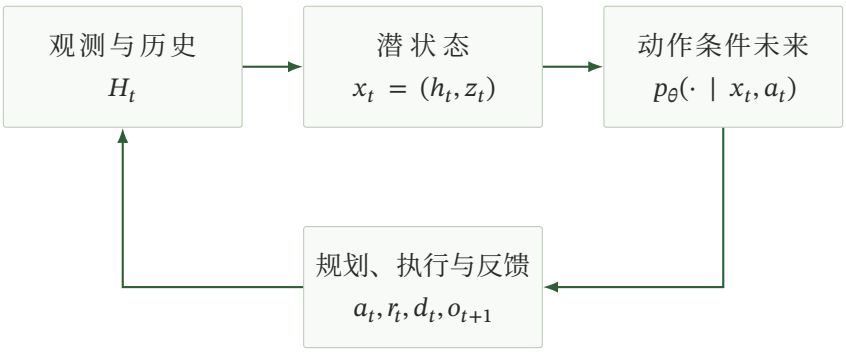
Jingfeng

世界模型

从潜状态动力学到可规划的交互系统

Jingfeng · 2026 年 10 月 22 日修订



世界模型把“接下来会看到什么”推进到“采取这个动作之后会发生什么”。本书从部分可观测性出发，推导潜状态学习、想象控制与模型评测，再用可以复跑的小实验检验这些概念。

这一版包含什么

完整的 belief 与 RSSM 推导；MPC、价值自举和策略学习；不同模型路线的接口对照；表示、因果与不确定性诊断；分层评测协议；已运行的两状态过滤和线性控制实验，以及尚未训练的 RSSM 教学实验规范。

Jingfeng Book · 原稿归档保留，修订源码独立维护。

本版资料核验截至 2026-10-22；文献报告、合成算例、本次实测与研究假设分别标明。

前言与阅读路线

本书面向掌握基础概率、线性代数和 Python 的读者。读者不必已经实现过强化学习算法，但应愿意区分条件概率、优化目标和实验指标。理解世界模型最常见的障碍，是把这三者混在一起：能拟合视频的目标，未必识别了动作的后果；能在内部生成高奖励的模型，未必能把机器人送到目标。

全书沿着一条信息链展开：观测历史提供信息，状态表示保存信息，动力学预测动作后果，规划器使用预测，真实执行再检验预测。第 2 章和第 3 章解释前两步的数学；第 4 章说明模型怎样进入决策；后续章节比较不同实现，并把局限落到可以检查的数据和日志上。

三条阅读路线

若希望先建立直觉，可先读第 1、5 章，再运行第 10 章的两状态过滤与线性控制实验。若希望读懂潜空间控制论文，应按第 2 至 4 章顺序推导，再回到实验中解释每个变量。若已有模型实现，优先读第 6、7、8 章，把自己的数据、终止标记、损失和闭环验证逐项对齐。

本书不把不同模型排成一列“先进程度”。RSSM、搜索模型、token 动力学、扩散轨迹和预测表征解决的问题有交集，但训练信号、可访问的动作和评测环境不同。第 9 章的基础模型与交互系统也必须放回这些条件理解。

证据怎样阅读

带文献引用的机制说明来自相应论文或官方资料；文献中的性能结果属于原作者的实验，不能视为本书复现。第 10 章明确标记为本次实测的部分，来自随书代码生成的合成环境，配置与结果保存在本主题的 `output/experiments/`。RSSM 训练规范是一份可执行研究设计，本次没有训练像素级 Dreamer、机器人控制器或大型视频模型。没有实验支持的建议写成待检验假设。

归档版本仍保存在 `source/`；本书独立项目位于 `manuscript/`，可以从同目录执行 `bash build.sh` 构建。数学符号和术语见附录 A；学习计划和研究问题见第 11 章。

与其他分册的关系

《强化学习》分册展开 MDP、策略梯度与价值学习的一般基础；本书只推导动力学模型进入决策所必需的部分。《具身智能》讨论传感器、机器人执行和硬件闭环；《NeRF 与 3D 高斯泼溅》展开辐射场与三维表示。本书提供理解这些接口所需的解释。

目录

前言与阅读路线 **ii**

第 1 章　先定义模型的契约 **1**

1.1　同一张画面，为什么可能需要不同动作 1
1.2　六个变量先于模型名称 1
1.3　四种用法，四种证据责任 1
1.4　动作条件化不等于动作因果 3
1.5　把世界模型放入三个循环 3
1.6　本章练习与提示 4

第 2 章　部分可观测性：为什么世界模型需要记忆 **5**

2.1　把一步交互的时间顺序写清楚 5
2.2　从初始观测到下一步后验 5
2.3　同一帧为什么可能需要不同动作 7
2.4　信念怎样变成可决策的状态 7
2.5　神经潜状态近似了什么 8
2.6　如何检查模型是否真的使用了历史 8

第 3 章　潜状态动力学：从概率分解到 RSSM 的训练目标 **10**

3.1　确定性记忆与随机潜变量各做什么 10
3.2　先写联合模型，再谈损失 10
3.3　滤波后验、平滑后验与信息泄漏 11
3.4　用 Jensen 不等式逐步得到 ELBO 11
3.5　似然的选择决定了损失的含义 12
3.6　KL 权重、free nats 与梯度平衡 13
3.7　表示塌缩和稳定性怎样诊断 13
3.8　从一步拟合走向可信的想象 14

第 4 章　让预测服务决策：想象、规划与模型偏差 **15**

4.1　Dyna 提供的基本分工 15
4.2　正确处理继续概率与末端价值 16

4.3 MPC：想一段，只执行第一步 16
4.4 用多步回报连接想象与价值学习 17
4.5 策略梯度怎样穿过想象 18
4.6 为什么规划长度没有统一的最优值 18
4.7 当优化器开始利用模型漏洞 20

第 5 章 模型家族：学习对象与规划接口 21
5.1 一段历史，几种持续存在的选择 21
5.2 概率模型：PETS、PlaNet 与 Dreamer 21
5.3 任务导向模型：MuZero 与 TD-MPC 22
5.4 离散 token 与序列动力学 22
5.5 两种扩散用途：轨迹计划与观测模拟 23
5.6 预测表征：JEPA 的价值与边界 23
5.7 把模型选择写成实验设计 23
5.8 练习：先固定条件再比较 24

第 6 章 从数据到可用的训练循环 25
6.1 先把一条转移存对 25
6.2 数据划分是模型定义的一部分 25
6.3 表示选择先问哪些信息可以丢掉 26
6.4 模型训练：损失、掩码与数值稳定性 26
6.5 从真实轨迹开始想象 26
6.6 短分支、长期记忆与多步训练 27
6.7 系统预算与复现实验包 27
6.8 练习与检查点 28

第 7 章 表示、因果与失败诊断 29
7.1 从一次失败追到可检验的假设 29
7.2 决策充分性：保留的是哪些未来 29
7.3 Bisimulation：何时可以合并状态 29
7.4 可观测性与可控性是不同条件 30
7.5 动作、干预与因果识别 30
7.6 不确定性：多样、未知与失配 31
7.7 误差与漏洞如何在闭环里放大 31
7.8 练习与提示 32

第 8 章　评测：从预测误差到闭环决策 **33**

8.1　先冻结问题，再选择指标 33

8.2　预测质量要沿时间展开 34

8.3　奖励、终止与不确定性不能附带处理 34

8.4　动作干预：从相关预测到可操作响应 35

8.5　闭环价值与模型利用漏洞 35

8.6　策略排序和统计不确定性 36

8.7　现有基准各自测到了什么 37

8.8　使用视觉语言模型裁判时 37

8.9　把预算写进结果表 37

练习与检查 38

第 9 章　基础模型、交互生成与三维世界的接口 **39**

9.1　基础模型首先是一个迁移承诺 39

9.2　交互生成：输入变化之后，环境怎样回应 40

9.3　预测表征怎样接上动作 41

9.4　三维、四维与持久记忆解决哪类问题 42

9.5　合成数据的价值要在下游测量 42

9.6　把展示能力变成可检验的接口 43

第 10 章　最小实验：从后验到闭环控制 **44**

10.1　先验证后验：同一帧不等于同一状态 44

10.2　把预测误差与控制回报分开测量 45

10.3　复现记录与预算怎样读 48

10.4　最小 RSSM：下一步应当怎样训练和验收 49

10.5　练习：尝试推翻一个看似合理的结论 51

第 11 章　研究工具箱：把直觉写成可证伪问题 **52**

11.1　阅读论文时先还原信息与计算流程 52

11.2　先区分三种误差，再设计替换实验 53

11.3　压缩误差为什么需要沿时间测量 53

11.4　六个研究问题及其反证条件 55

11.5　把泛化拆成具体变化 55

11.6　从基础到研究的十二周参考路线 56

11.7　提交一个别人能检查的研究方案 57

附录 A 符号、术语与关键练习提示 **58**

A.1 时间顺序与符号 58

A.2 重要术语的工作含义 59

A.3 信念、可观测性与因果练习 61

A.4 ELBO 与 KL 的核对步骤 61

A.5 规划与诊断练习 62

A.6 证据、实验与复现文件怎样对应 63

参考文献 **64**

第 1 章　先定义模型的契约

1.1　同一张画面，为什么可能需要不同动作

考虑一个小球穿过狭窄通道。当前图像里，小球恰好位于通道中央；它可能正在向左运动，也可能正在向右运动。若控制目标是把球停在中央，两个情形需要相反的制动动作。只根据这一帧生成一张清楚的下一帧，还没有解决状态估计，更没有解决控制。

世界模型需要回答一类带条件的问题：根据已经观察到的历史，如果现在施加动作，未来哪些量会怎样变化？这些量可以是像素、潜状态、奖励、终止、碰撞或某个任务的价值。模型不一定生成视频，但必须交代预测的对象、信息来源和使用方式。原始 World Models 工作展示了视觉压缩、循环动力学与控制器组合的一种实现；它不是所有世界模型的唯一结构 [1]。

本书使用的工作定义

世界模型是从数据学习的环境预测模型。讨论其决策用途时，必须说明：模型如何利用观测历史形成状态，动作以什么语义进入模型，预测哪些后果，以及外部控制或评测程序怎样使用这些预测。被动视频预测也可被相关文献称为世界建模；本书在这种用法下明确注明它尚未建立的动作和决策接口。

这个定义允许不同抽象层次，但不允许用抽象替代证据。例如一个游戏搜索模型只预测奖励、价值和搜索先验，仍可能很好地支持规划；它没有因此获得恢复现实世界几何的能力。反过来，一个逼真的视频生成器可能用于合成数据，却尚不具备可靠的控制接口。

1.2　六个变量先于模型名称

全书采用一致的时间顺序：在时刻 t 观察 o_t，据此选择 a_t；环境随后产生奖励 r_t、终止标记 d_t，到达 s_{t+1} 并给出 o_{t+1}。s_t 是真实环境状态，通常不可直接观察；模型内部的状态另记为 x_t，避免把估计当作真值。

某些系统没有奖励或终止头。这并非自动构成缺陷，但意味着这些量需要由外部程序定义，或该系统当前只承担感知、生成、表征用途。例如视频片段的末尾可能只是剪辑结束，不能直接作为“物理过程已终止”的监督标签。任务无关的动力学预训练可以暂不带奖励；一旦报告控制回报，就必须补齐奖励和终止语义。

1.3　四种用法，四种证据责任

第一类是**面向控制的潜空间模型**。模型接收历史和真实动作，预测后续潜状态、任务量及不确定性。PlaNet 使用模型预测控制，Dreamer 使用想象中的策略学习 [2, 3]。它们可以省掉控制时的图像解码，但需要真实执行结果检验潜空间预测。

第二类是**面向任务搜索的模型**。MuZero 的隐藏状态服务于奖励、价值和搜索先验，而不以

表 1.1 一份可以交给实现者的模型契约

字段	必须回答的问题	容易遗漏的条件
观测 o_t	输入是图像、深度、触觉还是状态量？	采样时间、遮挡、标定与传感噪声
动作 a_t	施加了什么控制？	单位、持续时间、限幅、延迟、坐标系
潜状态 x_t	保留哪段历史，怎样初始化？	是否重置记忆，是否含测试时不可用的真值
转移	预测哪个时刻的什么分布？	一步时长、随机性、是否自回归递推
奖励 r_t	哪种任务反馈被预测或另行计算？	是环境定义、偏好评分还是图像代理
终止 d_t	什么事件结束当前任务？	成功/失败、时间上限与数据缺帧不能混为一类

重建观测为目标 [4]。这里的状态首先是搜索算法所需的计算对象；只有在额外验证后，才可赋予它物理或因果解释。

第三类是**生成式视频环境或可交互模拟器**。动作条件视频、潜动作交互、场景条件生成都属于这一大范围，但它们的可控程度不同。可交互至少要求用户输入对后续结果有稳定影响；可靠模拟器还要验证多步状态一致、动作语义、重置与分支、错误频率及策略用途。Genie 的潜动作与真实机器人关节指令不是同一种动作 [5]。

第四类是**预测表征模型**。I-JEPA、V-JEPA 通过预测特征而非像素获得表示 [6, 7]。这可以是世界模型的有用组件，但静态图像遮挡预测并不自动构成时间动力学；视频表征也仍需动作条件化和控制验证。V-JEPA 2 将视频表征预训练和动作条件后训练连接起来，应分别阅读两个阶段的目标与证据 [8]。

把六个变量逐一代入，四种用法的契约会更清楚。表 1.2 中“未定义”表示必须由下游任务另行补充，不能把视频结束或特征误差直接当作终止与奖励。

表 1.2 四种用法中的观测、动作、状态、转移与任务反馈

用法	观测、动作与内部状态	转移、奖励与终止
控制型潜模型	o_t 是传感器或状态观测；a_t 是标定过的执行指令；x_t 压缩可用历史。	学习 $p(x_{t+1} \mid x_t, a_t)$；r_t, d_t 来自环境定义，可由模型预测或外部计算。
任务搜索模型	$o_{\le t}$ 形成搜索表示；a_t 是环境合法动作；x_t 服务搜索目标。	隐藏转移预测奖励并连接价值与策略先验；终止须由合法动作/环境规则或专门接口处理，不以潜状态几何解释。
生成式环境	o_t 是帧、视频片段或场景条件；a_t 可为真实指令、相机指令或潜动作；x_t 可为 token、缓存或场景状态。	预测下一帧、片段或场景分布；无奖励/终止监督时二者未定义，需由外部任务和判据提供。

用法	观测、动作与内部状态	转移、奖励与终止
预测表征模型	o 为图像/视频上下文；x 是特征；基础预训练通常没有执行动作，动作后训练才引入 a_t。	预测目标特征；静态遮挡预测无时间转移，视频预测也不自动含动作转移；r_t, d_t 通常需下游另建。

表 1.3 先看预测对象，再判断可用范围

当前证据	可以支持的陈述	还需要的实验
视频在观察分布内自然	学到了有用的视觉统计规律	动作互换、物理量与长时一致性
输入可改变生成结果	存在某种可操纵接口	控制方向、幅度、跨场景动作语义
固定策略在模型中得高分	该模型给出高评估值	独立真环境回测、漏洞搜索
模型中排序接近真实排序	在该策略集合上有评价效用	新策略族、不同域、置信区间
真实闭环回报提高	在报告条件下改善了决策	预算公平、泛化与失败切片

1.4 动作条件化不等于动作因果

一个模型可以在网络输入中接收动作，却只学会利用动作与场景的相关性。假设数据中“向左”总出现在红色房间，“向右”总出现在蓝色房间；模型通过房间颜色就能取得很低的预测损失，而无需理解动作。若测试时交换动作，这个捷径才暴露出来。

进一步地，潜动作可能吸收相机移动、遮挡或剪辑变化。让模型响应一个潜动作编码，与让它准确模拟指定扭矩的后果之间，存在识别与标定问题。需要在足够相似的历史下比较多个可执行动作，记录真实响应，而不能只看生成画质。第 7 章解释这些干预结论需要哪些假设。

1.5 把世界模型放入三个循环

信息循环把历史压缩为能预测未来的状态；**决策循环**比较动作、执行一步、吸收新观测；**学习循环**用真实执行更新模型与策略。三个循环的时间尺度不同。模型可以训练一小时却在毫秒级接受新观测；策略也可以离线训练，但执行时仍需要状态反馈。

想象轨迹只是模型输出，不能凭自身增加有关真实环境的证据。它能降低某些优化中的采样方差或复用已有知识，也会重复已有偏差。模型的样本效率收益必须在相同真实交互预算下测量，模型计算成本则另行记账。这也是为什么 Dyna 架构中的内部经验和真实经验需要分别统计[9]。

1.6 本章练习与提示

1. 为“机械臂根据摄像头推杯子”填写六变量契约。若奖励来自图像相似度，指出它与杯子真实到达目标之间可能出现的差异。提示：遮挡、相机运动与目标不可达都可改变解释。
2. 一个模型只能生成 5 秒视频，不能接收中途动作。它能否用于比较任意闭环策略？提示：区分预先固定动作序列与根据未来观测改变动作的策略。
3. 模型重建误差下降而控制回报下降，有哪些可能机制？至少提出一个表示原因和一个评测原因，并分别说明怎样验证。

下一章把“保留足够历史”变成概率递推。只有先说明什么信息可用，后续的潜状态与规划目标才有明确意义。

第 2 章　部分可观测性：为什么世界模型需要记忆

看见一张车道照片，不能确定前车是否正在刹车；看见同一个机械臂姿态，也不能确定被抓物体是否已经滑落。问题不一定出在图像分辨率，而在于单次观测没有包含影响下一步的全部信息。世界模型首先要回答：当前究竟知道什么，以及还不知道什么？本章用概率滤波把这个问题变成可以计算、可以检查的对象。

2.1　把一步交互的时间顺序写清楚

全书采用以下约定。环境具有不可直接访问的真实状态 s_t，智能体先接收观测 o_t，再选择动作 a_t；随后环境产生奖励 r_t、终止标记 d_t、下一状态 s_{t+1} 和下一观测 o_{t+1}。$d_t = 1$ 表示执行 a_t 后任务真正终止，而不是仅仅到达训练数据片段末尾。一个状态空间模型的基础部分是

$$T(s' \mid s, a) = p(s_{t+1} = s' \mid s_t = s, a_t = a), \qquad O(o \mid s) = p(o_t = o \mid s_t = s). \tag{2.1}$$

转移核 T 可以是离散概率表，也可以是连续状态上的条件密度。这里先讨论离散状态；连续状态时把求和替换为积分，必要时用粒子近似。

状态的马尔可夫性不是说世界没有记忆，而是说所选状态已经包含了预测未来所需的记忆。例如，仅以位置为状态的惯性系统通常不是马尔可夫的；加入速度后，才可能成为合适的状态模型。模型之外的温度、磨损和控制器内部变量，都可能使这种近似失效。

记可用的观测与动作历史为

$$H_t = (o_0, a_0, o_1, a_1, \dots, a_{t-1}, o_t), \qquad b_t(s) = p(s_t = s \mid H_t). \tag{2.2}$$

b_t 称为信念状态（belief state）。它是对真实状态的不确定性分布，不是一个“最可能状态”的别名。若两个状态概率相近而控制代价相反，保留整个分布可能直接改变动作选择。

奖励也是信息时怎么办？ 式 (2.2) 的 o_t 可理解为决策时实际可用的信息包。如果上一动作的奖励、接触信号或失败反馈能透露隐藏状态，就应把它们加入信息包，或明确扩充历史。不能一方面让滤波器忽略这些信息，另一方面声称 b_t 是相对于所有已知信息的后验。下面的简单递推采用式 (2.1) 的当前状态观测形式；信息包还依赖上一状态、上一动作时，应使用本节末尾的一般递推。

2.2　从初始观测到下一步后验

假设初始状态分布为 $\mu(s)$。即使还未执行动作，第一帧也应该更新初始信念：

$$b_0(s) = \frac{O(o_0 \mid s)\mu(s)}{\sum_u O(o_0 \mid u)\mu(u)}. \tag{2.3}$$

这里已经暴露出一个实验设计问题：若测试时把真实初始状态交给滤波器，而真实部署只有图像，测试任务就被悄悄简化了。

在时刻 t，先执行动作，再获得下一观测。第一步是**预测**：

$$\bar{b}_{t+1}(s') = \sum_s T(s' \mid s, a_t) b_t(s). \tag{2.4}$$

第二步是利用观测做**校正**：

$$Z_{t+1} = \sum_u O(o_{t+1} \mid u) \bar{b}_{t+1}(u), \tag{2.5}$$

$$b_{t+1}(s') = \frac{O(o_{t+1} \mid s') \bar{b}_{t+1}(s')}{Z_{t+1}}. \tag{2.6}$$

分母 $Z_{t+1} = p(o_{t+1} \mid H_t, a_t)$ 是当前模型给这次观测的预测概率。它既负责归一化，也能成为异常诊断信号：模型反复给真实观测极低概率，可能说明动力学参数改变、传感器故障或训练数据缺少这种情形。

推导只需要把联合概率分解：

$$\begin{aligned} &p(s_{t+1} = s', o_{t+1} \mid H_t, a_t) \\ &\quad = O(o_{t+1} \mid s') \sum_s p(s_{t+1} = s' \mid s_t = s, a_t) b_t(s). \end{aligned} \tag{2.7}$$

这一步有三个不能省略的条件。

1. 给定当前真实状态与动作，下一状态与更早历史条件独立。否则 T 的状态定义不够完整。
2. 给定下一真实状态，下一观测与更早状态、动作条件独立。若相机曝光取决于上一控制指令，应改写观测核，而不是继续套用当前公式。
3. 动作由已知历史及独立随机数选择，因此 $s_t \perp a_t \mid H_t$。若离线数据的操作者看到了记录中没有的隐藏变量，条件化于动作可能额外透露状态；此时不能自动用 b_t 代替 $p(s_t \mid H_t, a_t)$。

最后一点把概率滤波与因果问题接了起来。部署中的智能体可以主动选择动作；离线日志中的动作却可能携带未记录的专家知识。仅从相关性学习 $p(o_{t+1} \mid H_t, a_t)$，并不保证模型识别了实际执行动作干预的效果。

两个边界。 如果 $Z_{t+1} = 0$，当前模型认为观测不可能发生，公式没有定义。用一个极小数替换分母只能避免程序崩溃，不能恢复正确后验；需要检查观测模型、状态支持集或污染观测假设。另一个边界是数值下溢：长序列可以在对数域累加似然，并用 log-sum-exp 归一化，但仍应分别记录数值异常和真正的零概率。

带奖励与终止反馈的一般递推。 令 y_{t+1} 是新的传感器读数，定义联合环境核

$$K(s', r, d \mid s, a) = p(s_{t+1} = s', r_t = r, d_t = d \mid s_t = s, a_t = a).$$

若传感器读数只依赖新状态，则在收到 (y_{t+1}, r_t, d_t) 后，

$$b_{t+1}(s') \propto O(y_{t+1} \mid s') \sum_s K(s', r_t, d_t \mid s, a_t) b_t(s). \tag{2.8}$$

这不要求奖励、终止和下一状态彼此条件独立。例如“是否碰撞”既决定终止，也影响下一状态，强行拆成三个独立预测头就增加了模型假设。真正终止后可以进入吸收状态；下一条新回合数据必须重新初始化，不能延续上一回合的隐藏记忆。

2.3 同一帧为什么可能需要不同动作

考虑只有两个隐藏状态的环境，$s \in \{0,1\}$，一段短实验中状态保持不变：$T(s' \mid s,a) = \mathbf{1}\{s' = s\}$。观测也是一个比特，传感器以 0.8 的概率报告正确状态，各时刻噪声在给定 s 后独立。初始先验为 $\mu(1) = \mu(0) = 1/2$。动作暂时不影响状态；这个简化专门隔离记忆的作用。

用后验赔率表达 Bayes 更新更直观。若观测为 1，

$$\frac{b_t(1)}{b_t(0)} = \frac{0.8}{0.2}\frac{b_{t-1}(1)}{b_{t-1}(0)} = 4\frac{b_{t-1}(1)}{b_{t-1}(0)}; \tag{2.9}$$

观测为 0 时，赔率乘以 1/4。因此，下表两段历史最后看见的是相同的 $o_2 = 1$，结论却不同。

表 2.1 两状态手算示例；数字来自所设概率，不是训练得到的结果。

观测历史	后验赔率 $b_2(1)/b_2(0)$	$b_2(1)$	最可能状态
$(1,1,1)$	$4^3 = 64$	$64/65 \approx 0.9846$	1
$(0,0,1)$	$4^{-1} = 1/4$	$1/5 = 0.2$	0

若现在必须选择 $a \in \{0,1\}$，且猜中隐藏状态奖励为 1，否则为 0，则期望奖励分别是 $b_2(a)$。两段历史的最优动作相反。任何只接收最后一帧的确定性编码器 $z = g(o_2)$ 都会给出同一个表示，因此无法同时选对两种情形。这是**状态混叠**，不是简单地把编码器做大就能消除的问题。

不要把本例解读为“记忆越长越好”。这里隐藏状态恒定，旧证据不会过期；若状态以一定概率切换，预测步骤会先把过于确定的信念向不确定方向拉回。如果传感器噪声有时间相关性，直接连乘独立似然还会重复计算同一条证据。第 10 章给出表 2.1 中两状态算例的精确计算、机器可核验结果与可调整参数。

2.4 信念怎样变成可决策的状态

在模型已知、历史完整且上述条件成立时，b_t 足以预测未来的观测与回报。直观地说，过去只通过“现在可能处于哪些状态”影响未来；如果两个历史给出相同的信念，采用同一个从当前开始的观测反馈策略，就产生相同的未来分布。这使部分可观测问题可以转换成信念空间上的完全可观测问题 [10]。

设 $\rho(s,a) = \mathbb{E}[r_t \mid s_t = s, a_t = a]$，则信念上的即时奖励为

$$\rho(b,a) = \sum_s b(s)\rho(s,a). \tag{2.10}$$

把式 (2.6) 的更新记为 $\tau(b,a,o')$。在有限动作集、$0 \le \gamma < 1$ 且奖励有界时，最优折扣价值满足

$$V^*(b) = \max_a \left\{ \rho(b,a) + \gamma \sum_{o'} p(o' \mid b,a) V^*(\tau(b,a,o')) \right\}. \tag{2.11}$$

式中终止可以通过吸收状态处理，或显式乘继续概率。本式按只有 o' 提供新信息的简式书写；若奖励与终止额外透露状态，应把 o' 换成完整反馈包，并使用式 (2.8) 的更新与相应反馈分布。信念通常是高维分布，因而这条公式建立了概念上的充分性，却没有自动给出廉价求解器。

动作除了获取奖励，还能获取信息。机器人挪动相机也许没有立即增加抓取回报，却能区分两个被遮挡的状态，使后续选择更好。将信念压成一个均值、并假设均值就是真实状态的确定性规划，会丢掉这种信息价值。它在某些近线性、高信噪比场景可以是实用近似，但不是一般的 POMDP 最优解。

2.5 神经潜状态近似了什么

实际世界模型很少显式维护所有真实状态上的概率表，而是用递归编码器维护有限维内部量。需要分清三种对象：真实状态 s_t、真实状态上的信念 b_t，以及模型自己定义的潜变量 z_t。z_t 的坐标可以任意旋转、重标记或非线性变化；它通常与 s_t 不在同一个空间。

因此，不能在没有映射、共同测度和辨识假设时直接写“$q_\phi(z_t \mid H_t) \approx p(s_t \mid H_t)$”。更稳妥的要求是：由内部表示导出的可观测预测接近真实条件分布。例如，用 $q_\phi(dx_t \mid H_t)$ 表示内部状态的不确定性，对任务相关未来量 Y 和预先指定的未来动作串 A，希望

$$\int p_\theta(Y \mid x_t, A) q_\phi(dx_t \mid H_t) \approx p(Y \mid H_t, \mathrm{do}(A)), \tag{2.12}$$

确定性压缩 $x_t = e(H_t)$ 对应点质量分布；随机潜状态则需要对其后验预测积分，不能要求任意一个样本都携带整段历史的信息。右侧的干预含义需要动作数据覆盖和相应因果假设。若只讨论观测分布，应明确数据产生策略；对未来自适应动作串做普通条件化还可能引入选择偏差，不能直接替代固定动作干预。把两者混在一起，会高估被动视频数据能支持的结论。

预测充分性也要带上任务范围。一个表示可能足以预测道路方向，却不足以预测结冰路面的刹车距离；足以完成当前奖励任务，也未必能支持换一个奖励函数后的规划。第 3 章将构造一个可训练的潜状态模型，第 4 章再解释规划器实际消费哪些预测量。

能否从观测恢复隐藏状态？ 对线性系统 $s_{t+1} = As_t + Ba_t$、$o_t = Cs_t$，无噪声有限窗口下的可观测性由

$$\mathcal{O} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}, \qquad \mathrm{rank}(\mathcal{O}) = n \tag{2.13}$$

判定。这里 n 是状态维数，输入已知，系统矩阵已知。秩不足时，某些初始状态无论如何观察都不可区分；增加神经网络参数不会创造缺失的信息。带噪声时，还要关心数值条件与信噪比：理论满秩不等于有限数据下易于估计。非线性、多模态环境不能直接套用线性秩判据，但“哪些状态历史上不可区分”仍是正确问题。

2.6 如何检查模型是否真的使用了历史

仅看递归网络的结构不能证明它保留了有用记忆。可以设计以下相互补充的检验。

- **控制最后一帧**。构造末帧相同、早期证据不同的样本，检查后验、奖励预测和动作选择是否随历史变化。

- **打乱或截断历史**。在保持参数量和测试初始条件一致时比较多个窗口长度。若长窗口完全无效，需检查重置标记、梯度截断和任务是否真的部分可观测。
- **检测证据重复**。把一帧复制多次，不应无条件地被解释成多次独立测量。传感器有相关噪声时，检查后验置信度是否过度增长。
- **用真实状态做事后探针**。模拟器状态可用于评估记忆中保留了什么，但测试策略不应因此获得部署时没有的状态输入。探针可预测不等于模型已正确使用，也不等于形成了可干预的因果变量。

练习与提示。

1. 把两状态示例的观测正确率改为 p，推导长度为 n、其中有 k 次观测为 1 的后验。提示：赔率为 $[p/(1-p)]^{2k-n}$；检查 $p = 1/2$ 时是否退化为先验，$p < 1/2$ 时是否能反向使用传感器。
2. 若每步以 δ 的概率翻转隐藏状态，写出 $\bar{b}_{t+1}(1)$。提示：$\bar{b}_{t+1}(1) = \delta + (1-2\delta)b_t(1)$；再接观测校正。解释 $\delta = 1/2$ 时旧历史为何失效。
3. 令专家在知道真实隐藏状态时选择动作，观测日志却不记录该状态。举例说明 $p(s_t \mid H_t, a_t)$ 可以不同于 b_t。提示：令专家总选 $a_t = s_t$，然后区分“看见专家选了动作”与“自己强制执行动作”。
4. 在无噪声一维匀速系统中只观测位置，说明一帧与两帧的区别。提示：先把位置和速度组成状态，再写两行可观测性矩阵。

第 3 章　潜状态动力学：从概率分解到 RSSM 的训练目标

精确的信念滤波需要已知的转移与观测模型，也可能需要在极大的状态空间上积分。潜状态世界模型用一个可学习的内部系统替代这套计算：看到观测时修正内部状态，没有观测时按动作向前预测。递归状态空间模型（recurrent state-space model，RSSM）是这一思路的具体实现之一 [2, 3]。本章重点不是堆叠网络模块，而是逐项回答：联合概率是什么、后验近似了什么、损失究竟优化什么，以及训练成功能推出哪些结论。

3.1　确定性记忆与随机潜变量各做什么

沿用第 2 章的时间约定，真实环境状态始终记为 s_t。模型内部使用确定性记忆 h_t 和随机潜变量 z_t，并把二者合称为 $x_t = (h_t, z_t)$：

$$h_t = f_\theta(h_{t-1}, z_{t-1}, a_{t-1}), \qquad t \geq 1, \tag{3.1}$$

$$z_t \sim p_\theta(z_t \mid h_t), \tag{3.2}$$

$$o_t \sim p_\theta(o_t \mid x_t), \tag{3.3}$$

$$r_t \sim p_\theta(r_t \mid x_t, a_t), \qquad d_t \sim p_\theta(d_t \mid x_t, a_t). \tag{3.4}$$

h_0 是指定的初始记忆，z_0 由初始先验采样。h_t 可以累积过去的证据，z_t 则允许给定同一历史与动作后仍出现多种未来。这里的 h_t 是网络向量，不再用它表示完整历史；完整历史只记为 H_t。

观测到 o_t 时，表示网络构造近似滤波后验

$$q_\phi(z_t \mid h_t, o_t). \tag{3.5}$$

训练与真实控制都可以使用它；在想象未来时，未来观测尚不存在，只能使用先验 $p_\theta(z_t \mid h_t)$。先验与后验因此不是两个可随意互换的采样函数。后验重建好、先验滚动差，是完全可能发生的训练失败。

这套模型并不声称 x_t 等于真实物理状态。一个随机变量可以表示速度的不确定性，也可以编码难以解释的图像特征；是否形成决策充分表示，需要额外的预测与控制检验。确定性记忆也不能自动保证马尔可夫性，有限容量和有限训练窗口仍可能丢失信息。

3.2　先写联合模型，再谈损失

取一个完整序列片段

$$\mathcal{D} = (o_{0:T}, r_{0:T-1}, d_{0:T-1}), \qquad A = a_{0:T-1}.$$

本节把动作串 A 当作给定的外部输入来定义受控生成模型。它不是在没有因果条件时，把专家日志中的动作条件分布直接当作干预分布。离线数据的动作选择问题仍受第 2 章的限制。

给定潜变量路径，所有 h_t 都由确定性递推得到。简化的联合模型为

$$\begin{aligned} &p_\theta(\mathcal{D}, z_{0:T} \mid A) \\ &\quad = \prod_{t=0}^{T} \big[p_\theta(z_t \mid h_t) p_\theta(o_t \mid x_t)\big] \prod_{t=0}^{T-1} \big[p_\theta(r_t \mid x_t, a_t) p_\theta(d_t \mid x_t, a_t)\big]. \end{aligned} \tag{3.6}$$

这里隐含了给定 x_t, a_t 后奖励与终止条件独立、给定 x_t 后观测独立等建模选择。这些是假设，不是环境必须遵守的眞理。若“碰撞造成终止”和某类状态跳变具有强依赖，可以把相应输出改成联合头，或把奖励与终止条件化于 x_{t+1}；但整本实现的时间索引必须同时更改。

式 (3.6) 把最后一帧 o_T 作为有观测、无对应片段内动作的边界，因此只对 $t < T$ 计算奖励和终止似然。这样不会凭空多出 a_T、r_T 或 d_T。实际代码若统一存储长度，也应使用相应掩码。

确定性记忆若写成随机图模型，会对应 Dirac 条件分布。因为同一递推同时出现在生成模型和推断过程里，直接把 h_t 当作确定性计算最清晰；不要给每个 h_t 随意添加一个没有定义的概率密度项。

3.3 滤波后验、平滑后验与信息泄漏

完整模型的精确后验为 $p_\theta(z_{0:T} \mid \mathcal{D}, A)$，通常不可直接计算。选择因果的变分族

$$q_\phi(z_{0:T} \mid \mathcal{D}, A) = \prod_{t=0}^{T} q_\phi(z_t \mid h_t, o_t). \tag{3.7}$$

虽然左侧以整个片段作为条件，右侧刻意只使用当前和过去信息。h_t 依赖之前采样的 $z_{<t}$，因此并不是把各时刻后验无条件独立相乘。共享动力学参数经由 h_t 进入推断计算的依赖，在记号中予以省略，但自动微分实现中不能忽视。

这个族还没有显式利用奖励和终止反馈来编码后验，所以它是受限的近似；受限并不妨碍变分下界成立，只可能使界更松。若奖励本身是重要传感器，可以一致地把历史反馈加入编码器输入。关键是 r_t 在执行 a_t 后才到来，不能用它帮助决定同一时刻尚未执行的 a_t。

平滑后验可以利用未来观测 $o_{t+1:T}$，例如在离线分析中回推先前发生了什么。它往往给出更好的状态估计，但不能直接充当部署时的在线滤波器。若训练时用未来帧构造的状态作为策略输入，测试时却只给历史帧，这不是普通泛化误差，而是输入信息不一致。

3.4 用 Jensen 不等式逐步得到 ELBO

设 $q = q_\phi(z_{0:T} \mid \mathcal{D}, A)$，并假设相关密度和积分存在，$p_\theta(\mathcal{D} \mid A) > 0$，且 q 不给联合模型的零概率区域分配正质量。对隐变量积分，在 $q > 0$ 的区域乘除以 q：

$$\begin{aligned} \log p_\theta(\mathcal{D} \mid A) &\geq \log \int_{q>0} q(z_{0:T}) \frac{p_\theta(\mathcal{D}, z_{0:T} \mid A)}{q(z_{0:T})} \, dz_{0:T} \\ &\geq \mathbb{E}_q[\log p_\theta(\mathcal{D}, z_{0:T} \mid A) - \log q(z_{0:T})] \\ &\equiv \mathcal{L}_{\mathrm{ELBO}}. \end{aligned} \tag{3.8}$$

第一行的等号要求遗漏区域在联合模型下的质量为零；若遗漏了正概率区域，积分只会更小。第二行使用对数的凹性，即 Jensen 不等式。证据下界（evidence lower bound，ELBO）的松紧也可以通过恒等式理解：

$$\log p_\theta(\mathcal{D} \mid A) = \mathcal{L}_{\mathrm{ELBO}} + \mathrm{KL}(q \,\|\, p_\theta(z_{0:T} \mid \mathcal{D}, A)). \tag{3.9}$$

若变分族恰好包含精确后验，并且找到它，后面的 KL 为零，界才取等号。这是关于所定义概率模型的结论，不是“潜变量恢复了真实物理”的保证。

为缩短公式，定义每个时刻的对数似然和

$$\ell_t(x_t) = \log p_\theta(o_t \mid x_t) + \mathbf{1}_{\{t<T\}}\big[\log p_\theta(r_t \mid x_t, a_t) + \log p_\theta(d_t \mid x_t, a_t)\big]. \tag{3.10}$$

把联合分解代入式 (3.8)，整理潜变量密度比，得到

$$\begin{aligned}\mathcal{L}_{\mathrm{ELBO}} = &\sum_{t=0}^{T} \mathbb{E}_{q(z_{\le t})}[\ell_t(x_t)] \\ &- \sum_{t=0}^{T} \mathbb{E}_{q(z_{<t})}\left[\mathrm{KL}\big(q_\phi(z_t \mid h_t, o_t) \,\|\, p_\theta(z_t \mid h_t)\big)\right].\end{aligned} \tag{3.11}$$

在 $t = 0$ 时，外层历史期望为空，只有初始先验与初始后验之间的 KL。对其他时刻，h_t 随先前后验样本改变，所以 KL 外面的历史期望不可省略。一次采样路径给出了这个期望的 Monte Carlo 估计，并不意味着精确数学表达式里没有期望。

奖励和终止似然同样处在联合后验期望中。将奖励项写在期望之外，会让随机状态 x_t 在公式中成为一个未经说明的自由变量；只在明确“这一行表示固定采样路径上的估计量”时才成立。代码中可以用采样路径统一评估所有头，文稿中则应区分目标与估计量。

3.5 似然的选择决定了损失的含义

若观测模型是固定方差的独立高斯分布，

$$p_\theta(o_t \mid x_t) = \mathcal{N}\big(o_t; g_\theta(x_t), \sigma_o^2 I\big),$$

则负对数似然在一个与模型均值无关的常数之外，等于

$$\frac{1}{2\sigma_o^2}\|o_t - g_\theta(x_t)\|_2^2. \tag{3.12}$$

因此均方误差不是脱离假设的通用似然；其权重对应观测噪声尺度。把所有像素求和改成平均，等价于改变它相对于 KL、奖励和终止项的比例。报告损失时，必须说明是每维、每帧还是每序列的值。

对终止变量可用 Bernoulli 模型，预测 $\hat{d}_t \in (0, 1)$：

$$\log p_\theta(d_t \mid x_t, a_t) = d_t \log \hat{d}_t + (1 - d_t)\log(1 - \hat{d}_t). \tag{3.13}$$

预测继续概率时记 $c_t = 1 - d_t$，并始终检查软件接口返回的是终止还是继续。稀有终止样本下，准确率可能很高但召回极差；对规划更有意义的是事件概率校准、漏检率和终止时间误差。

奖励可以使用高斯、离散分桶或其他明确归一化的分布。若用任意加权回归、Huber 损失或自定义表示距离替代对数似然，这些仍可能是有效工程目标，但只有在补齐相应概率解释后，才能把总目标称为对应模型的 ELBO。不同预测头共享潜状态，也不意味着这些头的误差互相独立。

3.6 KL 权重、free nats 与梯度平衡

将似然和 KL 分别简记为 R 与 $K \geq 0$，标准 ELBO 是 $R - K$。实践中常优化

$$\mathcal{L}_\beta = R - \beta K. \tag{3.14}$$

当 $\beta = 1$ 时它等于上述 ELBO。当 $\beta \geq 1$ 时，$\mathcal{L}_\beta \leq \mathcal{L}_{\text{ELBO}} \leq \log p_\theta(\mathcal{D} \mid A)$，所以在相同的规范化似然、相同 q 和同一数据下，它仍是更松的证据下界，但不是 Jensen 推导所得的原始 ELBO。$0 \leq \beta < 1$ 则一般没有原证据下界保证：减轻 KL 惩罚可能把目标抬到对数证据之上。最稳妥的命名是“β 加权训练目标”，并把实际公式写出来。

减小 β 往往让后验保留更多当前观测信息，同时允许它更偏离想象时可用的先验；增大 β 强化先验与后验的一致性，也可能让随机潜变量不再承载有效信息。这些是有条件的训练倾向，不是对所有架构都单调成立的定律。

Free nats。 一种做法对每步 KL K_t 使用 $\max(\kappa, K_t)$，其中 $\kappa \geq 0$ 的单位是 nat（自然对数单位）。当 $K_t < \kappa$ 时，惩罚变为常数，其梯度为零，允许表示先使用一小部分信息容量。它不意味着模型凭空获得了 κ 的信息，也不是把 KL 估计中的噪声减掉。对总 KL 截断与对每个潜变量维度分别截断不是同一目标，应在配置中明确。

如果对精确 KL 设置这种惩罚下限，$R - \max(\kappa, K)$ 仍不超过原 ELBO；因此“不是原始 ELBO”不等于“一定不再是任何下界”。但它的优化目的已经改变，不能继续用式 (3.9) 的标准间隙恒等式来解释这一新目标。

KL balancing。 记 $\text{sg}[\cdot]$ 为停止梯度。一个通用的平衡形式是

$$K_t^{\text{bal}} = \alpha \,\text{KL}(\text{sg}[q_t] \,\|\, p_t) + (1 - \alpha)\,\text{KL}(q_t \,\|\, \text{sg}[p_t]), \qquad 0 \leq \alpha \leq 1. \tag{3.15}$$

第一项主要让动力学先验追赶固定的表示目标；第二项主要约束表示不要任意远离固定先验。在前向数值上两项加权后仍等于同一个 KL，但反向梯度不同。尤其 h_t 由共享网络产生时，哪些路径被截断必须以计算图为准，不能仅凭参数名猜测。平衡、free nats 和额外权重组合后的规则应视为指定的训练更新，不应声称它仍严格沿原 ELBO 的梯度上升。Dreamer 各版本的具体系数与分布设计并不相同 [11, 12]。

3.7 表示塌缩和稳定性怎样诊断

一个接近零的 KL 可能表示后验几乎不依赖当前观测，也可能表示当前信息已经能由历史预测，或确定性路径 h_t 已经承担了大部分工作。不能只看一个 KL 标量就宣布成功或塌缩。

在固定 h 时，令数据条件分布下的聚合后验为 $\bar{q}(z \mid h) = \mathbb{E}_{o|h} q(z \mid h, o)$，则

$$\begin{aligned}\mathbb{E}_{o|h}\,\text{KL}(q(z \mid h, o) \,\|\, p(z \mid h)) = I_q(z; o \mid h) \\ + \text{KL}(\bar{q}(z \mid h) \,\|\, p(z \mid h)).\end{aligned} \tag{3.16}$$

这里的互信息由给定 h 后的联合分布 $p_{\text{data}}(o \mid h) q(z \mid h, o)$ 定义。平均 KL 因而同时包含“当前观测带来的信息”和“聚合后验与先验的不匹配”。低 KL 只有结合观测预测、随机潜变量利用率

与控制表现才有解释力。高 KL 则可能来自有用信息，也可能只是先验没有学会后验所编码的动力学。

实际训练至少联合记录以下量。

- 每步重建、奖励和终止负对数似然；它们的求和维度、掩码和归一化方法。
- 原始 KL、截断后的 KL、先验熵和后验熵；离散潜变量可再记有效类别数与类别使用分布。
- 用后验更新的一步预测误差，以及从同一起点连续用先验滚动的多步误差。只报告后验重建会掩盖想象失败。
- 梯度范数、非有限值次数、预测方差上下界和真实闭环回报。降低训练损失并不保证控制回报上升。

常见的稳定化措施有明确的对应问题：对高斯标准差设正下界避免密度奇异；约束梯度范数缓解长递归链爆炸；使用完整回合边界掩码防止跨回合污染；从片段前缀做 burn-in 初始化记忆，减轻随机截段后 h_t 错配。它们仍需对照实验，不能代替正确时间索引。第 10 章提供最小实验的日志字段和预算记录方法。

3.8 从一步拟合走向可信的想象

训练中每一步都用真实观测校正后验，模型因此不断得到“答案提示”；想象时只有自身先验输出。若从真实后验状态出发，连续执行 k 个动作，得到先验预测分布，再与终点的后验或真实未来量比较，就能直接检验多步一致性。PlaNet 的 latent overshooting 是这类思想的重要实例 [2]。

多步观测损失、潜变量分布匹配和任务价值一致性有不同作用。观测损失受像素噪声影响；潜变量损失受表示坐标和停止梯度选择影响；价值一致性则依赖当前任务与价值函数质量。它们不能被不加条件地写成同一个 KL，也不保证长时误差消失。随机环境中，还应比较分布或适当统计量，不能把一条模型样本与一条真实样本的差异全部当成模型错误。

奖励监督可以迫使表示保留任务相关变量，却不会自动把相关表示变成因果表示。例如训练数据总在晴天成功、雨天失败，奖励头可能利用天气纹理，完全没有学到摩擦系数如何影响刹车。真正有说服力的证据包括在控制其他条件时改变动作、改变环境机制或改变视觉因素，并检验预测与回报是否按预期变化。奖励头既是规划接口，也是必须单独验证的预测器。

练习与提示。

1. 用 $T = 1$ 展开式 (3.6) 与式 (3.11)，列出两帧观测、一个奖励、一个终止标签和两个 KL。提示：检查没有使用不存在的 a_1。
2. 给出 $0 \le \beta < 1$ 时 $\mathcal{L}_\beta$ 超过对数证据的简单例子。提示：取只有一个观测事件的离散潜变量模型，令后验接近高似然潜变量而先验给它低概率；$\beta = 0$ 时目标只保留条件对数似然。
3. 对一维高斯 $q = \mathcal{N}(\mu_q, \sigma_q^2)$、$p = \mathcal{N}(\mu_p, \sigma_p^2)$ 写出 KL，并说明 $\sigma_p \to 0$ 的风险。提示：KL 包含 $[(\mu_q - \mu_p)^2 + \sigma_q^2]/(2\sigma_p^2)$，方差边界会影响梯度。
4. 假设总 KL 接近零但回报良好，设计实验判断随机潜变量是否必需。提示：保持确定性记忆容量不变，对比去除 z、打乱当前观测与改变遮挡程度；不要只比较重建图。

第 4 章　让预测服务决策：想象、规划与模型偏差

世界模型给出的未来必须被某个决策过程消费，才会影响真实行为。这个过程可以是搜索动作序列的规划器，也可以是通过想象训练出的策略。二者共享动力学接口，却把计算放在不同地方：前者每次决策都做优化，后者把部分优化提前到策略训练中。本章把真实数据、后验起点、先验想象、价值学习和执行后的观测校正连成一个闭环。

4.1　Dyna 提供的基本分工

Dyna 的核心是让同一份真实经验承担多种工作：学习价值或策略，拟合环境模型，再用模型生成内部经验更新价值或策略 [9]。它不要求模型生成像素，也不要求通过动力学反向传播。一个表格转移模型配合 Q-learning，同样属于这条路线。

学习的状态、奖励或终止模型一旦有偏，内部经验就不等价于新增真实样本。把“一步真实交互生成了一万步想象”写成“样本效率提升一万倍”，忽略了模型偏差和计算成本。至少应分别记录真实环境步数、模型更新数、想象步数、规划延迟与最终真实回报。

潜状态方法还多了一个接口：从真实回放片段过滤得到 $x_t = (h_t, z_t)$，然后只使用先验向前滚动。若想象从任意高斯向量出发，起点可能本来就不在模型训练过的状态区域；若每一步偷看未来真实观测，测到的则不是开放滚动预测能力。

算法 1 从真实经验到潜空间想象的通用训练框架

初始化回放库、RSSM、策略 π_η、价值函数 V_ψ
for 每个真实交互阶段 **do**
    用当前策略或规划器收集真实轨迹，保存观测、动作、奖励和边界类型
    从回放中抽取片段，利用前缀初始化记忆，按式 (3.11) 的目标更新模型
    从真实片段的后验采样起点 x_t，固定本轮模型参数
    for 每条想象轨迹的 $k = 0, \dots, H-1$ **do**
        采样 $a_{t+k} \sim \pi_\eta(\cdot \mid x_{t+k})$
        预测奖励、继续概率，递推 h_{t+k+1}，采样先验 z_{t+k+1}
    end for
    计算带终止掩码的回报目标，更新价值与策略
    用真实闭环回报和先验滚动误差验收本轮更新
end for

这里的“固定模型参数”指不对模型参数执行策略优化器更新。若使用路径梯度，仍需要保留输出对动作和内部状态的导数；把整个动力学计算放进禁止求导的上下文，会连同策略需要的梯度一起截断。不同算法如何更新策略与价值并不相同，这段框架也不意味着 Dreamer 与 TD-MPC 具有相同训练损失。

4.2 正确处理继续概率与末端价值

先写一个同时适用于真实或模型轨迹的回报。令

$$\Gamma_0 = 1, \qquad \Gamma_{k+1} = \Gamma_k \, \gamma(1 - d_{t+k}), \qquad 0 \le \gamma < 1. \tag{4.1}$$

从 t 开始向前看 H 步，以 V_ψ 补上剩余回报：

$$J_H = \mathbb{E}\left[\sum_{k=0}^{H-1} \Gamma_k r_{t+k} + \Gamma_H V_\psi(x_{t+H})\right]. \tag{4.2}$$

执行动作后终止的那一步，仍然获得 r_{t+k}，但不再获得其后的奖励或价值。遗漏这一区别，会让模型在“已经失败”的未来继续收获虚假的回报。

若在想象中使用预测的继续概率 $\hat{c}_k = 1 - p_\theta(d_{t+k} = 1 \mid x_{t+k}, a_{t+k})$，可用 $\gamma\hat{c}_k$ 替代单步折扣。这样做是对模型终止随机性的边缘化近似；只有与模型的条件独立结构一致时，才能把多个头的均值直接相乘。若终止与下一潜状态的噪声相关，应采样相应联合分布，不能假设独立头会自动保留这种相关性。

末端价值估计的是从末端开始、按某个后续策略行动的回报。若 V_ψ 仅在当前策略数据上训练，就不能无条件把它当成最优价值 V^*。短规划能否可靠，取决于到达末端时的状态是否在价值函数可信区域，以及规划后实际采用的后续策略是否与价值定义一致。

终止不等于截断。 任务成功、碰撞或其他定义内的结束是终止（termination）；数据分片结束、采样器暂停或人为时间上限通常是截断（truncation）。真正终止后，价值 bootstrap 必须为零；如果截断前任务本身还会继续，并且拿得到正确的最终观测，应保留价值 bootstrap。两者都可能要求重置环境或递归记忆，但价值目标不能因此使用同一个掩码。若任务本来就是有限时域，且剩余时间属于状态，那么到达任务时限可以是真正终止。自动重置接口返回的新回合首帧不能冒充截断前的最终观测。

4.3 MPC：想一段，只执行第一步

模型预测控制（model predictive control，MPC）每次在当前后验下寻找一个长度 H 的动作串，执行第一步，再用新观测校正状态并重新规划。设动作串为 $A = (a_t, \dots, a_{t+H-1})$，则典型问题是

$$A^* \in \arg\max_{A \in \mathcal{A}^H} \hat{J}_H(A \mid H_t). \tag{4.3}$$

估计 $\hat{J}_H$ 时，应同时考虑初始后验的不确定性与转移随机性。对多峰后验，均值可能对应一个根本不存在的状态；多粒子评估比只取均值更能保留“左侧有障碍”与“右侧有障碍”这类互斥可能。

交叉熵方法（cross-entropy method，CEM）是一种常用的无梯度采样优化器。下面给出连续、有界动作情形的版本。它与 PETS、PlaNet 等规划方法的思想有关，但并不声称复刻其中某篇论文的全部实现 [2, 13]。

高斯提议经过硬裁剪后不再是原高斯分布，可能在边界聚集；使用有界参数化或专门的截断分布时应写清楚。对离散动作，通常需要类别提议、树搜索或其他组合优化器，而不是直接把连续 CEM 生搬过去。

算法 2 带后验粒子与价值补偿的 CEM–MPC

输入：时域 H，候选数 N，粒子数 P，精英数 $K < N$，优化轮数 I

由真实历史过滤得到当前后验；初始化动作分布均值 μ 与尺度 σ
for 每个真实决策时刻 t **do**
  for $i = 1, \dots, I$ **do**
    从提议分布采样 N 个动作串，并映射到合法动作范围
    for 每个候选动作串 $A^{(n)}$ **do**
      从当前后验采样 P 个起点，分别用先验滚动 H 步
      按式 (4.2) 求各粒子回报，并计算平均分数
    end for
    保留分数最高的 K 个精英；估计各时刻动作的均值和方差
    平滑更新 μ, σ，并设置非零最小尺度以保留搜索能力
  end for
  执行所得动作串的第一个动作 a_t
  接收 r_t, d_t, o_{t+1}；用真实观测更新后验，处理回合边界
  将剩余均值序列左移作为下一次热启动，并补上末端提议
end for

这里优化的是开放环动作串：后面第 k 个动作在这次规划中不会根据那时尚未看见的观测分支。真实系统通过每步重新规划减少这一缺点，却没有自动解决完整的信念空间策略优化或主动感知问题。若当前动作的主要价值是为未来获取信息，简单开放环搜索可能低估它。

计算预算也属于算法定义。CEM 的模型转移调用量大致随 $INPH$ 增长；若只给一个方法更多粒子和更长规划时间，得到更高回报未必来自模型本身。比较时应同时固定或报告环境预算与决策计算预算。所有候选都达到很差分数时，继续增加采样不一定能补救失准的模型。

4.4 用多步回报连接想象与价值学习

单步 TD 依赖价值估计，完整多步回报依赖更多模型预测。λ 回报在这两者之间做一个可写清楚的混合。令想象单步折扣为 $\delta_k = \gamma \hat{c}_k$，在给定轨迹上定义

$$G_H^\lambda = V_{\bar{\psi}}(x_{t+H}), \tag{4.4}$$

$$G_k^\lambda = \hat{r}_k + \delta_k \left[(1-\lambda) V_{\bar{\psi}}(x_{t+k+1}) + \lambda G_{k+1}^\lambda \right], \quad k = H-1, \dots, 0. \tag{4.5}$$

$\bar{\psi}$ 可以是固定的目标网络参数；$\hat{r}_k$ 可以是采样奖励或模型条件期望。$\lambda = 0$ 对应一步 bootstrap，$\lambda = 1$ 对应到片段末端的 H 步回报加末端 bootstrap。它们都必须乘正确的继续折扣。

一个简单的价值回归目标为

$$\mathcal{L}_V(\psi) = \mathbb{E}\left[\sum_{k=0}^{H-1} w_k \left(V_\psi(x_{t+k}) - \mathrm{sg}\left[G_k^\lambda \right] \right)^2 \right], \qquad w_k = \prod_{j=0}^{k-1} \delta_j. \tag{4.6}$$

目标停止梯度，防止当前价值网络通过修改自己的目标而降低损失；具体实现还需决定是否冻结想象状态和权重。分布式价值预测可以用相应目标分布的交叉熵或似然替代平方误差，但不能省略目标分布的定义。

把 λ 调小，会缩短直接依赖模型奖励链的长度，却增加对价值估计的依赖。它不是消除偏差的开关。若价值已在模型分布外状态上过高，较小 λ 也可能更早引入乐观偏差。

4.5 策略梯度怎样穿过想象

策略参数记为 η，避免与编码器 ϕ、动力学 θ、价值 ψ 混用。先以式 (4.2) 的期望回报定义目标 $J(\eta)$，本次求导固定模型、初始状态分布与末端价值参数。若初始分布或末端价值也显式依赖 η，还须计入它们的导数。实际方法可能使用 G_k^λ 构造训练替代目标，并加入熵正则、优势标准化或裁剪；这些选择应逐一报告，而不是把所有 Dreamer 版本归结为一个“累加价值”的公式 [3, 11, 12]。

路径梯度。 若连续动作和随机动力学可以重参数化，写成

$$a_k = g_\eta(x_k, \epsilon_k), \qquad z_{k+1} = m_\theta(h_{k+1}, \xi_k),$$

其中基础噪声的分布不依赖 η。在可微与积分交换条件成立时，

$$\nabla_\eta J(\eta) = \mathbb{E}_{\epsilon,\xi}\left[\frac{d}{d\eta}\widehat{G}_0(\eta;\epsilon,\xi)\right]. \tag{4.7}$$

这里 $\widehat{G}_0$ 是式 (4.2) 对应的重参数化样本回报；任意替换为近似 λ 目标并不保留同一目标的无偏梯度解释。全导数通过动作、后续潜状态、奖励和末端价值传播。价值参数可以固定，但 $\partial V/\partial x$ 仍可用于策略改进。离散终止不能直接使用普通路径导数；若按模型结构正确边缘化终止，可对连续的继续概率求导，否则需要相应的离散梯度估计。

这种梯度低方差且能利用模型结构，也会放大错误的局部敏感性。模型即使在训练点预测准确，对动作的导数也可能不准确；策略沿着这个导数上升，就可能找到数据中从未验证过的高回报“缝隙”。

得分函数梯度。 不要求动作可微时，可以对策略采样概率求导。对固定动力学下的折扣回报，其典型形式是

$$\nabla_\eta J = \mathbb{E}\left[\sum_{k=0}^{H-1} \Gamma_k \nabla_\eta \log \pi_\eta(a_{t+k} \mid x_{t+k})\big(G_k - b(x_{t+k})\big)\right]. \tag{4.8}$$

这里 G_k 是从 k 开始的局部回报，含一致的末端价值；基线 b 不能依赖当前采样动作。作为策略损失实现时，回报或优势应停止梯度。该式的无偏解释要求用目标策略对应的轨迹分布与正确回报；把 G_k 换成近似的 λ 目标、截断优势或离策略数据后，还需要分别讨论引入的偏差与校正。

离散潜变量的 straight-through 估计器在前向使用离散样本、反向使用替代导数，一般是有偏的梯度估计。它可能是实用选择，却不能被描述成原始离散系统的精确路径梯度。得分函数、路径梯度和混合估计各有不同的偏差与方差，不能只凭“可以反向传播”判断优劣。

4.6 为什么规划长度没有统一的最优值

长规划可以看见延迟奖励，也会反复消费模型自己的预测。短规划少依赖动力学，却更依赖末端价值。下面给出两个有明确前提的误差界，用来理解这种交换，而不是给所有视频模型套用同一条结论。

4.6.1 同坐标确定性系统中的误差递推

本小节先假设没有提前终止；若使用吸收状态表示终止，扩充后的整个系统也必须满足下述正则条件。设真实系统和近似模型在同一个有度量的状态空间中分别为

$$s_{k+1} = F(s_k, a_k), \qquad \hat{s}_{k+1} = \hat{F}(\hat{s}_k, a_k), \qquad s_0 = \hat{s}_0.$$

固定相同动作串。在包含两条轨迹的区域内，假设真实 F 对状态是 L_F-Lipschitz，并且有一致模型误差

$$\|F(u,a) - F(v,a)\| \le L_F \|u - v\|, \qquad \|\hat{F}(u,a) - F(u,a)\| \le \varepsilon_F.$$

则令 $e_k = \|\hat{s}_k - s_k\|$，由三角不等式得到

$$e_{k+1} \le L_F e_k + \varepsilon_F, \qquad e_k \le \varepsilon_F \sum_{j=0}^{k-1} L_F^j. \tag{4.9}$$

$L_F < 1$ 时这个上界趋于 $\varepsilon_F/(1 - L_F)$；$L_F = 1$ 时上界线性增长；$L_F > 1$ 时上界允许几何增长。上界增长不是实际误差必定按该速度增长，方向抵消和局部稳定性都可能让实际误差小得多。

进一步假设真实奖励均值 ρ 是 L_r-Lipschitz，奖励模型一致误差不超过 ε_r；末端真实后续策略价值 V^π 是 L_V-Lipschitz，估计 $\hat{V}$ 的一致误差不超过 ε_V。对于前 H 步用该动作串、以后用 π 的目标，有

$$\begin{aligned} |\hat{J}_H - J_H| \le & \sum_{k=0}^{H-1} \gamma^k \left(\varepsilon_r + L_r e_k\right) \\ & + \gamma^H \left(\varepsilon_V + L_V e_H\right). \end{aligned} \tag{4.10}$$

这个式子把奖励误差、状态误差和价值误差分开：增加 H 会压低 $\gamma^H \varepsilon_V$，却增加需要经过模型预测的状态和奖励项。因此仅凭一步预测误差最小，不能选出最好的控制时域。

这些假设并不轻。接触、碰撞和硬终止可能使奖励或价值不满足所需 Lipschitz 条件；只在训练数据平均意义下小的误差，不能替代规划区域的一致界。RSSM 的潜变量与真实状态通常不共用坐标，也不能直接把二者欧氏距离代入式 (4.9)。第 10 章的线性小系统专门提供可控的共同坐标，用于验证递推和规划时域效应。

4.6.2 随机系统中的一个有界版本

另一种假设使用总变差距离。令真实与模型转移核在同一状态空间上满足

$$\sup_{s,a} \mathrm{TV}\big(P(\cdot \mid s,a), \hat{P}(\cdot \mid s,a)\big) \le \varepsilon_P,$$

其中 $\mathrm{TV}(p,q) = \sup_B |p(B) - q(B)|$。固定同一马尔可夫策略或同一开放环动作串，从相同初始分布出发，耦合论证给出第 k 步状态及动作联合分布距离不超过 $\min(1, k\varepsilon_P)$。若奖励均值绝对值不超过 $R_{\max}$、奖励均值模型的一致误差不超过 ε_r，且 $|V^\pi| \le V_{\max}$、末端估计误差不超过 ε_V，则

$$\begin{aligned} |\hat{J}_H - J_H| \le & \sum_{k=0}^{H-1} \gamma^k \left[\varepsilon_r + 2R_{\max} \min(1, k\varepsilon_P)\right] \\ & + \gamma^H \left[\varepsilon_V + 2V_{\max} \min(1, H\varepsilon_P)\right]. \end{aligned} \tag{4.11}$$

因本节采用 $\sup_B$ 的总变差定义，有界函数期望差的界带系数 2。终止可以加入吸收状态，并令其后奖励与价值为零；此时 ε_P 也必须覆盖终止机制误差，不能只检查未终止状态的动力学。这个结果同样依赖覆盖规划区域的一致误差；验证集平均似然无法直接给出 ε_P。

如果误差界 Δ_H 对所有被搜索的候选计划都成立，那么模型最优计划与真实最优候选的回报差至多为 $2\Delta_H$，再加规划优化器自身的次优误差。证明只需在真实值与模型值之间各加减一次。若界只在行为策略常见区域成立，搜索器恰恰可能主动跳出该区域，最优计划保证便不能使用。MBPO 对短模型分支的研究可作为进一步阅读，但其理论条件与本节简化推导应分别理解 [14]。

4.7 当优化器开始利用模型漏洞

模型利用漏洞（model exploitation）指策略或规划器通过优化找到模型高估回报、低估风险的动作，而不是找到真实环境里的好策略。它可能表现为接触前反复抖动以获得虚假奖励、驶入训练集中少见区域、选择模型低估的危险速度，或者把状态推到价值函数异常高的角落。

一个有效诊断流程应把“模型预测差”细化到闭环决策中的具体位置。

1. 在相同真实起点，保存规划器最高分动作串、一个行为策略动作串和若干随机对照；先比较预测分数，再比较真实执行后的回报。
2. 按模型评分分位数画出真实回报与预测回报的差。若优化越努力、模型分越高，真实结果反而越差，说明被优化的方向存在漏洞。
3. 把误差拆成奖励、终止、状态预测和末端价值；同时记录动作到数据支持集的距离、先验后验差异或集成分歧。任何一个不确定性量都需要用真实误差校准。
4. 固定真实交互预算，对比缩短 H、减少策略更新次数、增加真实校正、约束动作或引入保守惩罚。观察是哪一项改善了真实回报，而不只改善模型损失。

不确定性惩罚可以写成 $\hat{J}_H - \lambda_u U(A)$，但只有 U 与实际误差相关时才有帮助；多个同源模型共同犯错时，集成分歧也可能很低。更长 rollout、更多候选、更高想象更新比例都可能提高利用漏洞的能力。闭环实验必须保留真实评估，不能只在同一世界模型里给自己打分。

练习与提示。

1. 手算 $H = 2$、$d_t = 1$ 的式 (4.2)，确认终止步奖励保留而后续项消失。再把边界改为数据截断，说明为什么可能需要 bootstrap。
2. 对式 (4.5) 取 $H = 2$，展开 G_0^λ，核对 $\lambda = 0$ 与 $\lambda = 1$。提示：末端 $V(x_{t+2})$ 的系数包括两步继续折扣。
3. 对标量线性系统 $F(s, a) = \alpha s + a$，模型 $\hat{F}(s, a) = (\alpha + \Delta)s + a$，在有界状态域中给出 ε_F。提示：先规定 $|s| \le M$，于是可用 $|\Delta| M$；不规定状态范围便没有这个一致界。
4. 假设某次 CEM 计划的末端价值极高，删除该项后计划不再被选中。设计真实环境检查区分“确有长期收益”和“价值漏洞”。提示：在统一起点执行短前缀，再用定义 V 的后续策略继续，分别检查中间状态分布和剩余真实回报。

第 5 章　模型家族：学习对象与规划接口

5.1　一段历史，几种持续存在的选择

Dyna 将学习、规划和反应组织在同一架构中，模型生成的内部经验可以参与价值更新 [9]。此后系统辨识、概率动力学、潜变量学习和搜索形成了多条互相影响的路线；综述可提供分类框架，却不能替代各方法的任务条件 [15]。

2018 年的 World Models 展示了 VAE、混合密度循环网络和控制器组合；PETS 用概率动力学集成和轨迹采样进行控制 [1, 13]。PlaNet 将概率潜状态与在线规划连接；Dreamer 将行为学习放进潜空间想象 [2, 3]。MBPO 强调从真实状态分支的短模型 rollout，MuZero 强调搜索所需量的学习 [4, 14]。

后续的 token、扩散与预测表征路线没有消除这些问题，而是重新选择了压缩方式、预测目标及推理计算的位置。比较方法时先固定问题，再问模型把哪些困难转移给了训练数据、表示、搜索或价值函数。

5.2　概率模型：PETS、PlaNet 与 Dreamer

PETS 的模型集成用多个概率网络描述一步动力学，轨迹采样在规划中传播状态不确定性。集成成员之间的差异可作为参数不确定性的一种近似，而每个成员的输出方差表达条件随机性。二者并非严格的贝叶斯后验；若所有成员共享训练盲区，也可能共同自信地出错。PETS 的关键接口是给定候选动作序列，产生多个状态粒子并计算回报 [13]。

PlaNet 从像素学习 RSSM，在线规划时只在潜空间展开候选动作，再用预测奖励评分。确定性记忆与随机状态的分工、先验和后验的区别见第 3 章。多步潜状态一致性帮助训练时的过滤状态与开放环预测接近，但并不能保证所有动作序列上的预测正确。

Dreamer 进一步训练 actor，使执行时不必每一步都运行大规模动作序列搜索。策略训练使用模型中的奖励和价值目标，而非逐步生成图片。DreamerV2 采用离散随机表示并讨论 KL 平衡和策略梯度；DreamerV3 使用一组提高跨任务训练稳健性的设计 [11, 12]。这些版本的性能差别不能简单归因于某一种潜变量类型：损失、优化、归一化、计算预算和任务集也会变化。

DreamerV3 的预印本初版为 2023 年，2025 年以不同题名正式发表于 Nature[16]。论文版本、代码提交和权重版本应分别记录；正式出版年份不能直接用来给早期实验或实现重新定年。

怎样读 Dreamer 的计算图

先用三种颜色标出模型损失、价值损失、策略目标。再检查：策略更新时是否更新动力学参数？critic 目标是否停止梯度？随机状态和动作通过何种估计器传播梯度？只说“端到端反向传播”不足以复现具体算法。

SimPLe 则提供了另一种像素模型辅助 Atari 学习的实例 [17]。它提醒读者：模型生成经验

的用途可以是训练外部策略，不必与潜空间 actor 更新绑定。任务的动作离散性、帧跳跃和模型重置方式会影响误差传播，应在相同环境约定下比较。

5.3 任务导向模型：MuZero 与 TD-MPC

MuZero 将观测历史映射为隐藏表示 x_t，用递归动力学预测下一隐藏状态与奖励，再由预测头给出价值和策略先验。一个简化写法是

$$x_t = e_\theta(H_t), \tag{5.1}$$

$$(\hat{r}_t, x_{t+1}) = g_\theta(x_t, a_t), \tag{5.2}$$

$$(\hat{v}_t, \hat{p}_t) = f_\theta(x_t). \tag{5.3}$$

这里 $\hat{p}_t$ 是搜索的策略先验，不是环境转移分布。训练的价值与策略目标来自实际回报及搜索结果；搜索接口不要求图像解码 [4]。这个设计对特定决策任务很有效，但隐藏状态未必能回答未训练的几何或新奖励查询。

TD-MPC 用时序差分学习把短时潜动力学、奖励和终端价值接在一起；在线规划由短时动作候选与价值补偿长期信息 [18]。表示一致性的目标可抽象为

$$\mathcal{L}_{\text{cons}} = \left\| g_\theta(x_t, a_t) - \text{sg}\left[e_{\bar{\theta}}(o_{t+1})\right] \right\|^2, \tag{5.4}$$

另有奖励回归和 TD 价值损失。此式只表达共同思想，并非 TD-MPC2 的完整损失与实现。TD-MPC2 在多任务规模、训练稳健性和规划实现上继续改进 [19]。复现时应分别记录模型前向次数、动作候选来源、价值集成与动作约束，不能把所有采样式 MPC 都称为同一 CEM 实现。

MuZero 和 TD-MPC 都重视决策量，但一个以树搜索为核心，一个以连续控制中的短时 MPC 为核心。它们不是按“是否重建像素”就能互相替换的算法。

5.4 离散 token 与序列动力学

设一帧被 tokenizer 编码为 K 个离散 token $v_{t,1:K}$。动作 a_{t-1} 在 o_t 生成之前施加，时间顺序应写成

$$p_\theta(v_{0:T} \mid a_{0:T-1}) = p_\theta(v_0) \prod_{t=1}^{T} \prod_{j=1}^{K} p_\theta(v_{t,j} \mid v_{<t}, v_{t,<j}, a_{<t}). \tag{5.5}$$

训练通常最小化真实序列上的负对数似然。开放环生成时，把自己采样的 token 放回上下文；tokenizer 的微小误差和条件分布的偏差都会传播。更多上下文可以保存历史，却不能保证模型保留与控制相关的速度、接触和身份信息。

IRIS 将离散视觉 token、Transformer 动力学与模型内策略学习结合 [20]。iVideoGPT 将视频预训练及交互数据组织到可扩展的自回归建模框架 [21]。数据的动作格式、预训练阶段和后训练阶段应分别说明，不能因名称含“交互”就假定所有视频数据都带真实动作。

Trajectory Transformer 学习状态、动作和奖励序列的联合分布，再用序列搜索规划 [22]。Decision Transformer 则主要学习回报条件的动作生成 [23]。后者通常没有显式环境转移模型；把一个回报条件策略一概称为动力学模型，会混淆“选动作”和“预测动作后果”两种接口。

5.5 两种扩散用途：轨迹计划与观测模拟

对整段轨迹 τ，扩散训练构造

$$\tau_n = \sqrt{\bar{\alpha}_n}\tau_0 + \sqrt{1-\bar{\alpha}_n}\epsilon, \qquad \mathcal{L} = \mathbb{E}\,\|\epsilon - \epsilon_\theta(\tau_n, n, c)\|^2, \tag{5.6}$$

其中 n 为去噪时间，不能与环境时刻 t 混用；c 是初始状态或目标等条件。Diffuser 对轨迹分布建模，通过条件约束和引导生成行为计划 [24]。原论文的回报引导不应被泛化写成所有后续 classifier-free guidance 的实现。生成轨迹后还需按具体方法解码或提取动作，执行首步并重新条件化。

扩散也可直接建模动作条件的下一观测，形成能连续 rollout 的环境模型。DIAMOND 是这一路线的实例 [25]。轨迹扩散与扩散环境模型的训练对象不同：前者联合拟合数据里的行为轨迹，后者需要在动作干预下预测观测。两者都面临数据支持问题，但不能使用完全相同的因果解释。

扩散采样能表示多峰未来，却不直接给出模型“知道多少”。从同一个模型得到十个不同未来，可能只是已知随机性的多种结果。要解释为数据不足导致的不确定性，还需要模型后验、集成或其他可检验构造。去噪次数也会影响控制周期；报告质量时要同时报告采样延迟。

5.6 预测表征：JEPA 的价值与边界

JEPA 使用上下文编码器、目标编码器与预测器。一个示意损失是

$$\mathcal{L}_{\text{feat}} = \mathbb{E}\left\|g_\theta(e_\theta(o_{\text{context}}), m) - \text{sg}\left[e_{\bar{\theta}}(o_{\text{target}})\right]\right\|^2, \tag{5.7}$$

其中 m 表示目标区域或位置，目标编码器常用移动平均更新。目标特征不是固定真值，防止表示塌缩依赖结构、掩码、更新规则和训练条件，不能把停止梯度本身当作普遍保证。

I-JEPA 的目标来自图像区域，V-JEPA 来自视频中的时空区域；它们主要提供预测表征 [6, 7]。V-JEPA 2 进一步研究动作条件模型和规划 [8]。把预训练的理解指标和后训练的机器人控制结果分开，才能看清动作数据究竟补上了什么。

特征距离可以忽略无关纹理，也可能忽略关键的接触失败。应测试冻结表示后奖励、终止和动作后果是否仍可预测，并对比较对象使用相同训练/评测预算。一个小的特征误差并不自动具有毫米或牛顿等物理单位。

5.7 把模型选择写成实验设计

表 5.1 家族间比较的正确起点

家族	训练对象	决策接口	首要诊断
概率状态动力学	状态转移分布	粒子采样、MPC	覆盖、校准、分布外误差

家族	训练对象	决策接口	首要诊断
RSSM	观测与任务头似然、KL	先验想象、actor 或 MPC	后验/先验落差、长时漂移
任务搜索模型	奖励、价值、搜索策略	树搜索	搜索增益和真实回报
TD-MPC	潜一致性、奖励与 TD	短 MPC 与价值自举	末端价值偏差、动作边界
自回归 token	离散序列似然	token rollout、序列搜索	tokenizer 信息损失、上下文
扩散轨迹	行为轨迹联合分布	条件采样、计划执行	可行性、支持集、去噪成本
预测表征	特征预测	另接动作预测器或策略	塌缩、任务信息、动作对齐
生成式交互系统	视频/几何条件分布	用户交互或外部控制	状态持久、可控与功能效用

表中是问题导向的分类，不是互斥的架构清单。一个系统可以同时包含视觉 tokenizer、循环记忆、扩散预测器和任务价值头。先确认哪些组件真正参与测试，再做替换实验，才能判断改进来自哪里。

5.8 练习：先固定条件再比较

1. 给定同一离线机器人数据，比较特征预测与像素重建。至少固定哪些条件？提示：编码器大小、数据切分、动作头预算、控制器和评测环境都可能混淆结论。
2. 为什么回报条件动作序列不必包含转移模型？构造两种环境，它们在数据中行为相同，却对未出现的动作有不同后果。
3. 若 token 数减半、tokenizer 重建误差略升、控制变好，能否断言丢弃的信息无用？提示：计算节省也可能让策略在同一时限内搜索更多候选。

第 6 章　从数据到可用的训练循环

6.1　先把一条转移存对

模型训练之前，先检查数据中的一步究竟有多长。一条转移至少包含观测时间、动作施加时间与持续时间、后继观测、奖励和终止。如果机械臂命令在图像拍摄之后 80 毫秒才生效，却把它与当前图像的下一帧机械对齐，模型会把延迟误学为动力学。这种错误常比网络结构选择影响更大。

推荐存储

$$(\mathrm{id}, t, o_t, a_t, r_t, d_t, \tau_t^{\mathrm{cut}}, o_{t+1}, \Delta t), \tag{6.1}$$

其中 τ_t^{cut} 表示时间上限或采集截断标记。实现中用完整字段名 `truncated`；真实终止用 `terminated`。成功、失败、传感器丢包和人工停止最好再记录独立原因，不能全压成一个布尔变量。

数据序列必须按环境轨迹分段，RSSM 记忆应在真正重置时清零。训练切片若从轨迹中间开始，可用一段不计损失的历史作为预热；完全从零状态开始却把片段首帧当作已知完整历史，会产生训练与执行不一致。预热本身也消耗算力，应计入预算。

一条可以立即运行的数据审计

检查时间戳是否单调；同一轨迹最后一条记录的终止原因是否可解释；动作单位、限幅与频率是否固定；训练与验证是否包含同一段视频的重叠窗口。随机抽取“动作很大但下一帧几乎不变”和“动作不变但画面突变”的片段，人工确认延迟、遮挡、相机变化与重置。

6.2　数据划分是模型定义的一部分

连续视频的相邻帧高度相关。按帧随机划分会让验证集几乎重复训练集，掩盖对新轨迹的误差。至少按完整轨迹分割；若报告跨对象、跨场景或跨动力学泛化，还需在这些因素上留出。调参使用验证集，最终测试集应在结构与超参数确定后才用于报告。

多域数据还需记录观测与动作的映射。例如人的手部运动、机械臂末端位姿和关节扭矩处在不同空间；不能把它们拼接为同一数值向量并期待模型自行发现单位。可显式使用域、执行机构和时间步长条件，但条件 token 的存在不意味着已经消除域差异。应在目标执行机构上用真实动作校准。

离线数据由行为策略产生。数据中某类动作从未在某状态附近出现时，模型的反事实预测缺少直接约束。扩大模型容量不能凭空补足覆盖。离线强化学习的支持集问题也适用于基于模型的规划 [26]。可用行为约束、模型集成和保守目标降低风险，但必须通过独立回测检验，不能把“低集成分歧”当成覆盖充分的证明。

6.3 表示选择先问哪些信息可以丢掉

像素编码器有利于视觉重建，但可能把大量容量用于背景。离散 tokenizer 便于序列建模，却可能合并速度或接触差异。对象槽位可以支持组合，但对象发现错误会污染动力学。预测表征减少了像素生成负担，也可能忽略对某任务关键的小变化。没有一种表示能在不指定任务与数据的情况下保证“恰好保留需要的信息”。

表示验收应分成三层。第一层看信息：冻结编码器，能否从独立数据预测速度、接触、奖励和终止？第二层看动力学：在同一动作序列下，潜预测和可测输出能否保持一致？第三层看功能：固定控制预算，真实回报与失败率是否改善？线性 probe 的失败不等于信息完全缺失，可能是信息以非线性形式编码；probe 成功也不等于控制器能有效使用该信息。

不同模型的潜空间坐标没有天然对应。直接比较两个模型的 latent MSE 没有统一单位。较稳妥的是在同一冻结表示内比较先验与后验，或比较可测任务量；若引入对齐映射，需要在训练数据拟合，并在独立数据检验，不能用测试标签任意旋转坐标后再声称预测提升。

6.4 模型训练：损失、掩码与数值稳定性

第 3 章推导了严格的变分目标。实现中常采用加权目标

$$\mathcal{L} = \lambda_o \mathcal{L}_o + \lambda_r \mathcal{L}_r + \lambda_d \mathcal{L}_d + \lambda_{\mathrm{KL}} \mathcal{L}_{\mathrm{KL}} + \lambda_m \mathcal{L}_{\mathrm{multi}}. \tag{6.2}$$

这是一份训练设计，而不是无需说明便成立的概率等式。观测损失若对全部像素求和，与先取像素平均相差一个分辨率因子；改变图像大小却保留相同 KL 权重，会改变信息瓶颈的相对强度。奖励范围、终止类别比例、序列长度和有效掩码也会改变各项梯度尺度。

终止头可用 Bernoulli 负对数似然；若正例极少，应同时报告正例召回与概率校准。类别重加权会改变最优概率估计，需在独立验证集重新检查校准。将终止头与碰撞头合并并不总合适：一次轻微碰撞可能不终止任务，时间上限也可能没有危险。

对填充序列使用有效步掩码，保证填充区既不贡献重建损失，也不更新记忆。对跨轨迹的拼接块，重置掩码必须切断递归状态与梯度。批内平均分母应使用有效步数，避免短序列因填充比例而得到不同权重。

必要的稳定性记录包括梯度范数、被裁剪比例、学习率、数值非有限次数、先验/后验熵、每维 KL 和活跃潜维度。KL 接近零可能表示后验塌缩，也可能表示当前任务几乎可由确定性记忆预测。必须结合潜变量消融、重建与控制变化解释。单独追求“较大的 KL”同样可能只是后验携带了先验无法预测的纹理噪声。

6.5 从真实轨迹开始想象

一套可复现的训练循环应明确三个更新频率：每收集多少真实转移更新模型多少次；每次模型更新后更新策略多少次；每隔多少真实步做独立评测。只报告训练总步数，无法区分“用了更多数据”和“把同一数据重复优化得更多”。

想象期间动力学参数通常不由策略目标直接更新，但动力学对输入的导数仍可能参与 actor 的路径梯度。二者不能都简化成“冻结模型所以没有梯度”。具体梯度路径见第 4 章。

算法 3 分开记账的模型与行为更新

输入：真实交互上限 N_{env}、模型更新预算 N_m、策略更新预算 N_π

初始化模型、actor、critic、回放库与独立评测种子
while 真实交互与计算预算尚未用完 **do**
    用当前控制器执行真实动作，记录下一观测、奖励与终止原因
    从完整轨迹采样带预热的序列，过滤得到后验状态
    用带掩码的模型目标更新动力学，记录各项损失与稳定性
    从真实历史的后验状态采样起点，固定当前模型参数
    使用先验递推动作条件想象；记录继续概率和模型分歧
    按明确的价值目标更新 critic，按所选估计器更新 actor
    在独立真实环境种子上评测，计入评测交互与计算预算
    保存检查点、数据版本和模型内外表现差异
end while

6.6 短分支、长期记忆与多步训练

短 rollout 从真实过滤状态出发，减少连续消费错误预测的次数；长记忆则保留历史信息。这两个长度可以不同：一个模型可以用长历史形成状态，再只向前想象几步。训练时序列长度、记忆预热长度、规划长度和想象策略长度应分别记录。

多步一致性也有不同形式。用未来真实观测形成后验，再监督从当前状态出发的开放环先验，是潜空间多步训练的一种选择。若未来后验已经看过目标观测，它只能作为训练目标，不能在评测时被当成开放环预测。scheduled sampling 会改变训练输入分布与优化含义，不是严格似然训练的无条件替代。每次增加一项多步损失，都应检查一步质量、长时误差和闭环收益是否共同改善。

6.7 系统预算与复现实验包

表 6.1 训练与评测至少分开的四类预算

预算	推荐单位	必须补充的条件
真实数据	环境步、控制秒、完整轨迹	帧跳跃、重置、评测和探索是否计入
优化计算	更新步、样本数、设备小时	批大小、序列长、精度、硬件与软件版本
想象计算	模型转移数、候选数、去噪调用数	是否解码图片、是否共享缓存
执行与评测	单步延迟、峰值内存、回测步	同步/异步、预热、初始状态、失败重试

至少保存配置文件、代码版本或源码哈希、数据切分清单、随机种子、依赖版本、命令和逐次评测结果。均值与误差条要能从原始文件重新计算，不能只保留最后一张图。若实验因为预算停止，记录停止规则；挑选“最好种子”后当成平均表现会破坏比较。

本书的线性控制实验没有训练神经网络，不能用它的 CPU 用时推断 Dreamer 或视频模型的

成本。第 10 章把已运行的教学实验和未运行的 RSSM 训练规范分开，正是为了避免将不同规模的证据混用。

6.8 练习与检查点

1. 写出一个 batch 的张量形状与有效步掩码。哪些时刻必须重置 h_t？若序列结束只是采样窗口结束，是否应把 d_t 设为 1？
2. 把图像边长翻倍时，像素求和损失与 KL 的相对权重怎样变化？提示：像素数变为四倍，并检查通道数和分布方差是否也变了。
3. 同样 10 万环境步的两个系统，一个每步优化 1 次，另一个 100 次。怎样分别报告样本效率与计算效率？

第 7 章　表示、因果与失败诊断

7.1　从一次失败追到可检验的假设

一个控制器在模型里预测抓取成功，实际却把杯子推倒。可能的原因包括：单帧看不见杯子的速度；潜表示忽略了细小的接触；训练数据没有覆盖该推力；终止头漏掉了倾倒；价值函数对陌生潜状态过度乐观。仅凭最终回报无法确定哪一个解释正确。

诊断应固定其他环节，逐步替换。先以真实状态替换图像编码，判断感知和记忆是否是瓶颈；再用真实一步转移替换学习动力学，判断模型误差的贡献；再增加规划预算或使用已知可行控制器，区分搜索失败与模型失败。这些替换是诊断工具，使用了特权状态或真动力学时必须标明，不能与可部署系统混报。

7.2　决策充分性：保留的是哪些未来

先设 $x_t = e(H_t)$ 是历史的确定性压缩；随机表示应按式 (2.12) 检查后验预测。一个很强的充分性条件是：对于所有允许的后续动作干预，未来用于反馈的观测、奖励和终止的联合分布仅依赖 x_t，而不再依赖原始历史，并能以新反馈递归更新这个表示。仅要求当前奖励可预测则过弱；模型还可能把未来需要不同动作的历史合在一起。

任务相关的充分性并不要求恢复全部环境细节。若背景颜色从不影响任何奖励、约束或转移，可以忽略；若将来任务变成按颜色分类，原来的表示可能不再充分。因而“有用表示”必须带上任务族和允许的动作集。单任务价值监督无法自动保证新任务迁移。

可操作的训练信号包括奖励/终止预测、短期潜一致性、隐藏变量监督和多任务目标。它们只是代理约束。要验证充分性，应构造相同表示附近但历史不同的样本，检查动作后的结果是否分叉。第 2 章的两状态算例就提供了一种最小状态混叠：同一当前观测对应不同后验，单帧表示因此可能给出错误动作。

7.3　Bisimulation：何时可以合并状态

在有限、完全可观测、奖励有界的 MDP 中，记 $\rho(s,a) = \mathbb{E}[r_t \mid s_t = s, a_t = a]$，并考虑状态等价关系 $\sim$。若对任意 $s \sim s'$ 和每个可用动作 a，有相同的期望即时奖励，且转移到每个等价类 C 的概率相同，

$$\rho(s,a) = \rho(s',a), \tag{7.1}$$

$$\sum_{y \in C} P(y \mid s,a) = \sum_{y \in C} P(y \mid s',a), \qquad \text{对所有等价类 } C, \tag{7.2}$$

则这个划分具有 bisimulation 的结构：抽象后仍可保留相应的风险中性折扣决策问题。终止应作为吸收状态或明确的继续机制进入转移。对于风险敏感任务，只比较期望奖励还不够；通常还须按风险准则约束奖励与后继抽象状态的联合分布，单独匹配奖励边缘分布也未必足够 [27]。

度量形式用奖励差异和后继分布的距离递归定义近似相似性，例如

$$d(s,s') = \max_a \big\{ |\rho(s,a) - \rho(s',a)| + \gamma W_d\big(P(\cdot \mid s,a), P(\cdot \mid s',a)\big) \big\}, \tag{7.3}$$

其中 W_d 使用 d 作为运输代价。在有限状态、有限动作与 $0 \le \gamma < 1$ 的条件下，该算子是有界函数上的压缩映射，可由迭代得到唯一不动点。它可以是伪度量：不同但可合并的状态距离为零。神经表示中的欧氏距离并不会因使用这个名字就满足递归关系。Deep Bisimulation 的表示目标将相关思想用于学习对控制有用、对干扰相对不敏感的表示 [28]，但有限数据下的损失小不构成所有动作和分布外状态的保证。

一个具体反例

两个外观相同的位置都有即时奖励 0，但一个位置的“前进”通向目标，另一个通向陷阱。仅以当前奖励一致训练会把它们合并；检查后继分布才能发现区别。若数据从未在第二个位置执行“前进”，甚至转移一致性损失也可能漏掉问题。这既是表示问题，也是动作覆盖问题。

7.4 可观测性与可控性是不同条件

考虑已知线性系统

$$s_{t+1} = As_t + Ba_t, \qquad o_t = Cs_t. \tag{7.4}$$

在无噪声且已知输入的情形，扣除动作的已知贡献后，连续 n 个观测包含关于初始状态的线性方程。可观测矩阵为

$$\mathcal{O}_n = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}. \tag{7.5}$$

其列满秩意味着这段观测可以区分全部 n 维初始状态。若只观察位置而隐变量包含速度，连续位置观测可能恢复速度；只看一帧通常不能。若观测通道从不包含某个独立隐变量的影响，延长窗口也无济于事。

可控矩阵 $\mathcal{C}_n = [B, AB, \dots, A^{n-1}B]$ 则回答动作能否在该理想线性模型中到达任意状态方向。知道状态不等于能控制它；能够控制某个方向也不等于传感器能观察它。约束、噪声、非线性、时变性会改变实际问题，不能把一个局部线性秩检查当作复杂机器人全局可控的证明。

实际诊断可以改变历史长度、遮挡方式与可用动作集合，测量隐藏变量 probe、预测误差和真实回报。改变历史长度同时改变计算量时，需要另设等计算对照。主动感知还会让动作影响后续信息：照亮一个区域或移开遮挡物，短期奖励可能低，却让未来 belief 更确定。普通确定性开环计划可能忽略这种信息价值。

7.5 动作、干预与因果识别

一种结构方程写法是

$$s_{t+1} = f(s_t, a_t, u_t), \qquad o_t = g(s_t, v_t), \tag{7.6}$$

其中 u_t, v_t 是外生噪声。将 a_t 设为指定值，是对动作机制的干预；仅在视频中观察某种手势，则可能同时观察到了场景、意图和拍摄方式的变化。

若状态或历史足够充分，动作选择机制在给定这些信息后没有未控制混杂，并且数据覆盖待查询动作，条件转移可以用于相应的干预预测。缺少这些条件时，$p(o_{t+1} \mid o_t, a_t)$ 不能直接解释为“主动执行动作”的因果分布。离线数据中的人类意图尤其可能同时影响动作与结果。

潜动作还存在语义不可辨识：交换潜动作编号可能不改变生成似然；相机运动和对象运动也可能有多种解释。逆动力学能约束“相邻帧之间发生了什么”，却不能单独保证潜动作与扭矩一一对应。需要少量已标定真实动作、多视角同步、时间连续性与执行测试共同建立接口。

动作互换是一种诊断：固定相似初始历史，将模型输入从动作 a 换成 a'，再用真实重复试验比较结果。它只验证支持集内的局部响应；若无法从同一状态重置，初始状态差异必须进入不确定性分析。个体层面的反事实还要求对同一实例的隐噪声进行推断或耦合，不能把两次独立采样直接当作同一个世界的唯一反事实。

7.6 不确定性：多样、未知与失配

把 x 作为给定的查询输入，采用明确的参数后验 $p(\theta \mid \mathcal{D})$，并假设预测模型满足 $Y \perp \mathcal{D} \mid (x, \theta)$，则总方差有分解

$$\operatorname{Var}(Y \mid x, \mathcal{D}) = \mathbb{E}_\theta \operatorname{Var}(Y \mid x, \theta) + \operatorname{Var}_\theta \mathbb{E}(Y \mid x, \theta). \tag{7.7}$$

第一项是给定参数后的条件随机性，常称偶然不确定性；第二项来自参数未知，常称认知不确定性。这个等式需要有限二阶矩和明确的概率模型。若当前状态本身不确定，还应对其后验积分；固定 x 后的参数分歧不能替代状态估计的不确定性。一般深度集成只近似第二类变化，不能声称采到了真实后验；模型类别本身错误造成的失配也未必体现在这两项的估计中。

扩散或混合模型的多个未来可以表达行人左右转向的多峰结果，但同一模型多次采样不等于对模型未知性的估计。相反，所有集成成员都在没有见过的冰面上训练不足时，可能一致预测干燥路面的制动距离。这时低分歧与低风险完全不同。

可检查的指标有概率负对数似然、Brier 分数、预测区间覆盖与宽度、风险-覆盖曲线。应同时按距离训练分布的程度和失败类型分层。把高不确定样本拒绝后，必须报告被拒绝的比例、剩余风险以及拒绝后的真实执行策略；拒绝全部样本可得到低风险，却没有任务效用。

7.7 误差与漏洞如何在闭环里放大

对相同动作序列，误差可能沿动力学传播；对闭环策略，模型误差还会改变动作选择。第 4 章给出了带条件的误差递推。接触与终止是非光滑事件，状态误差即使很小，也可能让事件判断翻转。因而除了平均位置误差，还要测量穿透、碰撞漏检和终止时间偏移。

优化器还会主动寻找模型擅长“说谎”的区域。增加候选动作数或规划长度，可能提高模型内回报，同时降低真实回报。诊断时固定模型，画出真实回报、模型回报和预算的关系；这比只看一次最佳生成视频更能暴露模型利用漏洞。

表 7.1 从症状到干预的失败诊断

症状	竞争性解释	优先检查	可尝试的修正
同帧、不同历史需要不同动作	状态混叠或记忆不足	固定当前帧，交换前史	延长有效历史、潜状态过滤
一步准、长时漂移	误差传播或训练分布转移	匹配动作开放环、多时域误差	多步目标、短分支与真实重置
模型回报升、真实回报降	奖励偏差、价值外推、搜索漏洞	固定模型增加规划预算	行为约束、保守价值、失败回放
生成未来很多但漏掉失败	多样性不覆盖风险	事件召回、区间覆盖、条件校准	尾部数据、风险头、校准
动作互换后结果不变	动作输入被忽略或无可辨识效应	支持集内真实动作对照	动作标签、标定、补覆盖
新相机下崩溃	视角捷径或几何不一致	多视角同场景、遮挡后重访	几何约束、对象记忆
KL 低但控制稳定	可预测任务或随机潜变量无效	去掉随机状态后的性能	简化模型或重配表示
KL 高但先验很差	后验记住不可预测细节	先验/后验解码与预测对照	损失尺度、瓶颈与数据审计

上述修正都是待验证的设计选择，不构成安全保证。一个可执行的保护层需要规定作用域、可检查约束和回退策略，且它自身的传感与估计误差也要评测。对没有覆盖的危险场景，保守地报告证据不足比给一个缺乏校准的低风险分数更准确。

7.8 练习与提示

1. 构造两个期望奖励相同、奖励分布不同的状态。它们对风险中性与风险敏感控制分别能否合并？提示：比较高概率小损失与低概率大损失。
2. 对二维位置-速度系统，令 $C = [1\ 0]$，写出两步可观测矩阵。若采样间隔趋近零，满秩与数值条件良好是否相同？
3. 为“所有 ensemble 成员一致但都错误”设计一个数据分布。应如何让校准测试覆盖这种失败？
4. 为什么 action swap 通过不意味着任意新动作均可预测？提示：覆盖、混杂与动作幅度都可能变化。

第 8 章　评测：从预测误差到闭环决策

一个模型生成的杯子清楚、背景连贯，规划器却可能把夹爪穿过桌面；另一个模型重建模糊，却能保留目标位置和接触状态，完成抓取。两种现象都合理，因为评测测量的是特定能力。世界模型评测的起点应当是用途：为视频生成挑选样本、为策略比较提供代理环境，还是在固定真实交互预算下提高控制回报？用途决定测试分布、指标及可接受的误差，而不是反过来让一个现成的总分定义用途。

本章保留原稿的感知、预测、干预、功能四层结构，并把它扩展成能够执行的协议。每一层都提供下一层所需的部分证据，但不存在由低层分数自动推出高层能力的定理。尤其需要分清两种结论：在已经观察到的动作下预测准确，和在规划器挑选的动作下仍然可靠。后者会主动接近模型的薄弱区域。

8.1　先冻结问题，再选择指标

设一条测试轨迹为 $\tau = (o_0, a_0, r_0, d_0, o_1, \ldots)$。在时刻 t 执行动作 a_t 后产生 r_t, d_t, o_{t+1}；$d_t = 1$ 表示该次转移终止。真实状态 s_t 只有在可访问的仿真器或教学实验中才能用作评价标签。没有真实状态时，不能把视觉编码器的输出重新命名为真实状态，再宣称测得物理误差。

完整评测应先写一页协议，至少固定五个对象：模型及检查点、决策器、初始状态分布、交互接口、资源预算。决策器包括规划长度、候选数、优化迭代数、价值函数与风险惩罚。把同一个模型交给两个不同的规划器，控制回报变化可能来自搜索而非动力学；把一个模型配上更强的预训练视觉编码器，也应计入模型系统的资源和监督。

数据按**轨迹或采集会话**划分，不能把同一视频相邻帧随机分散到训练集和测试集。更严格的泛化测试按对象、场景、时间、任务、动力学参数或机器人实体留出。各划分回答的问题不同：同一场景的新轨迹测插值；新质量或摩擦系数测动力学迁移；新对象组合测组合泛化；新动作空间测控制接口迁移。论文若只测其中一种，就应只声称该种泛化。

表 8.1　四层评测对应的输入、测量和结论边界。

层次	推荐测量	所能支持的结论
感知	单帧与视频质量、时序闪烁、人类偏好；固定采样数	给定提示和采样设置下的可见质量
预测	多步状态或观测误差、奖励和终止概率、概率评分	指定动作与起始历史下的条件预测能力
干预	动作替换、相机变化、对象属性改变；成对轨迹	被测干预范围内的响应与一致性
功能	真实闭环回报、成功率、风险、策略排序、迁移	指定任务、决策器和预算下的决策效用

感知评测中的 PSNR、SSIM 或视频特征距离可以作为诊断，但随机未来不一定与录制的一条轨迹逐像素相同。对一团向两边散开的烟，只生成其中一个合理方向会产生很大的逐像素误差；输出两个方向的平均值反而可能降低均方误差。要同时看分布评分、事件统计和样本多样性。多采样后只报告最好的样本还会隐藏推理成本，因此必须写清是单次采样、平均结果，还是从多少次尝试中挑选。

8.2 预测质量要沿时间展开

开放环预测先用真实历史完成后验推断，再仅输入预定动作序列并递推模型，不能每一步偷看真实下一帧。对于可访问状态的系统，令 $\hat{s}_{t+k}^{(j)}$ 为第 j 个模型样本，在统一单位和坐标后，可报告

$$E_k = \frac{1}{N}\sum_{i=1}^{N}\left\|\bar{s}_{i,t+k} - s_{i,t+k}\right\|_2^2, \qquad \bar{s}_{i,t+k} = \frac{1}{M}\sum_{j=1}^{M}\hat{s}_{i,t+k}^{(j)}. \tag{8.1}$$

这是预测均值的误差，不是完整预测分布的质量。若能够计算似然，应另报测试负对数似然；若只能采样，可测覆盖率、事件频率或适用于该变量的样本评分。不同量纲应分别报告，或者使用**仅由训练集确定**的标准化尺度，不能让误差最大的变量任意支配总分。

选取多个物理时间点，例如动作周期为 0.1 秒时的 $k = 1, 5, 10, 20$，分别对应 $0.1, 0.5, 1, 2$ 秒。跨模型比较时若动作频率不同，只比较相同步数会误导。除均值曲线外，报告高分位误差、接触或碰撞等关键事件附近的误差，以及随初始历史长度变化的结果。后者能够检验模型是否确实使用了过去信息来消除部分可观测性。

闭环评测则每执行一步接收新的真实观测，重新估计潜状态并规划。它的误差来源同时包含观测编码、潜状态更新、动力学、价值估计和搜索。开放环长预测差而闭环控制可用，可能是反馈修正及时；开放环录像预测好而闭环失败，可能是测试动作过于接近数据策略。两个结果都应保留，不能用一个替另一个作结论。

对视觉任务，尽量给出可复核的中间事件，例如物体中心、抓取接触、目标进入容器、路口红绿灯与碰撞。事件提取器也会出错，应独立报告其误差；用与世界模型共享训练数据的提取器时，要检查共同偏差。没有传感器或标注支持的量，应标为代理测量。

8.3 奖励、终止与不确定性不能附带处理

模型的画面误差小，并不保证累计回报误差小。令预测的继续概率为 $\hat{c}_t = 1 - \hat{p}(d_t = 1)$。一个从 x_t 出发的有限长度目标通常形如

$$\hat{G}_{t,H} = \sum_{k=0}^{H-1}\gamma^k\left(\prod_{j=0}^{k-1}\hat{c}_{t+j}\right)\hat{r}_{t+k} + \gamma^H\left(\prod_{j=0}^{H-1}\hat{c}_{t+j}\right)\hat{V}(x_{t+H}). \tag{8.2}$$

空乘积取 1。错误的继续概率会使终止后的价值泄漏，也可能把仍可挽救的状态提前判死；误差作用于整个回报，而不只是某一帧。教学环境如果有时间上限，应区分环境真正终止和采样被截断，后者是否 bootstrap 必须遵循任务定义。

式 (8.2) 是给定想象路径的计分形式；随机模型还需对路径取期望。将独立预测头的均值直接相乘，额外假定它们保留了奖励、终止与下一状态所需的条件依赖。若这几种量共享随机结果，应使用一致的联合采样或相应的条件期望，具体边界见第 4 章。

奖励头要按任务量纲报告偏差、均方误差和高奖励区域的误差。对于成功与失败、碰撞和终止等二元事件，设概率预测为 p_i、标签为 y_i，可以使用 Brier 分数

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{8.3}$$

及二元对数评分；数值实现需对概率作预先公布的截断以避免 $\log 0$。还应给出事件基率、召回率、精确率及阈值。若失败率只有千分之一，总是预测不失败也有很高的分类准确率，却没有预警价值。

校准指预测概率与相应事件频率是否一致。把失败概率分桶后比较桶内平均预测与实际频率，可以画可靠性图；每桶同时写样本数和区间，避免把极少样本的折线当成规律。ECE 依赖分桶且不是严格适当评分，不宜单独优化或作为唯一指标。校准参数在验证集确定，在冻结的测试集上报告，不能用测试标签重新调温度。

对连续变量，给出例如 50%、80%、95% 的预测区间覆盖率与平均宽度：总覆盖率高但区间几乎无限宽没有实际价值。对风险筛选，画风险随保留样本比例变化的曲线，并和随机拒绝对照比较。分布内校准也不自动保证新场景中的校准，应按留出场景和高风险切片分别复测。模型集成的分歧是诊断信号，不能未经验证就称为真实风险概率。

8.4 动作干预：从相关预测到可操作响应

给定相同历史，分别施加 a_t 和 a'_t，比较模型预测的差异与环境实际差异，是最直接的动作条件化检查。例如同一小车状态下左右转向应产生相应的轨迹偏移；相机转动不应被误认为物体受力；机器人伸手与缩手不应只改变手臂纹理而保留相同接触结果。

在能够恢复状态与随机数的仿真器中，使用共同初始状态和共同外生随机噪声，构造成对干预；把对象位置、速度、接触和回报的变化分开比较。在真实环境中通常无法精确恢复同一个世界，应通过重复试验、受控实验或明确的识别假设限制结论。观察到的数据相关性不能取代干预证据。

动作置零、交换、延迟和打乱是廉价诊断。如果打乱动作几乎不改变测试误差，可能是动作被忽略，也可能是测试片段本来就不受该动作影响。要选择动作敏感的片段，并检查数据支持范围。对离散潜动作，需要先验证潜动作与真实执行命令之间的映射；从视频学到可控制的变化模式，仍不等于获得机器人关节的接口。

多视角一致性也应作干预测试。同一时刻的多个相机应看到可对应的对象身份和几何；相机离开房间再回来时，对象状态不应因未被观察而随意重置。可测重访定位误差、对象身份保持率、几何循环误差和隐藏后再出现事件。这里的相机位姿、时间同步与坐标变换误差都需校准，否则评测器本身会制造“不一致”。

8.5 闭环价值与模型利用漏洞

把模型接入规划器，才能看到搜索是否放大微小偏差。在相同初始状态集合上，至少比较无模型基线、使用学习模型的决策器和可用时的真实动力学上界。上界应保留同样的搜索预算；允许真实模型无限搜索会同时改变模型与算法两个因素。

逐回合记录真实回报、模型预测回报、成功、失败类型、真实交互数和推理时延。两种回报的差可定义为

$$\Delta(\pi) = \mathbb{E}_{\hat{P},\pi}[G] - \mathbb{E}_{P,\pi}[G]. \tag{8.4}$$

期望应使用相同的初始分布与回报定义。若模型中越成功的策略在真实环境越差，优先检查是否出现模型利用漏洞：例如无限反复获得一次性奖励、忽略碰撞、借终止误判取得虚假的 bootstrap，或利用训练集中未覆盖的极端动作。

固定模型、逐渐增加候选动作数和优化迭代数，是一种有用的压力测试。若预测回报继续升高而真实回报下降，更多搜索正在发现模型漏洞。此时不能仅用更强正则把下降藏起来，应公开两条曲线，并比较动作约束、较短预测长度、置信惩罚、补采数据和独立验证模型的效果。MBPO 对短模型 rollout 的论证解释了限制模型使用范围的价值，但不提供任意环境中最优长度的通用常数 [14]。

安全结果应按明确事件统计：发生接触不等于危险碰撞，失败不等于造成损害。报告事件定义、暴露时间、严重性及人工接管规则；“零次观测到事故”必须连同测试回合数或运行时长给出。有限测试不能证明风险为零，更不能把合成环境中的零事故直接转成真实系统的保证。

8.6 策略排序和统计不确定性

用世界模型筛选策略时，绝对成功率可能有偏，但相对排序仍然有用。WorldGym 的研究对象正包含模型代理环境中的策略评价和真实表现的相关性 [29]。评价这种用途应冻结一个**未用于挑选世界模型**的策略集合，覆盖不同算法、模型大小、检查点及失败模式；只选择几个非常相似的策略，会高估排序的稳健性。

除了 Spearman 或 Kendall 相关，报告错误排序对数、前若干名的真实表现、选择最优策略造成的遗憾，以及近似并列时的结论。对策略 π_i, π_j，可以直接报告成对差值

$$\delta_{ij} = \mathbb{E}[G(\pi_i) - G(\pi_j)] \tag{8.5}$$

的区间。差值区间跨零时，通常应称为证据不足，而不是强行排出第一与第二。若两个策略在同一初始状态上测试，按初始状态成对重采样；若有多个训练种子和任务，保留层次结构，按训练运行及任务重采样，不能把同一轨迹的所有帧当作独立样本。

一个手算警示。 某策略在 10 个独立回合中成功 8 次，点估计为 80%；在二项模型假设下，95% Wilson 区间约为 49% 到 94%。这是统计公式的**教学算例**，不是本书实测。与另一个成功 7 次的策略比较时，仅凭 80% > 70% 无法作可靠优劣结论；如果回合难度相关或训练种子很少，实际不确定性还可能更大。

平均回报应同时给出跨训练种子的离散程度和评价回合带来的不确定性。对多任务结果，预先确定任务权重，先报告各任务，再给汇总；避免在看到测试结果后改用对本方法最有利的均值、中位数或归一化方式。多次比较或反复试探测试集应如实说明。模型在有限策略集合中排序正确，只能支撑该集合及相应任务中的代理评价价值。

8.7 现有基准各自测到了什么

表 8.2 根据本次核验的论文摘要及原文定位整理，不把不同数据、交互接口和评分器的数值拼成总榜。基准名称中的 “world model” 也不能替代对协议的阅读。

表 8.2 现有评测的测量对象与使用边界。

基准	核心测量对象	需要补充的证据
SVIB[30]	因素组合划分下的一步视觉变换与系统泛化	长时 rollout、动作干预及真实闭环
WorldSim-Bench[31]	显式感知评价与隐式操控评价，含视频到动作一致性	接触真实性、长期任务和独立真实系统验证
WorldModel-Bench[32]	指令遵循、可见物理违规与人类偏好标注	不可见动力学、执行动作接口及控制回报
EWMBench[33]	场景一致性、运动正确性和语义对齐	按任务定义的奖励与终止、真实迁移
WorldGym[29]	动作条件视频环境中的策略 rollout 及排序	换策略族、换机器人、评分器偏差与外推校准

例如 WorldSimBench 的隐式操控评价包含从生成视频恢复控制信号的环节，该评价受到逆动力学模块质量影响。WorldGym 使用视觉语言模型提供奖励，世界模型的错误和奖励评估器的错误可能互相强化。读者应问：所见收益来自预测器、控制器、逆动力学，还是裁判？只换其中一个组件的消融，通常比堆叠一个总分更有解释力。

8.8 使用视觉语言模型裁判时

视觉语言模型（VLM）可把“杯子是否进入碗中”转成自动评分，降低人工查看成本，但不是物理真值。遮挡、视角、时间抽帧和语言偏见都能影响裁判。协议应保存模型提供方、可辨认版本、访问日期、提示词、输出解析、温度、输入帧选择和失败重试规则。远程服务发生版本变化后，旧结果不应无条件与新结果混合。

建立一个与开发集分开的人工标注子集，并覆盖容易失败的反例：夹爪经过物体但未抓住、对象暂时被遮挡、手臂挡住容器、画面好看但动作无效。采用盲评和重复标注，报告标注者一致性、裁判对每类事件的精确率及召回率。只报告与整体人类偏好的相关性，会掩盖稀有安全错误。

测试顺序偏差、提示词改写、视频长度和画质偏差；成对评价交换左右位置。对被评模型生成的数据，检查裁判是否偏好与自身训练分布接近的风格。若裁判奖励参与训练或筛选，则最终评测应有独立裁判和人工抽检，避免把优化裁判本身当成学习世界规律。遇到无法判断的帧应允许“不确定”，并公开弃权比例及其处理方法。

8.9 把预算写进结果表

真实交互预算、数据预算、训练计算与推理计算是不同坐标。真实环境转移数应包含初始数据收集、在线补采和用于调参的额外交互；环境帧数与动作决策数在 action repeat 下不同，需分

别声明。使用公开预训练权重时，给出可获得的预训练规模，并说明无法知道的资源；不能把“本次微调只用少量轨迹”写成“整个系统只见过这些轨迹”。

训练成本至少记录设备型号与数量、精度、批大小、优化更新数、有效序列长度、墙钟时间及峰值显存。模型训练、策略训练、验证与超参数搜索应分别列出。推理成本报告规划长度、候选数、模型采样数、扩散去噪步数、价值调用数和控制频率；必要时给出时延的中位数与高分位，而不是只报平均吞吐量。

表 8.3 可直接用于实验登记的最小结果字段。空白项应填“未测”或“不适用”。

类别	必须保留的记录
数据与划分	采集版本、轨迹划分哈希、泄漏检查、留出因素、动作单位与周期
真实预算	初始转移、在线转移、调参交互、评价回合和终止/截断规则
训练预算	模型与策略的更新数、独立种子、硬件、时间、显存、预训练来源
推理预算	候选数、规划长度、采样/去噪次数、重规划频率、时延分位数
预测与校准	多时间点误差、事件基率、概率评分、覆盖率、区间宽度
闭环与统计	各种子回报、成功与安全事件、成对差值、排序与置信区间
失败追踪	原始轨迹、预测轨迹、动作、奖励与终止头、评分器输出和代码版本

最后应保留预注册的主要指标，同时展示揭示失败的辅助指标。例如一个模型在平均回报上提高，却在稀有碰撞切片上恶化，后者应进入结论。世界模型的实用性来自可解释的证据链：在什么数据支持下、以多少资源、对什么动作和任务、在哪些统计不确定性内有效。下一章用这一证据标准阅读近期系统，避免把产品演示、论文报告与本书实测混为一谈。

练习与检查

1. 某模型下一帧误差低，但 20 步预测很差；闭环控制却与基线相当。提出至少两种不同解释，并设计能够区分它们的日志。提示：检查真实反馈的重规划频率和动作分布。
2. 失败概率预测全部为 0.01，测试集中恰有 1% 失败。说明为什么整体校准仍不能证明模型能识别危险状态。提示：构造两个失败率不同的子群。
3. 设计一个世界模型策略排序实验，明确训练种子、评价初始状态、留出策略、重采样单位和并列处理。说明增加评价回合为何不能替代增加独立训练种子。
4. 解释为什么截断轨迹时把终止标签全部置为 1 会改变式 (8.2)，并给出一个会受影响的持续控制任务。

第 9 章　基础模型、交互生成与三维世界的接口

前几章从一个任务、一个数据集和一套动作接口出发。基础世界模型尝试把规模更大的视频预训练转用于多个环境；交互生成系统则让用户或智能体在生成过程中持续行动。它们扩大了可研究的环境范围，也使“到底学到了什么”的证据更难解释。本章沿用第 1 章的六变量契约，分别检查预训练、动作适配、几何表示和下游验证。

本章的内容整理截至 2026 年 10 月 22 日。下面所列材料的逐项核验日期见来源核验记录，最晚为 2026 年 10 月 22 日；这是一组有明确来源边界的代表性材料，并非对该日前全部新工作的穷尽清单。模型家族名、论文版本、软件发布版和在线服务不是同一对象，复现时应继续锁定实际使用的权重、代码提交与接口。

9.1　基础模型首先是一个迁移承诺

基础世界模型（world foundation model）通常指通过大规模、多来源数据预训练，再适配到下游环境预测任务的模型。这个称呼没有规定唯一的潜状态、损失函数或动作空间。一个模型可能擅长按文字生成驾驶视频，另一个可能预测机器人未来特征；二者都可以采用预训练与后训练，但不能因此直接交换状态或控制器。

把迁移过程写开，会得到至少三个不同问题：

1. **观测迁移**。新相机、视角、光照或对象进入模型后，原有表示是否仍保留任务变量？
2. **动力学与动作迁移**。新机器人的动作单位、执行频率、延迟和约束是否与预训练条件兼容？是否需要新动作头或动作条件后训练？
3. **决策迁移**。即使短期预测可用，新奖励、终止条件和价值函数是否已被正确接入？

应分别报告这三层所用的数据和预算。仅冻结一个预训练编码器，并不意味着整个控制系统没有使用目标域数据；只在新背景上测试，也不能证明动力学迁移。

Cosmos 的平台论文把视频整理、tokenizer、预训练模型与领域后训练放在同一流程中，其定义强调可被定制的通用世界模型 [34]。这里引用的是 2025 年 1 月首次发布、2025 年 7 月修订的论文版本。它可以说明平台的组成与设计目标，不能代替某个后续软件发布版的模型卡，也不能证明所有下游任务都具有可靠的奖励或终止接口。

表 9.1　主要前沿材料的版本与证据类型

材料	固定的论文或发布日期	本书采用的证据范围
Cosmos	2501.03575v3；2025-07-09 修订	平台论文中的预训练、tokenizer 和后训练流程；不外推到同名后续产品
Genie	2402.15391v1；2024-02-23	潜动作、视频 tokenizer 和动力学模型，以及论文内的交互实验

材料	固定的论文或发布日期	本书采用的证据范围
Genie 2	官方研究介绍；2024-12-04	发布方披露的图像起点、键鼠交互、潜扩散机制与演示
Genie 3	官方研究介绍；2025-08-05	发布方披露的实时交互、重访与局限；不当作独立复现
iVideoGPT	2405.15223v3；2024-10-31 修订	token 序列、动作/奖励接口及论文指定的下游实验
WorldDreamer	2401.09985v1；2024-01-18	掩码 token、多模态提示与驾驶动作条件视频展示
V-JEPA 2	2506.09985v1；2025-06-11	无动作表征预训练与 V-JEPA 2-AC 动作后训练的分阶段证据
GAIA-1	2309.17080v1；2023-09-29	视频、文字和驾驶动作条件的序列生成
DriveDreamer	2309.09777v2；2023-11-27 修订	结构交通条件与驾驶视频生成的指定实验
TesserAct	2504.20995v1；2025-04-29	RGB-DN 生成、四维重建与论文内具身任务
Cosmos-Drive-Dreams	2506.09042v3；2025-06-18 修订	驾驶合成数据流程及论文报告的下游评测
Sora 技术报告	官方技术报告；2024-02-15	生成机制、展示能力及官方报告的失败；非独立闭环控制验证

表 9.1 的版本日期可以晚于首次预印本年份，因此文献表同时保留初版日期与核验版本。官方网页还可能持续更新；引用网页中的运行特性时，应保留访问日期，并区分发布方演示、论文实验和独立测量。本书没有对这些大型系统重新训练或进行性能复测。

9.2 交互生成：输入变化之后，环境怎样回应

Genie 使用时空视频 tokenizer、动力学模型和潜动作模型，从无真实动作标签的视频中学习可交互环境 [5]。其关键意义在于：训练时没有电机指令标签，也可以学到能稳定改变生成未来的控制变量。但潜动作的编号与物理控制量之间存在标定问题。一个编码能让角色向右移动，不等于它已对应某个机器人的速度或关节力矩；环境重力、碰撞和任务奖励也需要另行验证。

数据的探索策略也会改变可学到的交互。GenieRedux 相关工作比较了以训练后的智能体产生数据的方案，并使用可获得动作的变体区分潜动作推断带来的影响 [35]。这提示我们把数据采集策略纳入模型定义：随机动作可能到不了关键状态，熟练策略也可能只覆盖成功路径。更清晰的生成与更好的局部可控性，仍不能替代对未覆盖动作和失败区域的评测。

Genie 2 的官方介绍将系统描述为自回归潜扩散模型：给定起始图像，用户或智能体持续提供键鼠动作，模型据此前进。介绍中展示了物体离开视野后重现等记忆现象，并明确区分基础模型生成的示例与可实时运行、画质有所变化的蒸馏版本 [36]。因此，在比较延迟和画质时必须锁定具体版本，不能把两组特性拼成一个未经测量的系统。

Genie 3 的官方介绍进一步强调实时导航、较长交互一致性和文字触发的环境事件 [37]。发布方同时列出动作范围受限、多智能体交互仍有困难、真实地理复现不精确、交互时长有限等局

限。文字改变天气与智能体通过动作影响天气，是不同的干预接口；前者不能自动成为后者可执行的动作。官方演示提供了值得检验的能力线索，可靠模拟器仍需独立的重置、分支、长时误差、事件概率与策略测试。

“可以玩”还没有确定什么

交互环境至少有一个输入通道能够影响未来。用于训练或评测策略，还需要明确奖励、真正终止、时间步长、动作延迟和重置语义。若这些量由外部程序补上，应把外部程序也列为系统的一部分。判断成功若依赖另一个视觉模型，其误判会进入策略回报，不能把生成器的自然画面当作成功标签。

iVideoGPT 把视觉观测、动作与奖励组织成 token 序列，通过下一个 token 预测建立交互接口，并分别评测动作条件预测、视觉规划和基于模型的强化学习 [21]。它说明可扩展的序列模型可以连接到实际决策流程，但具体证据仍受训练数据、任务和预算约束。WorldDreamer 的掩码 token 建模展示视频生成、编辑与多模态条件能力，也在相同初始帧下展示了左转、右转等驾驶动作条件视频 [38]。这提供了动作条件响应的证据，仍需独立的闭环控制验证；多模态提示也不等同于持续的低层控制。

同样，Sora 官方技术报告标题中的“world simulators”应结合其视频生成证据阅读 [39]。视频能表现某些物体持续性和场景变化，值得进一步研究；要声称可用于任意闭环策略评测，仍需本书第 8 章所要求的动作、奖励、终止和模型外验证。

9.3 预测表征怎样接上动作

JEPA 路线把预测目标放在学习到的特征空间。它可能减少对难以预测的像素细节的建模负担，但“哪些细节可忽略”仍由表示、数据和训练目标共同决定。V-JEPA 的视频表征实验与 V-JEPA 2 的动作条件规划应分开看 [7, 8]。

V-JEPA 2 先进行无动作的视频/图像预训练，再以机器人交互轨迹训练 V-JEPA 2-AC 的动作条件预测器；论文报告了以图像目标进行规划的机器人实验。这里的机器人后训练轨迹包含状态与动作记录，不能把论文所说的“无标注”误读为“没有动作信号”。其目标不是给每一帧人工写语义标签，不意味着控制接口凭空从被动视频中辨识出来。

论文中目标环境的零样本部署，也应按它所说明的数据边界理解：没有在报告的目标机器人环境中再采集任务数据，不等于无需机器人交互预训练、动作表示、相机处理或执行约束。判断迁移时应列出哪些部分冻结，哪些部分训练，以及新实验究竟改变了对象、背景、动作机制还是机器人形态。

若以目标特征距离作为规划分数，例如

$$\hat{J}(A) = -\left\| \hat{y}_{t+H}(A) - y_{\text{goal}} \right\|^2, \tag{9.1}$$

其中 y 是固定编码器输出，那么规划器优化的是该编码器定义的距离。距离小不必然代表物体已稳定抓住：遮挡、相机视角或不可达目标都可能造成误导。需要用任务成功、接触与失败事件检验目标距离的含义，并与使用相同表征的较简单控制基线比较。

9.4 三维、四维与持久记忆解决哪类问题

二维画面把相机投影、对象几何、遮挡和运动混在一起。向左移动相机与向右移动对象，都可能让图像内容向右移动；单视角视频未必能唯一解释二者。三维表示引入深度、坐标和跨视角对应，四维表示再把三维结构随时间的变化纳入模型。这里的“四维”通常是三维空间加时间，不意味着已获得完整的力学状态。

一个接口层面的设计是令内部状态同时保存几何、对象、动力学和记忆分量，并分开写转移与成像：

$$w_t = (g_t, e_t^{\mathrm{obj}}, v_t, m_t), \tag{9.2}$$

$$w_{t+1} \sim p_\theta(w_{t+1} \mid w_t, a_t), \qquad o_t \sim p_\theta(o_t \mid g_t, e_t^{\mathrm{obj}}, c_t^{\mathrm{cam}}). \tag{9.3}$$

这里 g_t 表示几何，e_t^{obj} 表示对象属性，v_t 表示影响运动的变量，m_t 表示额外记忆，c_t^{cam} 是相机参数。w_t 是一种结构化内部状态的示意，区别于前面 RSSM 的 $x_t = (h_t, z_t)$。该分解不主张这些变量天然可辨识或彼此独立；也不能把相机参数与第 4 章的继续变量 $c_t = 1 - d_t$ 混用。

显式场景有利于检查重访和跨视角一致性，但重建出一个几何场不等于学会它如何响应动作。静态 NeRF 或三维高斯可以用于渲染观测；若要成为决策系统中的世界模型，仍需动态更新、动作接口、任务量和不确定性。相关表示与渲染的完整推导见《NeRF 与 3D 高斯泼溅》分册，本书关注这些模块向规划器提供什么信息。

TesserAct 通过联合生成 RGB、深度和法线（RGB-DN）视频，再将其转为四维场景，连接几何预测、逆动力学与具身任务 [40]。论文中的一部分真实数据深度和法线来自已有估计器，因此几何通道不是无误差真值。深度、法线和 RGB 由共享数据及模型产生时，它们的相互一致也可能只是共同偏差；应保留有独立几何真值的测试集，以及真实执行结果。

检验持久性可以从一个具体回路出发：看见桌上杯子，转向墙面，再回头。比较杯子的身份、位置、朝向和接触关系；再在遮挡期间施加已知扰动，检查模型是否把“不可见”误当作“静止”。对随机环境，只要求合理的条件分布，不能强迫模型预测未被观测、也无法推知的唯一结果。记忆的评价对象应是有证据支持的状态和不确定性，而非总能补出一幅清晰画面。

9.5 合成数据的价值要在下游测量

GAIA-1 将视频、文字和动作输入用于驾驶场景序列生成；DriveDreamer 将结构化交通约束接入驾驶视频建模 [41, 42]。这些工作中的场景条件、驾驶控制和任务接口需要逐项阅读。给生成器一条道路布局，不能直接解释为改变油门后的物理响应；同时生成一段合理动作，也不自动验证该动作能在真实车辆上执行。

Cosmos-Drive-Dreams 研究可控、多视角驾驶合成数据，并报告在指定感知与驾驶学习设置中的下游收益 [43]。这类证据比只展示视频更接近实际用途，但收益仍是数据配方、模型、任务和评测分布共同作用的结果。复用时至少建立以下对照：真实数据基线；真实数据加合成数据；在可行时增加等量真实数据；以及计算预算相近的普通数据增强。应单独标明难以做到的公平条件。

设下游评分为 S，在固定协议下可以报告

$$\Delta_{\mathrm{syn}} = S(\mathcal{D}_{\mathrm{real}} \cup \mathcal{D}_{\mathrm{syn}}) - S(\mathcal{D}_{\mathrm{real}}). \tag{9.4}$$

这个差值回答的是该配方的下游增益，不是合成场景符合真实风险频率的证明。若为覆盖罕见事故而刻意过采样，生成分布本来就不同于真实分布；训练可以受益，但评测风险时需要真实频率或明示的重加权，不能把合成事件占比当成事故概率。

9.6 把展示能力变成可检验的接口

表 9.2 基础与交互模型的功能检查

声明	最小检查	仍需控制的因素
动作有效	匹配历史下交换方向、幅度和持续时间，比较真实响应	数据支持集、坐标系、延迟与混杂
长期一致	固定轨迹长度，做遮挡、转视角、重访和分支	采样随机性、重置状态、评价器误差
几何可信	对照独立深度、相机姿态和接触真值	估计器共享偏差与尺度不确定性
能评价策略	在真实环境比较多类策略的回报与排序	策略族覆盖、漏洞搜索、排序区间
合成数据有用	固定真实数据与训练预算，测下游与长尾切片	额外算力、重复样本、测试泄漏
能迁移	明确更换的因素与允许的适配数据，测闭环结果	编码器、动作头、价值和执行器的变化

这些检验可能得到不同结论：一个模型可以很好地扩大视觉训练数据，却不适合给策略评分；也可以支持局部导航，却无法恢复精细接触。保留这种用途差异，比把所有系统放进一个总排名更有助于实验设计。

练习与提示。

1. 为一个文字生成、键鼠交互的环境填写六变量契约。奖励和终止若来自视觉裁判，设计一次裁判误判而策略“成功”的案例。
2. 一段 RGB、深度、法线视频完全相互一致，是否足以证明它是真实场景？提示：一起缩放场景与相机平移，或让三个通道共享错误几何。
3. 为“互联网视频预训练带来机器人迁移”设计四组对照，分开编码器初始化、动作后训练和目标域数据的作用。
4. 若长尾合成数据使事故识别召回提高，如何检查整体风险估计是否仍校准？提示：训练时的重采样比例不是部署时的事件频率。

下一章回到规模可控的实验。大型系统提出的动作、记忆和评测问题，可以先在有精确参考解的小环境中学习怎样验证，再进入昂贵的数据与模型。

第 10 章　最小实验：从后验到闭环控制

前面的推导说明模型怎样表示状态、怎样在想象中比较动作。本章把这些接口缩小到可以独立验算的规模：先用两个隐藏状态检查历史的作用，再在二维线性系统中分开测量预测误差和闭环回报。最后给出最小 RSSM 的训练与验收规范。

本章的证据边界

两状态过滤和线性 MPC 已在 2026 年 10 月 22 日复跑，是本书提供的合成教学实验；没有训练神经世界模型，没有运行 Dreamer，也没有接入真实机器人。RSSM 部分是明确标注的后续实验规范，尚无训练结果。固定解析模型的表现不能作为某种网络架构的性能证据。

10.1　先验证后验：同一帧不等于同一状态

10.1.1　完整地定义这个小环境

设隐藏状态 $s_t \in \{0,1\}$，初始分布为 $\mu = (1/2, 1/2)$。本例只用一个空动作，状态在 episode 内保持不变，转移与观测矩阵分别为

$$T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad O = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}, \quad O_{s,o} = P(o_t = o \mid s_t = s). \tag{10.1}$$

不同时间的观测在给定隐藏状态后条件独立；奖励恒为零、终止标记恒为零，本例只研究推断。T 的行对应旧状态、列对应新状态。先接收 o_0；之后每一步先转移、再接收下一观测。滤波程序因此区分第一次观测更新与后续预测 – 校正：

$$b_0(s) = \frac{O_{s,o_0}\mu(s)}{\sum_u O_{u,o_0}\mu(u)}, \tag{10.2}$$

$$\bar{b}_{t+1}(s') = \sum_s T_{s,s'} b_t(s), \qquad b_{t+1}(s') = \frac{O_{s',o_{t+1}}\bar{b}_{t+1}(s')}{\sum_u O_{u,o_{t+1}}\bar{b}_{t+1}(u)}. \tag{10.3}$$

考虑两段历史 $o_{0:2} = (1,1,1)$ 与 $(0,0,1)$。它们的当前观测都是 $o_2 = 1$，但前者支持 $s = 1$，后者的两次早期证据支持 $s = 0$。逐步后验见表 10.1。

表 10.1　本次合成实测：两条历史对同一当前观测给出不同后验。

观测历史	$P(s_0 = 1 \mid o_0)$	$b_1(1)$	$b_2(1)$
$(1,1,1)$	0.800000	0.941176	0.984615
$(0,0,1)$	0.200000	0.058824	0.200000

最后一列还可以不用滤波程序而直接手算：

$$P(s=1 \mid 1,1,1) = \frac{0.8^3}{0.8^3 + 0.2^3} = \frac{64}{65}, \qquad P(s=1 \mid 0,0,1) = \frac{0.2^2\,0.8}{0.2^2\,0.8 + 0.8^2\,0.2} = \frac{1}{5}. \tag{10.4}$$

只输入当前帧的确定性编码器必然给这两段历史相同表示，因此无法同时输出两个正确后验。这就是状态混叠的一个可定位实例：丢失的是历史证据，增加单帧分辨率不能补回它。若传感器完全准确，当前帧已确定状态；若传感器完全无信息，历史也不能创造证据。记忆的价值取决于环境与观测机制，而非“历史越长越好”的口号。

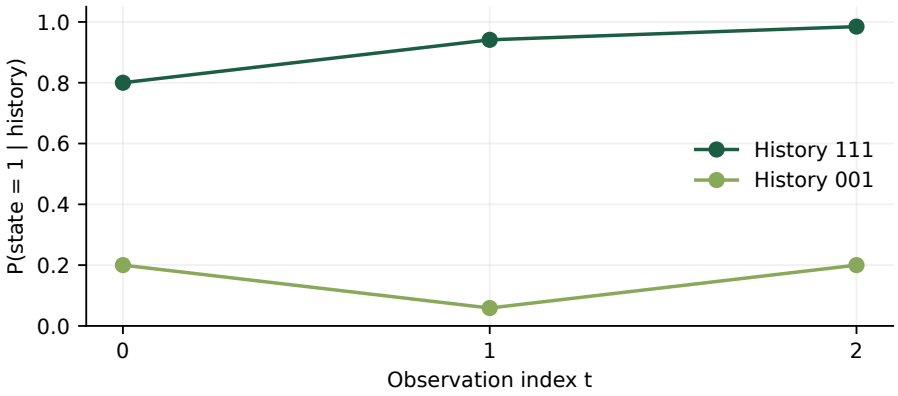


图 10.1 本次实测的后验轨迹。横轴从 o_0 开始；两条轨迹最后都观测到 1，但后验仍明显不同。图中英文标签对应历史 111 与 001。

10.1.2 独立参考解比自我复查更有用

程序另写了一份不调用滤波器的参考解，穷举长度 T 的全部 2^T 条隐藏状态路径，计算

$$P(s_{0:T-1}, o_{0:T-1}) = P(s_0) O_{s_0,o_0} \prod_{t=1}^{T-1} T_{s_{t-1},s_t} O_{s_t,o_t}, \tag{10.5}$$

再按末状态求和、归一化。除上述静态例子，还测试保持概率为 0.9 的转移，以及翻转概率为 0.9 的动作转移；三种转移各枚举全部 32 条五步二元观测历史，共 96 个交叉校验案例。最大容许绝对误差设为 10^{-13}，不是靠打印到三位小数判断相等。

边界检查还包括无信息观测、完美观测与确定性翻转动作，以及零证据历史。后者必须显式报错：若状态恒定、传感器完美，却收到 $(0, 1)$，该模型下观测概率为零，随意补一个“均匀后验”会掩盖模型与数据矛盾。这些检查验证实现和假设，不把数学上不可能的历史解释成普通不确定性。

10.2 把预测误差与控制回报分开测量

10.2.1 真环境、固定模型与计分规则

第二个实验观察完整真状态，刻意去掉感知与滤波误差，单独研究动力学与规划长度。令 $s_t = (p_t, v_t)^\top$，状态和动作均为无量纲合成变量：

$$s_{t+1} = A_* s_t + B_* a_t + \epsilon_t, \quad A_* = \begin{pmatrix} 1 & 0.2 \\ 0 & 0.96 \end{pmatrix}, \quad B_* = \begin{pmatrix} 0.02 \\ 0.2 \end{pmatrix}. \tag{10.6}$$

过程扰动为独立高斯 $\epsilon_t \sim \mathcal{N}(0, \mathrm{diag}(0.005^2, 0.01^2))$。观测为 $o_t = s_t$；动作不设幅值约束。奖励在执行动作后计算：

$$r_t = -\left(s_{t+1}^{\top} Q s_{t+1} + R a_t^2\right), \qquad Q = \mathrm{diag}(1, 0.1), \quad R = 0.08. \tag{10.7}$$

没有真正的失败终止，故 $d_t = 0$；每次 80 步仅是评测截断。回报为这 80 个奖励的未折扣和，越接近零越好。没有训练数据、拟合步骤、终端价值函数或隐含的奖励修正。

三组模型都是预先给定的解析对照。精确模型采用 (A_*, B_*)；“半增益”模型采用 $(A_*, B_*/2)$；“漂移失配”模型采用

$$\hat{A} = \begin{pmatrix} 1 & 0.28 \\ 0 & 1.12 \end{pmatrix}, \qquad \hat{B} = B_*. \tag{10.8}$$

它们的参数个数相同，因此本实验没有测量网络复杂度、参数学习能力或样本效率。设置失配只是为了制造可以解释的预测偏差，不预设哪一个规划长度应当获胜。

10.2.2 先做匹配动作的开环预测测试

对每个评测种子，用独立数据随机生成 128 个初始状态 $s_0 \sim \mathcal{N}(0, I)$ 和长度为 32 的标准高斯动作序列。真环境与每个模型共享同一组起点、同一组动作，且此处关闭过程噪声。模型只在起点看到真状态，此后消费自己的预测。对 $k \in \{1, 2, 4, 8, 16, 32\}$ 计算

$$\mathrm{RMSE}(k) = \sqrt{\frac{1}{2N} \sum_{i=1}^{N} \| \hat{s}_k^{(i)} - s_k^{(i)} \|_2^2}. \tag{10.9}$$

十个种子的 RMSE 再求均值与样本标准差。该指标排除了控制器选择不同动作造成的分布差异，测量的是固定动作下的模型递推误差。

表 10.2 本次实测的匹配动作开环状态 RMSE，表内为十个种子的均值。精确模型关闭过程噪声后为零是代码与真环境一致的结果，不是学习成果。

模型	$k = 1$	$k = 4$	$k = 8$	$k = 32$
精确	0	0	0	0
半增益	0.0698	0.1458	0.2431	1.0219
漂移失配	0.1239	0.6541	1.9221	66.4009

一步误差有限并不保证多步可用。漂移失配模型把速度衰减错写成增长；其 $\hat{A}$ 存在模长大于 1 的特征值，自身预测会放大速度偏差。这条受控曲线展示的是特定线性失配的后果，不能由此推断所有世界模型的误差都指数增长。

10.2.3 解析 MPC：排除搜索随机性

每次 MPC 在当前真状态上求解

$$\min_{a_{t:t+H-1}} \sum_{k=0}^{H-1} \left(\hat{s}_{t+k+1}^{\top} Q \hat{s}_{t+k+1} + R a_{t+k}^2 \right), \qquad \hat{s}_{t+k+1} = \hat{A} \hat{s}_{t+k} + \hat{B} a_{t+k}, \tag{10.10}$$

只执行第一动作，再用新观测重新求解。无动作约束且 $R > 0$，该问题是严格凸二次优化，可以精确求解，因而不会把候选采样不足误认为模型误差。

从末端 $P_H = 0$ 向后递推，令 $W_j = Q + P_{j+1}$，则

$$K_j = (R + \hat{B}^\top W_j \hat{B})^{-1} \hat{B}^\top W_j \hat{A}, \tag{10.11}$$

$$P_j = \hat{A}^\top W_j \hat{A} - \hat{A}^\top W_j \hat{B} K_j, \qquad a_{t+j} = -K_j \hat{s}_{t+j}. \tag{10.12}$$

注意 W_j 中的 Q 来自下一状态计分，而非直接套用当前状态计分的另一套公式。实现同时独立组装全部未来状态的块矩阵，把动作序列的一次项与二次项展开后直接解线性方程；在六种 H、每种五个随机状态上核对完整动作序列，共 30 个独立参考案例。

由于模型和 H 在 episode 内固定，增益序列可缓存。本次脚本仍逐步展开整段预测，以记录规划终点误差；实际控制动作只需要第一增益与当前状态相乘。预算应同时说明这种结构，不能把缓存后的线性 MPC 速度与每次重新采样的神经规划器比较。

10.2.4 相同起点与扰动下的闭环结果

对种子 0 至 9，分别采样

$$s_0 \sim \mathcal{N}\left(\begin{pmatrix} 2 \\ 0.4 \end{pmatrix}, \operatorname{diag}(0.5^2, 0.3^2) \right). \tag{10.13}$$

同一种子上的全部模型与全部 H 共用此起点及整段 80 步过程扰动。共运行 $10 \times 3 \times 6 = 180$ 个 episode。表 10.3 中“±”是 seed 间样本标准差，既不是标准误，也不是置信区间。

表 10.3 本次实测：80 步未折扣回报，十个种子的均值 ± 样本标准差。

H	精确模型	半增益模型	漂移失配模型
1	−58.192 ± 32.844	−94.965 ± 60.962	−57.291 ± 32.071
2	−25.927 ± 12.825	−37.610 ± 19.390	−23.129 ± 11.383
4	−17.390 ± 8.537	−18.864 ± 9.253	−18.145 ± 8.950
8	−16.916 ± 8.308	−17.279 ± 8.499	−19.343 ± 9.582
16	−16.915 ± 8.307	−17.612 ± 8.672	−19.164 ± 9.489
32	−16.915 ± 8.307	−17.617 ± 8.674	−19.165 ± 9.489

这些数字支持三个范围明确的观察。第一，精确模型的回报在 $H = 8$ 后几乎不变，在本任务继续增加预测长度收益有限。第二，失配模型的回报未随 H 单调提升：半增益模型从 $H = 8$ 到 16 的均值变差；漂移失配模型在 $H = 4$ 的结果好于 $H = 8$。第三，一步误差排名与短时控制排名也可能不同：$H = 1$ 时，漂移失配模型的一步开环误差较大，却在这组起点上取得略高于精确模型的平均回报。有限时域目标、模型失配与动作选择共同作用，不能只用一个预测指标排序控制器。

这里不据十个种子的均值宣称小差异具有普遍统计意义。初始状态差异造成的回报标准差较大，评估细微改进应进一步报告逐 seed 配对差值、更多种子与针对任务的置信区间。每次重

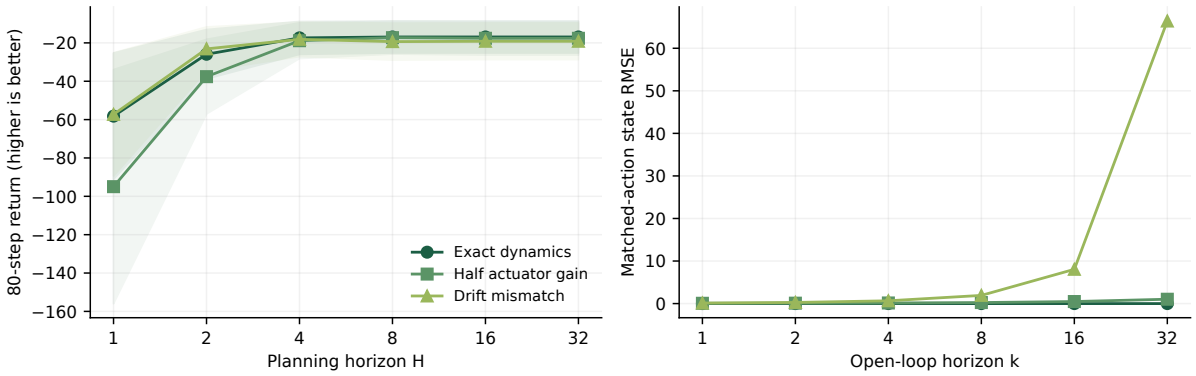


图 10.2 本次合成实测。左：不同规划长度的闭环回报，浅色带为 seed 间一个样本标准差。右：独立、匹配动作的开环误差；这不是左侧控制器各自产生轨迹上的误差。横轴均采用以 2 为底的对数刻度。

规划还会重置模型起点，因此右图的 32 步开环误差很大，并不意味着系统真的连续执行了这条错误的 32 步计划。闭环纠正能降低误差传播，但也不能保证在接触、约束或失效终止附近可靠。

脚本另记录闭环一步预测误差和规划终点误差。精确模型的闭环一步 RMSE 均值约为 0.00766，来自已开启的过程噪声；它与无噪开环测试的零误差并不矛盾。规划终点误差则把所选动作序列分别放入模型与无噪真动力学，不包含未来重规划的修正，应理解为当前计划的诊断量。

10.3 复现记录与预算怎样读

从本主题根目录运行下面的命令即可重算图表与全部 CSV：

```
python3 -m venv /tmp/jingfeng-wm-env
/tmp/jingfeng-wm-env/bin/python -m pip install \
  -r manuscript/experiments/requirements.txt
/tmp/jingfeng-wm-env/bin/python manuscript/experiments/run.py
```

本次实际环境为 Python 3.12.14、NumPy 2.3.5、Matplotlib 3.10.7、macOS arm64，使用 CPU；记录的脚本总墙钟时间约 0.96 秒，包含参考检查、数据收集、汇总与绘图，仅作为本机运行记录。十个种子、配置、源码哈希、完整命令、环境版本和预算写入本主题的独立实验输出目录。源码修改后重跑会更新哈希；默认会覆盖同名结果，保留历史运行时应指定另一个输出目录。

输出中的 `summary.csv` 为表格来源；`mpc_episodes.csv` 和 `mpc_trajectories.csv` 保留每个 seed 与每一步；`matched_cases.json` 保留起点与整段扰动哈希；`tests.json` 记录九组已通过的数值与边界检查。不是每一组都只测一个案例，例如路径穷举与二次优化参考解分别覆盖 96 与 30 个案例。时间日志包含诊断工作，不能当作纯规划器延迟；转移计数则把规划、诊断与评测分开列出，避免把廉价的合成调用混作学习样本。

表 10.4 本次实际计算预算。这里的“环境”是合成动力学函数，不是真实设备。

项目	数量
训练参数更新、训练环境交互	0
闭环评测 episode	180
闭环真环境步	14 400
规划模型转移	151 200
额外规划终点诊断真转移	151 200
开环评测真转移、模型转移	各 122 880
缓存增益的 Riccati 后向步	189

10.4 最小 RSSM：下一步应当怎样训练和验收

10.4.1 一个固定、可以实现的实验合同

适用边界

本节尚未执行训练。下列网络维度、种子和预算都是计划值，不属于本次实测。完整规范位于 `manuscript/experiments/rssm_protocol.md` 与 `rssm_protocol.json`；输出目录中的 `rssm_status.json` 明确记录训练更新数为零。

沿用线性真环境，但只向模型提供带 0.05 标准差观测噪声的位置。真速度只留给评估器，编码器与规划器均不可访问。行为动作独立均匀采样于 $[-2, 2]$；下一位置或速度绝对值超过 4 时设真正的 $d_t = 1$，附加奖励 -10 并重置；80 步时限是截断，不能混作失败终止。这是相对于上一实验另加的终止规则，必须重新生成数据，不能复用旧奖励标签。

分别以数据种子 20000、30000、40000 生成训练 16 384、验证 4096、测试 4096 个转移，按完整轨迹划分。训练窗口的重叠不能泄漏到测试集。取 $h_t \in \mathbb{R}^{16}$、$z_t \in \mathbb{R}^4$，用 GRU 将 (z_t, a_t, h_t) 映射到 h_{t+1}；先验只看 h_{t+1}，后验再看 o_{t+1}。观测头输入 $x_t = (h_t, z_t)$；奖励和终止头输入 (x_t, a_t)，分别预测 r_t, d_t，与全书时序约定一致。

先验、后验与输出头均使用 32 维单隐层 ELU 网络。随机潜状态使用对角高斯，标准差为 $\mathrm{softplus}(u) + 0.1$；观测与奖励似然标准差分别固定为 0.1 与 1。基线最小化有效位置平均的

$$\mathcal{L} = \mathcal{L}_o + \mathcal{L}_r + \mathcal{L}_d + \beta\ \mathrm{KL}(q_\phi \| p_\theta). \tag{10.14}$$

第一版不用 free bits 或 KL balancing，先确认普通目标与时序正确；否则一次加入多个稳定化技巧，难以解释各项变化。NLL 保留标准差与常数项，原始 KL 与加权 KL 分开写日志。普通负 ELBO 始终以 $\beta = 1$ 另算；$\beta < 1$ 的加权训练目标不称为严格负 ELBO。真正的重置、序列填充和不计分的 burn-in 都必须有独立掩码。

10.4.2 日志应能揭示失败，而不只留下一个总损失

重建损失低而先验预测差，常见于后验充分利用当前观测、先验却未学到可递推的状态；KL 极低也可能意味着后验忽略观测，不能直接称为表示更好。至少保留表 10.6 的信息，并按 seed

表 10.5 计划中的 RSSM 对照与训练预算，本轮未运行。

项目	固定方案
训练 seed 与 KL 权重	$\{0,1,2\}\times\{0.1,0.5,1.0\}$，共九次运行
更新预算	每次 1000 次；batch 为 16；每段 25 个计分转移，之前至多 5 步 burn-in
优化器	Adam，学习率 3×10^{-4}，梯度范数裁剪上限 100
验证与模型选择	每 100 步验证；按统一的验证负 ELBO（评估时取 $\beta=1$）选 checkpoint，测试集不用于选权重
预测评测	5 步真实历史确定起点，然后 prior-only 预测 $1,4,8,16$ 步；每个起点 32 个粒子
闭环评测	十个固定 episode 种子，动作限制 $[-2,2]$，最多各 80 步

单独出图。活跃潜维度可预先约定为后验均值在验证集上的方差超过 0.01 的维度数，同时报告连续方差值，避免只靠任意阈值判定塌缩。

表 10.6 RSSM 的必要日志及其诊断用途。

日志	用来回答的问题
观测、奖励 NLL 与终止 BCE	能否解释输入、预测任务量、识别真正失败；是否被类别不平衡掩盖
原始 KL、加权 KL、加权训练目标与普通负 ELBO	损失变化来自表示、先验对齐还是仅仅修改了权重
先验/后验标准差、活跃维度	是否出现数值退化、后验忽略观测或不确定性过小
分长度 prior-only RMSE、NLL、覆盖率	离开真实观测校正后还能预测多久；误差与不确定性是否匹配
逐 episode 真实回报、失败率	预测质量是否转化为控制能力；失败 seed 是否被隐藏
唯一真实转移、采样位置、想象转移	是否把重复训练样本或模型内部计算算作新的环境经验
实际更新数、秒数、设备、内存与哈希	是否在相同预算完成；能否定位具体 checkpoint 与软件环境

冻结表示与动力学后，用同一 CEM 规划预算比较“仅当前观测”“保留后验历史”和“真状态可见”三个控制接口。可固定 $H=8$，每轮 256 个候选、3 次迭代、32 个精英，每个候选用 8 个潜粒子；每步最多 $256\times3\times8\times8=49\,152$ 次潜转移。oracle 也应使用同一动作约束与 CEM 预算，而不是拿本章无约束解析控制器作直接对照。比较时匹配真实起点、过程噪声与规划随机数；不把较好的视频重建当作控制指标的替代品。

训练前先检验对角高斯 KL：$q=p$ 时结果应为零，再用独立 Monte Carlo 积分核对非平凡参

数；观测 NLL 与手算结果对齐。把未来观测替换为另一份文件，prior-only rollout 不应改变；修改 padding 的极端值不应改变有效损失；episode 重置后的轨迹不能反向改变前一 episode 的隐状态。最后分别测试真正终止与时间截断的回报掩码。这些检查与本章穷举后验、独立二次解遵循同一原则：先构造能推翻实现的参考证据，再讨论模型规模。

10.5 练习：尝试推翻一个看似合理的结论

1. 将静态两状态转移改为保持概率 0.5，再比较两条历史的最后后验。提示：此时转移会丢弃前一时刻的信息，解释“长历史总能改进估计”为何不成立。
2. 将奖励改为对当前状态计分，同时保留 Ra_t^2。提示：$H = 1$ 的最优动作变为零，原增益公式不能原样照搬；写出新的独立二次目标作核对。
3. 用已保存的逐 seed 回报计算半增益模型 $H = 16$ 相对 $H = 8$ 的配对差值；再将初始状态分布改为更大速度。提示：把结论限定在新的评测分布，不混用两次运行的均值。
4. 给动作加上 $[-2, 2]$ 约束，再比较较长规划长度。提示：直接裁剪无约束 LQ 动作一般不等于求解受约束 MPC，应使用独立约束优化器或动作序列穷举小例子检查。

本章的最小实验把后验、预测与控制连接起来，同时保留它们的区别。下一步研究可以更换模型和任务，但应继续保留可独立核验的时序、同条件的对照、完整的失败记录与分开的交互和计算预算。

第 11 章　研究工具箱：把直觉写成可证伪问题

读完一种模型的结构，容易产生“加入记忆会更好”“几何能改善泛化”“低比特可以加速”等直觉。研究的下一步是写出它在哪些条件下应成立，以及什么结果会让我们放弃它。本章把前面的数学、训练和评测组织成研究方案。这里提出的是后续工作设计，不是本书已经完成的神经模型实验。

11.1　阅读论文时先还原信息与计算流程

一张模型结构图常省略数据采集、环境重置和评价器。读论文时，先画出观测历史 H_t、模型状态、动作、预测量与真实执行之间的箭头，再将每个箭头对应到训练信号。以下问题可以分三次阅读回答。

第一次读清**问题与证据**：观测是什么，动作是否真实可执行，是否需要处理部分可观测性，回报和终止由谁定义；主要实验测量预测、干预、闭环控制还是迁移；所谓泛化究竟改变了哪个因素。第二次读清**学习与规划**：训练状态来自后验还是真实状态，想象时是否只用先验，模型学习的是像素、特征、奖励还是价值，规划器怎样消费这些量，末端价值对应什么策略。第三次读清**复现条件**：哪些预训练数据和权重被使用，真实交互和模型调用怎样记账，评价器是否独立，失败轨迹是否公开，关键结论是否有消融与随机种子支持。

若某一项无法从正文、附录或代码确认，把它记为未知。不能用该领域常用的做法替作者补齐实验事实。例如，看到“ensemble”不能默认其成员独立；看到“latent action”不能默认其单位等于电机指令；看到“zero-shot”也不能默认没有任何动作后训练。

一页论文记录

记录任务契约、预测对象、损失、规划接口、数据切分、真实交互预算、训练与推理预算、主要对照、失败条件，以及你无法确认的一项细节。最后用一句带条件的话重述结论，例如“在给定相机与动作范围内，加入多步一致性使指定时域误差下降，但长时控制收益尚未被检验”。这种记录比只写模型优缺点更便于后续复现。

进一步阅读可以先用综述定位问题，再回到原始论文核实机制和实验。《Understanding World or Predicting Future?》从表示理解与未来预测组织宽领域地图 [44]；物理模拟器与世界模型综述梳理外部仿真和内部预测在具身学习中的关系 [45]。机器人学习综述有助于查找模型与策略、模拟、数据生成和评测的接口 [46]，驾驶初始综述则提供领域问题与早期路线的入口 [47]。这些综述的覆盖范围与核验版本不同，它们是阅读导航，不能代替被引原论文中的实验条件。

11.2 先区分三种误差，再设计替换实验

系统表现变差，可能来自表示丢失了必要信息、动力学在当前分布上失配，也可能来自规划和执行。它们相互作用，不能一般地写成三个互不相关的误差项之和。一个很小的表示差异就可能改变动作，继而改变访问分布；终止事件又会使回报不连续。

更可操作的做法是设置受控替换：

1. 用历史或真状态替换单帧输入，在部署信息限制之外做诊断，判断记忆与感知的贡献。
2. 保持策略与初始条件，接入真一步转移，判断学习动力学是否是主要瓶颈。
3. 固定模型，增加规划预算或使用已知可行控制器，观察搜索误差与模型利用漏洞。
4. 固定上层指令，检查实际执行的动作、延迟和限幅，判断规划与执行是否在求解同一个问题。

这些结果提供机制证据，并不自动形成唯一的误差归因；替换环节可能同时改变分布和计算。应报告使用了哪些特权信息，以及替换后其余环节是否真的保持一致。

11.3 压缩误差为什么需要沿时间测量

世界模型的推理成本可能出现在图像编码、tokenizer、潜动力学、注意力缓存、集成成员和候选轨迹上。降低某一模块的位宽，既改变计算成本，也可能改变规划所见的未来。关于后训练量化（post-training quantization，PTQ）和量化感知训练（quantization-aware training，QAT）的具体算法，可结合《智能体模型量化压缩与部署》分册阅读；这里关注压缩后的动力学与决策。

11.3.1 固定动作下的局部误差递推

先考虑确定性潜空间动力学 F 及其压缩实现 F_q。为了比较，假定二者在相同的潜坐标、相同归一化与同一动作序列下递推，且潜状态从同一起点出发。定义

$$x_{k+1} = F(x_k, a_k), \qquad \tilde{x}_{k+1} = F_q(\tilde{x}_k, a_k), \qquad e_k = \|\tilde{x}_k - x_k\|. \tag{11.1}$$

这些 x_k 都是模型内部状态，不能据此把潜空间距离解释成真实物理误差。若压缩还更换了 tokenizer、潜维数或编码器，坐标可能不同，应通过明确对齐映射或共同的解码、任务预测量比较，不能直接相减。

假设在两条轨迹涉及的同一域内，F 对状态是 L-Lipschitz，且 $\|F_q(x,a) - F(x,a)\| \le \varepsilon_q$，则由三角不等式

$$\begin{aligned} e_{k+1} &\le \|F_q(\tilde{x}_k, a_k) - F(\tilde{x}_k, a_k)\| + \|F(\tilde{x}_k, a_k) - F(x_k, a_k)\| \\ &\le \varepsilon_q + L e_k. \end{aligned} \tag{11.2}$$

因此在 $e_0 = 0$ 时，

$$e_k \le \varepsilon_q \sum_{j=0}^{k-1} L^j = \begin{cases} \varepsilon_q (1 - L^k)/(1 - L), & L \ne 1, \\ k\varepsilon_q, & L = 1. \end{cases} \tag{11.3}$$

当 $L > 1$ 时表达式仍为正，可等价写为 $\varepsilon_q (L^k - 1)/(L - 1)$。域内统一误差界和轨迹留在该域的条件都很强，训练样本上的平均误差不能直接代入它们。这个推导解释可能的传播机制，而不是给一个深度模型提供自动可用的安全界。

若 F 在局部可微，可以进一步把向量误差写为近似式

$$\delta_{k+1} \approx J_F(x_k, a_k)\delta_k + \eta_q(\tilde{x}_k, a_k), \qquad \eta_q(x, a) = F_q(x, a) - F(x, a). \tag{11.4}$$

这里 $\delta_k = \tilde{x}_k - x_k$，近似来自对 F 的一阶展开，余项需要平滑性与局部小误差控制。量化误差可以有偏，也会随输入变化；不能在没有证据时把 η_q 当作独立、零均值白噪声。相同小偏差若反复作用在不稳定方向上，可能比同等一步均方误差的随机扰动更有害。

以上分析固定了动作。真实闭环中，压缩还会改变规划动作；若策略 π 可微或满足 Lipschitz 条件，应分析复合映射 $F(x, \pi(x))$。对离散动作、接触或终止事件，局部导数可能不足以描述动作切换，应直接检查事件与回报。随机潜状态模型则应比较条件分布或采用声明清楚的耦合，不把两次独立采样的差异全部计作压缩误差。

11.3.2 预测误差小，动作排序仍可能改变

考虑固定的有限候选动作串集合 $\mathcal{A}_0$。设参考规划分数为 $J(A)$，压缩系统给出 $\tilde{J}(A)$，并且对这个集合有

$$\sup_{A \in \mathcal{A}_0} |\tilde{J}(A) - J(A)| \le \varepsilon_J.$$

若压缩规划器返回的 $\tilde{A}$ 满足 $\tilde{J}(\tilde{A}) \ge \max_{A \in \mathcal{A}_0} \tilde{J}(A) - \varepsilon_{\text{opt}}$，令 A^* 最大化参考分数，则

$$\begin{aligned} J(A^*) - J(\tilde{A}) &\le [J(A^*) - \tilde{J}(A^*)] + [\tilde{J}(A^*) - \tilde{J}(\tilde{A})] + [\tilde{J}(\tilde{A}) - J(\tilde{A})] \\ &\le 2\varepsilon_J + \varepsilon_{\text{opt}}. \end{aligned} \tag{11.5}$$

这是相对于同一候选集、同一个参考分数的结论。若参考本身是有偏世界模型，它不能推出真实回报的同样保证；随机搜索更换候选集、奖励头失配和终止误判也需要另外处理。

特别地，若参考最优动作串与第二名之间的间隔大于 $2\varepsilon_J$，且压缩规划器精确最大化自己的分数，则最优动作不变。间隔很小时，即使预测变化很小也可能交换排序。实验因此应联合记录动作选择差异、候选分数间隔、终止事件和真实回报，而非只看平均潜状态漂移。

11.3.3 一份可以执行的压缩研究设计

固定一个已能可靠控制的小模型及其检查点，再逐模块压缩。先只改潜动力学，随后分别改编码器、奖励/继续头与注意力缓存；最后才评估组合压缩。否则无法判断误差来自状态起点、时间传播还是任务计分。校准数据应覆盖实际规划访问的状态和边界场景，并与测试集隔离；QAT 若增加训练数据或更新次数，应另记预算。

每组至少画四类曲线：位宽与真实部署延迟/内存；匹配动作的一步及多步预测误差；候选动作排序与事件校准；真实闭环回报及模型内外差距。比较时固定真实交互、候选数和规划长度，另设等延迟或等模型调用预算对照。硬件上没有相应低比特算子时，低位宽存储不必然带来推理加速，必须实际测量。

“任务敏感的维度应保留较高精度”是一项可检验假设。梯度大、方差大或重建重要的维度未必对动作最重要；可以用干预该维度后的动作变化、回报变化和多任务泛化来验证。若混合精度的收益仅来自多给了总体内存或校准数据，就不能把它归因于任务敏感分配。

11.4 六个研究问题及其反证条件

表 11.1 从方向判断到可证伪实验

问题	带条件的假设	最小实验与主要指标	什么结果削弱假设
决策对齐表示	固定任务与预算时，任务相关预测比均匀重建更利于控制	同一 RSSM 加/减任务损失；测先验误差、真实回报与新奖励任务	仅训练回报提高，换任务后信息丢失或真实回报无增益
选择性想象	校准的不确定性可帮助选择 rollout 长度	固定模型比较自适应时域与等预算固定时域；测风险、回报、拒绝率	不确定性对误差无预测力，或收益由更多计算解释
潜动作标定	少量覆盖充分的真实动作可改善可执行响应	固定视频预训练，改变动作标定量；测动作互换与真实执行	只改善画面差异，控制方向、幅度或跨场景语义仍错误
结构化几何	对有视角变化的任务，几何约束改善重访与控制	同参数/等计算对照；测跨视角、遮挡后重访和回报	仅使用额外深度真值才有收益，预测几何误差抵消收益
压缩与多时域	相同部署预算下，模块化精度分配优于统一位宽	PTQ/QAT、统一/混合精度；测漂移、动作排序、延迟与回报	优势不能跨种子、时域或任务重现
模拟器评测校准	经过独立校准的模型能改善指定策略族的排序	固定真实检查点划分校准/测试策略；测排序与模型外分差	新策略族排名崩溃，或校准使用了测试真回报

这些问题不保证产生正结果。较简单基线在匹配预算下更可靠、某个代理指标不能预测回报、一个假设仅在窄任务内成立，都是有价值的结论。实验前写出最小有意义差异、失败标准和计划分析，可以减少看完结果后再改写问题的自由度。

11.5 把泛化拆成具体变化

组合泛化不是任何测试集上的提高。可以从对象形状、颜色、质量、初速度、相机和动作规则中挑选有限因素，明确训练中见过哪些边缘取值与哪些组合，再构造未见组合。SVIB 提供了这类系统泛化研究的一个具体参考 [30]。若测试同时引入新颜色、新形状和新动力学，就无法仅凭性能下降判断是哪一类知识没有迁移。

一种有教学价值的设计是让颜色在训练中与动力学相关，在测试时解除相关；另设颜色完全随机的训练组。比较像素重建、速度预测与真实控制，检查模型是否依赖外观捷径。再把奖励换成颜色识别任务，观察原来希望忽略的变量是否变得必要。这样可以直接检验“对当前任务不变”与“对所有任务充分”的区别。

对于长时记忆，分别改变遮挡时长、相关噪声和环境变动速度。若对象在遮挡期间可能被外

部智能体移动，正确模型应该增加不确定性，而不是永远维持先前位置。对因果或反事实主张，应说明哪些机制被干预、哪些外生变量保持相同，以及数据能支持哪些比较；第 7 章的假设不会因为模型规模变大而消失。

11.6 从基础到研究的十二周参考路线

这是一份按依赖关系安排的参考计划，适合在小环境上建立实验能力。实际进度取决于编程经验和计算条件；大型模型训练不是每周必须完成的打卡项目。

表 11.2 阅读、实现与验收相互连接的学习路线

周次	核心问题	可审阅交付物
1–2	POMDP、Bayes 过滤、线性系统与 MPC	手推两状态后验；用独立路径枚举核验；解释线性实验的时域曲线
3	Dyna、PETS 与 MBPO	写清真实/想象数据的区别；在状态任务画短分支与模型误差曲线
4–5	RSSM、ELBO 与先验想象	按教学规范实现或阅读一个小型 RSSM；保存后验与先验对照、掩码测试和预算
6	Dreamer 与 TD-MPC 的价值接口	在同一小任务核对继续概率与 bootstrap；比较短时域和末端价值的作用
7	MuZero 与序列决策模型	写出各方法预测的量、搜索接口与训练信号；解释 Decision Transformer 的能力边界
8	token 动力学与扩散轨迹	固定小数据和任务，审计 tokenizer 或采样器；记录模型调用成本
9	JEPA 与动作条件后训练	分开表征探针、特征预测、动作响应和闭环验证，不合并成一个分数
10	交互生成、三维/四维接口	为一个公开系统填写六变量契约、版本表和重访测试方案
11	分层评测与泛化	冻结数据切分、策略集合与预算；建立失败切片及区间估计脚本
12	选择一个可证伪研究问题	提交假设、匹配基线、消融、种子计划、停止条件与复现包结构

对于不能在现有算力内训练的模型，可以先完成接口审计、公开检查点推理或实验协议，但必须准确标注做到了哪一步。没有运行的训练不能写成性能结果；只复现出可视化，也不能宣称复现了闭环控制。

11.7 提交一个别人能检查的研究方案

一份小而完整的方案应包括：一个主要问题；一个可反驳的机制假设；一个能分开竞争解释的最小环境；一个可信基线；一个预先确定的主要指标；以及完整预算。然后补上数据划分、随机种子、失败定义、统计单位与运行命令。把预测结果与控制结果放在同一个种子和配置索引下，才能追踪“这次回报变化对应哪一种模型误差”。

运行顺序可以从无噪声参考解开始，再加入过程随机性、部分观测和分布变化。每一步只引入能解释的复杂度。若一个方法在已知状态的线性问题上都无法通过独立参考检查，直接换成像素输入通常只会让错误更难定位。

研究与验证

世界模型研究的可用结论应能补全这句话：“在这些信息、动作、数据和计算条件下，这个模型改善了这些预测或决策量；这些变化会使它失败。”保留失败条件，结论才有可迁移的边界。

练习与提示。

1. 用两步标量动力学构造两个一步均方误差相同、长时偏差不同的压缩实现。提示：比较固定方向偏置与会交替抵消的偏置，明确动作和初始条件。
2. 给两个候选动作串指定参考分数，验证间隔小于 $2\varepsilon_J$ 时可以发生排序反转；再说明小间隔并不保证反转，而间隔大于 $2\varepsilon_J$ 是排序不变的充分条件，并非必要条件。
3. 若自适应时域方法提高回报，但平均模型调用数也翻倍，至少设计两个对照来判断改进来自哪里。
4. 为表 11.1 任一问题写出一个你愿意接受的负结果，并列出它仍无法排除的竞争解释。

附录 A　符号、术语与关键练习提示

本附录集中说明全书约定，并为容易混淆的推导提供较完整的检查步骤。它不替代正文中的假设：同一个公式只有在时间索引、信息来源和概率模型一致时才有明确含义。

A.1 时间顺序与符号

一次交互按以下顺序发生：接收 o_t，根据已知历史 H_t 选择 a_t，再收到 r_t, d_t，环境到达 s_{t+1} 并给出 o_{t+1}。奖励下标对应刚执行的动作；$d_t = 1$ 表示执行该动作后真正终止。若实现将奖励写为 r_{t+1}，可以使用另一套一致约定，但不能在损失、价值目标和日志之间混用。

表 A.1　全书主要符号

符号	含义	使用时应检查的条件
s_t	真实环境状态	通常不可直接访问；真状态探针属于诊断信息
o_t	决策前已获得的观测	若奖励或接触也提供信息，须扩充信息包或历史
a_t	执行动作	单位、坐标系、持续时间、限幅和延迟
r_t, d_t	动作后的奖励与真正终止	终止步仍有即时奖励，之后没有任务回报
τ_t^{cut}	采样、片段或时间上限截断标记	与真正终止 d_t 分开；有限时域任务需按定义判断
H_t	$(o_0, a_0, \dots, a_{t-1}, o_t)$	只含决策时可用的信息；扩充反馈时应明确写出
$b_t(s)$	$p(s_t = s \mid H_t)$	真实状态空间上的后验，不是一个神经向量
$T(s' \mid s, a)$	环境转移核	状态与动作条件充分；不要与片段长度 T 混淆
$O(o \mid s)$	当前状态观测核	复杂传感器依赖需改用更一般的核
h_t	RSSM 的确定性递归记忆	与完整历史 H_t 是不同对象
z_t	RSSM 的随机潜变量	不默认与 s_t 同坐标或一一对应
$x_t = (h_t, z_t)$	RSSM 的总内部状态	想象与后验更新应使用同一状态定义
w_t	几何接口示意中的结构化内部状态	第 9 章的候选设计，不等于真状态或 RSSM 状态

符号	含义	使用时应检查的条件
$p_\theta(z_t \mid h_t)$	生成模型的潜先验	想象时没有未来观测，只能依靠先验递推
$q_\phi(z_t \mid h_t, o_t)$	近似滤波后验	可用当前观测；不能偷看未来或尚未得到的奖励
$\ell_t(x_t)$	当步观测与任务头的对数似然和	片段末帧无对应动作时，奖励/终止项需掩码
$\mathrm{KL}(q\|p)$	相对熵	次序不可互换；自然对数下单位是 nat
β, κ	KL 权重与 free nats 阈值	说明作用于总 KL、每维还是每步
$\mathrm{sg}[u]$	停止梯度操作	前向值不变，反向计算图改变
π_η, V_ψ	策略与价值函数	价值必须对应指定的奖励、折扣与后续策略
$c_t = 1 - d_t$	是否继续的随机变量	与预测均值 $\hat{c}_t$、相机参数均不同
Γ_k	累积折扣与继续权重	$\Gamma_{k+1} = \Gamma_k \gamma (1 - d_{t+k})$
γ, λ	折扣因子与多步回报混合系数	无限折扣讨论通常取 $\gamma < 1$；有限实测可另定义
H, T	规划长度与观测片段末索引	H_t 带下标表示历史；$o_{0:T}$ 有 $T+1$ 帧

第 9 章的几何接口使用 c_t^{cam} 表示相机参数，和继续变量 c_t 不同。第 11 章比较的 $x_t, \tilde{x}_t$ 是参考与压缩实现的内部状态；只有坐标与归一化相同时，潜状态范数差才有可比较意义。

概率密度的对数值依赖所用变量与测度，不能把图像负对数似然、奖励均方误差和潜空间距离直接按数值大小比较。自然对数下 1 nat $= 1/\log 2$ bit；将 KL 从 nat 换算到 bit 时，数值除以 $\log 2$，不能只改单位标签。

A.2 重要术语的工作含义

表 A.2 从术语到可检查对象

术语	本书中的含义
POMDP	部分可观测马尔可夫决策过程。真实状态不能直接访问，决策依赖观测历史或其充分统计量。
信念状态	对当前真实状态的条件分布；在已知模型和适当信息条件下，可作为决策状态。
RSSM	递归状态空间模型，用确定性记忆与随机潜变量共同表示内部状态。

术语	本书中的含义
ELBO	某个规范化生成模型的证据下界。任意加权损失不能仅因含重建与 KL 就自动称为同一 ELBO。
想象 rollout	从给定内部状态按动作与模型先验递推。是否生成像素取决于下游接口。
MPC	模型预测控制：规划有限时段，只执行首个动作，再由新观测校正并重规划。
MCTS	蒙特卡洛树搜索：展开动作分支，以采样和价值/策略估计分配搜索，适用性依赖动作与模拟接口。
价值 bootstrap	用价值估计补足尚未显式展开的回报；其误差和分布适用范围需另行检验。
状态混叠	表示将需要不同预测或动作的历史合并。当前帧重建准确也可能发生。
Bisimulation	在指定动作集下，以奖励和抽象后继分布一致性定义状态等价；近似损失不自动等于全域保证。
动作忠实性	模型未来对指定动作的方向、幅度和时序响应是否符合真实接口。
认知不确定性	明确概率模型中，由参数或知识未知引起的预测变化；深度集成只是可能的近似。
偶然不确定性	给定模型参数后仍存在的条件随机性；与模型失配、采样多样性不能混为一谈。
开环预测	给定动作串，从起点向前预测且不利用后续真实观测校正。
闭环评测	策略根据后续观测持续行动，状态估计与后续动作随反馈更新；需说明是在学习模拟器还是真实环境中闭环。
组合泛化	在明确定义的未见因素组合上复用已学规律，而非笼统地在另一测试集上得高分。
PTQ / QAT	后训练量化与量化感知训练；需同时记录校准或额外训练数据以及实际部署算子。
Sim-to-real	将仿真或合成环境中获得的模型/策略用于真实系统，最终效用由真实执行检验。

A.3 信念、可观测性与因果练习

A.3.1 两状态后验的一般式

假设隐藏状态在整段内不变，先验 $\Pr(s=1)=\alpha\in(0,1)$，二元传感器正确率为 $p\in(0,1)$。给定状态后，各次观测独立。在 n 次观测中出现 k 个 1，则

$$L_1=p^k(1-p)^{n-k},\qquad L_0=(1-p)^k p^{n-k}, \tag{A.1}$$

$$\frac{\Pr(s=1\mid o_{0:n-1})}{\Pr(s=0\mid o_{0:n-1})}=\frac{\alpha}{1-\alpha}\left(\frac{p}{1-p}\right)^{2k-n}. \tag{A.2}$$

把赔率 u 变回概率用 $u/(1+u)$。取 $\alpha=1/2, p=0.8, n=3$，$k=3$ 得 64/65，$k=1$ 得 1/5。$p=1/2$ 时后验等于先验；$p<1/2$ 时观测仍有信息，只是应反向解释。$p=0$ 或 1 应用原始似然单独处理，避免在赔率公式中出现除零。

若每步以概率 δ 翻转隐藏状态，则先做

$$\bar{b}_{t+1}(1)=(1-\delta)b_t(1)+\delta[1-b_t(1)]=\delta+(1-2\delta)b_t(1),$$

再用新观测校正。$\delta=1/2$ 时，下一状态在观察新数据前与过去状态无关；历史不能继续提供原来静态环境中那样的累积证据。若噪声相关，前述独立似然乘积也必须改写。

A.3.2 一帧位置与两帧位置

令 $s_t=(p_t,v_t)^\top$，匀速系统 $A=\left(\begin{smallmatrix}1&\Delta t\\0&1\end{smallmatrix}\right)$，$C=[1\ 0]$，则两步可观测矩阵为

$$\mathcal{O}_2=\begin{pmatrix}1&0\\1&\Delta t\end{pmatrix},\qquad \det\mathcal{O}_2=\Delta t.$$

当 $\Delta t\neq 0$ 时，可由无噪声两帧恢复初速度；单帧只能确定位置。若测量有噪声，速度估计包含除以 Δt，小采样间隔会放大噪声。理论满秩与有限精度下估计稳定是两个问题。

A.3.3 看见专家动作与主动执行动作

令 $s_t\in\{0,1\}$ 等概率，H_t 不提供状态信息。专家能看见真实状态，总选 $a_t=s_t$，则观察专家选了 $a_t=1$ 后有 $p(s_t=1\mid H_t,a_t=1)=1$。但我们主动指定 $a_t=1$ 并不会改变动作之前的隐藏状态，因此 $p(s_t=1\mid H_t,\mathrm{do}(a_t=1))=1/2$。这说明动作标签可能携带额外专家信息；识别干预响应需要状态充分性、覆盖与无未控制混杂等条件。

A.4 ELBO 与 KL 的核对步骤

A.4.1 展开最短的有动作片段

取 $T=1$，数据为两帧 o_0,o_1，一个动作 a_0，以及对应的 r_0,d_0。令 $q_0=q_\phi(z_0\mid h_0,o_0)$，$q_1=q_\phi(z_1\mid h_1,o_1)$，$p_t=p_\theta(z_t\mid h_t)$。其中 $h_1=f_\theta(h_0,z_0,a_0)$，所以 q_1,p_1 都随 z_0 改变。式 (3.11) 变为

$$\begin{aligned}\mathcal{L}=&\mathbb{E}_{q_0}[\ell_0(x_0)]+\mathbb{E}_{q_0q_1}[\ell_1(x_1)]-\mathrm{KL}(q_0\|p_0)\\&-\mathbb{E}_{q_0}[\mathrm{KL}(q_1\|p_1)].\end{aligned} \tag{A.3}$$

ℓ_0 含观测、奖励与终止三个对数似然；ℓ_1 只有观测似然。没有 a_1, r_1, d_1。后一 KL 的外层期望不能去掉，因为 h_1 依赖先前样本。把这个两帧展开写成单元检查，往往比检查一个大批量总损失更容易发现时间错位。

A.4.2 为什么小于一的 KL 权重可能越过证据

取二元潜变量，令 $p(z=1)=\varepsilon\in(0,1)$；观测事件 $o=1$ 只在 $z=1$ 时发生，其条件概率为 1。在观察到 $o=1$ 后，精确后验集中在 $z=1$，因此

$$R=\mathbb{E}_q \log p(o=1\mid z)=0, \qquad K=\log(1/\varepsilon), \qquad \log p(o=1)=\log\varepsilon.$$

标准 ELBO $R-K=\log\varepsilon$ 在此恰好等于证据；若 $0\le\beta<1$，则 $R-\beta K=\beta\log\varepsilon>\log\varepsilon$。所以减少 KL 权重可以是工程选择，却不能继续声称它总是原模型的证据下界。

A.4.3 一维高斯 KL 与方差边界

若 $q=\mathcal{N}(\mu_q,\sigma_q^2)$、$p=\mathcal{N}(\mu_p,\sigma_p^2)$，二者方差均为正，则

$$\mathrm{KL}(q\|p)=\log\frac{\sigma_p}{\sigma_q}+\frac{\sigma_q^2+(\mu_q-\mu_p)^2}{2\sigma_p^2}-\frac{1}{2}.$$

当先验方差趋近零而后验未以相应速度收缩到同一均值时，KL 和梯度可能很大。设置正方差下界、记录方差分布与检查非有限值有明确作用；简单截断最终 KL 标量未必修复产生异常的密度参数。

A.5 规划与诊断练习

A.5.1 终止掩码应该乘在哪里

设某动作获得奖励 2 并真正终止，末端价值网络却给出 100。按照式 (4.2)，单步目标为 $2+\gamma(1-d_t)100=2$。若仅因数据片段结束而截断，任务仍会继续，且有正确的片段末观测，则目标可以保留 $2+\gamma 100$；当 $\gamma=0.9$ 时为 92。二者的差异来自任务语义，不是能否继续采样到当前文件中的下一行。有限时域任务自身的时限若定义为终止，则应按真正终止处理。

A.5.2 相同期望奖励不代表相同风险

一个状态确定获得 0，另一个以相同概率获得 -1 或 1，二者期望奖励都为 0。如果后继结构相同，风险中性的期望回报目标可能无法区分；关注损失概率、方差或尾部风险的目标会区分。将只匹配期望奖励的状态抽象用于风险敏感控制，需要重新检查所保留的奖励信息。

A.5.3 全体模型一致也可能全体错误

让所有集成成员仅在干燥路面数据上训练，再测试没有出现过的低摩擦路面。各模型可能一致地给出过短制动距离。分歧小只说明这些成员输出接近，不能证明环境机制未变。校准集应包含明确的分布变化切片，并检查拒绝率、剩余风险和真实回退行为；仅在同分布随机划分上画校准图，无法检验这一失败模式。

A.6 证据、实验与复现文件怎样对应

文献引用应回到 `references.bib` 固定的原论文或官方资料。`evidence/` 保存来源核验材料和编辑记录；归档的研究笔记有助于定位旧资料，但不应覆盖正式修订后的版本说明。论文报告、本书手算、本次实测与未运行方案的证据等级不同。

第 10 章的实测结果位于本主题 `output/experiments/`，代码与配置位于 `manuscript/experiments/`。复现时核对运行命令、解释器与依赖版本、随机种子、配置和脚本哈希，再读取逐回合结果与汇总。仅有一个 `passed` 标记不足以判断测试覆盖了什么，应阅读其独立参考解与边界检查。

RSSM 协议、研究章中的压缩方案和十二周路线均为后续设计，不能视为已得到的训练结果。若未来运行，应另建带配置标识的结果目录，记录真实交互、训练更新、模型调用与时间预算，并同步更新正文中的证据状态。

参考文献

[1] D. Ha and J. Schmidhuber. *World Models*. 核验版本 1803.10122v4，修订 2018-05-09. Mar. 27, 2018. DOI: 10.5281/zenodo.1207631. arXiv: 1803.10122. URL: https://arxiv.org/abs/1803.10122v4 (visited on 10/22/2026).

[2] D. Hafner, T. Lillicrap, I. Fischer, et al. *Learning Latent Dynamics for Planning from Pixels*. 核验版本 1811.04551v5，修订 2019-06-04. Nov. 12, 2018. arXiv: 1811.04551. URL: https://arxiv.org/abs/1811.04551v5 (visited on 10/22/2026).

[3] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. *Dream to Control: Learning Behaviors by Latent Imagination*. 核验版本 1912.01603v3，修订 2020-03-17. Dec. 3, 2019. arXiv: 1912.01603. URL: https://arxiv.org/abs/1912.01603v3 (visited on 10/22/2026).

[4] J. Schrittwieser, I. Antonoglou, T. Hubert, et al. *Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model*. 核验版本 1911.08265v2，修订 2020-02-21；2019 年预印本，所附 DOI 对应 Nature 2020 正式版. Nov. 19, 2019. DOI: 10.1038/s41586-020-03051-4. arXiv: 1911.08265. URL: https://arxiv.org/abs/1911.08265v2 (visited on 10/22/2026).

[5] J. Bruce, M. Dennis, A. Edwards, et al. *Genie: Generative Interactive Environments*. 核验版本 2402.15391v1，修订 2024-02-23. Feb. 23, 2024. arXiv: 2402.15391. URL: https://arxiv.org/abs/2402.15391v1 (visited on 10/22/2026).

[6] M. Assran, Q. Duval, I. Misra, et al. *Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture*. 核验版本 2301.08243v3，修订 2023-04-13. Jan. 19, 2023. arXiv: 2301.08243. URL: https://arxiv.org/abs/2301.08243v3 (visited on 10/22/2026).

[7] A. Bardes, Q. Garrido, J. Ponce, et al. *Revisiting Feature Prediction for Learning Visual Representations from Video*. 核验版本 2404.08471v1，修订 2024-02-15. Feb. 15, 2024. arXiv: 2404.08471. URL: https://arxiv.org/abs/2404.08471v1 (visited on 10/22/2026).

[8] M. Assran, A. Bardes, D. Fan, et al. *V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning*. 核验版本 2506.09985v1，修订 2025-06-11. June 11, 2025. arXiv: 2506.09985. URL: https://arxiv.org/abs/2506.09985v1 (visited on 10/22/2026).

[9] R. S. Sutton. "Dyna, an integrated architecture for learning, planning, and reacting". In: *ACM SIGART Bulletin* 2.4 (1991), pp. 160–163. DOI: 10.1145/122344.122377. URL: https://doi.org/10.1145/122344.122377 (visited on 10/22/2026).

[10] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. "Planning and acting in partially observable stochastic domains". In: *Artificial Intelligence* 101.1-2 (1998), pp. 99–134. DOI: 10.1016/s0004-3702(98)00023-x. URL: https://doi.org/10.1016/s0004-3702(98)00023-x (visited on 10/22/2026).

[11] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba. *Mastering Atari with Discrete World Models*. 核验版本 2010.02193v4，修订 2022-02-12. Oct. 5, 2020. arXiv: 2010.02193. URL: https://arxiv.org/abs/2010.02193v4 (visited on 10/22/2026).

[12] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap. *Mastering Diverse Domains through World Models*. 核验版本 2301.04104v2，修订 2024-04-17. Jan. 10, 2023. arXiv: 2301.04104. URL: https://arxiv.org/abs/2301.04104v2 (visited on 10/22/2026).

[13] K. Chua, R. Calandra, R. McAllister, and S. Levine. *Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models*. 核验版本 1805.12114v2，修订 2018-11-02. May 30, 2018. arXiv: 1805.12114. URL: https://arxiv.org/abs/1805.12114v2 (visited on 10/22/2026).

[14] M. Janner, J. Fu, M. Zhang, and S. Levine. *When to Trust Your Model: Model-Based Policy Optimization*. 核验版本 1906.08253v3，修订 2021-11-29. June 19, 2019. arXiv: 1906.08253. URL: https://arxiv.org/abs/1906.08253v3 (visited on 10/22/2026).

[15] T. M. Moerland, J. Broekens, A. Plaat, and C. M. Jonker. "Model-based Reinforcement Learning: A Survey". In: *Foundations and Trends in Machine Learning* 16.1 (2023), pp. 1–118. DOI: 10.1561/2200000086. URL: https://doi.org/10.1561/2200000086 (visited on 10/22/2026).

[16] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap. "Mastering diverse control tasks through world models". In: *Nature* 640.8059 (Apr. 2, 2025). DreamerV3 正式发表版；与 2023 年预印本题名及年份区分, pp. 647–653. DOI: 10.1038/s41586-025-08744-2. URL: https://www.nature.com/articles/s41586-025-08744-2 (visited on 10/22/2026).

[17] L. Kaiser, M. Babaeizadeh, P. Milos, et al. *Model-Based Reinforcement Learning for Atari*. 核验版本 1903.00374v5，修订 2024-04-03. Mar. 1, 2019. arXiv: 1903.00374. URL: https://arxiv.org/abs/1903.00374v5 (visited on 10/22/2026).

[18] N. Hansen, X. Wang, and H. Su. *Temporal Difference Learning for Model Predictive Control*. 核验版本 2203.04955v2，修订 2022-07-19. Mar. 9, 2022. arXiv: 2203.04955. URL: https://arxiv.org/abs/2203.04955v2 (visited on 10/22/2026).

[19] N. Hansen, H. Su, and X. Wang. *TD-MPC2: Scalable, Robust World Models for Continuous Control*. 核验版本 2310.16828v2，修订 2024-03-21. Oct. 25, 2023. arXiv: 2310.16828. URL: https://arxiv.org/abs/2310.16828v2 (visited on 10/22/2026).

[20] V. Micheli, E. Alonso, and F. Fleuret. *Transformers are Sample-Efficient World Models*. 核验版本 2209.00588v2，修订 2023-03-01. Sept. 1, 2022. arXiv: 2209.00588. URL: https://arxiv.org/abs/2209.00588v2 (visited on 10/22/2026).

[21] J. Wu, S. Yin, N. Feng, et al. *iVideoGPT: Interactive VideoGPTs are Scalable World Models*. 核验版本 2405.15223v3，修订 2024-10-31. May 24, 2024. arXiv: 2405.15223. URL: https://arxiv.org/abs/2405.15223v3 (visited on 10/22/2026).

[22] M. Janner, Q. Li, and S. Levine. *Offline Reinforcement Learning as One Big Sequence Modeling Problem*. 核验版本 2106.02039v4，修订 2021-11-29. June 3, 2021. arXiv: 2106.02039. URL: https://arxiv.org/abs/2106.02039v4 (visited on 10/22/2026).

[23] L. Chen, K. Lu, A. Rajeswaran, et al. *Decision Transformer: Reinforcement Learning via Sequence Modeling*. 核验版本 2106.01345v2，修订 2021-06-24. June 2, 2021. arXiv: 2106.01345. URL: https://arxiv.org/abs/2106.01345v2 (visited on 10/22/2026).

[24] M. Janner, Y. Du, J. B. Tenenbaum, and S. Levine. *Planning with Diffusion for Flexible Behavior Synthesis*. 核验版本 2205.09991v2，修订 2022-12-21. May 20, 2022. arXiv: 2205.09991. URL: https://arxiv.org/abs/2205.09991v2 (visited on 10/22/2026).

[25] E. Alonso, A. Jelley, V. Micheli, et al. *Diffusion for World Modeling: Visual Details Matter in Atari*. 核验版本 2405.12399v2，修订 2024-10-30. May 20, 2024. arXiv: 2405.12399. URL: https://arxiv.org/abs/2405.12399v2 (visited on 10/22/2026).

[26] S. Levine, A. Kumar, G. Tucker, and J. Fu. *Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems*. 核验版本 2005.01643v3，修订 2020-11-01. May 4, 2020. arXiv: 2005.01643. URL: https://arxiv.org/abs/2005.01643v3 (visited on 10/22/2026).

[27] N. Ferns, P. Panangaden, and D. Precup. "Metrics for Finite Markov Decision Processes". In: *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*. 2004 年会议论文；2012 年上传 arXiv 档案，二者不混同. 2004, pp. 162–169. arXiv: 1207.4114. URL: https://arxiv.org/abs/1207.4114v1 (visited on 10/22/2026).

[28] A. Zhang, R. McAllister, R. Calandra, Y. Gal, and S. Levine. *Learning Invariant Representations for Reinforcement Learning without Reconstruction*. ICLR 2021；预印本初版 2020-06-18，核验版本 2006.10742v2. 2021. arXiv: 2006.10742. URL: https://arxiv.org/abs/2006.10742v2 (visited on 10/22/2026).

[29] J. Quevedo, A. K. Sharma, Y. Sun, V. Suryavanshi, P. Liang, and S. Yang. *WorldGym: World Model as An Environment for Policy Evaluation*. 核验版本 2506.00613v3，修订 2025-09-30. May 31, 2025. arXiv: 2506.00613. URL: https://arxiv.org/abs/2506.00613v3 (visited on 10/22/2026).

[30] Y. Kim, G. Singh, J. Park, C. Gulcehre, and S. Ahn. *Imagine the Unseen World: A Benchmark for Systematic Generalization in Visual World Models*. 核验版本 2311.09064v1，修订 2023-11-15. Nov. 15, 2023. arXiv: 2311.09064. URL: https://arxiv.org/abs/2311.09064v1 (visited on 10/22/2026).

[31] Y. Qin, Z. Shi, J. Yu, et al. *WorldSimBench: Towards Video Generation Models as World Simulators*. 核验版本 2410.18072v1，修订 2024-10-23. Oct. 23, 2024. arXiv: 2410.18072. URL: https://arxiv.org/abs/2410.18072v1 (visited on 10/22/2026).

[32] D. Li, Y. Fang, Y. Chen, et al. *WorldModelBench: Judging Video Generation Models As World Models*. 核验版本 2502.20694v1，修订 2025-02-28. Feb. 28, 2025. arXiv: 2502.20694. URL: https://arxiv.org/abs/2502.20694v1 (visited on 10/22/2026).

[33] Y. Hu, S. Huang, Y. Liao, et al. *EWMBench: Evaluating Scene, Motion, and Semantic Quality in Embodied World Models*. 核验版本 2505.09694v2，修订 2025-05-18. May 14, 2025. arXiv: 2505.09694. URL: https://arxiv.org/abs/2505.09694v2 (visited on 10/22/2026).

[34] NVIDIA. *Cosmos World Foundation Model Platform for Physical AI*. 核验版本 2501.03575v3，修订 2025-07-09. Jan. 7, 2025. arXiv: 2501.03575. URL: https://arxiv.org/abs/2501.03575v3 (visited on 10/22/2026).

[35] N. Kazemi, N. Savov, D. Paudel, and L. Van Gool. *Learning Generative Interactive Environments By Trained Agent Exploration*. 核验版本 2409.06445v2，修订 2024-10-18. Sept. 10, 2024. arXiv: 2409.06445. URL: https://arxiv.org/abs/2409.06445v2 (visited on 10/22/2026).

[36] J. Parker-Holder, P. Ball, J. Bruce, et al. *Genie 2: A large-scale foundation world model*. 官方研究介绍；演示与发布方声明. Google DeepMind. Dec. 4, 2024. URL: https://deepmind.google/blog/genie-2-a-large-scale-foundation-world-model/ (visited on 10/22/2026).

[37] J. Parker-Holder and S. Fruchter. *Genie 3: A new frontier for world models*. 官方研究介绍；运行性能仅为发布方披露条件. Google DeepMind. Aug. 5, 2025. URL: https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/ (visited on 10/22/2026).

[38] X. Wang, Z. Zhu, G. Huang, B. Wang, X. Chen, and J. Lu. *WorldDreamer: Towards General World Models for Video Generation via Predicting Masked Tokens*. 核验版本 2401.09985v1，修订 2024-01-18. Jan. 18, 2024. arXiv: 2401.09985. URL: https://arxiv.org/abs/2401.09985v1 (visited on 10/22/2026).

[39] T. Brooks, B. Peebles, C. Holmes, et al. *Video generation models as world simulators*. 官方技术报告；2026-10-22 核对官方原文，非闭环控制评测. OpenAI. Feb. 15, 2024. URL: https://openai.com/index/video-generation-models-as-world-simulators/ (visited on 10/22/2026).

[40] H. Zhen, Q. Sun, H. Zhang, et al. *TesserAct: Learning 4D Embodied World Models*. 核验版本 2504.20995v1，修订 2025-04-29. Apr. 29, 2025. arXiv: 2504.20995. URL: https://arxiv.org/abs/2504.20995v1 (visited on 10/22/2026).

[41] A. Hu, L. Russell, H. Yeo, et al. *GAIA-1: A Generative World Model for Autonomous Driving*. 核验版本 2309.17080v1，修订 2023-09-29. Sept. 29, 2023. arXiv: 2309.17080. URL: https://arxiv.org/abs/2309.17080v1 (visited on 10/22/2026).

[42] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu. *DriveDreamer: Towards Real-world-driven World Models for Autonomous Driving*. 核验版本 2309.09777v2，修订 2023-11-27. Sept. 18, 2023. arXiv: 2309.09777. URL: https://arxiv.org/abs/2309.09777v2 (visited on 10/22/2026).

[43] X. Ren, Y. Lu, T. Cao, et al. *Cosmos-Drive-Dreams: Scalable Synthetic Driving Data Generation with World Foundation Models*. 核验版本 2506.09042v3，修订 2025-06-18. June 10, 2025. arXiv: 2506.09042. URL: https://arxiv.org/abs/2506.09042v3 (visited on 10/22/2026).

[44] J. Ding, Y. Zhang, Y. Shang, et al. *Understanding World or Predicting Future? A Comprehensive Survey of World Models*. 核验版本 2411.14499v4，修订 2025-12-10. Nov. 21, 2024. arXiv: 2411.14499. URL: https://arxiv.org/abs/2411.14499v4 (visited on 10/22/2026).

[45] X. Long, Q. Zhao, K. Zhang, et al. *A Survey: Learning Embodied Intelligence from Physical Simulators and World Models*. 核验版本 2507.00917v3，修订 2025-09-03. July 1, 2025. arXiv: 2507.00917. URL: https://arxiv.org/abs/2507.00917v3 (visited on 10/22/2026).

[46] B. Hou, G. Li, J. Jia, et al. *World Model for Robot Learning: A Comprehensive Survey*. 核验版本 2605.00080v1，修订 2026-04-30. Apr. 30, 2026. arXiv: 2605.00080. URL: https://arxiv.org/abs/2605.00080v1 (visited on 10/22/2026).

[47] Y. Guan, H. Liao, Z. Li, et al. *World Models for Autonomous Driving: An Initial Survey*. 核验版本 2403.02622v3，修订 2024-05-07. Mar. 5, 2024. arXiv: 2403.02622. URL: https://arxiv.org/abs/2403.02622v3 (visited on 10/22/2026).