

具身智能中的世界模型与世界动作模型：训练、推理与部署

从 DreamZero、Dream-1 到 VLA/WAM 算力比较：训练、推理与跨本体迁移

2026-10-22

目录

1. 为什么具身策略需要世界预测？	2
2. WAM 的四种建模范式	2
3. 训练数据：视频、动作与本体之间如何对齐？	3
4. WAM 的通用训练与推理接口	3
5. DreamZero：完整公开的 WAM 方法案例	4
6. Dream-1：零样本跨本体 WAM 的系统案例	5
7. DreamZero 与 Dream-1：技术演进与不可混淆的结论	8
8. VLA 与 WAM 的系统对比：能力、算力与部署	8
9. 高效推理：把模型时钟接到机器人时钟	11
10. 接触操作与失败恢复	11
11. 如何评价 WAM 的真实研究贡献？	11
12. 与现有课程的边界与阅读顺序	12
13. 论文与官方资料	12

[课程首页](#) · [English](#) · [DreamZero 论文](#) · [Dream-1 官方介绍](#) · [PDF](#)

项目	说明
主题	DreamZero、Dream-1 与跨本体迁移
修订日期	2026-10-22
许可	CC BY-NC-SA 4.0 · Xiong Zeng and contributors

1. 为什么具身策略需要世界预测？

本讲定位。承接本课程的 Flow Matching、模仿学习及 GR00T/SONIC，解释未来世界预测如何与可执行机器人动作耦合。课程保留通用训练、推理与部署基础，以 **DreamZero 为公开算法案例**、以 **Dream-1 为跨机器人部署案例**，并系统比较 **VLA 与 WAM 的能力、训练/推理算力和真机部署成本**。不重复连续性方程、PPO 或通用 Transformer 的完整推导，也不设置编程实验。

本次补充的核心问题。为什么 Dream-1 能在部分未见机器人和未见任务上零样本工作？答案不是“WAM 天生无需微调”，而是 **共享 EEF 动作接口、视频—动作联合世界建模、跨本体共享视觉表征** 三者配合；其中末端表示和 IK 同样可以用于 VLA。详细机制与可检验边界见第 6、8、11 节。

机器人动作的后果依赖物体位姿、环境几何、摩擦、接触模式、遮挡与扰动。常规视觉语言动作模型（VLA）直接拟合条件动作块：

$$\pi_{\theta}(A_t | h_t, \ell), \quad h_t = (o_{t-k:t}, q_{t-k:t}), \quad A_t = (a_t, \dots, a_{t+H-1}).$$

其中 o_t 是视觉/传感观测， q_t 是本体状态， ℓ 是任务指令。**VLA 并非完全不含世界知识**：它也可能隐式学习动力学。显式世界预测的额外价值在于时序监督、反事实推演和动作可行性验证。

世界模型（World Model, WM）可在观测或潜在状态空间预测未来：

$$p_{\phi}(o_{t+1:t+H} | h_t, A_t), \quad p_{\phi}(z_{t+1} | z_t, a_t).$$

世界动作模型（World Action Model, WAM）则把未来预测与**可执行动作**耦合，既可以同时生成视频与动作，也可以只在训练时把预测作为辅助监督。**WAM 不等于推理时必须生成完整未来视频，也不自动等于 MPC。**

类别	主要输出	主要学习信号	部署角色
VLA / 动作策略	动作块	动作监督、任务反馈	直接选择动作
WM	未来状态或观测	状态转移、视频时序	预测、打分、规划
WAM	与未来相关联的可执行动作	世界预测与动作的耦合监督	可执行策略，或其训练机制

视频在视觉上可信并不保证力学正确；尤其接触力、静摩擦边界和微小滑移无法单由 RGB 充分决定。

2. WAM 的四种建模范式

2.1 Imagine-then-execute。先生成目标未来 $\hat{O}^+$ ，再让动作策略对这个未来条件化：

$$\hat{O}^+ \sim p_{\phi}(O^+ | h_t, \ell), \quad A_t \sim \pi_{\theta}(A | h_t, \ell, \hat{O}^+).$$

优点是未来可解释；缺点是视频采样贵，且“看起来会成功”的未来不一定动力学可达。部分系统还对假设动作条件化未来，不能把此式当成全部 WAM 的定义。

2.2 Video-feature-conditioned。从视频世界模型读取中间表征，不必在部署时显式解码视频：

$$A_t \sim \pi_{\theta}(A | f_{\phi}(h_t, \ell), q_t).$$

必须通过真实任务成功率验证表征是否可控，而不是仅检查生成图像的质量。

2.3 Joint video-action。将未来视频与动作视为同一个条件联合分布：

$$p_{\theta, \phi}(O^+, A | h_t, \ell).$$

可以共用去噪骨干、使用跨模态 token，或采取因果分块生成。**DreamZero** 是这一类的典型公开实现。联合建模不要求所有模态采用相同的离散 token 次序。

2.4 Auxiliary video prediction. 训练时施加未来预测损失，执行时省略该分支：

$$\min_{\theta, \phi} \mathcal{L}_{\text{action}} + \lambda \mathcal{L}_{\text{future}}, \quad A_t \leftarrow \pi_{\theta}(h_t, \ell).$$

Fast-WAM 说明部分收益可以来自训练期的预测表征，而非测试时生成视频；这种结论依赖具体模型和任务，不能推广到所有 WAM。

3. 训练数据：视频、动作与本体之间如何对齐？

最小交互序列可表示为

$$\mathcal{D}_R = \{(o_t, q_t, a_t, o_{t+1}, r_t, d_t, \ell, \Delta t)\}_{t=0}^{T-1}.$$

必须记录机器人命令的控制模式、单位、坐标系、执行时间和观测时间戳； r_t 并非纯监督预训练的必要条件。区分期望命令与低层控制器实际执行的命令，对有饱和、滤波、通讯延迟的系统尤其重要。

数据来源	可学习的内容	固有限制
人类/网络视频	视觉运动与接触先验	缺可执行机器人动作标签
机器人遥操/示教	同步观测、本体状态与动作	本体和任务覆盖不充分
仿真轨迹	完整状态、接触、动作、奖励	视觉和接触 sim-to-real gap
在线交互	失败与恢复、分布修正	真实代价、安全与数据治理

混合训练的重要原则：没有动作标签的视频只能提供视频/潜变量预测监督，不能虚构“真实机器人动作”。必要时使用模态损失掩码，伪动作需明确区分重定向或逆动力学估计值。跨本体动作需说明是关节空间、末端 $SE(3)$ 、手部关节还是潜在动作。

4. WAM 的通用训练与推理接口

4.1 从预训练世界表征到动作生成

世界分支可用视频扩散、Flow Matching、latent regression、JEPA 式预测或离散 token 建模；相应监督目标并非完全相同。一个抽象的动作 Flow Matching 写为

$$A_{\tau} = (1 - \tau)\varepsilon + \tau A^*, \quad \mathcal{L}_{\text{act}} = \mathbb{E} [\|v_{\theta}(A_{\tau}, \tau, c_t) - (A^* - \varepsilon)\|_2^2],$$

其中 $\tau \in [0, 1]$ 是去噪生成时间，不是机器人物理时间 t 。更一般的训练框架可能同时包括世界预测、动作学习和跨模态一致性：

$$\mathcal{L} = \lambda_a \mathcal{L}_{\text{action}} + \lambda_v \mathcal{L}_{\text{video}} + \lambda_z \mathcal{L}_{\text{latent}} + \lambda_c \mathcal{L}_{\text{consistency}}.$$

上式是课程统一记号，不是某篇论文的原始目标。**DreamZero** 在下一节使用一个联合视频—动作速度场的 Flow Matching 目标，不能简单理解为训练两个互不关联的损失后相加。

4.2 三条推理路径

显式未来想象：生成未来视频/潜变量，再求动作。适合未来可解释或多候选评分，但可能串行叠加两类模型延迟。

联合视频—动作生成：由联合生成器一次性预测未来观测及动作块。**DreamZero** 采用这一机制，结合分块反馈避免纯开环滚动。

测试时只保留动作：未来预测只在训练中提供监督；部署时动作直接取自当前表征，有利于降低计算，但需要实测效果而不能默认与联合推理等价。

5. DreamZero: 完整公开的 WAM 方法案例

论文: Ye et al., *World Action Models are Zero-shot Policies* (2026) ·项目主页 ·代码与模型。本节讨论已公开的研究方法, 而非替 Dream-1 猜测架构。

5.1 核心思想: 未来视频为动作提供物理先验

DreamZero 从 Wan2.1-I2V-14B-480P 预训练图生视频扩散骨干出发, 将语言 c 、历史图像 $o_{0:l}$ 与机器人状态 q_l 作为条件, 同时输出未来视频 O^+ 和动作块 A 。由概率链式法则, 联合分布可以概念性地分解为

$$\underbrace{p_\theta(O^+, A \mid h_l, c, q_l)}_{\text{联合 WAM}} = \underbrace{p_\theta(O^+ \mid h_l, c, q_l)}_{\text{未来视频}} \underbrace{p_\theta(A \mid O^+, h_l, c, q_l)}_{\text{逆动力学相关映射}}.$$

这只是概率分解而非两个独立训练的模型。DreamZero 用共享的因果 Diffusion Transformer (DiT) 端到端联合去噪视频潜变量与归一化动作, 从而让预测运动与实际控制更紧密耦合。其意义不是“视频天然保证物理准确”, 而是视频基础模型已从大规模时序数据学到比静态 VLM 更丰富的运动先验。

5.2 架构与模态: VAE、状态编码、动作寄存器和因果 DiT

视觉帧先经 VAE 压缩为视频 latent z ; 语言使用文本编码器, 本体状态经 state encoder 编码; 动作由 action encoder 与 decoder 连接到 DiT。多相机视角通过拼接成复合视频帧对接视频骨干。模型将长轨迹划为连续 chunk, 当前 chunk 的联合噪声 token 能读取此前 chunk 的干净上下文。

需区分三个时间尺度: **物理时间** (控制周期与相机 FPS)、**生成时间** (Flow Matching 的 τ) 和 **chunk 序号** (自回归 k)。论文示例采用每 chunk $K = 2$ 个视频 latent 帧、训练窗口最多 $M = 4$ 个 chunk; AgiBot 视频约 5 FPS、动作 30 Hz、动作块 $H = 48$ 个控制步。字母 K 、 M 、 H 不能混用。

5.3 联合 Flow Matching: 为何不是“先视频后动作”

对第 k 个 chunk, 用 z_1^k, a_1^k 表示干净目标, z_0^k, a_0^k 表示独立高斯噪声, $\tau_k \sim \mathcal{U}[0, 1]$:

$$z_{\tau_k}^k = (1 - \tau_k)z_0^k + \tau_k z_1^k, \quad a_{\tau_k}^k = (1 - \tau_k)a_0^k + \tau_k a_1^k.$$

将二者拼接为 $x_{\tau_k}^k = [z_{\tau_k}^k, a_{\tau_k}^k]$, 目标速度为 $v^k = [z_1^k, a_1^k] - [z_0^k, a_0^k]$ 。抽象写出论文的联合目标:

$$\mathcal{L}_{\text{DZ}} = \mathbb{E} \left[\frac{1}{M} \sum_{k=1}^M w(\tau_k) \|u_\theta(x_{\tau_k}^k; \mathcal{C}_k, c, q_k, \tau_k) - v^k\|_2^2 \right], \quad \mathcal{C}_k = \{(z_1^j, a_1^j) : j < k\}.$$

这里 M 代表所采样的 chunk 数; 论文公式对求和索引的符号复用做了简化, 本讲统一使用 M 以避免与每 chunk 帧数 K 混淆。训练通过 **teacher forcing** 使用真实历史 chunk, 结合注意力因果掩码防止当前块读取未来目标。视频和动作在常规 DreamZero 训练中共享同一个噪声时间 τ_k 。

5.4 机器人数据: 大模型不意味着只需要几分钟数据

论文基于 AgiBot G1 收集约 500 小时、7,200 段、22 个环境的多样遥操纵迹, 并单独在 DROID/Franka 数据上训练比较。AgiBot 与 DROID 设置均报告 100K 训练步、global batch 128。训练更新 DiT、状态/动作编码与解码模块, 冻结文本编码器、图像编码器及 VAE; 论文尝试 LoRA 但其效果不如所用训练方案。

“30 分钟”仅指有强先验模型在新 YAM 本体上的少数数据适配, 不是从零训练 DreamZero 只需 30 分钟。重要消融表明, 在相同约 500 小时预算下, 多样轨迹比重复演示获得更强泛化 (其一组 PnP-Easy 设置的进度 50% vs 33%)。这支持“覆盖运动与状态变化的多样性”有价值, 但尚不构成所有任务的普适定律。

5.5 真实闭环：KV Cache 应怎样更新？

推理时联合采样下一段未来视频与动作块，执行部分动作后读取**真实**摄像头与本体观测，将其编码并更新作为下一轮条件的 KV Cache。于是

$$(\hat{O}_{t:t+H}, \hat{A}_{t:t+H}) \sim p_{\theta}(\cdot \mid h_t, c, q_t), \quad h_{t+1} \leftarrow \text{Update}(h_t, o_{t+1}^{\text{real}}, q_{t+1}^{\text{real}}).$$

这相当于不断用真实观察修正生成世界的条件历史，减少纯自回归生成视频的漂移；**但并不保证消除视觉估计误差、接触不确定性或长时控制误差**。KV Cache 这里主要缓存因果**历史上下文**的注意力键值，与下节在相邻去噪步复用 DiT 计算的 cache 是两回事。

5.6 从 5.7 秒到 150 毫秒：DreamZero-Flash 和系统优化

基础 14B 模型以 16 个去噪步推理约需 5.7 s/action chunk。论文组合了异步推理与执行、双 GPU classifier-free guidance 并行、DiT 重复计算缓存、`torch.compile/CUDA Graph`、特定硬件量化及内核调度优化。后续 DreamZero-Flash 在训练中**解耦视频和动作的噪声时间**：视频更常处于高噪声状态，动作仍在均匀时间采样，以学习“未来视频尚不干净也能预测有效动作”。例如论文用 $\eta \sim \text{Beta}(7, 1)$, $\tau_{\text{video}} = 1 - \eta$ ，使视频平均生成时间约为 0.125。

论文对比条件	去噪步数	Table Bussing 任务进度	单块推理延迟
DreamZero	4	83%	350 ms
DreamZero, 直接减步	1	52%	150 ms
DreamZero-Flash	1	74%	150 ms

论文报告在 **两张 GB200** 上综合达到约 **38 倍加速、150 ms/块、约 7 Hz 高层闭环**。这些不是消费级单 GPU 的保证值。机器人底层动作仍按约 30 Hz 执行； $H = 48$ 步对应 **1.6 s 动作块长度**，不能把 1.6 s 误当成网络/模型推理延迟。异步工作时必须管理新旧动作块的交接和安全限幅。

5.7 泛化证据与尚未解决的问题

AgiBot 未见环境的已见任务组，DreamZero 报告平均任务进度 62.2%，对比其最佳预训练 VLA 基线 27.4%；未见任务组约 39.5% vs 16.3%。这主要是**任务进度指标**而非全任务成功率。额外的 12 分钟人类视频或 20 分钟 YAM 视频（只加视频预测监督）可使一组未见任务进度由 38.3% 增至 54.3% 或 55.4%。另以 YAM 的 55 段、约 30 分钟 play data 进行了**有监督的本体适配**，其后才在新对象/任务上测试零样本泛化。

尚有限制：论文讨论的视觉历史约 6 s，并不等于具备可靠长时任务记忆；高质量视频也可能产生错误接触/可达性；7 Hz 高层闭环不等于毫秒级接触反馈。论文中的机器人间视频迁移、YAM 少样本适配与 Dream-1 的**真正零机器人数据**迁移是三种不同实验命题。

6. Dream-1：零样本跨本体 WAM 的系统案例

Dream Labs: *Zero-Shot to New Robots: Introducing Dream-1* (2026 年 10 月)。这是官方技术公告，**不是已披露全部训练配方的同行评审论文**。本节分清可确认机制、定量声称、未公开信息与适用边界。

6.1 从 few-shot adaptation 到 zero-shot embodiment transfer

DreamZero 的 YAM 结果是“先有 30 分钟本体数据并微调，再泛化到新任务”；Dream-1 则展示给**训练时未见过的双臂平行夹爪机器人**，不经该机器人专属示教/微调，直接根据新场景观测和语言指令执行。两者解决不同层级的泛化：

- **Unseen task / environment**：目标技能或场景不在训练集；机器人本体可能已见。
- **Few-shot embodiment adaptation**：新机器人提供少量数据更新模型，然后做未见任务。
- **Zero-shot embodiment generalization**：新机器人本身在部署前不提供训练数据，直接使用共享控制接口。

Dream-1 的突破是“机器人的形体变化不再必然要求新动作头/专属预训练数据”。但跨本体有效性仍受操纵形态和运动学分布限制，不代表适用于任意自由度、移动底盘等全部机器人。

6.2 关键接口：共享末端动作 + 机器人自身的 IK

Dream-1 可以在支持的本体上输出关节命令，也可以用更通用的**双臂末端执行器位姿（EEF pose）**及**夹爪开合**表示动作。对第 i 个机械臂，教学抽象可写为

$$a_t^{\text{shared}} = \{(T_{EE,i}^{\text{target}}, g_i)\}_{i \in \{L,R\}}, \quad q_{t+1}^{(r)} = \text{IK}_r(a_t^{\text{shared}}; q_t^{(r)}, \mathcal{C}_r),$$

$T_{EE,i} \in SE(3)$ 是目标末端位姿， g_i 为夹爪指令， IK_r 表示机器人 r 自身的运动学映射与约束处理。**该式是接口教学抽象，不是官方公开的精确 action tensor**。这种共享动作空间使不同机械臂能消费同一高层命令；与输出特定机器人关节角度的策略相比，降低了动作维度/本体耦合。**但是，EEF + IK 不是 WAM 独有的机制：采用统一笛卡尔动作空间的 VLA 也能做到这一步**。它解决“动作怎么传给新机器人”，并不能单独解决“新机器人在新任务上应该做什么”。

但 IK 只处理几何可达性与关节解，不保证动态可行、无自碰、无环境碰撞或接触稳定。官方也明确说明：性能依赖 IK 质量；策略指定“末端到哪里”，却不完全决定中间路径；运动范围或动力学相差过大时会失败。

6.3 视觉为何帮助跨本体？哪些机理尚未证实？

官方描述的输入是三路 RGB 视角：**头部一视角 + 左右腕部各一视角**。相较只依赖同款夹爪腕部视角的 UMI 迁移路线，WAM 联合建模视频与动作，让“不同机器人如何在像素里运动”成为共享的时序先验。头部视角提供全局物体几何与任务上下文，腕部视角有利于末端局部精度；共享 EEF 动作接口提供可执行的输出定义。**不同机器人的关节坐标和连杆形状并不相同，但物体接近、接触、随手移动等视觉变化有可迁移的结构**；视频预测提供的是对此类时空关系的额外监督，而不是传感器无关的精确物理定律。

其可解释的因果链是“共享视觉未来表征 $\rightarrow$ 与动作对齐的行为意图 $\rightarrow$ 通过 EEF/IK 投射到新机器人的关节命令”。**但仅凭官方博文尚不能证明该能力必然由某一种具体损失、注意力机制或内部动力学表征引起**；这些需要公开架构/训练消融才能验证。

6.4 三个机制如何共同支持跨本体与跨任务零样本泛化？

从部署机制看，Dream-1 的解释可归纳为三个**互补环节**：共享动作接口降低本体控制接口差异，视频—动作联合预测提供交互时序监督，视觉表征促进跨外观迁移。它们共同支持跨本体及跨任务泛化，但各自独立的因果贡献和缺一不可性仍须通过受控消融验证。

机制	主要缓解的泛化障碍	无法单独保证的能力
共享动作空间（EEF + IK）	6/7 自由度、连杆尺寸和关节命令不同；不同机械臂共享末端位姿语义	可达路径、避障和动力学
视频—动作联合建模	只拟合训练动作的分布，缺少动作后果的显式时序监督	接触力正确、未见任务必然成功
视觉跨本体表示（三相机）	新机器人外观、夹爪形状与训练机器人的图像差异	相机布置兼容性、完整执行接口

以教学抽象表示，它将“看到并预测交互”与“将动作落实到某一机器人”分离：

$$(o_{t-k:t}^{(r)}, q_t^{(r)}, \ell) \xrightarrow{\text{shared WAM}} (\hat{O}_t^+, A_t^{\text{EEF}}) \xrightarrow{\text{IK}_r} a_t^{(r)}.$$

视频与动作在此式中作为**耦合预测**出现，不意味着 Dream-1 官方公开了精确的两阶段推理图，也不意味着先渲染视频再运行 IK。跨本体泛化重点在于**共享控制语义 + 视觉动力学先验**；跨任务泛化则需要语言条件、具有变化性的交互数据以及可迁移的物体运动与接触模式。例如，在不同场景学习过接近、夹持、搬运和推动后，模型可能将这些行为规律组合到未见指令，但是否成功仍取决于训练覆盖与真实反馈。

与典型直接动作 VLA 相比，联合预测未来观测相当于引入更丰富的时序训练信号；这**不等于** VLA 不学习动力学，也**不等于** 只有 WAM 才能零样本跨机器人。反例是 **LAP-3B**：其使用语言形式的动作表示，在论文测试的新机器人与操作任务上报告**超过 50% 的平均零样本成功率**。因此，要验证“视频预测带来的独立增益”，必须在相同动作空间、相机、训练数据与控制协议下比较 VLA 与 WAM。

6.5 泛化不等于可靠执行：零样本和微调分别解决什么？

应将两种研究目标分开：**零样本能力衡量启动新任务/新本体所需的额外数据**；**闭环成功率衡量该能力是否可靠**。视频世界建模对接近、夹持、搬运等可见运动有助益，但插入、工具使用、精密对齐与滑移恢复可能因视觉歧义、接触力不可观测及可达性限制而失败。Dream-1 在多任务类别中也报告明显的性能差异；因此“可尝试未见任务”不能写成“未见任务均无须微调并可稳定完成”。若零样本表现不足，仍然需要任务数据、FaaS、低层控制器或强化学习后训练。

6.6 推理、API 与实时性：动作覆盖时长不等于推理耗时

官方提供云端 API，客户端选择 checkpoint、发送 observation、获得 action。Dream-1 和 Dream-1-flash 的公开标价分别约为 \$3/小时、\$2/小时；每次返回约 **1.6 秒的动作**，按持续请求估算每小时为 2,250 次调用。收费口径是按连续动作小时的等效价格，不能直接作为 GPU 推理成本或任务完成成本。

对于实际闭环，总延迟须测量

$$T_{e2e} = T_{\text{sense}} + T_{\text{encode}} + T_{\text{upload}} + T_{\text{model}} + T_{\text{download}} + T_{\text{dispatch}}$$

官方没有在该页面完整披露 Dream-1 的模型规模、去噪步数、真实网络 p95 时延、端侧缓存策略与低层控制频率。因而不能把 DreamZero 的 150 ms 或 7 Hz 宣称直接继承为 Dream-1 的 API 延迟。线上部署还需处理断网、过时动作块、状态同步、时间戳与安全停止。

6.7 已披露的评测：怎样读“27 台未见机器人”？

官方展示了 **27 种未见仿真机器人** 及多种真实双臂平行夹爪机器人上的 rollout，并给出一项仿真评测说明：每个本体 300 次 rollout，其中 150 次用于 50 类语言跟随任务，另 150 次用于 150 类未见任务。页面也展示 12 个合作方站点及不同任务类别的案例。应分清“视频示例”“单次 rollout”和“有定义的任务成功统计”，不能把挑选出的展示视频直接当作平均成功率。

官方给出的未见仿真机器人表现，是相对于已训练的 **YAM = 100%** 的归一化评分：例如双臂 PiPER **91.2%**、双臂 Franka FR3 **58.3%**、YAM-7 **38.0%**。这些值**不是绝对任务成功率**。它们同时说明了“可以迁移”和“机器人之间迁移质量差异显著”，不能凭平均演示视频断言所有机器人无需适配。

Dream Labs 引用第三方 RoboCurve 的 RoboDojo-RC Tier 1 评测；这一结果属于另一套任务/成本口径，**不能与 DreamZero 的任务进度百分比直接横向比较**。课程记录其存在，但不把宣传性榜单判断替代为统一控制基准上的证据。

6.8 FaaS：短数据本体适配的边界

对零样本能力不足的机器人，Dream Labs 提供 Fine-tuning-as-a-Service (FaaS)。官方展示：某两款本体分别采集约 **5 分钟** 数据后，在特定任务上表现改善；另一个更换不同末端器具的 YAM 案例约 **98 分钟**。12 台合作伙伴机器人的一项仿真对比中，有 10 台在 fine-tuning 后语言跟随得分提高；例如 Trossen 的归一化语言跟随分数从 71% 提升至 88%，Maker Arm 从 56% 提升至 75%。它们是特定仿真评分，**不是绝对成功率**。

这些数字是**示例所使用的数据时长**，并非“任何新机器人 5 分钟必然适配”的样本复杂度保证。官方没有披露每项适配的全部优化步骤、更新参数、训练算力、方差与安全验证细节；“one-shot”在这里也不应理解成严格只有一条示教轨迹。

6.9 尚未公开与现阶段不适用之处

Dream-1 的完整主干、参数规模、视频预测损失、动作 decoder、去噪策略、训练机器人/视频数据配比、zero-shot 是否依赖显式机器人描述、FaaS 更新范围等，在该官方技术文章中**未给出可重复实现的细节**。即使它延续 WAM 的联合视觉—动作理念，也不能断言它直接复制了 DreamZero 的 Wan2.1-14B 与 DreamZero-Flash。

当前官方主要将产品定位为**双臂平行夹爪操作策略**。

7. DreamZero 与 Dream-1：技术演进与不可混淆的结论

维度	DreamZero (论文)	Dream-1 (官方产品介绍)
主要问题	视频先验怎样成为动作策略	未见双臂机器人如何直接控制
公开架构	Wan2.1-I2V 14B、因果 DiT、联合 FM	WAM 思路与 EEf/IK 接口；内部结构未全面公开
动作接口	论文主要使用相对关节位置	可输出共享 EEf pose + 夹爪开合
训练证据	数据、损失、优化步骤、消融较充分	训练配方和消融披露有限
跨本体主张	30 min YAM 数据后适配并泛化	新双臂平行夹爪 零该本体数据 直接控制
高效推理	论文报告两张 GB200 下 150 ms/块	公布云 API 价格和每次动作覆盖 1.6 s；未披露同口径延迟
适用边界	大视频模型计算开销及长时/接触不确定性	IK、工作空间、路径可行性与手型限制
可复现程度	权重/训练代码部分公开	产品/API 可试用，完整模型训练尚不可复现

关键结论。 DreamZero 证明了“基于生成式世界先验的联合动作模型”可在特定条件下具备真机闭环能力；Dream-1 进一步展示了“共享末端动作接口 + 时序视觉先验”在多种未见双臂夹爪本体上的迁移潜力。这不是“WAM 一定优于 VLA”的普适定理，也不代表图像生成准确即可保证精密操作。

8. VLA 与 WAM 的系统对比：能力、算力与部署

本节把前面两种 WAM 案例放回具身基础模型的技术选型中。**不能把“VLA = 单步廉价推理，WAM = 多步视频生成”当作定义：**OpenVLA 采用离散动作 token， π_0 、 $\pi_{0.5}$ 和 SmolVLA 本身也可通过 Flow Matching 多步生成动作；Fast-WAM 则在训练时建模未来、部署时不显式生成未来视频。真正决定计算成本的是**参数与激活规模、视觉/视频 token、去噪步数，以及推理图中是否保留未来预测。**

8.1 对照的首先是学习目标，而非模型名称

维度	典型直接动作 VLA	典型 WAM (含联合/高效变体)
训练信号	图像/语言/本体状态到动作；BC、动作 Flow Matching 或动作 token 监督	额外施加视频/潜状态未来监督；可与动作联合训练
预训练先验	VLM 的视觉与语义知识，加机器人示教	视频模型的运动时序先验，加机器人动作对齐
执行输出	离散动作或连续动作块	最终仍需可执行动作块；视频只是中间量或辅助目标
推理路径	视觉编码 + 动作解码/采样	显式想象、联合视频—动作、或训练时预测/测试时直接动作
泛化优势假设	语言与物体语义、已有动作分布内迁移	动作造成的未来变化、陌生运动模式与时序结构
典型失败源	动作数据稀疏、分布外运动、感知偏差	加上视频幻觉、计算延迟、接触预测不可信

不能把这张表理解成“VLA 必须微调、WAM 不需要微调”。VLA 也可以使用共享末端控制接口（或语言形式的跨本体动作表示），**LAP** 已展示相应零样本跨本体证据。WAM 的关键可检验新增因素是**未来观测监督是否提升了任务与本体分布偏移下的成功率**，而不只是是否有 IK。

表中是架构层面的**常见趋势，不是逐模型性能定理**。是否零样本跨本体主要取决于动作空间、训练本体覆盖和控制接口；Dream-1 的共享 EEf/IK 思路不意味着所有 WAM 都天然跨本体，也不意味着 VLA 不能使用共享末端动作表示。

8.2 为什么视频分支往往更贵：一个可核算的成本模型

部署峰值显存不能只看参数量。以十进制 GB 表达，给出**权重的理论下界**与完整显存账本：

$$M_{\text{weights}}(\text{GB}) \simeq \frac{Pb}{8 \times 10^9}, \quad M_{\text{peak}} = M_{\text{weights}} + M_{\text{activations}} + M_{\text{KV}} + M_{\text{video}} + M_{\text{workspace}}.$$

P 为驻留参数数目， b 为每参数位宽。BF16 权重约为 2 GB/十亿参数；这是**加载权重的下界**，不是显卡容量建议值。量化还有 scales、zero-points、算子支持和运行时缓冲开销。对 Transformer，注意力矩阵的朴素计算量随 token 数近似二次增长；即使用 FlashAttention 减少显存访问，额外视频 token 仍会增加主干计算：

$$C_{\text{attention}} \propto LN^2d, \quad N_{\text{video}} \propto F \frac{H_{\text{img}}W_{\text{img}}}{s_H s_W}.$$

这里 L 是层数， N 是注意力 token 数， F 是未来帧数， s_H, s_W 是空间压缩/切块系数。式子是**量级解释**，不替代 DiT 完整 FLOPs、实际 VAE 压缩率和 FlashAttention 内核基准。

公开模型示例	参数规模	仅 BF16 权重下界	应怎样理解
SmolVLA	约 0.45B	约 0.9 GB	轻量直接动作 VLA；实际运行仍需视觉激活和缓存
π_0	约 3.3B	约 6.6 GB	VLM 约 3B + 动作专家约 0.3B
OpenVLA	约 7B	约 14 GB	7B 离散动作 VLA；训练远比权重下界贵
Fast-WAM	视频骨干约 5B + 动作专家约 1B	两部分约 12 GB	尚未计入其他编码器、缓存和工作区
DreamZero	视频骨干约 14B	约 28 GB	仅 BF16 核心权重已超过单张 24 GB GPU
Dream-1	未披露	无法可靠估算	云 API 标价不等于 GPU 计算成本

上述模型大小分别见 SmolVLA、 π_0 、OpenVLA、Fast-WAM、DreamZero 和 Dream-1。参数量不同、精度不同、分辨率不同、动作块不同，不能用表中的 GB 推出推理速度或成功率。

8.3 训练算力：从头预训练、全参微调与 LoRA 是三种预算

常见 BF16 + AdamW 全参训练，单纯参数/梯度/优化器就约有 $2 + 2 + 8 = 12$ 字节/参数；若另存 FP32 master 权重可达约 16 字节/参数，**还未计激活、反向图、视频 latent 和临时缓冲**。因此对 14B 参数，全参训练时这些张量粗略对应 **168–224 GB 的聚合内存下界**，并不表示有同等显存便能训练。梯度检查点、ZeRO/FSDP、CPU offload、混合精度和视频 token 数都会改写实际峰值。

OpenVLA 官方报告其约 970k 条机器人轨迹的原始预训练使用 **64 张 A100、15 天**；openpi 的单卡参考需求为推理 > 8 GB、LoRA 微调 > 22.5 GB、全参微调 > 70 GB（这些是项目经验阈值，不保证任意配置都刚好够用）。DreamZero 论文给出约 500 小时 AgiBot 数据、100k 更新步和大规模视频基础模型的联合训练，但不同论文的总预训练语料、分辨率、冻结模块、集群吞吐与 GPU 利用率不一致，**不能据此计算出 WAM 相对 VLA 的固定“几倍训练成本”**。

经验上，额外视频预测常使 I/O、GPU 激活与梯度开销更高，却允许利用缺少机器人动作标签的人类/网络视频。反过来，只有动作监督的 VLA 适配流程更容易在消费级 GPU 上快速迭代。LoRA 只减少可训练权重与优化器状态，**不消除冻结骨干的前向驻留、反向传播所需激活和视觉预处理**。

8.4 推理时延：比较真实决策周期，不比较“一个动作块有多长”

定义视觉编码时间 T_e 、动作生成步数 N_a 、视频联合采样步数 N_j 、对应单步时间 T_a, T_j 。忽略流水并行时，可以写出三种教学近似：

$$\begin{aligned} T_{VLA} &\approx T_e + N_a T_a + T_{IO}, \\ T_{WAM, joint} &\approx T_e + N_j T_j + T_{IO}, \\ T_{WAM, action \text{ only}} &\approx T_e + T_{\text{world features}} + N_a T_a + T_{IO}. \end{aligned}$$

对于离散自回归动作 VLA, $N_a T_a$ 应理解为生成动作 token 的累计时间；对于 Flow Matching 才是动作积分/去噪步数。KV Cache、视频—动作共享计算和异步流水线会改变实际时延；最关键的是**传感、编解码、网络、模型、控制派发全链路** $p50/p95/p99$ 。

有来源的测量	报告值与条件	可得出什么、不能得出什么
DreamZero-Flash	约 150 ms/块, 两张 GB200 ; 约 7 Hz 高层闭环	多卡强硬件上的优化结果; 不能迁移为 4090 指标
Fast-WAM (论文)	约 190 ms, 单张 RTX 5090D V2 32GB	不生成完整未来视频可显著减时延
Fast-WAM (代码库较新优化)	RTX 4090 约 110 ms (官方特定 LIBERO 推理配置)	是仓库实现报告, 不能直接当作跨任务端到端保证
Dream-1 API	每次返回约 1.6 秒覆盖长度的动作 ; 推理 $p95$ 未公开	动作覆盖时长不是 API 调用时延; 无法核算本地 GPU 成本

参见 DreamZero 项目、Fast-WAM 论文、Fast-WAM 实现 与 Dream-1 官方页面。动作块能以 30 Hz 或 50 Hz 被底层执行, 不代表大模型每 20–33 ms 都重新观测并纠错。**闭环决策频率、动作执行频率、动作块覆盖时长**是三个不同指标。

8.5 精度、泛化与速度不是单调关系

Fast-WAM 发现, 在其评测设置中, **训练时视频共训练**比部署时显式视频生成更影响结果; 因此只保留动作通路可以减少计算。不过, **Faster-WAM** 在其统一实验配方与 LIBERO-Plus 分布偏移评测下报告: Fast-WAM 49.14%, 引入稀疏未来条件的 Faster-WAM 73.57%, 后者还比其 Joint-WAM 基线快约 2.21 倍。这表明**推理时保留压缩的未来表征, 可能为长尾扰动与分布外环境带来值得付费的鲁棒性**; 数字只对各自实验协议成立。

因此, 技术路线不应只有“VLA 或完整视频生成 WAM”二选一, 而是形成连续谱: **直接 VLA → 训练时世界监督的动作策略 → 单次未来特征/稀疏未来条件 → 完整联合视频—动作采样**。需要按任务的分布偏移、接触容错与 GPU 预算选择工作点。

8.6 面向 24 GB GPU 和接触操作的工程选型

若资源是单张 RTX 3090/4090 (24 GB), 优先考虑 SmolVLA、 $\pi_0/\pi_{0.5}$ 推理、OpenVLA 的合适精度实现, 或使用 LoRA/部分冻结进行 VLA 任务适配; 但 openpi 参考的 LoRA > 22.5 GB 已很接近 24 GB 上限, 仍需实际测试 batch、相机数与峰值显存。**14B DreamZero 的 BF16 核心权重理论上需要约 28 GB, 单卡 24 GB 无法原样驻留**; 量化、模型切分、多卡并行可以尝试, 但不自动保留论文在双 GB200 上报告的闭环频率。四张 24 GB 卡也不是一张可直接寻址的 96 GB 显卡, 跨卡通信与模型划分需要另行设计。

以机械臂 + 头/腕 RGB-D 的精密操作为例, 建议把大模型作为中低频技能/目标生成器, 而将 **关节控制、IK、碰撞检测、触觉/电流接触监测和紧急停止**保留在本地高频闭环。若需利用人类视频或未来预测知识, 可优先评估“**WAM 视频监督训练 + 测试时动作直出**”或“**单次未来 latent 条件**”; 只有经消融证明想象视频在失败恢复/陌生接触上显著改善性能, 才承担多步视频生成的持续 GPU 成本。

公平比较应固定相机视角与分辨率、机器人与动作表示、训练示教数据量、任务与初始状态、动作块时长、量化精度和闭环更新协议, 同时报告: **全链路 $p50/p95/p99$ 、峰值显存、GPU·秒/动作块、真实任务成功率、分布外掉落/扰动恢复率以及安全停机次数**。不要只展示视频质量或单次平均推理延迟。

9. 高效推理：把模型时钟接到机器人时钟

机器人部署至少有三个频率：相机采样 f_s ，模型决策 f_m ，低层控制 f_c 。动作块长度为 H ，控制周期 Δt_c ，模型一次计算时间 T_m ，网络/调度延迟 T_n 。有时模型决策频率 f_m 显著低于 f_c ，但只要动作块连续交接，机器人仍能平滑执行；这不意味着扰动后能像高频反馈控制那样快速纠偏。

方法	主要减少什么	必须验证什么
因果 KV Cache	重算历史前缀注意力	真实观测如何更新、何时失效
去噪步间 DiT 缓存	重复的去噪骨干计算	缓存误差是否损伤动作
FlashAttention	注意力内存访问与临时张量	端到端收益，不只 kernel 时间
减少去噪步/蒸馏	模型函数评估次数	接触质量、长尾失败
动作块异步执行	模型阻塞控制器时间	时间戳、旧块接管、超时回退
量化/多卡并行	GPU 时延及显存预算	精度、通讯瓶颈、尾部延迟

不要混淆两种 Cache：因果 KV Cache 存放历史条件的注意力状态；DiT 缓存重用某些去噪时间步的中间计算。两者是否有效取决于结构与缓存失效逻辑。有关通用 FlashAttention、KV Cache 和算子层推导，转至本课程《具身基础模型的高效训练与实时推理》。

10. 接触操作与失败恢复

即使预测视频和执行动作视觉上吻合，纯 RGB 仍难以确定法向力、静摩擦裕度、物体是否真正被夹住。操作时建议同时建模可观测本体状态、电机电流/触觉或其置信度。一个任务级抽象为

$$z_t = (T_{\text{obj}}, T_{\text{hand}}, q_t, \hat{c}_t, \Sigma_t),$$

$\hat{c}_t$ 为估计接触模式， Σ_t 表示不确定性；若没有触觉，不能把视觉模型想象出来的“接触力”当作真实测量。

以“桌面任意位姿拾起螺丝刀、移动到作业位姿并旋转”为例：世界分支提供物体运动与遮挡先验，动作分支输出手臂目标，低层电机控制闭环维持抓持；检测到电流残差、滑移或位姿偏离后，必须允许重观测、停机和恢复，而不是持续执行旧动作块。

11. 如何评价 WAM 的真实研究贡献？

评价对象	有意义的指标	常见误判
世界预测	接触/运动一致性、不确定性校准	只看视频画质
策略能力	成功率、任务进度、跨对象/任务泛化	把进度当完成率
跨本体	本体训练数据量、全体机器人得分分布	只展示高光样例
闭环鲁棒性	扰动后恢复、遮挡、掉落重抓	开环演示替代反馈测试
系统可用性	端到端 p50/p95/p99、网络波动	用 action chunk 时长冒充推理延迟
安全性	速度/接触限幅、碰撞、超时回退	用生成视频替代安全控制

应分别验证视频预训练、联合训练、跨本体共享动作表示、实际闭环更新及 few-shot 适配各自的效果；不要只比较两个不相同数据、平台与任务口径的最终数字。**建议使用受控消融矩阵：**固定机器人训练/测试集合、相机视角、语言指令、示教时长、EEF/IK 接口和高低层控制频率；只改变“动作监督 VLA / 联合未来预测 WAM”，再单独加入或移除头部相机与未来预测损失，检验每项机制对未见任务、未见本体成功率和恢复能力的净贡献。

12. 与现有课程的边界与阅读顺序

本讲专注 WAM 从时序监督到可执行动作、从 DreamZero 的联合生成到 Dream-1 的跨本体接口。Flow Matching 数学细节参见同名课程；BC、DAgger、ACT、RTC 在模仿学习章节；强化学习后训练参见 RL Post-training；GR00T/SONIC 着重 System 1–System 0；Robot Agent 负责外层任务规划、验证和重规划。**DreamZero 与 Dream-1 首先是动作策略/基础模型，不应直接等同于带工具编排的 Robot Agent。**

推荐顺序：概念与四范式 → 数据对齐 → DreamZero 联合 FM/KV/Flash → Dream-1 EEf/IK 与 zero-shot → VLA/WAM 显存与算力对比 → 延迟、安全和接触限制。理解 Dream-1 之后，应重点追问“视觉预测的物理先验可以在多大程度上跨越末端器具差异、接触模型与执行接口差异”，而不只追求更长的生成视频。

13. 论文与官方资料

1. Zhang, Zeng, Zhang. *From World Models to World Action Models: A Concise Tutorial for Robotics*, 2026: 总体概念与研究范式。
2. Ye et al. *World Action Models are Zero-shot Policies (DreamZero)*, 2026: 14B 联合生成、训练公式、实时优化和迁移实验。
3. NVIDIA GEAR. *DreamZero* 项目主页 与 代码/模型仓库: 可复现实现与视频案例。
4. Dream Labs. *Zero-Shot to New Robots: Introducing Dream-1*, 2026-10: EEf/IK 迁移、API、未见机器人和 FaaS 案例。非完整技术论文。
5. Yuan et al. *Fast-WAM: Do World Action Models Need Test-time Future Imagination?*, 2026: 训练世界预测与部署时未来想象之间的取舍。
6. Chi et al. *Universal Manipulation Interface*, 2024: 同款夹爪与腕部视角的跨平台接口。
7. NVIDIA. *Cosmos*; Physical Intelligence. *openpi*: 世界生成与机器人动作专家的工程入口。
8. Lipman et al. *Flow Matching for Generative Modeling*, 2023: 速度场目标的数学基础。
9. Kim et al. *OpenVLA: An Open-Source Vision-Language-Action Model*, 2024; 项目主页: 7B VLA 与 64 张 A100 的预训练案例。
10. Black et al. π_0 : *A Vision-Language-Action Flow Model for General Robot Control*; *openpi* 项目: 3.3B 参数构成及显存需求的工程参考。
11. Hugging Face. *SmolVLA*, 2025: 450M 小模型与消费级 GPU 部署。
12. Yuan et al. *Fast-WAM* 论文 与 官方代码: 视频共训练、纯动作推理与优化版本的实测延迟。
13. Zhao et al. *Faster-WAM*, 2026: 稀疏未来条件与 LIBERO-Plus 分布外对照。
14. Zha et al. *LAP: Language-Action Pre-Training Enables Zero-shot Cross-Embodiment Transfer*, RSS 2026: 反例与对照, 说明 VLA 同样可能实现未见本体零样本迁移。

文档说明: 标注为“教学抽象”的公式不代表官方原始实现；本文不包含未执行的训练或真机实验声明。资料核对日期：2026-10-22。本文为教学材料，供公开资料学习与技术比较，不构成 Dream-1 的安全认证或性能保证。