tasks like pick-and-place, their lack of precision in positioning and task execution leads to failure when confronted with complex and fine-grained tasks, achieving a success rate of 0%. Only with human intervention can UniVLA complete a few steps. In contrast, π_0, known for its strong performance in deformable object manipulation, surpasses UniVLA and GR00T N1 in these areas. However, GE-Act significantly outperforms π_0 in complex, fine-grained deformable object manipulation tasks. This enhancement is primarily due to the GE-Base foundation model, which, through large-scale pretraining on real-world data, enables GE-Act to achieve better task adaptation and embodiment generalization. As a result, GE-Act delivers superior performance across a wide range of robotic platforms and manipulation scenarios.

As illustrated in Figure 10, our adapted GE-Act model generates coherent, instruction-conditioned multi-view videos for the cloth folding and box folding tasks. These videos accurately capture both rigid and non-rigid object dynamics with high fidelity. The results demonstrate strong consistency across different camera views and showcase GE-Act's robust handling of complex object deformations. Furthermore, as shown in Figure 12, we present real-world executions of the cloth folding and box folding tasks using the adapted GE-Act model. These results confirm GE-Act's ability to

4.1 少样本适应

为了使GE-Act能够在新型机器人本体上实现小样本适应，我们采用了两阶段的任务特定微调策略，如图7最后两行所示：

- 在第一阶段，我们通过使用一小批新收集的指令条件视频演示对视频DiT模块进行微调，从而将视觉生成组件适配到新的具身化领域。在此过程中，CLIP（Radford等人，2021）和视频编码器保持冻结状态，以保留预训练的语义和感知先验。这一步骤使模型能够合成逼真且与具身化一致的操作视频，并使其与新平台的视觉特征相匹配。
- 在第二阶段，我们从头开始训练一个新的动作DiT模块，使用特定任务的遥操作轨迹。由于实体结构和动作空间语义存在根本性差异，我们并未复用预训练的动作解码器。相反，我们保留了GE-Base视觉主干网络，并针对新型机器人平台的控制动态和接口特性，学习了一个全新的动作头部。

这一两阶段适应流程有助于感知与运动能力的有效迁移，在极少数据监督的情况下，实现高保真视频生成以及精准的、由指令驱动的策略推断。

图11：不同实现方式下任务特定操作性能的比较。

4.2 对 Agilex Cobot Magic 实体的泛化

我们使用两个复杂任务——*"box folding"* 和 *"cloth folding"*，在Agilex Cobot Magic平台上评估了GE的泛化能力。对于每个任务，我们利用基于Aloha的遥操作系统（Fu等人，2024），收集了250个高质量的遥操作演示数据，大约1小时的数据量。这些演示数据用作适应数据集，以对GE-Base和GE-Act进行微调。

如图11所示，我们将GE-Act与三种最先进的VLA模型进行了对比：GR00T N1（Bjorck等，2025）、π_0 (Black等，2024)，以及UniVLA（Bu等，2025b）。所有模型均在相同的数据集上针对这些任务进行了微调。实验结果清晰地表明，GE-Act的表现优于这三种模型。尽管UniVLA和GR00T N1在抓取与放置等简单任务中表现出较强的能力，但在面对复杂且精细的任务时，由于其定位和任务执行的精确性不足，导致完全无法完成任务，成功率为0%。只有在人工干预下，UniVLA才能完成少数几个步骤。相比之下，π_0以其在可变形物体操作方面的强大性能而闻名，在这些领域超越了UniVLA和GR00T N1。然而，在复杂且精细的可变形物体操作任务中，GE-Act的表现显著优于π_0。这一性能提升主要得益于GE-Base基础模型，该模型通过在真实世界数据上的大规模预训练，使GE-Act能够实现更好的任务适应性和具身泛化能力。因此，GE-Act在广泛的机器人平台和操作场景中均展现出卓越的性能。

如图10所示，我们所改进的GE-Act模型能够为叠布和叠盒任务生成连贯且受指令控制的多视角视频。这些视频以高保真度准确捕捉了刚性和非刚性物体的动力学特性。结果表明，不同摄像机视角之间具有很强的一致性，并展示了GE-Act对复杂物体变形的强大处理能力。此外，如图12所示，我们利用改进后的GE-Act模型演示了叠布和叠盒任务的实际执行情况。这些结果证实了GE-Act具备以下能力：

Figure 12: **Visualization of Real-World Demonstrations with GE-Act on Agilex Cobot Magic Platform.** This shows GE-Act adapted to a novel Agilex Cobot Magic embodiment, performing real-world robotic manipulation tasks, including cloth-folding and box-folding.

complete the tasks with high precision and reliability on a novel robotic embodiment, further reinforcing the capacity of GE-Base to transfer effectively to new embodiments. This experiment solidifies the potential of GE-Base as a scalable, adaptable foundation for real-world embodied intelligence.

4.3 Generalization to Dual Franka Embodiment

We further evaluate GE's cross-embodiment generalization on the Dual Franka platform by performing embodiment- and task-specific adaptation of GE-Act using 250 teleoperated episodes (approximately one hour) on the cloth folding task. Due to the absence of a dedicated teleoperation interface, data collection on Dual Franka is conducted using a simpler space-mouse-based control system. Consistent with the Agilex Cobot evaluation, we adopt GR00T N1 (Bjorck et al., 2025), π_0 (Black et al., 2024), and UniVLA (Bu et al., 2025b) as baselines, and fine-tune each on the 250-episode adaptation dataset. Figure 13 illustrates the cloth folding task on the Dual Franka platform, including both the future-space video predictions by GE-Base and the real-world manipulation results executed by GE-Act. The results indicate that GE effectively models task-relevant visual dynamics and generalizes to new embodiments for precise manipulation. As shown in Figure 11, GE-Act consistently outperforms task-specific baselines in real-world execution on the Dual Franka platform, mirroring trends observed on the Agilex Cobot Magic. Notably, while π_0 and GR00T N1 were extensively trained on large-scale data from the Franka embodiment, GE-Act achieves superior performance with only one hour of adaptation data.

4.4 Generalization to RoboTwin

We further evaluate the cross-embodiment generalization on the dual-arm simulator RoboTwin (Chen et al., 2025). We adopt an all-in-one strategy, jointly fine-tuning GE-Act on four tasks using 200 demonstrations (50 per task), and directly evaluating this unified model across all tasks. In contrast, baseline methods (Black et al., 2024; Bu et al., 2025a) perform task-specific adaptation. As shown in Figure 11, GE-Act achieves better performance than π_0 and GO-1 on

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GR00T 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平台上布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GR00T 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。

(a) Multi-View Video Generation on the Dual Franka

(b) Real-World Robotic Manipulation on the Dual Franka

Figure 13: **Visualization of Robotic Video Generation and Real-World Manipulation on Dual Franka via GE.**

three of the four tasks, despite not using a one-task-one-model setting, and is only slightly behind VLA methods on lift pot. This minor gap may be attributed to task interference introduced by joint training.

5 GE-Sim: World Simulator

To support real-world-aligned evaluation and closed-loop control, we develop a video-based world neural simulator that generates temporally coherent visual predictions conditioned on robotic actions. This neural simulator enables embodied policy models to interact with a consistent visual environment, decoupled from physics-based constraints, and serves as a unified testbed for policy learning and generalization across diverse tasks.

We realize this capability by extending the GE-Base foundation model into an action-conditioned simulator, GE-Sim. In this framework, action trajectories serve as the primary control signals driving video synthesis over time. To implement GE-Sim, we adopt two GE-Base architectures: the fast LTX-Video–based variant used in GE-Act, and the COSMOS2 2B–based variant for high-fidelity simulation and realistic video generation. To maintain visual consistency across generated frames, we incorporate a reference image encoded by a frozen CLIP image encoder as a lightweight style anchor. This reference is injected via cross-attention into each DiT block, complementing the spatial grounding provided by visual observations.

A fundamental challenge in this transformation is reconciling the semantic disparity between low-level control commands and the high-level latent representations encoded by the pretrained world model. To address this, as depicted in Figure 14, we introduce a hierarchical action-conditioning mechanism that integrates structured action representations directly into the token space of GE-Base. This architecture preserves the model's pretrained spatiotemporal semantics while enabling seamless interfacing with a wide range of policy models, thereby facilitating closed-loop, action-conditioned neural simulation with robust generalization to diverse robotic tasks.

5.1 Hierarchical Action-conditioning Mechanism

To ensure compatibility with diverse action policy models, we adopt a general representation of robotic trajectories. For a single manipulator, each control step is encoded as a 7-dimensional vector $[x, y, z, roll, pitch, yaw, o]$, where (x, y, z) denotes the end-effector position, $(roll, pitch, yaw)$ its orientation (roll, pitch, yaw), and o the gripper openness. In our dual-arm setup, the control signal per step is represented by a 14-dimensional vector formed by concatenating both arms' control vectors. Over a K-step horizon, the complete action trajectory is denoted as $\mathbf{A} \in \mathbb{R}^{K \times 14}$. To bridge this low-level control signal with the token-based input interface of the GE-Base foundation model, we propose a hierarchical action-conditioning mechanism that incorporates both spatial and temporal components.

(a) Multi-View Video Generation on the Dual Franka

(b) Real-World Robotic Manipulation on the Dual Franka

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。

(a) Action-conditioned GE-Sim Framework **(b) Simulator and Data Engine**

Figure 14: **Overview of the GE-Sim World Simulator.** (a) GE-Base is transferred into an action-conditioned video generator for simulating robotic behavior given predicted actions. Spatial pose conditions are projected into image space and fused with historical visual inputs, while temporal motion deltas are concatenated with a reference image to preserve style consistency and injected via cross-attention into the generation model. (b) GE-Sim enables closed-loop policy evaluation and controllable data generation by producing action-conditioned video rollouts, supporting instruction-following and consistent trajectory replay under diverse visual contexts.

Pose2Image Conditioning. At each timestep i, the pose vector $a_i = [x_i, y_i, z_i, r_i, p_i, y_i, o_i]$ encodes the spatial position, orientation, and gripper state. The position (x_i, y_i, z_i) is projected into pixel coordinates using calibrated camera intrinsics and extrinsics. The orientation (r_i, p_i, y_i) is converted into a rotation matrix, whose orthonormal axes are also projected into the image plane to indicate directionality. The gripper openness o_i is rendered on a unit circle, with shading intensity reflecting its state—lighter for open, darker for closed. Distinct color encodings differentiate the left and right arms. This process yields pose images $\mathbf{P}_i$ that are spatially aligned with the visual scene.

Each $\mathbf{P}_i$ is paired with its corresponding sampled history frame $\mathbf{I}_i$. Both are encoded using a shared video encoder $\mathcal{E}$, and their latent features are fused by element-wise addition:

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

The resulting fused token $\mathbf{v}_i$ captures both contextual visual semantics and explicit pose information, and is inserted into the visual token stream for downstream processing.

Motion Vector Conditioning. To capture temporal dynamics, we compute motion deltas between consecutive end-effector poses. Let $\mathbf{a}_i = [\mathbf{p}_i, \mathbf{r}_i]$ denote the 6-DoF pose at timestep i, with $\mathbf{p}_i \in \mathbb{R}^3$ as position and $\mathbf{r}_i \in \mathbb{R}^3$ as orientation. The delta is given by:

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

which encodes both positional and orientational change. These deltas are encoded into motion tokens via a learnable encoder, concatenated with the reference image style token, and injected via cross-attention into each DiT block. This temporally-aware representation provides coherent motion priors to guide action-conditioned video generation within GE-Sim.

5.2 Training Procedure

To ensure high-fidelity video simulation required for action-conditioned generation, GE-Sim is initialized from the high-temporal-resolution pretrained model GE-Base-MR, which offers fine-grained modeling of robotic dynamics. The model is subsequently trained on the full AgiBot-World-Beta dataset, using ground-truth action trajectories as conditioning inputs for video generation. To improve generalization and robustness, the training corpus is augmented with a diverse set of failure cases—including erroneous executions, incomplete behaviors, and suboptimal control

(a) Action-conditioned GE-Sim Framework **(b) Simulator and Data Engine**

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：

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

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：

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

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。

Figure 15: **Visualization of Action-Conditioned Video Generation by GE-Sim.** Given a ground-truth action policy, we generate the corresponding next-frame prediction using GE-Sim. For each sample, we visualize the head-view image by overlaying the projected action target onto the current frame, alongside the predicted next frame, to illustrate the model's spatial alignment with the intended control signal.

trajectories—collected from both human teleoperation and real-world robotic deployments. During this phase, the VAE and CLIP encoders are kept frozen to preserve pretrained semantic and spatial priors, while the remaining parameters are optimized via a flow-matching loss applied over the predicted video representations.

5.3 Action-conditioned Video Generation

To evaluate the precision of action-conditioned video generation, we visualize simulation outputs from GE-Sim based on ground-truth control sequences. As shown in Figure 15, each example presents the current observation frame overlaid with the projected target position of the next action, along with the corresponding predicted frame synthesized by the simulator. Across tasks and viewpoints, the generated end-effector motion consistently aligns with the spatial intent of the action input, demonstrating that GE-Sim can accurately translate low-level control commands into coherent visual predictions. Furthermore, we compare GE-Sim built on two base architectures under identical action-conditioning in Figure 16. The COSMOS2-based variant exhibits higher visual fidelity and stronger temporal consistency than the LTX-Video–based model, confirming its superior capability in generating high-quality, action-aligned robotic simulations.

5.4 Closed-Loop Simulation

To support closed-loop evaluation of arbitrary policy models, GE-Sim functions as a video-based world simulator. Given a language instruction and initial visual observations, the policy model first takes these as input and outputs an action trajectory. GE-Sim then conditions on both the initial observations and the predicted action policy to generate a video chunk simulating the outcome of the action. This generated video is fed back into the policy model, along with the original instruction, to produce the next action step. This iterative process continues until the instruction is completed, enabling closed-loop simulation and real-world-aligned evaluation of policy models in a consistent and controllable visual environment.

Beyond policy evaluation, GE-Sim also functions as a versatile data engine. By executing the same action trajectory under different initial visual environments, it can generate diverse manipulation sequences reflective of varied contexts.

This video world simulator, grounded in real-world data, offers a compelling alternative to traditional physics simulators, achieving high visual fidelity while significantly reducing deployment costs. Crucially, it enables scalable, flexible simulation without requiring manual environment modeling. As such, GE-Sim lays the foundation for a new class of general-purpose, realistic, and low-cost world models that bridge learning and evaluation in embodied intelligence.

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，直到指令执行完毕，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的动作轨迹，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。

Figure 16: **Comparison of Multi-View Action-Conditioned Generation Results between Two GE-Sim Variants.** Visualization of GE-Sim outputs based on different base models: COSMOS2 and LTX-Video, under identical action conditions.

6 EWMBench: Embodied World Model Benchmark

An effective evaluation framework functions as a navigational instrument for scientific progress—establishing standardized criteria and fostering meaningful comparisons across methodologies. In the context of robotic world modeling, the ability to systematically assess whether a model faithfully captures the structure, dynamics, and semantics of embodied environments is essential for advancing the field. To this end, we propose the embodied world model benchmark, EWMBench, a comprehensive evaluation suite designed to measure both representational fidelity and practical utility of video-based world models in real-world robotic manipulation.

Beyond conventional benchmarks for general video generation that focus on visual fidelity, language alignment, or human preference, robotic manipulation videos introduce stricter structural constraints. In this domain, background layouts, object configurations, and embodiment structures (*e.g.*, robot morphology) should remain invariant, while only the robot's pose and interactions evolve in accordance with the instruction. EWMBench is designed with these domain-specific properties in mind, providing task-oriented metrics that assess visual-scene consistency, motion correctness, and semantic alignment and diversity, enabling more faithful and practical evaluation of world models in manipulation-centric scenarios. To support this, EWMBench comprises a high-quality real-world benchmark dataset and a suite of open-source evaluation tools, establishing a standardized framework for rigorously assessing the capabilities of video-based world models in manipulation-centric tasks.

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 Benchmark Dataset

The benchmark dataset is curated from the AgiBot-World-Beta test set by selecting 10 representative tasks spanning household and industrial domains. These tasks are characterized by well-defined operational goals and strong sequential dependencies, requiring procedural reasoning over affordances and action ordering. To ensure a fair evaluation, all selected tasks are disjoint from those used during the 1M-scale pre-training phase. Each task is decomposed into 4–10 atomic sub-actions, with every sub-action annotated with a step-level caption, enabling fine-grained alignment between video segments, action labels, and linguistic descriptions. For each task, we uniformly sample 100 video instances to construct a balanced and comprehensive evaluation set.

To promote diversity within each task, we implement a trajectory selection strategy based on spatial variation. Specifically, dual-arm end-effector trajectories are extracted and voxelized into 3D grids. A pairwise similarity matrix is computed using 3D Intersection-over-Union (IoU), and a greedy algorithm is employed to iteratively select the least-overlapping trajectories. This approach ensures broad coverage of motion patterns and minimizes redundancy within each task's evaluation set.

6.2 Evaluation Metrics

We establish a unified evaluation framework to assess how accurately video-based world models capture the spatial, temporal, and semantic dynamics of robotic manipulation.

Scene Consistency. To evaluate the structural and visual coherence of generated videos, we introduce a scene consistency metric that assesses the stability of visual appearance, environment layout, and viewpoint alignment across time. Specifically, we propose a patch-level feature similarity metric computed over consecutive and initial frames. We first fine-tune a strong visual encoder, DINOv2 (Oquab et al., 2023), on a robotic manipulation dataset to align its representation space with the embodied domain. For each frame, we extract patch-wise embeddings using this encoder. Then, cosine similarity is computed across corresponding patches between frames to quantify temporal consistency. Higher similarity scores reflect greater preservation of scene structure and camera viewpoint throughout the video sequence, indicating stronger spatial-temporal fidelity.

Action Trajectories Quality. To evaluate the quality of action trajectories executed in response to instructions, we manually annotate a reference trajectory for each instruction as the GT. For each generated video, a trained EEF detector is used to localize the gripper across frames and reconstruct the trajectory. Three video samples are generated per instruction, and the corresponding trajectories are extracted. Spatial alignment (SA) is assessed using the Symmetric Hausdorff (symH) Distance, which measures the maximum point-wise deviation between the generated trajectory P and the ground truth G. To ensure higher scores indicate better alignment, we report the inverse of this value:

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

To account for generation variance, the trajectory with the lowest symH is selected for further evaluation.

Temporal alignment (TA) is then evaluated using Normalized Dynamic Time Warping (NDTW) (Ilharco et al., 2019), which captures consistency in both sequence and timing between the generated and ground truth trajectories. To produce a positively correlated metric, we report the inverse of the NDTW distance:

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

Additionally, we introduce a Dynamic Consistency (DYN) metric to assess the realism of motion dynamics by comparing velocity and acceleration profiles between predicted and ground-truth trajectories. Specifically, we compute the Wasserstein distance $W(\cdot)$ between the respective time series, capturing distributional alignment without requiring strict temporal correspondence. To account for variations in motion amplitude and prevent instability in low-dynamic cases, we normalize each component using amplitude-aware ratios. The final score is defined as:

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

where $\Delta v = \max(v) - \min(v)$, $\Delta a = \max(a) - \min(a)$, $\epsilon = 10^{-8}$, and $\alpha = 0.007$, $\beta = 0.003$. This formulation ensures that the metric reflects both dynamic fidelity and amplitude robustness. This multi-level evaluation provides a comprehensive measure of spatial, temporal, and dynamic fidelity.

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：

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

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：

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

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：

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

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。

(a) Fine-Grained Evaluation of Video Generation Models

Model	Scene	Motion	Semantics	Score
GE-Base	0.9427	1.6676	2.0907	4.7010
Kling	0.8888	0.9440	2.0370	3.8698
Hailuo	0.8577	0.5362	2.0186	3.4125
COSMOS	0.7963	0.7085	1.7824	3.2872
OpenSora	0.9210	0.3442	1.8739	3.1392
LTX	0.9156	0.4002	1.6518	2.9676

(b) Aggregated Evaluation Across Hierarchical Levels

Figure 17: **Comprehensive Evaluation of Video World Models for Robotic Manipulation.** Leveraging our EWM-Bench , we systematically evaluate a suite of video world models sourced from state-of-the-art general video generation and embodied world modeling approaches. All models are assessed under a unified text-and-image to video generation paradigm. Evaluation spans multiple levels, scene, motion and semantics, capturing visual fidelity, temporal coherence, and semantic grounding in diverse real-world robotic manipulation tasks.

Motion Semantics Metrics. We evaluate motion semantics from two perspectives: semantic consistency and behavioral diversity. Semantic consistency assesses whether the generated manipulation behavior aligns with the intended task instruction, while diversity measures the model's ability to produce varied yet valid trajectories. For semantic consistency, we adopt a multi-granularity evaluation framework based on the VLM, Qwen2.5-VL-7B-Instruct (Bai et al., 2025):

- *Global-level alignment:* A VLM generates a compact summary caption for each generated video, which is then compared to the original task-goal instruction using BLEU scores to assess overall alignment between the video and intended task semantics.
- *Key-step consistency:* To assess whether essential sub-tasks are correctly executed, the VLM generates step-by-step descriptions for both the generated and ground-truth manipulation videos. Consistency is measured by computing CLIP-based similarity between the corresponding steps in the two descriptions.
- *Logical correctness:* To identify violations of physical or commonsense constraints, we first prompt GPT to define a taxonomy of typical logical errors in robotic manipulation videos, such as hallucinated actions, object disappearances, or physically implausible motions. Then, a video-based VLM is used to detect the presence of these predefined error types in generated videos. Detected violations are explicitly penalized, encouraging the model to produce semantically accurate and physically coherent manipulation behaviors.

To assess the model's capacity for generating varied outputs, we measure semantic diversity using CLIP-based global video embeddings. Specifically, we compute pairwise CLIP similarities between generated videos conditioned on the same instruction, and define the diversity score as 1 - CLIP similarity. Higher scores indicate greater semantic variability, reflecting the model's ability to generalize beyond deterministic execution.

6.3 World Model Evaluation

To thoroughly evaluate the effectiveness of video-based world models for robotic manipulation, we establish a comprehensive evaluation framework, referred to as the "evaluation colosseum," enabling direct, comparative analysis across various model architectures. In this framework, we benchmark seven state-of-the-art video generation models, including Open-Sora (Zheng et al., 2024), Kling (Kuaishou, 2025), Hailuo (MiniMax, 2024), LTX-Video (HaCohen et al., 2024) and the scene-centric COSMOS (Agarwal et al., 2025). All models are evaluated under a standardized text-and-image-to-video generation paradigm, where natural language instructions and head-view visual observations condition video synthesis. Notably, the GE-Base model is built on the LTX-Video architecture, which enables it to focus on domain-specific tasks and leverage fine-tuned control.

As shown in Figure 17, **GE-Base** consistently outperforms the baselines across multiple evaluation dimensions, with notable strengths in temporal alignment and dynamic consistency, two core metrics for generating action-plausible and temporally stable robotic behaviors. While performance in motion semantics is comparable to generic video generation

(a) Fine-Grained Evaluation of Video Generation Models

Model	Scene	Motion	Semantics	Score
GE-Base	0.9427	1.6676	2.0907	4.7010
Kling	0.8888	0.9440	2.0370	3.8698
Hailuo	0.8577	0.5362	2.0186	3.4125
COSMOS	0.7963	0.7085	1.7824	3.2872
OpenSora	0.9210	0.3442	1.8739	3.1392
LTX	0.9156	0.4002	1.6518	2.9676

(b) Aggregated Evaluation Across Hierarchical Levels

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，

models, GE-Base demonstrates much stronger control-aware generation fidelity, offering more precise and reliable task execution. This advantage is attributed to GE-Base's pretraining on large-scale robotic manipulation data, which better equips it to capture task-relevant spatial-temporal dynamics.

In comparison, Kling (Kuaishou, 2025) achieves strong overall performance, particularly in robustness across general video generation tasks, but lacks the specialized understanding required for fine-grained control, which limits its performance on more complex robotic manipulation tasks. Hailuo (MiniMax, 2024), though proficient in zero-shot embodied scenarios, often generates cartoon-like outputs that compromise visual realism, limiting its applicability for real-world robotic manipulation. COSMOS (Agarwal et al., 2025) and LTX-Video (HaCohen et al., 2024) models, while effective in human hand-centric tasks, struggle with adapting their semantic understanding to robotic contexts, and often produce inconsistent task execution. Notably, LTX-Video experiences abrupt scene transitions and a tendency to generate stationary states during action sequences, while COSMOS struggles with maintaining consistent viewpoints and camera control. Lastly, OpenSora (Zheng et al., 2024) displays partial understanding of task scenes and action semantics, but frequently suffers from jittery robotic arm movements and generates static videos, particularly in more complex tasks.

These results highlight GE-Base's advantages in bridging high-level semantic understanding with low-level control execution. Its superior performance in temporal alignment, dynamic consistency, and task adaptation positions GE-Base as a leading model for real-world robotic manipulation.

6.4 Simulation Evaluation

In addition to instruction-conditioned evaluation, we further assess the fidelity and reliability of our video-based simulator in an action-conditioned setting across two base models. Given ground-truth action trajectories, the simulator generates visual predictions conditioned solely on these control sequences. Under the EWMBench framework (Table 2), GE-Sim consistently demonstrates high spatial accuracy, precise trajectory execution, and strong semantic coherence. The limited visual diversity under fixed action inputs indicates accurate action-to-video correspondence and robust alignment with control dynamics. While GE-Sim built on COSMOS2 and LTX-Video achieves comparable overall performance, the COSMOS2-based variant excels in dynamic consistency, reflecting superior embodied video generation fidelity and temporal coherence.

Figure 18: **Consistency and Validity Analysis of Evaluation Metrics.** We compare human preference with our proposed EWMBench and the general video benchmark VBench to assess the consistency and reliability of automated evaluation metrics across different video world models.

Model	BLEU	CLIP	DYN	Div.	PSNR
LTX	0.33	90.8	0.78	0.011	19.9
COSMOS	0.31	90.2	0.85	0.010	20.7
Model	SA	Log.	TA	Scn.	
LTX	0.94	0.97	0.98	0.90	
COSMOS	0.87	0.97	0.97	0.91	

Table 2: **GE-Sim Evaluation on EWMBench.** Comparison of LTX-Video and COSMOS2 as base models for action-conditioned video generation. Metrics evaluate spatial, temporal, and semantic alignment with ground-truth control trajectories.

These results highlight that, beyond its efficiency, low cost, and ability to generalize across diverse environments, our video-based simulator provides a reliable and semantically consistent platform for action-conditioned evaluation in robotic manipulation.

6.5 Metric-Human Consistency

To validate the reliability and task relevance of our proposed EWMBench, we conduct a comparative analysis against human preference ratings and the general video benchmark VBench. We collect human annotations on videos generated by four representative models, GE-Base, Kling-1.6 (Kuaishou, 2025), Hailuo I2V-01-live (MiniMax, 2024), and OpenSora-2.0 (Zheng et al., 2024), using a ranking protocol where annotators assign ordinal scores based on perceived

模型方面，GE-Base展现出更强的控制感知生成保真度，能够提供更精准、更可靠的任务执行。这一优势归功于GE-Base在大规模机器人操作数据上的预训练，使其更能捕捉与任务相关的时空动态。

相比之下，Kling（快手，2025）在整体性能上表现出色，尤其在通用视频生成任务中的鲁棒性较强，但在需要精细控制的特定理解方面有所欠缺，这限制了其在更复杂的机器人操控任务中的表现。Hailuo（MiniMax，2024）虽然擅长零样本具身场景，但其生成的输出往往带有卡通风格，视觉真实感不足，从而限制了其在现实世界机器人操控中的应用。COSMOS（Agarwal等，2025）和LTX-Video（HaCohen等，2024）模型虽然在以人手为中心的任务中效果显著，但在将语义理解适配到机器人情境时却存在困难，经常导致任务执行不一致。值得注意的是，LTX-Video在动作序列中会出现突兀的场景切换，并倾向于生成静止状态；而COSMOS则难以保持稳定的视角和相机控制。最后，OpenSora（Zheng等，2024）对任务场景和动作语义的理解尚不完全，且在较为复杂的任务中经常出现机械臂抖动以及生成静态视频的问题。

这些结果凸显了GE-Base在连接高层次语义理解与低层次控制执行方面的优势。其在时间对齐、动态一致性以及任务适应性方面的卓越表现，使GE-Base成为现实世界机器人操作领域的领先模型。

6.4 仿真评估

除了基于指令的评估之外，我们还在两种基础模型的行动条件设置下，进一步评估了基于视频的模拟器在保真度和可靠性方面的表现。在给定真实动作轨迹的情况下，该模拟器仅根据这些控制序列生成视觉预测。在EWMBench框架下（表2），GE-Sim始终展现出较高的空间精度、精确的轨迹执行以及强大的语义连贯性。在固定动作输入条件下，视觉多样性的有限性表明动作与视频之间具有准确的对应关系，并且与控制动态实现了稳健的对齐。尽管基于COSMOS2和LTX-Video构建的GE-Sim整体性能相当，但基于COSMOS2的变体在动态一致性方面表现更优，反映出其在具身视频生成方面的更高保真度和更强的时间连贯性。

图18：评估指标的一致性和有效性分析。我们通过将人类偏好与我们提出的EWMBench以及通用视频基准VBench进行比较，以评估自动化评估指标在不同视频世界模型中的一致性和可靠性。

Model	BLEU	CLIP	DYN	Div.	PSNR
LTX	0.33	90.8	0.78	0.011	19.9
COSMOS	0.31	90.2	0.85	0.010	20.7
Model	SA	Log.	TA	Scn.	
LTX	0.94	0.97	0.98	0.90	
COSMOS	0.87	0.97	0.97	0.91	

表2：GE-Sim在EWMBench上的评估。对比LTX-Video和COSMOS2作为动作条件视频生成的基础模型。各项指标用于评估与真实控制轨迹在空间、时间和语义上的对齐情况。

这些结果表明，除了高效、低成本以及能够在不同环境中实现泛化之外，我们的基于视频的模拟器还为机器人操作中的动作条件评估提供了一个可靠且语义一致的平台。

6.5 度量与人类的一致性

为了验证我们提出的EWMBench的可靠性和任务相关性，我们将其与人类偏好评分以及通用视频基准VBench进行了对比分析。我们针对由四种代表性模型生成的视频收集了人工标注数据，这四种模型分别是：GE-Base、Kling-1.6（快手，2025年）、Hailuo I2V-01-live（MiniMax，2024年）和OpenSora-2.0（Zheng等，2024年）。我们采用了一种排序协议，由标注者根据感知情况分配序数评分。

overall quality. Rankings are aggregated across annotators and samples, with multiple review rounds to ensure annotation consistency. As shown in Figure 18, empirical results demonstrate that EWMBench rankings exhibit strong concordance with human judgments, effectively capturing dimensions of temporal alignment, semantic fidelity, and visual coherence. In contrast, VBench exhibits misalignment, particularly in scenarios demanding embodied consistency and goal-conditioned reasoning. These results confirm that EWMBench provides a more faithful and task-grounded assessment of video-based world models in robotic manipulation.

7 Related Works

World Models for Robotic Manipulation. The concept of world models as internal predictive representations for perception, planning, and control has long played a central role in robotics (Chatila and Laumond, 1985; Sutton and Barto, 1981). Early approaches relied on analytical modeling and system identification (Murray et al., 2017), requiring task-specific engineering and limited generalizability. The introduction of neural world models (Ha and Schmidhuber, 2018) enabled learning compact representations of dynamics directly from sensory inputs. These models have since evolved to operate in both pixel space (Ebert et al., 2018; Finn et al., 2016) and learned latent spaces (Hafner et al., 2019; Hu et al., 2024; Wu et al., 2023), with applications in control and planning. However, most prior efforts remain task-specific, constrained by limited interaction data. Recent advances propose general video-based world models trained on large-scale datasets (Agarwal et al., 2025; Bruce et al., 2024; Jang et al., 2025; Russell et al., 2025), yet these primarily focus on visual synthesis and do not support closed-loop robotic control. In contrast, our work develops a unified framework that integrates video-based world modeling with an action decoding module (GE-Act) and a closed-loop simulator (GE-Sim), enabling direct application in real-world robotic manipulation.

Video Generative Models for Robot Learning. Progress in video generation has led to powerful models capable of synthesizing high-quality videos from text or image prompts (Blattmann et al., 2023; Ho et al., 2022; OpenAI, 2024). While these models achieve impressive visual quality (Blattmann et al., 2023; OpenAI, 2024; Yang et al., 2024), their application to robotics remains limited by the lack of action conditioning, temporal coherence, and multi-view reasoning. Robotic manipulation requires models that can predict future states conditioned on action instruction, maintain long-term temporal consistency, and reason over spatially distributed observations. Action-conditioned video models (Bruce et al., 2024) have shown initial promise, and increasingly sophisticated systems are being developed, including driving (Russell et al., 2025) and robotic models (Agarwal et al., 2025), with camera controllability (Wang et al., 2024) further enhancing manipulation capabilities. However, existing methods are limited to single-view predictions and often lack comprehensive task understanding. GE-Base addresses these limitations through multi-view synthesis and autoregressive decoding with a memory mechanism, improving spatiotemporal consistency and task relevance.

Vision-Language-Action Models. Vision-language-action (VLA) models have emerged as a dominant paradigm in instruction-conditioned robotics (Black et al., 2024; Brohan et al., 2023; Driess et al., 2023; Kim et al., 2024). These models typically initialize from large-scale vision-language pretraining and are fine-tuned on robot demonstrations to predict action sequences. Although this approach has shown strong performance in diverse tasks, it suffers from inherent limitations. Behavior cloning restricts the agent to imitation without the ability to recover from errors or explore alternative strategies. The absence of explicit world models prevents internal simulation or reasoning over potential outcomes. Moreover, collecting high-quality teleoperation data remains a major bottleneck. Alternative approaches attempt to use VLMs as frozen encoders (Nair et al., 2022) or high-level planners (Ahn et al., 2022; Huang et al., 2023). Our framework takes a different approach by using vision-language inputs to condition a generative world model, enabling predictive reasoning and planning through internal simulation.

Policy Evaluation in Robotics. Efficient policy evaluation is essential for scaling robot learning. Traditional physics engines such as MuJoCo (Todorov et al., 2012) and Isaac Gym (Makoviychuk et al., 2021) provide fast simulation but require extensive manual tuning and still face a gap when transferring to the real world. Real-world evaluations, while more accurate, are slow and resource-intensive (Zhou et al., 2025). Recent efforts incorporate generative models into simulators (Authors, 2024; Nasiriany et al., 2024), offering new possibilities for efficient and scalable evaluation. However, many of these approaches are limited to simplified settings or restricted observation modalities. GE-Sim addresses these challenges by embedding robotic models within a generative loop that supports long-horizon manipulation across multiple views and includes both successful and failure-mode trajectories to improve robustness and reliability.

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。

Evaluation of Embodied World Models. Assessing the quality of embodied world models requires metrics that reflect performance in realistic manipulation scenarios. Traditional video generation metrics such as MSE or FVD do not correlate well with real-world task success. Recent benchmarks introduce structured evaluation protocols (Huang et al., 2024a,b) with broader metric coverage, but many still emphasize visual realism over task relevance. Specialized frameworks such as PhyGenBench (Meng et al., 2024) and T2V-CompBench (Sun et al., 2024) assess physical understanding and compositionality, respectively, but lack alignment with control objectives. Our EWMBench addresses this gap by providing a comprehensive evaluation suite focused on visual fidelity, motion consistency, semantic alignment, and action-conditioned controllability (Yue et al., 2025). It is specifically designed to assess the capabilities of video-based world models in the context of embodied robotics.

8 Limitations

In this work, we present a systematic investigation into world models for real-world robotic manipulation, addressing core challenges in visuomotor representation, policy learning, and embodied evaluation. While our Genie Envisioner framework lays a foundational path toward scalable and generalizable robotic intelligence, several limitations remain:

- *Data Coverage and Source Diversity.* Although we conduct cross-embodiment transfer experiments, our training relies exclusively on the AgiBot-World-Beta dataset—a large-scale yet single-platform real-world corpus. No internet-scale or simulation-based data sources are incorporated, limiting the diversity of embodiment types, sensor modalities, and scene configurations encountered during pretraining. While Genie Envisioner demonstrates promising generalization via few-shot adaptation, its robustness across heterogeneous sources and low-resource domains remains underexplored. Future extensions incorporating large-scale simulated or web-derived demonstrations will be critical for further expanding transfer capabilities.
- *Embodiment Scope and Dexterity.* The current study is confined to upper-body tabletop manipulation using parallel-jaw grippers. More complex embodiment settings, including dexterous hand coordination and full-body locomotion, are not addressed. These capabilities are crucial for real-world general-purpose robotics and warrant further integration into the Genie Envisioner framework to support fine-grained, multi-contact interactions and whole-body behaviors.
- *Evaluation Methodology.* While our EWMBench provides a structured evaluation of visual fidelity, action consistency, and language grounding, it still relies on proxy metrics and partial human validation. Fully automated and reliable assessment of task success—particularly under diverse failure modes and ambiguous semantics—remains an open challenge. Building scalable evaluation protocols that align closely with human judgment will be essential for robust benchmarking and safe deployment in real-world scenarios.

While Genie Envisioner is not yet a complete solution, it represents a meaningful step toward Genie, embodied AI systems with the potential for AGI-level manipulation capabilities.

9 Conclusion

In this work, we introduce Genie Envisioner, a unified and scalable platform for dual-arm robotic manipulation, leveraging high-fidelity video generation. At its core, GE-Base provides a robust foundation, capturing the spatiotemporal and semantic dynamics of robotic interactions for instruction-aligned video synthesis. The integration of GE-Act enables high-precision task execution, demonstrating not only strong performance across diverse in-domain tasks but also exceptional cross-embodiment generalization. Through minimal adaptation, GE-Act successfully transfers to novel robotic platforms and excels in complex tasks such as cloth folding and box packing. GE-Sim enhances the framework further by supporting closed-loop simulation, allowing for continuous policy refinement. EWMBench provides a comprehensive evaluation suite, ensuring robust assessment across visual realism, semantic alignment, and policy consistency. Extensive real-world evaluations confirm the superiority of GE-Base, GE-Act, and GE-Sim, establishing Genie Envisioner as a powerful foundation for building general-purpose, instruction-driven embodied intelligence.

Acknowledgment

We gratefully acknowledge the foundational contributions of prior works, including EnerVerse (Huang et al., 2025), EnerVerse-AC (Jiang et al., 2025), and EWMBENCH (Yue et al., 2025), which provided the inspiration and foundation

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。

for this research. We appreciate the AgiBot Genie Team for their invaluable contributions to data collection, real-world evaluation, and the provision of both robotic hardware and software support throughout this project.

References

A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, J. Bao, A. Benhaim, M. Cai, V. Chaudhary, C. Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, P. Chattopadhyay, Y. Chen, Y. Cui, Y. Ding, D. Dworakowski, J. Fan, M. Fenzi, F. Ferroni, S. Fidler, D. Fox, S. Ge, Y. Ge, J. Gu, S. Gururani, E. He, J. Huang, J. Huffman, P. Jannaty, J. Jin, S. W. Kim, G. Klár, G. Lam, S. Lan, L. Leal-Taixe, A. Li, Z. Li, C.-H. Lin, T.-Y. Lin, H. Ling, M.-Y. Liu, X. Liu, A. Luo, Q. Ma, H. Mao, K. Mo, A. Mousavian, S. Nah, S. Niverty, D. Page, D. Paschalidou, Z. Patel, L. Pavao, M. Ramezanali, F. Reda, X. Ren, V. R. N. Sabavat, E. Schmerling, S. Shi, B. Stefaniak, S. Tang, L. Tchapmi, P. Tredak, W.-C. Tseng, J. Varghese, H. Wang, H. Wang, H. Wang, T.-C. Wang, F. Wei, X. Wei, J. Z. Wu, J. Xu, W. Yang, L. Yen-Chen, X. Zeng, Y. Zeng, J. Zhang, Q. Zhang, Y. Zhang, Q. Zhao, and A. Zolkowski. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

G. Authors. Genesis: A universal and generative physics engine for robotics and beyond, December 2024. URL https://github.com/Genesis-Embodied-AI/Genesis.

S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

D. Berenson, S. S. Srinivasa, D. Ferguson, and J. J. Kuffner. Manipulation planning on constraint manifolds. In *2009 IEEE international conference on robotics and automation*, pages 625–632. IEEE, 2009.

J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al. A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

J. Bruce, M. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, Y. Aytar, S. Bechtle, F. Behbahani, S. Chan, N. Heess, L. Gonzalez, S. Osindero, S. Ozair, S. Reed, J. Zhang, K. Zolna, J. Clune, N. d. Freitas, S. Singh, and T. Rocktäschel. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.

Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Huang, S. Jiang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025a.

Q. Bu, Y. Yang, J. Cai, S. Gao, G. Ren, M. Yao, P. Luo, and H. Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025b.

R. Chatila and J.-P. Laumond. Position referencing and consistent world modeling for mobile robots. In *Proceedings of the IEEE International Conference on Robotics and Automation*, volume 2, pages 138–145, 1985.

T. Chen, Z. Chen, B. Chen, Z. Cai, Y. Liu, Q. Liang, Z. Li, X. Lin, Y. Ge, Z. Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2023.03378*, 2023.

用于本研究。我们感谢AgiBot Genie团队在整个项目过程中为数据收集、实际评估以及机器人硬件和软件支持所提供的宝贵贡献。

参考文献科学论文

A. 阿布埃莱宁、A. 阿什法克、A. 阿特金森、H. 阿瓦达拉、N. 巴赫、J. 包、A. 本海姆、M. 蔡、V. 乔杜里、C. 陈等。Phi-4-mini 技术报告：通过 LoRA 混合实现紧凑而强大的多模态语言模型。*arXiv preprint arXiv:2503.01743*, 2025。

N. 阿加瓦尔、A. 阿里、M. 巴拉、Y. 巴拉吉、E. 巴克尔、T. 蔡、P. 查托帕迪亚、Y. 陈、Y. 崔、Y. 丁、D. 德沃拉科夫斯基、J. 范、M. 芬齐、F. 费罗尼、S. 菲德勒、D. 福克斯、S. 葛、Y. 葛、J. 顾、S. 古鲁拉尼、E. 何、J. 黄、J. 休夫曼、P. 詹纳蒂、J. 金、S. W. 金、G. 克拉尔、G. 蓝、S. 兰、L. 莱阿尔-泰谢、A. 李、Z. 李、C.-H. 林、T.-Y. 林、H. 林、M.-Y. 刘、X. 刘、A. 罗、Q. 马、H. 毛、K. 莫、A. 穆萨维安、S. 纳、S. 尼弗蒂、D. 佩奇、D. 帕斯卡利杜、Z. 帕特尔、L. 帕瓦奥、M. 拉梅扎纳利、F. 雷达、X. 任、V. R. N. 萨巴瓦特、E. 施默林、S. 施、B. 斯特法尼亚克、S. 唐、L. 查普米、P. 特雷达克、W.-C. 曾、J. 瓦尔盖斯、H. 王、H. 王、H. 王、T.-C. 王、F. 魏、X. 魏、J. Z. 吴、J. 徐、W. 杨、L. 严辰、X. 曾、Y. 曾、J. 张、Q. 张、Y. 张、Q. 赵和 A. 索尔科夫斯基。物理人工智能的宇宙世界基金会模型平台。*arXiv preprint arXiv:2501.03575*, 2025。

M. Ahn、A. Brohan、N. Brown、Y. Chebotar、O. Cortes、B. David、C. Finn、C. Fu、K. Gopalakrishnan、K. Hausman 等人。见贤思齐，而非言不由衷：在机器人可供性中 grounding 语言。*arXiv preprint arXiv:2204.01691*，2022年。

G. 作者. Genesis：一种适用于机器人及其他领域的通用生成式物理引擎，2024年12月。网址 https://github .com/Genesis-Embodied-AI/Genesis。

S. 白，K. 陈，X. 刘，J. 王，W. 葛，S. 宋，K. 邓，P. 王，S. 王，J. 唐等。通义千问2.5-vl技术报告。*arXiv preprint arXiv:2502.13923*, 2025。

D. 贝伦森、S. S. 斯里尼瓦萨、D. 弗格森和 J. J. 库夫纳。约束流形上的操作规划。载于 *2009 IEEE international conference on robotics and automation*，第 625–632 页。IEEE，2009 年。

J. 比约克、F. 卡斯塔涅达、N. 切尔尼亚杰夫、X. 达、R. 丁、L. 范、Y. 方、D. 福克斯、F. 胡、S. 黄等。Gr00t n1：一种面向通用人形机器人的开源基础模型。*arXiv preprint arXiv:2503.14734*, 2025年。

K. 布莱克、N. 布朗、D. 德里斯、A. 埃斯梅尔、M. 埃奎、C. 芬恩、N. 富赛、L. 格鲁姆、K. 霍斯曼、B. 伊希特等。一种用于通用机器人控制的视觉-语言-动作流程模型。*arXiv preprint arXiv:2410.24164*, 2024。

A. 布拉特曼、T. 多克霍恩、S. 库拉尔、D. 门德列维奇、M. 基利安、D. 洛伦茨、Y. 莱维、Z. 英格利什、V. 沃莱蒂、A. 莱茨等。稳定视频扩散：将潜在视频扩散模型扩展至大规模数据集。*arXiv preprint arXiv:2311.15127*, 2023。

A. 布罗汉、N. 布朗、J. 卡尔瓦哈尔、Y. 切博塔尔、X. 陈、K. 乔罗马斯基、T. 丁、D. 德里伊斯、A. 杜贝、C. 芬恩、P. 弗洛伦斯、C. 付、M. G. 阿雷纳斯、K. 戈帕拉克里希南、K. 韩、K. 豪斯曼、A. 赫尔佐格、J. 侯赛因、B. 伊赫特、A. 伊尔潘、N. 乔希、R. 朱利安、D. 卡拉什尼科夫、Y. 匡、I. 利尔、L. 李、T.-W. E. 李、S. 列文、Y. 卢、H. 米夏莱夫斯基、I. 莫达奇、K. 佩尔茨、K. 拉奥、K. 雷曼、M. 瑞欧、G. 萨拉查尔、P. 桑凯蒂、P. 塞尔马内特、J. 辛格、A. 辛格、R. 索里库特、H. 特兰、V. 万豪克、Q. 武翁、A. 瓦希德、S. 韦尔克、P. 沃尔哈特、J. 吴、F. 夏、T. 肖、P. 徐、S. 徐、T. 余、以及 B. 齐特科维奇。Rt-2：视觉-语言-动作模型将网络知识迁移至机器人控制。*arXiv preprint arXiv:2307.15818*, 2023。

J. 布鲁斯、M. 丹尼斯、A. 爱德华兹、J. 帕克-霍尔德、Y. 施、E. 休斯、M. 莱、A. 马瓦兰卡尔、R. 施泰格瓦尔德、C. 应用、Y. 艾塔尔、S. 贝希特勒、F. 贝赫巴哈尼、S. 陈、N. 海斯、L. 冈萨雷斯、S. 奥辛德罗、S. 奥扎伊尔、S. 里德、J. 张、K. 索尔纳、J. 克卢恩、N. d. 弗雷塔斯、S. 辛格和 T. 罗克塔谢尔。Genie：生成式交互环境。载于 *International Conference on Machine Learning*，2024年。

Q. 布，J. 蔡，L. 陈，X. 崔，Y. 丁，S. 冯，S. 高，X. 何，X. 黄，S. 蒋等。Agibot世界斗兽场：一个用于可扩展且智能具身系统的大型操控平台。*arXiv preprint arXiv:2503.06669*, 2025a。

Q. 布，Y. 杨，J. 蔡，S. 高，G. 任，M. 姚，P. 罗和H. 李。Univla：学习以任务为中心的潜在行动，在任何地方采取行动。*arXiv preprint arXiv:2505.06111*，2025b。

R. 查蒂拉和J.-P. 劳蒙德。移动机器人的位置参考与一致的世界建模。载于*Proceedings of the IEEE International Conference on Robotics and Automation*，第2卷，第138–145页，1985年。

T. 陈，Z. 陈，B. 陈，Z. 蔡，Y. 刘，Q. 梁，Z. 李，X. 林，Y. 葛，Z. 顾等。Robotwin 2.0：一种具有强大领域随机化的可扩展数据生成器与基准，用于实现稳健的双臂机器人操作。*arXiv preprint arXiv:2506.18088*, 2025。

Z. 陈，J. 吴，W. 王，W. 苏，G. 陈，S. 邢，M. 钟，Q. 张，X. 朱，L. 陆等。Internvl：扩展视觉基础模型并适配通用视觉-语言任务。载于*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*，第24185–24198页，2024年。

D. 德里埃斯、F. 夏、M. S. M. 萨贾迪、C. 林奇、A. 乔杜里、B. 伊希特、A. 瓦希德、J. 汤普森、Q. 武昂、T. 余、W. 黄、Y. 切博塔尔、P. 塞尔梅内特、D. 达克沃斯、S. 列文、V. 万豪克、K. 豪斯曼、M. 托桑特、K. 格雷夫、A. 曾、I. 莫达奇和 P. 弗洛伦斯。PaLM-E：一种具身多模态语言模型。*arXiv preprint arXiv:2023.03378*, 2023。

F. Ebert, C. Finn, S. Dasari, A. Xie, A. Lee, and S. Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.

C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *NeurIPS*, 2016.

Z. Fu, T. Z. Zhao, and C. Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *Conference on Robot Learning (CoRL)*, 2024.

D. Ha and J. Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Y. HaCohen, N. Chiprut, B. Brazowski, D. Shalem, D. Moshe, E. Richardson, E. Levin, G. Shiran, N. Zabari, O. Gordon, P. Panet, S. Weissbuch, V. Kulikov, Y. Bitterman, Z. Melumian, and O. Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019.

J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. In *NeurIPS*, 2022.

E. S. Hu, K. Ahn, Q. Liu, H. Xu, M. Tomar, A. Langford, D. Jayaraman, A. Lamb, and J. Langford. Learning to achieve goals with belief state transformers. *arXiv preprint arXiv:2410.23506*, 2024.

S. Huang, Z. Jiang, H. Dong, Y. Qiao, P. Gao, and H. Li. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv preprint arXiv:2305.11176*, 2023.

S. Huang, L. Chen, P. Zhou, S. Chen, Z. Jiang, Y. Hu, Y. Liao, P. Gao, H. Li, M. Yao, et al. Enerverse: Envisioning embodied future space for robotics manipulation. *arXiv preprint arXiv:2501.01895*, 2025.

Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Z. Huang, F. Zhang, X. Xu, Y. He, J. Yu, Z. Dong, Q. Ma, N. Chanpaisit, C. Si, Y. Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024b.

G. Ilharco, V. Jain, A. Ku, E. Ie, and J. Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping. *arXiv preprint arXiv:1907.05446*, 2019.

J. Jang, S. Ye, Z. Lin, J. Xiang, J. Bjorck, Y. Fang, F. Hu, S. Huang, K. Kundalia, Y.-C. Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

M. Janner, J. Fu, M. Zhang, and S. Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

Y. Jiang, S. Chen, S. Huang, L. Chen, P. Zhou, Y. Liao, X. He, C. Liu, H. Li, M. Yao, et al. Enerverse-ac: Envisioning embodied environments with action condition. *arXiv preprint arXiv:2505.09723*, 2025.

M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Kuaishou. Kling. https://app.klingai.com/cn/, 2025.

B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

M. T. Mason. *Mechanics of robotic manipulation*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262133962.

F. Meng, J. Liao, X. Tan, W. Shao, Q. Lu, K. Zhang, Y. Cheng, D. Li, Y. Qiao, and P. Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

MiniMax. Hailuo AI. https://hailuoai.video, 2024.

R. M. Murray, Z. Li, and S. S. Sastry. *A mathematical introduction to robotic manipulation*. CRC press, 2017.

A. Nagabandi, K. Konolige, S. Levine, and V. Kumar. Deep dynamics models for learning dexterous manipulation. In *Conference on robot learning*, pages 1101–1112. PMLR, 2020.

F. 埃伯特、C. 芬恩、S. 达萨里、A. 谢、A. 李和 S. 列文。视觉预见：基于模型的深度强化学习在基于视觉的机器人控制中的应用。*arXiv preprint arXiv:1812.00568*, 2018年。

C. 芬恩、I. 古德费洛和 S. 列文。通过视频预测进行物理交互的无监督学习。载于*NeurIPS*，2016年。

Z. 付、T. Z. 赵和C. 芬恩。移动阿罗哈：利用低成本全身遥操作学习双手移动操作。载于 *Conference on Robot Learning (CoRL)*，2024年。

D. 哈和J. 施密德胡伯。世界模型。*arXiv preprint arXiv:1803.10122*，2018年。

Y. 哈科亨、N. 奇普鲁特、B. 布拉佐夫斯基、D. 沙勒姆、D. 莫舍、E. 理查森、E. 莱文、G. 希兰、N. 扎巴里、O. 戈登、P. 帕内特、S. 魏斯布赫、V. 库利科夫、Y. 比特曼、Z. 梅卢米安和 O. 比比。Ltx-video：实时视频潜在扩散。*arXiv preprint arXiv:2501.00103*, 2024。

D. 哈夫纳、T. 利利克拉普、I. 费舍尔、R. 维列加斯、D. 哈、H. 李和 J. 戴维森。从像素中学习潜在动力学以进行规划。载于*International conference on machine learning*，第2555–2565页。PMLR，2019年。

科学论文<|startofcontentleakJ. Ho、T. Salimans、A. Gritsenko、W. Chan、M. Norouzi 和 D. J. Fleet。视频扩散模型。载于 *NeurIPS*，2022年。E. S. Hu、K. Ahn、Q. Liu、H. Xu、M. Tomar、A. Langford、D. Jayaraman、A. Lamb 和 J. Langford。利用信念状态变换器学习实现目标。*arXiv preprint arXiv:2410.23506*，2024年。S. Huang、Z. Jiang、H. Dong、Y. Qiao、P. Gao 和 H. Li。Instruct2act：利用大型语言模型将多模态指令映射到机器人动作。*arXiv preprint arXiv:2305.11176*，2023年。S. Huang、L. Chen、P. Zhou、S. Chen、Z. Jiang、Y. Hu、Y. Liao、P. Gao、H. Li、M. Yao 等。Enerverse：为机器人操作构想具身未来空间。*arXiv preprint arXiv:2501.01895*，2025年。Z. Huang、Y. He、J. Yu、F. Zhang、C. Si、Y. Jiang、Y. Zhang、T. Wu、Q. Jin、N. Chanpaisit 等。Vbench：视频生成模型的综合基准套件。载于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，2024a年。Z. Huang、F. Zhang、X. Xu、Y. He、J. Yu、Z. Dong、Q. Ma、N. Chanpaisit、C. Si、Y. Jiang 等。Vbench++：视频生成模型的全面且多功能基准套件。*arXiv preprint arXiv:2411.13503*，2024b年。G. Ilharco、V. Jain、A. Ku、E. Ie 和 J. Baldridge。基于动态时间规整的指令条件导航通用评估。*arXiv preprint arXiv:1907.05446*，2019年。J. Jang、S. Ye、Z. Lin、J. Xiang、J. Bjorck、Y. Fang、F. Hu、S. Huang、K. Kundalia、Y.-C. Lin 等。Dreamgen：通过视频世界模型解锁机器人学习中的泛化能力。 *arXiv preprint arXiv:2505.12705*，2025年。M. Janner、J. Fu、M. Zhang 和 S. Levine。何时信任你的模型：基于模型的策略优化。*Advances in neural information processing systems*，第32卷，2019年。Y. Jiang、S. Chen、S. Huang、L. Chen、P. Zhou、Y. Liao、X. He、C. Liu、H. Li、M. Yao 等。Enerverse-ac：以动作条件构想具身环境。 *arXiv preprint arXiv:2505.09723*，2025年。M. J. Kim、K. Pertsch、S. Karamcheti、T. Xiao、A. Balakrishna、S. Nair、R. Rafailov、E. Foster、G. Lam、P. Sanketi 等。Openvla：一个开源的视觉-语言-行动模型。*arXiv preprint arXiv:2406.09246*，2024年。快手。Kling。https://app.klingai.com/cn/，2025年。B. Liu、Y. Zhu、C. Gao、Y. Feng、Q. Liu、Y. Zhu 和 P. Stone。Libero：面向终身机器人学习的知识迁移基准测试。*Advances in Neural Information Processing Systems*，第36卷：44776–44791页，2023年。V. Makoviychuk、L. Wawrzyniak、Y. Guo、M. Lu、K. Storey、M. Macklin、D. Hoeller、N. Rudin、A. Allshire、A. Handa 等。Isaac gym：用于机器人学习的高性能基于GPU的物理仿真。*arXiv preprint arXiv:2108.10470*，2021年。M. T. Mason。*Mechanics of robotic manipulation*。麻省理工学院出版社，美国马萨诸塞州剑桥市，2001年。ISBN 0262133962。F. Meng、J. Liao、X. Tan、W. Shao、Q. Lu、K. Zhang、Y. Cheng、D. Li、Y. Qiao 和 P. Luo。迈向世界模拟器：打造基于物理常识的视频生成基准测试。*arXiv preprint arXiv:2410.05363*，2024年。MiniMax。海洛AI。https://hailuoai.video，2024年。R. M. Murray、Z. Li 和 S. S. Sastry。*A mathematical introduction to robotic manipulation*。CRC出版社，2017年。A. Nagabandi、K. Konolige、S. Levine 和 V. Kumar。用于学习灵巧操作的深度动力学模型。载于*Conference on robot learning*，第1101–1112页。PMLR，2020年。

S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2022.

S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

OpenAI. Sora, 2024. URL https://openai.com/sora/.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

L. Russell, A. Hu, L. Bertoni, G. Fedoseev, J. Shotton, E. Arani, and G. Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

M. Stilman. Task constrained motion planning in robot joint space. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3074–3081. IEEE, 2007.

K. Sun, K. Huang, X. Liu, Y. Wu, Z. Xu, Z. Li, and X. Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*, 2024.

R. S. Sutton and A. G. Barto. An adaptive network that constructs and uses and internal model of its world. *Cognition and Brain Theory*, 4(3):217–246, 1981.

E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

Z. Wang, Z. Yuan, X. Wang, Y. Li, T. Chen, M. Xia, P. Luo, and Y. Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH*, 2024.

P. Wu, A. Escontrela, D. Hafner, P. Abbeel, and K. Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR, 2023.

Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

H. Yue, S. Huang, Y. Liao, S. Chen, P. Zhou, L. Chen, M. Yao, and G. Ren. Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694*, 2025.

Z. Zheng, X. Peng, T. Yang, C. Shen, S. Li, H. Liu, Y. Zhou, T. Li, and Y. You. Open-sora: Democratizing efficient video production for all, March 2024. URL https://github.com/hpcaitech/Open-Sora.

Z. Zhou, P. Atreya, Y. L. Tan, K. Pertsch, and S. Levine. Autoeval: Autonomous evaluation of generalist robot manipulation policies in the real world. *arXiv preprint arXiv:2503.24278*, 2025.

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. Teng、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/hpcaitech/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。