第 1 章

1 世界模型概览



自20世纪中叶图灵测试被提出以来，人工智能的发展长期在逻辑推演与模式识别之间交替演进。生成式预训练Transformer（Generative Pre-trained Transformer，GPT）及其衍生的大语言模型（LLM）推动了自然语言处理的范式转变，但仅靠文本统计预测仍难以支撑可靠的物理交互：一个能够撰写量子力学综述的AI，也可能无法准确预测一个翻滚的杯子落地后的破碎形态。因此，世界模型的兴起并不是对语言模型的简单替代，而是把智能系统的建模对象从符号序列进一步推向可感知、可干预并随时间演化的环境。

在通用智能相关研究中，世界模型常被用来讨论常识缺失与规划脆弱这两个问题。本章作为全书导论，首先给出世界模型的定义边界、历史脉络、技术版图、核心路线以及主要驱动因素，为后续章节的原理剖析与工程实现提供统一的理论基础。

1.1 定义与边界

1.1.1 世界模型的核心内涵与形式化定义

界定世界模型（World Models）的内涵，首先需要打破将其等同于单纯视觉预测器的陈见。从本质上讲，世界模型是智能体对环境动力学所学习得到的一种内部近似模型，它允许智能体在不与真实环境发生交互的情况下，通过内部的想象进行推理、规划与决策[1]。从技术实现的视角来看，世界模型通常被表述为部分可观测马尔可夫决策过程（Partially Observable Markov Decision Process，POMDP）的近似器：由于智能体往往无法直接获取世界的真实全量状态s_t，而只能观测到带有噪声或信息缺失的观测值o_t，世界模型的核心是学习环境的状态转移方程$P(s_{t+1} \mid s_t, a_t)$，即在当前状态s_t下执行动作a_t后，世界将如何演化至下一状态s_{t+1}。

一个有效的世界模型并不必然追求像素级重构，而更关注与决策相关的状态结构和动态规律。例如，在模拟球类运动时，模型不需要预测每一处草坪阴影的变化，但应当稳定预测球的弹跳轨迹与碰撞规律。现代世界模型通常把确定性状态与随机变量相结合，同时描述相对稳定的物理规律和环境中的随机扰动，使智能体能够评估多种可能的未来，而不是只给出单一路径。

1.1.2　智能范式的拓展：从大语言模型到世界模型

大语言模型的核心逻辑在于通过海量语料库学习token之间的条件概率关系，其典型预测目标是下一个token的概率分布$P(x_{t+1} \mid x_{1:t})$。这种基于统计拟合的智能在处理抽象语义和逻辑推理时表现出色，但由于缺乏与物理环境持续交互所形成的具身经验，其物理常识与环境理解能力仍存在一定局限。相比之下，世界模型将建模对象从纯文本空间进一步拓展到可感知、可行动的环境空间，为认知科学中的符号接地问题提供了一条重要的技术探索路径[2]：LLM中的“苹果”主要表现为高维语义空间中的一种表征，而世界模型中的“苹果”则进一步关联其形状、质量、运动和碰撞等物理属性，以及随时间变化的状态，使AI从单纯的信息处理者逐渐转变为环境的观察者与干预者。

世界模型的兴起也可置于莫拉维克悖论的背景下理解。该悖论指出，对机器而言，一些高级推理任务所需的计算并不一定高于人类日常完成的感知与运动任务。大语言模型已经展示了符号与语义推理的可行路径，而世界模型关注的是另一类更贴近物理交互的问题——那些对人类而言自然、对机器而言却难以稳定完成的常识性感知与规划任务[3]。

1.1.3　理清概念边界：与模拟器、生成式 AI 的核心差异

为了准确把握世界模型的范畴，有必要将其与相近概念进行区分，其核心差异如表1-1所示。虽然视频生成模型（如Happy Horse）能够产生高度逼真的视觉序列，但若缺乏基于动作的稳定且可控的动作干预能力，则更适合作为视觉内容生成模型，而不能直接等同于具备规划与决策价值的世界模型；经典物理模拟器（如MuJoCo）主要依赖人工设定的物理方程和参数进行数值计算，其环境动力学通常并非直接从观测数据中学习得到，因此与数据驱动的世界模型在建模方式上存在明显差异。

表 1-1　智能系统层级与世界模型边界对比

模型类型	核心目标	核心输入维度	是否支持动作干预	物理接地程度
生成式大语言模型	文本符号与语义关系建模	纯文本/多模态token	通常否	较低，以语义关联为主
视频生成模型	视觉序列生成与时序预测	文本/图像等条件	较弱	中等，以视觉一致性为主
经典物理模拟器	基于显式物理规律进行数值计算	物理参数/状态变量	是	高，基于显式物理模型
世界模型	环境表征与动力学学习	时序观测+动作序列	是	较高，可以从数据中学习环境规律

世界模型的关键不仅在于生成未来画面，还要支持动作条件下的预测。只有当模型能够回答“若执行动作a_t，未来状态s_{t+1}将如何演化”时，它才具备真正的规划与决策价值。

1.2　五代演进谱系

世界模型并非近年来新近诞生的概念，其思想根源可以追溯至20世纪中叶的控制理论、状态空间建模与系统辨识，并随着神经网络、计算机视觉、强化学习以及大规模算力的发展不断演进。如果从环境建模方式和预测表征形式的变化来看，其发展大致可以划分为5个阶段，主要演进谱系如表1-2所示。

表 1-2　世界模型演进谱系一览

发展阶段	核心算法架构	关键技术特征	代表性成果	核心局限
G1	显式数学方程与状态空间模型	专家定义、显式建模、严格公式化	Kalman 滤波、状态空间模型等	难以覆盖高度复杂的非结构化环境
G2	VAE+RNN+Controller 三位一体架构	潜空间压缩、时序预测、内部“想象”	Ha&Schmidhuber World Models	长期预测易发生时序漂移
G3	RSSM+随机隐状态架构	确定性记忆与随机表征结合、隐空间规划	Dreamer V1~V3	高保真视觉细节重构能力有限
G4	Transformer 及生成式视频模型	长程时序建模、大规模视频学习、生成式环境建模	Genie 等	计算开销较高，物理一致性与可控性仍有限
G5	JEPA 类联合嵌入预测架构	表征空间预测、语义抽象、减少无关细节干扰	V-JEPA、I-JEPA 等	动作条件建模与通用规划能力仍在发展中

1.2.1　第一代：符号主义显式参数模型（符号主义时代）

严格来说，第一代方法尚未具备“从数据中学习”的世界模型特征，但它们为后续数据驱动方法奠定了数学基础。在人工智能的早期阶段，研究者主要通过手写规则和显式物理方程来描述环境，这类方法可视为现代世界模型思想的前身。它们在航天控制等边界清晰的任务中发挥了重要作用，但面对非结构化的日常环境时泛化能力受限。原因在于，现实场景中的变量数量、噪声类型和组合方式难以在事前被完整枚举；一旦环境超出预定义参数范围，模型便难以适应。学术界也因此逐步转向从观测数据中学习环境规律的技术路线。

1.2.2　第二代：连接主义隐空间初步探索（RNN/VAE 时代）

随着反向传播算法的成熟，以RNN、LSTM为代表的序列模型开始在时序模拟任务中得到广泛应用。2018年，Ha与Schmidhuber发表了奠基性论文*World Models*（世界模型），提出了由VAE压缩

感知、RNN时间记忆和Controller决策组成的三段式架构[1]。这项工作展示了智能体如何在世界模型生成的内部“梦境”中训练策略，再将策略迁移到真实环境中执行，从而减少强化学习对真实交互样本的依赖。第二代模型由此把世界模型研究从显式方程描述推进到数据驱动的潜空间建模阶段。

1.2.3　第三代：离散化表征与健壮性规划体系（RSSM/Dreamer 时代）

2020年前后，DeepMind的研究者通过引入递归状态空间模型（Recurrent State-Space Model，RSSM），将确定性的RNN隐状态与随机性的观测变量结合，有效缓解了长期预测中的误差累积和状态漂移问题。Dreamer系列算法的提出，标志着世界模型显著降低了对真实环境奖励信号的即时依赖[4]。其中，Dreamer V2将连续隐变量替换为离散分类变量，其设计受到生物神经元离散放电机制的启发，显著增强了模型对复杂时序逻辑的抗干扰能力，使世界模型在Atari游戏、Minecraft等具有长期依赖关系的复杂任务中表现优异。

1.2.4　第四代：Transformer 与扩散模型的架构重构

在大规模预训练浪潮下，世界模型开始在部分路线中替代传统的循环架构，转向基于Transformer的全局注意力机制；与此同时，扩散模型等生成式技术的发展，使世界模型具备了生成高保真视频的能力。在这一阶段，Google的Genie等模型展现了从大规模视频数据中学习交互式环境的潜力，它们不再是简单的预测器，而是逐渐演化为可交互、实时运行的神经驱动的交互式环境模拟器，无须显式输入物理方程，即可从海量数据中在一定程度上隐式学习摩擦力、重力乃至简单流体力学等自然规律。与此同时，WorldSimBench等评测基准开始系统检验视频生成模型充当世界模拟器的能力[6]。

1.2.5　第五代：联合嵌入架构与通用具身智能（JEPA 时代）

2023年以来的发展以Yann LeCun提出的JEPA（Joint Embedding Predictive Architecture，联合嵌入预测架构）为代表，该架构弱化了像素级重构目标，转而追求抽象特征空间内的预测一致性[7]。这一技术路线试图解决一个关键问题：模型不应把所有不可预测的视觉细节都作为主要目标，而应把模型容量集中于与任务相关的结构信息。通过在隐空间中进行表征预测，JEPA模型可以降低背景噪声（如被风吹动的树叶）对预测目标的干扰，更关注目标事件是否发生、物体是否发生接触等具有因果重要性的事件。这种对无关变化的有效过滤，被认为是具身智能进一步发展的重要基础。

1.3　核心技术版图

一个功能完备的世界模型系统并非单一的神经网络，而是一个由多个专门化模块构成的模块化软件体系。这些模块在逻辑上模拟了哺乳动物大脑中视觉皮层、海马体与前额叶等脑区的分工协同，其核心模块架构与信息流向如图1-1所示。

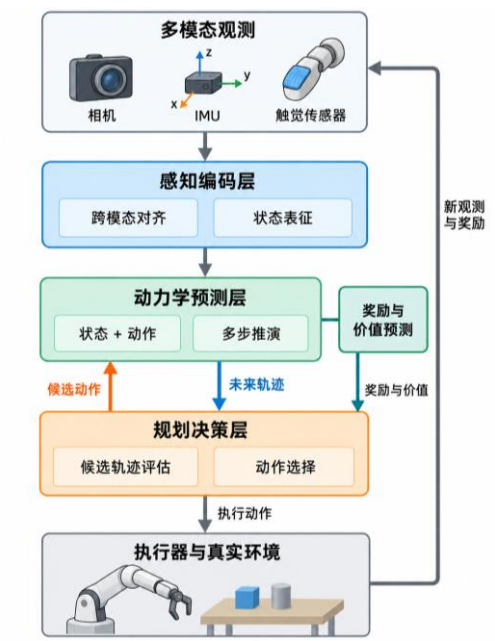


图 1-1　世界模型核心技术版图与模块协同架构

1.3.1　感知编码层：高维异构信息的数字抽象

感知模块承担高维信息压缩与表征学习的任务，需要把视频帧、惯性测量单元（Inertial Measurement Unit，IMU）加速度、机械臂力觉反馈等异构传感器数据映射为紧凑的特征张量。现代世界模型常集成DINOv2等大规模视觉预训练模型作为特征编码器，使模型在进入具体物理任务之前，已经具备一定的物体边界、遮挡关系和几何结构表征能力。因此，感知模块不仅是降维工具，也在很大程度上决定了后续动力学模型能够获取哪些环境结构信息。

感知模块的难点之一是跨模态对齐。例如，当机械手触摸金属物体时，压力传感器信号需要在隐空间中与该物体视觉外观对应的视觉特征建立关联，形成统一的实体表征。缺乏这种跨模态对齐时，后续动力学模型容易把同一物体在不同传感器下的观测误判为彼此无关的信号。

1.3.2　动力学预测层：世界演化的因果时序模拟

动力学预测器决定了智能体能够向前推演的时间跨度及预测稳定性，其数学本质是学习条件概率分布$p(z_{t+1} \mid z_{1:t}, a_{1:t})$。随着Transformer等序列模型进入世界模型，长程依赖建模不再完全依赖传统循环神经网络。时间记忆的长度直接影响规划深度：只能预测未来1秒的模型，很难处理烹饪、装配等具有长时序依赖的任务；而能够稳定预测更长时间尺度的模型，才有可能支持复杂的多步骤规划。

为了提高训练效率，世界模型通常支持并行化的梦境回放（Dream Rollout）：智能体可以在

一次GPU推理过程中并行模拟大量不同的行动路径，快速搜索潜在的最优策略。这种大规模隐空间并行模拟能力，使世界模型在样本效率方面显著高于传统的无模型（Model-free）强化学习算法。

注：Rollout在世界模型语境中可以译为“想象”或“推演”，它们在本质上描述的是同一过程，但“想象”更侧重于对人类心智活动的类比，而“推演”更侧重于技术实现层面的描述。本书将根据具体语境交替使用这两个术语，以描述世界模型基于内部状态预测未来轨迹的过程。

1.3.3　规划决策层：从虚拟想象到真实动作执行

模拟最终要服务于决策。规划模块利用世界模型的预测能力，在潜空间内执行蒙特卡洛树搜索（Monte Carlo Tree Search，MCTS）、模型预测控制（Model Predictive Control，MPC）等算法。这样，智能体能够先在内部模型中评估候选动作，再把较优的动作交由真实系统执行，从而有效降低真实环境中的试错成本。

一些世界模型还引入奖励或价值预测模块和内在动机机制：前者用于估计动作可能带来的收益或代价，后者则鼓励智能体主动探索预测不确定性较高的区域。这类机制通常用于提升探索效率和提升技能学习速度。

1.4　主要技术路线

尽管世界模型的技术底层在持续融合，但在应用层面上，目前已分化为三条各具侧重的主流技术路线。这些路线之间的差异，本质上体现了学术界与工业界对“世界模型应优先服务于预测、生成还是决策”这一问题的不同判断。

1.4.1　非生成式路线：以决策效率与轻量化为核心

这一路线的代表是Dreamer V3、TD-MPC2，目标是为强化学习提供更高效的环境建模与策略学习机制。它并不追求逐帧生成高保真的视觉效果，而更关注隐空间动力学的准确性及计算成本的可控性。

Dreamer V3在未使用人类示范数据的条件下完成Minecraft中的钻石获取任务，说明只要潜空间保留了与决策相关的环境结构，即使解码图像不够精细，也能够支撑复杂行为学习。

1.4.2　生成式路线：以视觉与常识统一为目标

以Google的Genie等模型为代表。这一路线强调，长视频或交互环境生成如果能够稳定保持物体恒常性、运动连续性和场景交互关系，就说明模型已经学习到了一部分环境结构。在这种定义下，世界模型更接近于一种通用的交互式视觉环境：Genie等模型可以根据草图、文本或视频片段生成可控制、可探索的虚拟环境。但这一路线也面临推理延迟高、计算开销大以及长期一致性和动作控制能力仍需提升等问题。

1.4.3　混合式路线：融合生成与预测的联合架构

这一路线以Meta的V-JEPA、NVIDIA的Cosmos等工作为代表，聚焦机器人操作、自动驾驶等真实物理场景，强调模型需要处理真实环境中的接触、运动和空间约束。基于V-JEPA表征的后续机器人系统展示了通过大规模视频学习物体运动规律，并在新的实验室环境中实现对未见物体零样本泛化抓取的能力。这类结果说明，视觉预训练表征有望降低传统工业机器人对任务专用代码的依赖。具身路线也越来越重视空间智能，即对三维结构、深度关系以及物体相对位置的显式建模与理解，这将直接影响世界模型能否走出受控实验环境。

1.5　核心发展驱动力

世界模型近年来快速发展并受到广泛关注，并非源于单一技术突破，而是由算力、数据和具身应用需求共同推动的。

1.5.1　规模定律的边界拓展

在LLM时代，我们见证了参数规模和预训练数据规模持续扩大所带来的性能提升[8]。而在世界模型语境下，规模定律（Scaling Law）正在向新的维度延伸：一是推理阶段计算的重要性不断提升，世界模型可以通过搜索、规划和内部模拟提高决策质量；当模型在作出决定前进行更多未来推演时，规划效果通常会随推演次数的增加而改善，但其收益会逐渐递减。二是合成数据的自我迭代：随着高质量真实数据获取成本不断上升，具有一定物理一致性的合成数据可以作为真实数据的补充，帮助模型覆盖更多长尾情形。

1.5.2　多模态基础模型的能力支撑

早期的世界模型由于缺乏较强的语义理解能力，其习得的技能往往难以通过自然语言进行指令控制。随着多模态大语言模型的成熟，世界模型与大语言模型（LLM）开始形成更加紧密的分工：大语言模型负责高层次的语义规划和任务拆解，世界模型负责底层的物理模拟和动作反馈。这种分层协同架构有助于缓解具身智能中“听得懂指令，却难以稳定执行”的问题，也为世界模型接入自然语言任务提供了统一的语义接口。

1.5.3　产业落地与安全对齐的刚需

在自动驾驶、工业制造、人形机器人等领域，安全性是系统设计中的首要约束。真实环境测试成本高、风险大，世界模型可以提供可反复测试的虚拟环境或数字孪生环境，用于构造长尾场景和开展压力测试。它不能替代真实验证，但可以把一部分高风险试错前移到仿真阶段，降低实车、实机测试中的风险和测试盲区。

1.6　本章小结

本书按照“概念辨析–理论奠基–机制剖析–应用聚焦–工程实践–前沿探索”的顺序组织，形成五个阅读板块：概念辨析与理论奠基合为基础篇，对应第1～4章，先界定世界模型与大语言模型和视觉语言模型（LLM/VLM）的能力边界，再建立数学形式化描述和表征学习基础；模型篇对应第5～8章，剖析RSSM、Transformer动力学路线、规划机制与奖励–价值学习；应用篇对应第9～12章，讨论具身智能与机器人、自动驾驶、智能体（Agent）内部模拟器与多模态世界模型；工程篇对应第13～15章，集中处理数据、训练与部署问题；展望篇对应第16章，讨论未来世界模型的发展方向。本书的组织方式，是先厘清世界模型与相邻技术概念的差异、明确其数学描述语言，再逐步进入模型机制、应用场景与工程实现。

本章作为全书导论，界定了世界模型的基本边界：它不是单纯的视频生成器，也不是手写物理模拟器，而是面向可观测、可干预且随时间演化环境的内部环境模型或动力学近似。本章从定义、演进、技术版图、主要技术路线和产业驱动力五个角度，为后续章节建立了共同语境。

概览之后，读者应把握一条主线：世界模型的关键能力在于把观测、动作和未来后果连接起来，使智能体能够在内部模拟中比较不同的行动方案。第2章将首先比较世界模型与LLM/VLM在生成、预测、行动条件和记忆结构等方面的差异，明确为什么世界模型不能被简单归入通用生成模型。

1.7　本章参考文献

[1] HA D, SCHMIDHUBER J. World Models[Z]. Zenodo, 2018. DOI:10.5281/ZENODO.1207631.

[2] HARNAD S. The Symbol Grounding Problem[J]. Physica D: Nonlinear Phenomena, 1990, 42(1–3): 335-346. DOI:10.1016/0167-2789(90)90087-6.

[3] MORAVEC H. Mind Children: The Future of Robot and Human Intelligence[M]. Cambridge, MA: Harvard University Press, 1988.

[4] HAFNER D, LILLICRAP T, BA J, 等. Dream to Control: Learning Behaviors by Latent Imagination[C]. International Conference on Learning Representations (ICLR). 2020. DOI:10.48550/arXiv. 1912.01603.

[5] OKADA M, TANIGUCHI T. DreamingV2: Reinforcement Learning with Discrete World Models without Reconstruction[C].2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2022: 985-991. DOI:10.1109/IROS47612.2022.9981405.

[6] QIN Y, SHI Z, YU J, 等. WorldSimBench: Towards Video Generation Models as World Simulators[C]. Proceedings of the 42nd International Conference on Machine Learning (ICML). 2025. DOI:10.48550/arXiv. 2410.18072.

[7] ASSRAN M, DUVAL Q, MISRA I, 等. Self-Supervised Learning from Images with a Joint-Embedding

Predictive Architecture[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 15619-15629. DOI:10.1109/CVPR52729.2023.01499.

[8] KAPLAN J, MCCANDLISH S, HENIGHAN T, 等. Scaling Laws for Neural Language Models[Z]. DOI:10.48550/arXiv.2001.08361.

第 2 章

世界模型与LLM/VLM的根本差异



本章导读

本章承接导论中的概念边界，深入辨析世界模型与LLM/VLM的根本差异。二者都可能采用大规模预训练、多模态输入和生成式接口，但目标函数、内部表示和行动闭环并不相同。本章首先比较生成与预测两种范式，再说明可执行表示和行动条件生成为何是世界模型区别于通用生成模型的关键，最后落到长期依赖与记忆结构，讨论两类模型在持续交互任务中的能力边界。

2.1 生成与预测的本质

当我们审视人工智能领域近十年的发展轨迹，会发现两条看似相似却走向截然不同方向的技术主线：一条是以GPT系列为代表的大语言模型（Large Language Model，LLM）[1]及其多模态扩展（Vision-Language Model，VLM）[5]，它们凭借强大的自回归生成能力重塑了自然语言处理与内容创作的格局；另一条则是以Dreamer、Genie为代表的世界模型[8]，它们通过学习环境状态及其随动作变化的动态规律，为具身智能与决策系统提供了重要支撑。表面上看，两者都涉及根据输入产生新输出的能力，但深入其底层机制便会发现，内容生成与环境预测在目标和约束上存在明显差异。这种差异不仅体现在技术实现上，也决定了它们各自的能力边界与适用场景。

2.1.1 自回归生成范式的统计本质

自回归生成范式的核心可以概括为一句话：给定历史序列，预测下一个token的概率分布，并据此进行采样。无论是处理文本、图像还是视频，LLM/VLM都会将输入统一编码为离散的token序列，再通过Transformer架构学习序列中token之间的统计依赖关系[12]。

LLM的生成过程具有两个显著特征。首先，它本质上是一个概率采样过程。虽然可以通过温度系数、Top-K采样等技术调节生成结果的多样性与确定性，但模型本身通常并不显式编码完整的物理定律或逻辑约束。因此，它可能生成语法正确但物理上不成立、逻辑上相互矛盾的内容，即通常所说的“幻觉”现象。其次，它主要依赖训练分布中的统计模式。LLM能够在已有语料覆盖的语义空间内进行有效组合，但面对分布外（Out-of-Distribution，OOD）输入时，其性能往往会下降，外推结果也更加不稳定。

2.1.2　世界模型预测范式的因果本质

与LLM的生成范式不同，世界模型关注的是动作条件下的状态演化预测。它的目标不是生成一段看起来合理的序列，而是学习一个能够近似环境动力学的函数。该函数以当前环境状态和智能体动作作为输入，输出下一时刻的环境状态。在这里，预测不仅是统计意义上的延续，更需要尽可能反映物体运动、相互作用、因果关系以及时间演化等物理约束[7]。

世界模型的预测过程可以理解为受约束的状态转移。与开放式文本生成相比，动力学模型通常具有更强的时序先验结构的归纳偏置。例如，在视频预测任务中，模型需要学习物体的连续性、运动的平滑性以及碰撞后的相互作用。这些约束未必来自显式编程，更多来源于交互数据和模型结构共同形成的统计规律。约束越稳定，模型越有可能在未见的组合中预测出物理上合理的未来状态。

为了更直观地理解两种范式的差异，我们可以对比它们在处理同一任务时的不同表现。假设我们有一个简单的物理场景：一个小球在桌面上滚动，前方有一个障碍物。LLM在描述这一场景时，可能会依据训练数据中的常见模式生成“小球撞到障碍物后停了下来”或“小球绕过障碍物继续滚动”等描述，但仅依赖LLM本身，通常难以稳定预测小球碰撞后的连续运动轨迹、速度变化和反弹角度。而世界模型则能够通过学习到的动力学函数，预测小球在后续时刻的位置、速度和加速度等状态变化，并对不同材质或碰撞条件下的结果进行条件推演。这种差异的根源在于，LLM主要学习描述物理过程的语言规律，而世界模型则更直接学习环境状态随动作演化的动力学规律。

两种范式的差异还体现在训练目标的设计上。LLM通常采用下一个token预测作为核心训练目标，并通过交叉熵损失优化序列建模能力；这一目标主要约束局部条件概率，而不直接保证长序列在事实、物理或逻辑层面的全局一致性。

在知识表示方面，LLM将知识隐式编码在Transformer模型的参数（权重）中。这种表示方式赋予了LLM较强的任务泛化能力，但也导致知识难以解释、难以编辑。我们通常无法直接从LLM的参数中提取某一具体的物理定律，也无法在不重新训练整个模型的情况下修改其中某一部分知识。相比之下，在许多基于潜空间动力学的世界模型中，世界模型通常采用结构化或潜在状态表示，将环境信息编码为紧凑的状态变量或潜在向量，其中可以包含与任务相关的物体位置、速度、形状等信息。这种表示方式使世界模型具有更好的可解释性和可干预性：既可以通过修改潜在向量来影响模型的预测结果，也可以通过分析状态变量的变化，更深入地理解模型的决策过程。

生成范式与预测范式的计算流程对比如图2-1所示。

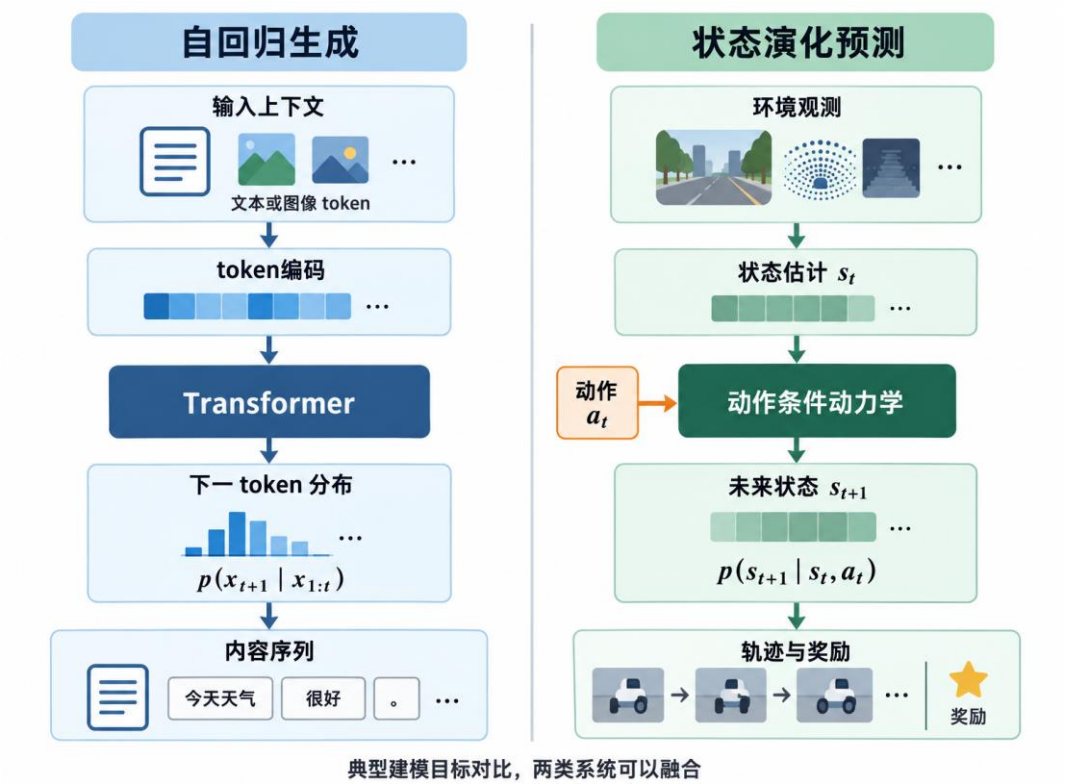


图 2-1　生成范式与预测范式的计算流程对比

2.1.3　范式差异的深层内涵与融合方向

生成与预测的差异，最终体现在推理方式上。LLM更擅长基于相似语境进行联想和补全。例如，它可以回答“玻璃杯掉在地上会发生什么”，因为训练语料中包含了大量类似描述。但当材料、环境或动作条件发生变化时，仅依赖语言共现关系，很难保证推理结果仍然符合物理规律。世界模型则更强调模拟式推理：在内部状态空间中预测动作发生后的环境变化，再据此比较不同行动可能带来的后果。这也是世界模型能够用于决策任务的重要原因。

这并不意味着LLM的价值有所降低。LLM在自然语言理解、知识检索和内容生成等方面仍具有明显优势；而世界模型则更适用于具身智能、机器人控制、自动驾驶等需要持续预测物理后果的任务。区分生成与预测两种范式，并不是为了强调孰优孰劣，而是为了针对不同任务选择更合适的建模方式。

从任务角度看，生成范式主要回答“如何产生符合上下文的内容”，预测范式则进一步回答“状态会如何演化”以及“动作会带来什么后果”。LLM已经证明了大规模语义建模的能力，但通常并不天然具备面向物理环境的状态估计和动作条件预测机制。世界模型的意义，在于把环境演化和行动后果纳入模型内部，使智能体能够在行动之前完成推演与评估。

当然，生成与预测并不是完全对立的两种范式，它们之间存在着相互融合的趋势。近年来，越来越多的研究开始探索将LLM的语义理解能力与世界模型的物理预测能力相结合，构建更加通用的人工智能系统。例如，可以利用LLM解析自然语言指令，将其转换为世界模型能够理解的目标状态，然后由世界模型规划实现这一目标的动作序列。这种融合不仅能够发挥两种范式各自的优势，还有望催生新的能力，如基于自然语言的物理推理、多模态反事实推理等。

总体而言，生成与预测代表了两种不同的计算范式：LLM侧重语义建模和内容生成，世界模型侧重状态演化和行动后果。理解这一根本差异之后，才能进一步讨论可执行表示、长期依赖和记忆结构等更具体的能力边界。

2.2　可执行表示与行动条件生成

2.1节从计算范式上区分了生成与预测。进一步看，两者的差异还体现在表示方式上：LLM/VLM学习的是面向描述的语义表示，而世界模型学习的是面向行动的可执行表示。后者把动作纳入状态转移过程，因此可以支持行动条件生成和多分支规划。

2.2.1　可执行表示的核心内涵与范式对比

可执行表示（Executable Representation）是指一种能够与动作结合，并通过状态转移预测未来状态的表示形式。与LLM/VLM中的离散token不同，世界模型的可执行表示通常采用潜在状态表示。它不仅编码环境的当前状态，还编码状态如何随动作发生变化。所谓“可执行”，是指在表示空间中输入行动后，模型可以推进状态，从而得到下一步甚至多步未来的表示。

相比之下，LLM/VLM的语义表示更偏向描述性：它可以表达“杯子在桌子边缘”，但通常并不显式维护“轻推杯子会如何改变状态”这样的动力学关系。在LLM中，“推”通常表现为token或语义片段；而在世界模型中，动作作为状态转移的输入，直接影响下一时刻的潜在状态。

为了更清晰地理解两种表示的差异，可以从数学形式上进行对比。LLM的自回归生成过程可表示为：

$$p(x_{t+1}|x_1, x_2, \cdots, x_t)$$

其中，x_t表示第t个token。该公式表明，LLM的下一个输出由历史token序列条件化，动作信息如果存在，通常也需要先以相应的输入形式编码到上下文中。相比之下，世界模型的预测过程则可表示为：

$$p(s_{t+1}|s_t, a_t)$$

其中，s_t表示第t时刻的潜在状态，a_t表示第t时刻的智能体行动。该公式表明，世界模型的未来状态不仅依赖当前状态，还显式依赖于智能体采取的行动。这种对行动条件的显式建模，正是可执行表示的重要特征。

语义表示与可执行表示的结构对比如图2-2所示。

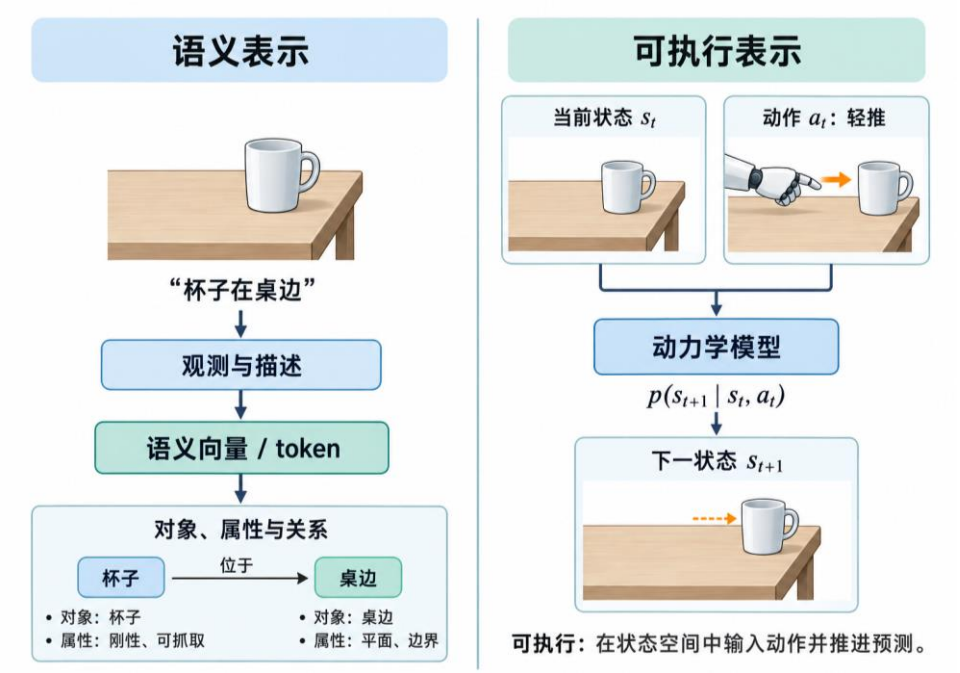


图 2-2　语义表示与可执行表示的结构对比

语义表示与可执行表示的核心特性对比如表2-1所示。

表 2-1　语义表示与可执行表示的核心特性对比

特性维度	LLM/VLM 语义表示	世界模型可执行表示
核心目标	描述世界是什么	预测世界会如何变化
表示形式	token 或多模态特征	状态或潜在状态表示
状态转移	主要依赖历史上下文	依赖当前状态和行动
生成特性	上下文条件生成	行动条件下的多分支预测
可干预性	通过上下文等条件引导	通过行动输入直接干预
知识类型	以陈述性知识为主	更接近程序性知识
适用场景	内容创作、知识问答	决策规划、机器人控制

行动条件生成与多分支预测流程如图2-3所示。

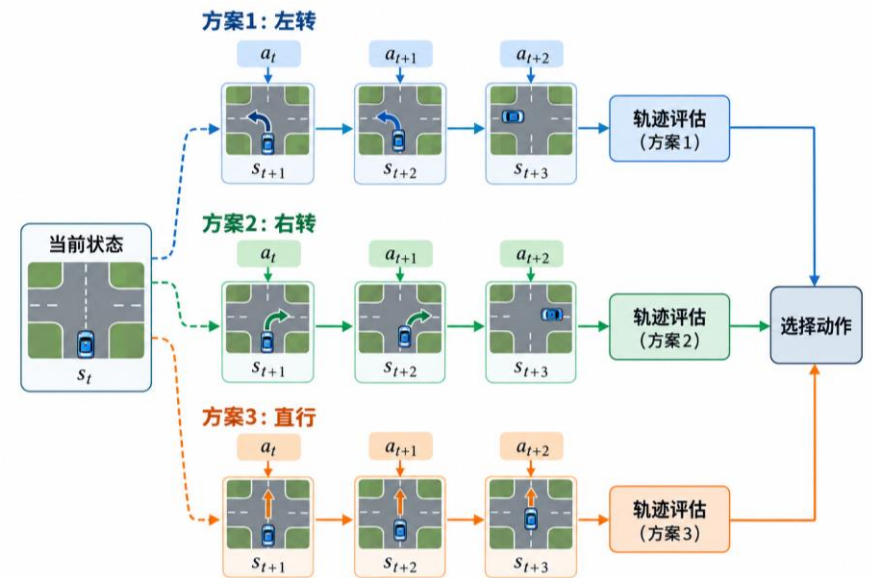


图 2-3　行动条件生成与多分支预测流程

2.2.2　行动条件生成的核心特性与决策支撑

行动条件生成（Action-Conditioned Generation）是可执行表示的自然产物，指世界模型能够根据不同的行动输入生成对应的未来状态序列。这一能力使世界模型成为一种“内部模拟器”：智能体可以在采取真实行动之前，先在该模拟器中预演不同行动的后果，从而选择较优的行动方案。与一般的上下文条件生成相比，行动条件生成具有三个关键特性：多分支性、行动一致性和可干预性。

多分支性是指，对于同一个当前状态，不同行动可以导致不同的未来状态轨迹。LLM的生成过程虽然也可以通过采样产生不同的输出分支，但这些分支通常不是由显式的行动变量所驱动，而是在给定上下文条件下生成不同的可能结果。例如，当LLM被要求续写“我站在十字路口”时，它可能会生成“我向左转”或“我向右转”的文本，但通常不会主动对这两种行为对应的后续状态进行并行推演和比较。而世界模型则可以同时模拟“向左转”和“向右转”两种行动对应的未来状态，生成两条并行的状态轨迹，并根据奖励函数评估轨迹。

行为一致性是指，行动条件生成的状态序列应当与输入行动保持合理的对应关系。在世界模型中，行动作为状态转移的重要条件，状态变化应当与所采取的行动相一致，这种关系主要通过大量交互数据中的“行为-结果”对应规律学习得到，而不一定意味着模型已经获得了严格意义上的因果结构。因此，世界模型生成的未来状态应当是行动的合理结果，而不是与行动无关的随机延续。例如，当智能体采取“向前走”的行动时，世界模型预测的下一个状态中，智能体的位置通常应表现出与向前运动相一致的变化。这种行为一致性是世界模型能够用于决策的重要基础，因为只有当行动与结果之间存在稳定的对应关系时，智能体才能通过选择行动来影响未来状态。

可干预性是指，可以通过改变行动输入直接干预世界模型的预测过程，从而获得不同的未来状态。在LLM中，我们通常只能通过修改上下文来引导生成过程，这种引导更多是通过输入条件影响生成结果；而在世界模型中，可以直接调整行动输入，从而改变状态演化的方向。这种可干预性使世界模型非常适合用于规划任务，因为智能体可以不断调整动作序列，寻找更能达到目标状态的路径。

当前主流的实现路径分为两类：一类是梦境规划范式，以Dreamer系列算法为代表[9,11]，在潜空间进行想象并学习策略，通过“策略网络-价值网络”（Actor-Critic）框架优化控制策略，直接输出端到端的控制行为，并提高真实环境交互的样本利用效率；另一类是搜索规划范式，以MuZero为代表[6]，将学习得到的动力学预测与蒙特卡洛树搜索（Monte Carlo Tree Search，MCTS）相结合，在内部模型中完成候选动作的搜索与评估。其他世界模型还可以结合模型预测控制（Model Predictive Control，MPC）进行轨迹优化，从而适配自动驾驶、复杂游戏等不同类型的决策和控制任务。

2.2.3　可执行表示的认知本质与落地挑战

可执行表示更接近程序性知识。陈述性知识回答“是什么”，例如水在100℃会沸腾；程序性知识回答“怎么做”，例如如何保持平衡、如何推门、如何抓取物体。LLM/VLM主要从文本和图像中学习可描述的知识，而世界模型则更强调从交互轨迹中学习动作和结果之间的关系。很多操作经验虽然很难完整写成语言规则，却可以通过状态转移模型加以学习。

这也可以联系到具身认知——其源头可追溯至Brooks提出的“无表征智能”论断[2]。具身认知认为，认知不仅来自抽象符号处理，还受到身体结构、感官输入和环境交互方式的影响。LLM/VLM通常缺少直接的物理行动接口，因此更偏向于学习关于世界的描述；世界模型则把动作、环境状态和反馈信息纳入状态转移过程，因而更适合处理真实环境中的操作与决策任务。

需要强调的是，可执行表示并不排斥语义信息，相反，它可以与语义表示有机结合。近年来，越来越多的研究开始探索将LLM的语义理解能力与世界模型的环境预测能力相结合，构建语义与行动相连接的表示。例如，可以使用LLM将自然语言指令转换为目标状态表示，然后由世界模型在可执行表示空间中规划达到目标状态的动作序列。这种结合不仅能够发挥两种表示各自的优势，还能够支持基于自然语言的物理推理、多模态任务规划等能力。

然而，构建高质量的可执行表示仍然面临诸多挑战。首先，可执行表示需要编码足够丰富的环境信息，包括物体的属性、空间关系、动力学特性等，这对编码器和状态表示的设计提出了较高要求。其次，行动条件生成需要保证长序列的状态一致性和预测稳定性，避免误差累积导致预测失真。最后，可执行表示需要具有良好的泛化能力，能够在未见过的场景中合理预测行动后果。这些挑战正是当前世界模型研究的热点问题，我们将在后续章节中详细讨论。

可执行表示与行动条件生成是世界模型区别于LLM/VLM的重要特征。LLM/VLM擅长描述世界，世界模型则尝试建模“行动如何改变世界”。这一差异决定了它更适合作为机器人、自动驾驶

和交互式智能体中的内部模拟器。

2.3节将进一步讨论长期依赖和记忆结构，这也是两类模型在实时交互任务中差异较为明显的方面。

2.3　长期依赖与记忆结构差异

前两节讨论了计算范式和表示方式。本节进一步讨论记忆结构：LLM/VLM主要依赖上下文窗口和全局注意力，而世界模型通常采用状态空间中的递归或增量式记忆。这一差异会直接影响长序列建模能力、实时性以及状态一致性。

2.3.1　LLM/VLM 的全局注意力记忆与长期依赖局限

长期依赖问题是序列建模领域的经典挑战，指模型能否捕捉序列中相距较远元素之间的依赖关系。对于人工智能系统而言，长期依赖能力直接影响它能否理解复杂的过程、记住历史关键事件，并作出连贯的决策。人类能够轻松记住几小时前发生的事情，并依据这些记忆指导当前行动，但对于大多数机器学习模型而言，长时间跨度下的信息保持和有效利用仍然具有挑战性。

LLM/VLM解决长期依赖问题的核心机制之一是Transformer的自注意力架构。自注意力机制允许模型直接计算序列中不同token之间的依赖关系，从机制上具有建模长距离依赖的能力。

然而，LLM的这种记忆机制主要依赖上下文的历史信息。历史信息通常被保留在上下文窗口中，模型并不天然具备独立于上下文的持久状态记忆机制。在自回归推理阶段，历史token对应的键和值通常存储在KV Cache中；每生成一个新的token，都需要与已有Cache中的信息进行注意力计算。

因此，LLM的长期依赖能力主要存在三个限制：第一，它依赖上下文窗口，信息一旦移出窗口，除非通过外部记忆、重新输入或Longformer等稀疏注意力长上下文方案[3]等方式补充，否则难以继续发挥作用；第二，它主要学习token之间的条件依赖关系，而不是显式的动作条件状态转移；第三，标准Transformer本身缺少独立于输入序列的持久状态，推理阶段的记忆主要来自当前上下文。

2.3.2　世界模型的增量式状态记忆与核心优势

与LLM的全局注意力记忆不同，世界模型通常可以采用增量式状态记忆。世界模型的核心思想是，将与未来预测相关的历史信息压缩到一个潜在状态中，这个状态承担模型内部记忆的作用。在每一个时刻，模型只需要根据当前观测和行动对状态进行增量更新，而无须重新处理整个历史序列。这种机制与马尔可夫决策过程（Markov Decision Process，MDP）的思想密切相关，在状态充分表征历史信息的条件下，未来状态可以由当前状态和行动进行预测。

循环状态空间模型（Recurrent State-Space Model，RSSM）是世界模型中常见的记忆结构，最早由PlaNet提出[10]。它把潜在状态分为确定性状态和随机状态：确定性状态由循环网络更新，用于

保留历史信息（如位置、速度以及已发生事件等）；随机状态通过概率建模和推断，用于表示当前状态中的随机性和不确定性。这种分离机制使RSSM能够同时兼顾记忆能力和不确定性建模。

LLM全局注意力记忆与基于RSSM结构的世界模型状态空间记忆的结构对比如图2-4所示。

世界模型的状态记忆具有三个实际优势。首先，压缩后的状态不再直接依赖完整的原始序列，系统可以保存状态，并在之后恢复推理。其次，它采用增量更新机制，每一步主要处理当前观测、动作和已有状态，因此对于基于循环状态空间（如RSSM）的世界模型来说，其计算量通常不会随历史长度持续增加。最后，状态更新通常遵循动作条件动力学，因而有助于维持时序一致性。这些特点使它更适合机器人和自动驾驶等需要长时间连续运行的实时系统。

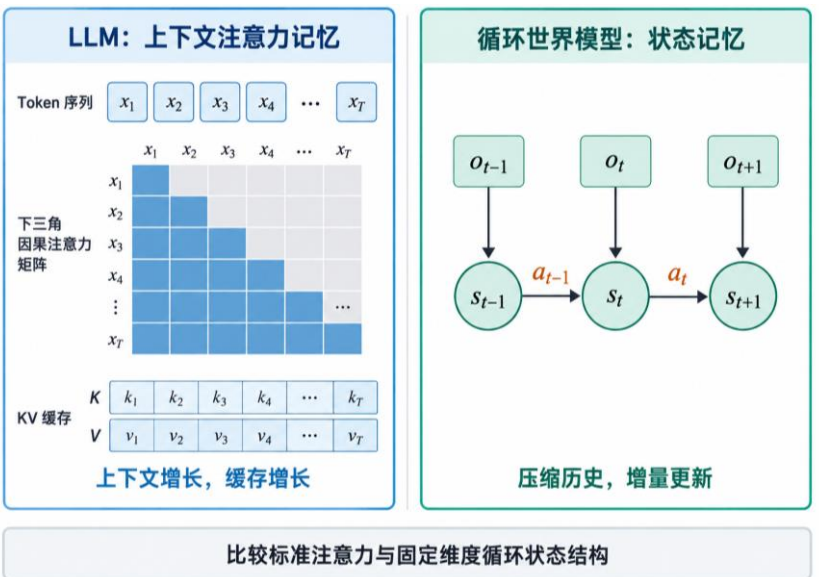


图 2-4　LLM 全局注意力记忆与世界模型状态空间记忆的结构对比

为了更直观地理解两种记忆机制的差异，可以对比它们在一个典型长序列任务中的表现。假设有一个机器人需要在大型办公楼中导航，从一楼大厅到十楼某间办公室。在这一过程中，机器人需要记住沿途路标、已走过的路线以及遇到的障碍物等信息，整个过程可能持续几十分钟，并产生大量帧视觉数据。对于仅依赖上下文窗口的LLM/VLM系统来说，如果需要持续保留大量历史观测，不仅会占用巨大的内存，还可能导致推理开销增加。当上下文窗口达到上限后，早期信息还需要通过压缩、外部记忆等方式保存。对于采用状态空间记忆的世界模型而言，则可以持续维护一个紧凑的潜在状态，在每一个时刻依据当前相机图像和控制指令更新状态，从而持续估计自身和环境状态。在状态维度固定的情况下，其单步状态更新开销通常不会随历史序列长度持续增加。

LLM与基于RSSM的世界模型的记忆结构与长期依赖能力对比如表2-2所示。

表 2-2　LLM 与基于 RSSM 的世界模型的记忆结构与长期依赖能力对比

对比项	LLM/VLM 全局注意力记忆	世界模型状态空间记忆
记忆载体	上下文窗口+KV Cache	潜在状态
更新方式	基于上下文进行注意力计算	增量式状态转移
计算复杂度	$O(n^2)$（自注意力）/$O(n)$（KV Cache）	$O(1)$（固定状态下单步开销相对稳定）
内存占用	通常随上下文长度增长	固定状态维度下相对稳定
长期依赖能力	受上下文窗口及有效利用能力限制	不直接受固定上下文窗口限制
状态一致性	长序列下可能出现前后不一致	通过状态转移维持时序一致性
实时性	长上下文计算开销增加	适合增量式实时更新
可中断性	通常需要保留上下文或 Cache	可保留状态并恢复

2.3.3　两种记忆机制的边界与技术演进趋势

世界模型的状态记忆也存在一定局限，主要问题是信息压缩带来的损失。由于历史信息被压缩到潜在状态中，一些细节可能会丢失。例如，模型可能记住房间里有一张桌子，却未必保留桌面上所有物品的细节。为缓解这一问题，研究人员提出了增加状态容量、引入外部记忆模块、采用分层状态表示等方法。其中，分层状态表示把状态拆分为高层语义状态和低层感知状态，分别编码长期抽象信息与短期细节信息，以平衡信息容量和计算效率。

近年来，LLM也在向状态化交互方向演进。从函数调用到实时多模态交互，系统层面开始引入外部状态、工具状态和增量更新机制。例如，思维树（ToT）等方法通过引入外部推理结构增强LLM的多步规划能力，体现了LLM向状态化、结构化交互演进的趋势[4]。这种演进趋势表明，状态记忆并非世界模型独有的特性，而是需要长期交互与持续决策的智能系统的共同需要。

然而，需要强调的是，LLM的状态化改造仍然建立在其原有的自回归生成框架之上，并未改变其基本的序列建模方式。而世界模型通常围绕状态空间和动作条件下的状态转移进行设计，它的记忆机制与决策和规划能力深度融合。这种原生的状态化设计，使世界模型在处理持续物理世界交互任务时具有天然优势。

长期依赖与记忆结构是世界模型与LLM/VLM之间的重要差异之一。全局注意力更适合处理给定上下文内的文本和图像任务；状态空间记忆则更适合持续运行的物理交互任务。选择哪种结构，取决于任务是否需要长期状态跟踪、低延迟更新和动作条件预测。

2.4　本章小结

本章从计算范式、表示方式和记忆结构三个维度对世界模型与LLM/VLM进行了比较。LLM/VLM主要学习符号或多模态表示之间的条件依赖关系，擅长语义生成与知识组织；世界模型则以动作条件下的状态转移为核心，关注环境如何随行动而发生变化，因此更接近可执行的环境模型。

这种差异并不意味着二者彼此替代。LLM/VLM可以提供语义规划、任务分解和人机交互接口，世界模型则更侧重环境预测、内部模拟和决策支持。明确二者的能力边界之后，第3章将进一步进入数学形式化部分，说明状态空间模型、POMDP、因果建模与反事实推理如何把这种差异表达为可训练、可推理、可规划的系统框架。

2.5　本章参考文献

[1] BROWN T, MANN B, RYDER N, 等. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[2] BROOKS R A. Intelligence without representation[J]. Artificial Intelligence, 1991, 47(1): 139-159. DOI:10.1016/0004-3702(91)90053-M.

[3] BELTAGY I, PETERS M E, COHAN A. Longformer: The Long-Document Transformer[Z]. arXiv, 2020. DOI:10.48550/arXiv.2004.05150.

[4] YAO S, YU D, ZHAO J, 等. Tree of thoughts: deliberate problem solving with large language models[C]. Advances in Neural Information Processing Systems, 2023, 36: 11809-11822. DOI:10.52202/075280-0517.

[5] GEMINI TEAM, ANIL R, BORGEAUD S, 等. Gemini: A Family of Highly Capable Multimodal Models[Z]. arXiv, 2023. DOI:10.48550/arXiv.2312.11805.

[6] SCHRITTWIESER J, ANTONOGLOU I, HUBERT T, 等. Mastering Atari, Go, chess and shogi by planning with a learned model[J]. Nature, 2020, 588(7839): 604-609. DOI:10.1038/s41586-020-03051-4.

[7] PEARL J. Causality: Models, Reasoning, and Inference (2nd ed.)[M]. Cambridge: Cambridge University Press, 2009.

[8] BRUCE J, DENNIS M, EDWARDS A, 等. Genie: Generative Interactive Environments[C]. Proceedings of the 41st International Conference on Machine Learning. PMLR, 2024, 235: 4603-4623.

[9] HAFNER D, LILLICRAP T, BA J, 等. Dream to Control: Learning Behaviors by Latent Imagination[C]. International Conference on Learning Representations (ICLR). 2020. DOI:10.48550/arXiv.1912.01603.

[10] HAFNER D, LILLICRAP T P, FISCHER I, 等. Learning Latent Dynamics for Planning from Pixels[C]. Proceedings of the 36th International Conference on Machine Learning. PMLR, 2019: 2555-2565. DOI:10.48550/arXiv.1811.04551.

[11] HAFNER D, LILLICRAP T, NOROUZI M, 等. Mastering Atari with Discrete World Models[C]. International Conference on Learning Representations (ICLR). 2021. DOI:10.48550/arXiv.2010.02193.

[12] VASWANI A, SHAZEER N, PARMAR N, 等. Attention is all you need[C]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008., 2017: 6000-6010. DOI:10.48550/arXiv.1706.03762.

第 3 章

世界模型的数学形式化

3



本章导读

第2章从能力边界上区分了世界模型与LLM/VLM：世界模型的核心不是单纯的统计续写，而是在行动条件下进行状态估计、动力学预测和规划评估。本章进一步从数学形式化的角度建立这些能力的基本语言。后续讨论中的潜在状态、部分可观测性、贝叶斯滤波、因果结构和反事实推演，都将在这里获得统一的定义。

本章首先从状态空间模型出发，说明如何将动态系统抽象为状态转移与观测过程；随后引入POMDP框架，解释智能体为何必须在不完整观测条件下维持信念状态；最后讨论因果模型与反事实推理，为后续表征学习、动力学建模、规划以及安全分析提供形式化基础。

3.1 状态空间模型

动态智能系统的重要能力之一，是对环境状态及其演化规律进行建模与预测，而这种能力需要建立在严谨的数学描述之上。状态空间模型（State Space Model，SSM）正是描述动态系统演化最常用的数学框架之一。它起源于20世纪60年代的控制论与信号处理领域，后来被广泛用于航空航天控制、导航定位、工业过程控制和现代机器学习模型。放到世界模型中来看，SSM给出了一条清晰的建模路径：先把传感器获得的高维信号表示为内部状态，再利用该状态预测未来演化并支持决策。因此，理解SSM的基本定义和性质，是后续讨论世界模型结构、训练目标和推理方法的起点。

3.1.1 状态空间模型的核心定义与马尔可夫性

状态空间模型的核心思想可以概括为一句话：用一个状态向量总结系统过去与未来演化相关的信息，使系统未来的演化主要依赖于当前状态和输入。这一思想的工程含义十分直接：系统不必保存所有观测细节，而只需保留那些决定未来行为的关键变量。在数学上，一个离散时间的状态空间模型由两个核心方程组成：状态转移方程和观测方程。状态转移方程描述系统状态如何随时间演

化，观测方程描述状态与可观测信号之间的映射关系。具体来说，对于任意时刻t，有：

$$x_{t+1} = f(x_t, a_t, w_t)$$

$$o_t = g(x_t, v_t)$$

其中，$x_t \in \mathbb{R}^d$是d维的状态向量，包含预测系统未来所需的相关信息；$a_t \in \mathbb{R}^m$是m维的控制输入或行动向量，表示智能体对系统施加的影响；$o_t \in \mathbb{R}^n$是n维的观测向量，表示智能体能够感知到的环境信号；w_t和v_t分别表示过程噪声和观测噪声，用于刻画系统中的随机性和不确定性。函数f称为动力学函数，用于描述状态的转移规律；函数g称为观测函数，用于描述从状态到观测的映射关系。

在满足马尔可夫假设时，状态x_t可以视为对历史信息的概括：给定当前状态x_t和后续输入后，未来状态$x_{t+1}, x_{t+2}, \cdots$和观测$o_{t+1}, o_{t+2}, \cdots$，与过去状态$x_0, x_1, \cdots, x_{t-1}$和观测$o_0, o_1, \cdots, o_{t-1}$之间不再需要通过完整历史直接关联（即条件独立）。这个性质就是马尔可夫性。它将具有复杂历史依赖的动态系统转化为一阶马尔可夫过程（即基于当前状态的递推过程），从而降低建模和推理的复杂度。在世界模型中，马尔可夫性意味着模型可以通过持续更新内部状态来概括历史信息，而无须在每一步重复处理全部历史观测，从而支持长序列的持续预测与决策。

状态空间模型的基本架构如图3-1所示。

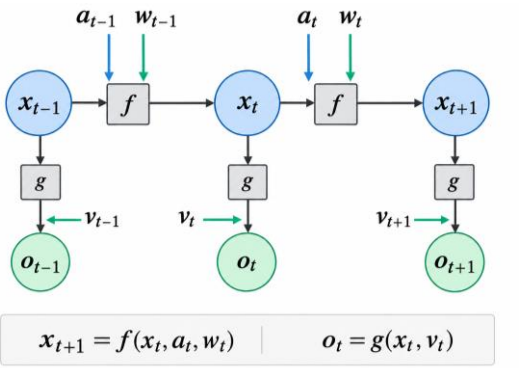


图 3-1　状态空间模型的基本架构

3.1.2　经典线性与非线性状态空间模型

根据动力学函数f和观测函数g的形式，可以将状态空间模型分为线性SSM和非线性SSM两大类。线性SSM是最简单、也是研究最为透彻的一类模型。它假设f和g均为线性函数，并且噪声w_t和v_t都是零均值高斯噪声。此时，状态空间模型可写为：

$$x_{t+1} = Ax_t + Ba_t + w_t$$

$$o_t = Cx_t + v_t$$

其中，A、B、C为固定的矩阵参数。线性SSM的重要优势在于，在线性高斯条件下存在解析的状态

估计算法——卡尔曼滤波（Kalman Filtering）[2]。卡尔曼滤波能够根据历史观测和动作序列，递归计算隐状态x_t的后验概率分布$p(x_t \mid o_1, \cdots, o_t, a_1, \cdots, a_{t-1})$。该后验分布仍是高斯分布，因此可由均值和协方差矩阵完全刻画。卡尔曼滤波被广泛应用于航空航天、导航定位和自动控制等领域，至今仍是许多工程系统中的重要状态估计方法。

然而，真实世界中的许多系统具有明显的非线性特征，线性假设往往只能在一定工作范围内成立。为了建模非线性系统，研究人员提出了多种近似推理方法，其中最经典的是扩展卡尔曼滤波（Extended Kalman Filter，EKF）和无迹卡尔曼滤波（Unscented Kalman Filter，UKF）。扩展卡尔曼滤波通过在当前状态点对非线性函数进行泰勒展开，将非线性模型局部线性化，然后应用卡尔曼滤波进行状态估计。无迹卡尔曼滤波通过选取一组称为西格玛点（Sigma Point）的确定性采样点，使这些样本点经过非线性函数变换后，近似传播状态分布的均值和协方差。这些方法在许多弱非线性系统中取得了不错的效果，但在强非线性和高维系统中，其近似误差可能会迅速增大，甚至导致滤波发散（即不稳定）。

粒子滤波是另一种非线性推理方法，它使用大量随机样本（粒子）表示后验分布，并通过重要性采样和重采样不断更新粒子的权重。粒子滤波能够处理较一般的非线性和非高斯分布，但它的计算复杂度随粒子数量线性增加，在高维状态空间中容易遭遇维度灾难问题。

线性与非线性状态空间模型的分布演化对比如图3-2所示。

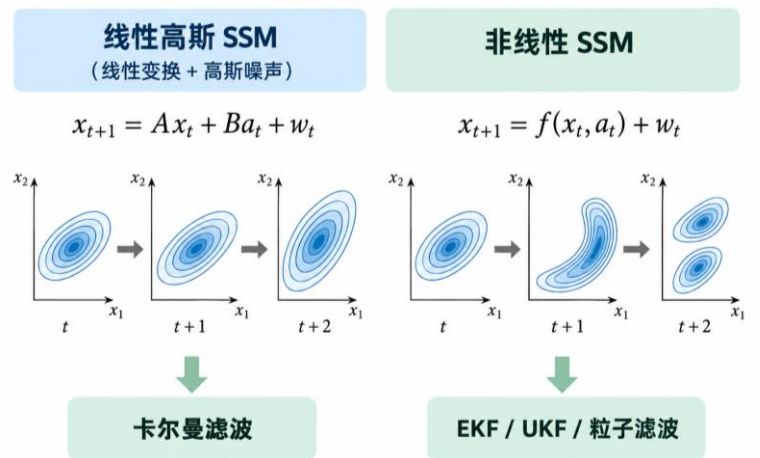


图 3-2　线性与非线性状态空间模型的分布演化对比

3.1.3　深度状态空间模型与世界模型范式

传统状态空间模型的一个重要局限性在于，其动力学函数和观测函数往往需要较强的人工建模假设，难以直接适应复杂多变的真实世界。深度学习的兴起为解决这一问题提供了新的思路：利用神经网络参数化f和g，并从大量数据中学习这些函数的参数。这种深度状态空间模型不仅保留了SSM的数学框架，同时增强了对非线性关系的建模能力和泛化能力。

在世界模型领域，深度SSM已经成为一种常见架构，其中具有代表性的是循环状态空间模型（Recurrent State-Space Model，RSSM）[15]，它也是Dreamer系列世界模型的核心。RSSM将隐状态分为确定性部分和随机性部分：确定性部分由循环神经网络（GRU）建模，用于保留历史信息；随机性部分通过概率分布建模，用于表示潜在状态的不确定性。这种混合架构结合了RNN的循环记忆和概率模型的不确定性建模能力，并被广泛用于基于潜在状态的环境动力学学习。

近年来，结构化状态空间模型（Structured State Space Model，S4）的提出[5]，为深度SSM的发展带来了重要突破。传统循环神经网络（RNN）在长序列建模中可能面临梯度传播困难；而Transformer的标准自注意力机制的计算复杂度随序列长度呈平方增长。S4模型通过对状态空间参数进行结构化设计，将递归形式转化为可高效计算的卷积形式，从而支持长序列的并行训练，其训练计算复杂度可以达到近线性或$O(n\log n)$量级。S4在语音识别、时间序列预测、视频建模等部分长序列任务上取得了与Transformer相当甚至更优的性能。在世界模型中，S4为长时序世界模型提供了潜在的架构选择，能够高效处理数千帧乃至数万帧的视频序列，也推动了后续结构化和选择性状态空间模型的发展。

S4结构化状态空间模型的核心计算单元如图3-3所示。

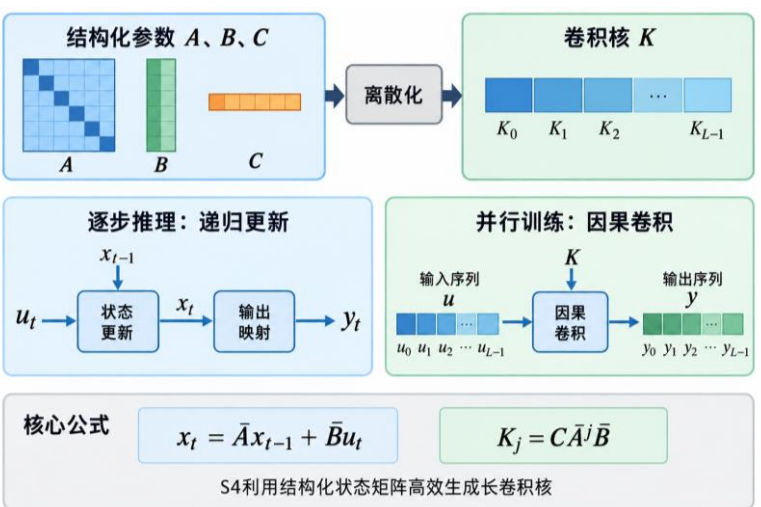


图 3-3　S4 结构化状态空间模型的核心计算单元

为了更清晰地展示SSM与其他主流序列模型之间的差异，表3-1对它们的核心特性进行对比。

表 3-1　SSM 与 RNN、Transformer 的核心特性对比

特性维度	状态空间模型/SSM	循环神经网络/RNN	Transformer
计算复杂度	依具体结构而定，部分模型为近线性，即$O(n\log n)$	通常为$O(n)$	标准自注意力通常为$O(n^2)$
长序列建模能力	较强	弱（梯度消失）	较强

（续表）

特性维度	状态空间模型/SSM	循环神经网络/RNN	Transformer
训练并行性	结构化 SSM 较好	较差（串行计算）	较好
推理速度	快（常数时间/步）	快（常数时间/步）	慢（随序列长度增长）
状态表示	状态和潜在状态	隐状态	隐式序列表示
不确定性建模	取决于具体模型	弱，通常不显式建模	弱，通常不显式建模
代表模型	S4, RSSM, Mamba	LSTM, GRU	GPT, BERT
世界模型应用	动力学建模、长时序预测	早期世界模型与递归动力学	大规模生成式世界模型

从表3-1可以看到，状态空间模型在世界模型中仍具有较强的实用价值：它能够以递推方式维持状态，并在部分结构中兼顾较高的计算效率和长序列建模能力。Transformer在表达能力和大规模预训练方面表现突出，但自注意力机制带来的计算开销以及隐式状态表示（内存开销），会限制它在长序列实时控制和资源受限场景中的应用。更现实的发展路线往往不是二选一，而是探索状态空间模型与注意力机制之间的融合。

采用SSM的语言重新描述世界模型，可以得到一个清晰的对应关系：编码器近似于观测模型g的逆映射，把高维观测o_t压缩为隐状态x_t；动力学模型对应状态转移函数f，根据当前状态和行动预测未来状态；解码器对应观测模型g，把隐状态映射回观测空间。整个系统通常以自监督学习训练，以重构损失、预测损失或表示学习等为目标，使潜在状态尽可能保留与未来演化相关的信息。

当然，标准状态空间模型也存在一些局限。首先，马尔可夫性假设要求当前状态包含概括预测未来所需的历史信息，而这一条件在复杂的长程依赖任务中并不容易满足。其次，非线性SSM的精确推理通常较为困难，现有各种近似方法都不可避免存在一定的误差。最后，标准SSM主要面向单模态时序数据，如何将多模态数据统一到同一状态空间中，如何有效处理视觉、语言和传感器等多模态数据，仍然是重要研究方向。

状态空间模型为动态系统建模提供了基础语言，也为世界模型提供了可计算的结构。从线性卡尔曼滤波到非线性深度SSM，再到结构化SSM，研究重点不断从解析推理转向高维数据中的可学习动力学。不过，标准SSM仍然隐含一个重要前提：系统状态需要能够被观测信息充分估计，而真实环境往往存在明显的部分可观测性。智能体通常只能通过有限的传感器观测环境的一部分信息，而无法直接获得完整状态。为了描述这种更加符合现实的情形，需要进一步引入部分可观测马尔可夫决策过程（Partially Observable Markov Decision Process，POMDP）。

3.2　POMDP 框架

在3.1节中，我们建立了状态空间模型（SSM）的数学基础，它为动态系统的演化提供了统一的描述框架。然而，标准SSM隐含着一个在真实世界中难以满足的强假设：系统状态可以被完整观

测或准确恢复。在现实场景中，智能体只能通过有限传感器获取局部信息，无法直接获取完整的系统状态，也无法知晓其他智能体的内部意图以及未被感知的环境变量。例如，自动驾驶汽车只能通过相机和激光雷达感知道路上的物体，无法直接知道被遮挡车辆的行驶意图；机器人只能通过触觉感知物体表面，无法直接获取物体的内部结构和质量分布。为了描述这种交互过程，需要引入部分可观测马尔可夫决策过程（POMDP）。POMDP是SSM与马尔可夫决策过程（MDP）的结合，它不仅描述了环境的动态演化，还考虑了智能体的决策过程和观测的不完全性（也就是将状态转移、部分观测和序列决策统一到同一框架中）。POMDP是世界模型常用的重要理论基础之一。

3.2.1　POMDP 的数学定义与核心要素

POMDP是用于描述智能体在部分可观测环境中进行序列决策的通用数学框架，可以用一个七元组来定义[1]：

$$\mathcal{M} = (S, A, T, R, O, Z, \gamma)$$

其中，每个元素都对应真实世界交互过程中的一个核心概念，下面逐一解释它们的含义和作用。

S是状态空间，它表示所有可能环境状态的集合，与SSM中的状态空间完全一致。每个状态$s \in S$包含预测环境未来演化所需的全部信息，并满足马尔可夫性。A是行动空间，它表示智能体所有可能采取行动的集合。在每个时刻t，智能体选择一个行动$a_t \in A$，该行动会作用于环境并影响环境状态。

T是状态转移函数，用于描述环境在行动作用下的动态演化规律，定义为条件概率分布：

$$T(s' \mid s, a)$$

表示在状态s下采取行动a后，环境转移到状态s'的概率。与SSM中的动力学函数f相比，T采用概率分布的形式，可以明确地刻画环境演化过程中的随机性。

R是奖励函数，可以定义为：

$$R(s, a, s')$$

表示智能体在状态s下采取行动a，并转移到状态s'后获得的即时奖励。奖励函数是智能体优化目标的数学表达，它告诉智能体什么是好的行为，什么是坏的行为，智能体的最终目标是最大化长期累积奖励的期望。

o是观测空间，它表示智能体所有可能接收到的观测集合。Z是观测函数，定义为条件概率分布：

$$Z(o \mid s')$$

表示当环境转移到状态s'时，智能体接收到观测o的概率。观测函数是POMDP区别于完全可观测MDP的重要特性：在完全可观测情形（MDP）下，观测函数是恒等函数，即$o = s'$，智能体能够

直接观测到真实状态；而在POMDP中，观测通常只提供关于真实状态的部分或带噪声信息，智能体无法直接获得s'，只能通过o推断s'的可能分布。

最后，$\gamma \in [0,1)$是折扣因子，用于表示未来奖励的现值系数（即重要程度）。通常来说，γ越接近1，智能体越重视长期奖励；γ越接近0，智能体越重视短期奖励。

POMDP的时序交互架构如图3-4所示。

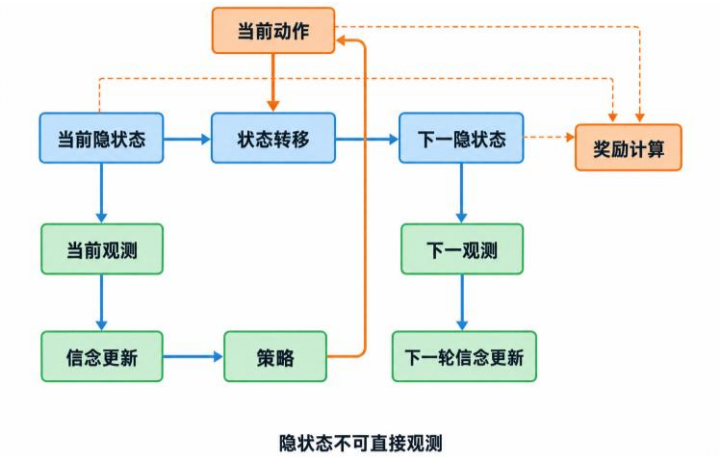


图 3-4　POMDP 的时序交互架构

为了更清晰地展示POMDP与MDP、SSM之间的关系，表3-2对它们的核心特性进行对比。

表 3-2　MDP、SSM 与 POMDP 的核心特性对比

模　型	状态可观测性	决策能力	核心目标	典型应用
MDP	完全可观测	有	最大化累积奖励	棋盘游戏、简单控制
SSM	完全可观测	无（本身不含决策优化）	建模与预测系统状态演化	信号处理、导航、天气预报、控制
POMDP	部分可观测	有	基于不完全观测最大化累积奖励	自动驾驶、机器人、对话系统

从表3-2可以看出，POMDP同时覆盖了不完全观测和序列决策两个核心问题。MDP可以看作POMDP在完全可观测条件下的特例；而SSM与POMDP在状态转移和观测建模上具有相近结构，但二者的建模目标并不完全相同。当观测函数为恒等函数时，POMDP退化为MDP；当不考虑动作和奖励时，POMDP退化为SSM。世界模型可以理解为一种从交互数据中学习环境动态及其相关预测模型的方法，在POMDP语境下，它可以学习状态转移、观测和奖励等模型，并在此基础上支持规划和决策。

3.2.2　信念状态与贝叶斯滤波

在完全可观测的MDP中，智能体的决策只依赖于当前真实状态s_t。但在POMDP中，智能体无法直接观测到s_t，只能获得一系列历史观测和行动[3]：

$$h_t = (o_1, a_1, o_2, a_2, \ldots, o_t)$$

因此，智能体的决策不能基于未知的真实状态，而必须基于它对当前状态的信念（Belief）。

信念状态b_t是定义在状态空间S上的概率分布，$b_t(s)$表示在给定历史h_t的条件下，当前环境处于状态s的后验概率：

$$b_t(s) = P(s_t = s \mid h_t)$$

信念状态是智能体对环境状态的概率估计，汇总了历史观测和行动中与当前状态有关的信息。它的重要性质在于自身满足马尔可夫性，即信念状态可以作为POMDP决策过程的充分统计量：给定当前的信念状态b_t，未来所有信念$b_{t+1}, b_{t+2}, \cdots$与过去的历史h_t条件独立。也可以这么理解：给定当前信念状态b_t和后续行动、观测后，后续信念状态可以递归更新，而无须重新保留完整历史。由此，可以将POMDP转化为以信念状态为状态的完全可观测MDP，即信念MDP。这是POMDP理论中的重要结论之一，使得MDP中的许多求解方法可以应用于POMDP问题。

信念状态的更新过程通常通过贝叶斯滤波实现，它是POMDP中状态估计的核心过程。贝叶斯滤波分为两个步骤：预测步和更新步。

在预测步中，智能体根据上一时刻的信念b_t和采取的行动a_t，预测下一时刻的先验信念$\bar{b}_{t+1}$：

$$\bar{b}_{t+1}(s') = \sum_{s \in S} b_t(s) T(s' \mid s, a_t)$$

这个公式表示，对于每一个可能的下一状态s'，其先验概率等于所有可能当前状态s的信念概率与从s转移到s'的概率乘积之和。

在更新步中，当智能体接收到新的观测o_{t+1}后，利用观测函数Z修正先验信念，得到后验信念b_{t+1}：

$$b_{t+1}(s') = \eta Z(o_{t+1} \mid s') \bar{b}_{t+1}(s')$$

其中，η是归一化常数，用于保证b_{t+1}是合法的概率分布，即：

$$\sum_{s' \in S} b_{t+1}(s') = 1$$

贝叶斯滤波是一个递归过程，它从初始信念b_0开始，在每个时刻重复执行预测步和更新步，不断修正智能体对环境状态的估计。对于线性高斯POMDP，贝叶斯滤波具有解析的递归形式，即3.1节中介绍的卡尔曼滤波。对于非线性或非高斯POMDP，则需要采用近似方法，例如扩展卡尔曼滤波、无迹卡尔曼滤波以及粒子滤波。其中，粒子滤波是较通用的方法之一，它利用一组随机样本

（粒子）近似表示信念分布，通过重要性采样和重采样更新粒子的权重，能够处理较一般的非线性和非高斯分布。

贝叶斯滤波的信念更新流程如图3-5所示。

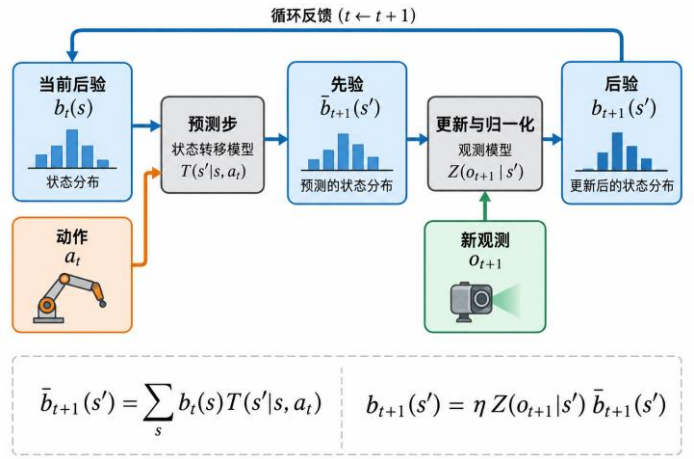


图 3-5　贝叶斯滤波的信念更新流程

3.2.3　POMDP 与世界模型的范式对应

POMDP不仅是一个抽象的数学框架，也是世界模型技术的重要理论基础。各种世界模型架构，从基于RNN的RSSM到基于Transformer的Genie，都可以在一定程度上用POMDP中的状态、观测、动作和转移等概念加以理解，其目标之一是解决部分可观测条件下的状态估计与决策问题。因此，可以将世界模型的各个模块与POMDP中的核心要素建立对应关系。

首先，世界模型的编码器对应于POMDP的观测模型Z的逆映射。编码器接收高维原始观测o_t，映射为内部状态表示，该状态表示可视为当前环境状态或信念状态b_t的参数化近似。例如，在RSSM中，编码器首先提取观测特征，并结合历史状态形成对潜在状态的后验估计；而在基于Transformer的世界模型中，观测通常被编码为一系列token或特征表示，用于表征当前环境信息。

其次，世界模型的动力学模型对应于POMDP的状态转移函数T。动力学模型接收当前信念b_t和行动a_t，预测下一时刻的信念b_{t+1}。这一预测过程对应于贝叶斯滤波中的预测步骤，即在没有新观测输入时，仅依靠当前状态和动作对未来进行推演。在世界模型中，动力学模型是实现想象能力的重要组成部分，使智能体能够在内部模拟环境演化过程，并预演不同行动可能产生的后果。

最后，世界模型中的奖励模型对应于POMDP的奖励函数R。奖励模型接收当前信念b_t和行动a_t，预测未来的即时奖励和长期累积奖励。奖励模型使智能体能够在想象轨迹上评估不同行动的价值，从而选择更优的动作序列。

POMDP组件与世界模型模块的对应关系如图3-6所示。

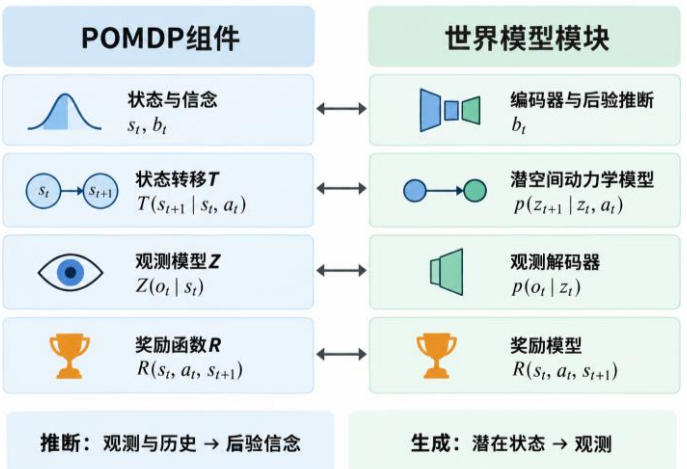


图 3-6　POMDP 组件与世界模型模块的对应关系

传统POMDP方法面临的主要瓶颈之一是维度灾难。当状态空间和观测空间的维度较高时，信念状态的表示和更新在计算上往往变得不可行。世界模型通过深度学习方法缓解了这一问题：利用神经网络对信念状态表示、状态转移和观测过程进行参数化，将高维原始观测压缩为更紧凑的低维隐状态表示，从而降低后续预测与规划的计算复杂度。例如，一幅256×256的RGB图像包含196 608个像素通道值，而世界模型可以将其压缩为一个256维的隐状态向量，并尽量保留与预测和决策相关的信息。

Dreamer系列可视为POMDP思想在深度世界模型中的一种典型实现。其核心组件RSSM把信念状态划分为确定性部分和随机性部分，分别表示环境中相对稳定的历史信息和具有不确定性的未来变化。Dreamer的想象训练过程，可以理解为在信念MDP中进行策略学习：智能体在内部状态空间中生成想象轨迹，并利用这些轨迹训练策略网络-价值网络（Actor-Critic）。这样可以减少对真实环境交互数据的依赖，并降低部分真实交互所带来的成本与风险。

POMDP为世界模型提供了清晰的理论接口：它定义了状态、动作、观测、奖励和信念更新，并解释了为什么在部分可观测环境中需要维护内部状态。不过，POMDP主要描述智能体如何在不完全信息条件下更新信念并选择动作，并不能直接解释环境变量之间为何会以特定方式相互作用。要进一步讨论主动干预和反事实推演等问题，还需要引入因果结构。

3.3　因果结构与干预视角

前两节分别给出了SSM对动态系统的数学表达，以及POMDP对“智能体-环境”交互的决策框架。二者都以历史交互数据为基础，学习变量之间的概率关系。这类关联建模可以很好地拟合训

练数据中的观测分布，但在具身决策中仍然存在局限：当智能体执行训练集中较少出现的动作，或环境进入训练数据未覆盖的状态时，仅依赖关联统计难以准确预测未来状态如何演化。因果结构与干预视角，正是为分析和缓解这一类问题提供的重要理论工具。

3.3.1　从关联预测到因果建模：世界模型的核心刚需

关联统计的核心局限在于其主要捕捉变量之间的观测相关性，而无法区分因果关系与伪相关。传统SSM的状态转移建模，本质上是学习条件概率分布：

$$P(s_{t+1} \mid s_t, a_t)$$

即基于当前状态与动作预测下一状态的概率分布。这种建模方式能够拟合训练数据中的状态转移，但并不意味着模型已经获得了环境变量之间的真实因果机制。

例如，在机器人抓取任务中，若训练数据均来自白天光照条件下的抓取样本，关联模型可能将光照亮度与抓取成功率建立强关联；但光照并非抓取成功的根本原因。当测试场景中的光照发生变化时，这种基于相关性的预测可能出现偏差。这类伪相关会削弱世界模型在长时序仿真中的泛化能力：当智能体执行的动作序列偏离训练数据分布时，关联模型的预测误差会随着时间步长累积，最终导致内部模拟结果失真。

要进一步理解这一问题，可以引入朱迪亚·珀尔提出的因果阶梯理论[8]，将世界模型的建模能力从最低的关联层提升至干预层与反事实层。因果阶梯的三个层级分别对应世界模型的三类核心能力：

第一层级：关联（seeing），对应传统生成式模型的观测预测能力，仅能基于已有观测数据生成符合分布的内容，这也是当前多数LLM/VLM（大语言模型/视觉语言模型）的能力边界。

第二层级：干预（doing），是世界模型的核心门槛，对应智能体通过主动动作改变环境，并准确预测干预后环境状态变化的能力，其数学表达通常通过do算子描述。

第三层级：反事实（imagining），对应高阶决策能力，可回答“如果当初执行了另一个动作会发生什么”问题的能力，为智能体的规划与决策提供反事实推理的理论基础。

世界模型要成为可交互的环境模拟器，关键不仅在于复刻历史数据中的统计规律，更在于能否对智能体的主动干预给予合理响应。本章将在数学形式化层面介绍因果结构、do算子与反事实问题的基本概念。

3.3.2　因果结构的形式化表达与状态空间的融合

结构因果模型（Structural Causal Model，SCM）是因果建模的核心数学框架，可表示为四元组：

$$M = \langle U, V, F, P(U) \rangle$$

其中，U为外生变量集合，对应系统中未被观测、不受模型内其他变量影响的随机因素；V为内生变量集合，对应模型中可观测或可控制的系统状态、动作、观测等变量；F为结构方程集合，每个

内生变量$v \in V$对应一个结构方程：

$$v = f_v(pa_v, u_v)$$

其中，pa_v表示变量v在因果图中的父节点，$u_v \in U$为对应v的外生噪声；$P(U)$表示外生变量的联合概率分布。

与SCM对应的因果图通常表示为有向无环图（Directed Acyclic Graph，DAG），每个节点对应V中的一个内生变量，有向边从父节点pa_v指向子节点v，表示变量之间的直接因果关系。

干预的数学表达由do算子定义：

$$P(Y \mid do\ (X{=}x))$$

表示通过外部干预将变量X设置为x后，结果变量Y的概率分布。它不同于条件概率：

$$P(Y \mid X{=}x)$$

条件概率只描述在观测数据中变量之间的统计关系，在观测数据中筛选满足$X{=}x$的样本，并不改变变量之间的因果结构；do算子则表示对变量施加外部控制，切断X与其父节点之间的因果边，会改变原有因果机制中的对应关系。将这一思想引入状态空间模型，就得到具有因果解释能力的因果状态空间模型（Causal SSM），从而为世界模型的状态转移赋予明确的因果语义。

具体而言，将SSM中的时序状态变量s_t、动作变量a_t和观测变量o_t纳入SCM的内生变量集合V，将环境中的未观测扰动和噪声因素纳入外生变量集合U。传统SSM的状态转移方程被重构为因果结构方程：

$$s_{t+1} = f_s(s_t, a_t, u_t)$$

其中，动作a_t作为干预变量（外部控制变量），描述智能体对环境施加的干预，由do算子进行外部赋值，直接驱动下一状态s_{t+1}的生成；观测过程可以表示为：

$$o_t = f_o(s_t, u_{o,t})$$

f_o描述潜在状态到观测信号之间的生成关系。

因果状态空间模型既继承了SSM的时序动态建模能力，同时通过显式因果结构明确了变量关系的解释能力，弥补了传统SSM中状态转移仅能拟合关联分布的不足。二者的核心差异如表3-3所示。

表 3-3　关联状态空间模型与因果状态空间模型的核心差异

维　　度	关联状态空间模型	因果状态空间模型
建模目标	拟合历史数据中的状态转移规律	建模具有干预解释能力的状态关系
状态转移本质	关联拟合	条件概率$P(s_{t+1} \mid s_t, a_t)$
核心算子	条件概率与贝叶斯推断	do算子与因果效应分析
分布外泛化能力	弱，依赖训练分布覆盖程度	强，理论上可以提升干预场景下的泛化能力
核心误差来源	分布偏移与伪相关	未观测混杂与因果结构误判

为直观呈现因果结构与时序动态的融合关系，因果状态空间模型的时序DAG结构如图3-7所示。

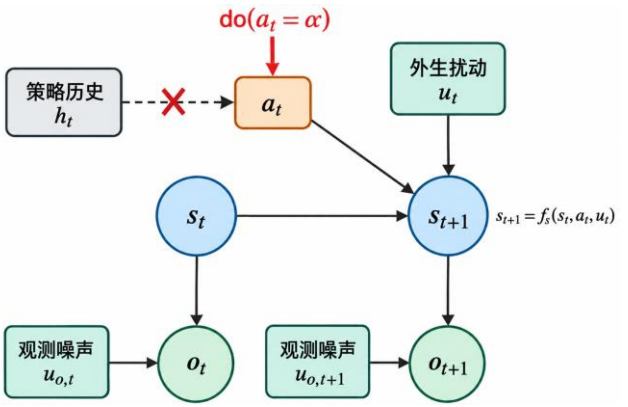


图 3-7　因果状态空间模型的时序 DAG 结构图

在此基础上，可以利用因果图中的后门准则和前门准则，对动作干预效应进行分析。以后门准则为例，当存在混杂变量同时影响动作a_t与下一状态s_{t+1}时，只需调整满足后门准则的变量集合，即可阻断混杂变量带来的伪相关，从而无偏估计动作与状态变化之间的因果效应，为3.2节POMDP框架中的部分可观测问题提供因果层面的解决方案。

同时，因果马尔可夫性的引入进一步强化了状态空间的马尔可夫假设（强化了因果结构与状态表示）：给定变量的父节点后，该变量独立于所有非后代节点，从而确保潜在状态s_t可以完整捕获所有对未来有因果影响的历史信息，并减少传统SSM中潜空间包含冗余关联信息的问题。这一性质为设计具有因果解释性的状态表示提供了理论依据。

3.3.3　因果干预的世界模型求解与能力边界

基于因果结构增强的世界模型，其核心求解目标之一是估计干预条件下的状态转移分布：

$$P(s_{t+1} \mid s_t, \mathrm{do}(a_t))$$

并据此实现长时序的干预模拟与决策优化。当前主流的求解路径可分为两类，分别适用于不同的世界模型应用场景。

- 第一类是显式因果图建模。该路径通过因果发现[6]或专家知识约束等方式，尝试从智能体与环境的交互数据中学习内生变量之间的因果DAG结构，并基于DAG构建对应的结构方程，实现对干预效应的分析。这类方法具有较强的可解释性，适用于变量定义较明确、可观测、结构相对稳定的场景，如工业控制、机器人操作和部分自动驾驶任务。
- 第二类是隐式因果表示学习[4]。该路径不显式构建完整因果DAG，而是通过环境不变性学习，表征解耦等方法，使模型学习在不同环境条件和干预变化下保持状态转移预测的一致性，从而隐式学习变量之间的因果关系。这类方法更适合处理高维、非结构化输入数据（如

视觉图像、多模态传感信号），也是当前通用世界模型（如IRIS、Genie）的主流建模思路。

因果干预建模也有助于缓解长时序预测中的误差累积问题。传统关联模型一旦依赖伪相关，当动作序列偏离训练分布后，预测误差便会在多步推演中逐渐放大。因果模型试图学习更加稳定的环境机制，因此在环境变化或动作分布变化条件下更不容易产生系统性偏差。不过，这种优势依赖于因果结构假设的正确性以及训练数据对相关机制的充分覆盖，并不能简单理解为能够自动消除所有预测误差。

与此同时，还需要明确因果建模在世界模型中的能力边界。

其一，因果效应的无偏估计通常依赖于无未观测混杂变量这一强假设。当存在同时影响动作与状态的未观测因素时，即使建立了因果模型，也可能难以准确估计干预效应，这正是POMDP框架中部分可观测性带来的核心挑战。这与POMDP中的部分可观测问题相关，但二者关注角度不同：POMDP强调状态不可直接获得，而因果建模关注变量关系是否能够被正确识别。

其二，因果建模的泛化能力通常建立在因果机制相对平稳的假设之上。当环境的底层因果结构发生突变（如物理规则改变、交互对象行为逻辑突变）时，模型仍可能出现预测失效。因此，如何实现因果结构的持续学习与动态自适应更新，仍是开放问题。

因果结构与干预视角弥补了关联统计在解释动作影响方面的不足，使世界模型能够进一步讨论动作、状态变化与反事实结果之间的关系。但在部分可观测环境中，状态估计、未观测混杂因素和长期规划仍然构成核心挑战。要把这一因果状态空间模型进一步用于长期决策，还需要讨论POMDP求解方法。

3.4　POMDP 求解器

前面几节已经给出了三层基础：SSM描述动态系统，POMDP描述部分可观测环境下的序列决策，因果结构刻画动作干预及其泛化边界。接下来的问题是如何进行求解。传统POMDP算法通常依赖显式信念表示和较强的状态建模假设，难以直接处理具身智能、自动驾驶等场景中的高维连续输入。世界模型的作用在于利用潜空间表征和学习得到的动力学模型，把高维POMDP转化为更易于计算的近似求解问题。

3.4.1　POMDP 求解的核心困境与范式分类

POMDP求解的目标是寻找一个策略π，使智能体在部分可观测随机环境中的长期折扣奖励期望最大化。由于智能体无法直接获得真实状态s_t，只能根据观测o_t和历史“动作-观测序列”估计信念状态：

$$b_t(s) = P(s_t = s \mid h_t)$$

其中：

$$h_t = (o_1, a_1, o_2, a_2, \dots, o_{t-1}, a_{t-1}, o_t)$$

求解难点主要来自以下三个方面。

1. 信念状态的维度灾难

信念状态是定义在整个状态空间S上的概率分布。对于离散状态空间，若状态数为：

$$n = |S|$$

则信念空间位于$\mathbb{R}^n$ 中的$(n-1)$维概率单纯形，并满足归一化约束：

$$\sum_{i=1}^{n} b(s_i) = 1$$

随着状态空间规模增加，信念空间维度快速增长。对于连续状态空间，精确贝叶斯更新通常涉及高维积分或复杂概率表示，其计算成本会随状态维度和模型复杂度呈指数增长，因此传统算法难以直接应用于大规模问题求解。

2. 部分可观测的信息缺失与误差累积

智能体的决策依赖于信念状态，而信念状态更新受到观测噪声、模型误差以及历史信息不完备性的影响。在长时序决策过程中，每一步状态估计误差可能逐步累积，使信念状态逐渐偏离真实环境状态，最终导致策略失效。这一问题与世界模型长时序预测中的误差累积具有一定相似性，也是传统POMDP方法难以直接用于长时序任务的重要原因。针对部分可观测条件下的价值估计难题，Hausknecht与Stone提出了深度循环Q网络（DRQN），利用循环结构聚合历史观测[11]。

3. 稀疏奖励下的长时序信用分配困境

真实世界中的许多决策任务具有奖励稀疏的特性，动作产生的正向或负向反馈可能需要数十步甚至数百步之后才能体现。传统强化学习方法需要在部分可观测的条件下[10]，将最终奖励归因于历史动作，因而容易面临策略学习效率低和收敛困难的问题。

POMDP求解算法通常可分为两类，其主要特性与适用场景见表3-4。

第一类是精确求解算法，其目标是在有限离散POMDP中求得全局最优策略，具有严格的理论最优性保障，但计算复杂度较高，通常只适用于小规模问题。

第二类是近似求解算法，通过近似信念空间、价值函数或策略空间，在可接受计算成本下获得性能较优的策略。世界模型属于基于模型的近似求解路线：它通过学习环境动力学，把高维观测空间中的POMDP转化为低维潜空间中的近似信念决策问题——MDP（Markov Decision Process，马尔可夫决策过程）[9]。

表 3-4　POMDP 求解核心范式的特性对比

求解算法	核心思想	状态空间适配	样本效率	最优性保障	核心适用场景
精确求解算法	显式表示信念空间，求解全局最优价值函数或策略	小规模离散空间	低	严格全局最优	理论分析、小规模工业控制场景
无模型近似求解算法	直接从交互数据中学习策略，不显式建立环境模型	中高维连续空间	依赖数据规模	局部最优，通常无严格保证	游戏 AI、机器人控制等密集奖励的决策任务
世界模型驱动的基于模型求解	学习潜在环境动力学，在内部模型中进行预测和规划	高维连续多模态空间	通常较高	近似最优或经验有效	具身智能、自动驾驶、长时序智能体

3.4.2　经典 POMDP 求解算法的核心框架

经典POMDP求解算法的发展始终围绕如何突破维度灾难，在更大规模的问题中实现有效求解这一核心命题展开。下面将从精确求解算法的理论基础出发，逐步延伸到适用于真实场景的近似求解算法，梳理各类算法的核心思想及其适用边界。

1. 精确求解算法：理论基础与局限

精确求解算法的核心前提是POMDP的状态、动作和观测空间均为有限离散集合，其理论基础在于：对于有限离散POMDP，最优价值函数关于信念状态是分段线性凸函数，可由有限个$\boldsymbol{\alpha}$向量组成的集合表示。

基于这一理论，学术界提出了两类经典的精确求解方法：价值迭代与策略迭代。

价值迭代的核心思想是通过迭代备份$\boldsymbol{\alpha}$向量集合，逐步逼近最优价值函数。每一步迭代均基于当前的$\boldsymbol{\alpha}$向量集合，计算各信念点上的最优价值，并生成新的$\boldsymbol{\alpha}$向量集合；策略迭代则包括策略评估与策略改进两个阶段，首先评估当前策略对应的价值函数，再依据评估结果优化策略，直至满足策略收敛条件。

两类算法均具有严格的理论最优性保证（即最优性收敛），但适用于状态数不超过1000的小规模问题。随着状态空间规模扩大，$\boldsymbol{\alpha}$向量的数量将呈指数级增长，计算成本完全不可控。在世界模型所面向的高维真实场景中，精确求解算法更多作为理论基础，而难以直接应用于工程实践。

2. 点基近似求解算法：信念空间的稀疏近似

为突破精确求解算法的规模限制，点基价值迭代（Point-Based Value Iteration，PBVI）算法被提出[12]。其核心创新思想是放弃对整个信念空间的精确表示，仅利用有限个具有代表性的信念点近似整个信念空间，并在这些采样信念点上进行价值函数的更新与备份，从而大幅降低计算复杂度。

PBVI算法的核心流程分为三个步骤：

首先，通过随机采样或启发式方法，在信念空间中生成具有代表性的信念点集合。

其次，在各信念点上执行价值函数备份，更新$\boldsymbol{\alpha}$向量集合。

最后，根据更新后的价值函数提取对应策略。

PBVI算法将POMDP求解的适用规模提升了一个数量级，成为中等规模离散或低维连续POMDP问题的重要近似求解方法。然而，在世界模型面对的高维视觉输入和复杂连续控制场景中，信念点的表示、采样与更新仍然具有较高计算成本，同时难以直接处理大规模连续状态与动作空间，因此存在明显的应用局限。

3. 策略搜索与采样式求解算法：高维空间的实用化突破

针对高维连续空间中的决策需求，学术界提出了策略搜索类方法。其核心思想是放弃对完整价值函数或信念空间的显式表示，直接对策略：

$$\pi_\theta(a_t \mid h_t)$$

进行参数化，并通过策略梯度等方法优化策略参数θ，最大化长期累积奖励期望[13]。

这类算法的代表包括REINFORCE算法（REINFORCE Algorithm，REINFORCE）和Actor-Critic（策略网络−价值网络）框架，其优势在于能够自然适配连续动作空间，并避免显式维护高维信念状态，因此在复杂控制任务中得到广泛应用。

与此同时，基于蒙特卡洛采样的启发式求解算法（如蒙特卡洛树搜索，Monte Carlo Tree Search，MCTS）也被广泛应用于序列决策问题。其核心思想是通过大量随机采样近似估计信念状态的价值与动作的收益，无须显式建立环境动力学模型，也无须完整表示信念空间，因此能够适用于大规模复杂决策任务。

然而，这类传统无模型策略优化方法与采样式搜索算法仍面临样本效率问题。它们通常需要大量环境交互数据才能学习策略（即与真实环境进行数百万甚至数千万次的交互才能收敛）。对于机器人、自动驾驶等真实交互成本较高的场景，直接应用存在巨大挑战。

因此，传统近似求解方法虽然扩大了POMDP的适用范围，但仍难以同时满足高维连续空间、长期规划和高样本效率的需求。世界模型的核心优势正在于利用学习得到的环境动力学，在内部模型中进行预测和规划，从而大幅减少真实环境的交互需求。

3.4.3　世界模型驱动的 POMDP 求解器架构

世界模型求解POMDP的基本思路，是通过潜空间表征学习[7]，将高维观测空间中的POMDP近似转化为低维潜空间中的信念MDP，从而降低直接在原始观测空间进行状态估计和规划的计算负担。在SSM框架下，编码器把高维观测o_t映射为潜在状态s_t，状态转移模型则学习潜空间中的动力学规律。由此，部分信念状态的表示、未来预测和规划过程均可在潜空间内完成，其计算成本远低于直接处理原始观测。

世界模型驱动的POMDP近似求解架构，可以看作SSM潜空间建模、因果动力学预测与近似规划或策略优化方法的有机组合。其整体架构如图3-8所示，主要包含以下4个模块。

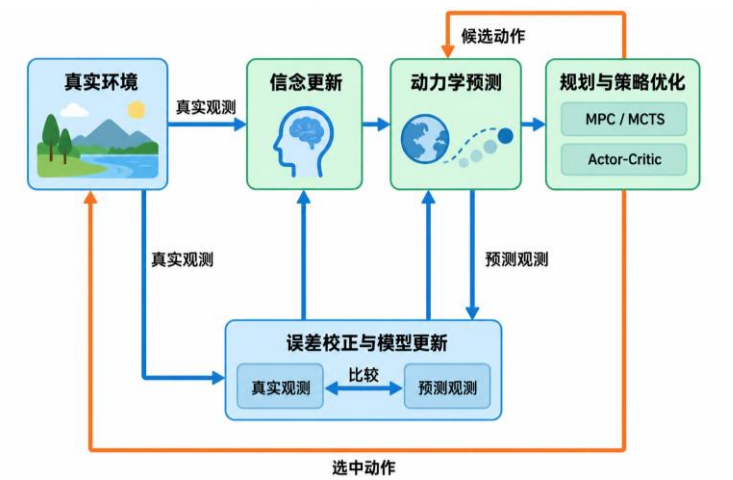


图 3-8　世界模型驱动的 POMDP 求解器整体架构图

1. 潜空间信念更新模块

该模块对应SSM中的滤波过程，负责依据历史“动作–观测”序列更新潜空间中的信念状态，即潜状态的后验分布。与传统POMDP需要显式表示和更新高维信念状态不同，神经网络求解器可以利用RNN或Transformer直接把历史序列映射为高斯分布或离散分布的参数，从而适应高维多模态输入条件下的状态估计。引入因果结构建模后，还可以减弱潜空间中伪相关因素的影响，提高信念更新的稳定性。

2. 因果动力学预测模块

该模块负责根据当前潜空间信念状态和候选动作预测未来状态、观测及奖励分布。若进一步引入do算子，可以把动作显式表示为干预变量，从而分析不同动作干预对未来状态的影响。这样，智能体便能够在内部模型中评估多条候选轨迹，从而减少对真实环境试错的依赖。

3. 规划与策略优化模块

该模块是求解器的决策核心，基于动力学模块生成的虚拟仿真结果，实现策略优化与动作选择。当前主流实现路径分为两类：一类是梦境规划范式，以Dreamer系列算法为代表[14]，在潜空间构建的虚拟环境中完成策略学习，并通过策略梯度算法优化“策略网络–价值网络”框架，直接输出端到端的连续控制策略，从而提高真实交互数据的利用效率；另一类是启发式搜索范式，将世界模型的动力学预测与MCTS、MPC等规划算法结合，通过虚拟仿真完成树搜索或轨迹优化，最终输出较优动作或动作序列，适用于自动驾驶、复杂游戏等具有“离散–连续”混合动作空间的场景。

4. 误差校正与模型更新模块

该模块负责监控内部仿真与真实环境之间的偏差，并在偏差过大时校正信念状态或更新模型参数。常见方法包括状态重初始化、增量学习以及保守策略等。它的作用并非保证模型预测始终正

确，而是尽可能减弱误差累积对后续决策的影响。

在工程实现中，需要重点关注三类问题：一是数值稳定性，避免潜空间分布坍缩以及信念更新过程中的数值不稳定；二是计算效率，确保信念更新、动力学预测和规划过程满足实时任务的延迟要求；三是分布外泛化能力，使求解器在训练数据未充分覆盖的场景中仍具有一定的稳定性。

POMDP求解器把前述数学形式化框架进一步转化为可实现的系统结构。SSM提供状态表示，POMDP提供决策语义，因果结构提供干预解释，而世界模型则利用潜空间动力学把这些组成部分有机连接起来。由此也能看出，世界模型与LLM/VLM的差异并不只是体现在生成形式上，更体现在前者更加专注行动条件下的状态估计、未来预测与决策规划能力。

3.5　本章小结

本章建立了世界模型的数学基础。SSM利用隐状态描述动态系统的时间演化，POMDP进一步把部分可观测环境中的动作、奖励和信念状态纳入统一决策框架，因果结构则补充了干预分析与反事实推理所需的变量关系。由此，世界模型可以理解为一种在潜空间中完成状态估计、动力学预测和策略评估的近似POMDP求解器。

本章的核心结论是：世界模型并非仅用于预测下一时刻的观测，而是在不确定观测条件下维护可用于决策的内部状态，并评估动作对未来状态的影响。第4章将进一步转向表征学习，讨论这些状态变量如何从高维观测中被压缩、离散化，并组织成可预测、可规划的潜空间表示。

3.6　本章参考文献

[1] KAELBLING L P, LITTMAN M L, CASSANDRA A R. Planning and acting in partially observable stochastic domains[J]. Artificial Intelligence, 1998, 101(1-2): 99-134.

[2] KALMAN R E. A new approach to linear filtering and prediction problems[J]. Journal of Basic Engineering, 1960, 82(1): 35-45.

[3] THRUN S, BURGARD W, FOX D. Probabilistic robotics[J]. Communications of the ACM, 2002, 45(3): 52-57.

[4] SCHÖLKOPF B, LOCATELLO F, BAUER S, 等. Toward Causal Representation Learning[J]. Proceedings of the IEEE, 2021, 109(5): 612-634. DOI:10.1109/JPROC.2021.3058954.

[5] GU A, GOEL K, RE C. Efficiently Modeling Long Sequences with Structured State Spaces[C]. International Conference on Learning Representations(ICLR). 2022. DOI:10.48550/arXiv.2111.00396.

[6] GLYMOUR C, ZHANG K, SPIRTES P. Review of Causal Discovery Methods Based on Graphical Models[J]. Frontiers in Genetics, 2019, 10:524. DOI:10.3389/fgene.2019.00524.

[7] HA D, SCHMIDHUBER J. World Models[Z]. arXiv, 2018. DOI:10.48550/arXiv.1803.10122.

[8] PEARL J. Causality: Models, Reasoning, and Inference (2nd ed.)[M]. Cambridge: Cambridge University Press, 2009.

[9] SIGAUD O, BUFFET O. Markov decision processes in artificial intelligence[M]. Hoboken, NJ: John Wiley & Sons, 2010.

[10] SUTTON R S, BARTO A G. Reinforcement Learning: An Introduction[M].2nd版. Cambridge, MA: The MIT Press, 2018.

[11] HAUSKNECHT M, STONE P. Deep Recurrent Q-Learning for Partially Observable MDPs [C]. AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents, 2015.DOI:10.48550/arXiv. 1507.06527.

[12] PINEAU J, GORDON G, THRUN S. Point-based value iteration: an anytime algorithm for POMDPs[C]. Proceedings of the 18th International Joint Conference on Artificial Intelligence, 2003: 1025-1030.

[13] SUTTON R S, MCALLESTER D, SINGH S, MANSOUR Y. Policy gradient methods for reinforcement learning with function approximation[C]. Advances in Neural Information Processing Systems, 1999, 12: 1057-1063.

[14] HAFNER D, LILLICRAP T, BA J, 等. Dream to Control: Learning Behaviors by Latent Imagination[C]. International Conference on Learning Representations (ICLR). 2020. DOI:10.48550/arXiv. 1912.01603.

[15] HAFNER D, LILLICRAP T P, FISCHER I, 等. Learning Latent Dynamics for Planning from Pixels[C]. Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 97: 2555-2565. DOI:10.48550/arXiv.1811.04551.