arXiv:2611.30879v1 [cs.GR] 23 Nov 2026

借助视觉扩散模型学习下一步最佳着法

Amir Razlighi[1,2,3] Ali Mahdavi Amiri[1] Yiorgos Chrysanthou[2,3] Andrey Voynov[4] Daniel Cohen-Or[5]西蒙菲莎大学 [2]塞浦路斯大学 [3]CYENS卓越中心 [4]Google DeepMind [5]特拉维夫大学



图 1. 无需显式前瞻的视觉推理。我们的模型逐步求解国际象棋谜题（上图），其中包括弃车以达成将死；它还能直接根据图像，找出穿越地形的最低成本路线，而地形成本从未显示出来（下图）。蓝方和红方分别标记每步棋的起点和终点。

摘要

本文展示了视觉扩散模型能够学习在需要序列推理和前瞻的任务中预测下一步最佳着法（*NBM*），而无需在推理时显式探索未来步骤或进行穷举搜索。先前研究表明，这类模型能够生成几何和推理问题的完整解法。在此，我们证明，只需针对视觉状态—下一步最佳着法对进行简单微调，预训练模型便能选择其价值取决于数步之后结果的动作。模型直接从当前状态的图像空间预测每一步，无需显式的符号表示或搜索过程。随后，反复进行 *NBM*预测，便能逐次决策构建解法。

我们在国际象棋谜题和基于方格的寻路任务上展示了这一能力，并证明它还能扩展到四子棋、围棋和台球。在国际象棋中，模型预测合法着法并找到最佳着法的频率，远高于通过回忆相似训练局面来做到这一点的频率，其中还包括那些收益只有在后续战术变化中才会显现的弃子。在寻路任务中，它能在统一成本地图上构建最短路径，并在通行成本各不相同的地形方格间找到成本最优路线；当较长路线的总体成本更低时，它会选择较长路线。这些成本是通过示例隐式学得的，无需数值成本监督。这些结果表明，直接根据视觉信息预测着法，能够支持具备全局信息的决策，并实现成功的序列问题求解。

1. 引言

“我只看一步棋，但这一步总是正确的。”—据称出自 *Jos´e Ra´ul Capablanca*

视觉生成模型展现出令人惊讶的能力：通过将解决方案生成为图像，攻克棘手难题。视觉扩散模型不仅能够合成逼真的外观，还能生成几何约束图形、完整谜题，以及构建满足复杂约束的路径 [5, 7, 8, 22]。这些发现表明——

视觉生成可以支持以往由专用算法处理的推理形式。在这项工作中，我们探究视觉扩散模型是否也能学会预测下一步最佳着法（NBM）：选择一个能推动任务走向解法的动作，而无需搜索未来状态。值得一提的是，我们的目标并非开发特定任务的求解器（例如，竞技级国际象棋棋手），而是展示并分析视觉扩散模型一项引人入胜的能力。我们以国际象棋、寻路和其他任务作为测试平台，来探究并更好地理解这项能力。

选择下一步着法带来了另一项挑战：着法的质量取决于之后可能发生的事情。一个有效动作只需满足当前状态下的规则；而最佳动作还必须考虑数步之后的后果。生成完整解法后，其中各个部分便可以一并表示和优化。相比之下，NBM 预测会直接选定一个动作，却不生成能够证明该动作合理性的后续着法。因此，挑战在于如何通过直接视觉预测获得前瞻的好处，同时不在推理时显式探索未来步骤。

国际象棋谜题为这一挑战提供了一个具体场景。仅凭棋盘图像，模型就必须预测一个既合法、又能在应对对手的最佳回应时朝着解开谜题迈进的着法。弃子尤其引人注目：其收益只有通过后续的结果才会显现。第 2 节将结合成本感知寻路展开这一示例，说明我们希望模型学会的决策。

我们表明，通过对视觉状态及其对应的下一步最佳着法这一对数据进行微调，预训练视觉扩散模型能够学会做出此类决策。给定作为图像的当前状态，模型会直接生成 NBM 的视觉预测，而无需显式搜索或符号推理模块。执行着法后，更新后的状态便作为下一次预测的输入。通过反复应用这一过程，模型一次做出一个决策，逐步构建解法，并选择其收益取决于后续结果的动作。同一方法还可扩展为一次预测一小段未来着法。预测的着法数量通常称为预测时域。视频扩散模型会标记连续的着法并生成其结果状态；在国际象棋中，着法由求解方与对手交替进行。

我们的研究表明，在训练预算相同的情况下，同时预测多个着法可以提高下一步着法的准确率，即使其余预测并未执行。生成后续内容因此不仅能带来单个动作。对于确定性任务，模型还可以在再次调用前执行多个预测着法。面对对手时，由于对手的应对可能与预测不同，我们只执行第一步棋，并在观察到

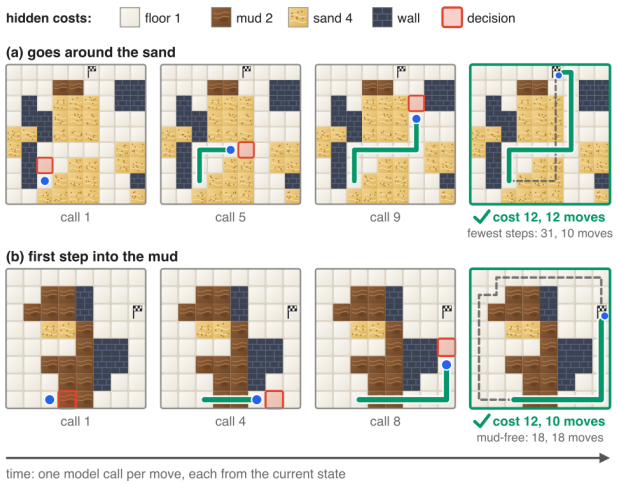

图2. 在两张留出地图上，通过连续预测下一步最佳着法（NBM）进行成本感知导航。*(a)* 当绕开高成本地形能够降低总通行成本时，模型会选择更长的路线；而*(b)* 当某些高成本动作能显著缩短路径时，模型也会接受这些动作。地形成本是从示例解法中隐式学习的，无需明确的数值监督。

应答。

我们对下一最佳着法（NBM）预测的研究揭示了一个出人意料的发现：根据第一步去噪结果预测着法，比使用完全去噪样本更准确。在这一早期阶段，单个图像中会出现多个相互重叠的标记，代表几种可能的着法。最明显的标记往往揭示了正确着法。随着去噪过程推进，模型会选定一个候选项，但它可能不同于第一步中指示最明显的着法。因此，我们可以在去噪完成之前读出模型的首选着法。

2. 行动中的视觉推理

我们首先通过国际象棋谜题和成本感知路径规划，展示模型观察到的内容、表达动作的方式，以及为何选择该动作需要考虑后续影响。

2.1. 国际象棋谜题

国际象棋谜题给出一个棋盘局面，要求棋手找出一系列着法，以实现将死或取得决定性的子力优势等战术目标。解题时必须选择能够应对对手最佳应答的着法。许多着法可能都合法，但只有少数能够推进解题进程。

我们的模型接收棋盘图像，并通过标记起始格和目标格来预测着法。它不会接收符号化棋盘编码、棋子身份，也不会接收

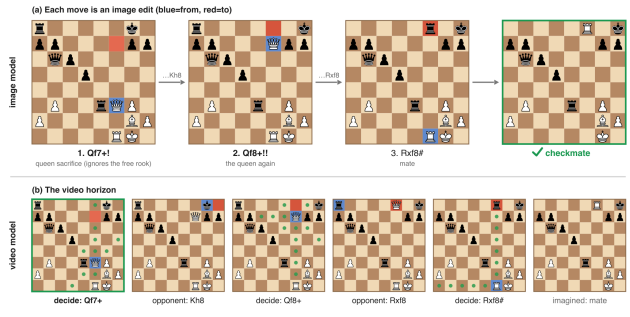

图3. 无需搜索即可找到的双后弃子，*(a)*由图像模型发现，该模型每次调用都根据当前棋盘预测一步着法；*(b)* 由视频模型发现，该模型会同时预测多个着法及其导向的棋盘状态。

明确的移动规则。例如，象仅通过其视觉外观来表示；模型必须从视觉样例中学习它如何移动，以及何时移动才有用。微调样例对展示棋盘图像及其对应的下一步最佳着法，为模型提供有关有效移动模式和有效动作选择的监督。

图3 展示了一个引人注目的例子：某项决策的价值只有在后续着法中才显现出来。在这道留出谜题中，白方可以用 Qxe3 吃掉一辆无人防守的车，立即获得子力优势。然而，模型预测走 Qf7+，弃出皇后，迫使黑王走向角落。走出应答 Kh8 并将新的棋盘局面输入模型后，它预测走 Qf8+，再次弃后，引诱黑车移至 f8；黑方走 Rxf8 后，白方以 Rxf8# 将死。这个弃子尤其能说明问题：尽管立即获得子力优势，模型仍放弃了价值最高的棋子。

每一步着法都根据当前棋盘单独预测。模型在生成任何后续着法之前，就决定弃后，不构建未来棋盘状态、不评估其他后续变化，也不进行显式的树搜索。连续的 NBM 预测还原出图示的战术序列，每次预测都与谜题规定的着法一致。这个例子说明，直接的视觉预测可以选出一个行动，其合理性取决于数步之后的协同后续着法，而推理时无需显式生成这一后续变化。这并非孤立个例：在标记为弃子的 1,001 道留出谜题中，视频模型有 47.9 % 的概率找到下一步最佳着法（投票时为 57.3%），而复制最近训练局面的着法这一方法的比例为 16.4%（表 1）。

2.2. 成本感知寻路

图2 展示了在未参与训练的网格地图上进行导航的两个样例。智能体在相邻的可通行网格之间移动，每种地形类型进入方格的代价都是固定的。

目标是在最小化路线上的地块代价总和的同时到达目标。当地块代价一致时，这相当于寻找最短路径；当地块代价取决于地形时，较长的路线也可能代价更低。这两种情况都可以用迪杰斯特拉算法求解，该算法会显式扩展候选状态并维护累计距离估计。相比之下，我们的模型每次只预测一个着法，不会显式探索未来路线。

训练期间，模型不会接收任何数值形式的地形地块代价，只从最优下一步的样例中学习。通过连续进行 NBM预测，它构建出成本最优路径，并在多走几步能够避开足够昂贵的方格时选择绕行。每一步都根据当前视觉状态进行预测，不会生成剩余路线，也不会比较未来的其他后续走法。这些样例展示了决策质量取决于数步之后代价的情形：进入低成本方格可能导致后续路线代价高昂，而一开始走间接路线则可能最终实现更低成本。这种现象并不局限于这两个样例：在留出地图中，最低成本路线要么绕开高成本地带，要么穿越该地带；对此，模型在 85.0% 的地图上做出了正确选择，而随机猜测的正确率为 50%；在87.5% 的测试地图上，模型找到了最低成本路线，而步数最少路径的这一比例为59.2%（表2）。

3. 方法

我们将一个预训练视觉生成模型转变为决策器，而不改变其架构：状态是一幅图像，动作是对该图像的编辑，决策则通过采样作出。

3.1. 下一步最佳着法预测

一个任务是一系列状态 $s_0, s_1, \ldots$，只能通过渲染结果 $x_t = R(s_t)$ 观察。处于状态s_t 时采取动作 a_t 会导向 $s_{t+1} = E(s_t, a_t)$；在双人游戏中，环境 E 还会执行对手的应答。训练数据为每个状态提供一个取自已知解的参考最佳下一步着法$a_t^\star$。我们用 $M(x, a)$ 表示绘有动作 a 的图像 x （第 3.2 节），并训练条件生成器 p_θ，使其根据 x_t 和固定的文本提示 c 生成 $M(x_t, a_t^\star)$。在推理时，确定性解码器D 从生成的图像中读取动作，

$$\hat{a}_t = D(\hat{y}_t, x_t), \qquad \hat{y}_t \sim p_\theta(\cdot \mid x_t, c), \tag{1}$$

环境返回 $s_{t+1} = E(s_t, \hat{a}_t)$，然后再次对 x_{t+1} 调用模型。因此，解法便在闭环中逐个决策地构建起来。

无需搜索。 每次调用都会抽取一个噪声样本，并执行固定数量的去噪步骤。模型从不枚举、模拟或评估其他动作，也不会在一次调用中查询环境或任何引擎。因此，决策对后续步骤的任何预判，都必须由一次生成来承载。

3.2. 作为图像编辑的动作

一个动作会将实体从网格的单元格 u 移至单元格 v，例如让国际象棋棋子在棋格之间移动，或让智能体移至相邻方格；也可以将实体放置在单元格 v上，例如在围棋中放置一枚棋子。 $M(x,a)$ 用蓝方标记 u，用红方标记 v（仅v 用于放置），并保持所有其他像素不变（图 3）。因此，输入与目标仅在动作发生的位置有所不同，输出空间就是图像本身，而非特定任务的动作词汇。每项任务的提示都是固定的，只包含图像无法呈现的信息，例如应由哪一方行棋，以及如何在图像上标记预测的着法。规则、着法生成器和成本从不作为输入提供；在包含地形的寻路任务中，各类方格的成本既不会绘出，也不会说明。模型对这些内容的了解，完全来自训练目标。

解码。 D 会反转标记。对于每个单元格，它将$\hat{}y_t$ 与 x_t 之间的差异拟合为朝蓝色色调和红色色调的混合，并给出强度 $\alpha_b(u)$和 $\alpha_r(v)$，然后返回蓝方最强单元格和红方最强单元格组成的一对。此读出不使用任何规则，无法执行的动作计为失败。

3.3. 从一步到多步视界

式（1）的闭环每次调用都会进行一次 NBM 预测。视界模型在单次生成中进行 H 次预测：它预测当前状态下的 NBM 、该着法所导向的状态、该状态下的 NBM，依此类推。对于参考后续序列 $a^\star_t$，……，$a^\star_{t+H-1}$，其视觉状态为$x^\star_{t+k} = R\,(s^\star_{t+k})$ 和 $x^\star_t = x_t$，目标为

$$\big(\,M(x^\star_{t+k-1}, a^\star_{t+k-1}),\; x^\star_{t+k}\,\big)_{k=1}^{H}, \tag{2}$$

渲染为一段视频，视频从 x_t 开始，并交替呈现决策帧：标记了下一步着法的当前状态，以及想象帧：该着法之后的视觉状态。对于单个智能体，视野由智能体连续的 NBM 组成；在双人游戏中，视频在双方之间交替，预测给对手的 NBM 就是其最佳应对。在国际象棋中，视频模型的决策帧还会用绿色圆点标出移动棋子的合法落点；解码器不会读取这些圆点（见附录B），它们用于检验模型是否学会了每种棋子的走法。每一步着法都从生成帧中解码得出，$\hat{a}_{t+k-1} = D(\hat{}m_k, \hat{}x_{k-1})$ 其中 $\hat{}x_0 = x_t$，因此每个预测的着法都是针对明确状态作出的、明确且可检验的决策。

执行多步视界。要执行多少个预测着法，取决于当前状态能够确定多少未来。在确定性单智能体任务中，可以直接执行这些着法：这样便可用一次调用替代 H 次调用；当 H 等于解答长度 L时，整个解法就会绘制在单张图像中。面对对手时，下一步之后的着法取决于对手的应对，而这些应对可能与预测的应对不同，因此我们采用滚动时域 [15]：只执行第一个预测着法，观察真实状态，然后再次调用模型。这样便恢复了方程（1）中的闭环，但现在每个决策都附带模型预期会随之发生的后续内容。因此，一种表述便涵盖了单步决策（$H = 1$）、带有明确后续内容的决策（滚动时域）以及完整解法（$H = L$）；第4 节对它们进行了比较。

3.4. 模型与训练

我们用两个预训练模型实例化 p_θ，并通过 LoRA [10], 对二者进行适配，不添加任何新模块。对于 $H = 1$ 和完整解法图像，我们使用图像编辑模型 Qwen-Image-2。1 [25], 该模型通过多模态文本编码器以及置于带噪目标潜变量之前的干净图像潜变量，以 x_t 为条件。对于 $H > 1$，我们使用文生图像到视频模型 Wan2.2-TI2V-5B [20], , 并将 x_t作为其第一帧；决策帧和想象帧会联合生成。两个模型都保留其整流流目标函数 [13,14]: 其中目标潜变量为 z_0，噪声为 $\epsilon \sim \mathcal{N}(0,\ I)$，而 $z_\sigma = (1-\sigma)\, z_0 + \sigma\epsilon$,

$$\mathcal{L}(\theta) = \mathbb{E}_{z_0,\epsilon,\sigma}\big\| v_\theta(z_\sigma, \sigma, x_t, c) - (\epsilon - z_0)\big\|^2. \tag{3}$$

测试时扩展。 NBM 是一项标准的条件生成任务，因此扩散模型现有的测试时技术可以直接应用。我们报告了采用和不采用其中两种技术的结果：多数投票，以及使用较少去噪步数进行采样（第 4 节）；这两种技术都不是我们主要论点成立的必要条件，也不会在推理时增加引擎、基于规则的模块或搜索。第三种技术是通过对比引导，远离一个学得的失败着法提示；其评估结果见附录 M。

多数投票。 在相同状态下，使用独立噪声对模型进行 n 次采样，然后执行出现频率最高的解码后的着法，

$$\hat{a}_t = \operatorname*{mode}_{i=1,\dots,n} D\big(\hat{y}_t^{(i)}, x_t\big), \qquad \hat{y}_t^{(i)} \sim p_\theta(\cdot \mid x_t, c), \tag{4}$$

这与语言模型的自一致性规则 [21] 相同：该规则汇总模型自身的预测，且不比较任何后续内容。

4. 分析与消融实验

设定。 国际象棋局面取自 Lichess 谜题数据库，并按棋类游戏划分，以防训练集与评估集重叠。我们的主要模型使用多达三百万道谜题进行训练，并在 10,000 个留出的棋局上进行评估（附录 A.1）。图像模型预测每道谜题解法中的下一步着法，而视频模型预测最多八个半回合，包括对手的应对着法。其余任务由精确求解器或引擎生成，仅用于构建训练目标并为预测评分（附录 A）。除非另有说明，我们使用20

	NBM 准确率（%）			对称棋盘（0 = 无变化）		一个棋子发生变化（%）		
	所有 ↑	弃子 ↑	远离训练 ↑	颜色 flip	镜像	保持走法：最佳着法不变 ↑	保持走法：最佳着法改变 ↓	找到新的最佳着法 ↑
Ours								
图像模型	56.0	46.3	50.4	+1.2	−1.9	82.3	35.3	39.5
+ 多数投票	63.8	54.9	58.7	+1.2	−1.5	89.7	37.7	42.5
+ 8个样本的预言机†	*85.4*	*83.5*	*82.3*	–	–	–	–	–
+ 一步	62.5	51.7	56.6	**−0.1**	−1.8	94.7	35.9	43.6
+ 一步 +多数投票 63.2		53.0	57.9	**−0.1**	−1.8	94.8	36.1	44.4
视频模型	56.4	47.9	51.0	**+0.1**	−2.7	94.3	42.0	38.4
+ 多数投票	**65.4**	**57.3**	**61.5**	**−0.1**	−2.9	94.5	39.7	42.6
+ 8个样本的预言机†	*86.6*	*82.1*	*84.1*	–	–	–	–	–
+ 一步	63.9	55.6	59.1	+0.7	−1.1	95.8	37.0	45.5
+ 一步 + 多数投票 64.7		56.2	60.0	+0.9	**−0.5**	**95.9**	36.6	**46.6**
从训练中复制								
最近的训练局面	18.6	16.4	8.7	+1.2	−4.3	36.2	**32.1**	6.7
16 个最近邻的多数票	27.0	27.0	16.8	+0.8	−6.9	56.5	48.1	12.5
随机合法着法	5.5	3.8	4.0	–	–	–	–	–

表1。 **回忆还是落子？** 我们在留出国际象棋谜题上比较模型：既比较单个样本，也比较多数票；并将其与从最相似的训练局面复制着法进行对比。测试既包括留出谜题（“远”：与最近的训练局面相差 13–16 个棋格），也包括经过对称变换或单个棋子变化后的同一谜题（这可能会显著改变 NBM）。 †至少有一个样本答对。所有模型均在三百万道谜题上训练。更多详情和变体见附录 E。

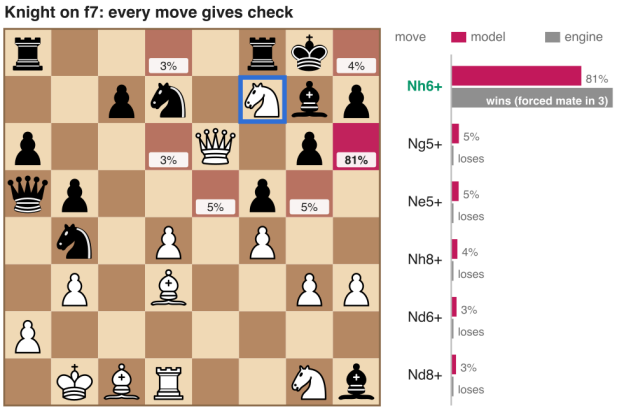

图4。**隐式策略**。这里每一步马步都能将军，但按模型的去噪误差为每步评分后，81% 的概率都分配给了 Nh6+，这是唯一能获胜的着法。

个欧拉步处理图像模型，并使用 30 个 UniPC 步处理视频模型；无效动作计为失败。对于国际象棋，我们报告下一步最佳着法准确率；对于寻路，我们报告最优路线所占比例。

回忆还是实战？ 为了检验这些决策反映的是局面本身，而不是回忆起的训练样例，我们将其与从训练集进行的检索结果进行比较，在经过棋类游戏精确对称变换的棋盘上进行评估，并改变单个棋子，使最佳着法发生变化：模型的选择会跟随这些变化，而检索结果不会。

模型依赖的是正确的棋子吗？ 图 5 展示了同一个留出谜题（评级为 1932）中的相同效应。在原始棋盘上，我们的模型走车弃子 Rxh6+，这会强制形成三步将死。在 c3 处添加一个白格象后，局面在视觉上几乎没有变化，但最佳着法发生了改变：此时白格象控制 h8，因此 Qh8# 可一步将死，模型也切换为这一着法。为了了解每个决策所依据的因素，我们每次移除一个棋子，并测量模型对其着法的置信度下降了多少。在原始棋盘上，这次弃子取决于构成后续杀网的棋子，即负责给出最后的将军的皇后，以及封锁王的逃跑路格的象和兵。添加白格象后，该决策最依赖于使新的将死成立的两个棋子：控制 h8 的象，以及黑方 g6 上自己的兵；后者封锁了王唯一的逃跑路格。移除其中任意一个，Qh8 确实就不再是将死。由此，模型按照局面的要求区分了仅相差一个棋子的两个棋盘，而表 1 显示，这一结论在数千个此类变化中都成立。

隐式策略。 扩散模型还可以通过去噪误差为它未采样的输出评分，这与扩散分类器 [3, 12] 中的做法类似。仅在此分析中，我们枚举所有合法着法，将每种着法绘制到输出图像中，在共享噪声下测量模型对其去噪的效果，再通过 Softmax 将误差转换为策略。它最可能的着法比同一模型任何采样读出所选的着法更常正确，而且它对其余着法的排序与

Task	视觉绘制的动作	指标（%，↑）	图像	视频	最强基线	
最短路径	网格上的路径	最优、留出地图	88.8	**94.6**	基础模型、零样本	0.0
		最优、更长的路径	47.0	**66.6**	基础模型、零样本	0.0
		最优、更大的地图	79.0	**79.8**	基础模型，零样本	0.0
成本感知路径（从不显示成本）	地形路线	最低成本路线	**87.5**	82.8	最少步数（广度优先搜索）	59.2
		绕行或穿越	**85.0**	76.8	机会	50.0
四子棋	列中的棋子	合法、正确的落子行	**100.0**	99.6		
		无需前瞻即可保持获胜	**67.5**	64.5		
		+ 一步	**76.4**	72.0	一步启发式	60.4
Go (19×19, KGS)	点上的棋子	合法着法	99.8	–		
		最佳着法	20.9	–	随机合法着法	0.7
		+ 一步	**31.1**	–	复制最近局面	1.7
台球	提示角度，力度	球入袋	18.6	**34.6**	复制最近的击球	19.0
		+ 一步	28.0	**47.9**	随机击球	0.7

表 2. **超越国际象棋**。 在以路线、棋子、棋子或连续击球表示动作的任务上，采用相同且未经修改的训练方案。规则、地块代价和物理规律均未提供；模型仅从示例解法中学习。路线是逐步构建的：图像模型每次调用标记一步着法，而视频模型则在每个视频片段中生成多个着法。

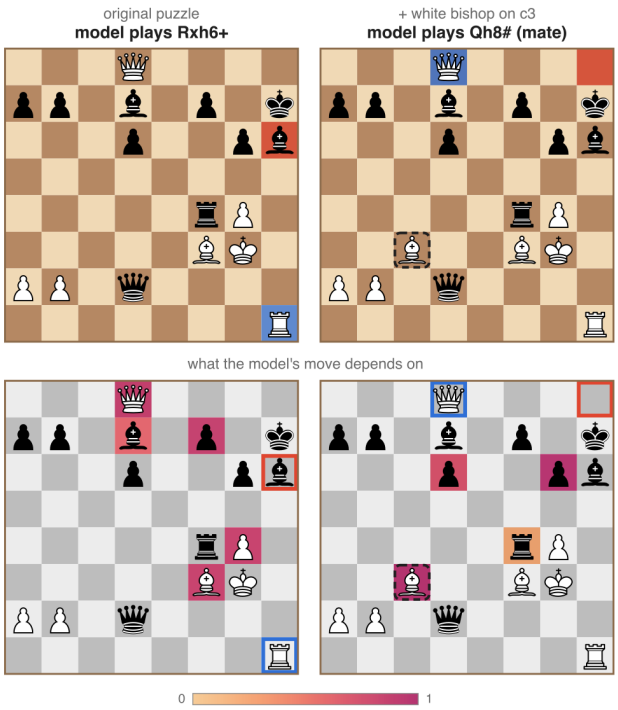

图5. “顶部”： 两个相似的棋盘配置，唯一的区别是在 c3 格新增了一枚白格象。“底部”： 移除每枚棋子后，模型对该着法的置信度会下降多少。

Stockfish（附录 D.2）。在图 4 中，每种马步都能将军，但策略的大部分概率却分配给了 Nh6+，这是唯一能获胜的着法。

一步读出决策。 训练使模型能够针对棋盘的着法 m 学得一个分布 $\pi(m \mid c)$ c，而去噪步数只决定如何读出这个分布。原因在于训练目标：根据带噪输出预测干净输出的最佳结果，是所有可能产生该带噪输出的干净输出的均值；在完全噪声下，输入不包含任何关于着法的信息，此均值就是各着法图像 I_m 按其概率加权后的平均值，

$$\hat{x}_0 \;=\; \mathbb{E}[\,x_0 \mid c\,] \;=\; \textstyle\sum_m \pi(m \mid c)\, I_m \qquad \text{（完全噪声）} \tag{5}$$

(Tweedie 公式 [4]；证明见附录 D.1）。因此，一步便能返回完整分布，而其中最显著的标记会指出最可能的着法。这就是为什么在所有要求单步决策的任务中，一步都是更准确的读出方式（图 6，表1），同时它也揭示了自身的局限：由于平均值几乎不受噪声影响，重复的一步输出彼此一致，而八个这样的输出仅能涵盖 69.2% 国际象棋谜题的答案。相比之下，完整采样会根据每次噪声抽样所对应的概率，追踪到一个单独的着法。单个样本答对的概率较低，但八个样本能涵盖 85.4% 的答案。因此，一步更适合给出答案，而采样更适合提供候选项；二者之间的差距代表着尚有提升空间，而任何尚未获得答案的读出方式都无法利用这一空间。平均值也解释了为什么路线需要完整采样：多条路线的平均值本身并不是一条路线。在所有任务中，决策都在去噪的前几步内确定（附录D.3），这与视频模型求解迷宫时的观察结果一致 [16]；剩余步骤只是细化一个候选项，而不是在多个候选项之间进行搜索（附录 D）。

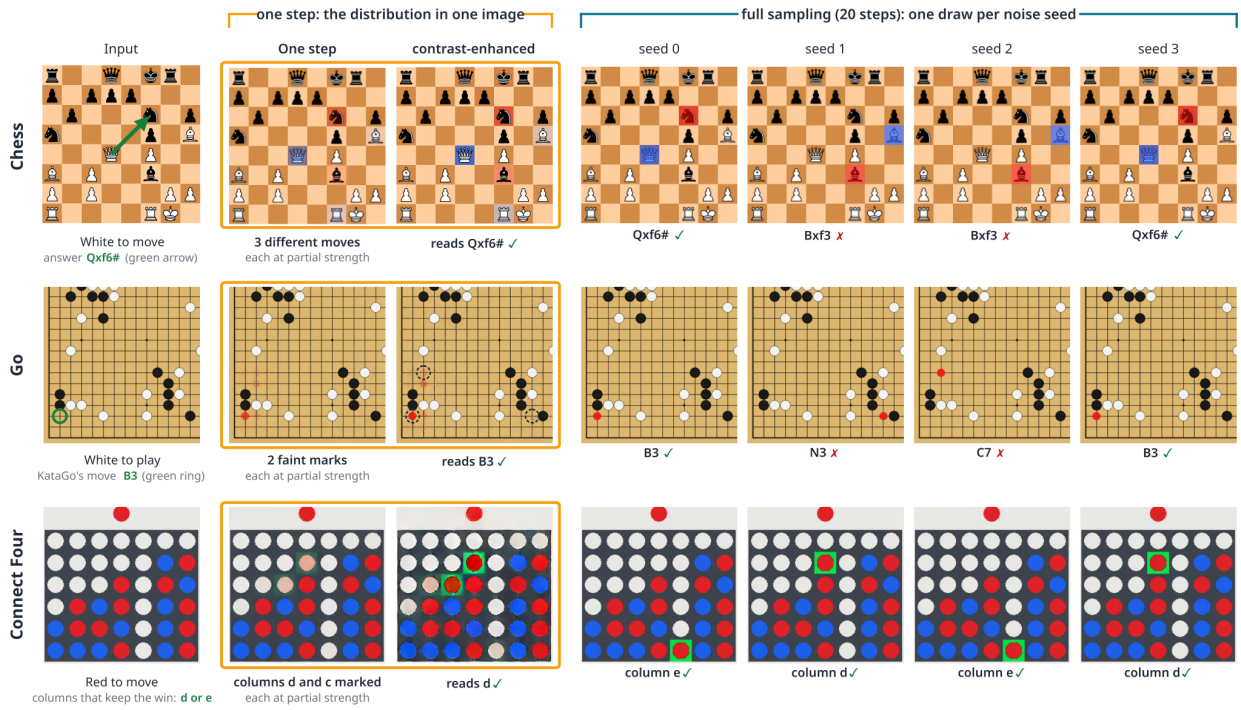

图 6。 **一步读取决策；完整采样则抽取一个决策。** 在留出国际象棋（上）、围棋（中）和四子棋谜题（下）中，单步输出以不同强度标记出多个候选走法，并将最强的标记读取为决策。相反，完整的 20 步采样会确定单个着法，而不同的噪声种子会确定不同的着法。在这道国际象棋谜题中，半数样本走 Bxf3，而不是单步输出所读取的将死着法 Qxf6#。更多定量结果见附录 D。

超越国际象棋。 相同的表述方式无需修改，即可应用于动作空间截然不同的任务（表 2）。这些任务的共同点恰好满足了该表述方式的要求：玩家观察到的内容可以呈现在一张图像中，而动作可以表示为对该图像的编辑，例如标记要移动的棋子及其终点、要落子的点，或击球的方向和力度。在这一类任务中，模型必须推断出各项任务中从未明说的特性。对于寻路任务，这种特性是地形成本：在成本信息被隐藏的情况下，闭环图像模型在 87.5% 的地图上找到了最低成本路线；而忽略地形的广度优先搜索仅为 59.2%。在均匀地图上，每个视频片段规划多个着法的视频模型则能更好地泛化到更长的路径。在四子棋中，这种特性是重力：在图像模型的 2,000 个测试局面中，只有一个局面的棋子没有落在正确的行；视频模型则只有八个。围棋中的特性是游戏规则：棋子有 99.8% 的时间落在合法落子点上；模型在 31.1% 的棋局中与 KataGo 的着法一致，而复制最近训练局面的比例为 1.7%。台球中的特性是球台物理特性：视频模型在 47.9% 的留出布局中成功将球打入袋中，而重复使用最近的训练击球的比例为 19.0%（附录A）。

从后续变化中学习。 描绘未来也有助于模型决定当前的行动。训练来绘制完整解题路线（包括对手回应）的视觉模型，比仅针对第一步棋训练的同一模型更准确地预测下一步最佳着法，尽管评分时只考察第一步棋（在国际象棋中分别为 24.6% 和 21.4%，使用相同的 400,000 道谜题和训练预算）；在地形任务中，两种模型也呈现相同的结果。针对专门设计的寻路任务所做的对照实验揭示了原因（附录K）：当第一步棋取决于沿着两条其他方面完全相同的走廊向下 k 若干格处是否有一堵墙时，在训练量相同的情况下，仅针对第一步棋训练的模型会忽略两格之外的墙，而训练来绘制路线的模型则会利用最远十二格之外的信息；这种增益来自训练，而非推理时读取绘制出的后续内容。绘制出的着法是否也应当执行，取决于任务。在最短路径等确定性任务中，执行整个视频片段与每走一步后重新预测一样准确，却能大幅减少模型调用。面对对手时，根据观测到的棋盘重新预测，优于遵循绘制出的后续内容（在相同棋盘上，下一步着法的准确率分别为 52.5% 和 41.2%）。预测未来能让每次决策都更好；而面对对手时，每次决策最好都基于棋盘的实际状态。

5. 相关工作

介绍了我们的表述和结果后，现在我们将其置于视觉问题求解、序列生成和学习型决策的背景下进行探讨。我们关注的是如何从视觉示例中学习任务规则和动作偏好，并利用这些知识来选择下一步最佳着法（NBM）。

Goren 等人 [7] 表明，视觉扩散模型可在像素空间中生成几何问题的完整解，包括内接方形、施泰纳树和最大面积多边形。我们将视觉解法生成扩展到序贯决策：在国际象棋和寻路任务中，即使某个动作的价值取决于其后续影响，连续预测的动作也能构成解法。

DiffThinker [8] 对预训练的图像编辑模型进行微调，以生成迷宫、旅行商问题（TSP）、数独和拼图等任务的完整视觉解法。视频模型也展现出一定的推理能力，既能以零样本方式 [23] 生成迷宫轨迹，也能在监督微调后 [26]生成。这些工作表明，视觉生成可用于产出完整解法或其逐步展开的轨迹。我们的单步模型采取了不同的方案：它选择一个动作，但不生成其未来的后续内容。我们的视界模型则将这两种方案联系起来，通过生成较短的后续内容，使我们能够比较在有无显式视觉轨迹的情况下做出的决策。

Wewer 等人 [22] 与我们的工作密切相关，他们展示了序列生成能够提升去噪模型中的推理能力。他们的空间推理模型利用不确定性，为视觉数独及其他约束任务中的空间变量确定生成顺序。他们的方法会选择下一步要解决的解法中的未知部分，并优先处理能够以更高置信度进行预测的变量。我们的实验则根据动作对最终结果的贡献来进行动作选择。国际象棋中的弃子，以及更长但更便宜的地形路线，都说明仅凭即时增益无法判断某些选择。

¨Ezquerro 和 Ozdenizci [5] 通过以对象为中心的关系抽象来改进逻辑约束满足，这些关系抽象引导完整解图像的生成。他们展示了在彩色数独、七巧板和基于逻辑的人脸生成等任务上的改进。他们的贡献核心在于加入对对象及其关系的表示。我们的模型通过微调预训练视觉扩散模型获得决策能力，无需额外的对象发现或关系模块。

无需在推理时进行搜索即可学习强大的动作，这一方向也曾在采用符号表示的国际象棋中得到研究。Ruoss 等人 [18] 使用引擎标注的棋局进行训练，无需在推理时搜索便达到特级大师水平。DiffuSearch [27] 使用离散扩散来预测未来状态和动作以及下一步，并将生成的后续内容用作一种隐式搜索。这些方法采用国际象棋专用编码，使用的引擎标注棋局数量多出几个数量级。我们探究通用视觉生成器能否仅从像素中学习有效且合理的动作，适用于从国际象棋到台球等各种任务。与受限的国际象棋输出词汇不同，我们的像素读出必须仅凭样例学习合法且有效的移动。我们的视野变体进一步将我们的方法与 DiffuSearch 联系起来，通过视觉方式表达预测的动作和未来状态。

我们基于时域的执行方式也与用于规划与控制的扩散模型有关。Diffuser [11] 通过采用不断后移的时域规划进行去噪来生成轨迹，而 Diffusion Policy [2] 则预测动作序列，并在再次进行预测之前执行每个序列中的一部分动作。我们的时界模型遵循相同的执行原则：执行预测出的第一个动作，观察结果状态，然后再次进行预测。在这里，状态和动作标记共享一种视觉表示，而动作则从生成的像素中解码出来。在对抗性任务中，生成的后续内容还包括预测的对手应着。这一联系使我们对单步着法预测与基于时域的预测的比较，置于通过生成模型学习行动这一更广泛的研究背景之中。

6. 讨论与结论

在这项工作中，我们展示了视觉扩散模型能够通过学习绘制下一步着法，来学习决定下一步该做什么。这一简单的表述支持范围出人意料地广泛的决策：为了数步之后的收益而牺牲一枚车，选择更长的路线以避开高代价地形，或是在台球中选择一次击球。在序列任务中，模型一次完成一个决策，逐步构建出一个解法，而无需显式地搜索未来延续。

一个自然的问题是：为什么我们要生成着法，而不是对其进行分类？在国际象棋中，输出专门用于对棋盘格对进行评分的分类器，表现可能优于我们的模型。生成的优势在于，同一种输出表示可以适用于不同任务：模型可以直接在图像上绘制国际象棋着法、路线或台球击球。这不仅仅是重复使用同一种架构。当我们用相同的提示同时训练一个模型来完成四子棋和寻路任务时，它在四子棋上的表现与单任务模型相当，在寻路上的表现也只比单任务模型低几分。附录详细报告了这些比较结果。

最引人注目的发现是，与完全去噪样本相比，第一步去噪对于选择着法可能更有用。它会以相互重叠的标记呈现几种可能的着法。在我们的实验中，读取最明显的标记，比绘制一个完全去噪样本更能准确地做出单步决策，而且无需额外训练。

当我们想要获得多样化的候选方案时，完全去噪仍然有用。对于通过绘制来学习行动的模型来说，第一张草图就已经能更好地指引下一步该怎么做。

参考文献

[1] Giulio Biroli、Tony Bonnaire、Valentin De Bortoli 和 Marc M´ezard。扩散模型的动力学机制。《自然·通讯》，15(1):9957，2024年。7[2] Cheng Chi、Zhenjia Xu、Siyuan Feng、Eric Cousineau、Yilun Du、Benjamin Burchfiel、Russ Tedrake 和 Shuran Song。Diffusion Policy：通过扩散动作实现视觉运动策略学习。《国际机器人研究期刊》，44 (10-11):1684–1704，2025年。8[3] Kevin Clark 和 Priyank Jaini。文本到图像扩散模型是零样本分类器。《神经信息处理系统进展》，36:58921–58937，2023年。5[4] Bradley Efro n。Tweedie 公式与选择偏差。《美国统计学会期刊》，106(496):1602–1614，2011年。6, 5[5] Ana Ezquerro 和 Ozan ¨Ozdenizci。利用扩散模型进行空间推理的关系抽象，2026年。1, 8[6] Fabian Gloeckle、Badr Youbi Idrissi、Baptiste Rozi`ere、David Lopez-Paz 和 Gabriel Synnaeve。通过多令牌预测打造更优、更快的大语言模型。arXiv预印本 arXiv:2404.19737，2024年。17, 19[7] Nir Goren、Shai Yehezkel、Omer Dahary、Andrey Voynov、Or Patashnik 和 Daniel Cohen-Or。视觉扩散模型是几何求解器，2025年。1, 8[8] Zefeng He、Xiaoye Qu、Yafu Li、Tong Zhu、Siyuan Huang 和 Yu Cheng。DiffThinker：迈向基于扩散模型的生成式多模态推理，2025年。1, 8[9] Jonathan Ho 和 Tim Salimans。无分类器扩散引导。arXiv预印本 arXiv:2207.12598，2022年。20[10] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。LoRA：大语言模型的低秩适配。arXiv预印本 arXiv:2106.09685，2021年。4[11] Michael Janner、Yilun Du、Joshua B Tenenbaum 和 Sergey Levine。基于扩散的规划，实现灵活的行为合成。arXiv预印本 arXiv:2205.09991，2022年。8[12] Alexander C Li、Mihir Prabhudesai、Shivam Duggal、Ellis Brown 和 Deepak Pathak。你的扩散模型其实是一个零样本分类器。载于2023年IEEE/CVF国际计算机视觉大会（ICCV），第2206–2217页。IEEE，2023年。5[13] Yaron Lipman、Ricky TQ Chen、Heli Ben-Hamu、Maximilian Nickel 和 Matt Le。用于生成建模的流匹配。arXiv预印本 arXiv:2210.02747，2022年。4[14] Xingchao Liu、Chengyue Gong 和 Qiang Liu。流动更直、更快：利用整流流学习生成和迁移数据。arXiv预印本 arXiv:2209.03003，2022年。4[15] David Q Mayne、James B Rawlings、Christopher V Rao 和 Pierre OM Scokaert。约束模型预测控制：稳定性与最优性。《自动化学报》，36(6):789–814，2000年。4

[16] Kaleb Newman、Tyler Zhu 和 Olga Russakovsky。视频模型能够提前推理：利用规划承诺求解迷宫。arXiv 预印本 arXiv:2603.30043，2026年。6[17] Gabriel Raya 和 Luca Ambrogioni。生成式扩散模型中的自发对称性破缺。《神经信息处理系统进展》，36：66377–66389，2023年。7[18] Anian Ruoss、Gr´egoire Del´etang、Sourabh Medapati、Jordi Grau-Moya、Li K Wenliang、Elliot Catt、John Reid、Can- nada A Lewis、Joel Veness 和 Tim Genewein。利用大规模变换器进行摊销规划：国际象棋案例研究。《神经信息处理系统进展》，37：65765–65790，2024年。8, 10, 21[19] Zhiqing Sun 和 Yiming Yang。Difusco：用于组合优化的基于图的扩散求解器。《神经信息处理系统进展》，36 3731，2023 3706 10 : – .[20] Team Wan、Ang Wang、Baole Ai、Bin Wen、Chaojie Mao、Chen-Wei Xie、Di Chen、Feiwu Yu、Haiming Zhao、Jianx- iao Yang 等。Wan：开放且先进的大规模视频生成模型。arXiv 预印本 arXiv:2503.20314，2025年。4[21] Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。自洽性提升语言模型中的思维链推理。arXiv 预印本 arXiv:2203 11171，2022 4 . .[22] Christopher Wewer 、 Bart Pogodzinski 、 Bernt Schiele 和 Jan Eric Lenssen。利用去噪模型进行空间推理，2025年。1, 8[23] Thadd¨aus Wiedemer、Yuxuan Li、Paul Vicol、Shixiang Shane Gu、Nick Matarese、Kevin Swersky、Been Kim、Priyank Jaini 和 Robert Geirhos。视频模型是零样本学习者和推理器。arXiv 预印本 arXiv:2509.20328，2025年。8[24] Noam Wies、Yoav Levine 和 Amnon Shashua。子任务分解使序列到序列任务中的学习成为可能。arXiv 预印本 arXiv:2204.02892，2022年。17[25] Chenfei Wu、Jiahao Li、Jingren Zhou、Junyang Lin、Kaiyuan Gao、Kun Yan、Sheng-ming Yin、Shuai Bai、Xiao Xu、Yilei Chen 等。Qwen-image 技术报告。arXiv 预印本 arXiv:2508.02324，2025年。4[26]Cheng Yang、Haiyuan Wan、Yiran Peng、Xin Cheng、Zhaoyang Yu、Jiayi Zhang、Junchi Yu、Xinlei Yu、Xiawu Zheng、Dongzhan Zhou 和 Chenglin Wu。视频推理：首次通过迷宫求解任务评估视频模型的推理能力。arXiv 预印本 arXiv:2511.15065，2025年。8[27] Jiacheng Ye、Zhenyu Wu、Jiahui Gao、Zhiyong Wu、Xin Jiang、Zhenguo Li 和 Lingpeng Kong。通过离散扩散实现隐式搜索：一项关于国际象棋的研究。载于《国际学习表征会议》，第100369–100390页，2025年。8

使用视觉扩散模型学习下一步最佳着法

补充材料

A. 任务与数据

每项任务都会呈现一个渲染后的状态，并要求模型绘制一个动作。视频模型还会绘制后续动作和状态。提示中从不说明规则、代价或物理规律。除国际象棋和围棋外，提示都只有一句“继续”。精确求解器或强大的引擎会提供标签，用于构建数据并为预测评分。下文将分别说明每项任务、其渲染方式、数据与数据集划分，以及评估方法。

A.1. 国际象棋

问题。 棋局取自Lichess谜题数据库。每道谜题都从对手走完最后一步后开始，并附有一条经引擎验证的解题路线；在这条路线中，只有行棋方的第一步棋被要求是唯一着法。目标是找出这第一步棋。

渲染。 棋盘从白方视角绘制，尺寸为256×256 像素，不带坐标。模型通过将起点格染成蓝色、终点格染成红色来标记着法。图像模型的提示是“白方行棋。标出最佳着法：起点格标为蓝色，终点格标为红色。”（黑方同理）。视频模型还会在移动棋子的合法目标格上绘制绿色圆点，随后展示走完这步棋后的棋盘，如第 B 节所述。

数据与数据集划分。 谜题根据其来源棋类游戏的哈希值分配到不同的数据集划分，因此同一棋类游戏的所有谜题都属于同一划分。凡是起点局面出现在验证集或测试集中的谜题（棋子、行棋方、王车易位和吃过路兵权利均相同），都会从训练集中移除，这也排除了局面转置。验证集和测试集仅保留经过充分验证的谜题（等级分偏差不超过 90，且至少有 200 次对局），并按八个等级分区间分层，从低于 1000 到高于 2500。最终得到 5,808,734 道训练谜题、2,000 道验证谜题和 10,000 道测试谜题。表 1 中的模型使用前三百万道训练谜题进行训练。除非另有说明，本补充材料中的变体和分析均使用前一百万道谜题。对于视频模型，每道谜题都转换为一段最多包含八个解题半回合的棋谱（平均为 3.6 个）。

引擎标签。 对于隐式策略分析和对比引导，我们使用 Stockfish 19 对每个合法着法进行评分（所有着法均搜索至深度 12，每个局面最多 4M 个节点，单线程）。分数通过 Lichess 公式转换为行棋方的胜率，强制将杀的排名高于所有厘兵分。引擎给出的最佳着法与 99.8% 的测试局面中的谜题答案一致。

指标和基线。当从生成图像中读出的着法（第C节）与谜题答案一致时，即视为预测正确。忽略升变棋子。我们报告两个子集：弃子（Lichess 标记为弃子的 1,001 道测试谜题）和远离训练集，即最近训练局面与之在 64 个棋格中有 13 至 16 格不同的测试谜题（相较于 300 万道训练谜题，共 1,757 道；相较于前 100 万道，共 2,403 道）。检索基线及其构建方法见第E节。

A.2. 最短路径和地形

问题。 一个智能体必须在 $N \times N$ 网格（$N \leq 12$）中到达目标。在均匀变体中，每一步的成本都相同，目标是找到一条最短路径。在地形变体中，单元格分为地板、泥地或沙地，进入这些单元格的成本分别为 1、2 或 4，目标是找到一条最低成本路径。墙壁不可进入。成本从不直接显示，必须根据样例推断。

渲染。 地图以 384×384 像素绘制，单元格大小为 32 像素，因此一个单元格恰好对应一个视频模型词元和 2×2 图像模型词元。智能体是一个蓝色圆盘，目标是一个红色圆环，已访问单元格为绿色。泥地带有条纹，沙地带有点状纹理，因此无需仅凭颜色也能区分二者。无规则解码器通过将每个单元格的中心、圆环和角落颜色与已知调色板进行匹配来读取单元格。

格式。 图像模型以闭环方式运行：它用红色圆盘标记下一个单元格，执行着法，然后将结果状态作为下一次输入。视频模型生成最多包含八步着法的片段，每个片段都包含一个决策帧和结果状态。更长的路线会串联这些片段。滚动时域变体只执行每个片段中的第一个决策（第 C节）。

数据。地图来自三个类别：随机区块（墙壁密度为 15% 至 32%）、带有一处或两处缺口的墙壁，以及通过递归划分生成的房间。地形地图中，泥地或沙地覆盖 25% 至 45% 的可通行单元格。目标通过广度优先搜索计算（地形地图则使用迪杰斯特拉算法），若有多条路径并列，则选择转弯次数最少的路径；评估接受任何最优路径。每种变体都包含 200,000 张训练地图和 600 张验证地图。均匀变体包含 1,000 张分布内测试地图，以及三组分布外测试，每组 500 张地图：更大地图（N = 11 或 12）、更长路径（21 至 32 步）和迷宫；迷宫是训练中从未出现过的类别。地形变体包含 1,000 张测试地图，以及一个由 500 张地图组成的探测集；在这些地图上，每条最低成本路径要么穿越一条高代价地形带（250 张地图），要么绕过它（250 张地图）。训练中不会出现任何评估地图，且各评估集两两互不重叠。

指标与基线。 如果一条路线连通、不触碰墙壁、不重复经过任何位置，且不改变路径之外的任何内容，那么它就是有效的。如果它有效且长度最优（地形：成本最优），那么它就是最优的。在探测集上，如果一条路线有效，

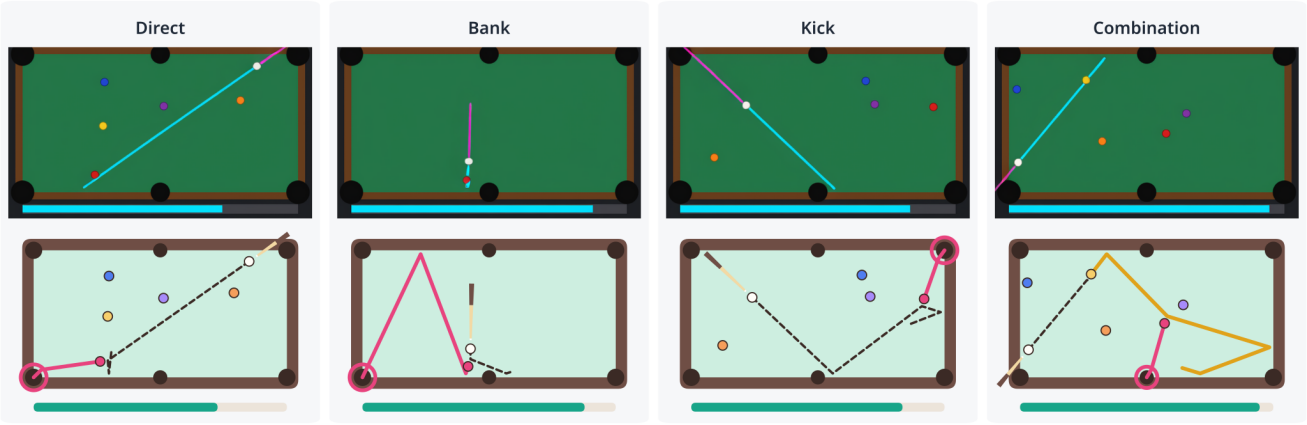

图 7. **台球决策与结果**。 每列展示一种击球类型的留出布局：直接击球、反弹球、借球或组合球。上： 视频模型未经修改的输出，用从母球出发的瞄准线、母球后方的球杆以及填充程度表示击球力度的力度条标出击球。下： 对应的模拟器展开。虚线描绘母球的轨迹，圆环标记接收红球的球袋。反弹球、借球和组合球的布局难度较高，因为不存在可以直接入袋的球。

并且当且仅当最低成本路径穿越该带状区域时，它也穿越该区域，那么这条路线就是正确的。地形盲基线遵循步数最少的最短路径。在 59.2% 的地形测试地图上，它的成本达到最优；根据探测集的构造，它在一半的探测地图上是正确的。未经微调的预训练模型从未生成过有效路线。在迷宫测试中，图像模型和视频模型分别在 78.4% 和 84.4% 的地图上达到最优。

A.3. 四子棋

问题。 标准 7×6四子棋：模型为当前行棋方选择一列。棋子会落到最低的空行这一规则从未被说明。

渲染。 棋盘由 256×256 像素组成，单元格为 32 像素。红方先行，蓝方后行，顶部边缘的棋子标明轮到哪一方行棋。着法绘制为行棋方的棋子落在其着陆单元格中，并染成绿色。

数据。 棋局来自自我对弈游戏：每位棋手以从 $\{0, 0.05, 0.1, 0.2, 0.35, 0.5\}$ 抽取的概率随机选择一个合法着法，否则选择最佳着法或价值保持走法。我们只保留关键局面，即并非所有列的博弈论价值都相同的局面。标签来自 Pascal Pons 的精确求解器。标签是一个最佳列，评估时接受任意一个最佳列。数据包含 200,000 个训练棋局、1,000 个验证棋局和 2,000 个测试棋局，另有一个包含 1,000 个局面的困难测试集，其中没有立即获胜的着法，也没有立即威胁。数据集划分内部及不同划分之间的棋局均按镜像对称去重，且每个评估集每盘棋最多取一个局面。

指标与基线。 合法且落在正确行：标记的列未满，棋子绘制在它落下时所在的行。保持获胜：在困难测试集上，所选着法保留该局面的博弈论价值，因此已获胜的局面仍然获胜，平局局面仍然是平局。在 1,000 个困难局面中，有 936 个是获胜局面。

一步启发式方法会在能够立即获胜时选择获胜着法；否则就阻止对手立即获胜；再否则选择最靠近中间的合法列。

A.4. 围棋

问题。 完整 19×19 围棋。模型会标记 KataGo 偏好的着法。

数据和标签。 我们从公开的 KGS 档案库中选取高水平业余棋手之间的对局（7段，或双方均至少为6段），保留截至2015年进行、至少有80手且没有机器人参与的让先棋局。棋局从每盘棋的中盘阶段抽取。每个局面都由 KataGo 搜索（b11c768 网络，400 次访问，中国规则，贴目 7.5），并将访问次数最多的着法作为标签。已能明确判断结果的局面会被筛除。较小的训练集包含来自 44,191 盘棋的 100,000 个局面，较大的训练集包含来自 91,253 盘棋的 1,000,000 个局面。两者共用 1,000 个验证局面和 2,000 个测试局面，且不同数据集划分之间不共享棋局或棋盘位置。

渲染与提示。棋盘由 384×384 个像素组成，着法是在选定位置放置一个红色圆盘。提示为“黑方下棋。用红点标出最佳着法。”（白方亦然）。

指标与基线。 当解码出的点位与 KataGo 的着法相同时，预测即为正确。每个局面约有 296 个合法点位，因此随机选择一个合法着法，预期正确率为 0.35%。表格中报告的单次随机抽取结果为：2 在 2,000 个局面中，有 15 个预测正确（0.7%）。复制最近训练局面的着法，正确率为 1.7%。作为参考，KataGo 的原始策略网络（不进行搜索）与其搜索所得着法相符的比例为 74.0%，而对局中的实际着法与其相符的比例为 39.5%。

A.5. 台球

问题。 一张球台上有母球、一个红色目标球和最多四个其他球。模型选择一次击球、球杆方向和力度；如果红球入袋且母球未入袋，则视为成功。

模拟器与求解器。 使用确定性的二维模拟器模拟击球，不考虑旋转：滚动摩擦、弹性球碰撞（恢复系数为 0.95）、库边（恢复系数为 0.8）和六个球袋。每种布局模拟 12,960 次击球（每 0.25° 以及九种力度），并将对力度的小幅变化具有鲁棒性的最大成功角度范围的中心作为标签。击球包括直接进袋、反弹球、借球、轻碰球和组合球。

渲染。 球台尺寸为 512×288 像素。击球轨迹绘制为一条青色瞄准线，从母球延伸至母球中心首次碰到库边的位置，并绘制在球的下方；此外，还会在母球后方绘制一根洋红色球杆，并在球台下方绘制力度条。无规则解码器根据瞄准线和球杆读取角度，并根据力度条的填充部分读取力度。在干净的渲染图中，其角度误差中位数为 0.003°。在视频模型的视频片段中，决策帧之后会显示球的模拟运动。

数据与指标。 共有 200,000 个训练布局、1,000 个验证布局和 2,000 个测试布局，并为每个测试布局提供一个镜像副本。根据棋盘的对称变换，不同数据集划分之间不存在共享布局。如果解码得到的击球在模拟器中成功，则该击球计为入袋 。最近击球基线会找到 cue 球和红球距离最近的训练布局（包括镜像布局），并复用其击球动作。它能使 19.0% 的测试布局中的球入袋，而随机击球的比例为 0.7%。在镜像布局上，一步图像模型和视频模型的入袋率分别为 28.2% 和 49.1%；在原始布局上则分别为 28.0% 和 47.9%。

A.6. N 皇后问题

问题与渲染。 在一个部分填充的 N 皇后棋盘上，模型放置下一枚皇后，使棋盘仍可完成，且任意两枚皇后都不共享同一行、列或对角线。带有 $N \le 12$ 的棋盘绘制在固定的 12×12 网格上。经过秩为 64 的 LoRA 微调后，Qwen-Image-2.1 接收提示“继续。”，并用红色标记下一枚皇后。动作根据最明显的红色标记读取，不设最低强度阈值。

数据与评估。 我们使用 200,000 个棋盘训练 24k 步，使用的尺寸为 $N = 8$，9、10。将通过旋转或反射相关联的完整解族整体从训练集中留出，用于验证集和测试集：在 $N = 8$ 中，12 个解族留出 4 个；在$N = 9$中，46 个解族留出 16 个；在 $N = 10$ 中，92 个解族留出 30 个。尺寸 11 和 12 从未用于训练。测试棋盘及其任何旋转或反射形式都不会出现在训练集中。标签由精确枚举提供：任何仍允许得到完整解的摆放方式都是正确的。我们在验证集上选择一个采样步。

结果。 对于留出解族，仅用第一步棋训练的模型有 85.2% 的棋盘可走出可完成着法，并有 99.4% 的棋盘能合法落子。随机选择一个未受攻击的单元格，有 71.2% 的概率可完成解法；从最近的训练棋盘复制着法，则为 62.3%。复制法在训练解族上的准确率为 96.8%，在留出解族上则降至 62.3% ；而模型相较于随机选择未受攻击单元格的优势基本不变（+15.5 和 +14.0 个百分点）。在从未训练过的 11 和 12 这两种尺寸上，模型的准确率达到 79.2%，高于随机基线的 72.6%，合法落子率为 97.0%。在困难棋盘上，模型准确率达到 62.0%，高于对照条件的 49.1%。采样 20 步时，留出解族准确率为 78.8%，低于经验证集选定的一步结果。受控时域测试见第 K 节。

B. 模型与训练

图像模型。 我们使用 LoRA（秩为 64，$\alpha = 64$）对 Qwen-Image-2.1 进行微调，应用于其 Transformer 的注意力投影层和图像多层感知机层。输入图像编码为条件，模型通过整流流速度损失进行训练，以生成编辑后的图像。

视频模型。 我们使用相同秩的 LoRA，对 Wan2.2-TI2V-5B 的注意力层和前馈层进行微调。输入状态占据第一帧，且不计入损失。每个视频片段包含 17 个由四帧组成的区块（共 65 帧）：区块 0 显示输入；对于最多八步着法中的每一步，决策区块显示标记着法，下一区块则显示其结果状态。最后一步着法之后，画面保持最终状态。在 256 像素分辨率下，一个潜在标记覆盖一个国际象棋格。在国际象棋中，每个决策区块还会用绿色圆点标出移动棋子的合法目标格。解码器仅根据蓝色和红色方格读取着法（第 C 节）。这些圆点用于测试图中标出的目标格是否与棋子的合法着法相符，而图像模型的训练不使用这些圆点。

Task	模型	步数	批次	分辨率
国际象棋，3M	图像	375k	8	256
国际象棋，3M	视频	187.5k	16	256
国际象棋，100万	图像	125k	8	256
国际象棋，100万	视频	6.25万	16	256
地形路径	图像	24k	8	384
路径	视频	5k	16	384
四子棋	图像	12k	16	256
四子棋	视频	10k	16	256
围棋，100k	图像	20k	8	384
围棋，100万	图像	125k	8	384
台球	图像	20k	16	512×288
台球	视频	12k	16	512×288

表3。主要结果中所用模型的训练配置。

训练。 所有运行均使用 AdamW，学习率为 10^{-4}，采用余弦衰减、200 个预热步数、权重衰减 10^{-4}，并将梯度裁剪设为 1。表 3列出了支撑主要结果的各次运行。主要表格中的图像模型仅接触

最佳着法提示。用于对比引导的模型（第 M 节），以及在一百万个谜题上训练的国际象棋模型和在 100,000 个棋局上训练的围棋模型，则采用混合提示进行训练：一半的样例要求给出最佳着法，其余样例则要求给出任意合法着法，或一个看似合理但实际是败着的着法。

C. 推理与评估

采样。 图像模型使用 20 步的欧拉流匹配采样器进行采样；视频模型则使用 UniPC 采样 30步（流偏移 3，不使用无分类器引导）。

读取动作。 我们仅从像素解码动作，不使用任务规则。在国际象棋中，我们拟合叠加在每个方格上的蓝色和红色着色，并将蓝色最强的方格选为起点，将红色最强的方格选为终点。其他任务则使用相同原理，并采用各自的调色板（第 A 节）。每种标记中最强的标记决定动作，无论其强度如何。我们不会将读出限制为合法动作，因此非法动作也算作失败。对于视频，我们对每个区块的两个中央帧取平均值，并根据模型刚刚想象的棋盘检查每个决策。

闭环与滚动时域。 要完成整段棋谱或路线，图像模型每步着法调用一次：执行模型给出的着法，并渲染真实的新状态，供下一次调用使用。视频模型每次最多生成八步着法。在滚动时域变体中，每个视频片段只执行第一个决策，并根据结果状态生成新的视频片段。

对比引导。 引导采样的描述和评估见第 M 节。使用 $w = 0$ 采样器为无引导采样器。

多数投票。 我们在一个批次中使用不同噪声抽取 8 个样本，并保留出现频率最高的解码着法；若票数相同，则按样本顺序决定。评分时，将抽样结果不是合法着法的样本合并为一个“无着法”投票，该票永远不可能正确，因此只要此票胜出，预测就算失败。因此，合法性仅在评估时检查；与公式（4）中对原始解码着法取众数相比，这种合并方式更为严格。在台球任务中，投票结果为解码后角度的圆形中位数和力度中位数。对于第 E 节中经过变换和变化后的棋盘，我们使用 4 个样本。

少步采样。 我们还使用更少的去噪步骤进行采样。对于表。1 和表。9，步数仅在验证集上选定：在 1、2、3、4 步和默认设置中，选择准确率最高的设置；若准确率相同，则选择步数较少的设置。由此，表中两个模型均选择了一步。1，而在一百万个谜题的设置下，图像模型选择两步，视频模型选择一步。表。2中的每一行都使用完整采样，而单决策任务（四子棋、围棋、台球）的“+ 一步”行恰好使用一步，不进行选择；表。7报告了所有步数。使用对比引导时，减少步数并无帮助（表。6），因此我们

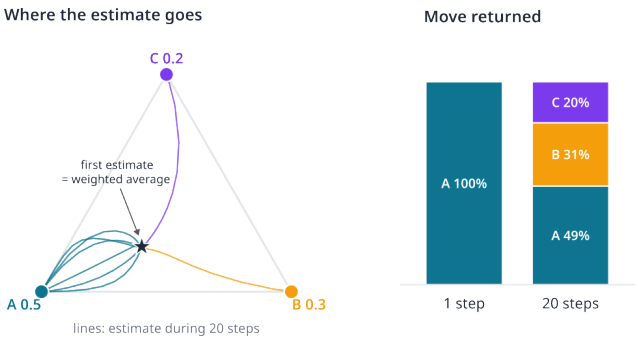

图 8. **一步读取决策，多步对其进行采样。** 一个精确的玩具模型包含三种着法（概率分别为 0.5、0.3 和 0.2），其目标图像各自为一个方格着色，因此这些着法构成三角形的三个顶点。左图：对于每次噪声抽样，干净输出的第一次估计都是概率加权平均（星号）。在 20 步内，该估计从此处移动到一个顶点（线条）。右图： 对 2,000 次噪声抽样返回的着法。

报告称，在引导下没有采用少步设置。

D. 通过扩散训练，一步决策

少步采样、多数投票和隐式策略的表现都优于单个完整采样（表1）。本节说明，读取同一个对象有三种方式：模型在训练中学到的动作分布。在完全噪声下，模型对干净输出的估计是各个动作图像的加权平均，每个图像的权重是其概率，因此最显著的标记对应最可能的动作。一个去噪步骤就能得到这个平均值。相反，许多步会将其转化为从分布中抽取的单个样本，而这种样本答对的频率较低。我们从训练目标推导出这一点（第D.1节），直接测量该分布（第D.2节），并在国际象棋（第 D.3节）和其他任务中（第 D.4节）检验其影响。第 D.5 节讨论了这对推理意味着什么。

图8 展示了一个精确的玩具模型中的这一思路，其中有三种着法，概率分别为 0.5、0.3 和 0.2。无论噪声抽样为何，第一次估计都是同一个点，即三种走法图像的加权平均，其最显著的标记对应最可能的着法。经过二十步后，每个噪声抽样都会转化为三种着法之一，其比例与各自的概率相符。如果答案遵循模型的概率，那么第一次估计有一半的概率正确，而 20 步采样正确的概率为 $0.5^2 + 0.3^2 + 0.2^2 = 0.38$，因为真实值和模型必须各自独立地选出同一个答案。

D.1. 为什么一步读出最可能的动作

设定。 对于棋盘 c 和噪声水平 $\sigma \in (0, 1]$, 训练会将干净输出 x_0 扰乱为 $x_\sigma = (1 - \sigma)\, x_0 + \sigma\epsilon$，噪声水平为 $\epsilon \sim \mathcal{N}(0, I)$，而网络 v 会预测速度$\epsilon - x_0$。它对干净输出的估计为 $\hat{} x_0 = x_\sigma - \sigma v$。我们用 I_m 表示动作 m 的目标图像，用 $m^\star$ 表示

正确动作，并将棋盘的目标视为一个混合：给定 c，干净输出为 I_M，概率为 $M \sim \pi(\cdot \mid c)$，其中$\pi_m = \pi(m \mid c)$是模型分配给动作 m的概率。这忽略了同一动作图像内部的变化。

引理 1（模型预测条件均值）。训练损失 $\mathbb{E}\,\|v(x_\sigma, c) - (\epsilon - x_0)\|^2$在 $v^\star = \mathbb{E}[\epsilon - x_0 \mid x_\sigma, c]$处达到最小值，然后

$$\hat{x}_0(x_\sigma, c) = \mathbb{E}[x_0 \mid x_\sigma, c]. \tag{6}$$

证明。 将 $X = (x_\sigma, c)$ 记作输入，将 $Y = \epsilon - x_0$ 记作目标，并将 $g(X) = \mathbb{E}[Y \mid X]$。对于任意预测器 v，

$$\mathbb{E}\|v(X) - Y\|^2 = \mathbb{E}\|v(X) - g(X)\|^2 + \mathbb{E}\|g(X) - Y\|^2 + 2\,\mathbb{E}\big[(v(X) - g(X))^\top (g(X) - Y)\big].$$

最后一项为零：在给定 X 的条件下，因子 $v(X) - g(X)$ 是固定的，而 $g(X) - Y$的均值为零。第二项与 v 无关，因此当 $v = g$ 时，损失最小。由于 $x_\sigma = (1-\sigma)\,x_0 + \sigma\epsilon$，目标可以写成$Y = (x_\sigma - x_0)/\sigma$，因此

$$v^\star = \big(x_\sigma - \mathbb{E}[x_0 \mid X]\big)/\sigma,$$
$$\hat{x}_0 = x_\sigma - \sigma v^\star = \mathbb{E}[x_0 \mid X]. \qquad \square$$

这是针对流匹配写出的 [4]Tweedie 公式。

命题 1（第一次估计是动作的加权平均值）。在 $\sigma = 1$处，对于每次噪声抽样，

$$\hat{x}_0(x_1, c) = \sum_m \pi_m\, I_m. \tag{7}$$

证明。 当 $\sigma = 1$输入 $x_1 = \epsilon$ 是独立于 x_0 给定 c的纯噪声时，根据公式（6）

$$\hat{x}_0(x_1, c) = \mathbb{E}[x_0 \mid c] = \sum_m \pi_m I_m. \qquad \square$$

单步采样器恰好输出这个估计值，$x_1 - v(x_1, c) = \hat{x}_0(x_1, c)$。图像模型的两步调度（$\sigma = 1 \to 0.02 \to 0$）返回的图像几乎相同，因为它的第二步起始点几乎就在 $\sigma = 0$处。

读取平均值。假设这些标记相加得到 $I_m = B + \beta\, e_{f(m)} + \rho\, e_{t(m)}$，其中 B 是棋盘，e_s 是方格 s的色调，而 $f(m)$和 $t(m)$分别是着法 m的起点和终点。根据公式（7），方格 s 将以强度 $\pi_{\text{from}}(s) = \sum_{m:\, f(m)=s} \pi_m$ 显示为蓝色，以强度$\pi_{\text{to}}(s) = \sum_{m:\, t(m)=s} \pi_m$显示为红色；只要该着法的概率超过一半，选取最强的蓝色方格和最强的红色方格即可读出最可能的着法。概率低于此值时，如果多个着法共用一个方格，读出结果可能会将一个着法的起点与另一个着法的终点配对。实际中这种情况很少见（D.3节）。

读取最可能的动作优于采样。集成到末端的采样器会从 π 中采样一个动作，其正确的概率为$\pi_{m^\star}$。当 $\operatorname{argmax}_m \pi_m = m^\star$ 时，读取最可能的动作是正确的。如果模型经过校准，即在模型为其分配概率 π 的棋盘中，正确动作在其中 π_m 的比例上为 m，那么在棋盘上的两种准确率分别为 $\mathbb{E}[\sum_m \pi_m^2]$和 $\mathbb{E}[\max_m \pi_m]$，并且由于每个 $\pi_m \le \max_{m'} \pi_{m'}$，

$$\sum_m \pi_m^2 \;\le\; \max_{m'} \pi_{m'} \sum_m \pi_m \;=\; \max_m \pi_m, \tag{8}$$

仅当 π 在其允许的动作上均匀分布时，等号才成立。对于确定性模型和均匀模型，这一差距都会消失，在两者之间时则最大。对样本进行多数投票会收敛到最可能的动作；当领先动作之间的差距很小时，收敛速度较慢。至少有一个 n 样本正确的概率为 $\mathbb{E}[1 - (1 - \pi_{m^\star})^n]$，即使 $m^\star$ 不是最可能的动作，该概率也会趋近于一。

D.2. 衡量分布：隐式策略

为了检验这些预测，我们需要 π 本身。条件扩散模型可以通过衡量输出的去噪效果，为未采样的输出打分。扩散分类器利用这一点，选出能让图像最容易去噪的提示所对应的类别 [3, 12]。我们用这种方式为每个合法着法打分。与第C节的读出不同，这项分析已知合法着法。

为着法打分。我们用变分自编码器将 I_m 编码为潜变量 $x_0^{(m)}$，添加噪声 $x_\sigma^{(m)} = (1-\sigma)\,x_0^{(m)} + \sigma\,\epsilon_k$，并对预测速度的误差取平均，

$$E(m) = \frac{1}{|\Sigma|\,K} \sum_{\sigma \in \Sigma} \sum_{k=1}^{K} \frac{1}{d} \left\| v\big(x_\sigma^{(m)}, c\big) - \big(\epsilon_k - x_0^{(m)}\big) \right\|^2,$$

对 $\Sigma = \{0.8, 0.9\}$ 个 $K = 8$ 噪声抽样，其中 d 是潜变量的维度。棋盘上的所有着法都使用相同的噪声抽样 ϵ_k 进行评分，因此 $E(m)$ 的差异仅由着法造成。隐式策略是在合法着法上按 $q(m) \propto \exp(-E(m)/T)$ 进行归一化。

误差的计算方式为 π。 固定一个着法 m 和一次噪声抽样 ϵ，将 $z = (1-\sigma)\,I_m + \sigma\epsilon$ 记为该着法图像的带噪版本。模型接受训练要预测的速度为 $\epsilon - I_m = (z - I_m)/\sigma$，而它预测的速度可写作 $v = (z - \hat{x}_0)/\sigma$，其中 $\hat{x}_0$ 是模型对干净输出的估计。将两者相减，着法的误差为

$$\big\|v - (\epsilon - I_m)\big\|^2 \;=\; \frac{1}{\sigma^2} \big\|I_m - \hat{x}_0\big\|^2, \tag{9}$$

对于任意网络：着法的误差就是该着法与模型估计值之间的距离。

对于最优模型，式（6）使该估计成为所有着法的加权平均值，

$$\hat{x}_0 = \sum_k w_k I_k, \qquad w_k \;\propto\; \pi_k \exp\!\Big(-\frac{\|z - (1-\sigma)I_k\|^2}{2\sigma^2}\Big),$$

其中，w_k 是给定含噪输入时着法 k 的概率，根据贝叶斯定理，它等于先验概率 π_k 乘以着法 k 对z的解释能力。

对于距离为 $d = \|I_m - I_k\|$ 的两个候选着法 m 和 k，m 的误差为 $(1-w_m)^2 d^2/\sigma^2$，并且

$$\log \frac{w_m}{w_k} = \underbrace{\log \frac{\pi_m}{\pi_k}}_{\text{belief}} + \underbrace{\frac{(1-\sigma)^2}{2\sigma^2} d^2 + \frac{1-\sigma}{\sigma} \langle \epsilon,\, I_m - I_k \rangle}_{\text{evidence in } z}. \tag{10}$$

对于固定的含噪输入，较大的 π_m 会带来较大的 w_m，因此误差更小。每个着法都基于各自的含噪输入进行评分，因此两个着法之间的比较中，最后一项的符号相反；这一比较是在噪声的平均意义上成立，而非对每次抽样都成立。这就是我们对 K 次抽样取平均值的原因，而平均误差最低的着法，就是模型最相信的着法。候选项超过两个时，共用一个方格的着法比孤立的着法更接近平均值，因此误差仅近似遵循 π，第 D.3节会衡量这种近似程度。噪声水平决定了两部分之间的平衡：接近 $\sigma = 0$ 时，证据占主导地位，每个着法都能被完美去噪；而在 $\sigma = 1$时，证据消失，误差就是到式（7）中平均值的距离。像我们使用的中间水平则兼顾两者。

结果。 在前 1,000 道测试谜题中，误差最低的着法是 67.0% 谜题的答案，并且在 88.1% 的谜题中位列前三（随机合法着法：5.6% 和 16.1%；使用四次噪声抽样评分的预训练模型：6.7% 和 15.5%）。在每道谜题中，误差也能按照与 Stockfish 一致的顺序对其他着法进行排名；即使只看安静着法，并且剔除棋子类型、吃子和将军的影响，结果依然如此。

D.3. 国际象棋中的证据

我们在前 1,000 道测试谜题上，检验表1所示结果对图像模型的影响。我们拥有该模型的单步输出、八个 20步样本以及隐式策略。为将策略与样本进行比较，我们根据样本所选的着法拟合其温度，不使用答案（$T = 0.0017$）。

第一个估计值是预测平均值。图 9展示了单步输出以及平均值旁边的结果 $\sum_m q_m I_m$，命题 1 根据隐式策略预测了该平均值。在 1,000 道谜题的全部 64 个棋格上，单步输出的蓝方和红方强度都紧随预测强度 π_{from}和 π_{to}（皮尔逊相关系数分别为 0.95 和 0.93）。正如命题 1 所预测的那样，输出几乎不依赖噪声：两次噪声抽样在 89.4% 的谜题中给出了相同的着法。在 80.2% 的谜题中，从输出中读出的着法是 q 下最可能的着法。第 D.1 节所述的错配仅在 1,000 道谜题中的 40 道里被预测出来，而在这些谜题中，输出显示出错配的



图 9. **一步即可呈现整个分布。** 上图：对于每道谜题，模型的单步输出中的标记与隐式策略加权后的所有着法平均值十分接近 q（命题 1）：每个候选项按其概率出现，读取最显著的标记即可得出最可能的着法，而单个 20步样本往往会漏掉该着法（条形图：q 与 8 个样本所占的比例）。下图：对 1,000 道谜题中的每个方格进行相同的比较：单步输出中每个方格的蓝色（红色）色调，对应于 π_{from}（π_{to}），即着法从该处起始（在该处结束）的概率。虚线表示完全匹配。

着法仅出现了 5 次，而最可能的着法出现了 17 次：相比加性标记，网络能更好地将起点和终点保持在一起。

样本是从分布中抽取的。八个样本选择某一着法的频率遵循其概率，依据为q（相关性 0.92），并且 q 可预测单个样本以及八个样本投票的准确率（表4，上）。

最可能的着法胜过一个样本。 隐式策略和两步输出都会读出最可能的着法，而且二者的准确率都高于20步样本（表 4，上）。两步输出也几乎是确定性的：在10,000个测试谜题中，八个两步样本全部一致的比例为76.9%，而20步时仅为27.6%。与单个样本的差距在中等置信度下最大（表 4，下），因为

读出	测量值	预期值
20步样本	57.1	55.6
对8个样本投票	63.8	64.7
两步样本	63.3	–
隐式策略（合法着法）	67.0	71.4

置信度最大值$_m$ q_m	谜题	最可能	样本
低于 0.4	91	39.6	30.4
0.4–0.6	228	48.2	34.9
0.6–0.8	243	63.8	51.3
0.8 及以上	438	84.2	77.5

表 4。 **一个分布，多个读出。** 表 1 中的图像模型在前1,000个测试谜题上的准确率（%）。 上： 实测值，以及如果样本来自隐式策略时的预期值 q。单个样本行取八次噪声抽样的平均值。下：隐式策略的最可能着法与单个 20 步样本的对比，并按策略置信度对谜题进行划分。

方程（8）给出预测。

模型	候选项 位于 σ=1	已提交 在第 1 步	中位数 提交 σ
国际象棋图像，无引导	3.7	47%	0.97
引导式国际象棋图像	2.0	65%	1.00
国际象棋视频	2.8	60%	1.00
路径图像	1.2	90%	1.00
路径视频	1.1	97%	1.00
地形图像	3.0	80%	1.00
地形视频	2.0	93%	1.00
围棋图像，引导	55.1	37%	0.93

表 5。决策会提前确定（验证项，包含一百万个谜题的国际象棋模型）。

决策很早就已确定。 在每个采样步，我们都会解码模型当前对干净输出的估计，并将其视为最终输出。为了描述它保留了多少候选项，我们将每个合法着法的起点蓝方强度与终点红方强度相乘，赋予该着法相应权重，再对所有合法着法进行归一化，并计算有效候选数量（熵的指数）。合法着法仅用于此项计数。决策在某一步之后其领先着法不再变化时，即视为确定。在每个模型和任务中，决策确定时的噪声水平中位数都达到 0.93 或更高（表 5）： 候选着法所在的棋格各不相同，相对于噪声彼此相距甚远，因此几乎每个样本都会立即归入其中一个候选项 [1, 17],，剩余步骤只会让图像更加清晰。不同任务在初始估计所包含的候选项数量上有所不同：寻路任务中的候选项很少，因为目标几乎是唯一的；而围棋任务中的候选项则很多。图 10 展示了表 1中的国际象棋图像模型，以及在一百万个棋局上训练的围棋图像模型的这一变化过程。在 $\sigma = 1$，每次噪声抽样都包含多个被部分标记的着法，经过几步后便会确定为其中一个。每一行展示的是记录到的、最终在其着法上。

采样 步	图像模型 无引导	图像模型 引导	视频 模型
1	38.0	–	**50.1**
2	**38.3**	38.8	48.3
4	34.8	41.5	47.4
8	32.2	**41.9**	44.6
15	–	–	42.5
20	31.4	41.5	–
30	–	–	41.9
40	31.5	–	–

表6。在4,000个测试谜题上，不同采样步数下的准确率（%）（在一百万个谜题上训练的模型）。下划线：默认值；粗体：最佳值。

步数更少，决策更好。 表6 在最初的 4,000 道测试谜题上，比较了针对一百万个谜题训练的模型在不同步数下的表现。不使用引导时，两种模型的步数越少，决策效果越好；减少步数后，每个谜题子集的表现都有提升（按解答深度、评级、合法着法数量和合理替代选项数量划分），其中简单谜题和单解谜题的提升幅度大于困难谜题和多解谜题。步数的分配方式也会产生同样的影响：在步数相同的情况下，以一个较大的步幅跨过高噪声范围的调度方案，优于细致解析该范围的方案（图像模型：36.8% 对 32.4%；视频模型：47.8% 对 42.8%），因为较大的首步会保留平均值，并落在最可能的着法上；而在高噪声下采用细步长，则会任由噪声决定结果。引导是个例外：它会用外推值替代条件均值，因此命题 1 不再适用，减少步数也不会带来帮助。此处，读取最强标记而不设最低强度要求（第 C 节）很重要：单步输出对着法的标记很淡，而要求标记清晰的读者，在视频模型 51.2% 的单步输出中找不到任何着法。

D.4. 超越国际象棋

每项任务所需的步数都更少。 表7 比较了每项任务和每个模型在不同步数下的结果。对于输出为单步决策的任务，一步或两步采样都比完整采样更准确：两个模型在国际象棋、围棋、四子棋和台球任务中均是如此。即使在动作连续的台球任务中，这一结论也成立：多个好击球的平均结果本可能是一次糟糕的击球，但实际并非如此。路线任务的表现则有所不同，而且两个模型之间也存在差异。图像模型每次调用标记一个着法，在均匀地图上的一步采样准确度与完整采样相当（最优路径占比为 89.0%，对比 88.8%）；在地形地图的最低成本路线指标上也如此，但在探测地图上则不然，因为最低成本路线可能会绕开高代价地形。视频模型绘制完整的路线段，在所有路线任务中，少步采样的表现都较差，需要大约八步才能达到完整采样的水平。不同的噪声也可能产生不同的正确解法（图11）。

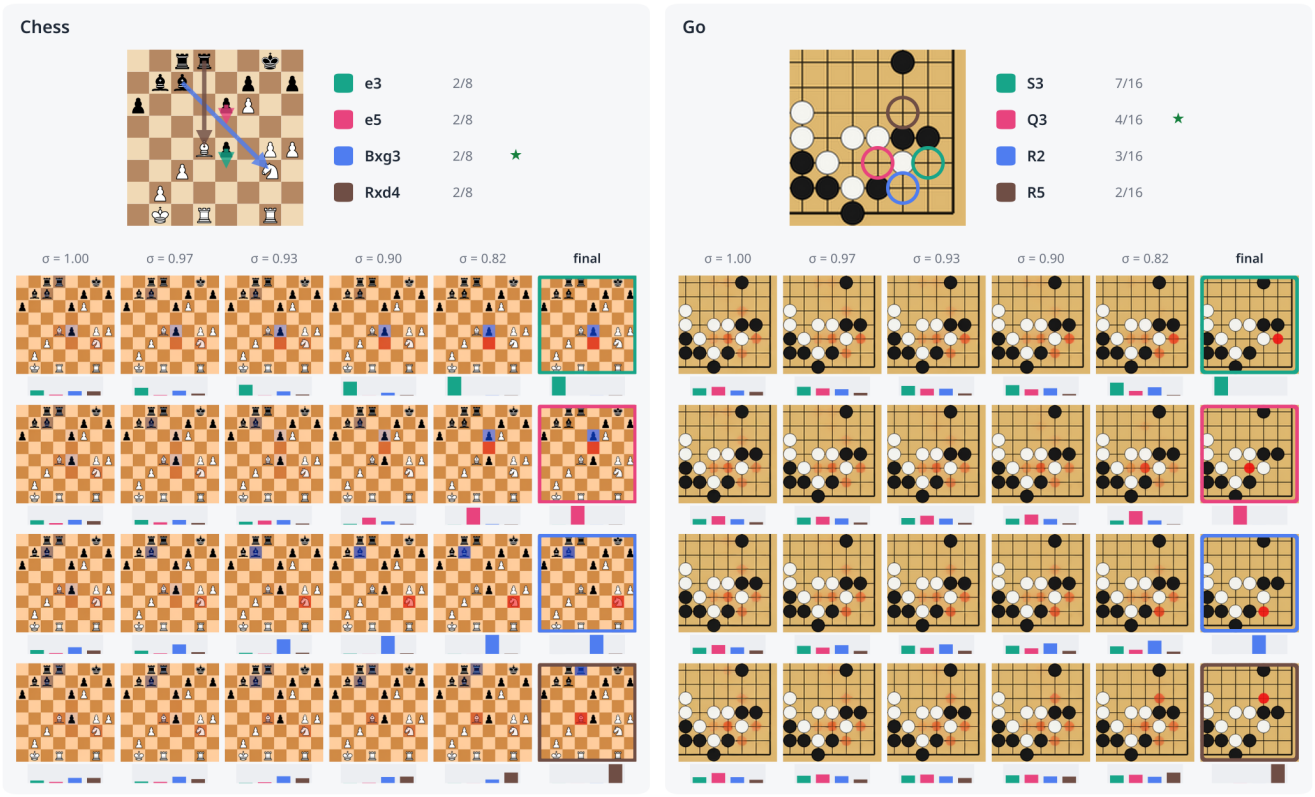

图 10. **决策很早就已确定**。国际象棋（左侧，黑方走棋）和围棋（右侧，白方走棋，放大视图）。每行对应一次噪声抽样，列展示标注噪声水平下解码得到的、未经修改的干净输出估计值$\hat{x}_0$，以及最终的 20 步输出。条形图显示候选标记强度。轮廓颜色标示最终着法，图例比例表示完整样本的频率，星号标记正确着法。

最短路径地图有 168 条由 17 步组成的最优路线。对图像模型进行六次闭环运行，每次调用生成一个着法，每个着法采样 20步，得到三条不同的路线；广度优先搜索证实六条路线均为最优。在困难的四子棋局面中，七列中有六列都能保持蓝方获胜，而前六个 20步样本选择了五个不同的保持价值的列。

为什么路线需要更多步。探测集区分了两种失效模式（表7）。对于图像模型，在需要绕行的地图上出现的每个误差，都是穿越高代价区域的路线。在穿越是正确的情况下，一步的准确率略高于完整采样（90.8% 对 88.0%）；而在绕行是正确的情况下，一步的准确率则低得多（67.2% 对 82.0%）。这正是读取命题 1 的平均值所预测的结果：当模型在完全噪声下仍不确定时，穿越是一条单一直线路线，因此其概率集中在少数几个单元格上；而绕行可以采取许多略有不同的路径，因此其概率会分散在许多单元格上。于是，即使绕行的总体概率更高，平均值中概率最高的单元格仍位于穿越路线上；而完整采样会先确定一条路线，然后沿着它前进（图12）。视频模型没有表现出这种偏差。在步数较少时，其片段会在到达目标之前停止（在一步采样时，250 张绕行地图中有 172 张出现这种情况，而完整采样时只有 8 张），因为片段的平均值并不是连贯的视频片段。在到达目标的路线中，正确选择所占的比例几乎不随步数变化。因此，在主要结果中，我们对所有路线任务都采用完整采样。

第一个估计是模型自身的分布。没有隐式策略，仍然可以用模型自身的样本来检验该理论：根据公式（7），某个动作在单步输出中的标记强度，应与完整采样器选择该动作的程度相当。对于围棋和四子棋，我们在每个局面上各抽取了 16 个完整采样样本，每种游戏共 1,000 个局面（表8）。在所有模型中，标记强度都符合样本频率；单步决策比单个样本更常与多数样本一致，而且其准确率与对全部 16 个样本进行投票的结果大致相当，成本却仅为其十六分之一。在围棋中，还可以将单步输出与外部参考——KataGo 的策略网络——进行比较：在 34.3% 的局面中，KataGo 的首选着法是单步输出中强度最高的着法，而 20 步输出的这一比例为 20.0%；单步输出的强度图与 KataGo 策略的相关性也约为其两倍（0.54 对 0.26）。图13 展示了示例——

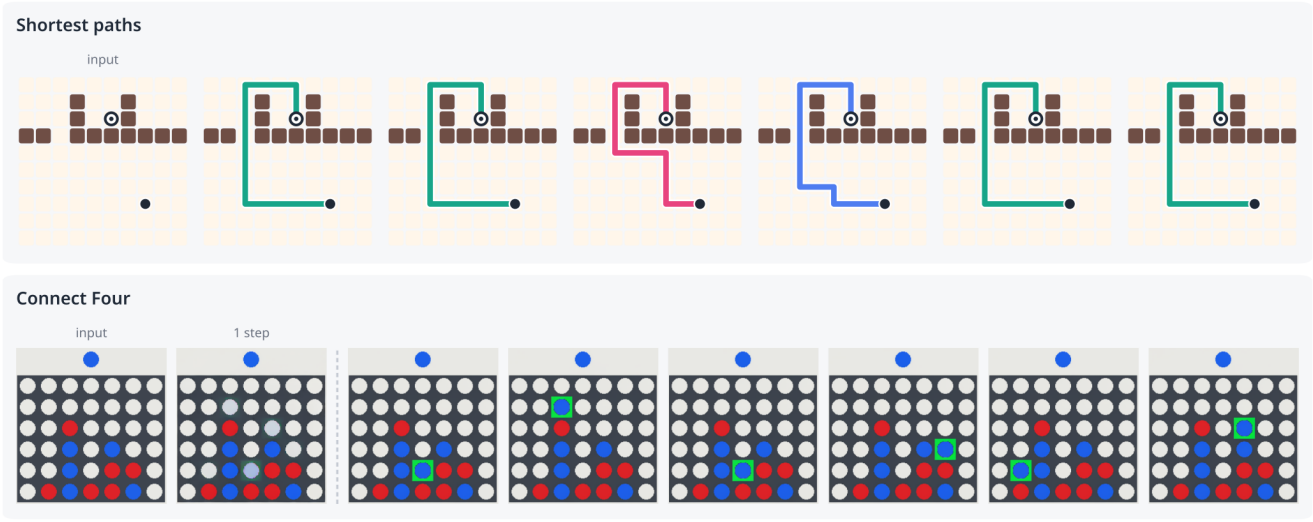

图 11. **不同噪声，不同解法**。上：同一张地图上的六条路线（圆点：起点，圆环：目标）。颜色相同表示路线相同。下：一个棘手的四子棋局面、其淡化的一步输出，以及虚线右侧六个未经修改的 20步样本。绿色方格标示下出的棋子。所有展示的路线都是最优路线，所有展示的四子棋着法都能保持获胜。

Task	模型	1	2	4	8	Full
国际象棋*	图像	**62.1**	62.0	59.9	–	57.3
国际象棋*	视频	**61.8**	61.6	61.1	–	55.6
Go	图像	**31.1**	**31.1**	22.8	21.5	20.9
四子棋	图像	**76.4**	76.3	70.8	68.4	67.5
四子棋	视频	**72.0**	70.2	67.5	65.6	64.5
台球	图像	**28.0**	27.9	24.0	20.2	18.6
台球	视频	**47.9**	43.9	39.6	35.4	34.6
地形路线	图像	87.3	87.2	86.5	**87.7**	87.5
地形路线	视频	62.6	71.6	80.4	**82.8**	**82.8**
最短路径	图像	89.0	89.1	**89.3**	**89.3**	88.8
最短路径	视频	83.0	88.9	94.0	**94.7**	94.6
绕行或穿越	图像	79.0	79.0	80.8	83.4	**85.0**
穿越是正确的	图像	**90.8**	–	90.4	88.4	88.0
绕行是正确的	图像	67.2	–	71.2	78.4	**82.0**
绕行或穿越	视频	30.2	51.6	66.0	74.0	**76.8**
穿越是正确的	视频	41.2	–	76.4	90.0	**92.8**
绕行是正确的	视频	19.2	–	55.6	58.0	**60.8**

表7. 准确率（%）与采样步数的关系（完整采样：图像模型使用 20 个欧拉步，视频模型使用 30 个 UniPC 步）。国际象棋*：表1中的模型，2,000 道验证谜题。围棋：2,000 个测试局面。四子棋：在 1,000 个困难测试局面上，统计价值保持走法的准确率。台球：在 2,000 个测试布局上，统计红球入袋的准确率。地形：在 1,000 张测试地图上，统计最低成本路线的准确率。最短路径：在 1,000 张留出地图上，统计最优路线的准确率。绕行或穿越：500 张探测地图；缩进的行按最低成本路线是穿越高代价区域还是绕行，将地图分开（各 250 张）。

ples.

D.5. 这对推理意味着什么

多数票。来自不同噪声的样本可能提出不同的着法，而投票会选出出现频率最高的

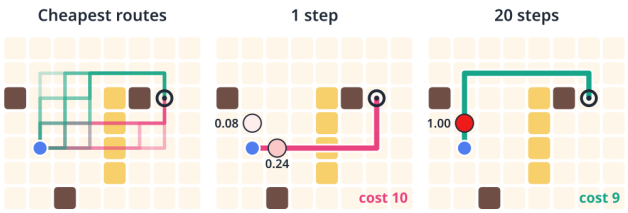

图 12. **为什么路线需要更多步**。一张地形探测地图：绕过高代价区域（黄色）是成本最低的走法。左：绕过该区域的所有最低成本路线（绿色，成本为 9）以及穿过该区域的最低成本路线（粉色，成本为 10）。它们中有许多都共享朝向该区域的第一步棋。中：经过一次采样步，图像模型输出这些路线的平均值，因此首步标记较淡且分散（朝向该区域的强度为 0.24，绕行方向为 0.08）。最强的标记落在共享单元格上，闭环最终穿越了该区域（成本为 10）。右：经过 20步，模型选定一条路线，绕过该区域的着法标记完整（1.00），而这条路线是最低成本路线（成本为 9）。标记强度根据模型的首次输出测量（0 = 无标记，1 = 完整标记）。

一个。因此，样本之间的差异越大，投票的帮助就越大。在一百万个谜题中，对于视频模型、图像模型和引导（引导会使样本更集中）而言，至少有一个样本正确的谜题比例分别为 79.0%、70.0% 和 62.4%；投票则分别带来 10.4、8.2 和 2.7 个百分点的增益（表 9）。少步采样是极端情况：它们读出的平均值都相同，因此彼此差异很小，投票带来的提升也很有限（表1）。

推理时不进行搜索。迭代去噪负责优化时，增加步数会改善解法，并且



图 13。**另外两项任务的首次估计**。上方，围棋：局面、未经修改的一步输出、其中每个点的强度、KataGo 的策略网络，以及模型在 20 步时的 16 个样本落子于各点的频率（地图绘制为按最大值缩放的红色圆盘）。下方，四子棋：局面、一步输出（标记了若干列），以及每列的一步标记强度占比与落子于该列的样本占比。W / D / L：求解器对每列的评估；绿色：保持评估值不变。行：在一步决策正确的局面中，样本分布最均匀的那些局面。

越难的实例需要越多步 [19]。但我们在这两方面都观察到了相反的结果：步数越多，准确率越低；局面越难，从增加步数中获得的增益也越少，而非越多。在计算量相同的情况下，增加样本数比增加步数更有价值：在 10,000 道测试谜题中，一步只需进行一次网络评估，准确率便达到 62.5%；20步达到 56.0%，需要进行 20 次网络评估；对八个20步样本进行投票，准确率则达到 63.8%，需要进行 160 次网络评估。前瞻是在训练中习得的，这与未经过搜索训练的国际象棋模型相同 [18]，，而推理时增加计算量只能从这些结果中进行选择

模型自身的候选项。这也正是剩余提升空间的大部分所在：在 85.4% 的测试集谜题中，正确着法出现在八个 20步样本之中；在 88.1% 的测试集谜题中，正确着法属于隐式策略最可能的三个着法，而任何不使用答案的读出方法，其比例都不超过 67.0%。完整棋题也有相同的情况：八个 20步样本中至少有一个能解出 69.0% 的完整棋题（第F节）。

仅在完全噪声下训练。 公式（7）提出了一个可以直接检验的预测。如果一步返回动作的平均值，那么训练损失中的完全噪声部分，也就是对标记图像进行普通回归，应该足以产生决策。

	强度 vs.	与多数意见一致		准确率	
模型	频率	1步	1个样本	1步	vote
国际象棋，图像†	0.95 / 0.93	80.9	76.8	63.0	63.8
国际象棋，视频	–	81.8	70.6	61.0	63.3
围棋，图像	0.85	62.5	46.2	32.1	29.3
四子棋，图像	0.87	80.5	57.0	76.4	76.0
四子棋，视频	0.79	72.5	52.5	72.0	71.3

表8。 **单步输出与模型自身样本的对比**（每项1,000个，在完整步数下有16个样本，且 †8 用于国际象棋图像）。 强度与频率：在所有项目和动作上，单步输出中某个动作的标记强度与选择该动作的样本占比之间的皮尔逊相关系数（国际象棋：起点／终点棋格；视频模型未测量）。国际象棋中的相关系数分别为0.954和0.932，接近第 D.3 节中与隐式策略的相关系数（0.954和0.929）。 与多数票一致： 单步决策或单个完整采样的动作与样本中频率最高的动作相同的项目百分比。 准确率： 单步决策与所有样本的投票结果相比的正确率。

我们在四子棋上检验这一点：使用相同的骨干网络、LoRA、数据、提示和步数，但让每个训练样本都处于 $\sigma = 1$。通过一步读出，该模型在 83.5% 的困难局面中保住了胜局，而扩散训练模型的比例为 76.4%；这符合预期，因为该模型的全部训练都集中在读出所用的那一噪声水平上。但它无法提供备选方案。对于 98.4% 的局面，该模型针对八次噪声抽样给出的结果一致，因此只有 83.9% 的局面包含价值保持走法；相比之下，扩散训练模型的八个 20步样本仅有 3.7% 的结果一致，却有 94.4 % 的局面包含价值保持走法。因此，单步决策来自目标函数中的完全噪声项，正如公式（7）所述；而在其他噪声水平上训练，则使同一模型拥有了可供抽取候选项的分布（图 11）。

适用范围。 推导假设模型最优，且标记之和为定值。实际中，标记会模糊，图像解码器也具有非线性，因此两步输出更常出现非法情况（8.6%），这比第 D.1 节所解释的错误配对更频繁。使用引导时，模型不再预测条件均值，减少步数也无济于事。隐式策略有些过度自信（预期值为71.4%，而测量值为67.0%），且我们只在一个模型上测量了校准。在所有模型中，样本意见不一致时，单步输出相较于单个样本的增益更大，但差异并不显著。

E. 回忆或对弈：详情

检索基线。 为了与记忆式方法进行比较，我们针对每个测试局面，在被比较的模型所训练过的所有训练棋谱局面中，找出与其棋格差异最少且行棋方相同的训练局面：Tab. 1使用三百万个局面，Tab. 9使用一百万个局面。最近的训练局面 会在测试棋盘上走出该局面的着法，而 *16* 个最近邻的多数票 则会走出这 16 个最近局面中最常见的着法。完全重复的情况十分罕见：10,000 个测试局面中，只有 17 个出现在三百万条训练棋谱中，且都作为解题棋谱中的后续局面出现。没有一个是训练谜题的起点，因此都不是任一模型的训练输入。最近局面的距离也展现了这两种策略性能下降方式的差异。当存在近似重复局面时（差异棋格不超过 4 个，共 763 道测试谜题），复制的正确率为 52.9%，16 个最近邻的多数票为 55.4%，而我们的图像模型为 61.9%。当差异棋格为 13 至 16 个时（共 1,757 道谜题），复制的正确率降至 8.7%，但图像模型仍能解出 50.4%。

对称变换。 将棋盘左右镜像，或交换颜色并将棋盘上下翻转，都会把一个谜题变换为等价谜题，其答案就是原答案经过相同变换后的结果。我们对 2,000 个测试谜题分别应用每种变换，并比较变换前后的准确率。对于所有模型，颜色互换都不会改变准确率（差异均在 1.2 个百分点以内）。镜像会使准确率降低 0.5 至 4.0 个百分点，因此模型对这种对称性的学习只是近似的。

单个棋子变动。 最有力的测试是改变局面，使最佳着法可能发生变化。对于每个测试谜题，我们移除一个非王棋子，或将其移至相邻的空棋格，同时不触及原答案涉及的棋格。随后由搜索深度18的 Stockfish 19 判断这种变化是 决定性的（即此时另一着法明显成为最佳着法），还是 对照条件（即原着法仍然明显最佳）。依赖模式回忆的模型在这两种情况下都应保持原着法；而评估新局面的模型应在对照棋盘上保持原着法，在决定性棋盘上放弃原着法。我们的图像模型表现出后一种行为（表 1）：它在 82.3% 的对照棋盘上保持原着法，但在决定性棋盘上仅有 35.3%；在决定性棋盘上，它有 39.5% 的概率找到新的最佳着法，而随机合法着法的这一比例为 5.1%。检索几乎没有反应：复制方法和 16 个最近位置的多数票在对照棋盘与决定性棋盘之间的差异分别为 4.1 和 8.4 个百分点；我们的图像模型为 47.0 个百分点，少步采样则为 58.8 个百分点。一百万个谜题规模下的视频模型表现相同（41.2 个百分点，表 9）。

路线中途改变地图。我们还在路线中途改变了一次地图。在前 250 张地形测试地图上，每个模型都从起点出发。走完三分之一的路线后，且距离目标至少还有两步时，我们要么在它接下来将进入的单元格上放置一堵墙，要么将泥地或沙地改成地板，从而出现一条严格更低成本、且旧路线不会经过的新路线。在所有最低成本路线之外进行同等幅度的改动，作为对照实验；随后在变更后的地图上继续该回合，并保持相同的采样噪声。图14展示了每种改动对应的一张地图。当墙出现时

	留出谜题的准确率（%）对称棋盘（0 =无变化）					一个棋子发生变化（%）		
	全部 ↑	弃子 ↑	远离训练 ↑	颜色 flip	镜像	保持走法：最佳着法不变 ↑	保持走法：最佳着法改变 ↓	找到新的最佳着法 ↑
我们的方法，一百万道训练谜题								
图像模型	31.8	30.9	26.3	**0.0**	−4.0	62.5	**42.7**	20.4
+ 多数投票	40.0	43.3	33.7	+0.2	−2.4	73.4	47.9	23.2
+ 少步	38.4	38.7	31.8	+0.7	−2.4	89.8	52.3	24.0
图像模型 +引导	40.9	42.2	34.5	+0.5	−1.6	82.1	49.8	29.5
+ 多数投票	43.6	**47.4**	37.4	+1.1	−1.8	87.0	53.4	29.9
视频模型	41.7	35.9	35.4	+0.2	−2.9	91.7	50.5	25.6
+ 多数投票	**52.1**	46.8	46.1	+0.4	−3.6	92.0	50.8	29.0
+ 一步	50.5	40.6	45.0	+0.7	**−1.0**	**94.5**	48.1	31.4
+ 一步 +多数投票	52.0	42.8	**46.4**	+0.4	−2.1	94.2	47.8	**32.5**
从同一批一百万个谜题中复制								
最近的训练局面	15.3	13.8	6.9	+0.4	−5.2	33.1	30.4	6.0
16 个最近邻的多数票	23.6	22.1	15.4	+0.2	−7.0	53.8	49.7	10.3
直接预测，在相同的一百万个谜题上训练								
Qwen3-VL-8B（棋盘图像）	42.0	43.1	34.1	−0.5	−5.5	91.8	61.1	21.6
Qwen3-8B（FEN文本）	40.9	41.4	32.7	+0.4	−7.1	90.5	65.6	18.2
随机合法着法	5.5	3.8	4.0	–	–	–	–	–

表9。一百万个谜题上的回忆或走棋结果，使用表 1 中的列。“引导”表示对比引导（第 M 节），无引导的图像行使用相同的混合训练模型，并采用 $w = 0$。“少步”采用图像模型在验证集上选出的两步。对于所有谜题，95% 置信区间至多为 ±1.0 个百分点；对于弃子谜题，至多为 ±3.1；对于远离训练数据的谜题，至多为 ±2.0；对于对称性差异，至多为 ±2.4；对于单棋子变化，至多为 ±3.8。

在下一个单元格上（左列），图像闭环和视频滚动时域会读取新状态，转向避开墙，并沿着变更后的地图上的一条最低成本路线前进；而执行已经生成的视频片段则会径直撞上墙。当捷径开启时（中列），闭环会选择这条捷径。在这个例子中，滚动时域沿着一条成本较高的路线到达目标。在路线之外出现的墙（右列）不会改变任何路线。在所有地图上（表10），墙出现后，两个闭环在 98% 的回合中都能到达目标，从未进入墙所在的单元格，并且最终沿最低成本路线行进的频率与地图未改变时大致相同。捷径开启时，它们分别在 82% 和 80% 的回合中选择捷径，而旧路线的这一比例分别为 10% 和 19%。执行完整视频片段则无法适应变化：每个回合中，生成的着法都会走进新出现的墙；而在捷径开启的回合中，只有 11% 最终沿着新的最低成本路线行进，因为模型要到下一个视频片段开始时才能看到变化。

路线中的变化	图像，闭环	视频，推荐视野	视频，clip
无（所有地图）	92.1	89.1	87.8
下一个单元格有墙	91.9	83.6	0.0
地图相同，没有变化	*91.2*	*89.6*	*91.0*
捷径开启	80.6	72.9	10.6
地图相同，没有变化	*77.6*	*71.4*	*74.2*
对照：路线外的墙	94.1	91.8	91.0
地图相同，没有变化	*91.2*	*89.6*	*91.0*
对照条件：地板偏离路线	72.7	69.6	75.4
地图相同，没有变化	*77.3*	*72.5*	*75.4*

表10。**在路线中途变更地图**，在前250张地形测试地图上的结果。各项数值表示：在到达变更点的回合中，从变更点出发的剩余路线在变更后的地图上属于最低成本路线的回合百分比。斜体行显示未变更时相同模型和地图的结果。由于该指标从变更点开始计算，因此第一行的数值高于表2中的全路线准确率。2。“片段”会完整执行每个绘制的片段。

一百万个谜题、引导和视频模型。表9重复了表1 在一百万道训练谜题时的结果，此时图像模型（有和没有对比式引导，见第M节）和视频模型均可用，同时还可从同一百万个谜题中进行检索。主表中的所有结论在这一规模下对所有模型变体均成立

蚂蚁。这些模型的准确率远高于复制，也远非训练所得（26.3% 至46.4%，而复制和训练分别为6.9%和15.4%）；它们不受颜色互换影响，并且在对照棋盘上保持其着法的频率远高于在决定性棋盘上保持的频率，

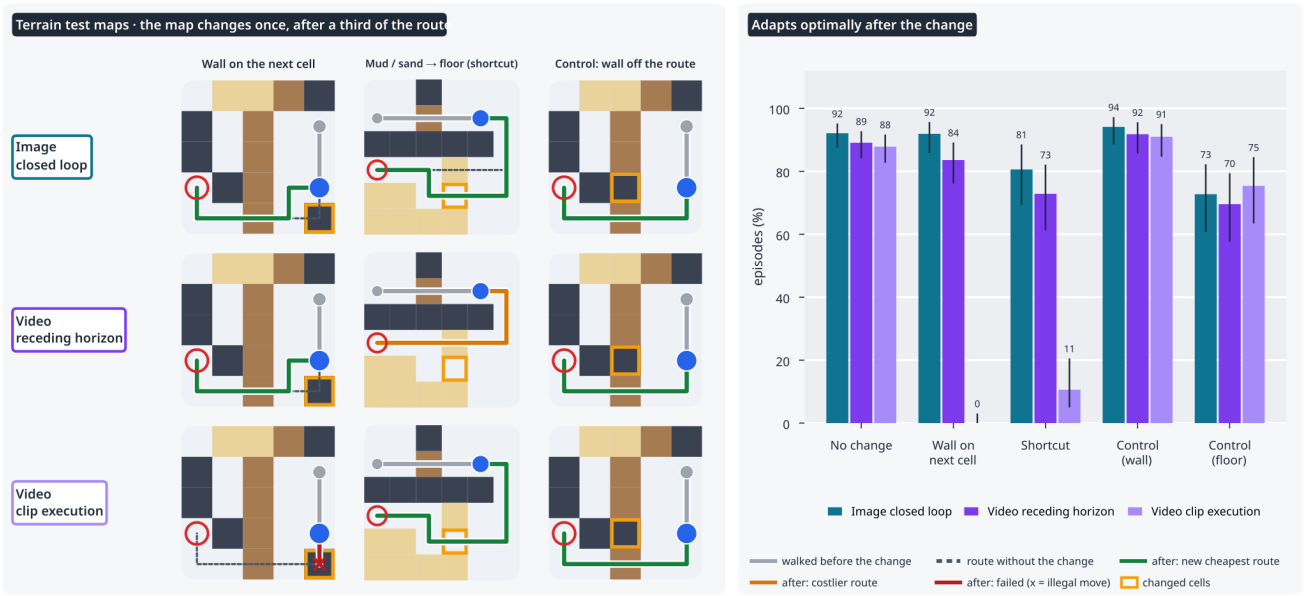

图14。**在路线中途更改地图**。路线行进三分之一后，对金色格子进行一次更改：下一个单元格出现一堵墙，原本的泥地或沙地中开启一条更便宜的捷径，或者在远离路线的位置出现一堵墙，作为对照条件。左：每种更改和每个模型的样例。每次着法后重新进行预测的模型会避开新出现的墙，而在更改前绘制的视频片段则会直接走进去。右：更改地图后的路线中，最便宜路线所占的比例（95%区间）。

而检索几乎无法区分二者。在数据量相同的情况下，视频模型是我们单个样本预测器中最强的模型（41.7%；相比之下，仅使用最佳着法提示训练的图像模型为39.9%，无引导和有引导的混合训练图像模型分别为31.8%和40.9%，见第M节），并且从投票中获益最多（52.1%），因为其样本更加多样（见第D.5节）。比较两个规模可以看出两种策略利用数据的方式有多么不同：将训练集扩大至三倍，使复制率从15.3%提高到18.6%（16票多数票从23.6%提高到27.0%），但仅使用最佳着法提示训练的图像模型则从39.9%提高到56.0%。

F. 完整谜题

Lichess 只有在解题者的每一步着法都正确时，才会将一道谜题计为已解决。我们按照这一规则评估全部 10,000 道测试谜题。第一次预测使用棋题局面，之后的预测则使用每一步记录的对手应答之后的真实棋盘，无论此前的预测结果如何。这样会新增 15,679 个后续输入局面。两个模型都没有接受从这些局面进行预测的训练：图像模型只在第一步局面上训练，而视频模型则从各自的第一步局面绘制整条棋谱。

一步对完整棋题和第一步棋的单次读出都最准确：求解率分别为 38.3% 和 40.5%，而完整采样时分别为 31.1% 和 31.9%（表 11）。尽管图像模型仅针对初始局面进行训练，但在后续着法和第一步棋上的准确度也相近

模型	读出	第一步棋	后续着法	完整棋题
图像，一步	单个	62.5	62.8	38.3
	8票投票 8个中的任意一个	63.2	63.3	38.8
		69.2	68.8	45.2
图像，20步	单个	56.0	57.0	31.1
	8票投票 8个中的任意一个	63.8	64.0	39.2
		85.4	**85.1**	**69.0**
视频，一步	单个	63.9	66.0	40.5
	4票投票 4个中的任意一个	64.4	66.7	41.0
		70.2	72.2	47.9
视频，30步	单个	56.4	57.6	31.9
	4票投票 4个中的任意一个	62.8	63.8	38.5
		79.4	80.9	60.2

表 11. 表1 中各模型在 10,000 道测试谜题上的整题结果（%）。后续着法根据记录的对手应着之后的真实棋盘进行预测。“4个中的任意一个 n”表示至少有一个 n 样本正确。

（62.8% 对比 62.5%）。完整棋题的成功率高于根据独立着法预测的比率（38.3% 对比 33.1%），这表明谜题的难度往往会贯穿整个棋题。评级低于 1000 时，解题率为 89.8%；评级高于 2500 时，则为 3.8%。完整采样能覆盖更广的候选项：在 20 步时，8 个样本中至少有一个解出谜题的比例为 69.0%；相比之下，8 个一步采样的比例为 45.2%，而两者读出的平均值几乎相同。

按记录的解法中我们着法的数量分组，单步视频模型分别解决了有一步、两步和三步的谜题中的92.2%、49.4%和25.8%

以及有四步或更多着法的谜题中的12.3%（图像数据分别为90.2%、47.0%、23.6%和11.4%）。尽管视频的首步准确率几乎没有变化（57.4%对57.1%），但从三步到四步或更多着法时，整题成功率进一步下降。

错误着法是替代解法吗? 记录的解法之外的着法仍可能同样出色。我们使用 Stockfish 19，在每个局面固定搜索 1600 万个节点，将每个此类着法与解法进行比较。如果解法存在强制将杀，则该着法只要保留强制将杀即可；否则，胜率最多降低 5 个百分点，即达到 Lichess 的不准确阈值，该着法便符合条件。替代着法很少见：在第一个局面中，仅有 0.2% 的其他合法着法符合条件；在后续局面中，这一比例为 0.7%。将同样出色的模型着法计为正确，最多只能使每着准确率提高 0.4 个百分点，并使整题谜题的求解率从图像模型的 38.3% 提高到 38.8%，从视频模型的 40.5% 提高到 40.9%。

G. 这些谜题需要多少前瞻?

所需深度	1	2–3	4–5	6–9	10–30
谜题	7,748	1,246	537	317	145
视频，一步	69.2	47.5	44.1	41.0	44.8
图像，一步	67.8	48.0	41.5	37.9	39.3
单层规则	42.3	23.4	23.6	18.9	24.1
16 个最近邻的多数票	29.2	19.4	20.3	18.9	15.9
随机合法着法	5.6	4.9	4.8	4.8	5.1

表12。表1 中各模型的首步准确率（%），按引擎的最佳着法成为谜题答案时所对应的最浅 Stockfish 搜索深度排列。单层规则优先选择将死，否则选择最佳吃子，再否则选择将军。在七道谜题中，引擎在不超过30的任何搜索深度下都没有走出答案，其中六道是一步将死且存在另一个同样好的将死着法；这些谜题被排除在外。

不同所需搜索深度下的准确率。评级和解答长度只能粗略反映谜题所需的前瞻深度。因此，我们将每道测试谜题标记为 Stockfish 19 的最佳着法与谜题答案相同的最浅搜索深度（单线程、固定哈希，搜索深度不超过 30），并按此深度统计首步准确率（表 12）。引擎在 7,748 道谜题中已于深度 1 找到答案。不过，深度 1 并不等于一步规则：引擎仍会持续吃子，直到局面平静下来，并用一个基于更深层搜索训练的网络评估结果；而只会走将死、最佳吃子或将军着法的规则，在这些谜题中的答案命中率仅为 42.3%，单步视频模型则为 69.2%。对于需要深度 6 或更深搜索的 462 道谜题——人类解题者也觉得这些谜题难得多（平均评级为 2232，对比 1649）——视频模型的准确率为 42.2%，图像模型为 38.3%；相比之下，单层规则为 20.6%，从 16 个最近训练局面中选出的多数着法为 18.0%，随机合法着法为 4.9%。当答案不再是引擎的即时首选时，准确率会下降，

但随着所需深度从 2 增至 30，准确率并未持续下滑，而且很大一部分降幅可归因于谜题评级：在同一评级分区间内，视频模型解出这些深层谜题的比例比其他谜题低 2.4 个百分点，图像模型则低 8.3 个百分点。标记为弃子的 139 道深层谜题表现相近（视频模型为 42.4%，任何基线最高不超过 22.3%）。在四子棋中，精确求解器会给出必胜所需的着法数，呈现出相同的模式：对于在我方走 1–3、4–7、8–11 步或 12 步及以上即可获胜的困难局面，单步图像模型分别能在 81.5%、84.9%、79.5% 和 72.5% 的局面中保持获胜；在每个区间都比一步启发式方法高出 14 至 21 个百分点。因此，对于只有通过深度搜索才能由引擎解出的局面，这些模型仍远胜于浅层规则，尽管它们并不进行搜索。

单层规则与弃子。在表1所示的三个测试集上，单层规则在所有谜题中找到38.0%的答案，在弃子谜题中找到26.8%，在远离训练集的谜题中找到33.6%；相比之下，单步视频模型的对应比例为63.9%、55.6%和59.1%。弃子标签描述的是整个解法，因此我们还检查着法本身。根据静态交换评估，在1,001道弃子谜题中的857道里，我方在解法的某处放弃了子力，其中699道在第一步棋本身就放弃了子力。在这699道中，视频模型和图像模型找到该着法的比例分别为52.6%和48.2%，而单层规则和复制基线最高仅为26.6%。

谜题结束于	谜题	图像	视频
一步将杀	853	90.4	92.4
两步将杀	944	85.3	87.2
三步将杀	313	69.6	76.7
四步或更多步将杀	90	64.4	64.4
无将杀	7,800	56.4	57.5

表13。模型在谜题结束方式下的首步准确率（%）（表 1 ）。

按谜题结束方式划分的准确率。 构成强制将杀的第一步棋比其他着法更常被找到，而且随着将杀所需着法变长，找到的频率会降低（表 13）。即使对于四步或更多步将杀的谜题，两个模型也能在64.4%的谜题中找到第一步棋，高于它们在无将杀谜题上的准确率。

H. 注意力与遮挡

图 5 展示了国际象棋决策所依赖的棋子。我们在一百万个谜题的模型上，针对大量棋局进行测试，使用两类谜题：可以赢得一个敌方棋子，或可以攻击敌方国王。对于视频模型，从决策帧指向可赢得棋子的注意力是机会的 1.13 倍，而指向其他敌方棋子的是 0.86 倍；在控制视觉显著性（哪些棋格上有棋子，以及哪些棋格被该着法使用）后，这一差异仍然存在。在将死谜题中，敌方国王的王区获得的注意力是机会的 1.00 倍，而在另一类谜题中为 0.89 倍；相比之下，预训练模型给出的数值为 1.10

在两类谜题中都是如此：微调会使注意力取决于谜题的内容。移除棋子进一步从因果上证实了这一点：移除可赢得的棋子，会使所选着法的强度降低 0.61；移除另一枚敌方棋子，则仅降低 0.02。对于图像模型，移除棋子显示出相同的依赖关系（0.67 对 0.05），但其注意力并非如此。可赢得的棋子获得的注意力高于其他棋子（分别为机会的 1.15 倍和 1.05 倍），但预训练模型也呈现出相同的差异，而且对王区的注意力在两类谜题之间没有区别。因此，图像模型的决策依赖于正确的棋子，但其注意力图并未显示出这一点。

寻路使测试更加清晰，因为相关性很容易界定。对于两种视频模型，从决策到所选路线各格的注意力都高于机会水平，也高于匹配的路线外单元格（比率为 1.10 至 1.14），且这一效应源于微调。对于图像模型，对路线的注意力处于机会水平。在地形地图上，三种模型对路线所避开的高代价单元格的关注都多于对匹配地板单元格的关注，这表明它们会查看那些影响路线的障碍物。第 E节的反事实测试揭示了因果层面：在所选单元格上设置一堵墙，或移除路线所避开的地形，都会改变决策；而在其他位置进行相同改变则不会。

I. 前瞻探测

隐式策略表明，图像模型能够区分好着法和坏着法。由于着法的优劣往往取决于后续发展，我们想知道，尽管训练要求模型只绘制自己的着法，它是否仍在内部表征未来着法。

特征。 我们在一百万个谜题上运行无引导图像模型，流程与评估时完全相同，并记录第 0、4 和 10 个采样步（σ = 1.00、0.86 和 0.60）时第 8、16、24 和 32 个 Transformer块（共 32 个）的输出。每个棋盘方格覆盖 2×2 个图像词元，我们将其平均，得到每个棋格对应的一个 4,096 维向量。

目标。 谜题的解法给出了三组棋格：模型所画出的我方第一步棋、对手的应答，以及我方的后续着法。对于应答，我们排除第一步棋所在的棋格；对于后续着法，则排除前两步棋所在的棋格。否则，探针可能仅凭找到棋盘上标出的记号就能成功，例如，应答在我方着法的终点方格上吃回棋子时。

探针与选择。 探针是一个由所有棋格共享的单个逻辑回归层，它将一个方格的标准化向量映射为该方格属于目标走法的概率。对于每个局面，它会对候选棋格进行排序，我们用其平均精度为该排序评分。探针在模型从未见过的训练行中的 20,000 个谜题上训练。区块、噪声水平和正则化在 1,834 个验证集谜题上选定，并且仅对选定的探针进行一次评估，使用 3,679 个谜题。

测试谜题。

基线。 探针也可能因为应答很容易猜中而成功。为排除这种可能，我们在另外四种输入上训练同一个探针。第一种描述棋盘：对于每个棋格，描述其周围一个 15×15 窗口内每个棋格上的棋子，以及行棋方。第二种在此描述中加入真实的第一步棋，让探针知道模型会走什么。第三种使用预训练模型的特征，第四种则将我们模型的特征与另一个局面的标签配对。

目标	机会	棋盘	+ 着法	预训练	Ours	Gain
应答	2.7	32.1	40.9	39.2	**47.8**	+7.0
后续	2.5	38.0	41.4	39.0	**47.1**	+5.7

表14。未来着法探针的测试集平均精度（%）。“增益”是相对于最佳基线测量的。95%区间为 [+6.0, +7.9]，对应应答；以及 [+4.7, +6.7]，对应后续。

结果。 微调模型的特征对未来两步着法的预测效果都优于任何基线，包括包含真实第一步棋的棋盘描述（表 14）。预训练特征的表现与棋盘描述大致相当，表明微调带来了预测信息。两步着法在网络中段的第16个模块处最容易读出，并且在第一个采样步时就已可读出。对于应答，在$\sigma = 1$处的平均精度为 47.8，在 $\sigma = 0.6$处为 46.9。

J. 视频模型想象中的未来

第I节中的探针检测图像模型中关于未来着法的信息。视频模型则将这些未来显式呈现出来：每次决策后，它都会绘制由此产生的棋盘以及对手的应答。我们评估这些想象状态和着法的准确率，以及它们与初始决策之间的关系。生成后续内容是否能改善该决策，将在第K节中探讨。

	第1回合 (我们的方法)	第2回合 (应答)	第3回合 (我们的方法)	第4回合 (应答)
正确	41.7	20.1	12.2	2.0
在前缀正确的情况下正确	–	40.9	39.5	17.1
想象棋盘完全准确	99.1	98.0	96.3	93.6

表15。视频模型在10,000道测试谜题上的想象线路准确率（%）。第2至第4半回合的评分对象，是解法恰好包含相应半回合数的谜题（第2和第3半回合为9,141道，第4半回合为4,594道）。只有当原始绘制在想象棋盘上合法，且与谜题走法相符时，该着法才算正确，因此提前结束的视频片段计为错误。想象棋盘在合法绘制的着法后进行评分。

想象线路有多好？ 在10,000道测试谜题上，

在一百万个谜题上训练的视频模型，有 41.7% 的概率能正确预测第一步棋（表 15）。后续着法是从模型自身想象的棋盘中解码得出的，因此误差会不断累积。在第一步棋正确的情况下，想象中的应答与谜题中的着法相符的比例为 40.9%。在连续两步棋都正确的情况下，模型的后续着法与谜题相符的比例为 39.5%。即使应答与谜题不同，也可能是合理的：在第一步棋正确后，51.4% 的应答是 Stockfish 给出的最佳应答，63.6% 的应答与其获胜概率相差不超过两点。想象棋盘也相当准确：在走出合法的和棋着法后，根据半回合不同，所有 64 个棋格均正确的比例为 93.6% 至 99.1%。

视频片段按顺序逐步显现。视频模型会联合去除每一帧中的噪声，不受预设时间顺序的约束，但较后的棋盘往往也较晚才稳定下来。与第一步棋后的棋盘相比，第二、第三和第四步棋后的棋盘在片段中较晚稳定下来的比例分别为 54%、74% 和 90%。这种模式源于帧与帧之间的依赖关系：较后的棋盘仍然不确定，直到较早的着法确定下来。事实上，当两个第一步棋在高噪声下相互竞争时，第一个想象棋盘会融合这两种着法所产生的局面；随着其中一种着法逐渐消退，与之对应的棋盘也会消退。随后，模型会确定下一个棋盘，再依次确定更后面的棋盘。这与第 D.3 节中较早作出决策承诺的现象相辅相成，表明视频片段是作为一条相互依赖的对弈线路生成的，而非由互不相关的帧组成。这只是一种倾向，并非严格的顺序：有些较后的棋盘会与第一步棋同时稳定，或比它更早稳定。

一厢情愿。走出错误的第一步棋后，模型往往会想象出一种应答，让自己的着法看起来比实际更好。我们在真实的着法后局面上评估合法的想象应答，并使用 Stockfish 19、深度12，对所有对手应着进行排名。第一步棋正确时，想象中的应答有 51.4% 是最佳应对；而第一步棋错误时，这一比例仅为 28.6%，此时最佳应对通常是反驳着法（按局面的强制程度分层的 Cochran–Mantel–Haenszel 检验，$p < 10^{-100}$）。与最佳应对相比，走出错误着法后，想象中的应答使行棋方的胜率高出 35.9 个百分点；而着法正确时则高出 3.8 个百分点。因此，错误着法往往伴随着一段想象中的后续棋谱，使该错误看起来似乎奏效，而不是对着法进行独立检验。

K. 预测时域

我们考察推演后续内容是否能改善第一步棋、带来增益的原因、模型在推理时是否会参考其想象的未来，以及推演出的着法应在何时执行。这扩展了第 4 节中“从后续变化中学习”的分析。

使用延续内容进行训练。我们比较仅包含第一个决策和结果状态（仅首步）的目标，以及包含最长八步参考后续序列（包括对手应着）的目标（式（2））。数据、训练步数、学习率和随机种子均相同，

Task	模型	第一步棋 only	With 后续内容	p
国际象棋，8 万条棋谱	视频	8.7	11.0	$6 \cdot 10^{-13}$
国际象棋，40万棋谱	视频	21.4	24.6	$5 \cdot 10^{-14}$
国际象棋，100万棋谱	视频	33.7	41.7	$3 \cdot 10^{-64}$
地形路径	视频	82.5	95.0	$3 \cdot 10^{-24}$
地形路径	图像	83.1	97.3	$5 \cdot 10^{-31}$
困难四子棋	视频	65.4	64.5	0.53

表16. 使用相同的数据和训练预算进行训练（有或没有后续内容）后的首步准确率（%）。国际象棋：10,000 个测试谜题。地形：1,000 张测试地图，按代价最优的第一步棋评分。四子棋：在 1,000 个困难局面中，保持价值不变的第一步棋。p 数值为精确的 McNemar 检验结果。

在国际象棋中，两种目标都会绘出绿色合法落点圆点（第B节）。评估时仅对第一个决策评分。对于图像模型，延续目标是一张完整路径图像（$H = L$）。我们执行第一步棋，渲染结果状态，然后再次调用模型，并将其与使用相同 8k 步数训练的单步模型进行比较。在每次训练访问中，单步模型都会绘制路线上的一个局面，并据此预测下一步，因此它看到的路线决策与完整路径模型一次绘制的相同。延续训练提升了国际象棋中各种数据规模下的首步准确率，也提升了两个模型在地形寻路上的首步准确率（表16）。即使训练量达到三倍（24k 步），单步地形图像模型的准确率也仅为 94.5%，低于完整路径模型在 8k 步时的 97.3%（$p = 4 \cdot 10^{-4}$）。四子棋是个例外：在困难测试集上，两种模型表现无法区分；而在标准集上，仅首步模型表现更好（69.1% 对 65.8%，$p = 2 \cdot 10^{-4}$）。由于其训练棋谱来自精确求解器，含噪延续无法解释这一性能上限。

对于在一百万个谜题上训练、并以 30 个采样步进行评估的视频模型，延续训练使记录解法中包含我们一着、两着、三着以及四着或更多着法的谜题，其首步准确率分别提高了 14.7、8.3、6.5 和 6.3 个百分点。这一增益覆盖全部四组，其中一着谜题的增益最大。

将相同决策作为独立样例。在后续内容上训练的模型，每个样例也能看到更多带标签的决策，因此我们将它与对照实验进行比较：这些对照实验接收相同的决策，但将其作为独立的首步样例。对于地形图像模型，单步模型本身就是这样的对照，因为它会在路线上的每个局面进行训练：在相同计算量下，其准确率为 83.1%，而对照为 97.3%；将训练量增加到三倍后，准确率为 94.5%。对于地形视频模型，我们将仅预测首步的模型在相同片段中的每个路线决策上进行训练，并将每个决策作为独立样例，训练步数同为 5k。其准确率达到 85.3%，高于仅首步的 82.5%；使用后续内容时则达到 95.0%（$p = 4 \cdot 10^{-18}$ 对比后续内容模型），弥补了约五分之一的差距。在国际象棋中，使用 80,000 个棋谱时，

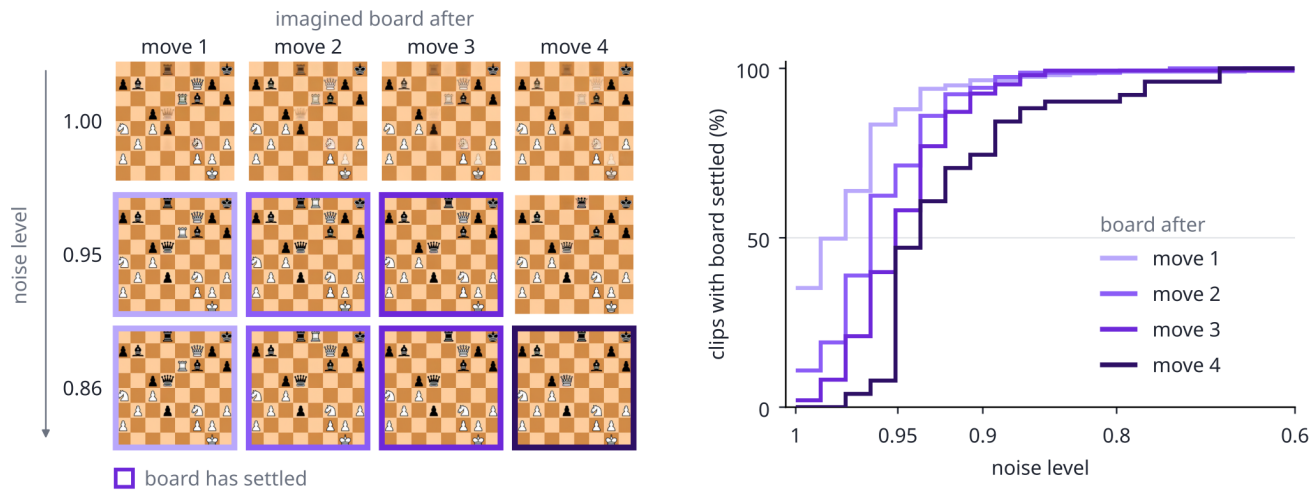

图15。**视频片段按顺序逐步确定。** 左图： 视频模型在三个噪声水平下生成的一个国际象棋视频片段，展示其对第1至第4着后棋盘的未经修改的估计。帧标记表示棋盘已稳定。右图： 各棋盘已稳定的视频片段占比与噪声水平的关系。与第1着后的棋盘相比，第2、第3和第4着后的棋盘分别在54%、74%和90%的视频片段中较晚稳定。

相同计算量下，同一对照实验的准确率为 6.3%，低于仅首步的 8.7% 和后续内容的 11.0%（$p = 4 \cdot 10^{-54}$）。将其训练量延长至 2.3 倍，使每个决策都恰好出现一次后，准确率达到 12.0%，比后续内容模型高一个百分点。下方的延迟提示分叉在受控环境中得出了相同结论。因此，在相同计算量下，生成后续内容比将相同决策分别提供给模型，能更有效地提供训练信号。在国际象棋中，独立样例需要 2.3 倍的训练量才能达到同等效果；而在地形任务中，即使训练量增加到三倍，效果仍不及后续内容。

续接内容为何有帮助：延迟提示分叉。 我们使用包含两条等长镜像走廊的地图，单独考察未来监督的作用。一堵墙 $k \in \{1, 2, 3, 5, 8, 12\}$ 位于其中一条走廊下方若干格，决定了第一步棋。每张地图都有一张孪生地图，墙位于另一条走廊中，因此忽略墙会得到随机准确率（50%）。四种视频模型在相同的 40,000 张地图、训练步数和随机种子下进行训练。它们的目标仅在正确的第一步棋之后有所不同：不提供后续内容、提供接下来的八步路线步数、提供接下来的三步路线步数，或提供七步偏离路线的续接。我们在每张 k 留出地图上评估第一步棋（图 16）。在这一训练预算下，仅预测首步的模型对一格和两格外的墙作出响应（准确率分别为 99.4% 和 92.8%）；墙距三格时，准确率降至 57.8%，距五格及更远时则处于随机准确率水平。采用路线目标时，墙距八格以内的准确率为 99% 至 100%，距十二格时为 91.6%，超出了绘制的后续内容范围。提供三步路线步数时，模型能处理三格以内的墙；等长的偏离路线续接则能帮助模型处理距三格的墙，但对五格及更远的墙没有帮助。因此，增益来自学会沿路线前进至决定性壁垒，而非仅仅来自视频片段长度。

更多标签还是更多训练？ 两种对照实验将后续内容与其伴随因素区分开来。首先，后续内容对每张地图中的八个决策进行监督，而非一个。我们将仅预测首步的模型用于训练：把同一批八步路线上的每个局面都作为独立示例，训练步数同为 2,500。对于距墙三至八格的情况，该模型的准确率仍处于随机水平（53.4%；仅首步模型为 54.2%，而提供后续内容时为 99.7%）。它的样例中只有八分之一展示了岔路决策，因此在近处的墙面前表现也较弱（距墙两格时为 43.1%）。将训练步数增加到八倍后，当模型见过每个岔路的次数与仅首步模型相同时，准确率达到 94.2%，但面对远处墙壁时仍不及后续内容模型（距墙八格时为 85.3%，而后者为 99.1%；距墙十二格时为 70.0%，而后者为 91.6%）。其次，第一步棋目标能够学会这条规则，但需要更多训练。在相同地图上将训练步数增加到两倍后，模型在距墙三至八格时的准确率达到 84.7%，距墙十二格时为 70.9%，但各个距离上仍低于后续内容模型；训练步数增加到八倍后，它则能解出所有测试岔路（图17）。后续内容的作用范围也受到绘制内容的限制：墙距 10、14 和 16 格，超出绘制的八步路线时，准确率分别降至 96%、87% 和 71%。这正是中间监督所预期的模式。逐步引导模型朝线索前进的目标，使每一帧都能由邻近证据进行监督，因此长程依赖被学成一连串局部依赖，复合规则也就能用少得多的样例学会 [6, 24]。因此，后续内容降低了远处线索所需的训练量，而不是让原本无法学会的内容变得可学。

N 皇后问题：以着法衡量的视野。 延迟线索的叉子战术衡量的是到达决定性壁垒的距离。我们可以用 N 皇后问题测量还需要放置多少个皇后

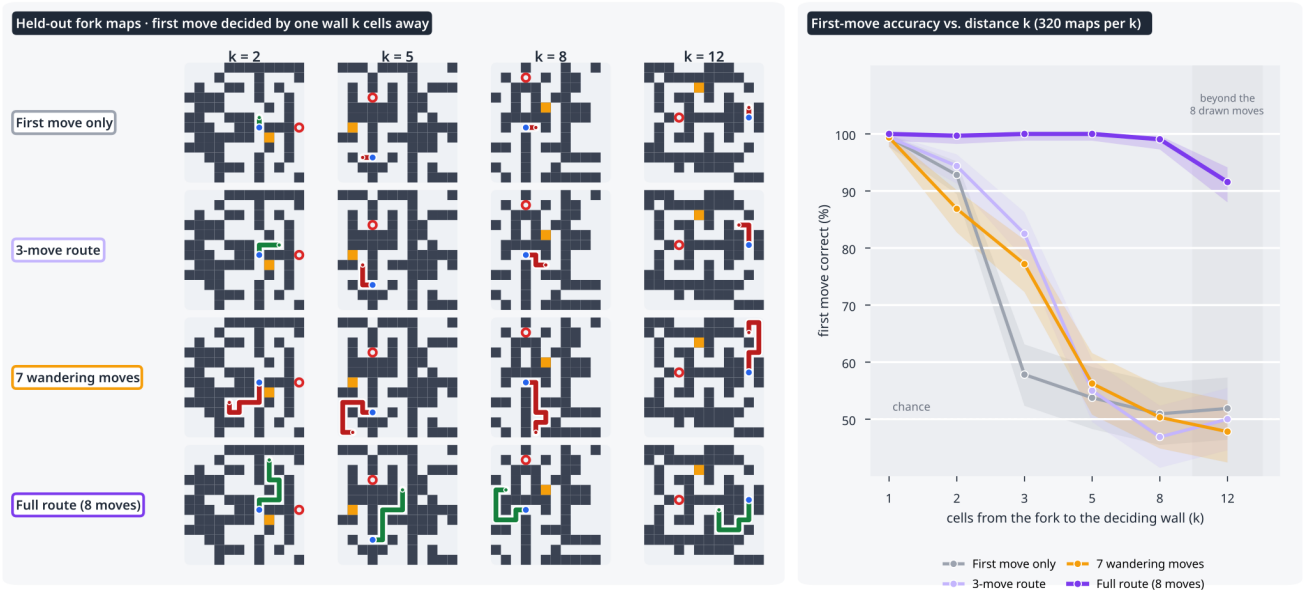

图16。**延迟提示分叉**。 左： 留出地图中，距离目标单元格 k 格的一堵墙（橙色）决定了第一步棋，并展示四个模型各自绘制的路线。右： 四种训练目标在 k 条件下的首步准确率（每个 k包含320张测试地图，95%区间）。

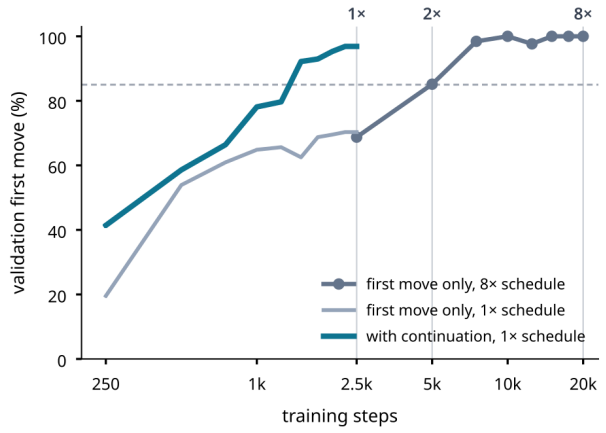

图17。 **学习延迟提示分叉**。 训练期间的验证集首步准确率。竖线标示图16训练预算的1倍、2倍和8倍。采用后续内容训练时，模型在1,500步内通过85%准确率。仅用第一步棋训练时，采用更长的训练计划需要5,000步；而采用原训练计划时，准确率始终未超过70%。

错误的第一步棋才会走入死路。在一个由 3,606 个棋盘构成的镜像平衡测试中，我们让下一行恰好有两个候选格：一个可以导向完整解，另一个则会在进一步放置皇后后走入死路。搜索深度为 0、1 和 2 时，各有 600 对镜像对（1,200 个棋盘）；搜索深度为 3 时仅有三对，因此下文仅作描述性报告。每个棋盘都会以其左右镜像，因此忽略未来的选择得分为50%。采用仅首步训练时，在搜索深度0、1和2下的准确率分别为71.7%、62.3%和48.5%（图18，右下）。搜索深度为3时，它在6个棋盘中的4个选择了可完成任务的单元格，样本太少，无法得出可比的估计。采样20步而非采用验证集选定的步数，并未扩大测量视野（搜索深度0至2下的准确率分别为58.5%、53.6%和46.6%）。

使用后续步骤训练。 在 带有后续内容的条件下，训练目标会画出所有缺失的皇后；而在仅首步条件下，只画出下一枚皇后。在这两种条件下，评估时都只读取第一步棋。在使用相同训练棋盘、训练方案和随机种子的情况下，延续训练将死胡深度为 0、1 和 2 时的首步准确率分别提升至 90.4%、80.2% 和 58.7%（图18，右下）。深度 2 时的增益为 10.2 个百分点（$p < 0.001$）。它还提升了留出解族上的第一步棋表现（92.6% 对 85.2%）以及困难棋盘上的表现（82.0% 对 62.0%）。因此，未来监督有助于图像模型应对超越下一次放置的后果，正如它有助于视频模型在叉子战术任务中应对远处的墙。

使用绘制的解。 我们通过两种方式评估使用后续内容训练的模型：一次绘制完整棋盘，或只执行第一个皇后的落子，然后从所得棋盘重新生成。在300个留出半填充棋盘上，每次调用采用两个采样步的闭环解出了55.7%的棋盘，而最佳单次通过结果（20步）解出了34.7%。在512次空棋盘试验中，闭环解出了38.1%，而最佳单次通过结果（50步，图 18，右上）。在左侧示例中，前四次循环绘图都不是完整解法，但每次都提供了一个可继续完成的下一步

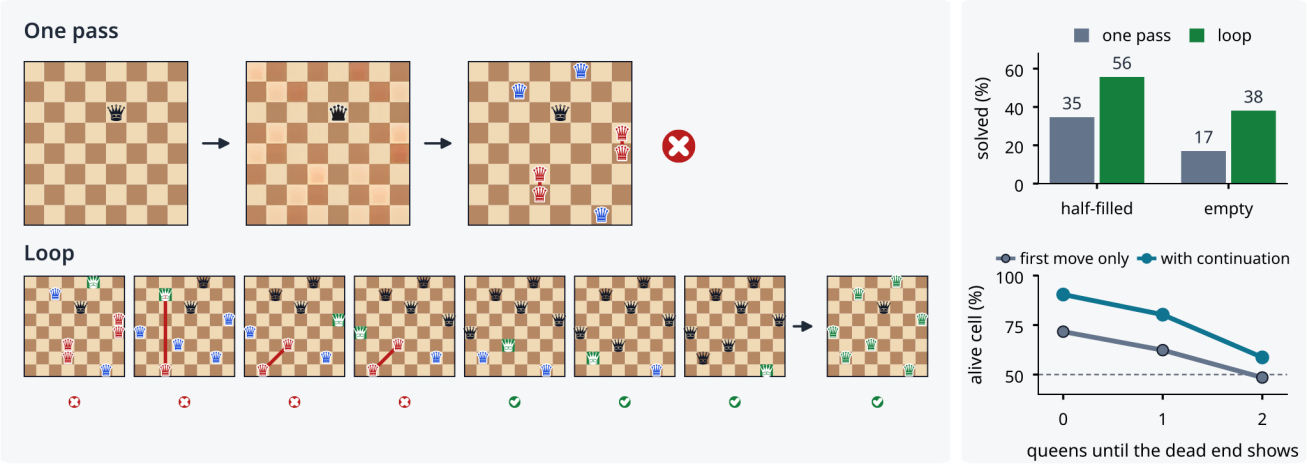

图18。**错误的绘图仍能引导出正确的着法**。左：在一个留出棋盘上，单轮补全未能成功，但先走出皇后的着法再重新绘图则成功了。深色皇后已放置在棋盘上，彩色皇后是绘制出来的，绿色标记表示所走的着法，叉号/勾号表示补全有效性。右：单轮遍历与循环的求解率（上），以及按死路深度划分的首步准确率（下，虚线表示机会水平）。

皇后落子。之后的绘图变得有效，闭环也解出了棋盘。这种比较不仅改变了绘图的用法，也改变了采样计划。它表明，与将最初的整幅绘图直接作为答案相比，只有在模型的第一个决策能够付诸实践并加以修正时，模型生成的后续内容才更有用。

改变棋盘。在另一项干预中，环境会在第一步棋之后添加一个皇后，与模型最初绘制的解法相矛盾。每次调用两步时，在300个半满棋盘中，有145个可能出现破坏计划的变化。循环在这些变化后的回合中求解了55.2%，而在相同但未发生变化的棋盘上，求解率为48.3%。对照条件是在模型原计划放置皇后的位置添加皇后。在两步设置下，半满棋盘上的破坏计划条件优于对照条件（55.2% 对 42.1%，$p = 0.011$），但在空棋盘上则逊于对照条件（43.2% 对 51.9%，$p = 0.0007$）。在20步设置下，对照条件的比较结果为：半满棋盘上46.3% 对 43.4%，空棋盘上43.5% 对 40.8%；两项差异均不显著。与未发生变化的情况相比，在相同的四种设置下，破坏计划条件的数值均不差（55.2% 对 48.3%、43.2% 对 38.5%、46.3% 对 44.1%，以及43.5% 对 40.8%）。这些结果表明，循环能够从意外状态继续进行，且未观察到平均表现受损，但并不表明破坏其计划会有所帮助。

推理时不会读取后续内容。 如果模型会参考自己生成的后续内容来选择第一步棋，那么改变后续内容就应该会改变决策。在视频生成过程中，我们固定对手回应帧，使其显示引擎对模型第一步棋的反驳着法或一个较弱的应答，同时保持这一步棋可见。在这两种条件下，第一步棋仅在1,000 道验证谜题中的 0.8% 会发生变化（精确 McNemar $p = 1$）。结合分叉实验，这表明训练带来了收益：预测未来状态会塑造学到的决策，但绘制出的后续内容并不会被用于作出决策。

与语言模型的一个相似之处。语言模型也会呈现一种相关效应，称为多令牌预测。在每个位置训练模型预测多个未来令牌，而不只是下一个令牌，可以提高样本效率；额外的预测则作为辅助训练任务 [6]。作者认为，增益来自选择点——其结果会影响后续发展的那些位置；同时预测未来会让这些位置获得更强的训练信号。下一步最佳着法就是这样一个点。我们的结果显示，同一原理也适用于视觉生成器：它预测的未来由绘制状态和对手应着构成，而非令牌；此外，这也让我们能够进行两项进一步测试：在不改变决策的情况下，可以在推理时改变绘制的未来；而在相同计算量下，将相同决策作为独立示例提供，并不能取代它。

执行绘制出的着法。 在确定性任务中，预测出的视频片段可以完整执行。在地形地图上，完整执行片段能在 82.8% 的测试地图上找到最低成本路线；每走一步都重新预测时，这一比例为 83.0%，探针集上的比例分别为 76.8% 和 78.8%。在均匀地图上，片段能解出 94.6% 的测试地图，而在前 500 张地图上采用滚动时域的方法则为 96.0%。每次调用时，每个片段最多执行八步，而非一步，因此每执行一步耗时为 3.6 秒，而不是 19.4 秒（表 21）。面对对手时，预测的应答可能与实际应答不同。在视频模型正确绘制出第一步棋和对手应答，并接着绘制出下一步的 1,411 道国际象棋谜题中，该步棋正确的比例为 41.2%，而

此时，基于同一组着法之后观测到的棋盘重新进行预测，正确率达到 52.5%。表 15 给出的绘制着法正确率为 39.5%，因为其中还计入了 62 个在对手应答后就停止的片段。面对精确的四子棋求解器，滚动时域博弈在获胜起点中取得胜利的比例为 28.7%；若不考虑对手的应答，直接执行绘制出的着法，则这一比例为 22.7%（$p = 0.004$，150 局游戏）。

L. 数据扩展

训练	图像，混合		图像，最佳着法	视频
样例	无引导	引导		
80k	–	–	–	12.7
160k	5.9	10.3	–	–
400k	14.6	19.9	–	26.4
1M	30.4	39.4	39.7	41.6
3M	–	–	56.7	54.8

表17. 国际象棋验证准确率（%）与训练集大小的关系，取三个随机种子的平均值（图像：2,000 道谜题，视频：1,000 道）。「混合」采用第 M 节的提示混合，「最佳着法」采用表1中的训练方案。

我们使用其他设置完全相同的训练方案，逐步增加数据量来训练国际象棋模型，并使用三个采样随机种子在验证谜题上评估每个模型（表17）。所有模型的准确率都稳步提升，而且在数据量最大时，增益仍未放缓，这表明限制这些结果的是数据，而非模型容量。采用第M节的提示混合时，图像模型的误差率大致按照样例数量的幂律下降，指数约为 0.17（无引导）和 0.21（引导）。视频模型也遵循类似曲线（0.15）。每次拟合仅使用三个点，因此这些指数只是粗略估计。仅使用最佳着法提示时，将数据量增加至原来的三倍，即三百万道谜题，可使验证准确率从 39.7% 提高到 56.7%。

有两种效应会随着数据量增加而增强。首先，随着模型性能提升，对比引导带来的帮助更大：在 16 万个样例时，相较于无引导采样，增益为 4.4 个百分点；在 40 万个样例时为 5.3 个百分点；在 100 万个样例时则为 9.0 个百分点。其次，与绘制单个决策相比，想象完整棋局变化的优势也在扩大：在 8 万、40 万和 100 万个视频时，分别多出 2.3、3.2 和 8.0 个测试点（第K节）。围棋也呈现出相同趋势：在 100万个局面上训练的模型，在 20.9% 的测试局面中选择了与 KataGo 相同的着法；而在 10万个局面上训练时，这一比例为 11.3%（$p < 10^{-20}$），采用多数投票时则为 27.9%。

M. 对比引导

方法。 条件扩散模型可以引导其避开已学会与不同提示相关联的输出。对于单步图像模型，LoRA 还会针对两个辅助提示进行训练：一个要求生成任意合法着法，另一个要求生成看似合理但会导致败局的着法。

着法 c^-，由与标注参考走法相同的引擎离线标注（第 A.1 节）。推理时，速度会被引导着避开败着，

$$\tilde{v} = v_\theta(\cdot, c) + w\left(v_\theta(\cdot, c) - v_\theta(\cdot, c^-)\right), \qquad (11)$$

即无分类器引导 [9]，其中无条件分支由学习得到的负面提示替代。我们使用$w = 3$。在围棋中，c^- 要求选择一个至少会使胜率降低 20 个百分点的着法。推理时不会查询任何引擎，每个去噪步骤只需额外进行一次网络评估（表 21）。视界模型始终使用其普通采样器，因为引导需要对整个视频进行第二次遍历。

结果。 对比引导提升了提示混合训练模型的表现，尽管主要结果采用的是无引导模型。在一百万道国际象棋谜题上，引导将单个样本的准确率从 31.8% 提高到 40.9%（表9）。这部分增益很大程度上抵消了提示混合的影响：仅使用最佳着法提示进行训练，无需引导即可达到 39.9%；使用三倍的数据进行训练则达到 56.0%（表1）。采用八个样本投票后，引导带来的增益从 9.1 个百分点缩小到 3.6 个百分点（40.0% 对比 43.6%）。在 10 万个围棋棋局上也呈现出类似趋势：引导将单个样本的准确率从 9.3% 提高到 14.0%，而仅使用最佳着法提示进行训练的准确率为 11.3%。采用八次投票后，这几种方法的准确率分别达到 14.1%、17.4% 和 17.8%。因此，我们将引导视为可选方法，而不在主表中使用。随着训练数据增多，引导相较于无引导采样的优势也随之增大（第L节）。

N. 结果是否取决于骨干网络?

Qwen-Image 使用 Qwen3-VL 对提示和输入图像进行编码，因此可能是视觉语言模型解决了任务，而扩散模型只是呈现答案。我们在地形寻路任务上对此进行测试：方格代价被隐藏，必须从样例中推断。这些图像模型会在一张图像中绘制完整路线，不同于表2中的闭环模型；后者在测试地图上同样达到 87.5%，但在探测地图上达到 85.0%。所有变体使用相同的数据、训练步数和学习率。

模型	文本编码器能看到图像吗?	最短路线	绕行或穿越
Qwen-Image-2.1	yes	87.5	87.6
Qwen-Image-2.1，盲编码器	no	88.1	82.6
FLUX.2 [klein] (4B)	no	86.1	83.8
Wan2.2 视频	no	82.8	76.8
最少步数（广度优先搜索）	–	59.2	50.0

表18。使用不同主干网络和文本编码器进行地形寻路（占测试集和探测地图的百分比）。图像模型在一张图像中绘制完整路线，而视频模型则绘制并执行最多八步的片段。

我们首先阻止 Qwen-Image 的文本编码器读取图像，使扩散变换器成为唯一的视

觉输入路径。最短路线准确率仍为 88.1% 对 87.5%（精确 McNemar $p = 0.59$）。在探测地图上，准确率下降了 5.0 个百分点，从 87.6% 降至 82.6%（$p = 0.008$），因此，编码器有助于处理困难的选择，但推断隐藏成本并非必需：盲模型的表现仍远高于机会水平。接下来，我们使用 FLUX。2 [klein], , 它采用较小的 40亿参数骨干网络和纯文本编码器。其准确率分别为 86.1% 和 83.8%，与 Qwen-Image 相比没有显著差异（$p = 0.29$ 且0。07），并且所有模型的表现都优于不考虑地形的规划器（表 18）。视频模型提供了第三种测试，因为它的文本编码器也仅处理文本。在同一百万道国际象棋谜题中，对于 10,000 道测试谜题，其准确率达到 41.7%，高于 Qwen-Image 的 39.9%（$p = 0.001$）。后者在采用对比引导时准确率达到 40.9%（$p = 0.16$）。这些结果共同表明，视觉决策能力并非 Qwen-Image 独有，也不局限于能够读取图像的编码器。

O. 为什么生成着法？直接预测基线

我们比较了直接根据渲染后的图像预测着法的模型，以及语言模型根据同一局面的文本描述预测着法的模型。它们使用相同的谜题或地图和验证集，并采用相同的无规则判读方式评分：不合法的答案计为错误（表 19）。强大的符号国际象棋变换器 [18]使用棋盘编码和国际象棋专用着法词汇表，因此不能直接用于其他任务。

视觉语言模型。 未经微调时，Qwen3-VL（8B 和 32B）的表现不如随机合法着法，无论输入的是棋盘图像还是 FEN文本。使用相同的一百万个谜题进行微调，并要求以 UCI记谱法作答后，8B 模型的准确率达到 42.0%，显著低于我们的单步视频模型的 50.5%（表 19）。在区分回忆与实战的测试中，差距进一步扩大（表 9）。在训练分布之外，准确率为 34.1%，而另一模型为 45.0%。镜像使准确率下降 -5.5 个百分点，而当一个棋子的变化使另一着法成为最佳着法时，模型在 61.1% 的情况下仍保留旧着法，对比之下另一模型为 48.1%。这些测试表明，与视觉语言模型相比，视觉生成模型能够更可靠地响应棋盘中的任务相关变化。

纯文本语言模型。 我们使用与视觉语言模型规模相近的 Qwen3-8B，在相同的一百万个谜题上，采用相同的训练方案进行微调。它读取 FEN，并使用 UCI记谱法作答；在答案前不进行推理，而是采用贪心解码。其 40.9% 的准确率显著低于单步视频模型（精确 McNemar$p < 10^{-60}$）。在远离训练分布的局面上，差距进一步扩大（32.7% 对 45.0%），镜像成本 -7.1 个百分点；而在变化后的棋盘上，当另一着法更佳时，它仍有 65.6% 的情况保留原来的着法。使用三百万个谜题时，其准确率达到 52.0%，而在相同数据上，单步图像模型达到 62.5%。无论输入是图像还是文本，这两个语言模型的准确率都相近，失败模式也相似。

国际象棋，一百万道训练谜题 模型	通过次数	测试（%）
Qwen3-VL-32B，零样本	0	0.8
Qwen3-VL-8B，微调	1	42.0
Qwen3-8B，微调的文本（FEN）	1	40.9
DINOv2 + 标签输出	1	21.8
DINOv2 + 标签输出	8	50.0
DINOv2 + 国际象棋走法预测头†	8	**78.5**
我们的方法，图像	1	45.5
我们的方法，视频	1	50.5

延迟提示分叉，第一步（%）	训练布局		直角
模型	1 次遍历	12 次遍历	fork
DINOv2，第一步棋	67.2	100.0	58.0
DINOv2 ，第一步棋 + 路线	58.7	100.0	49.5
我们的方法、视频模型，绘制路线	**98.4**	未运行	**94.0**

表19. **在相同数据上训练的直接预测基线。**上图：在训练谜题上的遍历次数不同，国际象棋测试准确率也不同；一次遍历的更新次数与我们的图像模型相同。文本模型读取 FEN，其余模型读取棋盘图像。我们的方法各行使用在验证集上选定的采样步数（图像模型为三步，其仅使用最佳着法提示进行训练；视频模型为一步）。 †国际象棋专用输出会为每个 64×64 起点–终点对评分。下图：延迟提示分叉。一次遍历后，分类器只利用附近的墙；十二次遍历后，它们便能解决训练布局中的每个叉子战术。

与用文本作答相比，在棋盘上画出着法能够在数据匹配的情况下获得更高的准确率，尤其是在正确动作取决于当前局面而非熟悉模式的情况下。

输出决定一切。 在冻结的 DINOv2-L 特征上使用通用 Transformer，在进行八次遍历并采用 4,096 类走法输出时，准确率达到 50.0%；当棋盘与其图像块未对齐时，准确率为 48.1%。仅将输出改为在起点和终点格上对着法进行评分，准确率便提升至 78.5%，几乎与使用完整棋盘结构的预测头（78.6%）相同。因此，空间输出解释了分类器很大一部分的优势。绘制路线能够跨任务提供空间输出，而无需为每个任务分别设计预测头。在与我们的图像模型相同、进行一次遍历且批次大小为 8 的设置下，通用分类器的准确率达到 21.8%，而我们的模型为 45.5%。

从每个示例中学习。 经过训练直至收敛后，分类器能够解决延迟提示分叉（100%），并且比我们的全路线模型更频繁地产生最优路线。在对岔路地图进行一次遍历后，按照与我们的模型相当的曝光量进行训练，它们只有在墙距离一格或两格时才会对墙作出响应（67.2%；采用路线监督时为 58.7%）。我们的路线绘制模型达到 98.4%。

迁移到新的布局。 我们还测试了走廊呈直角相交的岔路，这种布局在训练中并未出现。

任务（测试集）	步	单任务	单个模型基线	
四子棋，保持获胜	2	76.3	77.3	60.4
地形，最低成本路线	1	87.3	83.4	59.2
地形，最低成本路线	20	87.5	81.7	59.2
绕行或穿越	1	79.0	79.0	50.0
绕行或穿越	20	85.0	83.0	50.0

表20。**一个模型完成两个任务**（%）。从预训练模型开始，在四子棋和地形上共同训练一个图像模型。单任务模型使用相同的样例和训练方案。基线分别是四子棋的一步启发式方法、用于寻路的不考虑地形的规划器，以及绕行还是穿越探测任务中的机会基线。

收敛后的分类器得分分别为 58.0% 和 49.5%，并且经常选择训练布局中更受偏好的方向，结果走进墙里。我们的模型则保持在 94.0%。这表明，分类器学会了布局特定捷径，而路径绘制模型能够迁移到新的几何结构中。

一个模型完成两个任务。我们在四子棋和地形寻路上联合训练一个图像模型，两个任务使用相同的提示。模型必须根据输入图像识别任务，并且每个任务使用数量相同的样例，采用与其单任务对应模型相同的训练方案（表 20）。联合训练保持了四子棋的性能以及判断是否穿越高代价地形的能力，并且在最低成本路线上与单任务模型的差距保持在几分之内。因此，一个视觉生成器可以支持不止一个任务。

P. 绘制动作的合法性

由于没有模型接收到规则，其原始绘制可能不合法，所有报告的结果都将此类动作计为失败。在测试集上直接测量可见，几乎所有原始绘制都是合法的。在国际象棋中，使用三百万个谜题训练的图像模型，在 97.0% 的局面中绘制出了合法着法（使用一百万个谜题时为 88.3% 至 88.7%）；使用一百万个谜题训练的视频模型则达到 93.4%。在合法绘制走法之后，97.3% 的想象视频棋盘在全部 64 个棋格上都正确，单个棋格的正确率为 99.95%。在围棋中，被标记的点有 99.8% 的概率合法，而基础模型仅为 43%。在四子棋中，棋子在图像输出中的非满列和正确落子行均达到 100.0% 的正确率，在视频输出中则为 99.6%，尽管从未明确说明重力规则。对于图像模型和视频模型，路线分别在 94.7% 和 97.4% 的最短路径地图上有效，在地形地图上分别为 98.5% 和 95.0%；此外，台球输出中分别有 99.2% 和 98.2% 是可读的击球。

Q. 运行时间与内存

训练使用配备八块英伟达 RTX A5000 GPU（每块 24GB）的节点。图像模型及其文本编码器无法同时装入一块 GPU，因此每个进程使用两块：一块

训练	模型	实际耗时	GPU 小时
国际象棋，3M*	图像 / 视频	93 小时 / 79 小时	1,487 / 1,268
国际象棋，100万	图像 / 视频	51 小时 / 49 小时	407 / 389
围棋，100k / 100万	图像	19 小时 / 90 小时	152 / 722
地形路径	图像 / 视频	24 小时 / 9 小时	195 / 75
四子棋	图像 / 视频	9 小时 / 8 小时	75 / 66
台球	图像 / 视频	32 小时 / 22 小时	260 / 172

推理	步	图形处理器秒数 / 决策
国际象棋图像，无引导 / 有引导	20	3.1 / 6.1
国际象棋图像，单步	1	0.7
国际象棋视频，完整片段 / 单步	30 / 1	8.9 / 2.2
围棋图像，384 像素，无引导 / 引导	20	6.8 / 13.1
地形路径，图像闭环	20	每着法 3.3
地形路径、视频	30	每步 3.6
地形路径、视频、滚动时域	30	每步 19.4

表 21. **训练与推理成本**。上：在 A5000 GPU 上进行训练，包括定期验证。除使用两台机器、共十六个 GPU 的 * 次运行外，每次运行均使用八个 GPU。下：测量的推理成本，包括模型加载和解码。寻路条目给出每执行一步的实际耗时秒数。

用于运行扩散变换器，另一块用于运行 Qwen3-VL 和变分自编码器。在256 像素分辨率下，每个进程使用两个样例时，扩散变换器所在 GPU 的峰值显存占用为 18.3 GB，编码器所在 GPU 的峰值显存占用为 18.5 GB。视频模型可装入一块 GPU，峰值显存占用根据分辨率不同介于 15GB 和 18GB 之间。表 21报告了主要运行任务的训练成本。

推理成本主要取决于去噪步数，以及引导是否会增加第二次前向传播（表 21）。单步采样（第C节）成本最低；在没有引导时，准确度也更高。投票会使成本按样本数量倍增。在逐步执行的任务中，执行协议很重要：一次调用中，视频片段最多可执行八步，解决一张地形图所需时间与图像模型的闭环大致相同；而滚动时域每执行一步都会生成一个新的视频片段，速度约慢五倍。