arXiv:2611.32347v1 [cs.LG] 24 Nov 2026

超越硬标记：用于多模态蛋白质生成的预测记忆

Xiaoliang Shi[1], Zichen Wang[1],Zhijian Wei[1],Zhanpeng Shi[2,3],Zhenghong Zhou[1],Yunpeng Wang[1],Shuangjia Zheng[1†]

[1]上海交通大学[2]复旦大学[3]上海创新研究院

摘要

多模态蛋白质语言模型通过迭代掩码解码统一序列与离散结构标记，为联合蛋白质设计提供了一种颇具前景的范式。然而，标准解码器在每次迭代时都依赖无记忆掩码重置：只有少量新近确定的硬标记会被向前传递，而所有其余尚未解决的位置都会被重置为相同的 *[MASK]* 嵌入。在开放式联合生成中，由于没有任何模态提供固定的外部锚点，这种无状态转换限制了单步细化深度，也使两个轨道缺乏稳定的跨步软锚点，从而导致收敛缓慢或退化性的重复坍塌。我们观察到，尚未确定的预测会表现出显著的时间相干性，而非无规律的噪声。基于此，我们提出预测记忆（*PredMem*），这是一种无需训练的自条件递归机制，可在多个步次之间传递这些软信念。在每一步中，*PredMem* 将原始 *logits* 汇总为每个残基处按概率加权的 *Top-k* 期望嵌入，并在下一步将其仿射插值到符合条件的掩码输入中，且不会增加去噪器的额外成本。在 *DPLM-2*、*DPLM-2.1* 和 *ESM3* 上，*PredMem* 分别将无条件序列-结构协同生成成功率提高了 *3.6*、*17.7* 和 *48.7* 个百分点，同时改善了质量-多样性权衡并抑制了重复模式坍缩。受控诊断表明，这些增益主要源于无锚点跨模态协同演化中的早期轨迹稳定，而不是额外计算或简单的熵减小，由此确立了 *PredMem* 作为一种高效的多模态离散生成推理机制。

1. 引言

蛋白质是生命的分子机器，驱动着从催化和免疫防御到信号传导 [1]等过程。它们的功能受序列-结构-功能关系支配：氨基酸序列编码结构偏好，从而塑造三维构象 [2, 22], , 进而决-

决定生物学功能。计算蛋白质设计利用这一原理，工程化设计具有所需结构和功能特性的新型蛋白质 [20]。

近年来，生成模型极大地推动了这一目标的发展，最初主要采用单模态方法——包括基于结构的扩散模型，如 RFdiffusion [35],、Chroma [21], 和 FrameDiff [37], , 它们在连续的三维空间中生成骨架；以及蛋白质语言模型，如 ProtGPT2 [11], 、ProGen2 [24], 和 ProGen3 [4], , 它们生成氨基酸序列。然而，序列和结构通过共同的进化与生物物理约束内在耦合 [23, 26], , 因此孤立地对任一模态建模，会限制模型捕捉二者相互依赖关系的能力。以 ESM 3 [14] 和 DPLM-2 系列 [18, 34], 为代表的多模态生成模型，通过将三维构象表示为离散结构标记，并在统一标记空间中对两种模态进行建模，解决了这一问题。通过迭代掩码解码，这些模型在两条轨道上预测并确定标记，旨在生成彼此自洽的蛋白质序列和结构。

联合序列–结构生成的一个基础标准是生物物理自洽性：从生成序列中独立预测的结构是否与协同生成的主链 [14, 29, 34, 37]一致。然而，当两个轨道必须在没有固定对应项的情况下联合解析时，例如在无条件协同生成和基序支架构建中 [32,35, 40], 现有多模态模型仍难以维持这种自洽性。我们发现了<style id='4'>无记忆掩码重置：在每个步数中，去噪器都会对所有未解析位置预测完整分布，但采样器只传播新近确定的硬标记中的稀疏子集；其余所有位置——无论是保持掩码状态还是重新加掩码——都会被重置为

†通讯作者：shuangjia.zheng@sjtu.edu.cn

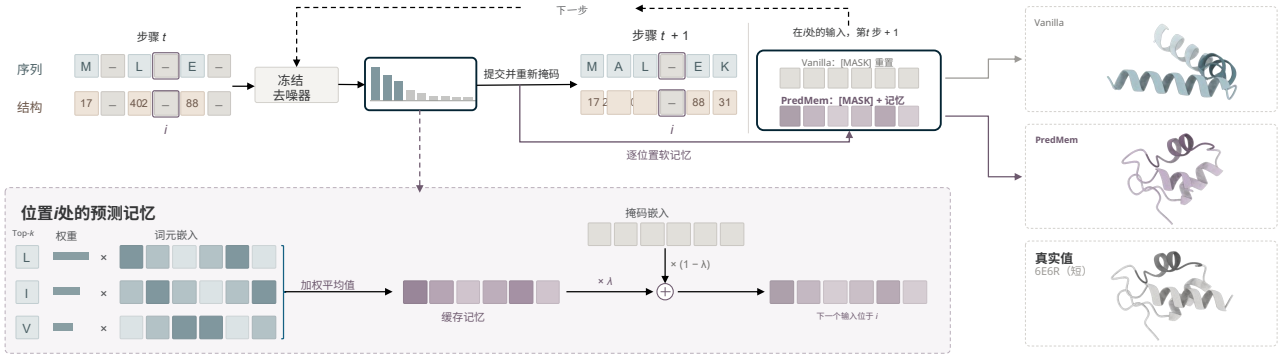

图1. **预测记忆概览。** PredMem 会为每个掩码位置提供软记忆，即按概率加权的 Top-k 词元嵌入，并将其与掩码嵌入插值融合，以取代原始的 [MASK] 重置。图中展示了 6E6R-短序列的基序支架构建输出，分别由原始解码和 PredMem 生成，并与真实值进行比较。

相同的 [掩码] 嵌入 [7, 12, 27]。这种无状态转换会抹去重新加掩码位置的局部上下文，并迫使持续处于掩码状态的位置在固定的单步前向深度内从头重建连续假设，而不是跨步进行递归细化。不同于具有稠密静态锚点的单模态生成，早期联合生成从两个轨道上的稀疏确定开始；如果没有跨步连续状态来稳定中间信念，任何一个轨道都无法充当另一个轨道的可靠软锚点，从而导致收敛缓慢或反复坍缩 [17]。这引出了一个核心问题：未确定的中间预测是否会以足够的时间相干性演化，从而在多个步数之间充当递归自条件状态?

检查解码轨迹可以发现，尚未确定的预测并非像杂乱噪声般波动；相反，在每个残基上，它们都表现出显著的时间相干性。在连续步数中，仍处于掩码状态的位置上的 Top-1 预测，比不考虑位置的基线更频繁地保持一致。这种平滑演变表明，存在可重复利用的特定位置信息。现有推理方法并未利用这一点：按轨道分类器自由引导（CFG）[40] 会在单个步数内重塑 logits，却仍会在迭代过程中重置未解析位点；而离散扩散中的自条件和软掩码 [6, 15] 则需要修改架构或重新训练，因而无法即插即用于冻结模型。我们提出预测记忆 （PredMem），这是一种无需训练的自条件机制，仅在去噪器的输入—输出接口上运行，以稳定早期协同生成。在每个步数，PredMem 会将原始、无引导的 logits 汇总为每个位置上按概率加权的 Top-k词元嵌入，并将其融入符合条件的未解析位点处模型原生的 [MASK] 嵌入，同时保持固定条件和受保护的基序不变，且不会增加去噪器的计算成本。

在 ESM3、DPLM-2 和 DPLM-2.1 上进行评估后，PredMem 持续提升无条件协同生成设计的成功率——在 DPLM-2 上从 **29.8% 提升至 33.4%**，在 DPLM-2.1 上提升幅度最大，从 **28.9% 提升至 46.6%**，在 ESM3 上则从 **33.4% 提升至 82.1%**。受控实验确认，位置打乱记忆会显著削弱这些增益，而熵受控基线和计算量匹配基线排除了不确定性降低或额外去噪步骤是唯一解释的可能。此外，收益主要集中在两条轨道都能自由共同演化的场景——联合协同生成和 ESM3 基序支架设计，表明**将位置特异性的软信念作为循环自条件状态传递**，能够增强无锚定多模态生成中的序列–结构自一致性。

我们的主要贡献总结如下：**1. 未提交信念的时间连贯性：** 我们表明，掩码解码中的中间未提交预测会跨步平滑演化，从而为反馈至冻结去噪器提供稳定的连续信号，且无需重新训练。**2.预测记忆：** 我们提出一种无需训练的推理时自条件循环，将概率加权的 Top-k 期望嵌入融合到符合条件的 [MASK]输入中，且不会增加去噪器的额外成本。**3. 系统验证与机制洞见：** 在 DPLM-2、DPLM-2.1 和 ESM3 上，PredMem持续提升协同生成成功率；受控诊断进一步指出，早期轨迹稳定化和无锚定的跨模态协同演化是其主要驱动因素。

2. 相关工作

多模态蛋白质生成。 离散多模态生成器通过掩码建模或扩散建模，将蛋白质序列与量化结构标记统一起来，代表性模型包括 ESM3 [14]，DPLM-2 [34] （在仅处理序列的 DPLM [33] 基础上扩展而来）和 DPLM-2。1 [18]，而多模态流则将连续主链与离散序列相结合 [5]

全原子扩散则协同设计主链、序列和侧链 [10]。我们不修改架构，而是研究冻结的离散生成器如何在联合解码期间重用中间预测，以保持序列–结构自洽性。

迭代掩码解码与推理时引导。 吸收态离散扩散 [3, 27]以及 Mask-Predict [12] 和 MaskGIT [7] 等基于置信度的并行解码器，会在重置剩余掩码的同时迭代提交标记，而重新掩码采样器 [31] 则进一步允许修改已提交的标记。在推理时，Zhou 等人 [40] 将无分类器引导（CFG） [16] 和奖励引导的束搜索应用于多模态蛋白质。离散扩散引导 [25]以及针对蛋白质序列生成器 [13, 36] 的引导，同样会将每个去噪步骤引向目标属性。不同于仅在步骤内引导 logits 或选择硬标记的方法，PredMem 会跨步骤传播未解析位点的软信念，从而补充按轨道分类器自由引导。

自条件与软掩码。 在连续扩散 [8, 28], 、迭代提示 [39], 和改编的掩码扩散 [6]中，自条件会将先前的去噪器估计输入后续步骤。与之最为相关的是软掩码 [15]和 DMax [9], ，它们会将先前的 top-k或中间嵌入融合到掩码输入中，但需要进行模型训练或适配。相比之下，PredMem 完全无需训练即可在冻结的多模态蛋白质生成器上运行，并在两条轨道上维护位置特定的分支局部缓存，且无需额外的去噪器前向计算。

3. 预测记忆

为克服迭代掩码解码中的无记忆掩码重置且无需重新训练的问题，PredMem 引入了一个轻量级、位置特定的缓存，用于存储模型先前的预测。在每一步中，PredMem 汇总每个残基处排名靠前的k候选词元嵌入，并将这些汇总结果插值到后续前向传播期间符合条件的 [MASK] 位置中，同时完全保持冻结的生成器、采样规则和引导机制不变。下文首先确立实证动机（第3.1节），然后形式化循环状态、读写操作以及兼容引导的缓存（第3.2节至第3.4节）。完整公式推迟至附录A。

3.1. 未提交信念的时间连贯性

由于预训练去噪器被冻结且从未接受训练在软 [MASK] 输入上，将不稳定或含噪的中间预测反馈回网络，可能会注入分布外扰动并使解码不稳定。因此，我们考察未确定的预测是否会在相邻步数之间平滑演变。在没有 PredMem 的情况下，我们选取在两个相邻步数中均被掩码的位置，并将其进行比较

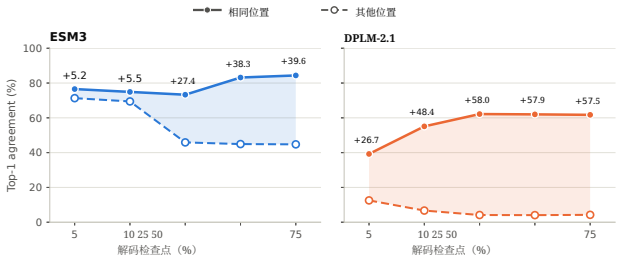

图2. **未提交的预测表现出位置特异性的时间相干性。**在没有 PredMem 的情况下，比较相邻解码步在相同位置与所有其他位置上的 Top-1一致性。

在相同位置的 Top-1 预测与一个对照进行比较，该对照将每个预测分别与其他位置的所有预测配对，从而反映仅由词元频率导致的一致性（图2；统计数据按轨道计算，并在序列和结构上取平均）。由此得到两个观察结果。

首先，在整个解码过程中，ESM3 和 DPLM-2.1 上相同位置的一致性都高于对照组，这表明未解析位点的预测会随着位置呈现特异性的时间相干性，而非随机波动。这启发我们采用按位置索引的循环缓存。其次，即便离散的 Top-1词元发生变化，连续词元在嵌入空间中的距离仍显著小于均匀随机抽取的不同词元对：ESM3 和 DPLM-2.1 上的余弦相似度分别为 0.49 和 0.53，而均匀随机抽取的不同词元对分别为 0.014 和 0.097。嵌入空间中的这种局部连续性表明，与其用单个硬标记概括预测，不如采用软期望嵌入。ESM3 解码早期的差距最小，此时预测高度重复，改进空间有限；然而，即便将这种微弱的位置特异性差异保留下来，也有助于打破全局对称性，防止早期出现重复坍塌。

3.2. 迭代掩码解码与状态

令 $z_t = (s_t,\ x_t)$ 表示长度为 L 的蛋白质在解码步骤 $t \in \{0, \dots, T-1\}$ 时的部分掩码序列和结构轨道，并令 c 表示固定的任务条件（例如结构基序、条件序列或固定主链）。对于每个模态 $m \in \{S, X\}$，令 $\mathcal{V}_m$ 表示其有效词元词汇表，令 E_m: $\mathcal{V}_m \cup \{\text{MASK}\} \to \mathbb{R}^{d_m}$ 表示其原生词元嵌入映射。

在步骤 t，冻结去噪器 f_θ 会在一个或多个解码分支 $b \in \mathcal{G}$ 上进行评估（其中 $\mathcal{G} = \{+\}$ 用于无引导解码，$\mathcal{G} = \{+, u_S, u_X\}$ 用于按轨道分类器自由引导；参见第3.4 节）。每个分支 b 接收其分支特定输入 $(z_t^b,\ c^b)$，并针对每个位置 i 和词元 $v \in \mathcal{V}_m$ 输出原始的、无引导的模态分数 $\ell^{b,\,t}_{i}{}^{,\,m}(v)$。

在标准迭代掩码解码中 [7, 27], ，任何位置

tion i在步骤 t 仍未解析的位置，仅由常数嵌入 E_m(MASK)表示——这种无记忆掩码重置 会丢弃此前的预测分布 $\ell^{b,\,m}_{t-1,\,i}$。相反，PredMem 会维护一个瞬时记忆缓存 M_{t-1} (初始化为 $M_{-1}=\varnothing$)，其中每个条目 $\mu^{b,\,m}_{t-1,\,i}\in M_{t-1}$ 均由分支 b、模态 m和残基位置 i进行索引。由于每次前向传播都会使用更新后的摘要覆盖缓存 M_t，过去的预测便会沿轨迹递归传播，而无需存储多步历史记录。

3.3. 记忆写入与读取

写入预测摘要。在步骤 t 的前向传播之后，PredMem 会从原始分数 $\ell^{b,\,m}_{t,\,i}$ 中提取紧凑摘要，然后再进行无分类器引导和采样温度缩放。令 $K^{b,\,m}_{t,\,i}\subseteq\mathcal{V}_m$表示在 $\ell^{b,\,m}_{t,\,i}$ $(1\le k\le|\mathcal{V}_m|)$ 下的前 k 个标记的集合。我们在单位温度下，对这一截断支持集上的原始概率进行归一化：

$$\widetilde{p}^{b,m}_{t,i}(v)=\frac{\exp\ell^{b,m}_{t,i}(v)}{\sum_{u\in K^{b,m}_{t,i}}\exp\ell^{b,m}_{t,i}(u)},\quad v\in K^{b,m}_{t,i},\qquad(1)$$

并将所得的概率加权平均嵌入写入缓存：

$$\mu^{b,m}_{t,i}=\sum_{v\in K^{b,m}_{t,i}}\widetilde{p}^{b,m}_{t,i}(v)E_m(v).\qquad(2)$$

在引导之前以单位温度计算确定性期望 $\mu^{b,\,m}_{t,\,i}$，可使缓存的信念与外部采样调度计划解耦，使摘要保持在合理的前 k 个嵌入的凸包内，并避免将采样方差注入下一步。摘要会同时写入掩码和可见残基位置；因此，如果某个已提交的位置随后被调度器重新掩码，其最新的预测摘要即可立即使用。

在符合条件的掩码位置读取。在步骤 t，仅将来自 M_{t-1} 的缓存摘要读入尚未确定、指定用于生成的位置。令 $\mathcal{I}^{\text{gen}}_m$表示任务允许解码器在模态 m中生成的位置集合，并令 $\mathcal{I}^{\text{fb}}_m\subseteq\mathcal{I}^{\text{gen}}_m$表示符合记忆反馈条件的位置子集（不包括固定条件残基、受保护的基序坐标以及特殊 BOS/EOS标记）。对于活跃的分支–模态对 $(b,\ m)\in\mathcal{B}$，在步骤 t接收反馈的位置集合为

$$\mathcal{A}^{b,m}_t=\mathcal{I}^{\text{fb}}_m\cap\{i:z^{b,m}_t(i)=\text{MASK}\}.\qquad(3)$$

当 $M_{t-1}\neq\varnothing$ 时，PredMem 会将每个活跃位置 $i\in\mathcal{A}^{b,\,m}_t$ 处的掩码嵌入替换为仿射插值——

tion

$$\widetilde{e}^{b,m}_{t,i}=(1-\lambda)E_m(\text{MASK})+\lambda\mu^{b,m}_{t-1,i},\qquad i\in\mathcal{A}^{b,m}_t,\ (4)$$

其中，$\lambda\in[0,1]$ 控制反馈强度，而所有其余位置 $i\notin\mathcal{A}^{b,\,m}_t$ 保留其原生输入。通过这种插值，每个尚未解析的位置都会在各步中携带一个柔性的、位置特定的假设，而不会过早确定为硬标记。将 $\lambda=0$ 或 $M_{-1}=\varnothing$ 设置为即可精确恢复基线前向传播；因此，步骤 $t=0$ 与未经修改的基线保持一致。

特定架构的实例化。虽然方程（4）定义了词元空间中的标准读取操作，但不同的预训练骨干网络在自注意力之前对嵌入进行不同的预处理。在 ESM3 [14], 中，PredMem 直接将方程（4）应用于序列和结构标记嵌入。在 DPLM-2 和 DPLM-2.1 [18, 34], 中，其输入层会先进行掩码比例重缩放，再进行层归一化，PredMem 会在层归一化后的注意力输入处进行插值，并匹配原生缩放因子。此外，由于 DPLM-2.1 将每个结构标记分解为 13 个二进制头（$2^{13}=8192$完整编码），PredMem 会评估所有 8192 种完整编码的联合对数概率，以选出排名靠前的k 个结构候选项，并通过原生码本投影映射这些候选项。详情见附录 A.4。

3.4. 引导兼容递推

按轨道分类器自由引导下的分支局部记忆。为在无需重新训练的情况下引导多模态生成，按轨道分类器自由引导 [40] 会评估一个完全条件化分支 $+$ 以及条件丢弃分支 u_m，其中目标模态 m 的条件输入会被屏蔽。模态 m 的引导逻辑值为

$$g^m_{t,i}=\ell^{u_m,m}_{t,i}+w_m\big(\ell^{+,m}_{t,i}-\ell^{u_m,m}_{t,i}\big),\qquad(5)$$

其中，$w_m\ge1$ 是引导权重（当 $w_m=1$ 时可恢复无引导条件化逻辑值 $\ell^{+,\,m}_{t,\,i}$）。

关键在于，在各分支之间共享单个缓存会将条件信息从 $+$ 泄漏到 u_m，或破坏有意丢弃的轨道。因此，PredMem 始终维护严格的分支局部缓存。具体而言，在使用 $\mathcal{G}=\{+,u_S,u_X\}$进行联合生成时，每个分支仅会接收其生成的目标轨道的反馈，具体由活动配对集合定义

$$\mathcal{B}=\{(+,S),(+,X),(u_S,S),(u_X,X)\}.\qquad(6)$$

在单模态引导生成中，$\mathcal{B}$ 简化为{(+，m)，(u_m，m)}；在无引导生成（$\mathcal{G}=\{+\}$）中，$\mathcal{B}$仅包含 +的生成模态。每个分支仅从由其自身无引导分数构建的缓存中读取 $\ell^{b,\,m}_{t-1}$，从而确保记忆不会被注入任何丢弃条件中。

循环转移与计算成本。在每一步 t，解码器通过式（4）从 M_{t-1} 读取，针对所有活跃分支评估一次冻结去噪器 f_θ，通过式（2）写入更新后的缓存 M_t，并将生成的引导分数 g_t 传递给未经修改的原生采样器。由于 PredMem 仅在去噪器的输入—输出接口处运行，因此总步数 T 和分支批量大小 $|\mathcal{G}|$ 均保持不变：无论是否使用 PredMem，一条轨迹都需要 T 次批量去噪器调用（$T|\mathcal{G}|$ 次分支评估）。唯一的计算开销来自 top-k 选择、嵌入聚合和向量插值，它们需要 $O\left(L\sum_{(b,\ m)\ \in\mathcal{B}} d_m\right)$ 的记忆，与 T 无关。

4. 实验

4.1. 实验设置

模型与任务。我们评估了三个冻结的多模态蛋白质生成器：ESM3、DPLM-2 和 DPLM-2。1 [14,18, 34]遵循 Zhou 等人 [40], 的做法，正文考察两项序列-结构协同生成任务：无条件协同生成和基序支架构建；而单模态条件任务（结构预测和逆折叠 [19]）则推迟至附录中讨论。

受控比较。对于两项联合任务，我们比较四种配置——Vanilla、PredMem、CFG 和 CFG+PredMem——以评估在有无继承引导的情况下记忆反馈的效果。在每个模型和任务内，四种配置均共享解码计划、步数预算、位置选择规则和采样温度，而 CFG 和 CFG+PredMem 保留了 Zhou 等人 [40]所采用的相同引导设置。因此，将 PredMem 与 Vanilla 进行对比、将 CFG+PredMem 与 CFG 进行对比，即可在去噪器调用次数和分支评估预算匹配的条件下，单独考察记忆反馈的作用。

指标与报告。对于联合任务，ESMFold [23]会独立地重新折叠生成的序列。我们使用骨架均方根偏差（scRMSD）和 TM-score（scTM）比较重新折叠后的结构与生成的结构 [38], ，并将重新折叠后的 pLDDT 作为置信度指标报告。这些质量指标是对所有生成输出取平均得出的，而 Foldseek [30]聚类（在 TM-score 阈值为 0.5 时，在每种长度或基序变体内进行）则严格基于成功子集计算。聚类数量衡量成功子集的广度，同时取决于产出率和多样性。报告值为五个随机种子均值；标准差、精确的超参数设置、针对 λ 和 k的消融实验，以及单模态结果见附录D 和 E。

4.2. 无条件协同生成

流程。两条轨道均从完全掩码状态开始，不进行外部条件输入。每次运行生成 500 对样本：在 $\{100, 200, 300, 400, 500\}$ 的每个长度下生成 100 对。当样本满足 scRMSD < 2 ˚A. 表 1 报告成功率、结构质量和成功簇数量。

无论是否使用引导，PredMem 都能提升无条件协同生成效果。表 1 显示，在三种生成器上，无论是否使用 CFG，PredMem 都提高了平均成功率。无引导时效果最为显著：在 ESM3 上，PredMem 将成功率从 33.4% 提升至 82.1%；在 DPLM-2.1 上，则从 28.9% 提升至 46.6%；在 DPLM-2 上的提升较小，为 +3.6 个百分点。这些增益并不局限于阈值化的成功率：在 ESM3 上，跨步传递软预测抑制了无引导原始解码中的重复模式坍缩——否则会生成低复杂度序列和延展且折叠不佳的主链——使所有输出的平均 scRMSD 从 31.9 降至 2。3 ˚A，平均 scTM 从 0.545 提升至 0.944，重新折叠后的 pLDDT 从 67.9 提升至 90.5。如图中一条匹配的 $L = 100$ 轨迹所示（图3），原始 ESM3 会坍缩为近乎同聚物的聚丙氨酸序列（仅含 2 种残基类型），其生成的棒状结构与 ESMFold 重新折叠得到的结构明显偏离（scRMSD $= 34.1$ ˚A，TM $= 0.19$）；而 PredMem 生成的序列组成多样（含 18 种残基类型），且序列与结构高度吻合（scRMSD $= 0.38$ ˚A，TM $= 0.99$，pLDDT $= 85.4$）。DPLM-2.1 也呈现相同趋势，平均 scRMSD 从 5.2 降至 3。3 ˚A。即使在 CFG 基础上，PredMem 仍能带来增益，使各模型的成功率提升 2.3 至 16.5 个百分点。在 ESM3 和 DPLM-2.1 上，PredMem 单独使用的效果已经优于 CFG，分别高出 17.6 和 13.7 个百分点，同时每步仅评估一个分支，而非三个分支。因此，在继承的 CFG 设置下，对这两个生成器而言，跨步传递预测似乎比在单步内调整 logits 更有效。

更高的成功率不一定会扩大结构覆盖度。成功簇的数量并不会随成功率同步增加。没有引导时，PredMem 在 ESM3 和 DPLM-2 上增加了簇，但在 DPLM-2.1 上减少了 17.0。使用 CFG 时，PredMem 在 ESM3 和 DPLM-2.1 上都减少了约 30 个簇，而在 DPLM-2 上仅变化了 1.2 个簇。这些结果表明，自洽性提升与成功结构覆盖度扩大是两回事。第 5 节通过更多控制实验考察了这一权衡。

4.3. 基序支架构建

协议。基序固定残基身份和结构条件，同时生成其余位置。每次运行会生成 24 个基序变体中的 2,400 个支架，每个变体采样 100 个样本。成功率要求

表 1。**无条件协同生成**。每次运行中，每个模型生成 500 个序列-结构对，并在五个随机种子上进行测量。彩色数字显示 PredMem 相对于 Vanilla 的变化，以及 CFG+PredMem 相对于 CFG 的变化。PredMem 提高了三个生成器的平均成功率，其中对 ESM3 的提升最为显著。

		质量				多样性
模型	方法	成功率（%）↑	pLDDT ↑	scTM ↑	scRMSD ↓	#簇 ↑
DPLM-2	Vanilla	29.80	82.00	0.852	4.95	78.0
	PredMem	33.40 +3.60	82.47 +0.47	0.852 +0.000	4.79 -0.16	**81.0** +3.0
	CFG	31.88	82.42	0.859	4.42	73.4
	CFG+PredMem	**34.20** +2.32	**82.99** +0.57	**0.862** +0.003	**4.37** -0.05	74.6 +1.2
DPLM-2.1	Vanilla	28.88	84.91	0.836	5.17	92.6
	PredMem	46.60 +17.72	87.77 +2.86	0.889 +0.053	3.32 -1.85	75.6 -17.0
	CFG	32.88	85.14	0.853	4.37	**98.0**
	CFG+PredMem	**48.32** +15.44	**87.85** +2.71	**0.895** +0.042	**3.13** -1.24	68.4 -29.6
ESM3	Vanilla	33.36	67.91	0.545	31.91	62.0
	PredMem	**82.08** +48.72	90.53 +22.62	**0.944** +0.399	**2.32** -29.59	73.0 +11.0
	CFG	64.48	88.36	0.891	6.62	**114.4**
	CFG+PredMem	80.96 +16.48	**91.59** +3.23	0.943 +0.052	2.47 -4.15	84.4 -30.0

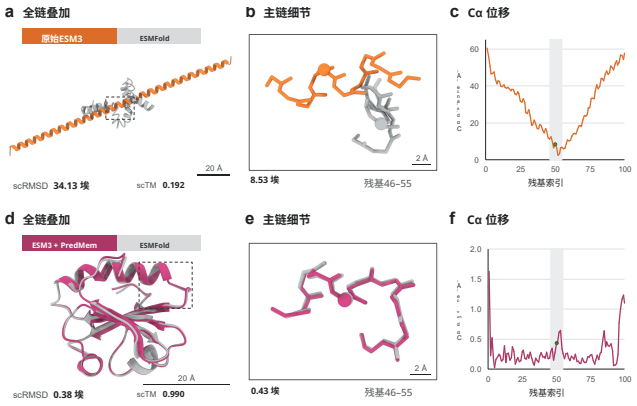

图3。**PredMem 可防止 ESM3 中的重复坍塌**。在长度匹配为 100 的案例中，对比原始 ESM3 和 ESM3+PredMem：全链叠加、主链细节，以及相对于 ESMFold 重新折叠结果的逐残基 Cα 位移。

scRMSD < 2 ˚A 以及生成基序的均方根偏差 < 1 ˚A（基于解码后的主链），遵循 Zhou 等人 [40]。如果某个变体的 100 个支架中有任意一个成功，则认为该变体已解决。表 2 报告成功率、自洽性、已解决变体数量和多样性。

PredMem 改善了全部三个生成器的基序支架构建。在表 2 中，无论是无引导设置还是在 CFG 基础上添加 PredMem，都会提高每个模型的平均成功率。在基序支架构建中，ESM3 受益最大——在无引导设置下提升 12.0 个百分点，从 30.4% 升至 42.4%，并将平均 scRMSD 从 7.0 降至 4。3 ˚A；而 DPLM-2 和 DPLM-2.1 的增益较小，为 0.8 至 2.0 个百分点。在全部三个生成器中，无引导的 Pred-Mem 单独使用时的平均成功率也超过了 CFG。关键在于，与第 4.2 节中观察到的集中现象不同，Pred- Mem 在全部六种模型—引导比较中都扩大了成功簇数量——在 CFG 基础上，ESM3 增加了 11.4 个簇，DPLM-2.1 增加了 6.8 个簇——并且在六种设置中的五种下解决了更多基序变体。因此，在基序条件下，保留未提交的预测能够同时提高设计产率和结构广度。

5. 分析

第 4 节表明，PredMem 无论是否使用 CFG，都能改善协同生成。现在，我们将考察其反馈如何改变解码、更简单的替代方案能否复现这一增益，以及它如何权衡质量与多样性。

5.1. 早期信念整合

PredMem 将轨迹收敛提前至终止词元附近。PredMem 并非在后期发挥润色作用，而是作为早期信念整合器，在已确定上下文仍然稀疏时，引导未解决位置趋向自洽的最终状态。为直接观察这一机制，我们在图 *Top-3* 覆盖率中追踪4——即仍处于掩码状态的位置中，其最终词元已出现在当前 Top-3 预测里的比例。在两种骨干模型上，PredMem 都加速了解码早期的内部共识形成，但具体动态有所不同。在 ESM3 上，原始解码在初始阶段的覆盖度较低，陷入停滞；而 PredMem 在最初的 10% 步数内打破了这一平台期，使覆盖度最多提高 22 个百分点，并在第一个四分之一阶段始终保持两位数的领先优势。在 DPLM-2.1 上，原始解码本身收敛较快，而 PredMem 则在其

表2. **基序支架构建**。每次运行生成2,400个支架，成功率为五个随机种子的平均值。PredMem提高了所有三种生成器的平均成功率，其中对ESM3的提升最为显著。

模型	方法	质量			覆盖度	多样性
		成功率（%） ↑	scTM ↑	scRMSD ↓	已解决 / 24 ↑	#簇 ↑
DPLM-2	Vanilla	20.60	0.666	**4.95**	16.8	37.8
	PredMem	**21.71** +1.11	0.668 +0.002	**4.95** +0.00	**17.0** +0.2	**41.0** +3.2
	CFG	20.48	0.661	5.03	16.2	37.6
	CFG+PredMem	21.25 +0.77	**0.669** +0.008	**4.95** -0.08	16.4 +0.2	38.2 +0.6
DPLM-2.1	Vanilla	18.33	0.659	5.38	15.2	30.2
	PredMem	20.29 +1.96	**0.671** +0.012	**5.16** -0.22	15.4 +0.2	31.2 +1.0
	CFG	19.07	0.652	5.48	14.8	30.0
	CFG+PredMem	**21.08** +2.01	0.664 +0.012	5.23 -0.25	**16.4** +1.6	**36.8** +6.8
ESM3	Vanilla	30.43	0.711	6.97	19.4	57.4
	PredMem	**42.40** +11.97	0.767 +0.056	4.33 -2.64	21.8 +2.4	64.8 +7.4
	CFG	35.81	0.754	4.29	**22.6**	83.6
	CFG+PredMem	41.93 +6.12	**0.776** +0.022	**3.75** -0.54	22.0 -0.6	**95.0** +11.4

自身收敛之前，在最早的几个步数中便迅速推动共识形成——进度达到 5% 时，覆盖度提高了 11 个百分点——随后原始解码才迎头赶上。

早期建立共识可以解释大部分质量增益。 由于 PredMem 主要在许多硬标记被确定之前稳定跨模态假设，因此其影响集中在解码的初始阶段。如图 5所示，仅在前 25%的步数中启用 PredMem，随后在原始解码下完成整个轨迹，已经能够在 ESM3 上获得完整运行成功率增益的约 94% ，并在 DPLM-2.1上获得约 79%。一旦足够多的硬标记锚定了序列—结构上下文，软反馈的边际价值便会降低；当将 PredMem 的启用范围延长至前 25%之后时，各次运行的不确定性区间也大幅重叠。这也解释了基序支架构建中较小的增益（第 4.3节），因为固定基序从第一步起就已经提供了确定的跨模态锚点。

5.2. 排除替代解释

我们检验额外的去噪步骤、非特异性记忆反馈或温度缩放，能否在无引导解码下复现 PredMem 的增益（表3，采用第4.2节的五个随机种子预算）。在单张 A800 GPU 上，以批量大小1 测量全部五种长度下的实际解码时间（每种长度重复10次，不计模型加载和折叠过程）。

额外的解码步骤并不能复现这一增益。在 Vanilla 中将步数预算加倍，而不增加内存 $2T$ 会使解码时间加倍，但在两个骨干模型上，成功率变化均处于不同随机种子带来的波动范围内，几乎可以忽略（表3）。相比之下，

表3。**无引导解码下的对照实验：成功率与成本**。五个随机种子的平均成功率（第4.2节的实验方案）及解码时间；短横线表示未测量的条目。标准差、差值和对照定义见附录F。

方法	ESM3		DPLM-2.1	
	成功率（%） ↑	时间（秒） ↓	成功率（%） ↑	时间（秒） ↓
Vanilla	33.4	17.86	28.9	24.83
Vanilla $2T$	34.5	35.74	30.4	49.59
PredMem+偏移	48.8	—	35.1	—
PredMem+下限	72.5	—	36.0	—
PredMem	**82.1**	18.42	**46.6**	28.02

PredMem 仅增加 3.1%（ESM3）和 12.8%（DPLM-2.1）的解码时间，即可实现完整增益；峰值内存开销最多仅增加 0.12 GiB，从而排除了额外计算量是其驱动因素的可能性。

位置对应关系对两种骨干模型都很重要。在 PredMem+shift 中，跨未解析位点对记忆进行循环移位，在保留缓存摘要多重集的同时打破残基级对齐。这会消除 PredMem 在两种骨干模型上约三分之二的增益，仅剩下源自全局正则化的较小残余增益。

PredMem 对词元排序的调整超越了熵和温度缩放的作用。为检验仅靠分布变尖是否就能带来增益，当 PredMem 的熵低于无记忆参考时，PredMem+floor 会提高温度，使熵与无记忆参考相匹配，同时保持词元排序不变。两种骨干模型仍显著优于 Vanilla，

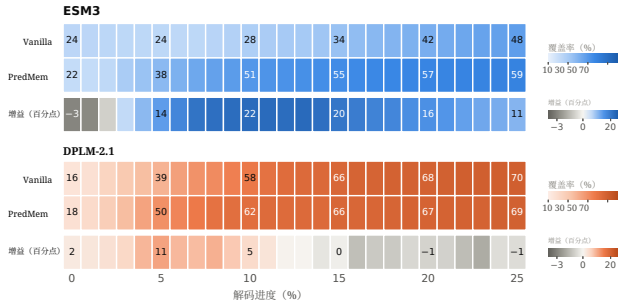

图 4. **PredMem 加速了对终止标记的早期共识形成**。解码前 25% 的 Top-3 覆盖率。增益行报告 PredMem 相对于 Vanilla 的百分点差异。

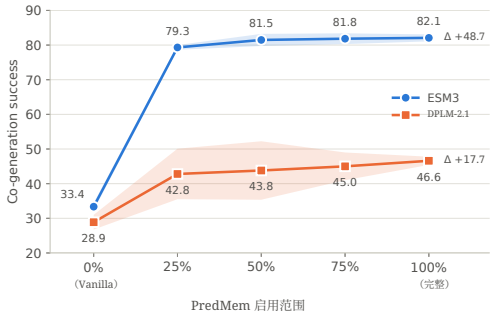

图 5. **不同解码窗口中的功能贡献**。启用 PredMem 时，最终协同生成成功率与初始步数比例的关系。

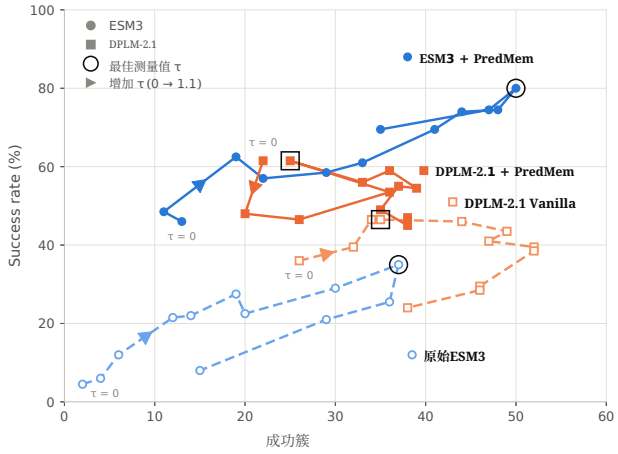

图 6。**不同解码温度下的成功率–多样性帕累托前沿**。 无引导解码在 12 个温度 $\tau \in [0, 1.1]$ 下分别对原始解码和 PredMem 进行扫描；圆形表示 ESM3，方形表示 DPLM-2.1。

保留了 PredMem 增益的约 80% （ESM3）和 40% （DPLM-2.1）。一致地，温度扫描表明，虽然重新调整 Vanilla 的温度可恢复 DPLM-2.1 在默认温度下部分增益，但 PredMem 在各种温度下始终明显领先（第 5.3 节）。此外，在相同解码状态下对比 logits，发现其排名变化不可能由正标量温度缩放造成：在进度约为25% 时，PredMem 会改变 ESM3 的 Top-1 预测，影响17.3% 的序列位置和 62.9% 的结构位置。总之，这些对照实验表明，尽管熵降低对 DPLM-2.1 的增益有一定贡献，PredMem 的主要机制仍是位置对齐的信念细化，从而重塑词元排序（详见附录F）。

5.3. 质量、多样性与帕累托前沿

生成式蛋白质设计面临结构可行性与输出多样性之间固有的权衡。通过引导早期解码步数朝向自洽状态，Pred-Mem 将轨迹引入高可行性折叠盆地，在保持较高结构覆盖度的同时，大幅提升设计成功率。

推动成功率与结构多样性的帕累托前沿。 为考察这种权衡关系，我们扫描固定采样温度 τ，遍历12个运行点，范围从 0。0到 1。1，采用无引导解码，并通过 Foldseek 聚类评估每个设置下的200 个候选序列，TM-score 阈值为0。5（图6）。在两种骨干模型上，PredMem 均显著将质量–多样性前沿向上推移：

•ESM3 上的严格帕累托支配。 在所有温度下，ESM 3+PredMem 的轨迹都严格位于 ESM3 原始解码的上方和右侧。PredMem 并非以多样性换取质量，而是同时提升二者：在各自成功率峰值温度下，PredMem 将成功率从 35。0% 提高至 80。0%，增幅超过一倍，同时将成功聚类产量从 37 提升至 50。

• **DPLM-2.1 上的高成功率区间转变**。在 DPLM- 2.1 上，PredMem 将成功率提升至 60% 以上，而原始解码的成功率最高仅为 46.5%。尽管在高温度下，较早的盆地收敛限制了聚类数量的上限，但在所有重叠的多样性水平上，DPLM-2.1+PredMem 的成功率都显著更高；当可设计性是主要瓶颈时，它提供了更高效的运行区间。

6. 结论

我们引入了预测记忆（PredMem），这是一种无需训练的推理时循环机制，用于解决迭代式掩码多模态蛋白质生成中的无记忆掩码重置问题。PredMem 在解码步骤之间传递概率加权的 Top-k 期望嵌入，并在无分类器引导下严格执行分支局部隔离，从而显著提升了序列–结构自洽性和设计

成功率，且适用于 ESM3、DPLM-2 和 DPLM-2.1，同时几乎不会增加计算和内存开销。除了显著提升无条件协同生成和 ESM3 基序支架设计的增益外，PredMem 还能抑制重复性坍缩，并将质量–多样性帕累托前沿向上推进。尽管我们的研究在冻结模型上采用固定的循环调度，并依赖 *insilico* 自洽性基准测试，但未来工作将探索不确定性感知门控，并将这种循环软信念机制扩展到实验性功能设计，例如针对特定靶标的 *denovo* 抗体序列–结构协同设计。

大语言模型使用声明

我们仅以两种有限方式使用了大语言模型工具。首先，GPT 和 Codex 帮助润色英文写作，使其表达清晰、语法正确。其次，Codex 帮助我们调试代码。除此之外，所有研究内容均为作者原创。这包括预测记忆的构想、数学推导，以及针对 ESM3、DPLM-2 和 DPLM-2.1 的推理钩子。还包括实验设计和对所有结果的解读。为确保准确性，我们审阅了大语言模型对文本和代码提出的每一处修改，并根据原始实验输出和原始参考文献核实了这些修改。因此，作者对整篇稿件承担全部责任。

参考文献

[1] Bruce Alberts. 细胞分子生物学. Garland 科学出版社, 2017年。 1[2] Christian B. Anfinsen. 决定蛋白质链折叠的原理。《科学》，181(4096):223–230，1973年。 1[3] Jacob Austin、Daniel D. Johnson、Jonathan Ho、Daniel Tarlow 和 Rianne van den Berg。离散状态空间中的结构化去噪扩散模型。载于 神经信息处理系统进展，2021年。 3[4] Aadyot Bhatnagar、Sarthak Jain、Joel Beazer、Samuel C. Curran、Alexander M. Hoffnagle、Kyle Shan Ching、Michael Martyn、Stephen Nayfach、Jeffrey A. Ruffolo 和 Ali Madani。规模扩展解锁更广泛的生成能力与更深入的蛋白质功能理解。载于 神经信息处理系统进展，2025年。 bioRxiv 预印本 10.1101/2025.04.15.649055。 1[5] Andrew Campbell、Jason Yim、Regina Barzilay、Tom Rainforth 和 Tommi Jaakkola。离散状态空间上的生成流：通过蛋白质协同设计应用实现多模态流。载于 第*41*届国际机器学习会议论文集，第5453–5512页，2024年。2[6] Michael Cardei、Huu Binh Ta 和 Ferdinando Fioretto。掩码扩散模型的简单自条件适应。 *arXiv* 预印本 *arXiv:2604.26985*，2026年。 2, 3[7] Huiwen Chang、Han Zhang、Lu Jiang、Ce Liu 和 William T. Freeman。MaskGIT：掩码生成图像变换器。载于 *IEEE/CVF* 会议论文集《计算机视觉与模式识别》, 第11315–11325页, 2022年。 2, 3[8] Ting Chen、Ruixiang Zhang 和 Geoffrey Hinton。模拟比特：使用带自条件的扩散模型生成离散数据。arXiv 预印本 arXiv:2208.04202，2022 年。3[9] Zigeng Chen 、 Gongfan Fang、Xinyin Ma、Ruonan Yu 和 Xinchao Wang。DMax：面向离散大语言模型的激进并行解码。arXiv 预印本 arXiv:2604.08302，2026年。3[10] Alexander E Chu、Jinho Kim、Lucy Cheng、Gina El Nesr、Minkai Xu、Richard W Shuai 和 Po-Ssu Huang。全原子蛋白质生成模型。《美国国家科学院院刊》，121(27):e2311500121，2024年。3[11] Noelia Ferruz、Steffen Schmidt 和 Birte H¨ocker。ProtGPT2：用于蛋白质设计的深度无监督语言模型。《自然·通讯》，13(1):4348，2022年。1[12] Marjan Ghazvininejad、Omer Levy、Yinhan Liu 和 Luke Zettlemoyer。Mask-Predict：条件掩码语言模型的并行解码。载于《2019年自然语言处理经验方法会议暨第9届国际自然语言处理联合会议论文集》，第6112–6121页，2019年。2, 3[13] Nate Gruver、Samuel Stanton、Nathan Frey、Tim G J Rud-ner、Isidro Hotzel、Julien Lafrance-Vanasse、Arvind Rajpal、Kyunghyun Cho 和 Andrew G Wilson。使用引导式离散扩散进行蛋白质设计。载于《神经信息处理系统进展》，2023年。3[14]Thomas Hayes、Roshan Rao、Halil Akin、Nicholas J. Sofroniew、Deniz Oktay、Zeming Lin、Robert Verkuil、Vin-cent Q. Tran、Jonathan Deaton、Marius Wiggert、Rohil Bad-kundri、Irhum Shafkat、Jun Gong、Alexander Derry、Raul S. Molina、Neil Thomas、Yousuf A. Khan、Chetan Mishra、Car-olyn Kim、Liam J. Bartie、Matthew Nemeth、Patrick D. Hsu、Tom Sercu、Salvatore Candido 和 Alexander Rives。使用语言模型模拟5亿年进化历程。《科学》，387(6736):850–858，2025年。1, 2, 4, 5, 13[15] Michael Hersche、Samuel Moor-Smith、Thomas Hofmann 和 Abbas Rahimi。软掩码扩散语言模型。载于《国际学习表征会议》，2026年。2, 3[16] Jonathan Ho 和 Tim Salimans。无分类器扩散引导。arXiv 预印本 arXiv:2207.12598，2022年。3[17] Ari Holtzman、Jan Buys、Li Du、Maxwell Forbes 和 Yejin Choi。神经文本退化的奇特案例。载于《国际学习表征会议》，2020年。2[18] Cheng-Yen Hsieh、Xinyou Wang、Daiheng Zhang、Dongyu Xue、Fei Ye、Shujian Huang、Zaixiang Zheng 和 Quan-quan Gu。阐明多模态蛋白质语言模型的设计空间。载于《第42届国际机器学习会议论文集》，第24156–24175页。PMLR，2025年。1, 2, 4, 5, 13[19] Chloe Hsu、Robert Verkuil、Jason Liu、Zeming Lin、Brian Hie、Tom Sercu、Adam Lerer 和 Alexander Rives。从数百万个预测结构中学习逆折叠。载于《第39届国际机器学习会议论文集》，第8946–8970页，2022年。5, 13

[20] Po-Ssu Huang、Scott E Boyken 和 David Baker。从头蛋白质设计的成熟。《自然》，537(7620):320–327，2016年。1[21] John B Ingraham、Max Baranov、Zak Costello、Karl W Bar-ber、Wujie Wang、Ahmed Ismail、Vincent Frappier、Dana M Lord、Christopher Ng-Thow-Hing、Erik R Van Vlack 等。利用可编程生成模型探索蛋白质空间。《自然》，623(7989):1070–1078，2023年。1[22] John Jumper、Richard Evans、Alexander Pritzel、Tim Green、Michael Figurnov、Olaf Ronneberger、Kathryn Tunyasuvu-nakool、Russ Bates 等。利用 AlphaFold 实现高精度蛋白质结构预测。《自然》，596(589):583–597，2021年。1[23] Zeming Lin、Halil Akin、Roshan Rao、Brian Hie、Zhongkai Zhu、Wenting Lu、Nikita Smetanin、Robert Verkuil、Ori Kabeli、Yaniv Shmueli、Allan dos Santos Costa、Maryam Fazel-Zarandi、Tom Sercu、Salvatore Candido 和 Alexan-der Rives。利用语言模型进行进化尺度的原子级蛋白质结构预测。《科学》，379(6637):1123–1130，2023年。1, 5, 14[24] Erik Nijkamp、Jeffrey A Ruffolo、Eli N Weinstein、Nikhil Naik 和 Ali Madani。ProGen2：探索蛋白质语言模型的边界。《细胞系统》，14(11):968–978，2023年。[25] Hunter Nisonoff、Junhao Xiong、Stephan Allenspach 和 Jennifer Listgarten。解锁离散状态空间扩散模型与流模型的引导。载于《第十三届国际学习表征会议》，2025年。3[26]Alexander Rives、Joshua Meier、Tom Sercu、Siddharth Goyal、Zeming Lin、Jason Liu、Demi Guo、Myle Ott、C Lawrence Zitnick、Jerry Ma 等。无监督学习扩展至2.5亿条蛋白质序列后，生物结构与功能自然涌现。《美国国家科学院院刊》，118(15):e2016239118，2021年。1[27] Subham Sekhar Sahoo、Marianne Arriola、Yair Schiff、Aaron Gokaslan、Edgar Marroquin、Justin T. Chiu、Alexander Rush 和 Volodymyr Kuleshov。简单而有效的掩码扩散语言模型。载于《神经信息处理系统进展》，2024年。2, 3[28] Robin Strudel、Corentin Tallec、Florent Altch´e、Yilun Du、Yaroslav Ganin、Arthur Mensch、Will Grathwohl、Niko-lay Savinov、Sander Dieleman、Laurent Sifre 等。用于文本生成的自条件嵌入扩散。arXiv 预印本 arXiv:2211.04236，2022年。3[29] Brian L Trippe、Jason Yim、Doug Tischer、David Baker、Tamara Broderick、Regina Barzilay 和 Tommi Jaakkola。面向基序支架构建问题的三维蛋白质骨架扩散概率建模。载于《第十一届国际学习表征会议》，2023年。1[30]Michel Van Kempen、Stephanie S Kim、Charlotte Tumescheit、Milot Mirdita、Jeongjae Lee、Cameron LM Gilchrist、Johannes S¨oding 和 Martin Steinegger。使用 Foldseek 快速准确地进行蛋白质结构搜索。《自然·生物技术》，42(2):243–246，2024年。5, 14[31] Guanghan Wang、Yair Schiff、Subham Sahoo 和 Volodymyr Kuleshov。利用推理时扩展对离散扩散模型进行重新掩码。《神经信息处理系统进展》，38:147282–147339，2026年。3

[32] Jue Wang、Sidney Lisanza、David Juergens、Doug Tischer、Joseph L Watson、Karla M Castro、Robert Ragotte、Amijai Saragovi、Lukas F Milles 等。《利用深度学习搭建蛋白质功能位点支架》。《科学》，377(6604):387–394，2022年。1[33] Xinyou Wang、Zaixiang Zheng、Fei Ye、Dongyu Xue、Shujian Huang 和 Quanquan Gu。《扩散语言模型是多功能的蛋白质学习器》。arXiv 预印本 arXiv:2402.18567，2024年。2[34] Xinyou Wang、Zaixiang Zheng、Fei Ye、Dongyu Xue、Shujian Huang 和 Quanquan Gu。《DPLM-2：多模态扩散蛋白质语言模型》。载于国际学习表征会议，2025年。1、2、4、5、13[35] J oseph L. Watson、David Juergens、Nathaniel R. Bennett、Brian L. Trippe、Jason Yim、Helen E. Eisenach、Woody Ahern、Andrew J. Borst、Robert J. Ragotte、Lukas F. Milles、Basile I. M. Wicky、Nikita Hanikel、Samuel J. Pellock、Alexis Courbet、William Sheffler、Jue Wang、Preetham Venkatesh、Isaac Sappington、Susana V´ azquez Torres、Anna Lauko、Valentin De Bortoli、E mile Mathieu、Sergey Ovchinnikov、Regina Barzilay、Tommi S. Jaakkola、Frank DiMaio、Minkyung Baek 和 David Baker。《利用 RFdiffusion 从头设计蛋白质结构与功能》。《Nature》，620 (7976):1089–1100，2023年。1[36]Junhao Xiong、Ishan Gaur、Maria Lukarska、Hunter Nisonoff、Luke M Oltrogge、David F Savage 和 Jennifer Listgarten。《ProteinGuide：对蛋白质序列生成模型进行即时属性引导》。arXiv 预印本 arXiv:2505.04823，2025年。3[37] Jason Yim、Brian L. Trippe、Valentin De Bortoli、Emile Mathieu、Arnaud Doucet、Regina Barzilay 和 Tommi Jaakkola。《用于蛋白质主链生成的 SE(3)扩散模型》。载于第40届国际机器学习会议论文集，页码40001–40039，2023年。1[38] Yang Zhang 和 Jeffrey Skolnick。《用于自动评估蛋白质结构模板质量的评分函数》。《蛋白质：结构、功能与生物信息学》，57(4):702–710，2004年。5、14[39] Kun Zhou、Yifan Li、Wayne Xin Zhao 和 Ji-Rong Wen。《Diffusion-NAT：用于非自回归文本生成的自提示离散扩散》。载于欧洲计算语言学协会第18届会议论文集（第1卷：长篇论文），页码1438–1451，圣朱利安，马耳他，2024年。计算语言学协会。3[40] Yi Zhou、Wang Qipeng、Yunqing Liu、Jun Xia、Qing Li 和 Wenqi Fan。《在推理时解锁多模态蛋白质语言模型》。载于2026年自然语言处理实证方法会议论文集，页码35132–35151，2026年。1、2、3、4、5、6、11、13、14、17

A. 实现细节：分支输入与读取接口

A.1. 分支输入与受保护位置

分支构造是一个特定于任务的映射 $(z_t^b,\ c^b) \ =\mathcal{D}_b\ (z_t,\ c)$，由每个解码器及其 PredMem 增强版本共享。参考的无条件协同生成构造为

$$z_t^{+} = (s_t, x_t), \qquad z_t^{u_S} = (s_t, \text{null}), \qquad z_t^{u_X} = (\text{null}, x_t), \tag{7}$$

其中 空值 表示原生的丢弃条件表示形式，而非移除残基位置 [40]。对于每个分支，只有式（6）中的配对会接收反馈。在单目标任务中，仅使用对应的两组配对；若 CFG 被禁用，则仅使用 $(+, m)$。

生成资格与反馈资格彼此不同。例如，坐标条件可以保护基序的输入坐标，同时仍将其结构标记作为生成目标；这类受保护位置可以属于 $\mathcal{I}_X^{\text{gen}}$，同时仍处于 $\mathcal{I}_X^{\text{fb}}$ 之外。对于仅坐标的基序条件化，具有有限条件坐标的位置、固定氨基酸以及 BOS/EOS标记仍处于 $\mathcal{I}_X^{\text{fb}}$ 之外，并且会从记忆反馈和读取中排除。输入坐标会被保留，解码后的基序几何结构将在不钳制到参考结构的情况下进行评估。

A.2. ESM3 的标记空间读取

在符合条件的掩码位置，公式（4）会替换原生的模态特定标记嵌入组件，同时保留其他输入组件以及后续的原生处理。可见标记和受保护的条件输入保持不变。Top-k 选择、softmax 归一化和嵌入聚合均以 FP32 计算，并且缓存会在读取接口处分离并转换为输入数据类型。

A.3. DPLM变体的读取映射

为清晰起见，省略分支、模态、位置和时间索引，令 h_0 表示符合条件的位置处原生的丢弃后输入，并令 $\mu_\kappa = \kappa\mu$ 表示公式 μ（2）经过适当原生缩放后的缓存摘要。层归一化前读取形成

$$\widetilde{h} = (1-\lambda)h_0 + \lambda\mu_\kappa. \tag{8}$$

~第一个注意力块接收 LN(h)，而残差~流使用 h_0。后层归一化读取则形成

$$\widetilde{u} = (1-\lambda)\,\text{LN}(h_0) + \lambda\,\text{LN}(\mu_\kappa), \tag{9}$$

它替代经过层归一化的注意力输入，同时保留 h_0 在残差路径上的内容。组合读取对注意力使用公式（9），对残差流使用公式（8）。这些是不同的读取映射；DPLM 实验使用公式（9）中的层归一化后读取（附录 D.6）。

原生缩放。 当原生标记丢弃路径使用掩码比例重缩放时，写作

$$s(I) = \frac{1-\rho_{\text{train}}}{1 - n_{\text{mask}}(I)/n_{\text{nonpad}}(I)}, \tag{10}$$

对于分母为正的输入，其中$\rho_{\text{train}} = 0.12$ 为原生训练掩码率常数。令s_2 表示在拼接输入上计算的缩放值，s_1 表示在序列半段上计算的缩放值。DPLM-2 的嵌入路径给出 $\kappa_S = \kappa_X = s_2^2$，而对于 DPLM-2.1，$\kappa_S = s_1 s_2$ 和 $\kappa_X = s_2$。当词元丢弃被禁用时，这些缩放因子均为 1。乘数 κ 结合了缓存词元嵌入与所选读取接口之间的缩放阶段，并在读取时刻基于当前分支输入进行计算。

归一化的影响。 对于层归一化增益 γ、偏置β_{LN}以及 $\epsilon_{\text{LN}} > 0$，令 $\bar{\mu}_\kappa$ 和 $\text{Var}(\mu_\kappa)$分别表示 μ_κ各元素的均值和方差，并考虑一个满足$h_0 = 0$的位置。在公式（8）的层归一化前读取下，注意力输入满足

$$\text{LN}(\lambda\mu_\kappa) - \text{LN}(0) = \gamma \odot \frac{\lambda(\mu_\kappa - \bar{\mu}_\kappa \mathbf{1})}{\sqrt{\lambda^2\,\text{Var}(\mu_\kappa) + \epsilon_{\text{LN}}}}. \tag{11}$$

仿射偏置会相互抵消，并且对于正的 λ且$\lambda^2\text{Var}(\mu_\kappa) \gg \epsilon_{\text{LN}}$，归一化会近似抵消注意力输入中的标称幅度缩放。在公式（9）的后层归一化读取下，相同位置给出

$$\widetilde{u} - \text{LN}(0) = \lambda\,\gamma \odot \frac{\mu_\kappa - \bar{\mu}_\kappa \mathbf{1}}{\sqrt{\text{Var}(\mu_\kappa) + \epsilon_{\text{LN}}}}, \tag{12}$$

其随 λ 线性缩放，并且不会被归一- 化抵消。因此，两种读取对注意力输入中的 λ响应不同，这解释了为何必须明确指定读取接口，也构成了附录 D.6 中消融实验的动机。

A.4. DPLM-2.1 的完整码记忆

DPLM-2.1 使用 13 个二值分量表示一个结构标记。设 $d_j(v) \in \{0,1\}$ 为完整码 v 中分量 j 的标签，$p^{b,X}_{t,i,j}(a)$为该分量输出头中标签 a的无引导概率。对于所有 2^{13} 个完整编码，定义

$$\ell^{b,X}_{t,i}(v) := \log q^{b,X}_{t,i}(v) = \sum_{j=1}^{13} \log p^{b,X}_{t,i,j}(d_j(v))\,. \tag{13}$$

这是由输出头指定的因式分解分布。我们根据这些分数选出排名前 k 的完整编码，并使用1) 和 (2) 式，其中

$$E_X(v) = \text{quant2emb}(\mathbf{d}(v)), \tag{14}$$

其中 $\mathbf{d}(v) = (d_1(v) \ldots, d_{13}(v))$ 是该位模式的原生输入表示，包括其符号约定和缩放，并保留投影偏置。完整代码分数使用 FP32 计算，并禁用自动混合精度转换，同时采用原生采样器的编码到位的排序。记忆构建使用 CFG 之前且缓存温度为一时的概率，同时保留原生硬标记采样器。

B. 预测记忆的性质

以下分析刻画了仿射令牌空间读取接口处预测记忆的近似保真度、方差降低行为和分布稳定性。

B.1. Top-k 截断

为清晰起见，省略分支、模态、位置和时间索引。令 $p = \text{softmax}(\ell)$为 $\mathcal{V}_m$ 上的单位温度分布，该分布构成公式（1）和（2）的基础。令 K为 top-k 集合，令 $1 - \varepsilon_K = \sum_{v \in K} p(v) > 0$ 为其保留质量，使得 $\widetilde{p}(v) = p(v)/(1 - \varepsilon_K)$ 在 K上。定义完整词汇表均值和缓存摘要

$$\mu^{\text{full}} = \sum_{v \in \mathcal{V}_m} p(v) E_m(v), \qquad \mu = \sum_{v \in K} \frac{p(v)}{1 - \varepsilon_K} E_m(v). \tag{15}$$

对于嵌入直径 $D_m = \max_{u,\, v \in \mathcal{V}_m} \|E_m(u) - E_m(v)\|_2$,

$$\|\mu - \mu^{\text{full}}\|_2 \le \varepsilon_K D_m. \tag{16}$$

证明。 若 $\varepsilon_K > 0$，令 μ_{tail} 为K之外的条件均值。则 $\mu^{\text{full}} = (1 - \varepsilon_K)\mu + \varepsilon_K \mu_{\text{tail}}$。两个条件均值都位于嵌入凸包内，其直径至多为 D_m。尾部概率为零的情形显然成立。

对于方程（4）中的仿射读取，top-k 输入与完整均值输入之间的差异因此至多为 $\lambda \varepsilon_K D_m$。该界将截断误差的决定因素归结为被舍弃的尾部概率质量 ε_K，而非词汇表的绝对大小；这解释了为何只要预测分布足够集中，固定且较小的 k仍能保持忠实。

B.2. 均值与采样预测输入

固定硬状态、缓存分布和符合条件的位置。将活跃三元组 $(b,\ m,\ i)$ 按 $(b,\ m) \in \mathcal{B}$和 $i \in \mathcal{A}^{b,\ m}_{t}$ 以 j为索引，并写出 $\widetilde{p}_j = \widetilde{p}^{b,\ m}_{t-1,\ i}$、 $\mu_j = \mu^{b,\ m}_{t-1,\ i}$以及 $E_j = E_m$。令 $V_j \sim \widetilde{p}_j$，使得 $\mu_j = \sum_v \widetilde{p}_j(v) E_j(v)$

如公式（2）所示。对于公式（4）中的仿射令牌空间读取，定义

$$\begin{aligned} u_j^{\text{draw}} &= (1 - \lambda) E_j(\text{MASK}) + \lambda E_j(V_j), \\ \bar{u}_j &= (1 - \lambda) E_j(\text{MASK}) + \lambda \mu_j. \end{aligned} \tag{17}$$

这里 $\bar{u}_j = \bar{e}^{b,\ m}_{t,\ i}$ 是方程（4）中的读取。令 U^{draw} 和 $\bar{U}$ 拼接这些分量，同时保持所有其余输入不变。它们的一阶与二阶矩为

$$\begin{aligned} &\mathbb{E} U^{\text{draw}} = \bar{U}, \\ &\mathcal{V}_{\text{draw}} := \mathbb{E}\|U^{\text{draw}} - \bar{U}\|_2^2 \\ &\qquad = \lambda^2 \sum_j \left[\sum_v \widetilde{p}_j(v) \|E_j(v)\|_2^2 - \|\mu_j\|_2^2 \right]. \end{aligned} \tag{18}$$

方差恒等式仅使用边缘矩，不要求各分支之间的抽样相互独立。

对于标量网络输出 F，假设沿着从 $\bar{U}$ 到任一可能的 U^{draw} 的每条线段，$\|\nabla^2 F(U)\|_{\text{op}} \le H_F$。那么

$$\left| \mathbb{E} F(U^{\text{draw}}) - F(\bar{U}) \right| \le \frac{H_F}{2} \mathcal{V}_{\text{draw}}. \tag{19}$$

证明。 线性泰勒项的期望为零，因为$\mathbb{E}(U_{\text{抽样}} - \bar{U}) = 0$。绝对余项至多为$H_F \|U^{\text{draw}} - \bar{U}\|_2^2/2$；取期望即可得到结论。

方程（18）和（19）表明，以确定性摘要 $\bar{U}$ 替代随机抽样 U^{draw} 在一阶近似上结果一致，同时消除了由抽样方差 $\mathcal{V}_{\text{draw}}$驱动的二阶扰动偏差。在非线性后层归一化读取下，对归一化嵌入求平均与先求平均嵌入再进行归一化并不严格可交换：

$$\mathbb{E}\,\text{LN}(\kappa E_j(V_j)) \ne \text{LN}(\kappa \mu_j) \quad \text{in general.} \tag{20}$$

尽管如此，在归一化之前聚合嵌入（$\text{LN}(\kappa \mu_j)$）相当于进行方向池化，类似地抑制单次随机抽样 V_j 引入的离散方向跳变方差。

B.3. 同状态分布变化

考虑固定历史下的单个分类输出分布。设 $p = \text{softmax}(g^0/\tau)$ 和 $q = \text{softmax}(g^F/\tau)$ 分别为同一有限词汇表上的无记忆分布和启用记忆的分布，其 logits 有限，且 $\tau > 0$。定义

$$\delta_v = \frac{g^F(v) - g^0(v)}{\tau}, \qquad r = \max_v \delta_v - \min_v \delta_v. \tag{21}$$

概率比和分布界为

$$q(v) = \frac{p(v) e^{\delta_v}}{\mathbb{E}_p e^{\delta}}, \tag{22}$$

$$\text{TV}(q, p) \le \tanh(r/4), \tag{23}$$

$$\max\{\text{KL}(q\|p), \text{KL}(p\|q)\} \le r^2/8. \tag{24}$$

其中 $\mathrm{TV}(q, p) = \frac{1}{2}\sum_v |q(v) - p(v)|$，且所有对数均为自然对数。

证明。公式（22）通过对 softmax 概率相除即可得到。令 $X = e^{\delta - \min_v \delta_v} \in [1, e^r]$ 且 $Z = \mathbb{E}_p X$。对于 $r > 0$，$|X - Z|$ 的弦界给出

$$\begin{aligned}\mathrm{TV}(q,p) &= \frac{\mathbb{E}_p|X-Z|}{2Z} \\ &\le \frac{(e^r - Z)(Z-1)}{Z(e^r-1)} \le \tanh(r/4).\end{aligned} \tag{25}$$

中间表达式在 $Z = e^{r/2}$处达到最大值。对于 KL 散度，令 $\phi(s) = \log \mathbb{E}_p e^{s\delta}$。其二阶导数是指数倾斜下的方差，因此 $\phi''(s) \le r^2/4$。具体而言，KL 散度$(p\|q) = \int_0^1 (1-s)\phi''(s)\,\mathrm{d}s$ 且 KL 散度$(q\|p) = \int_0^1 s\phi''(s)\,\mathrm{d}s$，由此得到公式（24）。若 $r = 0$，则 $p = q$。

对于任意续写质量指标 $Q : \mathcal{V} \to [0,1]$，公式（22）给出

$$\mathbb{E}_q Q - \mathbb{E}_p Q = \frac{\mathrm{Cov}_p(Q, e^\delta)}{\mathbb{E}_p e^\delta}. \tag{26}$$

该协方差恒等式揭示了记忆反馈改善生成效果的确切机制：只要所诱导的 logit 偏移 δ 与高质量的连续生成结果（协方差$_p(Q,\ e^\delta) > 0$呈正相关，预期质量就会提升，从而为置信度门控和分支耦合反馈提供理论依据；同时，公式（23）和（24确保单步分布偏移始终被平滑地限制在 r以内。

C. 模型、任务与评估

C.1. 预训练生成器

我们按照 Zhou 等人对任务定义和解码术语的规定，评估了三个冻结的多模态蛋白质生成器。 [40]所评估的配置（Vanilla、PredMem、CFG，以及用于联合任务的 CFG+PredMem）及其超参数详见附录 D。

DPLM-2。DPLM-2 通过掩码扩散联合建模氨基酸序列和离散结构标记 [34]。我们使用 650M 参数的检查点`airkingbd/dplm2_650m` 及其结构标记化器 `airkingbd/struct_tokenizer`。序列标记和结构标记共享统一去噪轨迹，同时保留模态专属词汇表：序列条件以氨基酸标记的形式提供，结构条件则使用原生标记化器。PredMem 在匹配原生缩放因子后，于后置 LayerNorm 注意力输入处执行插值（附录 A.3）。

DPLM-2.1。我们使用 6.5 亿参数的二进制结构变体 airkingbd/dplm2 bit 650m -[18],-，其结构头通过 13 个二值分量表示每个标记，从而生成 $2^{13} = 8192$完整编码。Pred-Mem 在因子化输出分布下对完整编码进行排序，并通过原生结构输入投影映射每个选定的编码；因此，k 表示的是完整结构编码，而不是独立选取的比特（附录 A.4）。读取接口与 DPLM-2 使用的层归一化后注意力读取相同。

ESM3。我们使用开放的 14 亿参数 ESM3 检查点 `EvolutionaryScale/esm3-sm-open-v1` [14]。尽管 ESM3 支持多个对齐的蛋白质轨道，但我们的实验生成序列和结构，并使用其原生坐标条件接口进行逆折叠和基序支架构建。PredMem 将标记空间读取直接应用于序列和结构标记嵌入（附录 A.2），其他所有输入组件和冻结生成器参数均保持不变。

C.2. 任务和数据集

无条件协同生成。两个轨道均从完全掩码状态开始，不使用外部序列或结构条件。每次运行针对每种配置生成 500 个序列—结构对（在 $\{100, 200, 300, 400, 500\}$ 中每种长度生成 100 个），并在主要实验中同步更新两个轨道。

基序支架构建。我们评估 24 个基序变体，每次运行中每个变体生成 100 个支架，每种配置共计 2,400 个支架。基序规定固定的残基身份和结构约束，其余位置则由模型生成。DPLM变体接收标记化的基序条件，而 ESM3 使用坐标条件。固定序列位置和提供的坐标在整个解码过程中均予以保留，并直接评估解码后的基序几何结构，而不将其钳制到参考坐标。

结构预测。此单模态条件任务固定氨基酸序列并生成结构。我们在完整的 CAMEO2022 数据集（183 个目标）和 PDB 日期划分数据集（449 个目标）上进行评估，并将预测结构与参考结构直接比较。

逆折叠。此任务 [19] 固定主链结构，并在相同的两个目标数据集上生成序列。我们测量氨基酸恢复率，并对生成的每条序列分别进行重新折叠，以评估其与条件结构的一致性。在这两种单模态条件任务中，CFG 均处于禁用状态。

C.3. 指标

结构质量与自洽性。 ESMFold [23]会对所有生成的序列分别进行重新折叠。对于联合任务，scTM（TM-score [38]）和 scRMSD 用于比较重新折叠后的结构与生成的主链；对于逆折叠任务，则将其与条件结构进行比较。结构预测 TM-score 和 RMSD 则用于比较预测结构与参考结构。质量指标取所有生成输出的平均值。对于无条件协同生成和基序支架构建，pLDDT 表示独立 ESMFold 重新折叠输出的逐残基置信度均值，取值范围为 0–100（而非生成器结构解码器提供的内部置信度）。TM-score 根据完整的 Cα 迹线计算，而 RMSD 则遵循各基准的主链掩码和比对约定。

成功率和基序覆盖率。 当 scRMSD < 2 ˚A 时，无条件协同生成即视为成功，而基序支架构建还额外要求生成基序均方根偏差 < 1 ˚A。令 x^{gen} 表示生成器解码坐标，x^{EF} 表示所生成序列的 ESMFold 预测，x^{ref} 表示参考基序，$\mathcal{I}^{\text{motif}}$ 表示基序残基索引。生成基序均方根偏差为

$$R_{\text{gen}} = \text{RMSD}(x^{\text{gen}}_{\mathcal{I}^{\text{motif}}}, x^{\text{ref}}), \tag{27}$$

在拼接不连续的基序片段后，对对齐的主链原子计算得到。因此，支架构建成功指标为

$$\mathbf{1}[\text{scRMSD}(x^{\text{gen}}, x^{\text{EF}}) < 2\,\text{Å}]\,\mathbf{1}[R_{\text{gen}} < 1\,\text{Å}]. \tag{28}$$

重新折叠的 pLDDT 与 R_{gen} 一并报告，作为补充的质量诊断。若某一基序变体的 100 个支架中至少有一个成功，则将该基序变体计为已解决，从而得到至多为 24 的运行级计数。

结构多样性。 Foldseek [30]聚类在 TM-score 阈值为 0.5 时严格仅对成功子集进行计算。对于无条件协同生成，我们在每个长度内进行聚类；对于支架构建，则在每个基序变体内进行聚类，然后对一次运行中的聚类数量求和；该总和报告为 #Clusters。所有运行均使用 Foldseek 比对类型 1、覆盖模式 0 以及最小序列一致性 0。

序列恢复率。 对于长度为 L 的目标，设 s_i 和 s_i^{ref} 分别表示位置 i 上生成的氨基酸和参考氨基酸，则氨基酸恢复率为 $100L^{-1}\sum_{i=1}^{L}\mathbf{1}[s_i = s_i^{\text{ref}}]$。我们报告每个数据集上目标级恢复率的未加权（宏观）平均值。

聚合与变化。 对于五个随机种子对应的运行级数值 $y_1, \ldots, y_5$，我们报告

$$\bar{y} = \frac{1}{5}\sum_{r=1}^{5} y_r, \qquad s_y = \sqrt{\frac{1}{4}\sum_{r=1}^{5}(y_r - \bar{y})^2}. \tag{29}$$

成功率、已解决的变体数量和聚类数量均在每次运行内、聚合前计算。附录 ±中的数值E 表示运行级别的样本标准差（而非标准误或逐蛋白变化）；确定性解码配置自然会产生零运行级别变化。主表条目按显示的精度进行舍入。

D. 解码与超参数设置

按任务整理设置，使用表1、2和10中的模型和方法名称。 默认表示标准基础配置，而*Vanilla* 表示不使用记忆或引导的已配置解码器。PredMem 会禁用引导，CFG 会禁用记忆，而 CFG+PredMem 会同时启用二者；当记忆被禁用时，不适用表示参数未激活。

设 L 表示总序列长度。支持数 k表示记忆中保留的预测数量（而非 token 采样的 top-k），λ 表示记忆插值权重，而$(w_S,\ w_X)$ 表示序列和结构 CFG 强度。禁用 CFG 时使用单个条件分支，并直接使用条件 logits，这对应于方程（5）中的$w_m = 1$，而不是设置 $w_m = 0$。

D.1. 无条件协同生成

表4 列出了表 1中的四种臂。两种模态均采用重新掩码进行同步解码。在每个模型中，所有臂使用完全相同的步数预算、位置选择规则和采样温度调度。

D.2. 基序支架构建

表 5 列出了表 2中的设置。为确保在所有三种生成器和四种臂之间进行统一比较，我们的主要实验始终采用同步解码（而 Zhou 等人 [40]的参考默认方案使用 ESM3 顺序解码）。DPLM-2 和 ESM3 将步数预算设为序列总长度 L。

D.3. 结构预测

表6 列出了表10中的结构预测设置，其中序列是固定的，CFG 处于禁用状态。由于预测记忆贯穿迭代细化步骤运行，DPLM-2 和 DPLM-2.1 PredMem 采用标准的 100 步默认调度（确定性解除掩码、零温度并启用重新掩码），而 ESM3 PredMem 则采用其 8 步 Vanilla 调度。
ESM3 在

设置	臂 / 数据集	DPLM-2	DPLM-2.1	ESM3
顺序	All	同步	同步	同步
步数	All	500	500	L
解除掩码	All	随机的	随机的	随机的
重新掩码	All	启用	启用	启用
温度参数	All	$2.0 \to 0.1$	$1.1 \to 0.1$	$1.0 \to 0$
记忆（λ， k）	原始版 / CFG	（0，N/A）	（0，N/A）	（0，N/A）
记忆（λ， k）	PredMem	(0.03, 3)	(0.3, 3)	(0.3, 3)
记忆 (λ, k)	CFG+PredMem	(0.03, 3)	(0.3, 3)	(0.3, 3)
CFG $(w_S$， $w_X)$	Vanilla / PredMem	禁用	禁用	禁用
CFG $(w_S$， $w_X)$	CFG / CFG+PredMem	(2.0, 1.5)	(2。0，1。5)	(2.0, 3.0)
读取接口	PredMem / CFG+PredMem	后置层归一化注意力	后置层归一化注意力	词元嵌入

表 4。**无条件协同生成设置**。每个模型的解码、PredMem 和 CFG 超参数。温度条目列出采样器参数；由此产生的调度计划定义于第 D.5 节。

设置	臂 / 数据集	DPLM-2	DPLM-2.1	ESM3
顺序	All	同步	同步	同步
步数	All	L	500	L
解除掩码	All	随机的	随机的	随机
重新掩码	All	启用	启用	启用
温度参数	All	$0.5 \to 0$	$0.5 \to 0$	$0.7 \to 0$
记忆（λ， k）	原始版 / CFG	（0，N/A）	（0，N/A）	（0，N/A）
记忆 (λ, k)	PredMem	(0.01, 3)	(0.3, 3)	(0.3, 3)
记忆 (λ, k)	CFG+PredMem	(0.01, 3)	(0.3, 3)	(0.3, 3)
CFG $(w_S$， $w_X)$	Vanilla / PredMem	禁用	禁用	禁用
CFG（w_S， w_X）	CFG / CFG+PredMem	（1。5，1。0）	(1.5, 1.0)	(2.0, 2.0)
读取接口	PredMem / CFG+PredMem	后置层归一化注意力	后置层归一化注意力	词元嵌入

表 5。**基序- 支架设置**。与表2对应的超参数；固定条件在各臂之间共享。

CAMEO2022 没有 PredMem 条目（—），因为其在此处的基础解码仅使用单步，因此没有可供预测记忆发挥作用的迭代轨迹。为完整起见，我们还报告了原生的最优采样 Vanilla 配方（DPLM-2 使用 $L/8$ 步，DPLM-2.1 使用 8 步）。

D.4. 逆折叠

表7 列出了表10中的逆折叠设置，其中结构固定，仅生成序列。在全部三个模型中，PredMem 均与默认配置完全匹配：步数预算、解掩码策略、温度调度和重新掩码设置均保持一致（DPLM-2/2.1 采用 100 步确定性解码；ESM3 采用 8 步随机解掩码，并使用 $1.0 \to 0$退火），在 λ =0.5 的 $k = 1$ 两个数据集上均适用于 ESM3。更快的原生 Vanilla 配方（DPLM-2 为 1 步，DPLM-2.1 和 ESM3 为 8 步）单独列出。

D.5. 温度、位置选择和引导

对于 DPLM 退火采样器，

$$\tau_t = \tau_{\min} + (\tau_{\max} - \tau_{\min})(1 - t/T), \tag{30}$$

其中 $t = 0$、.....、$T - 1$，因此最后一步仍略高于名义下限端点。PredMem 保留所有原生的位置选择和重新掩码规则。

对于 ESM3 多步退火，参数 θ 转换为

$$\tau_t = \left[\max\left(\theta - \frac{t}{\max(1, T-1)}, 0.001\right)\right]^2. \tag{31}$$

因此， $\theta = 0.5$ 开始时的实际温度为 0.25；在八个步数中，该调度依次取值约为 0.25、0.12755、0.04592、0.00510，随后四个步数取值为 10^{-6}。采样温度不同于内存缓存温度，后者始终固定为 1。随机位置选择与置信度无关，随机的选择使用原生的依赖置信度的随机化规则，而

设置	臂 / 数据集	DPLM-2	DPLM-2.1	ESM3
步数	默认	100	100	1
解除掩码	默认	确定性	确定性	确定性
温度	默认	0	0	0
重新掩码	默认	启用	启用	禁用
步数	Vanilla	$L/8$	8	8
解除掩码	Vanilla	随机	确定性	随机的
温度参数	Vanilla	$0.1 \to 0$	$0.1 \to 0$	$0.5 \to 0$
重新掩码	Vanilla	启用	启用	启用
步数	PredMem	100	100	8
解除掩码	PredMem	确定性	确定性	随机的
温度参数	PredMem	0	0	$0.5 \to 0$
重新掩码	PredMem	启用	启用	启用
记忆 $(\lambda,\ k)$	PredMem / CAMEO	$(0.5, 1)$	$(0.1, 1)$	—
记忆 (λ, k)	PredMem / PDB	$(0。3,\ 1)$	$(0.03, 3)$	$(0.5, 3)$
记忆 $(\lambda,\ k)$	默认 / Vanilla	(0，N/A)	(0，N/A)	(0，N/A)
CFG	All	禁用	禁用	禁用

表 6. **结构预测设置**。 默认设置和原生 Vanilla 最优配方与 PredMem 一并列出；破折号（—）表示未评估该设置。

设置	臂 / 数据集	DPLM-2	DPLM-2.1	ESM3
步数	默认	100	100	8
解除掩码	默认	确定性	确定性	随机
温度参数	默认	0	0	$1.0 \to 0$
重新掩码	默认	启用	启用	禁用
步数	Vanilla	1	8	8
解除掩码	Vanilla	随机	随机的	确定性
温度参数	Vanilla	$0.1 \to 0$	$0.1 \to 0$	$0.5 \to 0$
重新掩码	Vanilla	启用	启用	启用
步数	PredMem	100	100	8
解除掩码	PredMem	确定性	确定性	随机
温度参数	PredMem	0	0	$1.0 \to 0$
重新掩码	PredMem	启用	启用	禁用
记忆 (λ, k)	PredMem / CAMEO	$(0.5,\ 1)$	$(0.01, 3)$	$(0.5,\ 1)$
记忆 (λ, k)	PredMem / PDB	$(0.3, 1)$	$(0.5, 1)$	$(0.5,\ 1)$
记忆 $(\lambda,\ k)$	默认 / Vanilla	(0，N/A)	(0，N/A)	(0，N/A)
CFG	All	禁用	禁用	禁用

表 7。**逆折叠设置**。 解码和记忆设置；PredMem 与默认解码计划一致，而 Vanilla 列出原生最优方案。

确定性选择或熵选择遵循原生的置信度排序。

在每个联合任务中，CFG 和 CFG+PredMem 使用相同的引导权重。无条件协同生成的引导会在条件被丢弃的分支中移除互补模态，而基序引导则采用原生的跨模态丢弃基序分支构造。记忆会在 CFG 之前写入各分支的原始对数几率，且缓存严格保持在各自分支内。在主要条件任务中，CFG 处于禁用状态，以便单独评估预测记忆的效果。

D.6. 记忆实现与计算

所有DPLM实验均采用公式（9）中的后层归一化注意力读取方式，残差流插值权重为零，而 ESM3 采用词元嵌入插值。

在联合任务中，两个生成轨道都会接收记忆；在条件任务中，只有目标轨道会接收记忆（附录A）。

每次前向传播都会通过概率加权的 top-k 嵌入均值，更新样本级、分支级和位置局部缓存。top-k 选择、归一化和聚合均以 FP32 计算，分离梯度的缓存会在读取时刻转换回原生数据类型。虽然所有位置都会缓存预测性摘要，但仅在符合条件的掩码位置读取记忆；固定残基、提供的基序坐标和特殊标记均受到保护。初始前向传播从空缓存开始。

对于 T 步和分支集合 $\mathcal{G}$，生成需要 T 次批量去噪器调用和 $T|\mathcal{G}|$ 次分支评估（无引导解码使用一个分支，联合 CFG 使用三个分支）。PredMem 在不增加任何去噪器调用（NFE）的情况下，保持完全相同的评估次数，并返回单个终端样本，无需测试时候选搜索或重排序。缓存的内存占用与 T 无关。独立的 ESMFold 重新折叠和可选的终端似然诊断属于评估步骤，不计入生成预算。

E. 详细结果

表 8和9 按照相同的模型和方法顺序扩展了表1 和2 。数值报告的是五次等权重运行的均值和 ±样本标准差。标准差为零表示记录的运行级数值完全相同（例如在确定性解码下），而不是表示单个蛋白质之间不存在变化。比率以百分比表示，RMSD 以 ˚A.

E.1. 无条件协同生成

成功率、pLDDT、scTM、scRMSD 和成功簇对应于表1。

E.2. 基序支架构建

成功标准和基序内聚类与表2相同。报告的基序 RMSD 在生成器解码的结构上测量，因为该数值直接用于成功标准。

E.3. 结构预测与逆折叠

配置与实验方案。 对于单模态条件任务，禁用引导，以便单独评估记忆反馈。我们报告 Default（发布的参考配置）、Vanilla（遵循 Zhou 等人所述的配置的采样器） [40]以及 PredMem 在 CAMEO2022（183 个目标）和 PDB日期划分（449 个目标）上的表现。如附录D所述，PredMem 与其调度匹配的多步基线配对：对于两个任务中的两个 DPLM 模型均使用 Default（对于逆折叠中的 ESM3 也使用 Default），而在结构预测中的 ESM3 使用 Vanilla。所有其余配置均保留在表 10 中，作为参考基线。结构预测从固定序列生成结构，并使用均方根偏差和 TM-score 将其与参考结构进行评估；而逆折叠则从固定主链生成序列，并衡量重新折叠后的 scTM 和天然氨基酸恢复率（AAR）。

结构预测。 与调度匹配的基线相比，PredMem 在所有接受评估的模型和数据集上均降低了均方根偏差并提高了 TM-score（表 10）。ESM3 在 CAMEO2022 上没有 PredMem 结果，因为其基础解码仅需一步，因而不存在可供记忆反馈的迭代轨迹。在 ESM3 上，PDB 均方根偏差从4.83 降至 4.27 ˚A，同时 TM-score 也优于 8 步 Vanilla 基线和 1 步默认参考值。DPLM-2 和 DPLM-2.1 在两个基准测试中同样优于各自的默认调度，表明即使具有完整的序列锚点，也能实现一致的结构细化。

逆折叠。 相较于默认调度（在此任务中为三个模型提供 PredMem 的解码调度），重新折叠后的 scTM 在所有模型和数据集上均有所提升（表 10），ESM3 在 CAMEO2022 上从 0.846 提升至0.854，在 PDB 上则从0.923 提升至 0.930 。ESM3 的 AAR 在两个基准测试中也均有所提高（CAMEO2022 上从 46.51% 提升至48.19%；PDB 上从 49.62% 提升至 51.82% ）。在 DPLM-2 和 DPLM-2.1 上，PredMem 提高了结构自洽性（scTM），同时使 AAR 与默认基线保持高度一致。

零条件结构生成。 为了比较相同目标集上的有锚点生成与无锚点生成，表 11 评估了未提供序列时仅生成结构的情况（$f = 0\%$）。移除序列锚点后，内部自一致性（ΔscTM $= +0.4565$ 在 CAMEO2022 上，以及$+0.3619$ESM3 在 PDB 上； $+0.0737$ DPLM-2.1 在 CAMEO2022 上）的增益显著大于外部参照恢复方面的增益，这证实了预测记忆的主要优势体现在无锚点的跨模态场景中。

E.4. 混合权重、支持集大小与读取接口

中等混合权重表现最佳。使用双流 PredMem 和 k=3，我们扫描插值权重 λ （图 7 左半部分）。在两种骨干模型上，无论采用无引导还是引导解码，成功率都随着 λ上升，并在默认值 λ=0.3 处达到峰值（ESM3：无引导时为 33.4% $\rightarrow$ 82.1%，引导时为 64.5% $\rightarrow$ 81.0%；DPLM-2.1：为 28.9% $\rightarrow$ 46.6% 和 32.9% $\rightarrow$ 48.3% （来自 λ=0 ）。插值过于激进会降低生成效果： λ=0.5已使 ESM3 的成功率降至 11.6 %（无引导）和 18.0%

模型	方法	成功率（%）	pLDDT	scTM	scRMSD	成功簇
DPLM-2	Vanilla	29.80 ± 2.76	82.00 ± 0.24	0.852 ± 0.0046	4.95 ± 0.21	78.0 ± 6.6
DPLM-2	PredMem	33.40 ± 1.89	82.47 ± 0.33	0.852 ± 0.0054	4.79 ± 0.21	81.0 ± 7.2
DPLM-2	CFG	31.88 ± 3.16	82.42 ± 0.29	0.859 ± 0.0049	4.42 ± 0.08	73.4 ± 6.5
DPLM-2	CFG+PredMem	34.20 ± 1.98	82.99 ± 0.44	0.862 ± 0.0063	4.37 ± 0.23	74.6 ± 3.9
DPLM-2.1 Vanilla		28.88 ± 2.12	84.91 ± 0.44	0.836 ± 0.0017	5.17 ± 0.17	92.6 ± 3.6
DPLM-2.1 PredMem		46.60 ± 1.19	87.77 ± 0.27	0.889 ± 0.0032	3.32 ± 0.18	75.6 ± 5.9
DPLM-2.1 CFG		32.88 ± 0.72	85.14 ± 0.22	0.853 ± 0.0045	4.37 ± 0.22	98.0 ± 4.9
DPLM-2.1 CFG+PredMem		48.32 ± 2.05	87.85 ± 0.18	0.895 ± 0.0054	3.13 ± 0.16	68.4 ± 4.2
ESM3	Vanilla	33.36 ± 1.08	67.91 ± 1.11	0.545 ± 0.0151	31.91 ± 0.94	62.0 ± 5.1
ESM3	PredMem	82.08 ± 1.01	90.53 ± 0.21	0.944 ± 0.0024	2.32 ± 0.18	73.0 ± 1.9
ESM3	CFG	64.48 ± 2.62	88.36 ± 0.71	0.891 ± 0.0113	6.62 ± 0.97	114.4 ± 2.3
ESM3	CFG+PredMem	80.96 ± 1.94	91.59 ± 0.22	0.943 ± 0.0022	2.47 ± 0.16	84.4 ± 0.9

表8。**无条件 协同生成结果**。与表1对应的完整结果，报告为五个随机种子均值 ± SD。

模型	方法	成功率（%）	pLDDT	scTM	scRMSD	基序 RMSD 已求解 /24	簇
DPLM-2	Vanilla	20.60 ± 0.22	72.21 ± 0.12	0.666 ± 0.0019	4.95 ± 0.11	2.54 ± 0.04	16.8 ± 0.4	37.8 ± 2.8
DPLM-2	PredMem	21.71 ± 0.62	72.23 ± 0.15	0.668 ± 0.0025	4.95 ± 0.11	2.53 ± 0.04	17.0 ± 0.5	41.0 ± 4.8
DPLM-2	CFG	20.48 ± 0.81	71.72 ± 0.17	0.661 ± 0.0028	5.03 ± 0.08	2.52 ± 0.03	16.2 ± 0.8	37.6 ± 1.8
DPLM-2	CFG+PredMem	21.25 ± 0.78	71.69 ± 0.10	0.669 ± 0.0022	4.95 ± 0.07	2.54 ± 0.02	16.4 ± 0.5	38.2 ± 1.8
DPLM-2.1 Vanilla		18.33 ± 0.80	69.86 ± 0.18	0.659 ± 0.0024	5.38 ± 0.05	2.55 ± 0.02	15.2 ± 0.8	30.2 ± 3.3
DPLM-2.1 PredMem		20.29 ± 0.49	70.01 ± 0.12	0.671 ± 0.0030	5.16 ± 0.04	2.58 ± 0.02	15.4 ± 0.5	31.2 ± 1.1
DPLM-2.1 CFG		19.07 ± 0.60	68.47 ± 0.09	0.652 ± 0.0032	5.48 ± 0.03	2.49 ± 0.03	14.8 ± 0.8	30.0 ± 2.5
DPLM-2.1 CFG+PredMem		21.08 ± 0.79	69.09 ± 0.18	0.664 ± 0.0041	5.23 ± 0.07	2.56 ± 0.21	16.4 ± 0.9	36.8 ± 4.3
ESM3	Vanilla	30.43 ± 1.49	76.25 ± 0.52	0.711 ± 0.0066	6.97 ± 0.28	0.61 ± 0.06	19.4 ± 0.5	57.4 ± 4.4
ESM3	PredMem	42.40 ± 1.49	77.89 ± 0.34	0.767 ± 0.0053	4.33 ± 0.16	0.73 ± 0.10	21.8 ± 1.3	64.8 ± 5.4
ESM3	CFG	35.81 ± 1.29	75.64 ± 0.18	0.754 ± 0.0050	4.29 ± 0.17	0.65 ± 0.04	22.6 ± 0.5	83.6 ± 6.2
ESM3	CFG+PredMem	41.93 ± 1.31	76.73 ± 0.32	0.776 ± 0.0037	3.75 ± 0.17	0.70 ± 0.11	22.0 ± 1.0	95.0 ± 6.3

表 9. **基序支架构建结果**。 与表2对应的完整结果，每次运行输出 2,400 个结果（24 个基序各采样 100 个）。成功要求 scRMSD < 2 Å 且基序 RMSD < 1 Å，其中基序 RMSD 比较生成器解码的基序与参考基序；pLDDT 是 ESMFold 在 0–100 量表上的置信度。簇统计每个基序中的成功输出数量，再对所有基序求和。

		结构预测				逆折叠			
		CAMEO2022		PDB日期划分		CAMEO2022		PDB日期划分	
模型	方法	RMSD	TM-score	RMSD	TM-score	scTM	氨基酸恢复率（%）	scTM	氨基酸恢复率（%）
DPLM-2	默认	7.48 ± 0.00	0.686 ± 0.0000	5.26 ± 0.00	0.771 ± 0.0000	0.806 ± 0.0031	48.34 ± 0.04	0.875 ± 0.0017	55.59 ± 0.09
DPLM-2	Vanilla	7.45 ± 0.07	0.687 ± 0.0025	5.26 ± 0.02	0.769 ± 0.0011	0.808 ± 0.0031	48.32 ± 0.04	0.880 ± 0.0017	55.92 ± 0.09
DPLM-2	PredMem	7.34 ± 0.00	0.692 ± 0.0000	5.15 ± 0.00	0.773 ± 0.0000	0.808 ± 0.0036	48.19 ± 0.07	0.880 ± 0.0020	55.83 ± 0.09
DPLM-2.1	默认	6.20 ± 0.07	0.747 ± 0.0011	2.79 ± 0.02	0.882 ± 0.0006	0.816 ± 0.0018	53.34 ± 0.08	0.901 ± 0.0013	78.25 ± 0.03
DPLM-2.1	Vanilla	6.13 ± 0.05	0.749 ± 0.0005	2.75 ± 0.02	0.883 ± 0.0004	0.820 ± 0.0018	53.30 ± 0.08	0.903 ± 0.0013	78.01 ± 0.03
DPLM-2.1	PredMem	6.12 ± 0.06	0.750 ± 0.0007	2.75 ± 0.02	0.883 ± 0.0004	0.819 ± 0.0022	53.33 ± 0.06	0.904 ± 0.0011	77.98 ± 0.02
ESM3	默认	5.25 ± 0.00	0.794 ± 0.0000	4.02 ± 0.00	0.840 ± 0.0000	0.846 ± 0.0027	46.51 ± 0.20	0.923 ± 0.0024	49.62 ± 0.06
ESM3	Vanilla	5.60 ± 0.10	0.804 ± 0.0017	4.83 ± 0.05	0.842 ± 0.0017	0.846 ± 0.0015	48.51 ± 0.05	0.928 ± 0.0008	51.94 ± 0.02
ESM3	PredMem	—	—	4.27 ± 0.04	0.845 ± 0.0002	0.854 ± 0.0010	48.19 ± 0.07	0.930 ± 0.0009	51.82 ± 0.01

表 10。**结构预测和逆折叠结果**。在 CAMEO2022 和 PDB日期划分上的结果。默认是发布的参考配置，Vanilla 是配置的采样器。均方根偏差单位为 Å；AAR 为百分比。各项数据为五次运行的均值 ± 样本标准差。短划线（—）表示 ESM3 在 CAMEO2022 上的结果，此时单步基础解码对 PredMem 没有影响。

（引导），而 λ=1.0使 DPLM-2.1 的成功率降至 1.2% 和 2.2%，并使引导下的 ESM3 降至 0.4%，这是因为过度反馈会覆盖冻结去噪器在未解析位点所预期的原生 MASK 表示。

成功簇。成功簇数量（底行）揭示了相同的质量与覆盖度权衡。在引导解码下，将 λ 从0（CFG）提高到 0.3，会使 ESM3 的成功簇从 114 个降至 84 个，DPLM-2.1 的成功簇从 98 个降至 68 个，同时提高成功率；权重更大时，覆盖度则会大幅下降（引导解码下的 ESM3 仅保留 2

表11。**零条件输出的自洽性和参考恢复**。在 CAMEO2022 和 PDB日期划分上的五种子结果，每个数据集和随机种子包含 20 个目标。两个指标均在相同的 $f = 0\%$ 输出上进行评估，其中 f 是已揭示序列比例（不揭示任何序列，仅生成结构）。

模型	数据集（$f = 0\%$）	$\Delta\text{TM}_{\text{ref}}$ （外部参考）	ΔscTM（内部自洽性）	配对差距 ΔscTM $-$ $\Delta\text{TM}_{\text{ref}}$ [95% CI]
DPLM-2.1	CAMEO2022	$+0.0011 \pm 0.0082$	$\mathbf{+0.0737 \pm 0.0496}$	$\mathbf{+0.0727}$ [+0.0105, +0.1348]
	PDB日期划分	-0.0168 ± 0.0246	$\mathbf{+0.0076 \pm 0.0175}$	$\mathbf{+0.0244}$ [+0.0012, +0.0476]
ESM3	CAMEO2022	$+0.0900 \pm 0.0163$	$\mathbf{+0.4565 \pm 0.0583}$	$\mathbf{+0.3665}$ [+0.3105, +0.4225]
	PDB日期划分	$+0.0818 \pm 0.0250$	$\mathbf{+0.3619 \pm 0.0709}$	$\mathbf{+0.2801}$ [+0.2138, +0.3464]

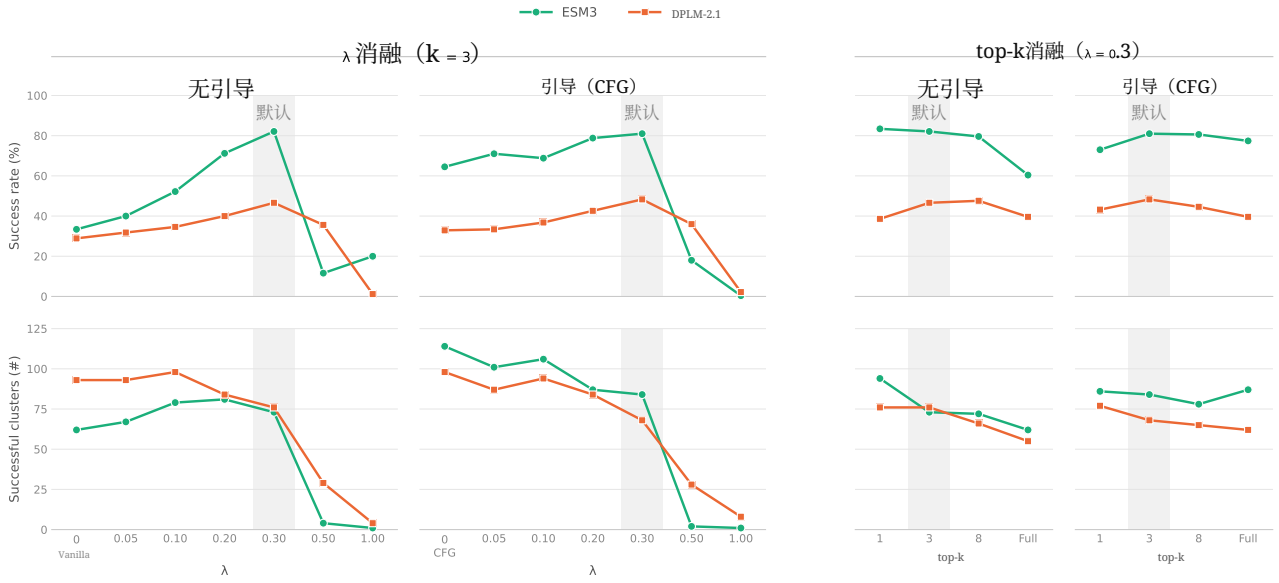

图 7。**混合权重和支持集大小消融**。左：在 k=3 下进行 λ 消融；右：在 λ=0.3 下进行 top-k 消融。顶行：成功率；底行：成功簇数量。每个面板比较 ESM3 和 DPLM-2.1 在无引导和引导（CFG）解码下的表现。阴影带表示默认设置；λ=0 对应于 Vanilla（无引导）或 CFG（引导）。

个簇（λ=0.5处）。不同骨干模型在无引导解码下表现各异：ESM3 在默认权重下的簇数量增加（从 62 个增至 73 个），因为成功的输出大幅增加；而 DPLM-2.1 的簇数量则从 93 个降至 76 个。

支持集大小。在固定 λ=0.3 的情况下，我们改变支持集大小k（图 7 右半部分）。在引导解码下，k=3在两种骨干模型上都取得了最高成功率（ESM3 上为 81.0%，DPLM-2.1 上为 48.3 %）；而扩展到完整词汇表后，成功率分别降至 77.4% 和 39.6%。在无引导的情况下，ESM3 在 $k \leq 8$ 范围内保持接近最优（k=1、3、8 时分别为 83.4%、82.1% 和 79.6%），之后在完整支持集时降至 60.4%；DPLM-2.1 则在 k=8 时达到峰值（47.6%），k=3 与峰值相差不超过 1.0 个百分点。较大的支持集也往往会降低覆盖度：随着 k 从 1 增至完整支持集，无引导时成功簇的数量在 ESM3 上从 94 降至 62，在 DPLM-2.1 上从 76 降至 55；引导时 DPLM-2.1 则从 77 降至 62（而引导时 ESM3 保持在 78–87）。由此可见，默认 k=3 在每种设置下的成功率都与最佳值相差不超过 1.3 个百分点

，k 可作为一个平滑的调节旋钮，在峰值成功率与结构广度之间进行权衡。

读取接口。表 12比较了在无引导无条件协同生成中、读取权重设置相同的情况下，将 Post-LN 读取替换为 Pre-LN 读取的结果。在 DPLM-2 上，成功率从33。40% 降至 0。44%，簇数量从 81。0 降至 1。4；在 DPLM-2.1 上，成功率从 46。60%降至 28。0%，簇数量从 75。6 降至3。0。这些结果与附录 A.3 中分析的归一化抵消效应一致。

F. 温度、熵与开销分析

本附录扩展了第5 节的内容，提供完整的温度扫描、熵下限与位置控制，以及运行时间和内存开销测量。每项分析均使用匹配基线、样本预算和随机种子集合。

表12。 **PredMem 在无引导无条件协同生成中的读取接口消融。** Post-LN（见表 8）和 Pre-LN 条目均报告五个随机种子均值 ±SD，采用相同的读取权重设置。将 Post-LN 替换为 Pre-LN 会严重降低协同生成成功率和结构覆盖度，从而证实附录 A.3 中的归一化抵消效应。

模型	读取接口	成功率（%）↑	pLDDT ↑	scTM ↑	scRMSD（Å）↓	簇 ↑
DPLM-2	Pre-LN	0.44 ± 0.36	53.35 ± 0.71	0.330 ± 0.0033	27.24 ± 0.51	1.4 ± 1.1
	Post-LN	$\mathbf{33.40 \pm 1.89}$	$\mathbf{82.47 \pm 0.33}$	$\mathbf{0.852 \pm 0.0054}$	$\mathbf{4.79 \pm 0.21}$	$\mathbf{81.0 \pm 7.2}$
DPLM-2.1	Pre-LN	28.0 ± 7.60	80.07 ± 0.96	0.871 ± 0.0022	$\mathbf{2.95 \pm 0.16}$	3.0 ± 0.0
	Post-LN	$\mathbf{46.60 \pm 1.19}$	$\mathbf{87.77 \pm 0.27}$	$\mathbf{0.889 \pm 0.0032}$	3.32 ± 0.18	$\mathbf{75.6 \pm 5.9}$

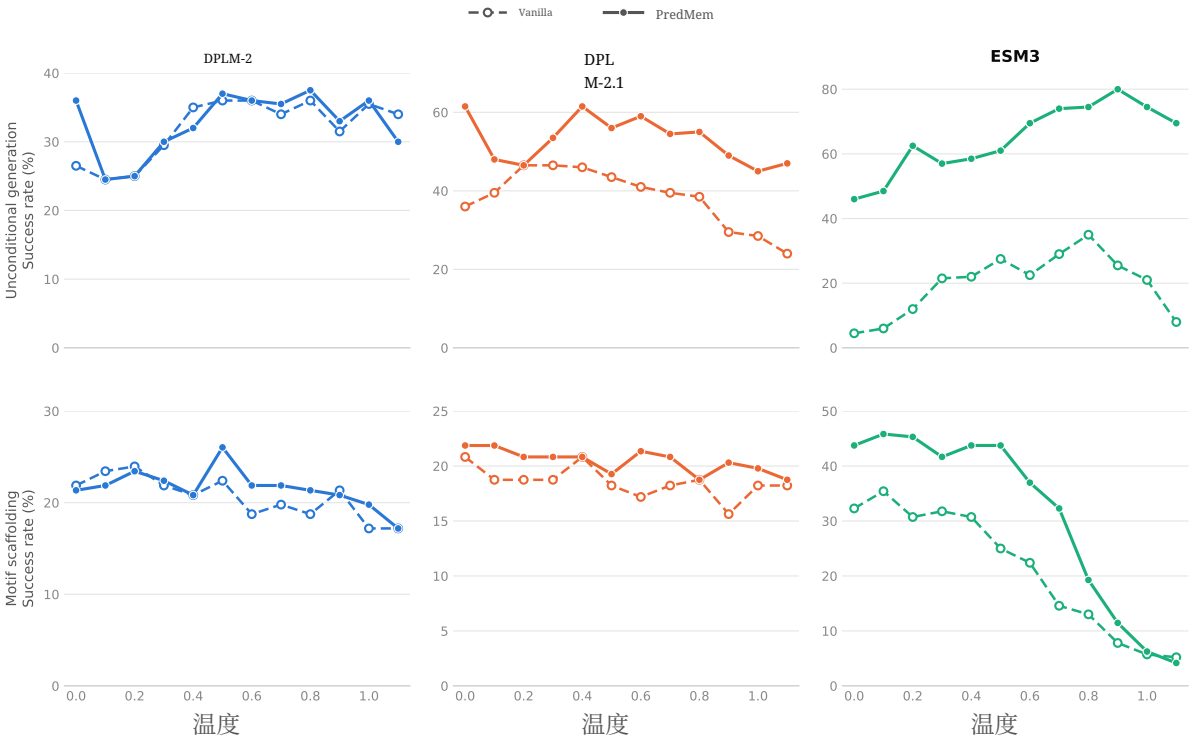

图 8. **完整温度扫描。** 上：无条件协同生成；下：基序支架构建。列：DPLM-2、DPLM- 2.1、ESM3。虚线和空心标记表示 Vanilla，实线和实心标记表示 PredMem；每种方法均在 12 个恒定温度下运行，且未使用 CFG（每个点的数据来自单次运行，包含 200 个无条件输出或 192 个基序输出）。

F.1. 完整恒温扫描

我们评估了三个模型、两项任务、两种方法，以及 12 种恒定采样温度 $\tau \in \{0.0, 0.1, \ldots, 1.1\}$（在单次运行中共计 144 种配置、 28,224 个输出，且禁用 CFG ）。无条件协同生成在每种长度下生成 40 个输出
$L \in \{100, 200, 300, 400, 500\}$（每种配置 200 个），而基序支架构建则针对 24 种变体各生成 8 个输出（每种配置 192 个）。对于每个模型–任务组合，解码顺序、重新掩码、步数预算和 PredMem 设置在整个温度网格中保持不变，并遵循表4 和 5；记忆缓存温度则保持为 1。这些恒定温度运行补充了退火的主——附录中的表格配方D。

图8 表明，PredMem 的增益在不同采样温度下均能保持，而不是依赖于单一调优后的运行点。在 ESM3 上，在无条件协同生成中，PredMem 在全部 12 个温度下均优于 Vanilla（+45.00 个百分点的峰值增益）；在基序支架构建中，PredMem 在 12 个温度中的 11 个上优于 Vanilla（+10.42 个百分点的峰值增益），且每个无条件 PredMem 数据点都超过了 Vanilla 的全局最大值。在 DPLM-2.1 上，在无条件协同生成中，PredMem 在 12 个温度中的 11 个上胜出（+15.00 个百分点的峰值增益）；在基序支架构建中，PredMem 在全部 12 个温度下均达到或超过 Vanilla（10 次胜出、2 次持平； +1.04 个百分点的峰值增益）。在 DPLM-2 上，由于无引导基线在不同温度下本已更加稳定，PredMem 将网格最大

值在无条件协同生成中提高了 +1.50 个百分点（7 次胜出、3 次持平、2 次落败），在基序支架构建中提高了 +2.08 个百分点。

F.2. 熵下限与内容控制

为排除简单的分布锐化是 PredMem 增益的主要驱动因素这一可能性，我们将其与相同状态的熵下限控制（PredMem+表 13中的下限）进行比较。在每个 PredMem 访问的硬状态下，令 g_0 和 g_M 分别表示给定位置和模态下的无记忆参考逻辑值与 PredMem 逻辑值。在原始采样温度τ下，我们定义目标熵下限

$$h^\star = \max\{H(\mathrm{softmax}(g_0/\tau)), \\ H(\mathrm{softmax}(g_M/\tau))\}, \qquad (32)$$
$$H(p) = -\sum_v p(v)\log p(v),$$

并求解每个位置的逆温度因子 $0 < \beta \leq 1$，使得 $H(\mathrm{softmax}(\beta g_M/\tau)) = h^\star$（当 PredMem 已达到或超过参考熵时保持$\beta = 1$）。控制器使用 float64 进行 48 次二分迭代，并将 float32 采样熵验证至2×10^{-4} 纳特以内，同时在保留 PredMem 的 Top-1 身份和 token 排名的情况下提高 $\tau/\beta \geq \tau$有效温度。

在无引导解码（表13）下，PredMem+下限在 ESM3 上保留了 PredMem 相对于 Vanilla 增益的约 80%（PredMem 为 72。5%，而 Vanilla 为 33。4%），在 DPLM-2.1 上保留了约 40%（*36*。0% 对 46。6%，以及 28。9%）。因此，熵减小仅解释了 ESM3 上增益的一小部分，却解释了 DPLM-2.1 上更大的一部分；与此同时，两者相较于 Vanilla 仍均有明显改进。

表13 和14提供了第5节14中总结的无引导反事实控制和同一状态温度拟合诊断的完整数值细分。在表中，即使拟合出一个最优正标量温度，使 Vanilla 在相同状态下与 PredMem 对齐，仍有 64.76%–99.56% 的 KL 散度无法解释，并且所有设置中的 Top-1 预测都发生了变化（2.57%–94.28%），这证实 PredMem 改变了词元排序，而不只是重新缩放温度。变化幅度因模型和阶段而异：DPLM-2.1 在早期最多改变 94.28% 的 Top-1 预测，ESM3 在中前期阶段最多改变 62.87%；而 DPLM-2 和 DPLM-2.1 在中前期阶段最多改变 6.38%（KL 散度低于 0.004 纳特）。

F.3. 运行时间和内存开销

表 15报告了 PredMem 相对于相同但不使用记忆的解码器，在各序列长度下的开销。无引导时，PredMem 在 ESM3 上增加 3.0–3.2% 的解码时间，在 DPLM-2.1 上增加 12.4–13.2%—— 这与

表 13. **无引导解码下的控制实验**。 五个随机种子均值± 在表 1的样本预算下，各随机种子间的标准差； Δ 各列给出每个模型内的百分点差异。Vanilla $2T$ 将步数加倍。PredMem+循环移位会按 $\lfloor n/2 \rfloor$ 残基顺序，在每个流内分别重新分配未解决位置的记忆。PredMem+下限会将 PredMem 的词元熵提升至至少等于相同状态下的无记忆参考值，同时保留词元排序。

模型	Arm	成功率（%）	Δ 对比 Vanilla	Δ 对比 PredMem
ESM3	Vanilla	33.4 ± 1.1	—	−48.7
	Vanilla $2T$	34.5 ± 1.2	+1.1	−47.6
	PredMem+移位	48.8 ± 1.2	+15.4	−33.3
	PredMem+向下取整	72.5 ± 2.0	+39.1	−9.6
	PredMem	$\mathbf{82.1 \pm 1.0}$	+48.7	—
DPLM-2.1	Vanilla	28.9 ± 2.1	—	−17.7
	Vanilla $2T$	30.4 ± 5.2	+1.5	−16.2
	PredMem+移位	35.1 ± 6.1	+6.2	−11.5
	PredMem+下取整	36.0 ± 6.0	+7.1	−10.6
	PredMem	$\mathbf{46.6 \pm 1.2}$	+17.7	—

表14。**无引导解码下的同状态温度拟合**。在相同的解码状态下，为每个位置拟合一个正标量温度 T，以最小化 KL（$p^{\text{PredMem}} \| \mathrm{softmax}(g^{\text{Vanilla}}/T)$）。对于 DPLM-2.1 的因子化13位结构头，在每个位置对全部13个二进制头拟合一个共享标量 T。在 T=1 处测量完整 KL，并报告跨位置均值；未解释比率则报告逐位置中位数残差比。正标量缩放无法产生 Top-1变化。中前期表示解码进度的大约25%。结构词表：13位（DPLM-2.1）、8192路（DPLM-2）、4096路（ESM3）。

模型	阶段 / 模态	Top-1 变化（%）	完整 KL（纳特）	未解释比例（%）
DPLM-2.1	早期 / 序列	63.74	0.1721	84.38
	早期 / 结构	94.28	0.4541	79.43
	中早期 / 序列	2.57	0.0034	64.76
	中早期 / 结构	6.38	0.0038	90.63
DPLM-2	早期 / 序列	25.45	0.0104	96.79
	早期 / 结构	29.62	0.0207	98.49
	中早期 / 序列	2.93	0.0026	78.76
	中早期 / 结构	3.04	0.0012	95.12
ESM3	早期 / 序列	8.56	0.0003	89.79
	早期 / 结构	40.60	0.0492	99.56
	中早期 / 序列	17.26	0.1380	74.15
	中早期 / 结构	62.87	0.5969	91.33

第 5 节报告的平均额外开销为：ESM3 为 3.1%，DPLM-2.1 为 12.8%——额外显存最多分别为 0.007GiB 和 0.026 GiB。在 CFG 下，ESM3 的额外开销为 3.8–4.5%，DPLM-2.1 为 8.9–13.2%，额外显存最多分别为 0.014 GiB 和 0.12 GiB（其中 0.12 GiB 的峰值是在批量大小为 10 时测得）。

表15。**PredMem 的运行时间和显存开销**。对于每个序列长度 L，时间表示解码时间的增加量，显存表示 PredMem 相对于 Vanilla（无引导）以及 CFG+PredMem 相对于 CFG（有引导）所增加的 GPU显存（以 GiB 计）。

		PredMem / Vanilla		CFG+PredMem / CFG	
模型	L	时间（%）	显存（GiB）	时间（%）	显存（GiB）
ESM3	100	3.16	0.0012	3.84	0.0023
	200	3.17	0.0024	3.85	0.0049
	300	3.03	0.0039	3.87	0.0078
	400	3.14	0.0049	4.13	0.0098
	500	3.12	0.0068	4.47	0.0137
DPLM-2.1	100	12.69	0.0224	12.40	0.0265
	200	12.36	0.0234	13.23	0.0322
	300	12.89	0.0263	12.76	0.0298
	400	13.24	0.0249	10.76	0.0462
	500	13.19	0.0243	8.87	0.0489

G. 代码可用性

代码可在 https://github.com/XL-S224/PredMem 获取