《多模态大模型原理与应用》

Lecture 12
具身世界模型

刘阳
中山大学
人机物智能融合实验室（HCP Lab）
liuy856@mail.sysu.edu.cn

为什么具身智能需要"先想后做"

核心问题

具身智能面对真实世界的动态交互，单靠**反应式控制**难以处理：

- 长时任务规划
- 风险决策与后果评估
- 复杂物理交互预测

解决思路

需要能够**预测未来状态**的模型来辅助规划与控制：

- 提前模拟动作后果
- 选择最优行动方案
- 避免危险操作

对比示意

直接抓取（失败）

观察 → 立即动作

不考虑物体滑动、碰撞

先预测再抓取（成功）

观察 → 预测未来 → 规划动作

预判物体运动，调整抓取策略

从传统机器人到具身基础模型

规则控制	模仿学习	强化学习	VLA	WAM
基于人工编写的规则	从专家示范中学习	试错学习最优策略	视觉-语言-动作	世界模型+动作
硬编码行为逻辑	行为克隆	奖励驱动	多模态策略	预测式智能体

研究趋势：将世界模型嵌入策略，使机器人从"反应式"走向"预测式" —— 这是具身智能的关键范式转变。

世界模型的两种核心理解

理解世界

Understanding World

核心目标：构建内部表征去理解世界

学习紧凑、抽象的隐变量表示，抓住环境中对决策有用的结构

关注"世界是什么样的"

预测未来

Predicting Future

核心目标：预测未来状态去支持模拟、规划和决策

从当前观察出发预测未来状态，生成后续帧或未来轨迹

关注"世界会怎样变化"

人类直觉类比：快思考与慢思考

系统一：快思考

本能反应，快速自动

不显式依赖世界模型

如：手碰到热物立即缩回

系统二：慢思考

审慎推理，模拟未来

显式依赖世界模型

如：规划如何稳定地搬运杯子

对应到机器人系统

Reactive：感知 → 直接动作

反射式控制

Predictive：感知 → 预测 → 动作

世界模型驱动的规划

最低限度直觉：什么是具身世界模型

世界模型

= **环境变化的预测器**

预测"如果我做了X，世界会变成什么样"

具身世界模型

= **预测 + 动作生成**

既预测环境如何变，又根据未来来产生动作

核心链路

关键：动作不是由当前观察**直接**决定，而是由对未来状态的**预测**来指导

具身场景为什么更难

具身任务的世界建模必须兼顾**空间、时间和动作因果关系**，挑战包括：

接触与力

物理接触产生力和反作用力，需要建模力觉反馈

遮挡与形变

物体可能被遮挡，柔性物体会形变，增加预测难度

视角变化

机器人运动时视角不断变化，需要不变的表征

本体差异

不同机器人形态差异大，模型需要跨本体泛化

时序依赖

动作效果随时间展开，需要建模长时动态

因果耦合

动作与观测强耦合，需要建模动作-观测因果链

典型任务：抓取、导航、双臂与人形

机械臂操作

抓取、放置、推物、装配等桌面操作任务

核心：理解"我做了什么，物体位置/姿态怎样变"

移动导航

室内/室外路径规划、避障、目标导向移动

核心：预判地形变化和障碍物运动

双臂协调

双手配合完成复杂操作，如折叠衣物、拧瓶盖

核心：协调两臂动作预测各自及联合效果

人形控制

全身运动控制、平衡维持、动态行走

核心：预测身体姿态变化与接触力

为什么只做 Observation-to-Action 不够

VLA

观察 (Observation) → 动作 (Action)

- ✗ 不显式建模动作后的物理世界演化
- ✗ 需要前瞻推理时受限
- ✗ 难以处理长时序后果评估

WAM

观察 → 未来预测 → 动作

- ✓ 显式建模世界动态变化
- ✓ 支持前瞻物理推理
- ✓ 可评估动作后果

关键区别：VLA 学习 "看完就动" 的映射，WAM 学习 "看完再想，想完再动" 的闭环

核心判断标准：模型是否真的在"想未来"

分析各种方法时，始终问一个问题：**这个模型是在直接输出动作，还是在借助对未来世界的预测来决定动作？**

Reactive 反应式

✗ 直接输出动作

✗ 不显式预测未来

✗ 仅依赖当前观察

代表：传统VLA、端到端策略

Predictive 预测式

✓ 显式或隐式预测未来

✓ 基于预测生成动作

✓ 动作与未来状态耦合

代表：WAM、World Model + Policy

世界模型研究为什么突然变热

近年多模态大模型和视频生成模型的突破，让"用生成模型做世界模拟器"成为更现实的方向：

多模态大模型

GPT-4V等模型提升视觉-语言理解能力，为跨模态世界建模奠定基础

视频生成模型

Sora等模型展现强时空建模能力，接近"未来世界演化"表征

视频理解模型

V-JEPA等模型证明抽象预测可服务世界建模和控制

关键趋势：视频生成模型天然接近"未来世界演化"的表征，使"用生成模型做世界模拟器"从理论走向实践。机器人操作、自动驾驶等领域开始大规模应用世界模型。

视频生成为什么推动了世界模型

视频生成模型擅长建模**时空连续变化**，天然接近"未来世界演化"的表征：

- **时序建模能力**：学习帧与帧之间的连续变化规律
- **空间一致性**：保持物体位置、外观的跨帧一致
- **物理直觉**：隐式学习到部分物理规律（重力、碰撞）
- **条件生成**：支持动作/文本条件控制视频生成

因此，很多新世界模型直接借力视频基础模型（如Sora、VDM、LVDM）作为骨干架构，在其上增加动作条件化和物理约束。

但视频生成不等于真正的世界模型

视频生成能做到

✓ 强视觉一致性

✓ 一定的物理直觉

✓ 流畅的时序过渡

视频生成做不到

✗ 精确的因果交互建模

✗ 严格的物理规律遵循

✗ 动作条件的可靠响应

核心区别

视频生成模型优化的是**视觉逼真度**，而世界模型需要的是**物理正确性和动作可控性**。Sora可以生成逼真的杯子掉落视频，但无法保证杯子遵循正确的重力加速度或响应特定的推力条件。

世界模型与模拟器的关系

物理模拟器

特点：

- 显式、外部、可控
- 基于牛顿物理方程
- 精确但计算昂贵

代表：MuJoCo, Isaac Sim, Gazebo

世界模型（学习式）

特点：

- 数据驱动、内部、可学习
- 从经验中学习动力学
- 高效但可能不精确

代表：RSSM, Transformer WM

互补关系，非替代

本节内容

CONTENTS

一、世界模型基础概念

二、世界动作模型的定义、分类与核心架构

三、WAM的数据来源、训练策略与应用场景

四、WAM的评测体系、核心挑战与未来方向

什么是世界模型?

语言模型

学习的是「文本」的统计结构

机器学会了「谈论」世界，但世界并非由文字构成。

世界模型

学习的是「空间与时间」的统计结构

光线如何落在表面、物体从未见过的角度是什么样、受力后如何运动——如何遵循物理规律。

「世界并不是由文字构成的。」 *— 路德维希·维特根斯坦《逻辑哲学论》*

一个框架：智能体—世界交互闭环

源自强化学习的「部分可观测马尔可夫决策过程」(POMDP)——一切「世界模型」都是这一闭环的不同投影

↻ **闭环：观测反馈回智能体，循环往复**

三类世界模型：闭环的不同投影

今天各种被称作「世界模型」的系统，各自输出闭环中的不同部分

渲染器
Renderer

输出：观测（像素 / 视频）

核心指标：视觉保真度

代表

文生视频、Google Genie 3、World Labs RTFM

模拟器
Simulator

输出：状态（几何/物理/动态）

核心指标：结构性准确

代表

训练 RL 智能体、机器人控制器的训练环境

规划器
Planner

输出：行动

核心指标：能指导下一步

代表

视觉-语言-行动模型、World Action 模型

世界模型的基本定义

世界模型 = 能内化环境动力学和动作影响的预测模型

学习环境如何随动作而变化，从而支持模拟、规划和决策

核心状态转移

三大功能

模拟 Simulation

在想象中推演未来

规划 Planning

基于预测选择最优动作

决策 Decision

评估不同动作的后果

在MBRL里，世界模型是什么

模型式强化学习（Model-Based RL）中，世界模型是**"先想象再决策"**的核心机制：

Dynamics Model
动力学模型
s', r = f(s, a)

+

Reward Model
奖励模型
r = R(s, a)

+

Planner
规划器
选择最优动作序列

MBRL 核心闭环

第一类：理解世界的内部表征

学习**紧凑、抽象的隐变量表示**，抓住环境中对决策有用的结构，而非逐像素记忆世界。

核心优势

信息压缩

从高维观测中提取低维关键信息

噪声过滤

忽略与决策无关的视觉细节

结构学习

发现环境中的有用结构

第二类：预测未来的生成模型

从当前观察和条件出发**预测未来状态**，在视频世界模型中体现为生成后续帧或未来轨迹。

生成链路

输出形式

像素级生成

直接生成未来图像/视频帧

特征级生成

预测未来隐空间表示

轨迹级生成

预测未来状态序列

动作条件 vs 语言条件世界模型

动作条件

$$(o_t, a_t) \rightarrow o_{t+1}$$

输入：当前观察 + 具体动作

输出：执行该动作后的未来观察

特点：

- 动作是低层控制信号
- 精确预测特定动作后果

语言条件

$$(o_t, l) \rightarrow future$$

输入：当前观察 + 语言描述

输出：符合语言描述的未来

特点：

- 语言是高层语义约束
- 强调高层目标导向

显式世界模型 vs 隐式世界模型

显式世界模型

Explicit World Model

做法：直接预测像素、视频或几何结果

优点：

✓ 可解释性强，可视化直观

✓ 容易接入逆动力学模块

代价：

✗ 计算量大

✗ 长时预测误差累积

隐式世界模型

Implicit World Model

做法：在潜空间中建模动力学

优点：

✓ 计算高效，紧凑表示

✓ 聚焦控制相关信息

代价：

✗ 解释性较弱

✗ 依赖隐空间质量

像素级预测为什么重要

直观可视化

未来计划可以直接"看"到，便于理解和调试

易于对接

可接入逆动力学或视觉规划模块

人机交互

生成的视频可作为人机沟通媒介

可解释性强

便于分析模型预测的对错

典型流程

生成未来视频帧 → 视觉分析 → 提取动作参数 → 执行控制

但像素级预测为什么又不够好

计算量大

逐像素生成需要大量计算资源，推理速度慢

误差累积

长时滚动预测中，小误差逐步放大导致预测失真

关注错重点

过度关注视觉细节而非真正影响控制的物理因果

解决方向

- ✓ 使用**隐式表示**替代像素级预测，在潜空间中建模
- ✓ 采用**层次化预测**，粗到精逐步细化
- ✓ 引入**物理约束**，让模型学习真正的因果结构

RSSM：经典潜动力学框架

PlaNet和Dreamer系列使用RSSM，结合**确定性记忆**和**随机状态**，在潜空间中学习动力学：

确定性状态 h_t

RNN风格的循环隐藏状态

负责记忆历史信息

随机状态 z_t

变分推断的隐变量

捕捉环境的不确定性

组合状态 s_t = [h_t, z_t]

在潜空间中进行预测和规划

Dreamer系列为什么重要

Dreamer证明世界模型不仅能做预测，还能直接支持RL中的**imagined rollouts**：

核心贡献

样本效率提升

在想象中生成大量训练数据，减少真实交互

端到端学习

世界模型与策略联合优化

SOTA性能

在多个连续控制基准上达到最优

Transformer世界模型的出现

随着序列建模能力增强，研究者用**Transformer替代RNN**建模长时依赖：

RNN vs Transformer 时序建模

RNN 系列

- 顺序处理，难以并行
- 长时依赖容易遗忘
- 梯度消失/爆炸问题

代表：RSSM, PlaNet

Transformer 系列

- 自注意力捕获全局依赖
- 可并行处理，训练高效
- 长距离依赖建模能力强

代表：TSSM, TransDreamer

Trend: Transformer-based WM 成为主流，与LLM技术栈统一，便于Scale Up

JEPA：不一定要重建像素

JEPA（Joint Embedding Predictive Architecture）**直接预测抽象表示**而非重建原始输入：

与传统方法对比

传统方法（重建）

预测原始像素 → 关注视觉细节

计算开销大，易过拟合纹理

JEPA（预测表示）

预测抽象表示 → 关注可预测结构

高效，关注真正重要的信息

V-JEPA对世界模型的启发

V-JEPA将JEPA扩展到视频，用**潜表示预测替代像素重建**：

核心机制：视频块掩码预测

关键启发

抽象预测足够：不重建像素也能学到强大的世界表示

可服务控制：学到的表示可用于下游决策和规划任务

视频世界模型的技术底座

视频世界模型与视频生成技术紧密相关，经历以下技术演进：

- **GAN 时代**

  生成对抗网络，训练不稳定但开创视频生成先河

- **U-Net 扩散**

  扩散模型取代GAN，生成质量大幅提升，训练更稳定

- **ViT / DiT**

  视觉Transformer将视频视为时空token序列，统一架构

- **Latent Diffusion**

  在潜空间做扩散，效率与质量兼得，Sora等模型的基础

VLA是什么

VLA = Visual-Language-Action Model

将机器人控制表述为多模态序列建模问题

VLA 基本架构

VLA是当前**通用机器人策略**的重要范式

核心思想：将控制问题转化为"条件生成问题"

输入视觉+语言 → 输出可执行动作

VLA的典型融合方式

早期融合策略

特征调制

用语言特征调制视觉特征

交叉注意力

视觉-语言交叉注意力融合

简单拼接

直接拼接视觉和语言特征

动作生成头

自回归动作头

逐个预测动作token

适合离散动作空间

扩散动作头

生成连续动作分布

适合精细连续控制

VLA的优势在哪里

语义泛化能力

借助大规模视觉语言预训练，理解开放词汇指令和新物体语义

跨任务迁移

同一策略可执行多个不同语言指令，具备任务泛化能力

数据效率高

利用互联网规模的视觉-语言数据预训练，减少机器人数据需求

统一框架

将感知、理解和控制统一到多模态序列建模框架中

VLA的核心短板在哪里

核心短板：缺乏显式世界动态建模，在需要前瞻物理推理、长时计划和环境后果评估时容易失效

"Can a policy act without imagining consequences?"

具体表现

物理推理弱

无法理解推物体会使其移动

长时规划差

难以完成多步骤任务

后果评估缺失

不能预判危险操作

这正是 **WAM（World Action Model）** 出现的背景和动机

本节内容

CONTENTS

一、世界模型基础概念

二、世界动作模型的定义、分类与核心架构

三、WAM的数据来源、训练策略与应用场景

四、WAM的评测体系、核心挑战与未来方向

机器人策略学习的三类范式

直接模仿
Behavior Cloning / VLA

从专家演示直接回归动作；VLA 把视觉-语言-动作统一到大模型里。

- 数据效率高、上手快
- 泛化弱、缺乏对未来的显式预测
- 代表：RT-2、OpenVLA

想象中学习
Model-Based RL / World Model

学习环境动力学模型，在"想象"中规划或训练策略（Dreamer 系）。

- 样本效率高、可规划
- 像素/状态预测与控制解耦
- 代表：Dreamer、DayDreamer

联合预测
Generative World Model / WAM

用生成式（视频扩散）世界模型，联合预测未来世界状态与动作。

- 视频作稠密监督、强泛化
- 推理成本高、物理一致性待解
- 代表：DreamZero、DreamGen

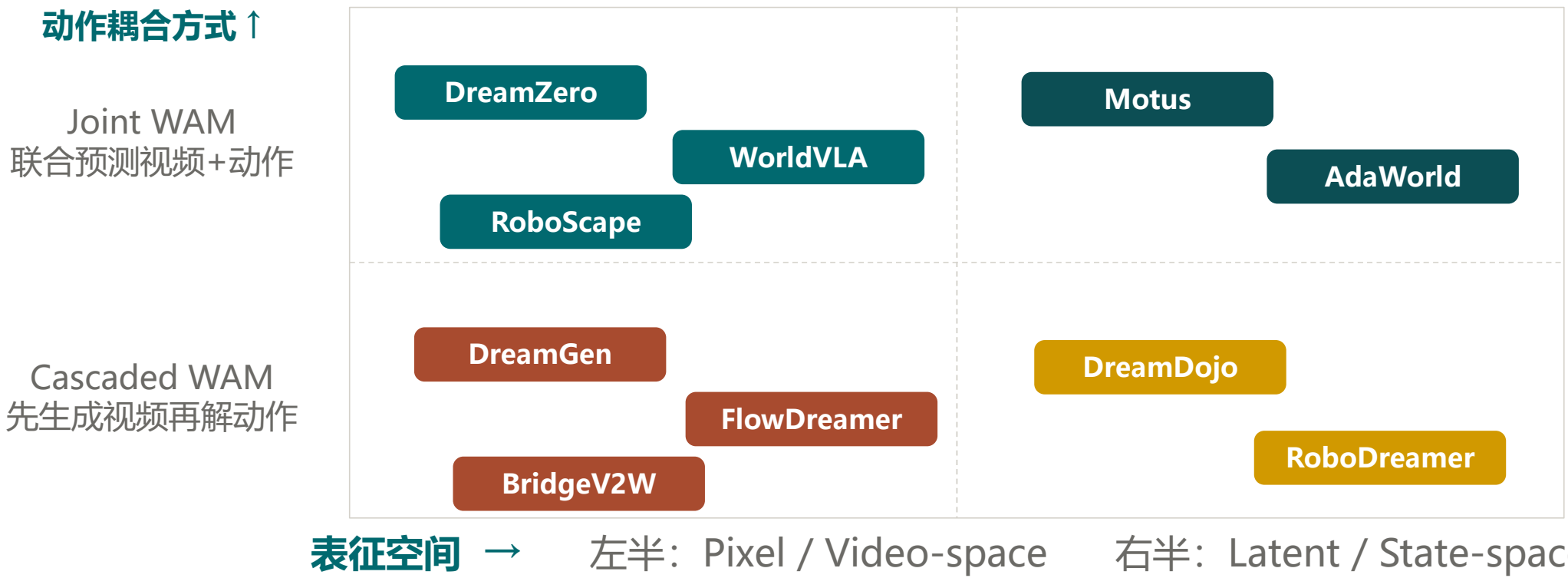
World Model、VLA、WAM基本概念

World Model (WM)	学习"如果这样做会怎样"——预测未来观测/状态的生成或预测模型。	*可在想象中规划，但本身不直接输出可执行动作。*
Vision-Language-Action (VLA)	把图像、指令直接映射为动作的大模型策略。	*动作直接、但缺乏对未来的显式建模，泛化依赖海量动作标注数据。*
World Action Model (WAM)	在预训练视频扩散骨干上，联合预测未来世界状态与机器人动作。	*视频提供稠密学习信号；动作与视觉共享表征——但推理需实时性。*

WAM 把"想象未来"（WM）与"输出动作"（VLA）耦合进同一个生成模型——收益是否真的来自测试时的未来想象，还是训练时的视频表征？

Source: NVIDIA Glossary: World Action Model

WAM 的设计空间



第三维（推理模式）： imagine-then-execute（测试时先想象再执行） vs representation-only（仅训练时用视频、推理跳过想象，如 Fast-WAM）

发展时间线：从 Dreamer 到 WAM

2018–22

MBRL 奠基

Dreamer • DayDreamer

世界模型在想象中训练策略，DayDreamer 首次搬到真实机器人。

2024

机器人想象

RoboDreamer • FLIP • DreamGen

视频生成驱动机器人想象，合成神经轨迹解锁泛化。

2025

统一与适配

WorldVLA • AdaWorld • RoboScape

动作-世界统一建模、潜动作可适配、物理信息注入。

2026

WAM 成形

DreamZero • Motus • Fast-WAM • DreamDojo

联合视频-动作、实时闭环、并开始质疑测试时想象。

Source: DayDreamer · RoboDreamer · DreamGen · DreamZero

WAM：世界动作模型的正式定义

WAM = 同时统一环境动态建模与动作生成的具身基础模型

不仅要预测未来状态，还要让动作严格对齐于这些预测未来

核心公式

$$\text{future, action} = \text{WAM}(\text{observation, instruction})$$

关键特征

统一建模

世界动态 + 动作生成

预测驱动

动作基于未来预测

对齐约束

动作与未来状态耦合

WAM成立的两个必要条件

1 Forward Predictive Modeling

模型必须具备**前向预测建模**能力

能够生成或利用可量化的未来状态表征

不只是特征提取，而是真正的未来预测

例：预测"推杯子后杯子位置的变化"

2 Coupled Action Generation

动作生成必须和**未来状态耦合**

不是独立生成动作，而是基于预测未来来生成

动作与未来预测之间有显式联系

例：根据预测到的杯子位置计算抓取轨迹

WAM与普通世界模型的区别

普通世界模型

角色：环境预测器
输入：当前状态 + 动作
输出：未来状态

局限：
- 只负责"预测"
- 不直接生成动作

WAM

角色：预测 + 动作生成
输入：当前观察 + 指令
输出：未来状态 + 动作

优势：
- 预测服务于动作
- 端到端可学习

核心区别：WAM 把世界预测和动作控制放进**同一范式**，强调"预测未来是为了生成行动"，而非仅仅预测。

WAM与VAM、Video Policy的区别

Video Policy

从视频直接输出动作

可能借用视频骨干特征

不一定有显式世界预测

VAM

视频动作模型

生成视频+动作

但局限于视频形式

WAM

最宽泛的概念

不要求视频形式

但要求显式世界预测承诺

WAM的总体分类

Cascaded WAM 级联式

先想象，再解码动作

流程:
预测未来状态 → 动作模块解码

特点:
- 结构清晰，两阶段解耦
- 可分别优化各模块

Joint WAM 联合式

预测和动作一体化

流程:
统一架构共同建模

特点:
- 共享表征，共同优化
- 强调因果耦合

Cascaded WAM：先想象，再解码动作

Stage 1 **预测未来**

输入：当前观察 + 指令

输出：未来状态/计划表示

形式：视频帧 / 隐表示 / 轨迹

↓

Stage 2 **解码动作**

输入：预测的未来表示

输出：可执行动作

模块：逆动力学 / 动作解码器

结构特点

优点：

✓ 结构清晰模块化

✓ 两阶段可独立优化

✓ 易于调试和替换

缺点：

✗ 两阶段耦合较松

✗ 预测误差影响动作

✗ 非端到端训练

Joint WAM：预测和动作一体化

统一架构

在一个统一架构中共同建模未来状态和动作，使两者共享表征并共同优化。

强调"未来预测"与"动作生成"的因果耦合。

架构示意

级联式的直观理解

可以把级联式理解为：**"先脑补未来视频，再从视频里倒推出怎么做"**

1 脑补未来

想象执行动作后的场景变化

2 分析视频

从脑补的画面中分析物体运动

3 执行

倒推出具体动作并执行

本质

级联式 WAM 更像**显式规划流程**：先规划（预测未来），再执行（解码动作）。

与人类"先想清楚再动手"的直觉一致。

联合式更像真正的一体化代理

联合式 WAM 不把预测和控制拆开，而是在**同一模型内部**共同学习：

核心优势

两个任务互相促进	端到端梯度传播	更接近统一智能体

级联式中的显式规划

显式像素或视频规划路线：把**未来视频当作中间计划载体**

为什么用视频作为Plan Carrier

可解释

人类可以直接观看理解

直观

视觉化的计划表示

易展示

便于人机交互和调试

流程

级联式中的隐式规划

不生成完整视频，只预测**潜空间未来表征**，再从表征中解码动作：

隐式规划流程

相比显式规划的优势

更高效

无需生成完整视频帧

更聚焦

只保留控制相关信息

更紧凑

低维表示节省存储

显式规划的优点与代价

优点

1. 可解释性强

生成的视频可以直接观看和验证

2. 可视化强

便于调试和展示

3. 容易对接

可与逆动力学模块直接对接

代价

1. 计算量大

生成视频需要大量计算资源

2. 时延高

视频生成速度慢，影响实时性

3. 误差累积

长时预测误差逐步放大

隐式规划的优点与代价

优点

1. 紧凑高效

低维表示，计算开销小

2. 聚焦控制

重点放在控制相关表征上

3. 推理快速

潜空间操作速度快

代价

1. 解释性弱

隐表示难以直接理解

2. 质量依赖

依赖潜空间是否保留物理因果

3. 调试困难

难以可视化中间结果

联合式WAM的两类生成头

联合式 WAM 中，根据生成方式可分为两大类：

→ 自回归范式

Autoregressive

把未来和动作看作token序列
逐个token自回归生成

优势：
✓ 与LLM范式兼容
✓ 序列建模清晰

局限：
✗ 长序列效率问题

扩散/流匹配

Diffusion / Flow Matching

联合生成未来状态和动作
通过去噪过程生成

优势：
✓ 建模多模态分布
✓ 生成质量高

局限：
✗ 推理时延较大

Autoregressive Joint WAM

自回归式方法：将未来和动作表示为token序列，**逐个生成**：

[obs_token] → [future_token_1] → [future_token_2] → [action_token_1] → ...

优势

LLM范式兼容

可直接使用大语言模型架构

序列建模清晰

因果关系明确

易于扩展

支持多模态token统一

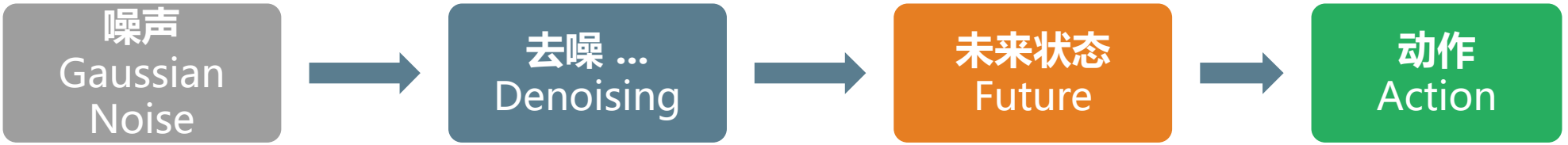
限制

长序列生成效率低，误差累积问题，推理速度受限

Diffusion-based Joint WAM

扩散式方法：通过**去噪过程**联合生成未来状态和动作：

去噪流程



优势

多模态建模

擅长建模多模态未来分布

高质量生成

连续动作生成质量高

局限：推理时延

多步去噪过程导致推理慢

Unified Stream：一个主干包办两件事

Unified Stream 把未来预测和动作生成编码进**同一预测流**：

核心特点

结构统一

单一网络处理所有任务

共享充分

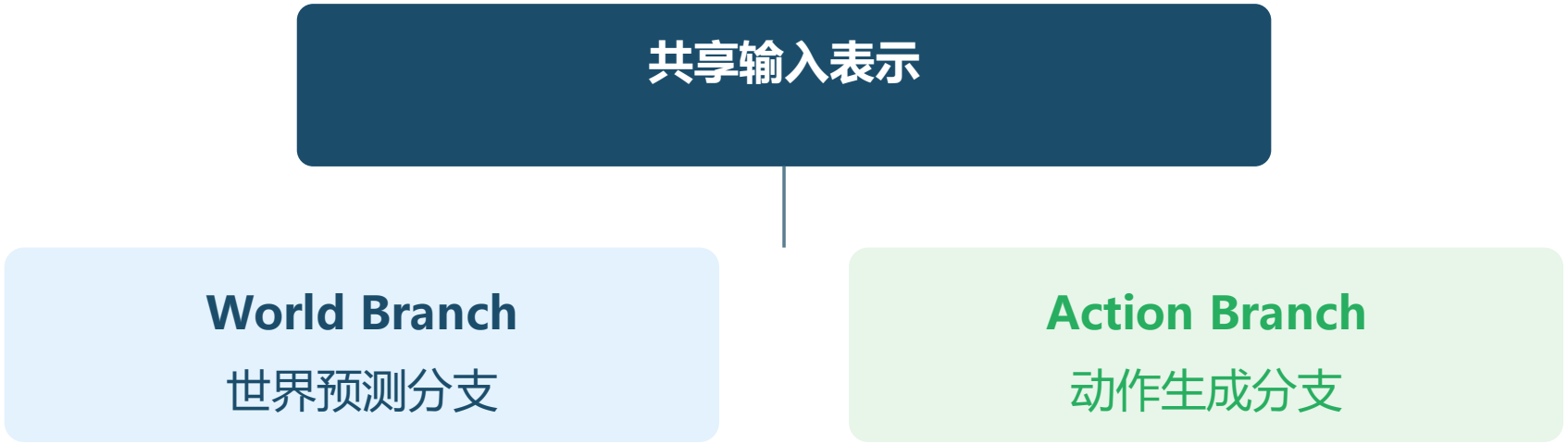
所有参数共享

目标关键

训练目标设计更重要

Multi-Stream：多分支协同

Multi-Stream 给 world branch 和 action branch **分开建模**，再交换信息：

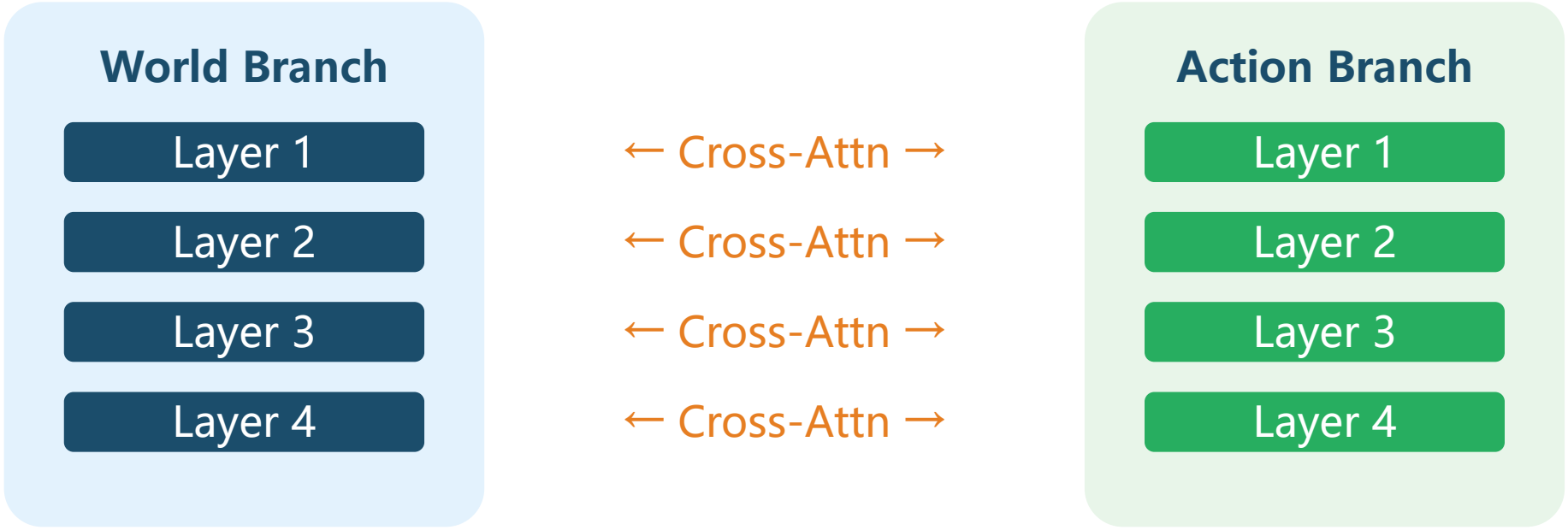


信息交换机制

交叉注意力 Cross-Attention　　隐藏状态 Hidden States　　共享表示 Shared Repr

Cross-Attention Coupling

世界分支和动作分支结构独立，在多个层级通过**交叉注意力**交换信息：



Hidden-State Coupling

世界分支生成**隐藏状态**，作为条件提供给动作分支：

优势

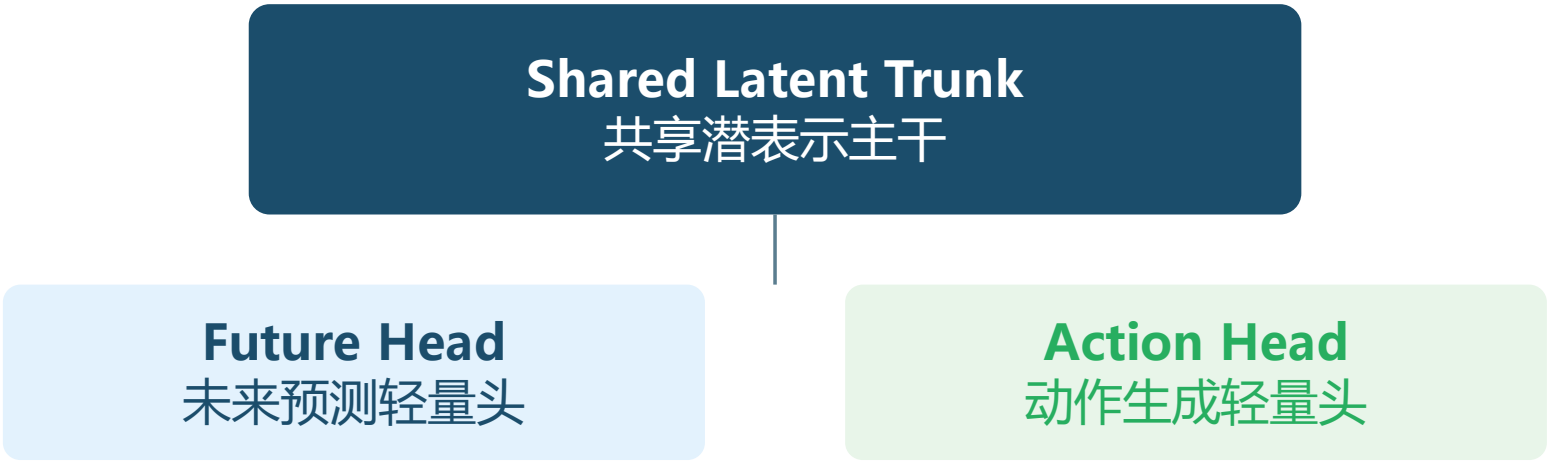
降低推理成本：无需生成完整视频

信息压缩：隐藏状态是紧凑的表示

Shared Representation Coupling

世界预测与动作生成**共享同一潜表示**，再由不同轻量头解码：

架构



优势

训练更统一 | 推理更灵活 | 参数更高效

为什么"预测未来"能帮助"生成动作"

动作不再只依赖当前观察，而是依据**"未来可能会发生什么"**来做选择：

物理理解

预测未来需要理解物理规律，这种理解自然提升动作质量

泛化能力

理解"什么导致什么"后，更容易泛化到新场景

安全性

能预判危险后果，避免执行有害动作

规划能力

预测使系统能进行"如果...就..."的推理

多模态未来：视觉不是全部

当前 WAM 大多局限于RGB预测，但接触丰富操作还需要更多模态：

RGB

视觉图像

Depth

深度信息

Force

力觉反馈

Tactile

触觉传感

Proprio

本体感知

为什么需要多模态

接触丰富的操作（如抓取、装配）仅靠视觉不足以判断：

- 力觉告诉"抓得够不够紧"
- 触觉告诉"表面是什么材质"
- 本体感知告诉"手臂在哪里"

代表性WAM：DreamZero（一）：问题动机

VLA 的瓶颈：泛化依赖海量、昂贵的机器人动作标注，面对未见任务/物体/具身时表现脆弱。

海量"无动作"视频（其他机器人、人类示范）蕴含丰富的物理与任务知识，却难以直接训练策略。

核心问题：能否把预训练视频生成模型的世界知识，直接转化为零样本可执行的机器人策略？

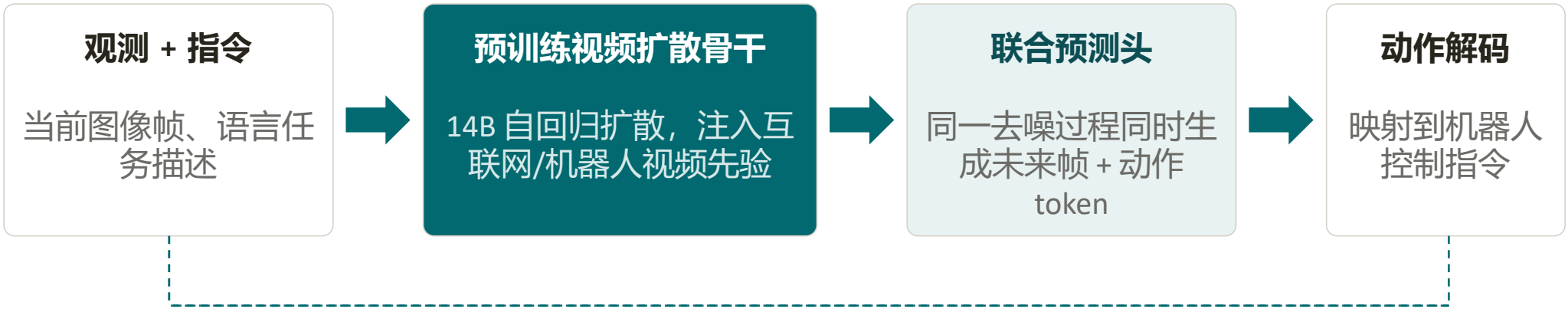
DreamZero 的主张

在预训练视频扩散骨干上联合预测未来世界状态与动作，使视频成为稠密的学习信号——把"看会做"转化为"零样本去做"。

> 2×	真实机器人实验中相对 SOTA VLA 的泛化提升
7 Hz	14B 自回归视频扩散模型实时闭环控制频率
+42%	仅 10–20 min 他者视频示范带来的未见任务提升

Source: DreamZero, arXiv:2602.15922

代表性WAM： DreamZero（二）：架构与联合建模



联合而非级联

视频帧与动作在同一生成过程中被预测，动作与视觉表征共享，避免级联误差。

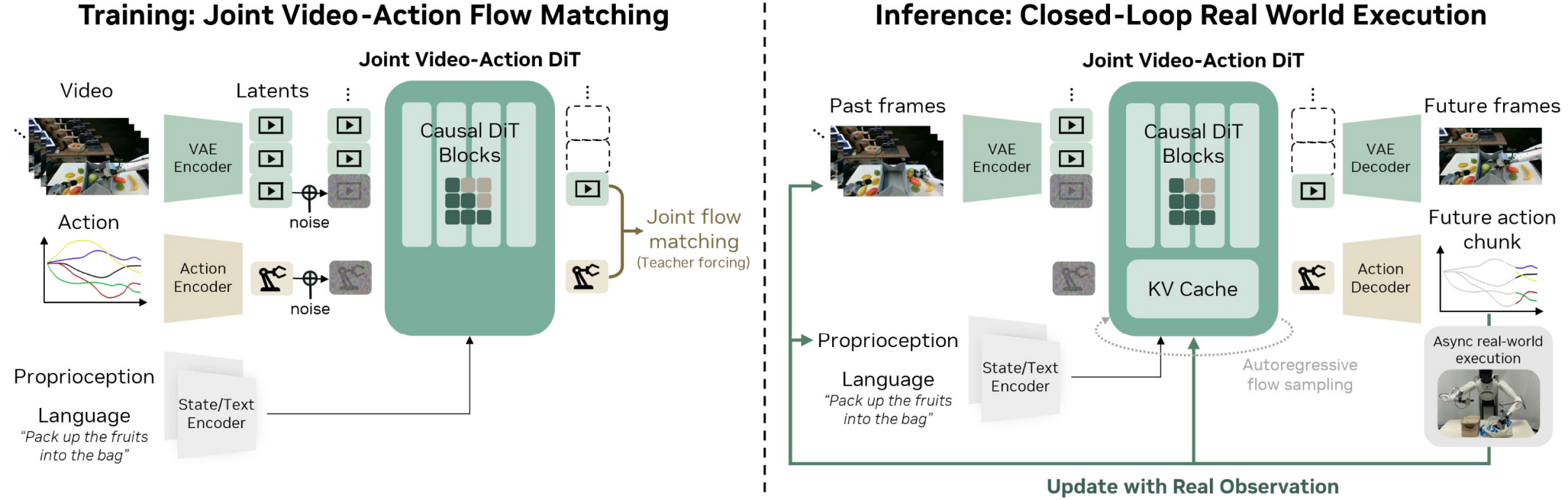
视频作稠密监督

未来帧预测为动作学习提供逐像素的稠密信号，降低对动作标注的依赖。

零样本策略

世界模型的预测能力直接被解读为可执行策略，无需任务特定微调即可迁移。

Source: DreamZero, arXiv:2602.15922 · NVIDIA WAM Glossary


Training: Joint Video-Action Flow Matching
Joint Video-Action DiT
Video
Latents
VAE Encoder
noise
Causal DiT Blocks
Joint flow matching
(Teacher forcing)
Action
Action Encoder
noise
Proprioception
Language
"Pack up the fruits into the bag"
State/Text Encoder
Inference: Closed-Loop Real World Execution
Joint Video-Action DiT
Past frames
VAE Encoder
Causal DiT Blocks
KV Cache
VAE Decoder
Future frames
Future action chunk
Action Decoder
Async real-world execution
Proprioception
Language
"Pack up the fruits into the bag"
State/Text Encoder
Autoregressive flow sampling
Update with Real Observation

代表性WAM： DreamZero（三）：训练数据与目标

预训练

在大规模互联网与机器人视频上预训练视频扩散骨干，习得通用物理与场景动力学先验。

联合目标

在带动作标注的机器人数据上联合优化：未来帧的扩散去噪损失 + 动作预测损失。

无动作视频

其他机器人、人类示范视频以"仅视频"形式参与，提供任务语义与运动模式监督。

数据信号的层级

互联网视频

最丰富 · 通用物理先验

他者机器人 / 人类视频

中等 · 任务语义、运动模式

目标机器人动作数据

最稀缺 · 精确动作对齐

Source: DreamZero, arXiv:2602.15922

代表性WAM： DreamZero（四）：实验结论与评价

> 2× **泛化提升** 真实机器人实验中相对 SOTA VLA	+42% **未见任务** 仅 10–20 min 视频示范 (video-only)	30 min **新具身适配** 用约 30 min play data 适配新机器人	7 Hz **实时闭环** 14B 自回归视频扩散完成实时控制

优点

- 零样本/弱监督泛化显著优于 VLA 基线
- 可吸收无动作视频，数据来源大幅拓宽
- 联合建模实现实时闭环（7 Hz），具落地性
- 低成本跨具身适配（30 min play data）

局限

- 14B 模型推理成本高，实时性靠工程优化勉强达成
- 生成视频的物理一致性、长时序漂移仍受质疑
- "想象未来"是否真带来收益尚待消融（见 Fast-WAM）
- 结果多为论文自报真实机器人实验，复现门槛高

Source: DreamZero, arXiv:2602.15922 · Project Page

代表工作横向对比

论文	表征 / 动作接口	数据来源	推理模式	核心贡献 • 解决的问题
DreamGen	Pixel · 伪动作(IDM/潜动作)	视频世界模型生成的合成轨迹	Cascaded	4 阶段管线生成神经轨迹；单一抓放遥操作数据→人形机器人 22 种新行为
DreamDojo	Latent · 连续潜动作代理	44k 小时第一视角人类视频	Cascaded	从大规模人类视频学通用机器人世界模型；实时 10.81 FPS，支持遥操/评估/规划
WorldVLA	Pixel · 自回归动作 token	机器人图像-动作序列	Joint	统一动作与图像理解/生成；注意力掩码策略改进动作块生成
AdaWorld	Latent · 自监督潜动作	无标注视频	Joint	从视频自监督学潜动作；少量交互/微调即可适配新世界
RoboScape	Pixel + 物理先验	RGB 视频 + 深度/关键点	Joint	物理信息具身世界模型；时序深度与关键点动力学；服务策略训练/评估
FlowDreamer	RGB-D · 3D 场景流	机器人 RGB-D 序列	Cascaded	显式 3D 场景流再扩散未来帧；提升语义相似度/像素质量/成功率
BridgeV2W	Pixel · 具身掩码控制	URDF/相机渲染的具身掩码	Cascaded	坐标动作→像素对齐具身掩码(ControlNet 式)；策略评估与目标条件规划
Motus	Latent · 光流潜动作	多任务理解/视频/动作	Joint	Mixture-of-Transformer 统一专家；UniDiffuser 式调度支持 WM/VLA/IDM/联合预测
Fast-WAM	Pixel · 视频协同训练	机器人视频 + 动作	Repr.-only	质疑测试时想象；保留视频协同训练但推理跳过未来预测；190ms，>4× 提速

Source: arXiv 各论文 · Fast-WAM Project · FlowDreamer RA-L

本节内容

CONTENTS

一、世界模型基础概念

二、世界动作模型的定义、分类与核心架构

三、WAM的数据来源、训练策略与应用场景

四、WAM的评测体系、核心挑战与未来方向

为什么数据是WAM的瓶颈

动作对齐数据

需要严格动作对齐的机器人演示数据，用于学习精确控制

物理先验数据

希望吸收互联网视频的大规模物理先验，学习通用物理知识

矛盾与挑战

机器人数据：动作精确但规模小、场景有限

互联网视频：规模大但缺少动作标签

WAM 需要同时利用两者，数据需求比 VLA 更复杂

四类核心数据来源总览

1 机器人遥操作

精确动作对齐的 (o_t, a_t, o_{t+1}) 三元组

质量高但采集昂贵

2 人类示范

便携式设备采集的人类操作数据

场景丰富但需动作重定向

3 仿真数据

程序生成的合成数据

易扩展但存在sim-to-real gap

4 互联网视频

大规模人类视频（Ego4D等）

物理先验丰富但无动作标签

机器人遥操作数据

提供高频、精确、动作对齐的 **(o_t, a_t, o_{t+1}) 三元组**：

o_t（当前观察） + a_t（执行动作） → o_{t+1}（下一观察）

精确的动作条件动力学数据

为什么是最可靠的数据来源

动作精确

遥操作记录真实控制信号

因果对齐

动作与观测严格对应

高质量

人类专家级操作示范

遥操作数据的优点与限制

优点

1. 动作真实

记录真实控制信号

2. 控制精确

高频精确采集

3. Sim-to-real gap小

直接在真实机器人上采集

限制

1. 采集昂贵

需要人类操作员和设备

2. 环境封闭

通常固定场景

3. 形态多样性不足

针对特定机器人类型

便携式人类示范数据（UMI路线）

UMI（Universal Manipulation Interface）通过**便携设备**在真实环境中采集人类操作：

核心作用

连接"**真实世界丰富性**"和"**可执行动作**"的重要桥梁

在真实环境中采集 → 对齐到机器人动作 → 用于训练WAM

为什么UMI风格数据很重要

保留场景多样性

在真实野外场景中采集，保留环境的丰富性和多样性

接近控制约束

提供比纯互联网视频更接近机器人控制的动作约束

对WAM学习的关键价值

UMI风格数据让WAM能够学习**现实物理变化**，而非仅限于仿真环境。

这是连接"人类世界"和"机器人世界"的关键数据类型。

仿真数据的作用

仿真数据的优势

易扩展

可程序生成大量数据

可控

精确控制场景参数

特权信息

深度、法向、分割等

特权几何监督

仿真环境可提供真实世界难以获取的**3D/4D监督信号**：

● 深度图 Depth ● 表面法向 Normal ● 点云 Point Cloud ● 语义分割 Segmentation

这些特权信息对3D/4D世界建模至关重要

互联网和第一视角人类视频的作用

大规模人类视频缺少精确动作标签，但提供**海量物理先验**：

代表性数据集

Ego4D | EPIC-KITCHENS | HowTo100M | Something-Something

核心价值

物体交互先验

人类如何与物体交互

物理常识

重力、碰撞等物理规律

开放场景

多样化的真实环境

动作缺失怎么办：推断Latent Action

互联网视频没有动作标签，如何从中学习？**推断潜在动作**：

Latent Action 推断流程

代表性工作

Genie

从视频中学习潜在动作

DreamDojo

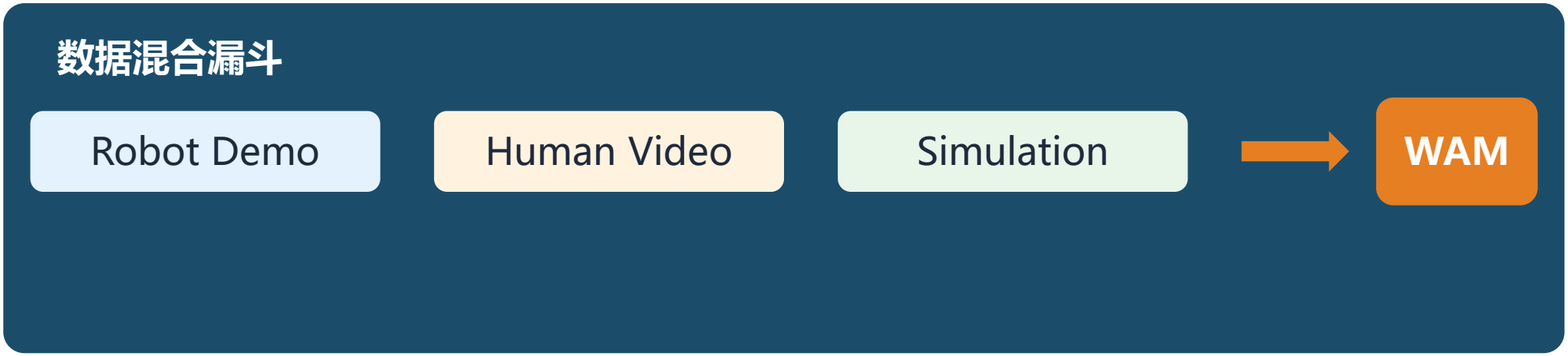
大规模视频动作学习

SWIM

无监督动作发现

数据混合为什么是WAM的独特优势

WAM 既能用高质量动作数据学习控制，也能吸收无动作视频学习物理先验：



相比VLA的优势

VLA 只能使用动作标注数据，而 WAM 可以**天然利用多源数据联合训练**

数据混合的关键问题

数据混合仍是**开放问题**，核心挑战包括：

多少数据?

每种数据源的边际价值是什么?

什么顺序?

最佳课程学习路径是什么?

怎么过滤?

如何筛选高质量数据?

这些问题目前尚无定论，是 WAM 领域的重要研究方向。

需要大量实证研究来探索最优数据混合策略。

世界模型如何帮助模仿学习

世界模型可以**合成或筛选训练轨迹**，提升训练样本的多样性与覆盖范围：

世界模型辅助模仿学习

核心作用

数据增强

合成多样化训练样本

覆盖扩展

探索专家未覆盖的状态

安全探索

在想象中安全试错

世界模型如何帮助强化学习

世界模型作为 **surrogate environment**，供 agent 在想象空间中试错：

真实环境 vs 想象环境

真实环境交互

✗ 成本高（硬件损耗）

✗ 有风险（可能损坏）

✗ 速度慢（物理时间）

想象环境交互

✓ 成本低（纯计算）

✓ 无风险（虚拟环境）

✓ 速度快（并行推演）

核心收益：大幅减少真实环境交互成本和风险，同时提升样本效率

世界模型如何充当奖励模型

利用世界模型的**预测一致性**来构造奖励，减少手工奖励设计：

奖励构造流程

可使用的信号

预测一致性

预测与实际的一致性

分布熵

生成分布的多样性

目标对齐

与目标状态的接近度

世界模型如何帮助策略评测

世界模型作为**数据驱动模拟器**，不上机即可批量测试策略：

评测流程

评测能力

批量测试

大量场景快速评估

OOD场景生成

测试分布外泛化

安全红队

主动发现策略缺陷

物理模拟器和世界模型如何协同

物理模拟器

强项：显式物理、可控实验

弱项：建模真实分布有限

世界模型

强项：数据驱动、拟合真实分布

弱项：物理精确性有限

协同 → 缩小 Sim-to-Real Gap

应用一：机械臂操作

世界模型在机械臂操作中的应用是**最活跃的方向**之一：

抓取

预测抓取后果

推物

生成训练数据

装配

辅助长时规划

机械臂操作是 WAM 的**核心应用场景**

世界模型用于预测抓取后果、生成训练数据、辅助长时操作规划

桌面操作任务是最常见的评估场景

应用二：人形与移动机器人

世界模型在人形和移动平台中的关键应用：

地形预判

预测地面情况，提前调整步态

导航规划

预测路径上的动态变化

全身控制

协调多关节运动

长时任务

执行复杂多步骤任务

本节内容

CONTENTS

一、世界模型基础概念

二、世界动作模型的定义、分类与核心架构

三、WAM的数据来源、训练策略与应用场景

四、WAM的评测体系、核心挑战与未来方向

为什么评测比"视频好不好看"更难

WAM 不仅要生成**看起来合理**的未来，还要保证这些未来对**动作决策真正有用**。因此不能只用视觉质量衡量好坏。

评测维度

视觉保真度 | **物理正确性** | **动作可执行性** | **下游成功率**

四个维度缺一不可，共同构成 WAM 的完整评测体系

第一类评测：视觉保真度

评估生成未来的**视觉质量**，常用指标包括：

常用指标

PSNR	SSIM	LPIPS	FVD	DreamSim
峰值信噪比	结构相似性	感知相似性	视频分布距离	语义相似性

局限性

视觉保真度只衡量"看起来真不真"，不衡量"物理对不对"或"动作可不可行"。

高视觉质量 ≠ 好的世界模型

第二类评测：物理常识与物理正确性

模型应符合**重力、碰撞、材质和因果**等物理常识：

代表性评测基准

VideoPhy

视频物理正确性评测

PhyGenBench

物理生成基准

Physics-IQ

物理推理评测

评测维度

重力遵循 | 碰撞合理性 | 材质一致性 | 因果关系 | 时间连续性

第三类评测：动作可执行性

WAM 生成的未来必须包含足够的**行动信息**，能否反推出可执行动作：

Action Plausibility 评测流程

评判标准

动作合理性

解码的动作是否在机器人可达范围内

执行成功率

动作执行后是否达到预期效果

第四类评测：下游任务成功率

最终看下游操作、导航、装配等**任务是否成功**完成：

代表性评测基准

MetaWorld	RLBench	LIBERO	RoboCasa
元操作任务	机器人学习基准	长时操作基准	家庭机器人基准

核心指标

Success Rate = 任务成功次数 / 总尝试次数

这是衡量 WAM 实用价值的最终标准

现有评测的根本问题

核心问题："世界预测"和"动作生成"常被**分开评测**

这违背了 WAM 强调两者因果耦合的初衷

好的世界预测 + 好的动作 ≠ 好的 WAM

世界预测排行榜 ←→ 动作生成排行榜

?

两者之间的因果一致性被忽略

未来评测方向：联合评估因果一致性

理想评测应同时检验 imagined future 和 generated action 是否**因果一致**：

Counterfactual Consistency

反事实一致性：改变动作，未来是否相应变化？

Foresight-Conditioned Success

前瞻条件成功率：动作是否遵循所想象的未来计划？

愿景：建立统一的 WAM 评测框架

同时衡量世界预测质量、动作生成质量、以及两者的因果一致性

挑战一：长时任务与层次规划

当前 WAM 多在**短时操作任务**上评估，通用具身智能需要解决长时任务：

当前：短时抓取

单步或几步操作

误差累积有限

挑战：长时整理桌面

多步骤复杂任务

误差累积严重

核心难题

分布漂移

长时执行中状态偏移

误差累积

每步小误差逐步放大

层次分解

任务如何分解为子目标

挑战二：物理规律与反事实推理

大模型可能学到**相关性而非真正因果**，如何学到稳健的物理规律？

核心问题：模型知道"球通常会下落"，但不知道"如果推得更重，球会飞得更远"

缺乏反事实推理能力

反事实推理示意

"轻推杯子"

→ 杯子移动一点

"重推杯子"

→ 杯子移动很多？

"不推杯子"

→ 杯子不动？

挑战三：效率与时延

把世界预测纳入闭环控制带来**严重延迟负担**：

控制频率对比

常规 VLA

控制频率：10-50 Hz

实时性较好

扩散/视频 WAM

控制频率：1-5 Hz

难以达到实时控制

瓶颈来源

- 多步去噪扩散
- 高分辨率视频生成
- 长序列自回归

挑战四：安全、可靠性与可验证性

WAM 能预测未来，但如果**预测错了**，可能更自信地执行危险动作：

风险悖论：错误的未来预测 → 错误的动作信心 → 危险执行

模型"以为"自己知道未来，但实际上预测有误

解决方向

不确定性估计

量化预测不确定性

安全验证

推理时验证动作安全性

保守策略

不确定时选择安全动作

问题和讨论