

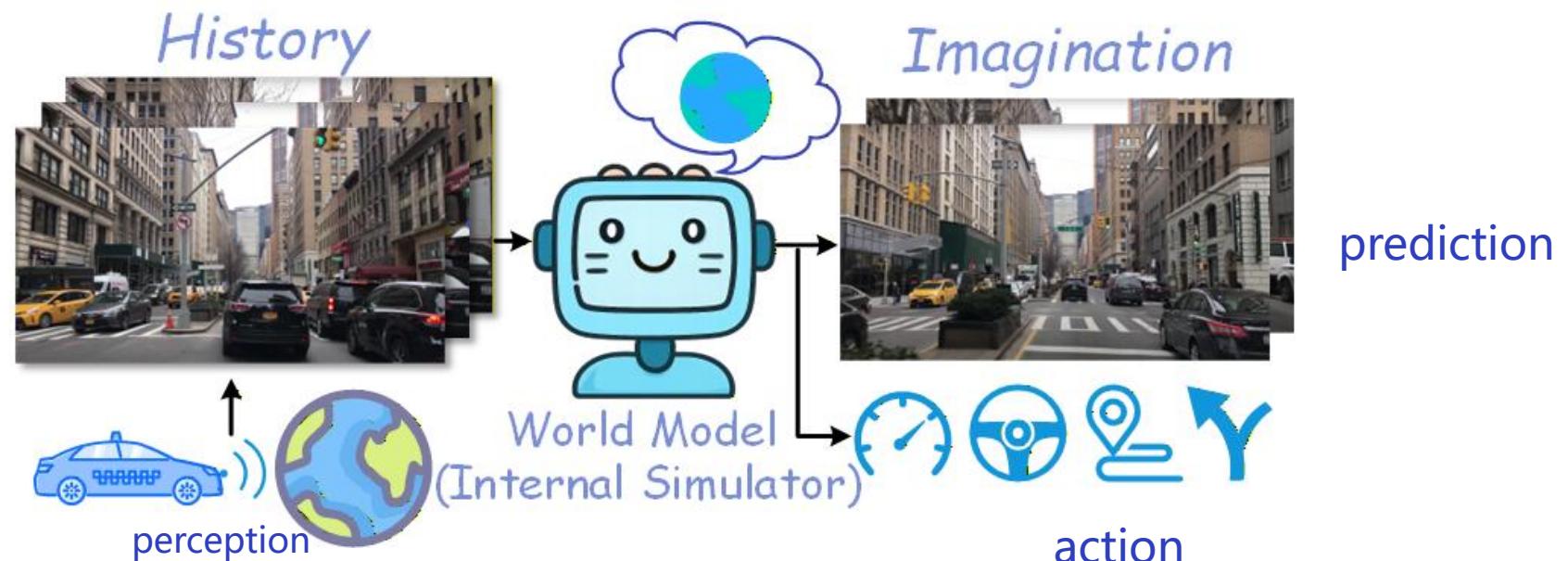
A Comprehensive Survey on World Models for Embodied AI

A three-axis taxonomy across functionality, temporality, and spatiality

李鑫庆

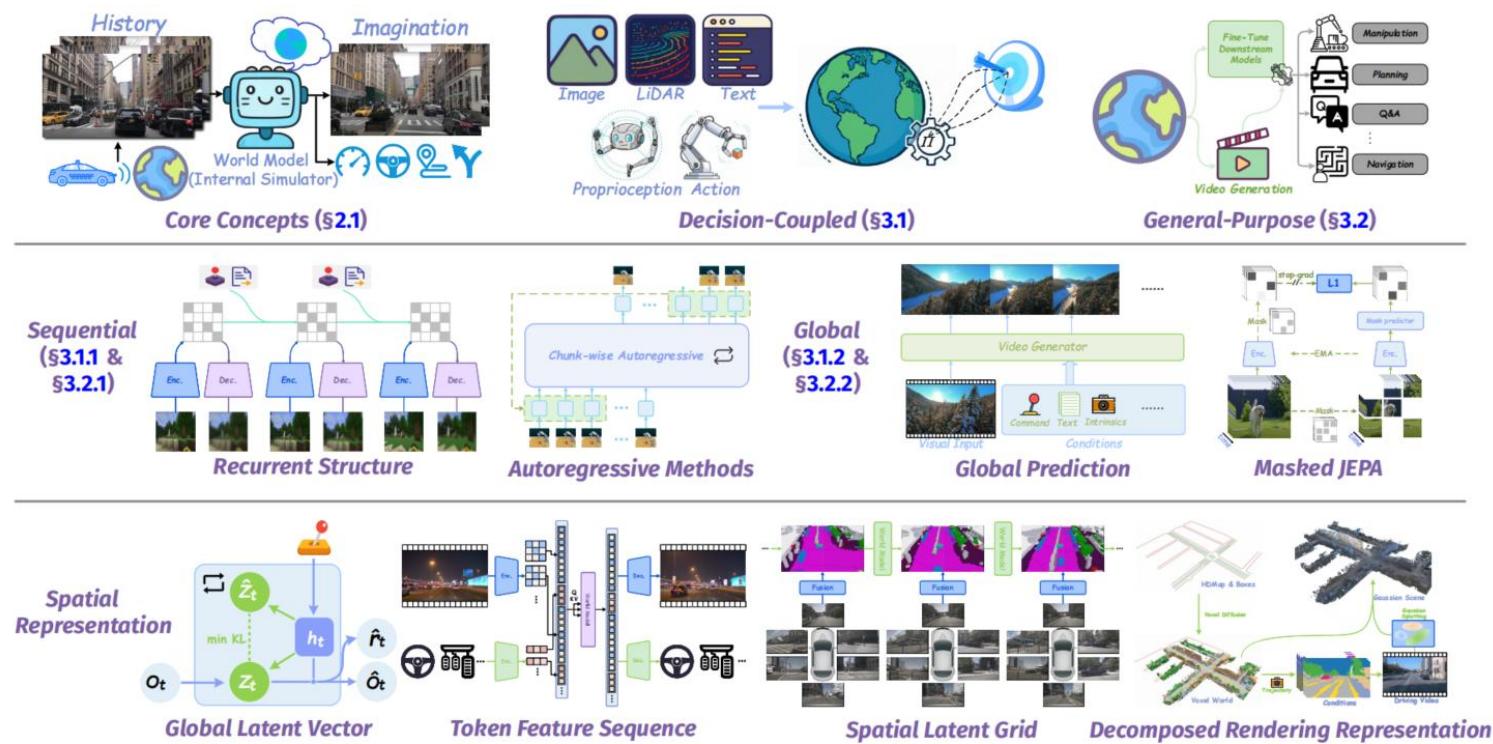
World Model:

- 核心：感知→行动→预测（具身智能的闭环）
- 起源：Model-based RL (Ha & Schmidhuber, 2018)，通过RNN学习环境动态实现内部模拟。
- 发展现状与问题：生成模型推动了多种架构的涌现，但缺乏统一梳理与系统分类。



三轴统一框架

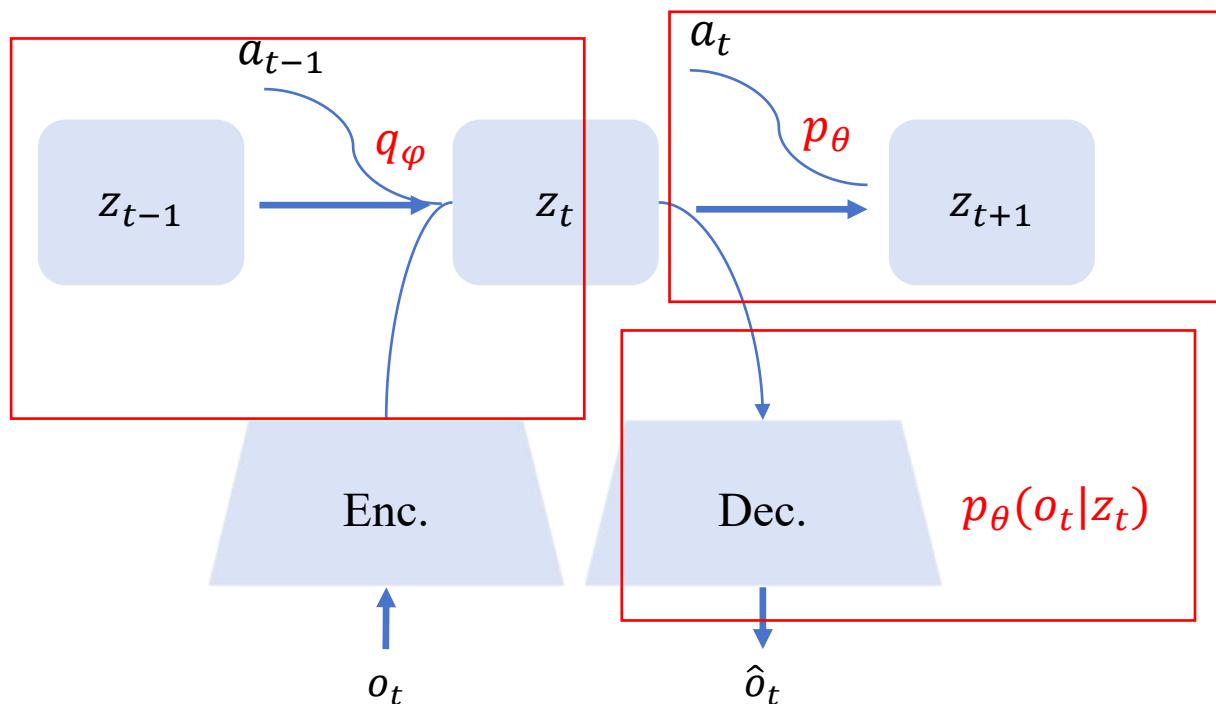
- 功能：决策耦合 \leftrightarrow 通用目的 —— 决定模型的优化目标
- 时序：顺序模拟 \leftrightarrow 全局预测 —— 决定时间建模范式
- 空间：全局潜变量、Token特征序列、空间潜在网格、分解式渲染 —— 决定世界状态的表征方式



数学框架与训练目标

- 任务(POMDP)：在部分可观测条件下学习隐状态 z_t 、动力学 $p_\theta(z_t|z_{t-1}, a_{t-1})$ 和观测生成 $p_\theta(o_t|z_t)$ 。
- 推断：用近似后验 $q_\phi(z_t|z_{t-1}, a_{t-1}, o_t)$ 估计潜在状态，“基于当前的观测与历史，对 z_t 的信念分布”。
- 目标(ELBO)：

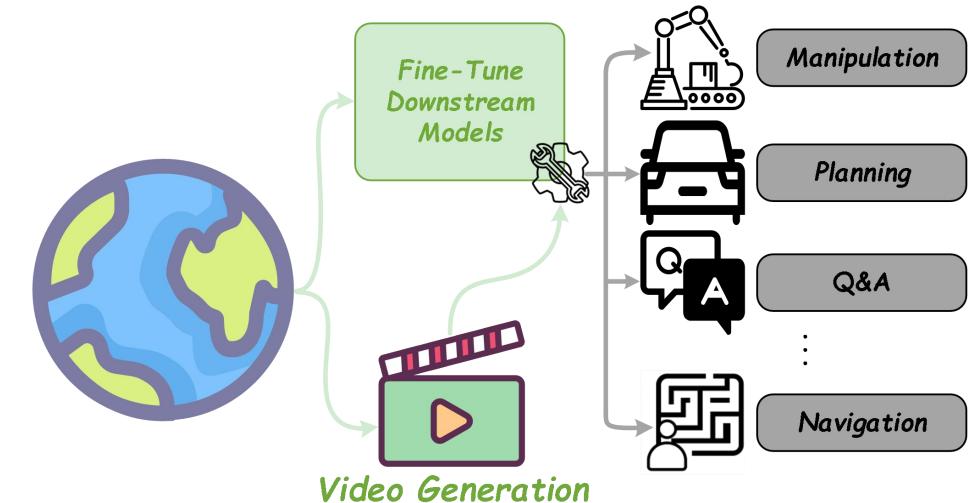
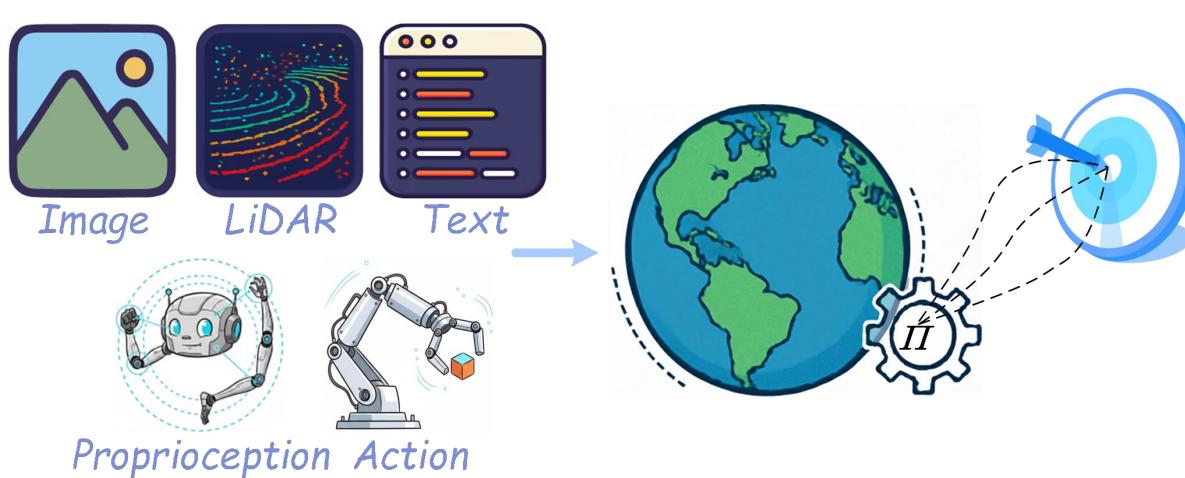
$$\mathcal{L}(\theta, \phi) = \underbrace{\sum_{t=1}^T \mathbb{E}_{q_\phi} [\log p_\theta(o_t | z_t)]}_{\text{重建}} - \underbrace{D_{\text{KL}}(q_\phi(z_t | z_{t-1}, a_{t-1}, o_t) \| p_\theta(z_t | z_{t-1}, a_{t-1}))}_{\text{正则}}$$



推理阶段没有观测 o_t ，需要依赖先验动力学 p_θ 滚动预测，需要利用KL散度对齐先验和后验。

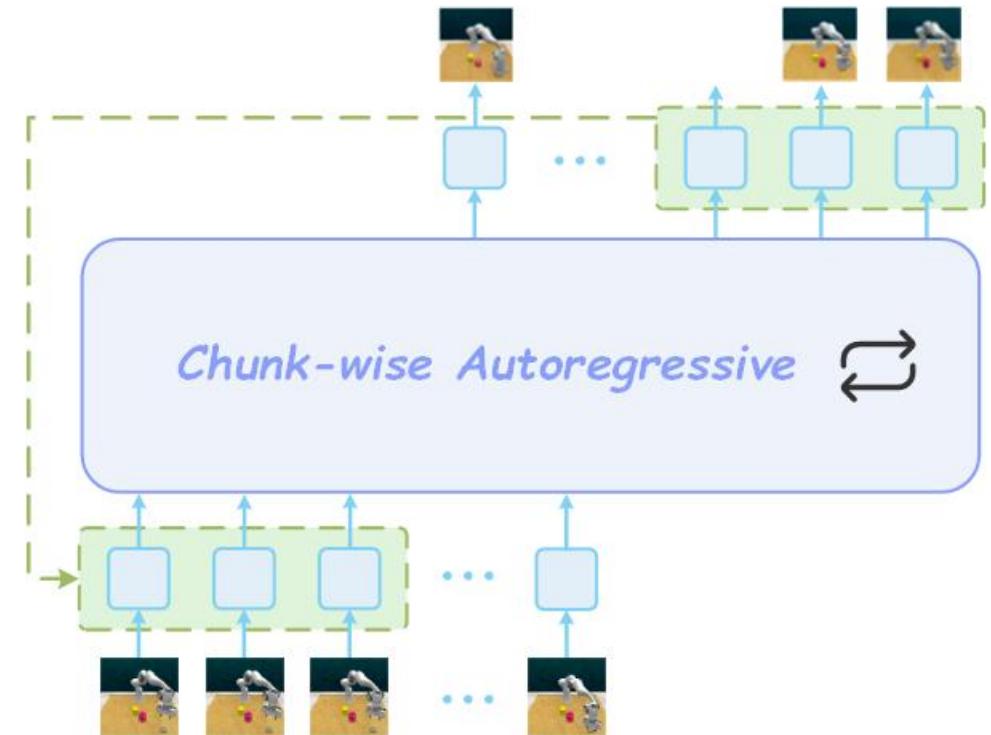
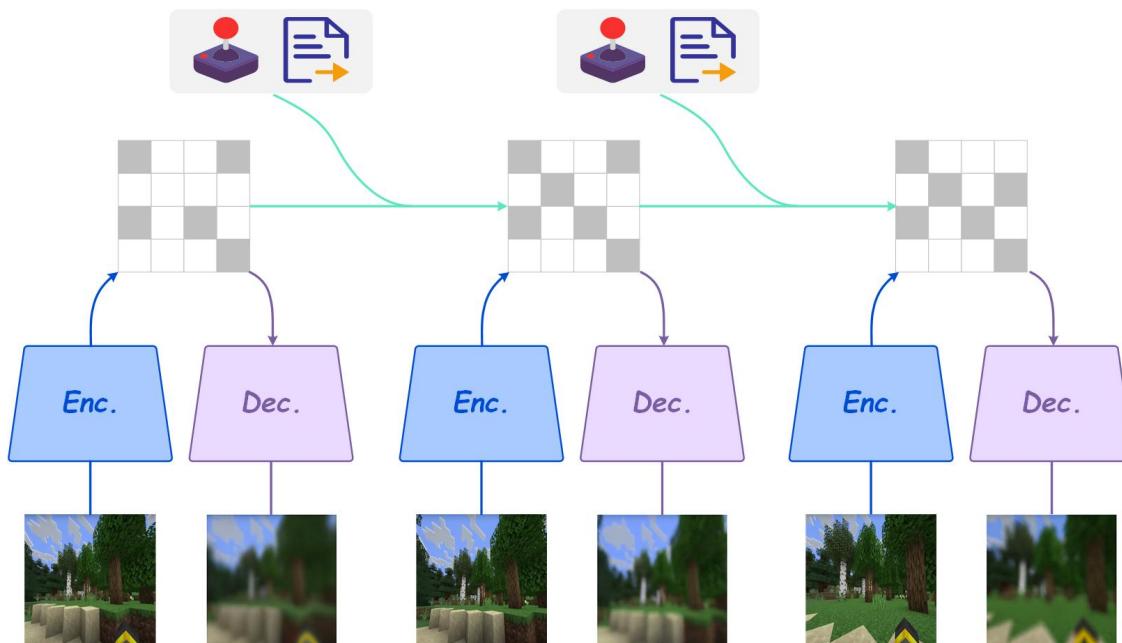
Decision-Coupled vs. General-Purpose

	Decision-Coupled	General-Purpose
定位	面向具体任务、紧耦合策略	任务无关的环境建模
优势	即插即用、任务表现强	跨任务可迁移、覆盖广
局限	表征黏任务、迁移弱	训练复杂、下游需适配
例子	MPC模型、任务特化WM	视频/通用WM、后续微调



Sequential

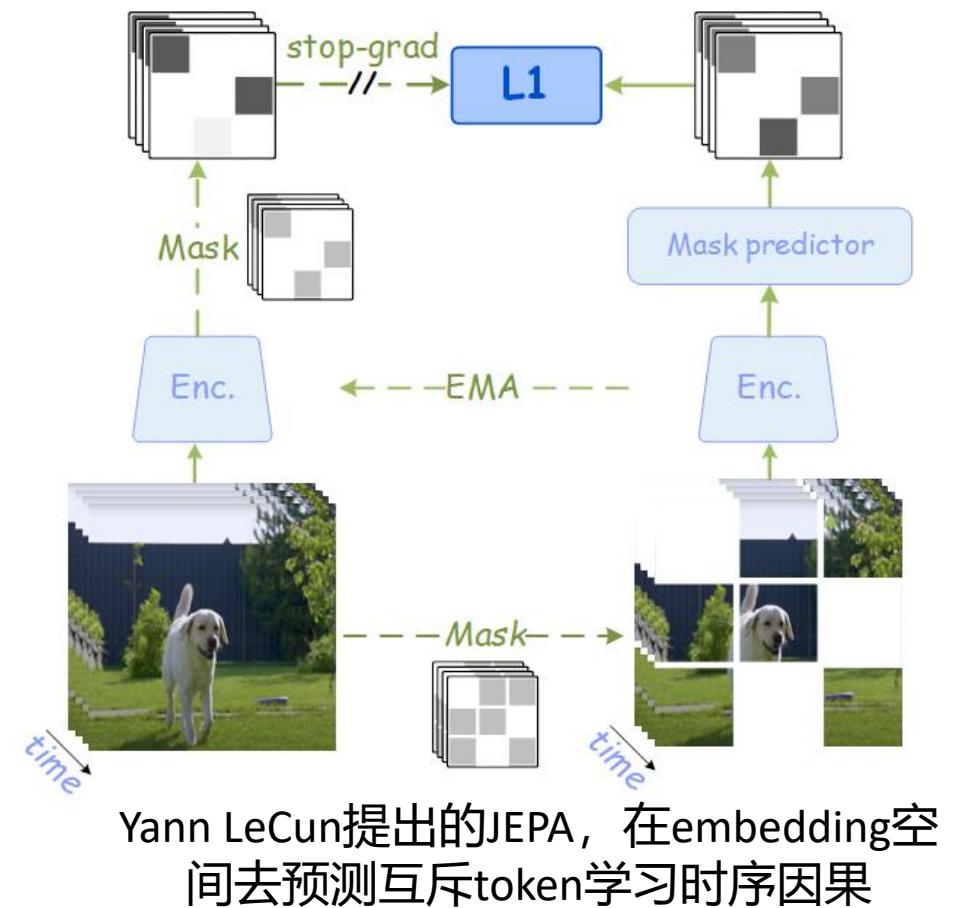
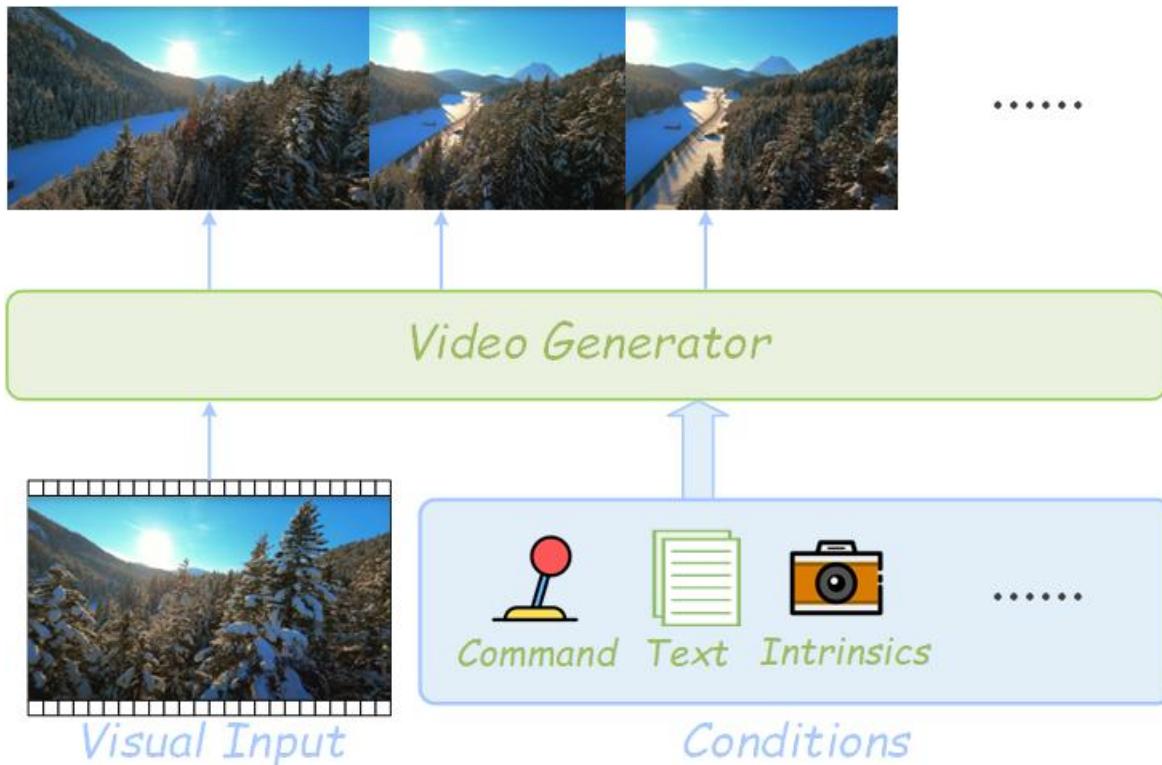
- 定义：按时间逐步（或分块）自回归地模拟未来演化。典型架构：RNN/SSM（如Mamba/自回归Transformer；也可配合目标分解/CoT做规划）。
- 优势：因果一致，适合闭环控制；接口简单，便于时序对齐。
- 局限性：长程误差累积；并行计算受限。



分块自回归来提高吞吐

定义：把整段未来作为目标一次性建模，常见做法：并行扩散/流式视频生成或掩码/JEPA特征预测。

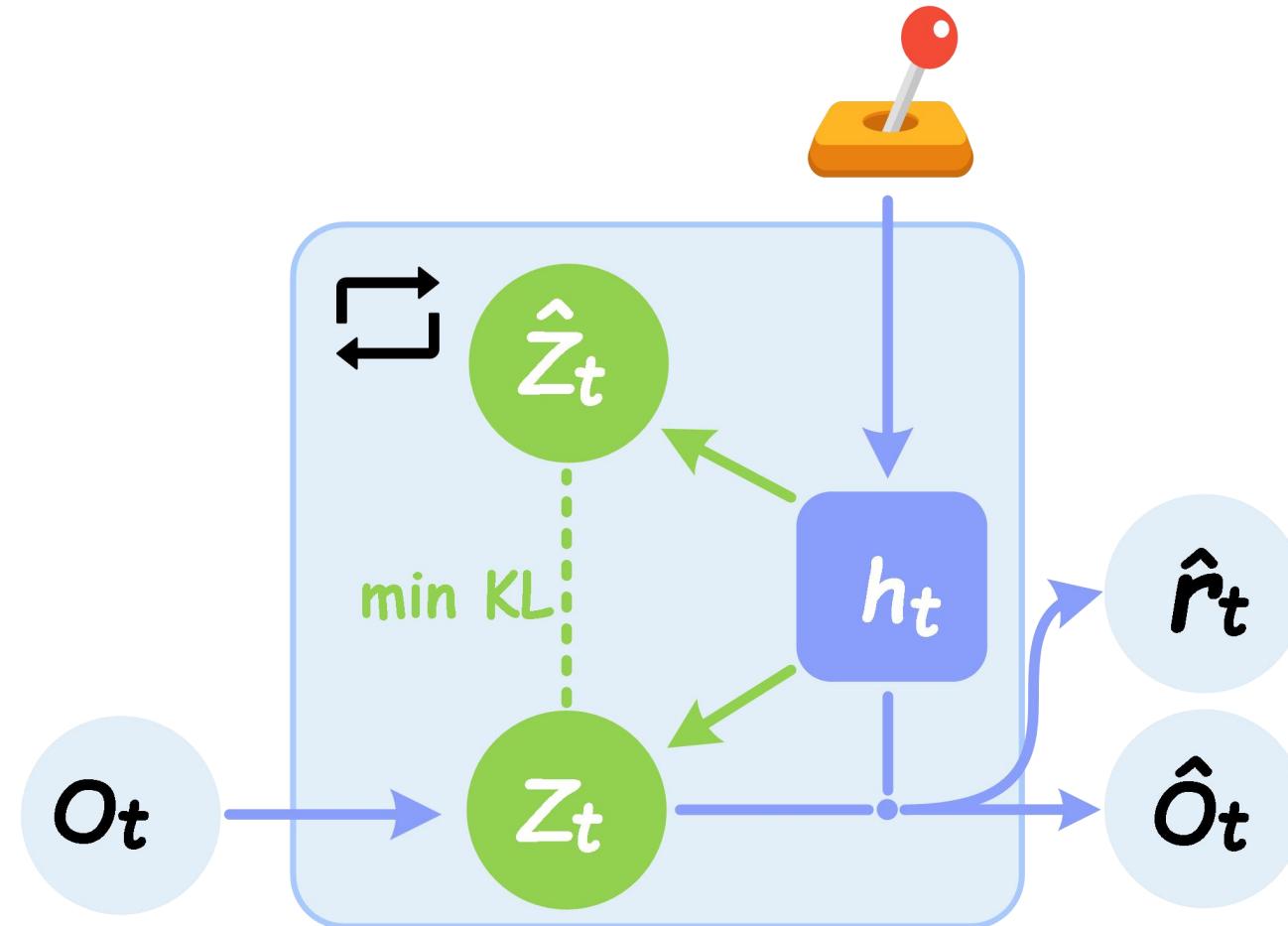
- **优势：**误差不易累积，平行效率高；易引入全局约束/条件。
- **局限性：**闭环交互性弱（缺动作反馈回路）；局部动力学细节不足。



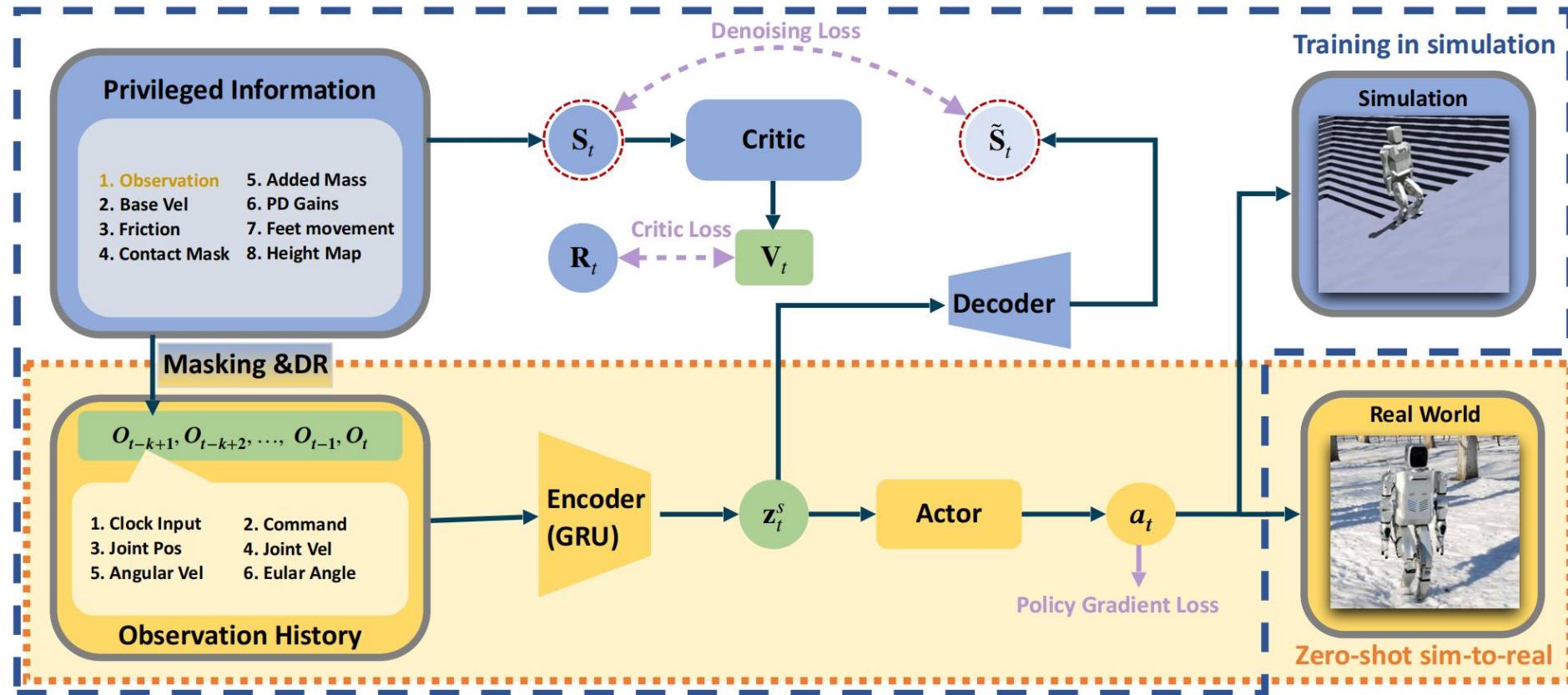
Global Latent Vector

定义：将场景/世界状态压缩为单个低维向量 z_t ，并在该紧凑表示上进行动力学建模。

- **优势：**计算与延迟友好、轻量化便于部署。
- **局限性：**细粒度时空信息损失。



Global Latent Vector

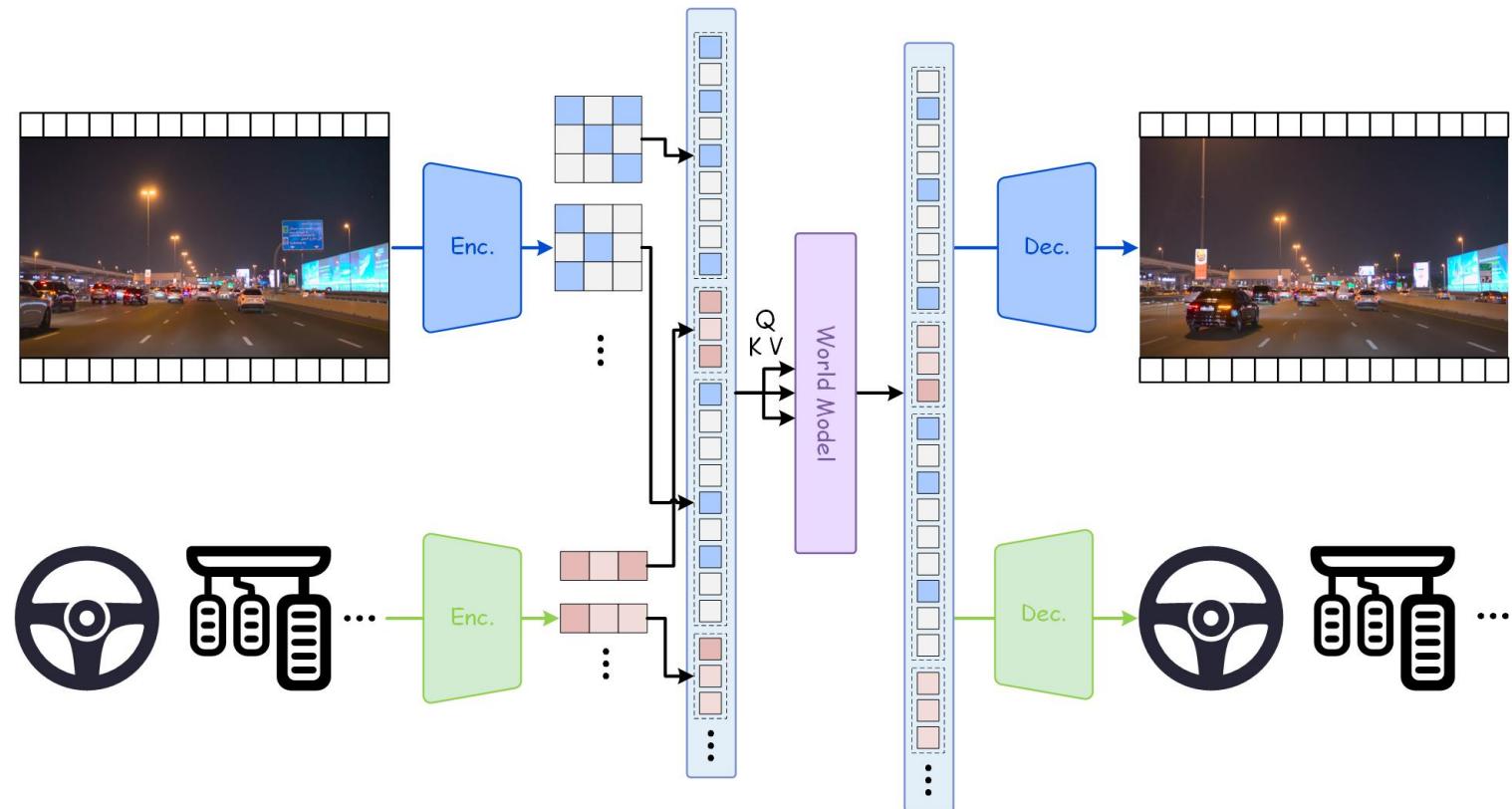


这类表征在工程部署中很常见。以RSS'24最佳论文DWL为例：训练期利用仿真中的特权信息训练critic预测价值 V_t ，以此评估并指导actor优化策略；部署期不再依赖特权信息，而是用观测历史经GRU推断潜状态 Z_t^s 作为短期记忆，由actor直接输出动作，实现从仿真到真实（S2R）迁移。

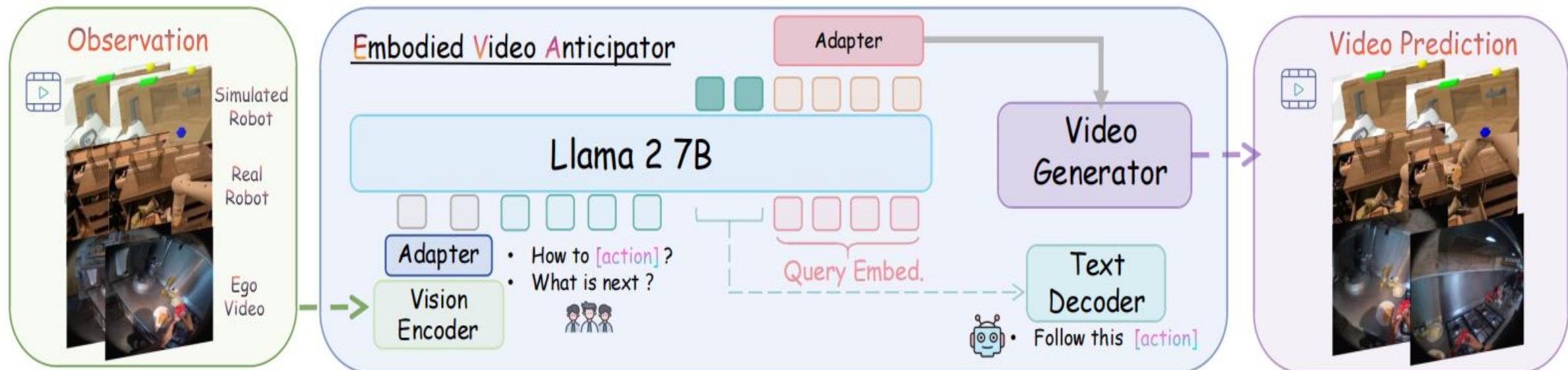
Token Feature Sequence

定义：将世界状态离散为一组token的特征序列，便于建模token间的依赖关系。

- **优势：**与注意力机制耦合，能细粒度地表示复杂场景和多模态信息。
- **局限性：**需要大量数据训练，且常依赖大参数量模型，推理开销高。



Token Feature Sequence

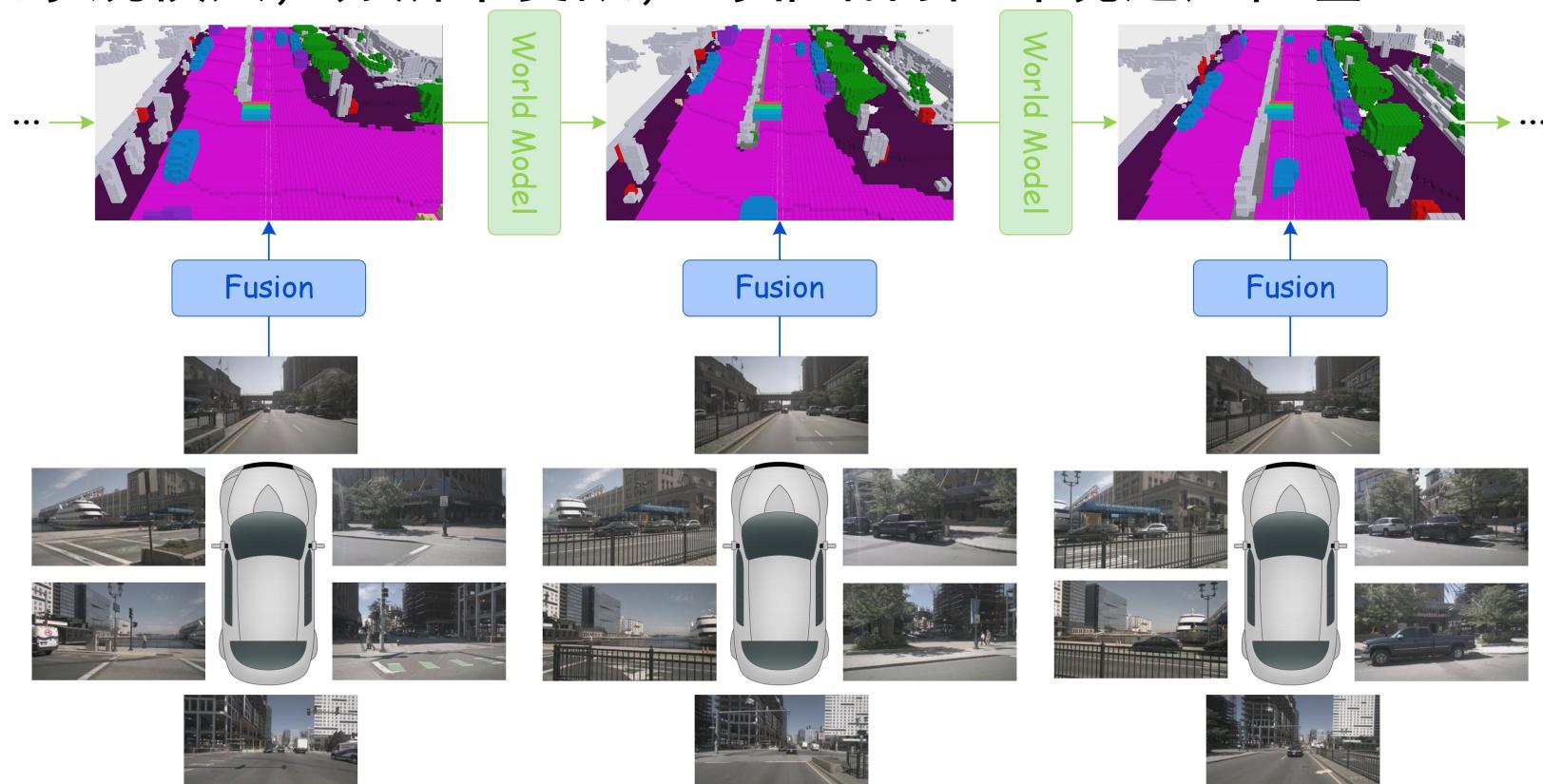


EVA (ICML'25)：该工作将仿真与真实机器人的第一人称视频作为输入，先经视觉编码压缩为visual tokens；用户文本指令被编码为查询。模型使用LLaMA进行多模态信息融合，在统一的token序列上完成推理，并通过双路解码分别生成动作文本与未来视频的预测。

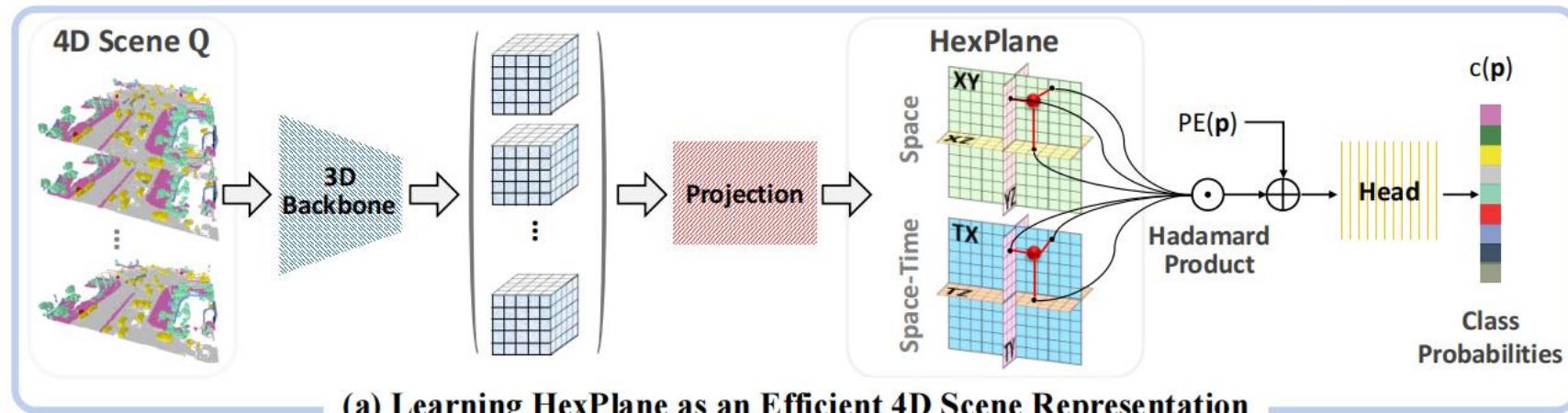
Spatial Latent Grid

定义：将空间先验信息注入场景或将场景编码到空间网格中，是自动驾驶领域的主流方式。

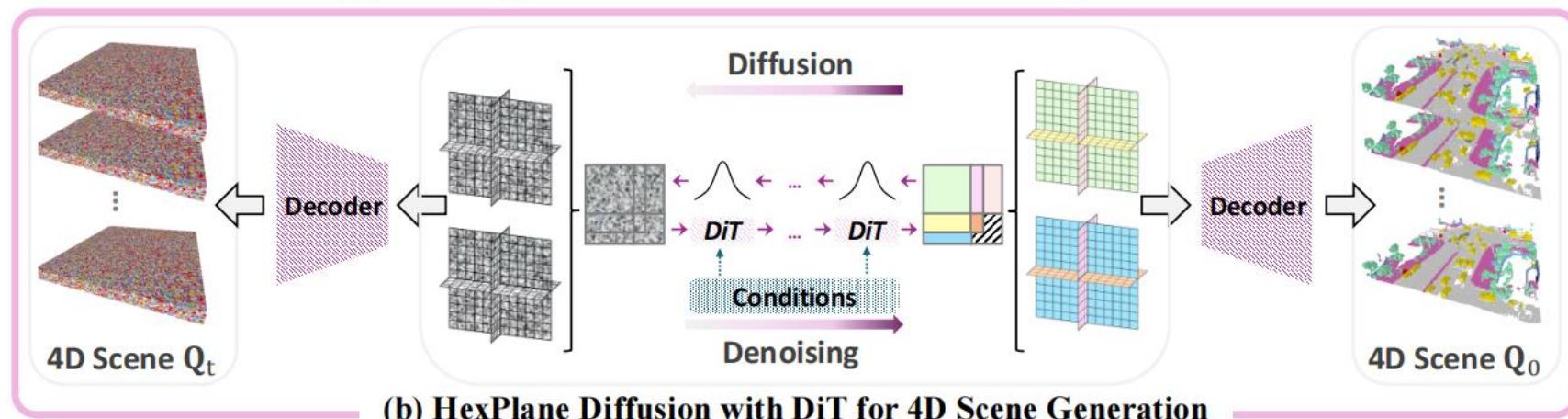
- **优势：**保留了空间局部拓扑，易于多视角融合和地图生成。
- **局限性：**表示规模大，分辨率受限，对非结构化环境适应性差。



Spatial Latent Grid



(a) Learning HexPlane as an Efficient 4D Scene Representation



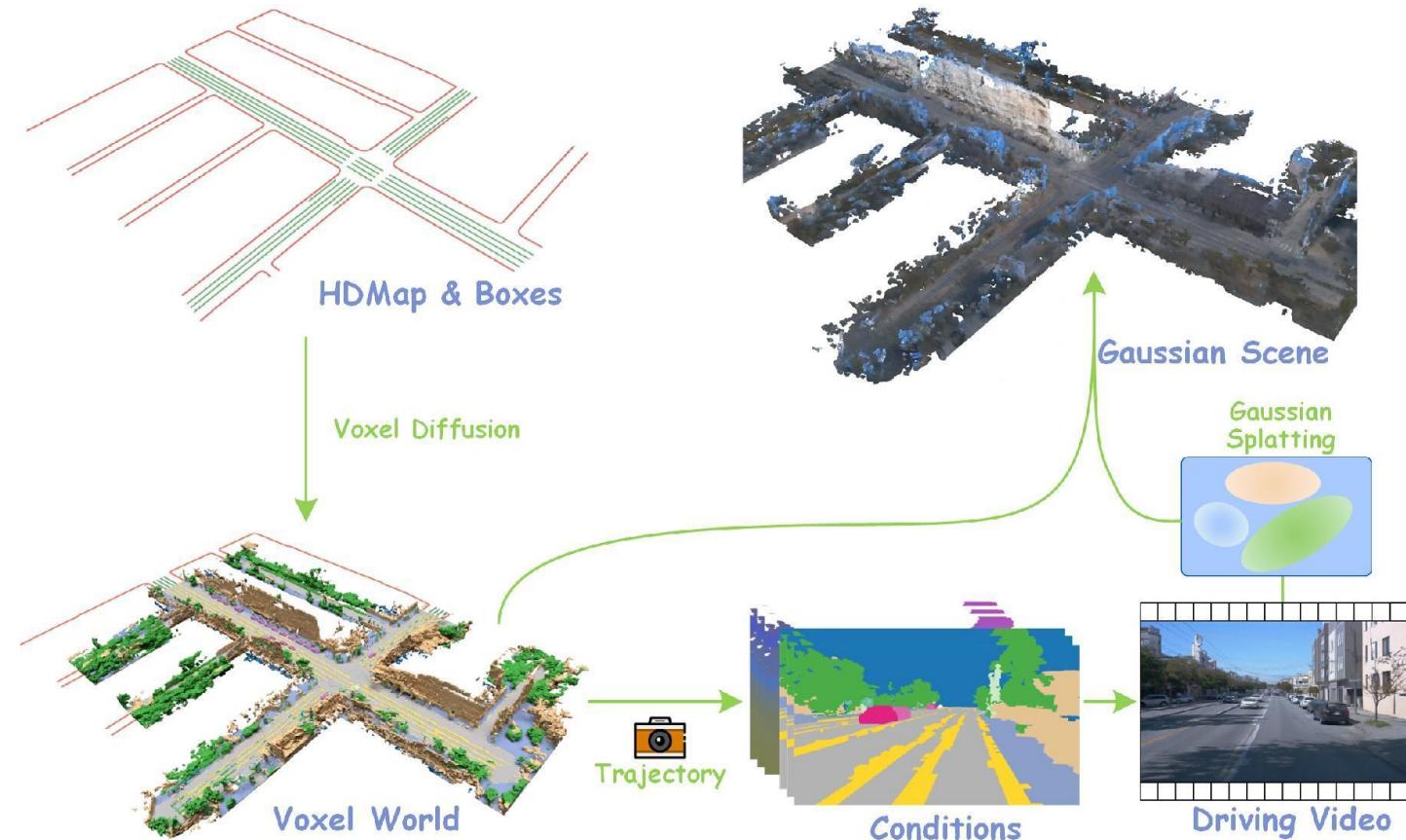
(b) HexPlane Diffusion with DiT for 4D Scene Generation

DynamicCity (ICLR'25 Spotlight) 将四维动态场景编码为六个二维 HexPlane (xy, yz, zx, tx, ty, tz)。针对每个查询位置，分别从六个平面提取特征，进行逐元素组合，并与位置编码融合，再送入轻量级预测头以输出占据/类别概率。在生成阶段，方法向六个子空间注入噪声，利用Diffusion Transformer去噪，随后由解码器重建完整场景。该设计在内存占用紧凑的同时保留空间拓扑，从而实现高效的 4D 推理与生成。

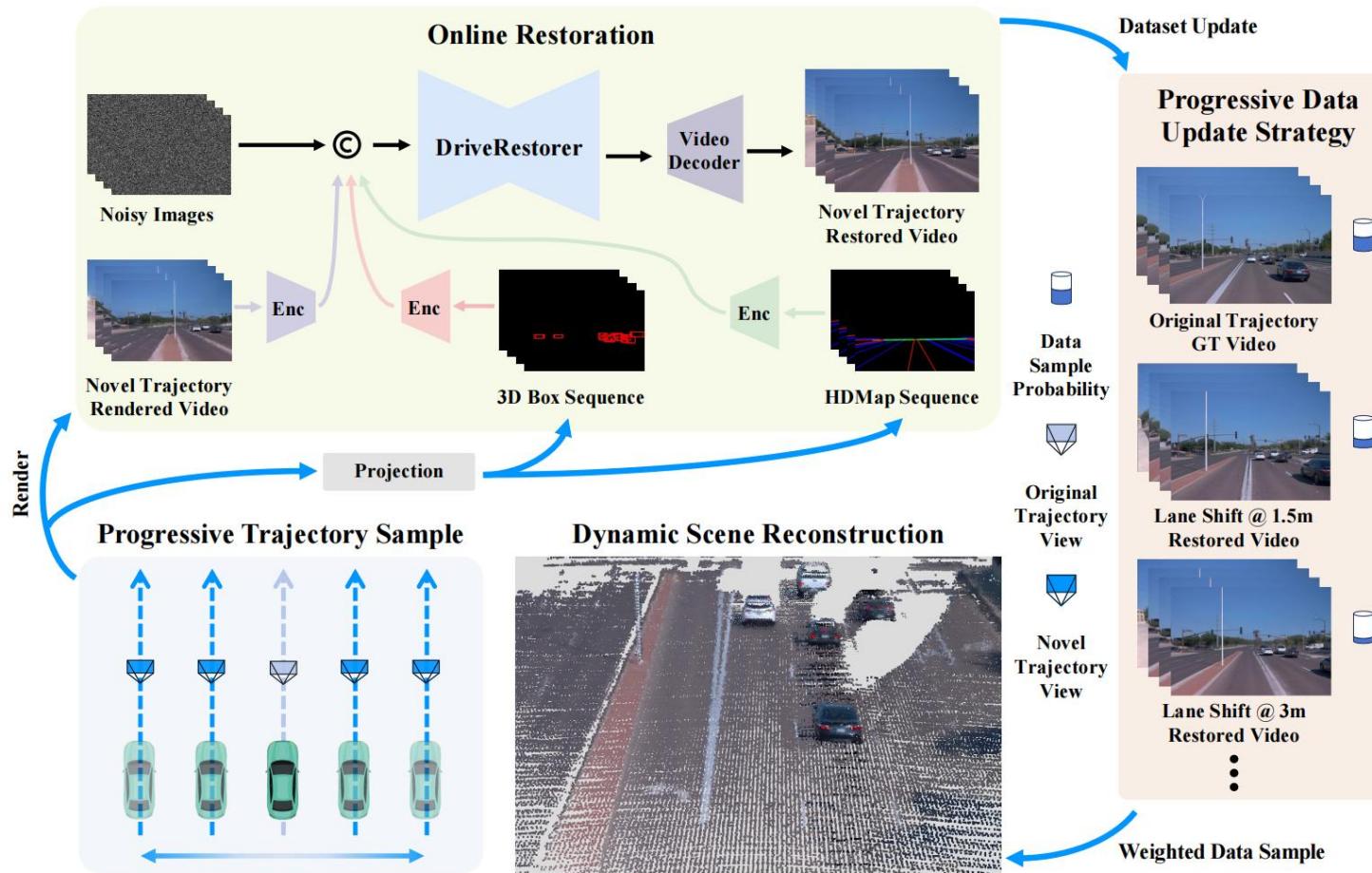
Decomposed Rendering Representation

定义：将场景拆解为一组可渲染基本要素，再通过渲染流程或生成模型构建观察空间。

- **优势：**几何一致&高保真；支持物体级别操作。
- **局限性：**训练开销大；动态/拓扑变化难实时更新。



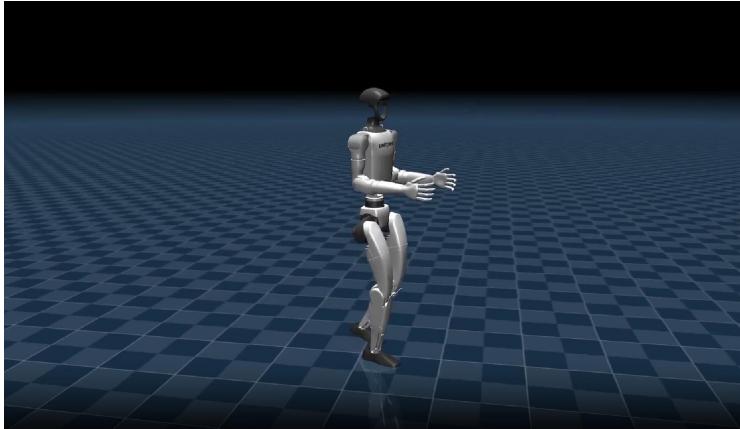
Decomposed Rendering Representation



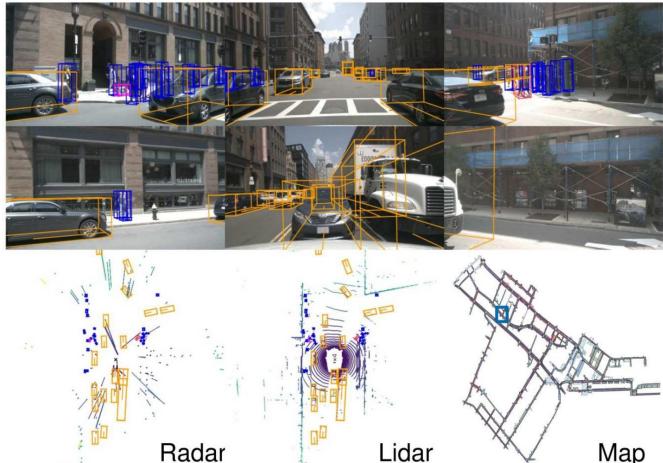
该工作属于分解渲染表征的代表案例CVPR'25的ReconDreamer。现有模型只在原始轨迹附近渲染可靠，一侧移就模糊/破碎。它用渐进式侧移采样生成“新轨迹”，先渲再用DriveRestorer在线修复，把原始与修复视频加入数据池，按从小到大侧移的加权采样闭环训练，逐步提升偏视角下的重建稳健性。

数据资源

具身智能的数据资源因其任务和形态的多样性，可被划分为四类：仿真平台、交互式基准、数据集与真实机器人平台。

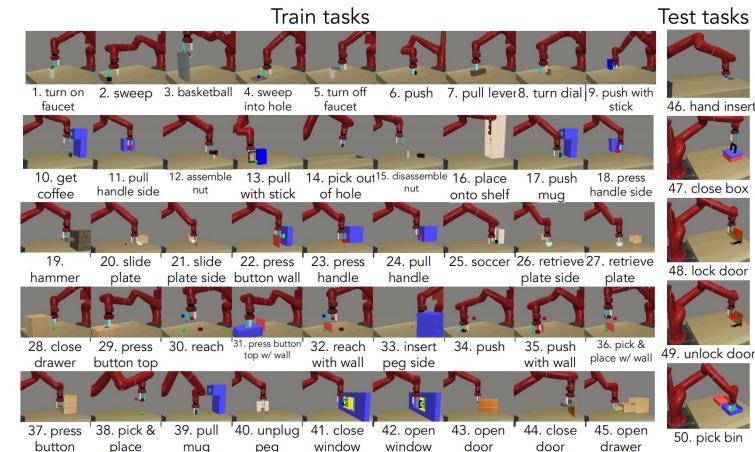


MuJoco仿真平台



"Ped with pet, bicycle, car makes a u-turn, lane change, peds crossing crosswalk"

nuScenes数据集



Meta-world交互基准



Unitree G1机器人

数据资源

TABLE 3
An overview of data resources for training and evaluating embodied world models.

Category	Name	Year	Task	Input	Domain	Scale	Protocol ¹
Platform	MuJoCo [218]	2012	Continuous control	Proprio.	Sim	-	-
	CARLA [219]	2017	Driving simulation	RGB-D/Seg/LiDAR/Radar/GPS/IMU	Sim	-	✓
	Habitat [220]	2019	Embodied navigation	RGB-D/Seg/GPS/Compass	Sim	-	✓
	Isaac Gym [221]	2021	continuous control	Proprio.	Sim	-	-
	Isaac Lab [222]	2023	Robot learning suites	RGB-D/Seg/LiDAR/Proprio.	Sim	-	-
Benchmark	Atari [223]	2013	Discrete-action game	RGB/State	Sim	55+ Games	✓
	DMC [224]	2018	Continuous control	RGB/Proprio.	Sim	30+ Tasks	✓
	Meta-World [225]	2019	Multi-task manipulation	RGB/Proprio.	Sim	50 tasks	✓
	RLBench [226]	2020	Robotic manipulation	RGB-D/Seg/Proprio.	Sim	100 tasks	✓
	nuPlan [227]	2021	Driving planning	RGB/LiDAR/Map/Proprio.	Real	1.5k hours	✓
	LIBERO [228]	2023	Lifelong manipulation	RGB/Text/Proprio.	Sim	130 tasks	✓
Dataset	SSv2 [229]	2018	Video-action understanding	RGB/Text	Real	220k videos	169k/24k/27k
	nuScenes [230]	2020	Driving perception	RGB/LiDAR/Radar/GPS/IMU	Real	1k scenes	700/150/150
	Waymo [231]	2020	Driving perception	RGB/LiDAR	Real	1.15k scenes	798/202/150
	HM3D [232]	2021	Indoor navigation	RGB-D	Real	1k scenes	800/100/100
	RT-1 [233]	2022	Real-robot manipulation	RGB/Text	Real	130k+ trajectories	-
	Occ3D [234]	2023	3D occupancy	RGB/LiDAR	Real	1.9k scenes	600/150/150; 798/202/-
	OXE [235]	2024	Cross-embodiment pretraining	RGB-D/LiDAR/Text	Real	1M+ trajectories	-
	OpenDV [90]	2024	Driving video pretraining	RGB/Text	Real	2k+ hours	-
	VideoMix22M [14]	2025	Video pretraining	RGB	Real	22M+ samples	-
Robot	Franka Emika [236]	2022	Manipulation	Proprio.	Real	-	-
	Unitree Go1 [237]	2021	Quadruped locomotion	RGB-D/LiDAR/Proprio.	Real	-	-
	Unitree G1 [238]	2024	Humanoid manipulation	RGB-D/LiDAR/Proprio./Audio	Real	-	-

¹ **Protocol:** For interactive benchmarks, a check mark (✓) indicates available evaluation protocols. For datasets, it indicates official data splits are provided.

评估指标

针对世界模型的不同侧重，有三层抽象水平：

层级	含义	指标	用途
像素级	视觉逼真/连贯	FID, FVD, SSIM, PSNR	视频生成、重建观感
状态级	几何/语义/轨迹一致	mIoU, mAP, ADE/FDE, BEV-IoU	占据/语义、规划一致性
任务级	能否完成目标	Success, Return, Collision	闭环控制、端到端评测

理想情况下，我们希望指标不仅关注像素级误差，更关心物理一致性和任务成功。因此，新近的评测倾向于设计诸如物理合规性、因果一致性等指标来弥补传统指标的不足。

主要挑战与未来方向

主要挑战：

1. **数据与评估**：缺乏统一的大规模**多模态**数据集以及衡量物理合理性、因果关系的评估指标。
2. **实时性**：先进模型（如DiT、VLM）**开销巨大**，难以满足**实时控制**的要求。如何在不**牺牲性能**的前提下提速是关键。
3. **建模策略与平衡**：顺序模拟和全局预测、不同**空间表示**之间寻求最佳折中。

未来方向：

1. **统一的数据基准**：构建能评估**物理一致性的**跨领域数据基准。
2. **提升推理效率**：模型压缩（量化/剪枝/蒸馏/稀疏）与更高效的**时序建模**（如Mamba）。
3. **策略与平衡**：结合顺序+全局、显式记忆与CoT式任务分解来支持长程推理，走粗到细、局部精化的混合范式，减少长程误差累积与错误循环。

Thanks