

Qwen-RobotWorld 技术报告：通过语言条件视频生成统一具身世界建模

Qwen-RobotWorld Technical Report: Unifying Embodied World Modeling through Language-Conditioned Video Generation

Qwen Team

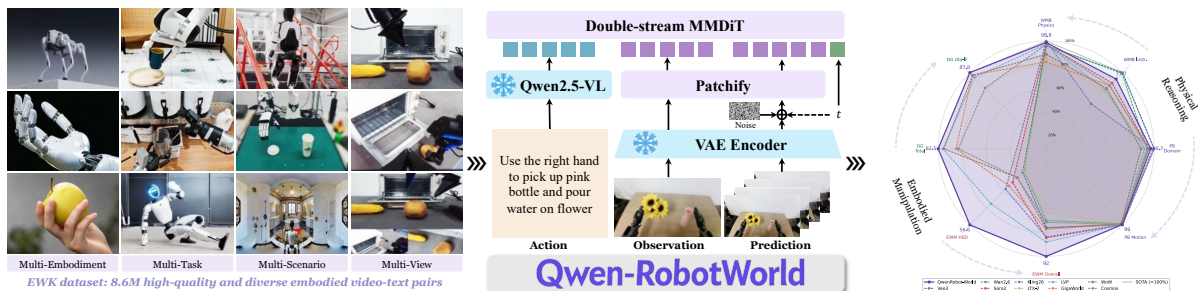
<https://qwen.ai/blog?id=qwen-robotworld>

摘要

引入 **Qwen-RobotWorld**，一个具身智能的语言条件视频世界模型。以自然语言为统一的动作接口，从当前观察预测物理合理的未来视觉轨迹——涵盖机器人操作、自动驾驶、室内导航和人-机转移。这个统一框架开启三个应用方向：为策略训练生成合成数据、为策略评估提供可扩展虚拟环境、为下游机器人控制提供语言引导的规划信号。

设计包含三部分。(a) **双流 MMDiT 与 MLLM 动作编码**：60 层双流扩散 Transformer 通过逐层联合注意力耦合冻结的 Qwen2.5-VL 语义与视频-VAE 潜在向量。(b) **具身世界知识 (EWK)**：8.6M 条视频-文本对话料库 (200M+ 帧)，涵盖 20+ 具身形态与 500+ 动作类别的动作-语言对齐。(c) **通用 + 专家渐进式课程**：两阶段训练策略，前期学习通用视觉先验，后期在共享语言接口下注入具身专门化。

广泛评测展现强竞争力：EWMBench 与 DreamGen Bench 整体排名第一，WorldModel-Bench 与 PBench 超越所有开源模型。RoboTwin-IF 零样本分析进一步证实稳健的泛化与多视图一致性。



1 引言

具身智能要求智能体在物理环境中感知、推理和行动——涵盖台面操作级的机器人操纵、城市交通中的自主导航和室内空间寻路。在真实世界直接训练此类系统成本高、效率低且风险大。世界模型提供了可扩展的替代方案：通过从观测数据学习环境动态，它作为交互训练平台，让具身智能体无需实际部署就能习得和优化行为。

世界模型可形式化为状态转移函数：给定当前状态 s_t 和动作 a_t ，它预测结果状态 $s_{t+1} = f(s_t, a_t)$ Ye 等 (2026)。在视频世界模型中，状态是视觉观测（视频帧或其潜在表示），模型基于当前观测和动作信号生成未来视觉轨迹。动作 a_t 可采取多种形式——低阶运动指令、高阶路径点轨迹或自然语言指令。其中，**自然语言是最通用、最易用的动作表示** Ye 等 (2026)：单个指令如“拿起红杯子放在架子上”隐含地编码了完整动作序列、目标状态和物理约束，无需机器人特定的控制接口。语言动作有两条互补的利用途径：作为显式输入，融入条件信号控制状态转移；或作为输出，事后从生成视频中推断出动作标签。这种灵活性使语言条件世界模型成为通用仿真骨干，跨具身平台泛化而无需重新设计接口。

然而，根本性张力目前限制了世界模型效果。通用视频生成模型 OpenAI (2024); Google DeepMind (2025) 在互联网规模数据上学得丰富视觉先验，但难以准确建模具身物理——接触动态、刚体结构约束和

本文为 AI 辅助翻译版本，仅供学习交流；引用请以英文原文为准：arXiv:2606.17030。

动作-结果关系都对物理合理性至关重要。相反，领域特定具身模型 Agarwal 等 (2025); Chen 等 (2025a); Team 等 (2025) 针对特定场景定制（如台面操作或驾驶），依赖机器人特定的动作表示如关节角或路径点。这类表示跨具身类型或任务类别无法泛化，根本限制了其跨平台仿真能力。

弥合这一差距，关键是以自然语言为统一动作接口，在通用视觉先验之上融合多样的具身经验。不同具身域贡献互补物理知识：操纵领域教授细粒度接触物理和工作空间内的状态变换；自动驾驶贡献大规模多智能体动态和场景几何；室内导航教授空间推理，要求语言指令转化为长视野的连贯视觉轨迹。这些域共享语言接口，联合训练时各自知识相互强化。此外，通过视频编辑将人类示范转换为机器人执行，为扩展具身训练数据超越物理机器人采集限制开启了实用途径。

我们提出 **Qwen-RobotWorld**——Qwen 系列首个语言条件视频世界模型，通过架构、数据和训练的协同创新实现上述愿景。除了高保真动作条件预测，该模型作为统一骨干，通过任务特定的适配，可支持三类代表性具身世界模型应用：合成数据引擎、策略评估环境和动作规划器。

架构：双流 MMDiT 和 MLLM 动作编码器 (§3)。实现语言条件状态转移，我们采用双流多模态扩散 Transformer (MMDiT) 骨干。理解流处理由冻结 Qwen2.5-VL 编码器提取的丰富语义特征，表示动作 a_t ；生成流处理视频兼容 VAE 的视觉潜在，表示视觉状态 s_t 。两流在每层通过联合注意力交互，实现整个去噪过程中的双向跨模态融合。用 MLLM 作动作编码器——而非 T5 Raffel 等 (2020) 或 CLIP Radford 等 (2021) 这样的轻量编码器——产生两个关键优势：(1) 其深层语言理解准确解析复杂组合指令为精确条件信号，控制细粒度状态转移；(2) 其内化的世界知识（如机器人臂是刚体，连杆长度固定、关节受约束）划定了物理上合理的转移空间。再配合 T2I 联合训练，模型无需显式几何提示就能避免帧间目标形变，而这正是缺乏语义扎根的模型最常见的失败。

数据：具身世界知识数据集 (§2)。为训练跨具身域泛化的状态转移函数，我们构造具身世界知识 (EWK) 数据集——约 860 万视频-文本对，包含超过 2 亿观测帧。语料跨越四个具身域及通用视频数据（占 30%）：**操纵**（约 590 万样本，20+ 机器人形态，1300+ 技能）提供核心具身基础；**自动驾驶**（约 20 万样本来自 Waymo、NVIDIA PhysicalAI-AD、Bench2Drive 和 Sekai）贡献大规模自我运动和多智能体动态；**室内导航**（VLNVerse 中 6000+ 语言引导集）提供扎根于连续轨迹的房间尺度空间推理；**人机转移数据**——通过自动化 MANO Romero 等 (2017) 到机器人流程跨 14 个机器人形态生成——支持跨具身视频编辑。我们的动作-语言映射框架作为核心方法论贡献，标准化 20+ 机器人具身类型和 500+ 动作类别为统一自然语言接口，产生约 860 万高质量跨场景、跨任务具身视频-文本对。这通过任务感知时间分割（确保每个样本捕获完整、明确定义的状态转移）和分层五层视点感知标注流程补充，标注的具体性和下游指令遵循能力都因此明显提升。

训练：从通用先验到具身特化 (§4)。我们采用两阶段渐进式训练课程。预训练中，跨 T2I、T2V 和 TI2V 任务的联合训练基于通用域数据构建基础视觉先验，T2I 特别锚定几何正确的目标形态转移到视频生成。在 SFT 阶段，具身数据通过四阶段混合计划渐进引入（70% 具身，30% 通用）：单视图操纵 → 多视图扩展 → 多视图连接生成 → 复杂任务和跨域数据。在具身部分中，操纵以 ~90% 采样权重主导以确保物理扎根深度，而多视图连接和导航/驾驶数据各获 ~5% 提供广度。这种通用 + 专家联合训练范式——在自然语言动作接口下统一——支持跨多样场景和任务的稳定联合训练，每个域的物理知识相互强化。非对称 3D RoPE 位置编码和多视图连接训练实现跨同步相机视图的几何一致合成而无需架构修改。

在四个既有基准上评估，QWEN-ROBOTWORLD 在跨场景和跨任务设置中达成竞争性能。它在 WorldModelBench 上超越所有开源模型（8.99，总体第 3），在牛顿定律、质量守恒、流体动力学和重力四项上物理遵循均为满分，与领先闭源模型相当，指令遵循也有 2.33/3.0 的成绩。在 EWM Bench 上排名**总体第 1**（4.60），HSD 中运动保真度明显领先（**0.566**，超领跑者 +33%）且场景一致性最佳（0.914）。在 DreamGen Bench 上，模型跨三个机器人具身子集排名**总体第 1**（4.952），在目标级组合泛化中出色。在 PBench 上，它超越所有开源模型（0.804），域理解总体排名第 3（0.857）且开源模型中运动平滑度排名第 2（0.990）。定性结果还展示了跨任务的泛化：人机转移中，模型不用机器人特定提示，就能从人类示范视频合成逼真的机器人执行；自动驾驶场景合成与房间尺度的室内导航生成同样成立。RoboTwin-IF 基准上的零样本成绩进一步印证，这种迁移在复杂指令下依然稳健。

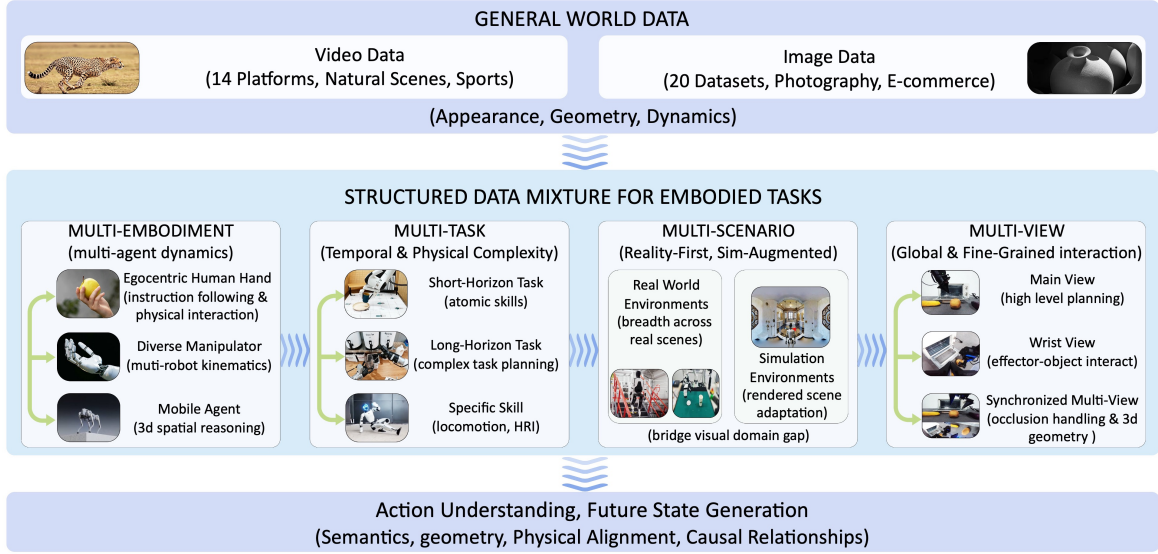


图 1: 具身世界知识 (EWK) 训练语料库概览。**通用世界数据** (上) 从互联网规模的视频和图像集合中提供外观、几何和动态的基础先验。**结构化具身数据** (中) 沿四个互补维度组织。每个维度针对不同的物理变异来源: 多具身 (人手、多样化机器人操纵器、移动智能体); 多任务 (短视界原子技能、长视界合成规划、特定技能如移动和人-机交互); 多场景, 采用现实优先、仿真增强的设计, 连接真实采集与下游 VLA 策略训练评估所用的仿真器; 多视图 (主视图、腕部视图、同步多视图流覆盖全局规划和细粒度末端执行器-物体交互)。这些信号共同提供语言条件动作理解和未来状态生成所需的语义、几何、物理对齐与因果关系 (下)。

我们的贡献总结如下:

- **框架。**我们提出 QWEN-ROBOTWORLD, 语言条件视频世界模型, 将自然语言视为通用动作接口统一跨场景和跨任务具身能力。通过在共享语言接口下联合训练操纵、驾驶、导航和人机转移, 模型达成任何单域模型都无法匹敌的互补物理泛化。
- **数据。**我们提出动作-语言映射框架, 标准化 20+ 机器人具身类型和 500+ 动作类别为统一自然语言接口, 并构造约 860 万高质量、跨场景、跨任务具身视频-文本对组成 EWK 数据集。
- **训练。**我们提出通用 + 专家联合训练范式, 在统一自然语言接口下, 装备模型兼具广泛世界建模能力和深层具身域专长, 支持跨多样场景和任务的稳定可扩展联合训练。
- **性能。**QWEN-ROBOTWORLD 在跨场景和跨任务具身评估指标上达成全面改进, 在 EWM Bench 和 DreamGen Bench 上排名总体第 1, 在 WorldModelBench 和 PBench 上超越所有开源模型。

2 数据

训练通用具身世界模型的核心挑战不仅是数据规模, 更在于表示异质性。机器人操纵动作表示为关节角或末端执行器路点, 驾驶表示为转向命令和速度轮廓, 导航表示为航向向量——每个都需要领域特定的模型或接口。我们通过**动作-语言映射**框架解决这一问题, 将来自 20+ 种机器人具身和 500+ 类动作的异质表示转换为统一的自然语言接口。在此统一接口下, Franka 夹爪、自主车辆和室内导航智能体的视频都成为同一语言条件视频生成任务的实例, 使单一模型能进行跨场景和跨任务联合训练, 无需任何领域特定控制接口。如图 1 所示, 该框架生成约 600 万高质量、跨场景、跨任务的具身视频-文本对, 我们进一步用通用视频数据 (占总体的 30%) 进行扩充, 构成具身世界知识 (EWK) 数据集: 共 860 万视频-文本对, 包含超 2 亿帧观察。

2.1 动作-语言映射

动作-语言映射框架针对具身数据的根本性不对称: 视觉状态 (视频帧) 已在共同的像素空间中, 但动作表示在不兼容的模态间碎片化。我们的框架通过将所有动作信号投影到共享的自然语言空间来解决这个

问题，使得同一个扩散 Transformer 能学习 $s_{t+1} = f(s_t, a_t)$ ，与底层物理域无关。

为何选择自然语言作为统一动作接口。 低级动作表示（关节角、末端执行器路点、力-矩命令）是硬件特定的，每个具身需要独立控制接口。相比之下，自然语言提供通用、具身无关的动作接口。单条指令“抓取红杯并竖直举起”隐含编码完整动作序列、目标状态和物理约束，无需了解运动学链。只以语言动作 a_t 为条件训练模型预测下一视觉状态 s_{t+1} ，我们便得到可跨具身泛化的仿真骨架——无论是 Franka 夹爪、Aloha 双臂系统还是人形机器人——无需重新训练或改造机器人特定接口。但这种泛化对标注质量要求极高：每条标注必须成为完整、自洽的动作规范，精确到模型能仅从 a_t 和 s_t 预测 s_{t+1} ，无需任何机器人元数据或本体感觉信号。

分层五级标注。 为在 20+ 种机器人具身和 500+ 类动作上一致生成动作丰富的标注，我们设计分层标注框架，包含五个递进层级。前三层构成结构化思维链，将每个视觉状态转移分解为可解释的部分：

1. **任务目标层**——推断转移的高层意图（ s_t 和 s_{t+1} 间应发生什么），融合外部指令与观察视频内容；
2. **动作细节层**——将动作 a_t 分解为时空轨迹、微动作、速度和力，**强制显式声明视点信息**（自中心主视图、腕部视图、外视图或多视图组合）；
3. **物理反馈层**——描述动作对环境的可观察后果（物体位移、变形、接触状态改变），使每个转移接地于可验证的物理结果。

基于此分析，生成两粒度的动作描述：

4. **详尽描述**（50–100 字）——完整指定视点-智能体-动作-反馈四元组，提供丰富动作信号用于精确状态转移预测；
5. **简洁描述**（15–30 字）——仅保留关键的视点-智能体-动作元素，使模型能处理推理时的简短高层命令。

我们执行四项质量控制原则：操作焦点（仅智能体动作和物体交互）、视点定义（显式视点类型和语义角色）、客观性（仅可见动态）和物理可验证性（仅可视验证结果）。训练时，我们等概率采样详尽和简洁描述（各 50%），使模型学会执行详细轨迹规范和简短任务级命令。

覆盖：20+ 种机器人具身，500+ 类动作。 该框架应用于所有数据域。在具身轴，它覆盖人手、七种机器人臂配置（单臂夹爪、双臂夹爪、单臂灵巧手、双臂灵巧手、移动双臂、半人形、全人形）、自中心车辆（环视摄像头）、行人/无人机和移动导航智能体——总共代表 20+ 种不同机器人具身，来自 RoboCoin（三大结构类 15 个模型）、Robomind（4 种形态）、InternData-A1（4 个模型）、Groot-XE 和其他数据集。在动作轴，它跨越 500+ 动作类别，源自训练数据集间的显式运动原语词表——仅 Agibot-World 就定义 84 种不同操纵原语（抓取、推、倒、折、擦、切等）——补充来自其他操纵数据集的独特原语和移动/导航动作（转向、换道、路点跟随、避障等），组织为四层：(1) 操纵原语、(2) 长视界组合、(3) 移动和导航、(4) 动态和可变形交互。这种系统覆盖确保生成的具身视频-文本对跨越任何单一领域都无法提供的语义丰富、物理多样的动作空间。

2.2 数据采集

2.2.1 通用数据

通用世界数据为模型掌握基本物理规律和形成精确视觉表示奠定基础。这一类包含互联网中多样化的视频和静态图像。视频数据标准化为 24 FPS，支持多种分辨率和宽高比（1:1、2:3、3:2、3:4、4:3、9:16、16:9 等）。图像数据整合高质量静态照片，作为视觉质量锚点，建立物体外观、材质纹理和空间组成的精确表示。所有通用数据用 Qwen2.5-VL [Bai 等 \(2025\)](#) 生成的自然语言描述进行标注；标注省略视点特定信息以保持灵活性和泛化性。对 AI 生成内容（AIGC）我们采取保守态度：通用数据排除 AI 生成的图像和视频，因为这些常引入视觉伪影、物理不一致和隐性偏差，可能削弱模型的泛化能力。

表 1: 具身世界知识（EWK）训练数据混合的详细清单，按领域组织。

数据集	具身	视图	任务	贡献
A. 操纵				
EgoHOD Pei 等 (2025), EPIC-Kitchens Damen 等 (2018), Egocentric-10k Build AI (2025)	人手	自中心视角	日常抓取与厨房	灵巧性与协调先验
Bridge V2 Walke 等 (2023), RH20T Fang 等 (2024), Droid Khazatsky 等 (2024)	单臂夹爪	外视图 + 腕部	桌面拾取放置	交互原语
Robomind Wu 等 (2025a), RoboCoin Wu 等 (2025b)	单/双臂、人形	自中心 + 外视	刚性与可变形物体	跨具身泛化
Agibot-World AgiBot-World-Contributors (2025), Galaxea Galaxea AI (2025)	单臂 (夹爪 + 灵巧手)	同步自中心 + 腕部 + 外视	长视界序列	时间与多视图一致性
Qwen-Aloha (internal)	双臂夹爪	头部 + 双腕部	多样抓取	多视图抓取先验
ActionNet Fourier ActionNet Team & Mu (2025), OpenLoong OpenLoong Baihu Team (2025)	灵巧手	腕部 + 外视	工具使用与手内操作	细粒度灵巧性
InternData-A1 Tian 等 (2025), Robotwin Chen 等 (2025b), Groot-XE Bjorck 等 (2025), RT1 Brohan 等 (2023)	混合臂 (仿真)	可变	流体与可变形物体	仿真到现实对齐
B. 自动驾驶				
Waymo E2E Waymo Team (2024), NVIDIA PhysicalAI-AD NVIDIA (2025b)	自中心车辆	5-8 环视	城市驾驶与交通	大规模运动与 3D 几何
Bench2Drive Jia 等 (2024)	自中心车辆 (仿真)	6 环视	9.8K 交通场景	仿真多样性与真值标注
Sekai Sekai Team (2025)	行人/无人机	自中心视角	城市行走	行人尺度移动
C. 室内导航				
VLNVerse Lin 等 (2025)	移动智能体	自中心视角	134 室内场景、语言引导	3D 推理与语言-轨迹对齐
D. 人-机迁移				
配对的人-机数据集	人 → 14 个机器臂	自中心双臂	跨具身操纵	视频编辑监督

2.2.2 具身操纵数据

为使世界模型在多场景多任务间获得接地的物理理解，我们构建结构化数据混合，涵盖操纵、驾驶、导航和跨具身迁移领域，见表 1。对于核心操纵领域，我们围绕四个维度组织数据：多具身、多任务、多场景和多视图。

多具身。 操纵数据涵盖人手、单臂夹爪、双臂灵巧系统、移动操纵器和全身人形等多种具身，教模型将任务级意图与具身运动学分开。人类操纵数据（EgoHOD Pei 等 (2025)、EPIC-Kitchens Damen 等 (2018)）示范了人类灵巧性的上限：模型观察物理上优雅的交互方式，学得眼手协调和工具使用的先验。机器人数据则展示这些意图在不同机械形态上的实现。通过同时学习重叠任务的人类示范和机器人执行（Robomind Wu 等 (2025a)、RoboCoin Wu 等 (2025b)），模型习得具身无关的动作语义——二指夹爪还是七指灵巧手，都能预测接下来该发生什么。

多任务。 操纵语料库涵盖从原子接触层动作到扩展多步过程的技能层级，教模型在多种时间粒度上操作。短视界数据集（Bridge V2 Walke 等 (2023)、RH20T Fang 等 (2024)）提供基础交互原语——抓取、推、插入——的密集覆盖，接地模型对接触物理和物体可供性的理解。长视界数据集（Agibot-World AgiBot-World-Contributors (2025)、Galaxea Galaxea AI (2025)）将这些原语链接成连贯序列，强制模型跨数十步维持状态跟踪和因果推理。另外，动态交互数据集（Humanoid Everyday Zhao 等 (2025)）引入高速度、全身运动，测试模型在显著动量和平衡约束下预测结果的能力。总体而言，这样的范围确保模型能推理“按这里会发生什么”和“经过十次顺序决策后会发生什么”。

多场景。多场景覆盖沿两个互补方向推进：真实环境的多样性和仿真渲染场景的补充。一方面，物理交互在不同情境下表现各异——厨房台面的光照、杂乱程度、材质特性不同于工厂车间或露天工地。因此我们的操纵数据主要源自真实场景，覆盖家庭厨房、工作坊、实验室和非结构化室外环境，让模型接触真实存在的照明变化、遮挡、材质多样和背景复杂度，防止过度适应单一环境。另一方面，我们加入光度逼真的仿真数据（InternData-A1 Tian 等 (2025)）作为重要补充。这源于 VLA 应用现状：大量策略模型在仿真环境中训练和评估，广泛采用 LIBERO Liu 等 (2023)、SimplerEnv Li 等 (2024)、RLBench James 等 (2020) 等基准。通用世界模型需在仿真风格的视觉和物理下可靠生成，消除真实与合成观察间的视觉差异，同时支持仿真到现实的迁移和闭环评估。仿真部分还能精确控制光照、物体位姿和摄像机位置的变化，强化视觉鲁棒性。

多视图。单视图数据教模型从固定视角预测合理的未来，但许多物理关键事件从任何单摄像机都被部分或完全遮挡。同步多视图记录（Agibot-World AgiBot-World-Contributors (2025)、Robomind Wu 等 (2025a)）让模型同时从头部、腕部和外部视角观察同一事件。这有两个目的：训练时，跨视图对应作为几何正则化器，隐含教模型物体形状、深度和空间关系；推理时，模型能从任何单个视角生成，或组合保持相互一致的多视图输出。我们 600 万具身样本中约 160 万包含同步 2-4 视图拼接，提供大量多视图监督而不支配语料库。

2.2.3 自动驾驶数据

操纵数据在有限工作空间内捕获细粒度物体交互，而自动驾驶数据使模型接触远大得多的运动空间，多样化机动（转向、换道、加速）跨越更大速度和轨迹范围。驾驶场景还包含丰富多智能体动态——周围车辆、行人和骑行者在交通规则下交互——要求世界模型学习多个物体如何随时间移动、遮挡和相互影响。此外，大摄像机位移通过视差和透视变化提供密集的 3D 场景几何监督信号，增强模型的视图合成和空间推理能力。

我们汇聚四个大规模驾驶数据集：Waymo E2E Waymo Team (2024)（真实驾驶，8 环视摄像机，7044 段 / 11.3 小时）、NVIDIA PhysicalAI-AD NVIDIA (2025b)（真实驾驶，5 摄像机 30°–120° 视角，1,342,418 段 / 1715.9 小时）、Bench2Drive Jia 等 (2024)（CARLA 仿真，9881 个交通场景，6 摄像机，384,948 段 / 511.2 小时）和 Sekai Sekai Team (2025)（行人和无人机第一人称视频，9995 段 / 166.6 小时，标注了场景和天气）。合计 1,744,405 段共 2405 小时。处理分三个阶段：(1) 逐帧提取与轨迹统一为路点格式；(2) 按机动事件分割为 2–8 秒片段；(3) 生成标注，融合轨迹描述与可选的 VLM 增强。

2.2.4 自中心室内导航数据

自中心室内导航数据从另一角度补充操纵和驾驶数据。操纵专注工作台内的细微交互，驾驶在开放场景中决策方向和速度，室内导航则要求模型理解房间布局、规避障碍、把文本命令转化为连贯的视觉运动。

遵循 VLNVerse Lin 等 (2025)，我们用 NVIDIA Isaac Sim NVIDIA (2022) 采集物理接地的自中心导航数据，具有光度逼真渲染和连续控制。我们在 134 室内场景跨收集 6,064 次成功导航情节，每个包含自中心 RGB 视频（256 × 256 分辨率 10FPS）配对自然语言导航指令。轨迹平均长度约 8.2m（范围 4 到 17.5m），累计总遍历距离约 49.8km 和约 5.8 小时连续第一人称导航视频。指令以两种格式提供：单字符串逐步指令（3,031 情节、平均 67.2 词）和多粒度描述覆盖正式、自然、随意三种语体（3,033 情节）。

在此类遍历数据上训练的视频生成模型能学到涌现的 3D 一致性和帧间空间一致性 Gao 等 (2026); Bar 等 (2025)，同时每个序列的物理接地、动作条件特性鼓励模型内化深度推理、几何一致性和避障规划 Shang 等 (2025); Han 等 (2025); Zhen 等 (2025)。通过在连续自中心遍历中接地语言指令，此数据使世界模型能在室内环境中联合学习语言理解、3D 空间推理和具身动作预测。

2.2.5 人-机迁移数据

为在无需物理机器人采集的情况下训练跨具身视觉对应模型，我们策划两个互补的人-机迁移数据源。首先是大规模人-机配对数据集，由自中心双臂操纵录制通过自动化管道构成：3D 手部关键点通过 MANO Romero 等 (2017) 重建和重定向到机器人末端执行器轨迹提取，人手通过视频修复移除，14 个

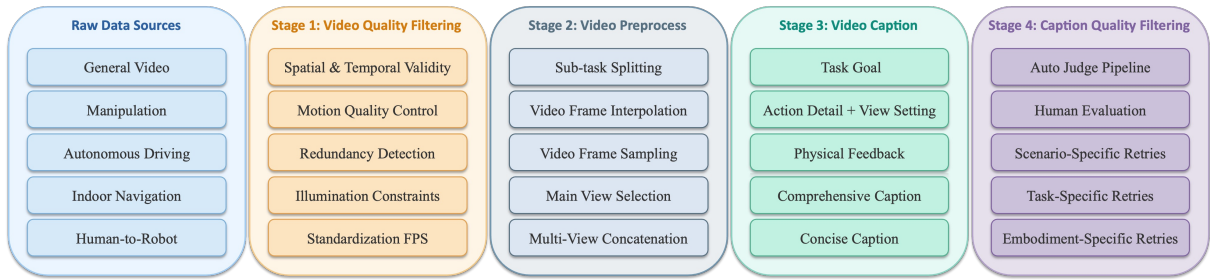


图 2: 统一数据处理管道概览。阶段 1（原始数据收集）从跨通用和具身领域的五个源类别收集异质数据。阶段 2（视频预处理）应用领域自适应操作——帧提取、帧插值、子任务分割、主视图选择和多视图拼接——产生均匀结构的片段。阶段 3（分层标注）通过五层框架生成视点感知标注：任务目标、动作细节、物理反馈、详尽标注和简洁标注。阶段 4（标注质量筛选）结合自动化 LLM 判断器与人工评估；表现不佳的标注路由回以进行场景、任务或具身特定的迭代提示改进。

机器人臂模型用 MuJoCo Todorov 等 (2012) 逆运动学渲染到修复场景，产生每个情节四个对齐视频流（原始人类视频、去手场景、纯仿真、机器人覆盖场景）。14 种具身在共享场景中的多样性确保编辑能力跨机器人形态泛化。

第二个源解决直接渲染的根本限制：简化渲染器忽略场景照明、投射阴影和材质相关镜面反射，在渲染和真实观察间创建光度差。为弥合这一差距，我们基于开源 InternA1 数据集构建 Tian 等 (2025)，使用 NVIDIA Isaac Sim NVIDIA (2022) 提供具有环境照明和准确阴影的光度逼真 RGB 观察。使用相同动力学参数和机器人 URDF，我们在 MuJoCo Todorov 等 (2012) 中渲染匹配的自中心视图——不含照明或阴影效果——产生在相同几何和视点下共享但光度逼真度不同的配对样本。这个配对数据使模型能学习简化渲染和光度逼真观察间的视觉映射，覆盖 Franka Emika Panda、AgileX Split Aloha、ARX Lift2 和 AgiBot Genie1 跨单臂、双臂、移动双臂和人形配置，约 80K 情节跨拾取放置、关节物体操纵和多物体重新排列任务。

2.3 数据处理

我们设计统一数据处理管道，将来自多样具身和通用视频源的异质原始数据转换为高质量、一致格式的训练样本。如图 2 所示，管道包含四个阶段：(1) 原始数据收集、(2) 视频预处理、(3) 分层标注和 (4) 标注质量筛选与迭代提示改进。阶段 2 和 3 根据源数据特性应用领域自适应操作，阶段 4 形成闭环反馈，将表现不佳的标注路由回以进行针对性重新标注。

2.3.1 阶段 1：原始数据收集

管道通过从跨通用和具身领域的五个源类别摄入原始视频数据开始。**通用视频**从纪录片、专业素材库和精选网页片段提供互联网规模的视觉多样性。**操纵**数据覆盖广谱机器人具身——单臂夹爪、双臂系统、灵巧手、移动平台和人形——来自包括 EgoHOD、Bridge V2、DROID、RoboMind、Agibot-World 等数据集。**自动驾驶**贡献来自 Waymo、Bench2Drive、NVIDIA PhysicalAI-AD 和 Sekai 的大规模自中心运动和多智能体动态。**室内导航**从 VLNVerses 跨 134 室内场景提供语言引导的空间推理情节。**人-机迁移**通过我们跨 14 种机器人类型的自动化 MANO 到机器人管道提供配对的人类示范和机器人执行数据。

2.3.2 阶段 2：视频预处理

原始视频经历领域自适应预处理以产生适合训练的均匀结构片段。我们根据源数据特性应用五个互补操作：

帧提取。对于短视界任务视频（通常持续 2–8s 的单步操纵），我们按目标速率提取捕获交互基本阶段——接近、接触、操纵和结果——的帧，确保每个样本包含原子动作的完整因果链。

帧插值。当源视频帧率不足以进行平滑运动学习时，我们应用时间插值增加帧密度，保留对建模细粒度接触动力学和物体状态转移至关重要的连续运动轨迹。

子任务分割。对于涉及多步过程的长视界情节（如顺序拾取放置、复杂组装），我们将视频分解为语

义连贯的子任务段。每个段捕获具有清晰开始和结束状态的完整原子动作，避免盲目均匀截断带来的半截动作伪影。

主视图选择。对于仅需要主视点的多摄像机录制（如单视图操纵训练），我们选择最具信息量的摄像机流——通常是最好捕获交互区域的自中心或外视——丢弃冗余角度。

多视图拼接。反之，对于多视图联合训练，我们将来自 2-4 摄像机视角的同步片段拼接为单一水平布局，保持跨视图的时间对齐。这使模型能学习跨视图几何一致性和同步状态转移而无需架构修改。

2.3.3 阶段 3：分层标注

分层标注提示模板

你是一位具身 AI 标注专家。给定 `{{embodiment_type}}` 从 `{{viewpoint_type}}` 视点执行操纵任务的视频片段，生成以下五层标注。

— 分析阶段 —

第 1 层-任务目标：识别此交互的高层意图。智能体试图实现什么？用一句话描述从当前观察到目标状态的期望状态转移。

第 2 层-动作细节：将智能体的动作分解为逐步序列。对每一步，指定：(a) 运动轨迹和方向、(b) 微动作（接近、抓取、抬起、旋转、释放等）、(c) 估计速度和力级别。**你必须显式说明视点**（自中心/腕部/外视/多视图拼接）。

第 3 层-物理反馈：描述每个动作对环境的可观察物理后果：物体位移、变形、接触状态改变和任何二级效应（如液体晃动、布料折叠）。仅包含可验证的结果。

— 生成阶段 —

第 4 层-详尽标注（50–100 词）：将第 1-3 层合成一个连贯段落，完整指定视点-智能体-动作-反馈四元组。包含摄像机视角、具身身份、完整动作序列和物理结果。

第 5 层-简洁标注（15–30 词）：凝聚为指令风格摘要，仅保留关键视点-智能体-动作元素，适合作为世界模型的直接语言命令。

质量约束：

- 操作焦点：仅描述智能体动作和物体交互；省略背景叙述。
- 视点定义：显式命名视点类型及其语义角色。
- 客观性：仅报告可见动态；不推断隐藏状态。
- 物理可验证性：每个声称的结果都必须从视频中可验证。

预处理视频通过我们的五层分层标注框架（第 2.1 节）传递，该框架在两种粒度——详尽（50–100 词）和简洁（15–30 词）——生成视点感知标注，在训练期间等概率采样。标注模型使用的提示模板如上所示。

2.3.4 阶段 4：标注质量筛选

为确保语料库中多样场景、任务和具身的标注质量，我们实现结合自动化评估与人工监督的闭环质量筛选系统。未通过质量检查的标注路由回阶段 3 进行针对性重新标注，形成迭代改进环。

判断管道。自动化 LLM 判断器沿多个维度评估每条标注，包括事实准确性、特异性、指令清晰度和视点一致性。具体而言，它评估标注是否正确描述视频内容、是否提供超出通用描述的足够细节、是否能作为可行命令、是否维持与摄像机视角一致的空间参考。不满足任何这些标准的标注被标记进一步审查。

人工评估。标注子集——特别是那些接近判断阈值或来自代表不足领域的——经历人类标注员的人工审查，验证正确性、识别系统性失败模式，提供后续提示改进的基准真值校正。

迭代提示改进。若判断管道在某类别发现一致性问题，我们沿三个方向重新设计提示：**针对场景**（如户外光照、厨房环境）、**针对任务**（如关节物体操纵、液体倾倒）、**针对具身**（如人形双臂配合、灵巧手操作）分别重试。每次重用针对性强的提示模板，改进后的标注重新经过判断管道，直至通过质量关。这个循环保证每类场景、任务、具身都能获得高质量标注，而不是沿用模板。

最终语料库统计。完整四阶段管道后，最终训练语料库包含约 860 万视频-文本对（超 2 亿观察帧），

具身数据占 70% 而通用数据占 30%。在具身部分中，单视图操纵数据构成多数约 430 万样本，其次是约 160 万多视图拼接样本带同步 2-4 摄像机视图，和约 20 万导航和驾驶样本。

3 模型

3.1 模型架构

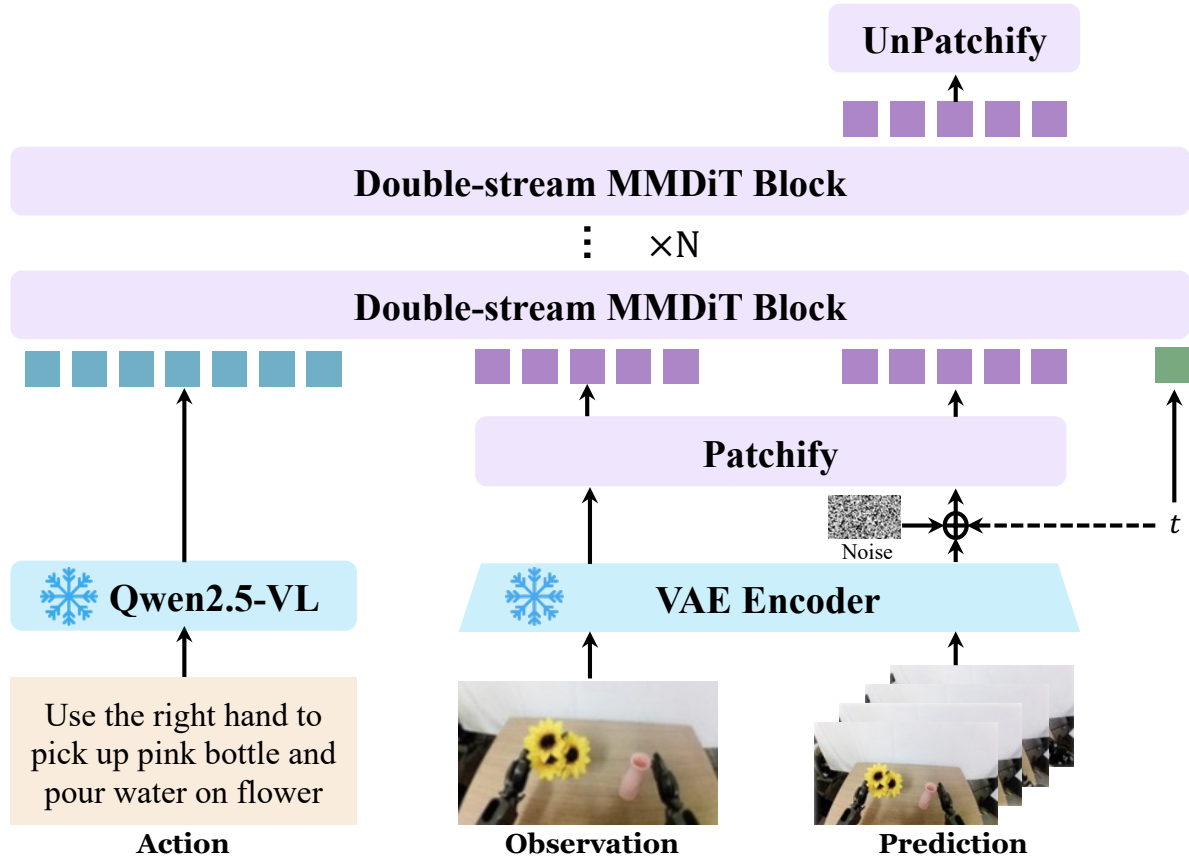


图 3: 我们的视频生成架构概览，包含 60 层双流 MMDiT 骨干。

如图 3 所示，模型包含三个核心组件：作为动作编码器的 MLLM、作为状态编码器/解码器的 VAE，和作为转移函数的 MMDiT Esser 等 (2024)，采用双流架构组织。

MLLM——动作编码器。我们采用冻结的 Qwen2.5-VL Bai 等 (2025) 将用户输入编码为条件信号。对于给定输入文本 S ，它提取最后层隐状态 $\mathbf{h} = \phi(S)$ ，作为动作条件。

VAE——状态编码器/解码器。VAE 将视频帧编码为潜在表示 $\mathbf{z} = \mathcal{E}(\mathbf{x})$ 并将预测潜在解码回视觉观测。我们采用 Wan-VAE Wan 等 (2025) 架构，处理图像和视频模态。

MMDiT——转移函数。MMDiT 采用双流架构：理解流接收 MLLM 编码 $\mathbf{h}$ （通过可训练连接器投影），生成流接收 VAE 的含噪状态潜在。在每个块处，两流通过联合注意力交互。骨干包含 60 个双流块，24 个注意力头（头维度 128）、隐层尺寸 3072 和补丁大小 2×2 。总参数：MLLM 7B、VAE 127M（编码器 54M + 解码器 73M）、MMDiT 20B。上下文长度支持多达 48360 视频 token。

3.2 三维旋转位置编码

我们采用 3D RoPE Su 等 (2024); Heo 等 (2024) 独立编码时间、空间高度和空间宽度维度。我们使用非对称分割而非均匀分配维度：时间轴 16 维，高度和宽度各 56 维，共 128 维（pe_axes_dim = [16, 56, 56]）。时间轴接收较少维度，因为相邻帧强相关；空间轴接收更多维度以捕获目标位置和场景布局的更大

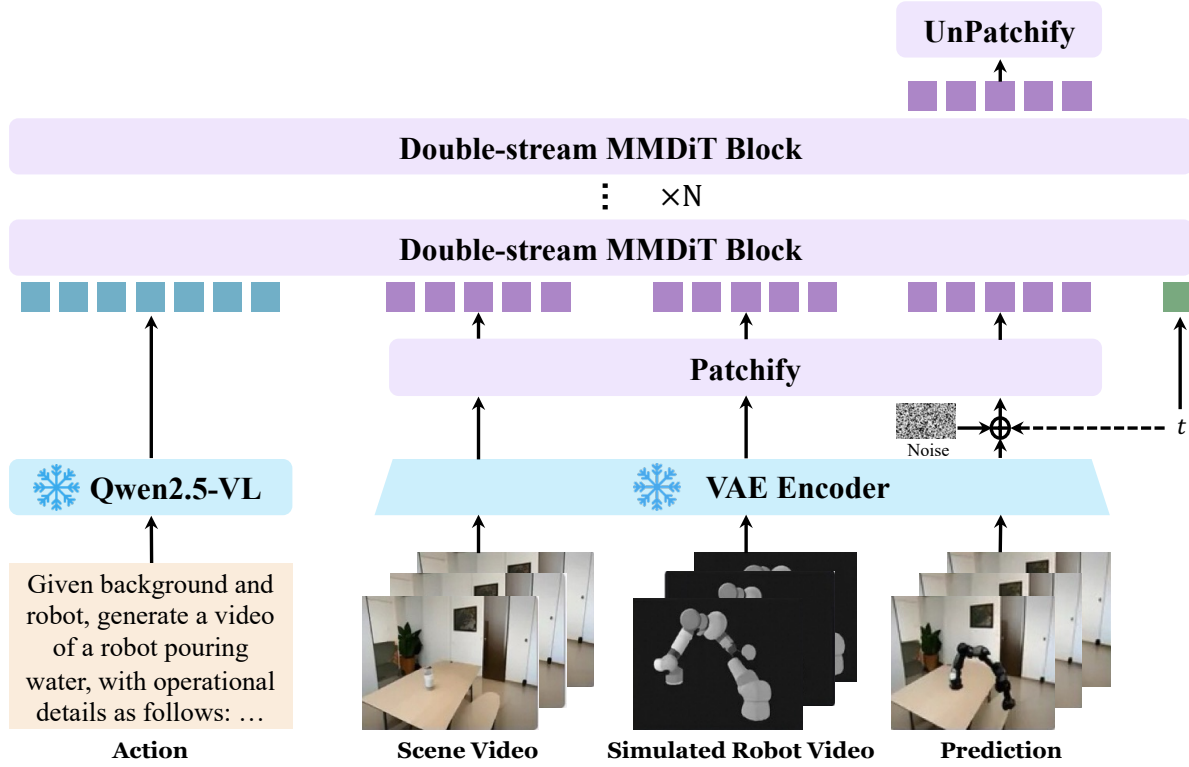


图 4: **Scene2Robot: 跨具身视频合成的多段条件**。输入序列组织为三个连续段——场景条件 (F 帧)、机器人参考 (F 帧) 和生成 (F 帧)。基于索引的机制将条件 token 分配到时步 $t = 0$ 并从损失计算中排除它们，因此只有生成段可训练。每个 MMDiT 块处的联合注意力让生成段同时关注场景外观和机器人运动轨迹，产生语义连贯的跨具身合成。

多样性。我们还应用可扩展 RoPE Wan 等 (2025) 以支持推理时到不同分辨率和时长的泛化。

3.3 Scene2Robot

基于双流 MMDiT 架构 (§3.1) 和非对称 3D RoPE 编码 (§3.2)，我们设计 Scene2Robot——多段条件机制，在相同骨干上实现跨具身视频合成，如图 4 所示。

首帧条件 (TI2V 基线)。标准文本-图像-到-视频任务中，首帧充当固定视觉条件：其 VAE 潜在在生成流设定为时步 $t = 0$ 并从去噪损失中排除，同时冻结 Qwen2.5-VL 处理文本指令进入理解流。双流联合注意力 (§3.1) 在各层融合这两个信号，使生成 token 既关注视觉内容又关注动作语义，产生与语言指令一致的时间连贯视频。

人机转移的多段扩展。人机转移提出视频编辑问题：模型必须参考场景上下文（背景、目标布局、光照）和来自模拟示范的目标机器人运动轨迹。我们通过将首帧条件扩展为三段输入序列来解决，全部在同一 VAE-MMDiT 流程内处理而无任何架构修改：

1. **场景条件** (F 帧)：原始人类示范视频（人手已遮挡），由 VAE 编码提供外观、空间布局和目标状态信息。
2. **机器人参考** (F 帧)：通过 MuJoCo 渲染的模拟机器人执行，由 VAE 编码，提供目标具身的运动轨迹和形态。
3. **生成** (F 帧)：含噪潜在向量经去噪得到最终的逼真机器人执行视频。

段 (1) 和 (2) 与首帧条件共享同样的 $t = 0$ 分配并从损失计算中排除；只有段 (3) 在训练时接收梯度更新。3D RoPE 编码 (§3.2) 为每个段分配其自己的时间索引范围，让模型区分跨段的时间位置。每个 MMDiT 块的联合注意力随后让生成 token 同时关注来自段 (1) 的场景外观、来自段 (2) 的机器人运动和来自理解流的 MLLM 动作语义。这种三部分条件让模型得以合成逼真机器人执行，忠实保留场景上下

文和指导操纵行为。

4 训练

4.1 训练策略

我们提出联合训练范式：通用场景生成和机器人操纵预测统一在单一自然语言接口下，作为同一条件视频生成任务，模型从两个数据体系同步接收梯度更新。这种统一表述使通用和具身先验通过共享骨干相互强化，支撑稳定的跨场景、跨任务联合训练。课程分两阶段：预训练建立广泛视觉基础，SFT 阶段强化具身特化同时维持通用-专家平衡。

4.1.1 预训练阶段：建立通用世界基础

通用世界先验。我们从 14 个高质量视频平台策划超过 2 亿真实观测样本，覆盖自然场景、日常生活和体育。这个广度让模型内化领域无关的世界先验——目标运动、光照变化、碰撞动态——这些形成后续具身泛化的通用骨干。我们进一步加入多相机同步观测和 3D RoPE 空间编码，建立初步跨视图几何一致性作为多视图具身生成的空间基础。

人类交互先验。我们引入大规模第一人称手部操纵数据 (Ego4D [Grauman 等 \(2022\)](#)、EPIC-Kitchen [Damen 等 \(2018\)](#) 等)。人类示范作为通用和具身之间的自然桥梁：通过从日常人类行为学习抓取、工具使用和目标操纵，模型构建动作先验和可供性理解，直接转移到后期机器人操作。

多任务联合训练。T2I、T2V 和 TI2V 任务在共享骨干上联合训练，作为通用和具身能力在一个模型中共存的核心机制。T2I 任务从通用图像数据学习尖锐视觉表示，作为视觉质量锚，其目标形态知识通过共享骨干自动转移到视频生成任务，防止形变和身份不一致。任务比例从纯 T2I 逐步转向完整三任务联合训练，模型在预训练结束时稳定跨多个生成模式运行。

4.1.2 SFT 阶段：具身特化

SFT 阶段逐步深化具身专长，每个训练批中持续混入通用世界数据，使具身特化和通用建模能力协同推进。

渐进具身知识注入。我们采用四阶段数据混合计划。初期：多具身机器人数据和人类手部操纵数据共主导，人类先验指导跨具身操作共性，机器人数据强化执行表示。其次：逐步引入腕部视角和第三人称视角数据，拓宽覆盖范围。再次：多视图连接训练——多相机同步首帧级联为单一输入，模型需同时生成所有视图的后续帧，迫使注意力层建立跨视图对应关系，达成几何一致合成。最后：补充稀缺的复杂任务（倒液、折叠、双臂协调、多材料交互）和长视野推理数据。全程中，通用数据与具身数据共同作用于同一骨干，确保两类能力协同发展。

4.2 训练目标与基础设施

我们采用流匹配目标 [Lipman 等 \(2023\)](#); [Liu 等 \(2022\)](#)，输入视频通过 VAE 编码器编码到潜在空间，噪声从标准正态分布采样。Qwen2.5-VL 将文本输入编码为引导信号。时步从对数正态分布采样，基于视频序列长度的自适应偏移 [Esser 等 \(2024\)](#)。对于 TI2V 任务，首帧时步固定在 0 以确保生成过程基于给定观测帧的条件。训练使用 Megatron-LM [Shoeybi 等 \(2019\)](#) 进行，采用混合并行策略，我们对双流块的子集应用选择性激活重计算 [Korthikanti 等 \(2023\)](#)，以平衡显存占用和训练吞吐量。

5 实验

我们在四个基准上进行全面评估，涵盖具身操作、物理推理和视频质量三个方面。模型在各基准上均领先：具身世界建模基准 EWMBench 达到最先进水平（总体分数 **4.60**，超过 LVP **0.55**），DreamGen Bench 和 WorldModelBench 上均排名开源模型第 1（分别为 **4.952** 和 **8.99**）。

量化评估 (§5.1)。我们对两类基线：(1) 通用视频生成模型——Sora2 OpenAI (2024)、Veo3 Google DeepMind (2025)、Wan2.6 Wan Team (2025)、Kling Kuaishou Technology (2024)、LTX-2 Lightricks (2025)；(2) 具身世界模型——Cosmos Agarwal 等 (2025)、WoW Chi 等 (2025)、LVP Chen 等 (2025a)、Vidar Feng 等 (2025)、GigaWorld Team 等 (2025)。

定性分析 (§5.2)。我们从三个递进维度评估操作能力：细粒度语言基础性、跨具身/任务/视点的泛化，及对强基线的零样本鲁棒性。

交叉域泛化 (§5.3) 另涵盖人机迁移、自动驾驶、室内导航等补充任务。

5.1 Quantitative Evaluation

除另行说明，定量表格中用 **加粗** 表示各列最优值，用 下划线 表示次优值。

5.1.1 EWMBench：具身运动保真度

基准。EWMBench Yue 等 (2025) 从三个维度评估具身世界模型：场景一致性 (SceneC)、运动正确性 (HSD、Dyn、nDTW)、语义对齐 (Diversity、BLEU、CLIP、Logics)。基准包含 21 个样本，分布在 7 个任务上，具有明确的动作顺序约束。

表 2: EWMBench 性能对比。

类型	方法	场景	运动			语义				总体
		SceneC	HSD	Dyn	nDTW	Diversity	BLEU	CLIP	Logics	
通用	Veo3	0.8415	0.2130	0.1932	0.1613	0.0221	0.2139	0.8965	0.9474	3.49
	Wan2.6	0.6712	0.2034	0.0900	0.1715	0.0502	0.1616	0.8743	1.0000	3.22
	Kling26	0.8211	0.3272	0.1822	0.3423	0.0173	0.2591	<u>0.9014</u>	1.0000	3.85
	LTX-2	0.7850	0.2076	0.1283	0.2443	0.0120	0.1425	0.8869	0.5000	3.01
	Sora2	0.8526	0.2807	0.3494	0.2754	0.0314	<u>0.2466</u>	0.9100	0.9474	3.89
具身	Cosmos	0.7963	0.2500	0.2052	0.2533	0.0803	0.1230	0.8458	0.7333	3.29
	GigaWorld	0.8707	0.3050	0.0849	0.2783	0.0278	0.2048	0.8873	0.9000	3.56
	LVP	0.8795	<u>0.4248</u>	0.0433	<u>0.6226</u>	0.0093	0.2179	0.8995	<u>0.9524</u>	<u>4.05</u>
	Vidar	0.7341	0.1877	0.1520	0.1769	<u>0.0653</u>	0.1607	0.8821	0.9411	3.30
	Wow	<u>0.8866</u>	0.2494	0.0529	0.2566	0.0266	0.1932	0.9001	<u>0.9524</u>	3.52
我们		0.9142	0.5660	<u>0.3429</u>	0.6708	0.0114	0.2079	0.8834	1.0000	4.60

结果。表 2 显示我们的模型总体排名 **第 1**，得分 **4.60**，超过亚军 LVP (4.05) **0.55** 分。运动保真度领先——HSD 指标 (**0.566**) 相比 LVP (0.425) 提升 33%——同时在场景一致性 (SceneC: **0.914**) 和逻辑约束满足 (Logics: **1.00**) 上都达到最优。

5.1.2 DreamGen Bench

基准。DreamGen Bench Zhou 等 (2025) 评估视频世界模型生成的机器人视频质量，在 GR1 机器人具身体的三个子集上测量指令跟随 (IF) 和物理对齐 (PA)：环境泛化 (GR1-Env)、对象泛化 (GR1-Object)、行为泛化 (GR1-Behavior)。IF 评估采用 Qwen2.5-VL Bai 等 (2025) 作为评估器。

表 3: DreamGen Bench 性能对比。

方法	GR1-Env		GR1-Object		GR1-Behavior		总分
	PA	IF	PA	IF	PA	IF	
Cosmos-sft	0.709	0.655	<u>0.775</u>	0.720	0.649	0.621	4.129
LVP	<u>0.810</u>	0.772	0.745	0.829	0.713	0.889	<u>4.758</u>
Vidar	0.445	0.647	0.478	0.726	0.394	0.651	3.341
GigaWorld	0.621	0.933	0.500	<u>0.852</u>	0.426	<u>0.884</u>	4.216
Wow	0.793	<u>0.826</u>	0.755	<u>0.849</u>	0.809	0.696	4.728
我们	0.828	0.793	0.840	0.878	<u>0.781</u>	0.832	4.952

结果。表 3 显示我们的模型达到最高总分 **4.952**，总体排名 **第 1**。GR1-Object IF (**0.878**) 排名第

一，展现了强大的对象级组合泛化能力。物理对齐在所有子集上都表现稳定（PA：0.828/0.840/0.781）。GR1-Behavior IF（0.832）略低于 LVP（**0.889**）和 GigaWorld（0.884），表明长时程行为泛化是后续改进方向。

5.1.3 PBench：物理行为评估

基准。 PBench NVIDIA (2025a) 从两个方面评估模型：(1) 领域分数，通过 Qwen2.5-VL 评估的 QA 对在六个领域（自动驾驶、机器人、工业、物理、人体、常识）上测量物理行为理解；(2) 质量分数，通过八个 VBench Huang 等 (2024) 指标测量视觉质量，包括图像到视频的一致性、美学质量、运动平滑度和主体一致性。总体分数是两者的平均值。

表 4: PBench 性能对比。

类型	方法	质量指标 (VBench)								质量	领域	总体
		I2V-Bg	I2V-S	Aes	Img	Bg-Con	Mot	Sub-Con	O-Con			
通用	Veo3	0.975	0.980	0.526	0.698	0.938	0.994	0.927	0.128	0.771	0.882	0.827
	Wan2.6	0.856	0.843	0.514	0.719	0.906	0.978	0.843	0.136	0.724	0.832	0.778
	Sora2	0.981	0.973	0.487	0.672	0.961	0.994	0.954	0.129	0.769	0.841	0.805
	Kling26	0.982	0.979	0.521	0.699	0.920	0.990	0.927	0.124	0.768	0.874	0.821
	LTX-2	0.948	0.955	0.506	0.622	0.932	0.986	0.904	0.118	0.746	0.845	0.796
具身	Cosmos	0.974	0.973	0.470	0.663	0.940	0.989	0.931	0.160	0.763	0.840	0.802
	LVP	0.979	0.981	0.515	0.679	0.954	0.991	0.962	0.116	0.772	0.812	0.792
	GigaWorld	0.957	0.944	0.495	0.641	0.925	0.984	0.892	0.128	0.746	0.841	0.794
	Vidar	0.935	0.922	0.501	0.573	0.912	0.982	0.863	0.120	0.726	0.810	0.768
	Wow	0.967	0.957	0.517	0.689	0.941	0.980	0.929	0.111	0.761	0.786	0.774
我们		0.956	0.943	0.455	0.649	0.956	0.990	0.933	0.124	0.751	0.857	0.804

结果。表 4 显示，我们的模型在开源模型中表现最优，总体分数 **0.804**。领域理解能力是最强维度（0.857，总体 **第 3**），超过大多数闭源模型。运动平滑度也表现突出（0.990，开源模型中 **第 2**），体现了生成过程中的时间连贯性。美学质量（0.455）和成像质量（0.649）相对较低，这是因为我们的模型针对具身任务优化，输出分辨率低于通用视频生成器，从而降低了 VBench 的像素级质量评分。但这个分辨率对下游机器人控制任务已足够。

5.1.4 WorldModelBench：物理推理与指令跟随

基准。 WorldModelBench Li 等 (2025) 从三个维度评估模型：指令跟随（0-3 量表）、常识推理（帧质量和时间质量）、物理合理性（5 种违反类型：牛顿定律、质量守恒、流体动力学、穿透、重力）。基准包含 350 个实例，分布在 7 个领域、56 个子领域。

表 5: WorldModelBench 性能对比。

类型	方法	指令	常识推理			物理合理性					物理	总分
		(0-3)	帧	时间	总体	牛顿	质量	流体	穿透	重力	总体	
通用	Veo3	2.52	0.98	0.95	1.93	1.00	0.89	0.99	0.91	1.00	4.80	9.25
	Wan2.6	2.50	0.99	0.95	1.94	1.00	0.89	0.99	0.94	1.00	4.83	9.27
	Sora2	2.21	0.96	0.93	1.88	1.00	0.91	0.99	0.95	1.00	4.84	8.93
	Kling26	1.59	0.97	1.00	1.97	1.00	1.00	1.00	1.00	1.00	5.00	8.55
	LTX-2	1.97	0.69	0.62	1.32	0.99	0.60	1.00	0.73	1.00	4.32	7.61
具身	Cosmos	2.14	1.00	0.94	1.94	1.00	0.92	1.00	0.94	1.00	4.86	8.94
	LVP	2.01	0.89	0.91	1.80	1.00	0.93	0.99	0.95	1.00	4.87	8.67
	GigaWorld	2.13	0.59	0.46	1.05	1.00	0.48	0.99	0.69	0.98	4.13	7.31
	Vidar	1.62	0.54	0.45	0.99	1.00	0.56	1.00	0.85	1.00	4.40	7.01
	Wow	2.05	0.76	0.65	1.41	1.00	0.65	0.99	0.81	1.00	4.45	7.91
我们		2.33	0.87	0.85	1.72	1.00	1.00	1.00	0.94	1.00	4.94	8.99

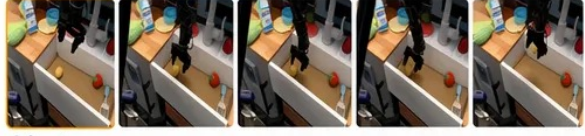
结果。表 5 显示我们的模型在开源模型中表现最优（**8.99**，总体第 3），仅次于闭源的 Wan2.6 和 Veo3。模型在四类物理合理性上都达到满分（1.00），指令跟随能力强（2.33/3.0）。常识推理方面的差距

(a) Contrastive Instruction Following

Pick up the **red strawberry** from the table.



Pick up the **yellow potato** from the table.

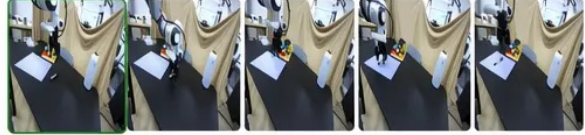


Target object

Pick up the pen from the table and place it on the **wooden tray**.



Pick up the pen from the table and place it on the **white paper**.



Destination

Take the glue stick and **hand it to the person**.



Take the glue stick and **put it in the penholder**.



Action type

(b) Complex Instruction Following

Grasp each Rubik's cube in turn and stack them into a stable two-layer structure on the tray.



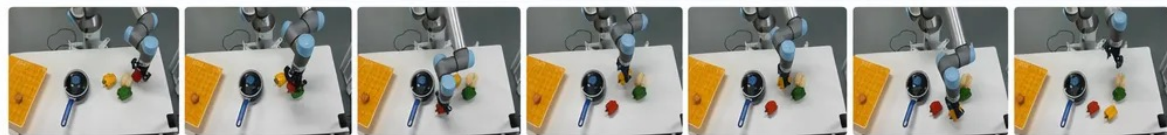
① Reach cube 1

② Place on tray

③ Reach cube 2

④ Stack on top

Pick up the red and yellow bell peppers in sequence and place them on the table from left to right from the camera's perspective.



① Pick red pepper

② Place left slot

③ Pick yellow pepper

④ Place right slot

图 5: 细粒度语言基础性。(a) 对比对: 每对列共享相同初始帧 (彩色边框); 两条指令间仅高亮关键词不同。对 1: 目标对象。对 2: 目标位置。对 3: 动作类型。在所有情况下生成的动作都精确映射到区分关键词。(b) 复杂指令: 需要多步执行或抽象目标推理的两个例子。彩色标签标记每个生成序列内的关键动作里程碑。

源于输出分辨率较低。

5.2 定性分析

5.2.1 细粒度语言基础性

精确的语言到视觉动作映射是 QWEN-ROBOTWORLD 作为语言条件世界模型的核心设计。图 5 从两个层面评估了这种能力。(a) 对比对: 初始帧相同, 仅改变一个关键词 (目标对象、动作类型或空间位置) 时, 模型生成完全不同的视频, 展现了细粒度的语义区分——远超通用操作先验。(b) 复杂指令: 模型能处理具有多步依赖关系的长步长顺序任务, 以及需要从上下文推断操作序列的抽象目标指令, 在无显式子任务提示的情况下分解为时间连贯的执行。

(A) Cross-Embodiment

same instruction · four morphologies

Pick up the object and place it in the container. Same instruction across four morphologies.



(B) Cross-Task × Cross-Environment

task-appropriate contact dynamics

Kitchen
Move bowl to sink



Living room
Fold & organize



Tabletop
H-R handover



(C) Multi-View Consistency

same episode · three cameras

Main



Wrist-L



Wrist-R

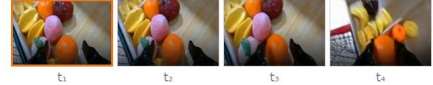


图 6: 跨具身、任务与视点的泛化。(A) 跨具身: 单条指令驱动四种形态 (单臂、双臂、人形、灵巧手); 每个单元格三个关键帧。(B) 跨任务 × 跨环境: 初始帧 (橙色边框) 后跟四个任务中生成的四个帧。(C) 多视角: 主摄像和腕摄像从与 (B, 第 1 行) 相同的片段联合生成。

5.2.2 跨具身、任务与视点的泛化

图 6 展示了操作能力的三个互补维度。(A) 跨具身: 单条指令驱动四种机器人形态 (单臂夹爪、双臂系统、人形、灵巧手), 无需机器人特定的调整, 验证了自然语言作为通用动作接口的有效性。每个单元格显示初始、中间、最终三个关键帧。(B) 跨任务 × 跨环境: 生成涵盖水果取放、碗盛取、布料折叠、人机协作——每个任务展现任务适配的接触动力学, 反映了对多样现实环境的物理知识理解; 每行显示一个初始帧后跟四个均匀分布的生成帧。(C) 多视角一致性: 三个同步摄像机视角 (主摄像、左腕摄像、右腕摄像) 从与 (B, 第 1 行) 相同的超市取放任务联合生成, 对象身份和运动轨迹在所有视角上保持几何一致性。

5.2.3 RoboTwin-IF 上的零样本鲁棒性

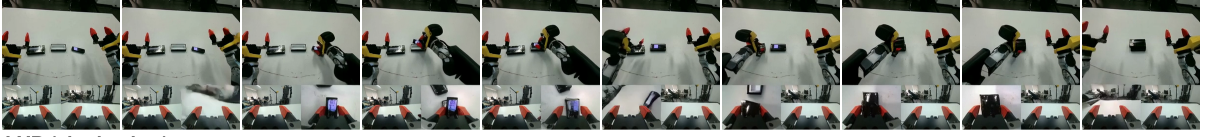
在展示单模型能力的基础上, 我们进一步验证这些改进在严格的模型间对比下是否保持。具身世界模型的综合评分可能混合三种失败源: 指令不匹配、多视角不一致、视觉质量下降。为了分离这些因素, 我们在四个 Unitree G1 任务上做零样本并排对比, 对手是两个强基线 LVP 和 **Cosmos2.5-14B**。图 7 显示 QWEN-ROBOTWORLD 执行更贴合语言指令 (正确的对象/动作对应和更干净的目标完成) 同时维持连贯的多视角轨迹。两个基线展现了不同的失败模式: **LVP** 更容易产生不完整的任务执行, **Cosmos2.5-14B** 在更复杂的情况下往往表现出指令与生成操作结果间较弱的对齐。

我们进一步在 **RoboTwin-IF** (指令跟随) 基准上评估零样本性能。这个基准基于 RoboTwin 模拟器构建, 包含许多新创建的复杂任务。尽管训练中仅使用了少量开源 RoboTwin 数据, 模型在 RoboTwin-IF 上仍展现强大的零样本性能, 并保持跨摄像视角的一致性。这说明改进不是针对少量定性样本的过拟合, 而是真正推广到了更具挑战性的未见具身任务。总体而言, QWEN-ROBOTWORLD 通过在统一生成框架中更好地对齐指令跟随、动作真实性和多视角连贯性, 展现了相比以往基线更强的零样本鲁棒性。

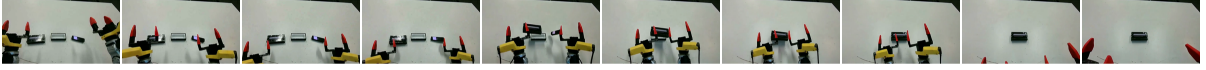
Task 01

Assemble a small electronic device by inserting a black rectangular component into a matching black housing.

Ours (multi-view)



LVP (single-view)



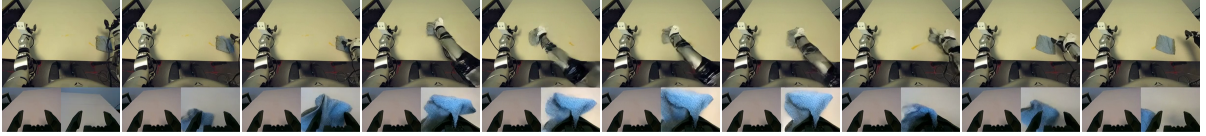
Cosmos (single-view)



Task 02

Wipe yellowish liquid stains from a light table using a gray rag until clean and dry.

Ours (multi-view)



LVP (single-view)



Cosmos (single-view)



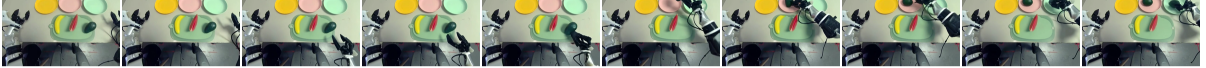
Task 03

Place banana, watermelon slice, and avocado onto color-matched plates from the cutting board.

Ours (single-view)



LVP (single-view)



Cosmos (single-view)



Task 04

Clean the whiteboard using a blue eraser and remove the red wavy line/markings.

Ours (single-view)



LVP (single-view)



Cosmos (single-view)

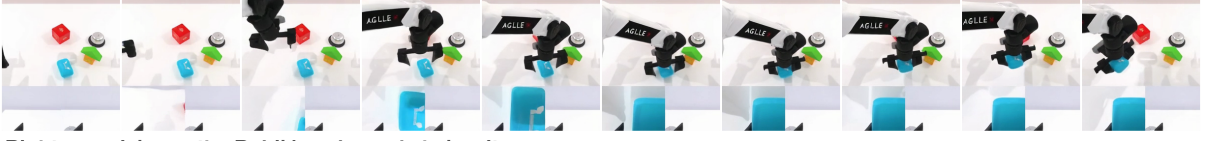


图 7: 语言-动作对齐与多视角连贯性的零样本定性对比。在相同条件下（相同初始帧、提示、摄像机布局）进行并排对比，QWEN-ROBOTWORLD 对阵 LVP 和 Cosmos2.5-14B。

图 8 提供了这个基准结果的代表性 RoboTwin-IF 零样本案例作为定性证据。每个任务用十个均匀采样的帧可视化，以第一帧和最后一帧为锚点，使中间进展和最终完成直观可见。在这些新构造的复杂任务中，QWEN-ROBOTWORLD 保持了连贯的执行和多视角一致性，与量化 RoboTwin-IF 结果一致。

Basic Manipulation

Left arm picks up the blue tissue box from the table.

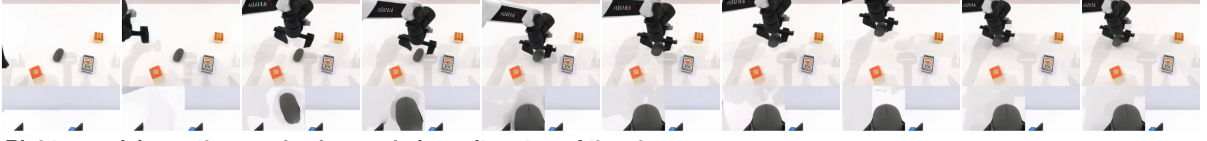


Right arm picks up the Rubik's cube and shakes it once.



Relative Positioning

Left arm picks up the mouse and places it at the upper-left area of the table.



Right arm picks up the wooden box and places it on top of the phone.



Bimanual Collaboration

Left arm picks up the wooden box, handovers to right arm, and right arm places it on the soap.

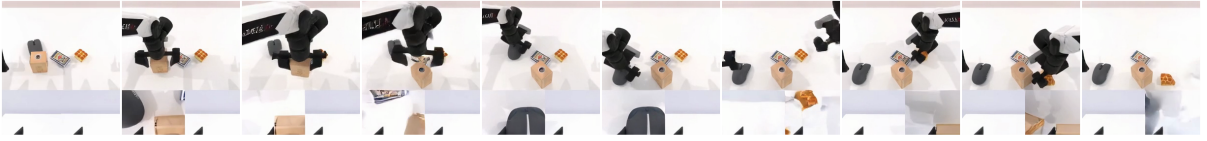


Left arm picks up the Rubik's cube while right arm picks up the bread.



Long-Horizon / Complex Instructions

Place wooden box in front, mouse to left, then bread to right in a three-step sequence.



Place yellow box at center, bell to right, then mouse to left in sequence.



图 8: RoboTwin-IF 零样本定性案例。基准基于 RoboTwin 模拟器构建，包含许多新构造的复杂任务。

5.3 交叉域泛化

除了以操作为中心的评估外，我们还评估了模型在核心操作领域外补充任务族上的泛化能力。图 9 展示了在八个目标具身体上的人机迁移，模型从人类示范中保持任务意图，同时将运动适配到具身特定的运动学约束。图 10 涵盖了移动场景，包括来自 Bench2Drive、NVIDIA PhysicalAI-AD、Sekai 和 Waymo 的自动驾驶片段，以及来自 VLNVerse 的自主导航第一人称视角片段。这些结果表明，学到的语言条件动态模型能泛化到单个具身体和场景族之外。



图 9: **人机迁移**。在八个目标具身体上，每行对比同一任务的人类示范（左）与合成机器人执行（右），每个视频采样五个均匀分布的帧。生成的轨迹保持任务意图，同时适配具身特定的运动学约束。

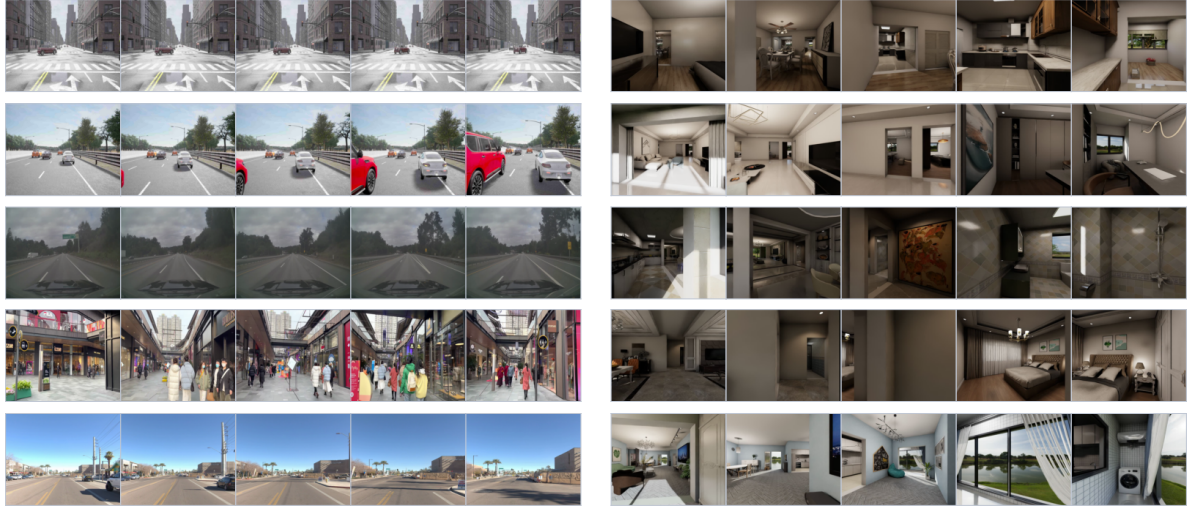


图 10: **移动能力生成**。配对列，五行：(左) 自动驾驶来自 Bench2Drive、NVIDIA PhysicalAI-AD、Sekai 和 Waymo 的片段；(右) 自主导航来自 VLNVerse 的第一人称视角语言引导遍历。每个片段采样五个均匀分布的帧。

6 结论

本报告介绍 QWEN-ROBOTWORLD，一个具身智能的语言条件世界模型框架，在共享的自然语言动作接口下统一机器人操作、自动驾驶、室内导航和人-机转移。为此开发三部分系统：(a) 双流 MMDiT 架构与 MLLM 动作编码，实现语义精确且物理合理的生成；(b) 具身世界知识（EWK）数据集，提供大规模跨具身动作-语言对齐；(c) 通用 + 专家渐进式课程，耦合宽泛的视觉先验与具身专门化。

这个设计支持统一的骨干网络，可向三大代表性具身世界模型应用适配——合成数据生成、策略评估和动作规划。基准评测与零样本分析均证实 QWEN-ROBOTWORLD 性能强劲稳定，稳健的多视图指令跟随泛化能力。本工作为具身世界模型的构建奠定实践基础——不仅感知能力强，也能为下游机器人学习和

控制发挥实用价值。

Authors

Jie Zhang^{*}, Xiaoyue Chen^{*}, Anzhe Chen, Dayiheng Liu, Deqing Li, Gengze Zhou, Hale Yin, Haoqi Yuan, Haoyang Li, Jiahao Li, Jiazhaoh Zhang, Jingren Zhou, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Pei Lin, Qihang Peng, Shengming Yin, Tianhe Wu, Tianyi Yan, Xiao Xu, Yan Shu, Yanran Zhang, Ye Wang, Yi Wang, Yilei Chen, Yixian Xu, Yiyang Huang, Yuxiang Chen, Zekai Zhang, Zhendong Wang, Zixing Lei, Zhixuan Liang, Zihao Liu, Zikai Zhou, Chenxu Lv[†], Xiong-Hui Chen[†], Chenfei Wu[†]

^{*}Equal contribution.

[†]Corresponding author.

参考文献

- Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.
- AgiBot-World-Contributors. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15791–15801, 2025.
- Johan Bjorck, Fernando Castaneda, Linxi Fan, Dieter Fox, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.
- Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, et al. RT-1: Robotics transformer for real-world control at scale. In *Robotics: Science and Systems (RSS)*, 2023.
- Build AI. Egocentric-10k. Hugging Face Datasets, 2025. URL <https://huggingface.co/datasets/buildai/Egocentric-10K>.
- Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, Caiyi Zhang, Peihao Li, William T. Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, Vincent Sitzmann, and Yilun Du. Large video planner enables generalizable robot control, 2025a. URL <https://arxiv.org/abs/2512.15840>.
- Tianxing Chen, Zhanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu 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*, 2025b.
- Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, Zezhong Qian, Anthony Chen, Qiang Zhou, Yueru Jia, Jiaming Liu, Yong Dai, Qingpo Wuwu, Chengyu Bai, Yu-Kai Wang, Ying Li, Lizhang Chen, Yong Bao, Zhiyuan Jiang, Jiacheng Zhu, Kai Tang, Ruichuan An, Yulin Luo, Qiuxuan Feng, Siyuan Zhou, Chi min Chan, Chengkai Hou, Wei Xue, Sirui Han, Yike Guo, Shanghang Zhang, and Jian Tang. Wow: Towards a world omniscient world model through embodied interaction, 2025. URL <https://arxiv.org/abs/2509.22642>.
- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *European Conference on Computer Vision (ECCV)*, 2018.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

-
- Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.
- Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation, 2025. URL <https://arxiv.org/abs/2507.12898>.
- Fourier ActionNet Team and Yao Mu. Actionnet: A dataset for dexterous bimanual manipulation. 2025.
- Galaxea AI. Galaxea open-world dataset and G0 dual-system VLA model. *arXiv preprint arXiv:2509.00576*, 2025.
- Zelin Gao, Qiuyu Wang, Yanhong Zeng, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.
- Google DeepMind. Veo 3. <https://deepmind.google/technologies/veo/veo-3/>, 2025. URL <https://deepmind.google/technologies/veo/veo-3/>.
- Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- Mingfei Han, Liang Ma, Kamila Zhumakhanova, Ekaterina Radionova, Jingyi Zhang, Xiaojun Chang, Xiaodan Liang, and Ivan Laptev. RoomTour3D: Geometry-aware video-instruction tuning for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2024.
- Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. RLBench: The robot learning benchmark. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.
- Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Li. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *NeurIPS Datasets and Benchmarks Track*, 2024.
- Alexander Khazatsky, Karl Pertsch, Suraj Nair, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems (RSS)*, 2024.
- Vijay Anand Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. Reducing activation recomputation in large transformer models. *Proceedings of Machine Learning and Systems*, 5:341–353, 2023.
- Kuaishou Technology. Kling: A progressive framework for video generation. <https://klingai.com>, 2024. URL <https://klingai.com>.

-
- Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models, 2025. URL <https://arxiv.org/abs/2502.20694>.
- Xuanlin Li, Kyle Hsu, Jiayuan Liu, Ken Goldberg, and Sergey Levine. Evaluating real-world robot manipulation policies in simulation. In *Conference on Robot Learning (CoRL)*, 2024.
- Lightricks. LTX-Video: Realtime video latent diffusion. <https://github.com/Lightricks/LTX-Video>, 2025. URL <https://github.com/Lightricks/LTX-Video>.
- Sihao Lin, Zerui Li, Xunyi Zhao, Gengze Zhou, Liuyi Wang, Rong Wei, Rui Tang, Juncheng Li, Hanqing Wang, Jiangmiao Pang, Anton van den Hengel, Jiajun Liu, and Qi Wu. VLNVerse: A benchmark for vision-language navigation with versatile, embodied, realistic simulation and evaluation. *arXiv preprint arXiv:2512.19021*, 2025.
- Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.
- Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.
- NVIDIA. NVIDIA Isaac Sim. <https://developer.nvidia.com/isaac-sim>, 2022.
- NVIDIA. PBench: A physical ai benchmark for world models, 2025a. URL <https://huggingface.co/datasets/nvidia/PBench>.
- NVIDIA. Nvidia physicalai autonomous driving dataset. 2025b. <https://developer.nvidia.com/physicalai>.
- OpenAI. Sora: Creating video from text. <https://openai.com/sora>, 2024. URL <https://openai.com/sora>.
- OpenLoong Baihu Team. OpenLoongData-v1.0. <https://www.openloong.org.cn/en/datasets/baihu>, 2025.
- Baoqi Pei, Yifei Huang, Jilan Xu, Guo Chen, Yuping He, Lijin Yang, Yali Wang, Weidi Xie, Yu Qiao, Fei Wu, and Limin Wang. Modeling fine-grained hand-object dynamics for egocentric video representation learning. In *International Conference on Learning Representations (ICLR)*, 2025.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.
- Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 36(6), 2017.

-
- Sekai Team. Sekai: Real-world egocentric walking videos for world model training. 2025. <https://huggingface.co/datasets/sekai>.
- Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. RoboScape: Physics-informed embodied world model. *arXiv preprint arXiv:2506.23135*, 2025.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- GigaWorld Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jiagang Zhu, Kerui Li, Mengyuan Xu, Qiuping Deng, Siting Wang, Wenkang Qin, Xinze Chen, Xiaofeng Wang, Yankai Wang, Yu Cao, Yifan Chang, Yuan Xu, Yun Ye, Yang Wang, Yukun Zhou, Zhengyuan Zhang, Zhehao Dong, and Zheng Zhu. Gigaworld-0: World models as data engine to empower embodied ai, 2025. URL <https://arxiv.org/abs/2511.19861>.
- Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, Yaping Li, Ping Wang, Junhao Cai, Jia Zeng, Hao Dong, and Jiangmiao Pang. InternData-A1: Pioneering high-fidelity synthetic data for pre-training generalist policy. *arXiv preprint arXiv:2511.16651*, 2025.
- Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5026–5033, 2012.
- Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.
- Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
- Wan Team. Wan: Open and advanced large-scale video generative models. <https://wanxai.com>, 2025. URL <https://wanxai.com>.
- Waymo Team. Waymo open dataset: End-to-end driving. 2024. <https://waymo.com/open/>.
- Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, et al. RoboMIND: Benchmark on multi-embodiment intelligence normative data for robot manipulation. In *Robotics: Science and Systems (RSS)*, 2025a.
- Shihan Wu et al. RoboCOIN: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025b.
- Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies, 2026. URL <https://arxiv.org/abs/2602.15922>.

-
- Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. EWMBench: Evaluating scene, motion, and semantic quality in embodied world models, 2025. URL <https://arxiv.org/abs/2505.09694>.
- Zhenyu Zhao, Hongyi Jing, Xiawei Liu, Jiageng Mao, Abha Jha, Hanwen Yang, Rong Xue, Sergey Zakharov, Vitor Guizilini, and Yue Wang. Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation. *arXiv preprint arXiv:2510.08807*, 2025.
- Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. TesserAct: Learning 4D embodied world models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.
- Joel Zhou, Jordan Juravsky, Sanja Fidler, Umar Bhatt, and Nima Fazeli. DreamGen: Unlocking generalization in robot learning through neural trajectories. *arXiv preprint arXiv:2505.12705*, 2025.