一种用于类扩散模型的高效概率硬件架构：完整翻译与深度解读

摘要与引言：直面人工智能的能源危机

原文翻译 (第1页)

一种用于类扩散模型的高效概率硬件架构

AndrazŸelincĶ,ŸOwen Lockwood, Akhil Garlapati, Guillaume Verdon, and Trevor McCourt *, †
Extropic Corporation (日期: 2025年10月29日)

概率性人工智能的激增推动了专用随机计算机的提案。尽管这些提案展现了可观的效率提升前景，但由于它们依赖于存在根本性限制的建模技术和奇特且不可扩展的硬件，未能获得广泛关注。在这项工作中，我们通过提出一种全晶体管概率计算机来解决这些缺点，该计算机在硬件层面实现了强大的去噪模型。系统级分析表明，基于我们架构的设备可以在一个简单的图像基准测试中，以大约少用10,000倍的能量，实现与GPU相当的性能。

近年来对大规模人工智能系统前所未有的投资，将很快对世界的能源基础设施构成压力。每年，美国公司在以人工智能为核心的数据中心上花费的金额，超过了经通胀调整后的阿波罗计划成本 。到2030年，这些数据中心可能会消耗全美总发电量的10% 。

尽管在扩展当今人工智能系统上押下了如此巨大的赌注，但它们在能源效率方面可能远非最佳。基于自回归大语言模型（LLMs）的现有AI系统是白领领域的宝贵工具 [4-9]，并且正以比互联网更快的速度被消费者采用 。然而，LLMs是专门为GPU设计的 ，这种硬件最初用于图形处理，其对机器学习的适用性是在几十年后被偶然发现的 。

如果在过去的几十年里流行的是一种不同风格的硬件，人工智能算法的演进方向将会完全不同，而且可能是一种更节能的方向。算法研究和硬件可用性之间的这种相互作用被称为“硬件彩票” ，它固化了那些可能远非最优的硬件-算法配对。

因此，审慎的规划要求我们系统性地探索其他类型的AI系统，以寻找节能的替代方案。当前的积极努力包括混合信号内存计算加速器 、光子神经网络 ，以及模仿生物脉冲的神经形态处理器 。

为AI开发更高效的计算机是一项挑战，因为它不仅需要在组件层面进行创新，还需要在系统层面进行创新。仅仅发明一种能在孤立状态下高效执行某种数学运算的新技术是不够的；还必须知道如何将多个组件组合起来以运行一个实用的算法。除了这些集成挑战之外，GPU的每焦耳性能每隔几年就会翻一番 ，这使得前沿计算方案很难获得主流应用。

- 这些作者对这项工作有同等贡献。 † 通讯作者: trevor@extropic.ai

深度解读

这篇论文的开篇就直击了当前人工智能领域最核心的痛点之一：**能源消耗**。对于高三的你来说，可能感觉AI是一个纯软件和算法的问题，但这篇论文告诉你，它的根基深深地扎在物理世界中，并受其限制。

概率性AI与随机计算机：首先，让我们理解两个关键概念。你熟悉的人工智能，比如ChatGPT，很大程度上是“确定性的”，即输入相同的问题，它会（大致）给出相同的逻辑推理路径。而“概率性AI”则从根本上拥抱不确定性，它不给出一个“唯一答案”，而是给出一个答案的“概率分布”。这在处理模糊、不确定的现实世界问题时非常强大。为了高效运行这种AI，科学家们提出了“随机计算机”——一种其基本运算就包含随机性的计算机。想象一下，你的电脑CPU的一个核心功能不是做加法，而是“掷骰子”。对于某些特定问题，这种基于随机性的计算方式，其效率可能远超传统计算机。这篇论文的核心，就是设计一种先进的“骰子”，并用它来构建一台全新的计算机。

能源危机与阿波罗计划的对比：作者用了一个非常震撼的类比：美国公司每年投入AI数据中心的钱，比当年把人类送上月球的阿波罗计划还要多。这不仅仅是钱的问题，更是能源的问题。预测到2030年，AI可能消耗美国10%的电力，这是一个足以影响国家能源安全和社会发展的数字。这为你揭示了一个深刻的现实：AI的进步不能仅仅依赖于更大的模型和更多的数据，否则我们很快会撞上能源这堵物理世界的“墙”。这篇论文的研究动机，正是要在这堵墙上开一扇门。

硬件的“偶然”与算法的“锁定”：这里提出了一个非常有趣且深刻的观点——“硬件彩票”。想象一下，你是一位天才厨师，但世界上唯一的炊具是一台微波炉。久而久之，你所有的食谱都会是围绕微波炉设计的，你会成为世界上最厉害的微波炉烹饪大师。但你可能永远不会想到去“炒”菜，因为你连“锅”这种工具都没见过。AI领域就发生了类似的事情。GPU（显卡）最初是为游戏设计的，擅长并行处理大量简单的图形计算。后来研究者偶然发现，这种能力恰好也非常适合训练神经网络。于是，GPU就像那台“微波炉”，赢得了“硬件彩票”。过去十几年，所有最成功的AI算法，比如驱动ChatGPT的Transformer模型，都是为GPU这碟“醋”包的“饺子”。作者大胆地提出一个问题：如果当初赢得彩票的是另一种硬件，我们今天的AI会不会是另一番完全不同、甚至更高效的景象？

系统性创新的挑战：最后，作者点出了解决这个问题的难度所在。这不像简单地发明一个更快的晶体管。你需要同时在三个层面思考：

1. **新组件**：发明一种能效极高的基础计算单元（后面会讲到，就是他们的“全晶体管随机数生成器”）。
2. **新架构**：设计一种全新的计算机体系结构，把成千上万个新组件有效地组织起来。
3. **新算法**：创造一种能够在这种新架构上高效运行的全新AI算法。 这三个层面必须紧密配合，即所谓的“软硬件协同设计”。这就像设计一辆F1赛车，你不能只想着引擎，还必须同时考虑空气动力学、轮胎和车身结构，三者必须完美契合才能达到极致性能。而他们要挑战的对手——GPU，性能还在飞速提升，这使得这场竞赛异常艰难。

总而言之，这篇论文的引言为你描绘了一幅宏大的图景：当前的AI发展路径正面临严重的能源瓶颈，而这个瓶颈很大程度上是历史偶然性的产物。要突破这个瓶颈，需要一场从最底层的物理硬件到最顶层的AI算法的系统性革命。而他们，正准备提出这场革命的蓝图。

原文翻译 (第1页 图1部分)

图1. 在超高效AI系统中利用CMOS概率硬件。 本文的核心成果：一台运行去噪热力学模型（DTM）的全晶体管概率计算机，可以在一个简单的建模基准测试中达到与GPU相当的性能，同时能耗降低约10,000倍。所有模型都在二值化的Fashion-MNIST 数据集上训练，并使用Fréchet Inception Distance (FID) 进行评估。DTM变体具有递增的深度，链接了2-8个顺序的能量基础模型（EBMs）。GPU基线涵盖了单步VAE 和GAN ，以及不同步数的DDPM 。我们还将DTM与一个在多种混合时间限制下的单体EBM进行了比较。横轴显示了使用训练好的模型生成一张新图像（推理）所需的能量。

概率计算是一种有吸引力的方法，因为它可以通过能量基础模型（EBMs）在系统层面直接与AI连接。 EBMs是当代深度学习中一个成熟的模型类别，在图像生成和机器人路径规划等任务中一直与最先进的技术保持竞争力 。EBMs的硬件实现需要使用遵循物理约束（如局部性、稀疏性和连接密度）的特殊模型家族。得益于这些约束，概率计算机可以利用专门的随机电路来高效、快速地……

表格1：AI模型性能与能耗基准测试（源自图1）

模型类别	硬件平台	每样本能耗 (焦耳/样本)	性能 (FID分数，越低越好)
VAE (变分自编码器)	GPU	~10^{-5}	0.06
GAN (生成对抗网络)	GPU	~10^{-4}	0.07
DDPM (去噪扩散概率模型)	GPU	~10^{-2} - 10^{0}	0.02 - 0.03
MEBM (单体能量基础模型)	概率计算机	~10^{-4}	0.02
DTM (去噪热力学模型) [本文]	**概率计算机**	**~10^{-8}**	**0.02 - 0.04**

深度解读

这张图表是整篇论文的“战绩公示板”，它用最直观的数据展示了作者们宣称的革命性成果。让我们像分析实验报告一样，一步步拆解它。

理解坐标轴：

- **横轴 (Energy Consumption)**：代表“成本”。具体来说，是模型生成一张新图片所消耗的能量，单位是焦耳（J）。这个坐标轴是**对数尺度**的，意味着每向左移动一个大格，能量消耗就减少到原来的十分之一。这是一个关键细节，微小的位置差异可能代表着巨大的能效差距。
- **纵轴 (Performance)**：代表“质量”。它使用一个叫做**FID (Fréchet Inception Distance)** 的指标来衡量生成图像的质量。你可以把它理解成一个专业的“图像鉴赏家”，它会对比模型生成的图片和真实的图片，给出一个“差距分”。分数越低，意味着生成的图片越逼真、质量越高。

分析图中的“玩家”：

1. **GPU阵营 (蓝色和绿色)**：

- **VAE 和 GAN**：这是两种经典的生成模型。你可以将它们理解为两种不同的“画家”。它们在GPU上运行，能耗在 10^{-5} 到 10^{-4} 焦耳的量级，但生成的图像质量一般（FID较高）。
- **DDPM (扩散模型)**：这是近年来非常火爆的模型，也是Stable Diffusion等文生图工具的技术基础。它像一个“修复大师”，从一张纯噪声图片开始，一步步地“去噪”，最终修复成一张清晰的图片。它的优点是质量非常高（FID很低），但代价是极高的能耗，因为它需要进行很多步修复，能耗可以达到 10^{-2} 甚至1焦耳。

2. **概率计算机阵营 (橙色和红色)**：
   - **MEBM (单体EBM)**：这代表了“旧”的概率计算方法。虽然它使用了新型的概率计算机硬件，但由于算法本身的缺陷（后面会详细解释），它的能耗依然很高（与GAN相当），尽管性能还不错。
   - **DTM (本文提出的模型)**：这是本文的主角。请注意它在图中的位置——它位于**左下角**。这意味着它同时实现了**极低的能耗**（约 10^{-8} 焦耳）和**顶级的性能**（FID与表现最好的DDPM相当）。

核心结论：数量级的碾压 现在，让我们进行一次关键对比。将DTM（橙色叉号）与性能相似的DDPM（绿色虚线上的点）进行比较。它们的纵坐标（性能）几乎在同一水平线上，但横坐标（能耗）却相差了**四个数量级**。10^{-8} 与 10^{-4} 的差距是**一万倍**。

这是什么概念？这意味着，当一台GPU为了生成一张高质量图片而消耗掉一块大电池的电量时，这台新型的概率计算机只用掉了其中万分之一的电量就完成了同样质量的工作。或者说，在GPU处理一张图片的时间和能耗内，新机器可以处理一万张。这已经不是简单的优化，而是质的飞跃，是不同技术范式之间的代差。

EBMs与硬件的天然联系：图表下方的文字解释了为什么概率计算这条路有希望。它提到了**能量基础模型 (EBMs)**。这是一个非常物理化的概念。你可以想象，任何一个系统（比如一堆分子）都倾向于处于能量最低的状态。EBMs就借用了这个思想，它为每一种可能的数据（比如每一张可能的图片）分配一个“能量值”。真实的、看起来合理的图片能量低，而随机的、混乱的图片能量高。模型的任务就是学习这个“能量函数”。这种模型的优美之处在于，它的形式与物理定律非常相似，因此我们可以设计出能够**物理上直接模拟**这个过程的硬件，从而实现极高的效率。这就是所谓的“在系统层面直接与AI连接”。

能量基础模型的挑战

原文翻译 (第2页)

从玻尔兹曼分布 中产生样本。根据所使用的具体硬件类型，这种采样可能是设备自然动力学的一部分 [28-32]，也可能是使用像吉布斯采样 [33-36] 这样的算法来精心安排的。使用概率硬件来加速EBMs属于热力学计算 的广义范畴。

过去在EBM加速器上的尝试在架构和硬件层面都存在问题。所有先前的提案都将EBMs用作数据分布的单体模型，而这种方式已知难以扩展 。此外，现有的设备依赖于奇特的组件，例如磁隧道结，作为产生强热噪声以用于随机数生成（RNG）的来源 。这些奇特组件尚未在商业CMOS工艺中与晶体管紧密集成，目前并不构成可扩展的解决方案 [41-43]。

在这项工作中，我们解决了这些问题，并提出了一个商业上可行的概率计算系统。我们的贡献从宏观的架构选择一直延伸到设计和制造新颖的混合信号RNG电路。

在顶层，我们引入了一种新的概率计算机架构，它运行**去噪热力学模型 (Denoising Thermodynamic Models, DTMs)** 而非单体EBMs。顾名思义，DTMs并非直接使用硬件的EBM来为数据分布建模，而是顺序地组合许多硬件EBMs来模拟一个逐步对数据进行去噪的过程。扩散模型 也遵循这种去噪程序，并且其能力远超EBMs。这一关键的架构变革解决了先前方法的一个根本性问题，并代表了将概率硬件应用于机器学习的第一个可扩展方法。

此外，我们通过实验性地展示一种快速、节能且小巧的全晶体管RNG，证明了我们的新架构可以使用当今的CMOS工艺进行大规模实现。通过仅使用晶体管作为我们RNG电路的构建模块，我们避免了在不同技术接口之间可能出现的显著且不明确的通信开销。此外，没有这种通信开销也使得我们能够在本工作中提出对设备性能的原则性预测。我们的RNG利用了亚阈值晶体管网络的随机动力学，我们最近在文献 中对此进行了详细研究。

我们的系统级分析表明，将我们的新架构与我们的全晶体管概率计算电路相结合，可以在概率建模中实现前所未有的能源效率。图1比较了这样一个系统的预测性能和能耗，与几种在GPU上运行的标准深度学习算法，以及一个传统的基于EBM的概率计算机。基于DTM的概率计算机系统在性能上与最高效的基于GPU的算法达到了同等水平，同时能耗减少了大约四个数量级。

本文的其余部分将证实图1中呈现的结果，这些结果是基于真实电路的测量、物理模型和模拟的结合。首先，我们将介绍使用EBMs作为独立数据模型时固有的一个基本妥协，我们称之为**混合-表达能力权衡 (mixing-expressivity tradeoff)**。然后我们将讨论如何通过将EBMs作为去噪过程的一部分而非单体地使用来避免这种妥协。接下来，我们将概述如何使用DTMs构建一个硬件系统，在非常低的层次上实现这个去噪过程。然后，我们研究这个硬件系统的模拟，进一步证明图1中显示的结果，并强调DTMs相比现有方法的一些实践优点。最后，我们通过讨论如何通过将概率加速器与传统神经网络合并来扩展其在机器学习中的能力来作结。

I. EBMs的挑战

机器学习的根本问题是推断出数据背后的概率分布 。一个早期的方法 是使用一个**单体EBM (monolithic EBM, MEBM)**，通过塑造一个参数化的能量函数 E 来直接拟合数据分布：

$$P(x;\theta) \propto e^{-\mathrm{E}(x,\theta)}$$

(1)

其中 x 是一个代表数据的随机变量，θ 代表EBM的参数。拟合一个MEBM相当于为数据丰富的 x 值分配低能量，为远离数据的 x 值分配高能量。真实世界的数据通常聚类成不同的模式 ，这意味着一个能很好拟合数据的MEBM将拥有一个复杂、崎岖的能量景观，有许多深邃的山谷被高耸的山脉所环绕。这种复杂性在图2(a)的示意图中得到了说明。

与我们提出的系统不同，基于MEBMs的现有概率计算机在处理真实世界数据的多模态性时会遇到困难，这妨碍了它们的效率。也就是说，如果计算机的能量景观非常崎岖，那么它为了从MEBM的分布中抽取一个样本所必须消耗的能量可能是巨大的。

具体来说，在高维空间中运行的采样算法（如吉布斯采样）是局部信息的迭代过程，这意味着它们通过基于低维信息在空间中随机进行小幅移动来对景观进行采样。当使用这样的过程从方程(1)中采样时，迭代移动到某个能量更高的状态 $X[k+1]$ 的概率，与当前状态 $X[k]$ 相比，在能量增加上是指数级小的，即：

$$P(X[k+1]=x' \mid X[k]=x) \propto e^{-(\mathrm{E}(x')-\mathrm{E}(x))}$$

(2)

深度解读

在这一部分，作者开始深入技术细节，解释为什么之前基于EBM的硬件尝试失败了，以及他们是如何从根本上解决这个问题的。

旧方法的两大缺陷：

1. **架构问题**：以前的方法试图用一个巨大而复杂的“单体EBM”（MEBM）来一次性地学习整个数据的分布。作者指出，这条路“已知难以扩展”。我们很快就会看到为什么。
2. **硬件问题**：以前的硬件依赖于“奇特组件”，比如“磁隧道结”。你可以把这理解为一种非常规的、在实验室里才能很好工作的电子元件。它们很难像我们手机里的晶体管那样，被数十亿个地集成在一块小小的芯片上。这使得它们“不构成可扩展的解决方案”。一个无法大规模生产的技术，无论在实验室里表现多好，都很难产生实际影响。

新方案的两大创新：

1. **架构创新：从“单体”到“去噪”**：这是本文最核心的算法思想。作者没有试图用一个EBM去完成整个任务，而是像扩散模型（DDPM）那样，将任务分解成一系列简单的步骤。他们不再直接对最终的清晰图片建模，而是对“如何将一张稍微有点噪声的图片变得更清晰一点”这个**过程**进行建模。他们用一连串（a chain of）简单的EBMs，每个EBM只负责一步去噪工作。这个从对“状态”建模到对“过程”建模的转变，是解决后面提到的“混合-表达能力权衡”问题的关键。
2. **硬件创新：回归主流，拥抱晶体管**：作者放弃了那些奇特的组件，回归到了最成熟、最可靠的半导体技术——CMOS晶体管。他们设计了一种“全晶体管”的随机数生成器（RNG）。这是一个非常务实的工程决策。这意味着他们的设计可以利用全球现有的、最先进的芯片制造工厂（晶圆厂）来进行生产，从而确保了方案的**可扩展性**和**商业可行性**。这让他们的性能预测变得非常可靠，因为它基于的是经过数十年验证的工业级模型，而不是实验室里的理论推测。

核心挑战：能量景观的“地形”问题 现在我们来理解EBMs的根本挑战。公式(1) $P(x;\theta) \propto e^{-\mathrm{E}(x,\theta)}$ 是核心。这里的 $P(x)$ 是数据 x 出现的概率，$\mathrm{E}(x)$ 是它的“能量”。这个公式源于物理学中的玻尔兹曼分布，它告诉我们，一个系统处于某个状态的概率与其能量成反比。能量越低，出现的概率就越高。

让我们用一个生动的比喻来理解。假设我们要为一个数据集建模，这个数据集里只有两种图片：“猫”和“狗”。

- **能量函数** $\mathsf{E}(x)$：可以想象成一个三维的地形图。
- **数据点** x：地形图上的一个(经度, 纬度)坐标。
- **能量值** $\mathsf{E}(x)$：该坐标对应的海拔高度。

为了很好地拟合数据，模型需要把所有“猫”的图片都放在一个深邃的“山谷”里（能量低=海拔低），把所有“狗”的图片放在另一个深邃的“山谷”里。而那些既不像猫也不像狗的随机噪声图片，则应该被放在高耸的“山脉”上（能量高=海拔高）。这样，我们就用一个地理上的“能量景观”来描述了数据的概率分布。

采样的困境：被困在山谷里的“盲人登山者” 模型训练好之后，我们需要用它来生成新的图片，这个过程叫做“采样”。采样就像是把一个蒙着眼睛的登山者随机空投到这个地形图的某个位置，然后让他自己去探索。这个登山者（采样算法，如吉布斯采样）非常“短视”，他每次只能探测自己脚下和周围一小步范围内的地形。他会遵循一个简单的规则：倾向于往“下坡”走（走向能量更低的地方），但也有很小的概率会往“上坡”走（走向能量更高的地方）。

现在问题来了。为了把“猫”和“狗”这两个类别分得足够清楚，模型必须在“猫之谷”和“狗之谷”之间建立起一道非常高耸的“山脉”。当我们的“盲人登山者”被空投到“猫之谷”后，他会很容易地在谷底附近探索，生成各种各样的猫。但是，因为他几乎不可能有足够的“运气”连续向上攀爬，翻越那道分隔两个山谷的巨大山脉，所以他几乎永远也到不了“狗之谷”。他被“困住”了。

这就是公式(2)所描述的困境。从状态 x（比如在山谷里）移动到能量更高的状态 x'（比如在山坡上）的概率，随着能量差 $\mathsf{E}(x') - \mathsf{E}(x)$ 的增大而**指数级减小**。山脉越高，翻越的可能性就越趋近于零。因此，一个为了精确区分数据而变得“崎岖”的能量景观，反而使得采样过程变得极其低效和困难。这就是接下来要讲的“混合-表达能力权衡”。

原文翻译 (第3页)

对于能量上的巨大差异，比如试图在被一个重要障碍隔开的两个山谷之间移动时遇到的情况，这个概率可能非常接近于零。这些障碍会使迭代采样器陷入停顿。

混合-表达能力权衡 (MET) 总结了现有概率计算机架构的这个问题，反映了对于MEBMs来说，建模性能和采样难度是耦合的。具体来说，随着一个MEBM的表达能力（建模性能）的增加，其混合时间（从MEBM的分布中抽取独立样本所需的计算量）会变得越来越长，导致昂贵的推理和不稳定的训练。

MET对基于MEBM的概率计算系统效率的经验性影响在图2(b)中得到了说明。混合时间随着性能的提升而非常迅速地增加，从而极大地增加了从模型中采样所需的能量。这种增加的混合时间的影响反映在图1中：尽管基于MEBM的解决方案使用了与基于DTM的解决方案相同的EBMs和底层硬件，但由于其极其缓慢的混合速度，其能耗要高出几个数量级。

II. 去噪热力学模型

MET清楚地表明，MEBMs存在一个使其难以扩展且能源成本高昂的缺陷。然而，这个缺陷是可以避免的，许多类型的概率机器学习模型已经被开发出来，用于解决分布建模问题，同时规避了MET。

去噪扩散模型被明确设计用来规避MET，它通过一系列简单的、易于采样的概率变换来逐步构建复杂性 。通过这样做，它们允许在给定的计算预算下表达更复杂的分布，并极大地扩展了生成模型的能力 [54-56]。

DTMs将EBMs与扩散模型相结合，为概率计算提供了一条缓解MET的替代路径。DTMs是深度学习从业者近期推动EBM性能前沿工作的一个轻微推广 [57-60]。

DTMs不是试图用单个EBM来建模数据，而是链接许多EBMs来逐步建立起数据分布的复杂性。这种复杂性的逐步建立使得链中每个EBM的景观都能保持相对简单（且易于采样），而不会限制整个链所建模的分布的复杂性；见图2(b)。

去噪模型试图逆转一个将数据分布 $Q(x^0)$ 逐渐转化为简单噪声的过程。这个**前向过程**由马尔可夫链给出：

$$Q(x^0, ..., x^T) = Q(x^0) \prod_{t=1}^{T} Q(x^t \mid x^{t-1})$$

(3)

前向过程通常被选择为具有一个独特的平稳分布 $Q(x^T)$，该分布形式简单（例如，高斯或均匀分布）。

前向过程的逆转是通过学习一组分布 $P_\theta(x^{t-1} \mid x^t)$ 来实现的，这些分布近似于方程(3)中每个条件概率的逆转。通过这样做，我们学习了一个从简单噪声到数据分布的映射，然后可以用它来生成新数据。

在传统的扩散模型中，前向过程被设计得足够精细（使用大量的步数T），以至于逆向过程中每一步的条件分布都呈现某种简单的形式（如高斯或分类分布）。

图2：混合-表达能力权衡

原文翻译

(a) 一幅卡通图，说明了EBMs中的混合-表达能力权衡。它展示了一个拟合简单数据集的能量景观的投影。 “飞机”模式与“狗”模式被很好地分开了，两者之间的数据非常少。EBM对数据拟合得越好，模式之间的能量壁垒 ΔE 往往就越大，使得从EBM中采样变得越来越困难。(b) 一个例子，展示了混合-表达能力权衡对在Fashion-MNIST数据集上测量的模型性能的影响。图中的蓝色曲线显示了在限制混合时间的情况下对MEBMs进行实验的结果。性能和混合时间密切相关。混合时间是通过将指数函数拟合到自相关函数的大延迟行为来计算的；见附录K。相比之下，一个DTM（橙色叉号）尽管采样要求低得多，却具有更高的性能。

表格2：混合-表达能力权衡的量化展示（源自图2b）

模型类型	采样难度 (混合时间，任意单位)	性能 (FID分数，越低越好)
MEBM (单体模型)	~10^4	~0.04
MEBM (单体模型)	~10^5	~0.025
MEBM (单体模型)	~10^6	~0.015
DTM (本文模型)	**远低于** 10^4	**~0.012**

深度解读

这一页正式提出了困扰传统EBM的核心矛盾——**混合-表达能力权衡 (Mixing-Expressivity Tradeoff, MET)**，并给出了解决方案的理论基础。

理解“混合-表达能力权衡” (MET) 这是一个非常关键的概念，让我们再次使用“地形图”的比喻来深入理解它。

- **表达能力 (Expressivity)**：指的是你的地形图能多精确地描绘真实的数据地貌。一个高表达能力的地形图，能把“猫之谷”和“狗之谷”的形状刻画得惟妙惟肖，并且在它们之间建立起高耸陡峭的山脉，以确保两者绝不混淆。这对应于图2(b)中更低的FID分数（更好的性能）。
- **混合时间 (Mixing Time)**：指的是你的“盲人登山者”需要走多少步，才能充分探索整个地图，而不是只被困在一个山谷里。如果登山者可以自由地在“猫之谷”和“狗之谷”之间穿梭，我们就说系统“混合得很好”，混合时间短。反之，如果他一旦掉进一个山谷就再也出不来，那混合时间就是无限长。这对应于图2(b)中的“采样难度”。

MET的核心矛盾：图2(b)的蓝色曲线完美地展示了这个矛盾。为了提升性能（纵轴向下移动），你必须让模型的能量景观变得更崎岖（增大能量壁垒 ΔE），但这不可避免地导致混合时间（横轴向右移动）**指数级增长**。你的模型越“聪明”（表达能力强），就越“固执”（难以采样）。这是一个两难的境地：要么你满足于一个模糊的、性能不佳的模型，它很容易采样；要么你追求一个精确的、高性能的模型，但你几乎无法用它来生成任何东西，因为它内部的“山脉”太高了。

图1中的MEBM点之所以能耗高，就是这个原因。尽管它使用了高效的硬件，但为了达到不错的性能，它的混合时间变得非常长，硬件必须迭代计算极多次才能得到一个合格的样本，总能耗自然就上去了。

解决方案：化整为零，分而治之 作者提出的解决方案——**去噪热力学模型 (DTMs)**，其灵感来源于近年来大获成功的扩散模型。这个思想的精髓在于“**不求一步到位，但求步步为营**”。

想象一下，你不是要一次性雕刻出一尊完美的大卫像（MEBM的做法），而是从一块纯粹的噪声大理石开始，进行一系列微小的、简单的雕刻步骤。

1. **前向过程 (Forward Process)**：这是“破坏”的过程。从一张清晰的图片（x^0）开始，我们给它加上一点点噪声，得到 x^1。再在 x^1 的基础上加一点点噪声，得到 x^2……一直重复T次，直到图片变成一片完全没有信息的随机噪声（x^T）。这个过程由公式(3)描述，它是一个马尔可夫链，意味着每一步只依赖于前一步的状态。这个过程是固定的、已知的。
2. **逆向过程 (Reverse Process)**：这是“创造”的过程，也是模型需要学习的部分。我们从纯噪声 x^T 开始，学习如何“撤销”最后一步加噪操作，得到一个稍微清晰一点的 x^{T-1}。然后，再学习如何从 x^{T-1} 得到更清晰的 x^{T-2}……一步步地，最终从纯噪声还原出一张清晰的图片 x^0。

DTM如何规避MET？ 关键在于，逆向过程中的**每一步**都是一个非常简单的任务。模型 $P_\theta(x^{t-1} \mid x^t)$ 需要学习的不是从噪声到图像的巨大鸿沟，而只是从“噪声等级为t的图像”到“噪声等级为t-1的图像”的微小变化。

回到地形图的比喻。MEBM试图学习的是从随机噪声（一片高原）直接跳到“猫之谷”或“狗之谷”的最终地形。这个地形必须非常崎岖才能保证最终结果的正确性。

而DTM则学习一系列**中间地形图**。第T步的地形图非常平缓，因为它只需要把纯噪声稍微变得有序一点点。第T-1步的地形图也相对平缓，因为它只需要在前一步的基础上再变得有序一点点。每一张中间地形图的“山脉”都不高，所以“盲人登山者”可以在其中轻松探索。虽然我们需要探索T张地图，但每张地图的探索成本都很低。

如图2(b)中的橙色叉号所示，DTM通过这种方式，以极低的采样要求（短混合时间），达到了比MEBM更优的性能。它巧妙地将一个极其困难的“全局优化”问题，分解成了一系列简单的“局部优化”问题，从而彻底绕开了MET这个根本性的障碍。

降噪热力学计算机：将算法铸成硅片

原文翻译 (第4页)

这个简单的分布由一个神经网络参数化，然后训练该网络以最小化联合分布 Q 和 P_θ 之间的 Kullback-Leibler (KL) 散度，

$$\mathrm{L}_{DN}(\theta) = D_{KL}(Q(x^0, ..., x^T) \mid\mid P_\theta(x^0, ..., x^T))$$

(4)

其中模型的联合分布是学习到的条件概率的乘积：

$$P_\theta(x^0, ..., x^T) = Q(x^T) \prod_{t=1}^{T} P_\theta(x^{t-1} \mid x^t)$$

(5)

更多细节请参见附录A.2。

基于EBM的去噪模型从一个不同的角度来解决这个问题。在许多情况下，将前向过程重写为指数形式是直接的，

$$Q(x^{t-1} \mid x^t) \propto e^{-\mathrm{E}^f_{t-1}(x^{t-1},x^t)}$$

(6)

其中 E^f_{t-1} 是与给 x^{t-1} 添加噪声的前向过程步骤相关的能量函数。然后我们使用一个具有特定能量函数的EBM来对条件概率建模，即，

$$P_\theta(x^{t-1} \mid x^t) \propto e^{-(\mathrm{E}^f_{t-1}(x^{t-1},x^t)+\mathrm{E}^\theta_{t-1}(x^{t-1},\theta))}$$

(7)

方程(7)允许在逆向过程近似的步数和每一步的采样难度之间做出权衡。随着前向过程中步数的增加，每一步加噪的影响变小，这意味着 E^f_{t-1} 更紧密地将 x^t 与 x^{t-1} 绑定在一起。这种绑定可以通过施加一个能量惩罚来简化方程(7)中给出的分布，防止其呈现强烈的多模态；进一步的讨论见附录A.4。

如图3(a)所示，方程(7)形式的模型将简单噪声重塑为数据分布的近似。在保持EBM架构不变的情况下增加T，会同时增加链的表达能力并使每一步更容易采样，从而完全绕过了MET。

为了最大限度地利用概率硬件进行EBM采样，DTMs通过引入**潜变量** $\{z^t\}$ 来推广方程(7)：

$$P_\theta(x^{t-1} \mid x^t) \propto \sum_{z^{t-1}} e^{-(\mathrm{E}^f_{t-1}(x^{t-1},x^t)+\mathrm{E}^\theta_{t-1}(x^{t-1},z^{t-1},\theta))}$$

(8)

引入潜变量使得概率模型的大小和复杂性可以独立于数据维度而增加。

DTMs的一个便利特性是，如果对逆过程条件的近似是精确的 ($P_\theta(x^{t-1} \mid x^t) \to Q(x^{t-1} \mid x^t)$)，那么人们也学习到了在 $t-1$ 时的边缘分布，

$$Q(x^{t-1}) \propto \sum_{z^{t-1}} e^{-\mathrm{E}^\theta_{t-1}(x^{t-1},z^{t-1},\theta)}$$

(9)

更多细节请参见附录A.6。请注意，此属性依赖于与方程(6)中分布相关的归一化常数独立于 x^{t-1}。

图3：去噪热力学计算机架构

原文翻译

(a) 传统扩散模型具有简单的条件概率，在近似逆过程时必须采取小步骤。由于EBMs可以表达更复杂的分布，DTMs可以潜在地采取大得多的步骤。(b) 一张草图，展示了基于DTCA的芯片如何链接硬件EBMs来近似逆过程。每个EBM由不同的电路实现，其中一部分专用于接收输入并有条件地对输出和潜变量进行采样。(c) 一个硬件EBM的抽象图。状态变量 x^t 和 x^{t-1} 映射到由蓝色和绿色节点分别代表的不同物理自由度上。这两组节点之间的耦合实现了前向过程能量函数 $\mathsf{E}_t^f(x^{t-1}, x^t)$。橙色节点集代表一组潜变量 z^{t-1}。这些节点之间以及与 x^{t-1} 节点之间的耦合实现了 $\mathsf{E}_{t-1}^\theta(z^{t-1}, x^{t-1})$。

深度解读

这一部分深入探讨了DTM的数学心脏，并首次展示了如何将这个数学模型转化为物理的芯片架构。

EBM在去噪中的新角色 首先，我们来看公式(6)和(7)，它们是理解DTM的关键。

- 公式(6) $Q(x^{t-1} \mid x^t) \propto e^{-\mathsf{E}_{t-1}^f}$：这只是用能量语言重新描述了前向（加噪）过程。它说，给定一张噪声更多的图片 x^t，得到一张噪声稍少的图片 x^{t-1} 的概率，可以用一个能量函数 E_{t-1}^f 来描述。这个能量函数是已知的、固定的。
- 公式(7) $P_\theta(x^{t-1} \mid x^t) \propto e^{-(\mathsf{E}_{t-1}^f + \mathsf{E}_{t-1}^\theta)}$：这是模型学习的逆向（去噪）过程。模型在已知的“前向能量” E_{t-1}^f 的基础上，增加了一个自己需要学习的“模型能量” E_{t-1}^θ。

这里的思想非常精妙。模型 E_{t-1}^θ 的任务不是从零开始学习整个去噪过程，而是在一个已经由 E_{t-1}^f 设定的“物理约束”下，学习一个“修正项”。E_{t-1}^f 确保了 x^{t-1} 不会离 x^t 太远，这极大地简化了学习任务。就像给你一张模糊的照片，并告诉你这张照片是通过某个特定的模糊滤镜处理得到的，让你去还原。有了滤镜的信息（相当于 E_{t-1}^f），你的还原工作（学习 E_{t-1}^θ）会容易得多。

DTM vs. 传统扩散模型：步幅的差异 图3(a)形象地展示了DTM相比传统扩散模型的一个巨大优势。传统扩散模型（如DDPM）为了保证逆向过程的每一步都足够简单（可以用一个简单的高斯分布来近似），必须把加噪过程切分得非常细碎，比如分成1000个小步骤。这就像一个想下山的人，因为害怕摔倒，只能每次挪动一小寸，虽然安全，但效率低下。

而DTM中的每一步都由一个强大的EBM来建模。EBM本身就能处理比高斯分布复杂得多的概率分布。因此，DTM可以采取**大得多**的步幅。它可能只需要8步就能完成传统模型需要1000步才能完成的去噪过程。这就像一个专业的滑雪者，可以自信地从山上大段大段地滑下，效率极高。这是DTM在算法层面实现高效的关键之一。

潜变量：模型的“想象空间” 公式(8)引入了一个至关重要的概念：**潜变量 (latent variables)** z^{t-1}。如果说数据变量 x^{t-1}（比如图像的像素）是模型需要处理的“现实世界”，那么潜变量 z^{t-1} 就是模型为了处理现实而开辟出的一个高维度的“内部思考空间”或“草稿纸”。

为什么需要潜变量？因为现实世界的变量（像素）之间可能存在非常复杂、非局部的关联。比如，识别一只猫，你需要同时理解它的眼睛、耳朵、胡须之间的空间关系。一个只在像素层面进行局部连接的模型很难捕捉到这种抽象的“猫”的概念。

通过引入大量的潜变量（橙色节点），模型可以在这个“想象空间”里进行复杂的计算。这些潜变量节点之间可以相互连接，形成复杂的网络，学习代表各种抽象特征（比如“毛茸茸的质感”、“尖尖的耳朵”等）。然后，这些学到的抽象特征再作用于数据变量（像素节点），从而生成一个整体上协调、真实的图像。

公式(8)中的 $\sum_{z^{t-1}}$ 符号（对所有潜变量求和）意味着，我们最终关心的只是像素 x^{t-1} 的概率，而潜变量 z^{t-1} 只是中间过程，在计算完成后就被“积分掉”了。它们是脚手架，帮助我们建好大楼，但大楼建成后脚手架就可以拆掉。

从数学到芯片：DTCA架构 图3(b)和(c)展示了如何将上述数学模型变成一块真实的芯片，即**去噪热力气计算机架构 (DTCA)**。

- **链式结构 (图3b)**：芯片上可以物理地实现多个EBM模块，像流水线一样串联起来。数据从一个模块流向下一个模块，每经过一个模块，就完成一步去噪。
- **单个EBM模块 (图3c)**：
  - **蓝色节点 (x^t)**：代表输入的、噪声更多的图像。在采样过程中，它们的值是**固定不变**的，作为“条件”或“引导”。
  - **绿色节点 (x^{t-1})**：代表输出的、噪声更少的图像。它们是需要被采样的变量。
  - **蓝色和绿色节点间的连接**：物理上实现了已知的“前向能量” E^f_{t-1}。这种连接非常简单，通常只是对应像素之间的一对一连接。
  - **橙色节点 (z^{t-1})**：代表潜变量，是模型的“计算核心”。
  - **橙色与绿色节点间的连接**：物理上实现了需要学习的“模型能量” E^θ_{t-1}。这些连接构成了复杂的网络，是模型表达能力的主要来源。

这个架构的精妙之处在于，它将一个复杂的算法完美地映射到了一个物理结构上。算法中的每一个变量、每一个能量项，都在芯片上有了对应的物理实体。这种“算法即硬件”的设计，是实现极致能效的根本原因。

原文翻译 (第5页)

III. 去噪热力学计算机

去噪热力学计算机架构 (DTCA) 将DTMs紧密集成到概率硬件中，从而实现了对EBM辅助的扩散模型的高效实现。

DTCA的实际实现利用了那些表现出稀疏和局部连接性的、易于实现的EBMs，这在文献中是典型的做法 。这一约束使得EBM的采样可以通过大规模并行的原始电路阵列来执行，这些电路实现了吉布斯采样。关于硬件架构的进一步理论讨论，请参考附录B和C。

DTCA的一个关键特性是 E^f_{t-1} 可以使用我们受约束的EBMs高效实现。具体来说，对于连续和离散扩散，E^f_{t-1} 都可以通过 x^t 和 x^{t-1} 中对应变量之间的单个成对相互作用来实现；详见附录A.1和C.1。这种结构可以反映在芯片的布局方式上，以在不违反局部性约束的情况下实现这些相互作用。

至关重要的是，方程(8)对 E^θ_{t-1} 的形式没有施加任何约束。因此，我们可以自由地使用我们的硬件实现起来特别高效的EBMs。在最低层次上，这对应于高维、结构规整的潜变量EBM。如果需要更强大的模型，这些硬件潜变量EBMs可以通过将它们组合成软件定义的图形模型来任意扩展。

DTMs的模块化特性使得各种硬件实现成为可能。例如，链中的每个EBM都可以使用同一芯片上的不同物理电路来实现，如图3(b)所示。或者，各种EBMs可以分布在多个通信的芯片上，或者由同一个

硬件在不同时间用不同的权重集重新编程来实现。

对于链中的任何一个给定的EBM，数据变量 x^t、x^{t-1} 和潜变量 z^{t-1} 都被物理地体现在采样电路中，这些电路以一种简单的方式连接，反映了方程(7)的结构。这种变量结构在图3(c)中被示意性地展示。

为了理解未来硬件设备的性能，我们开发了一个DTCA的GPU模拟器，并用它在Fashion-MNIST数据集上训练了一个DTM。我们使用FID来衡量DTM的性能，并利用一个物理模型来估计生成新图像所需的能量。这些数字可以与传统的算法/硬件配对进行比较，例如在GPU上运行的VAE；这些结果显示在图1中。

图1中产生结果的DTM使用了**玻尔兹曼机 (Boltzmann machine)** EBMs。玻尔兹曼机，在物理学中也称为伊辛模型 (Ising models)，使用二元随机变量，是最简单的离散变量EBM类型。玻尔兹曼机在硬件上是高效的，因为从它们中采样所需的吉布斯采样更新规则很简单。玻尔兹曼机实现的能量函数形式为

$$\mathrm{E}(x) = -\beta\Big(\sum_{i\neq j} x_i J_{ij} x_j + \sum_i h_i x_i\Big)$$

(10)

其中每个 $x_i \in \{-1, 1\}$。从相应EBM中采样的吉布斯采样更新规则是

$$P(X_i[k+1] = +1 \mid X_{\neg i}[k] = x) = \sigma\Big(2\beta\Big(\sum_{j\neq i} J_{ij} x_j + h_i\Big)\Big)$$

(11)

这可以通过一个适当偏置的随机比特源简单地评估。

使用玻尔兹曼机实现我们提出的硬件架构特别简单。一个设备将由一个规则的伯努利采样电路网格组成，其中每个采样电路实现单个变量 x_i 的吉布斯采样更新。采样电路的偏置（产生1而不是-1的概率）被约束为一个输入电压的S型函数（sigmoid function），这使得方程(11)中的条件更新可以使用一个简单的电路（如电阻网络）来实现电流相加（见附录D.1）。

具体来说，本工作中使用的EBMs是稀疏的、深的玻尔兹曼机，由 $L \times L$ 的二元变量网格组成，其中在大多数情况下使用 $L = 70$。每个变量根据一个简单的模式连接到它的几个（大多数情况下是12个）邻居。随机地，一些变量被选择来代表数据 x_{t-1}，其余的被分配给潜变量 z_{t-1}。然后，一个额外的节点被连接到每个数据节点，以实现与 x_t 的耦合。关于玻尔兹曼机架构的更多细节，见附录C。

由于我们选择的连接模式，我们的玻尔兹曼机是二分的（bipartite，即可二着色）。由于每个色块可以并行采样，单次吉布斯采样迭代对应于在第二个色块的条件下对第一个色块进行采样，然后反之亦然。从某个随机初始化开始，这个块采样过程可以重复K次迭代（其中K长于采样器的混合时间，通常 $K \approx 1000$），以便为逆过程近似中的每一步从方程(7)中抽取样本。

为了实现DTCA的近期、大规模实现，我们利用了亚阈值晶体管的散粒噪声动力学 来构建一个快速、节能且小巧的RNG。我们的全晶体管RNG是可编程的，并且对控制电压具有期望的S型响应，实验测量结果如图4(a)所示。从RNG输出的随机电压信号具有一个近似指数衰减的自相关函数，在约100 ns内衰减，如图4(b)所示。如文献 所示，这个时间约束远大于我们晶体管中噪声相关时间所施加的下限。因此，通过改进设计，RNG可以变得快得多。附录J提供了关于我们RNG的更多细节。

我们的全晶体管RNG的一个实践优势是，可以使用由代工厂提供的详细且经过验证的模型来研究制造变化对我们电路设计的影响。

深度解读

这一部分将抽象的架构蓝图与具体的电子实现连接起来，解释了他们选择的“积木”——玻尔兹曼机，以及制造这些积木的核心技术——全晶体管随机数生成器。

玻尔兹曼机：AI模型的“磁铁阵列” 作者选择了一种最简单、最经典的EBM——**玻尔兹曼机**。对于学物理的同学来说，它就是**伊辛模型 (Ising Model)**。你可以把它想象成一个由微小的、只能朝上（自旋+1）或朝下（自旋-1）的磁铁组成的网格。

- **变量** x_i：代表第 i 个磁铁的朝向（+1或-1）。
- **权重** J_{ij}：代表磁铁 i 和磁铁 j 之间的相互作用强度。如果 J_{ij} 是正的，它们倾向于朝向相同；如果是负的，则倾向于朝向相反。
- **偏置** h_i：代表一个外部磁场，它会使第 i 个磁铁倾向于朝上或朝下。

公式(10)就是这个磁铁阵列的总能量。系统的总能量取决于所有磁铁的朝向以及它们之间的相互作用。

吉布斯采样：简单高效的“翻转规则” 那么，如何对这个系统进行采样呢？答案是**吉布斯采样**，其更新规则在公式(11)中给出。这个规则非常简单直观：

1. **选中一个磁铁** x_i。
2. **看看它的所有邻居** x_j **的当前朝向**。
3. **计算一个“合力”**：将每个邻居的朝向 x_j 乘以它与 x_i 的连接强度 J_{ij}，然后全部加起来，最后再加上外部磁场 h_i 的影响。
4. **根据这个“合力”的大小和方向，以一定的概率决定是否要翻转磁铁** x_i。这个概率由S型函数 σ 给出。如果合力非常强地指向“上”，那么 x_i 就有很大概率翻转到+1；反之亦然。

这个过程的硬件实现效率极高。每个变量（磁铁）只需要一个简单的电路，该电路能接收来自邻居的信号（电流或电压），将它们加权求和，然后用这个和来控制一个随机源，最后输出一个+1或-1。通过在芯片上成千上万次地并行重复这个简单的“翻转”操作，整个系统就能自然地演化到符合其能量函数定义的概率分布，从而完成采样。

硬件实现：DTCA的物理基础

- **稀疏和局部连接**：为了让硬件实现变得容易，模型被设计成每个节点只与它附近少数几个节点连接。这就像在城市里铺设网络，如果每家每户都只需要连接到邻近的几家，而不是直接连接

到城市另一端的每一家，那么布线成本和通信延迟会大大降低。

- **二分图与并行计算**：他们设计的连接模式还有一个巧妙的特性，即“二分性”。你可以把所有节点涂成两种颜色（比如黑色和白色），使得任意两个直接相连的节点颜色都不同。这有什么好处呢？这意味着所有白色节点之间没有直接连接，所有黑色节点之间也没有。因此，我们可以**同时更新所有白色节点**（因为它们的决策只依赖于黑色邻居），然后再**同时更新所有黑色节点**。这种“块采样”方式极大地提高了计算的并行度，是硬件加速的关键。
- **全晶体管RNG：核心技术突破**：这是将理论变为现实的最关键一步。如前所述，作者没有使用实验室里的“屠龙之技”，而是用最常见的晶体管，巧妙地利用其在“亚阈值”工作区（一个非常低功耗的状态）时表现出的内在随机性（散粒噪声），构建了一个微型、快速、节能的随机数生成器（RNG）。这个RNG就是实现吉布斯采样中“概率性翻转”决策的物理核心。它是一个可编程的“电子硬币”，你可以通过一个控制电压来精确地控制它出现正面（+1）或反面（-1）的概率。

制造可行性：作者特别强调，他们可以使用芯片代工厂提供的标准设计工具（PDK）来模拟和预测这种RNG在真实制造过程中的表现。这意味着他们已经考虑到了芯片制造中不可避免的微小瑕疵和变化，并证明了他们的设计足够“鲁棒”，可以在大规模生产中可靠地工作。这展示了从学术研究到工业应用的严谨思考过程。

原文翻译 (第6页)

电路设计。在图4(c)中，我们使用这个工艺开发套件(PDK)来研究我们的RNG的速度和能耗随系统性晶圆间晶体管参数偏移（工艺角）以及单个芯片内预期变化函数的变化情况。我们发现，尽管存在这些非理想因素，RNG仍能可靠工作，这意味着它可以很容易地扩展到DTCA所需的大规模网格中。

图1中给出的概率计算机的能量估算是使用一个全晶体管玻尔兹曼机吉布斯采样器的物理模型构建的。该模型的主要贡献由以下公式捕获：

$$E = TK_{mix}L^2E_{cell}$$

(12)

$$E_{cell} = E_{rng} + E_{bias} + E_{clock} + E_{comm}$$

(13)

其中 E_{rng} 来自图4(c)中的数据。E_{bias} 项是使用一个可能的偏置电路的物理模型估算的，而 E_{clock} 和 E_{comm} 则是根据对时钟和单元间通信成本的物理推理得出的。K_{mix} 是为推理所需、使链充分混合的采样迭代次数，这通常少于训练期间使用的迭代次数。对于DTM，使用了 $K_{mix} = 250$（见附录D.4），而对于MEBM，则使用了图2中测量的混合时间。这个模型是近似的，但它捕捉了真实设备的基本物理原理，并为实际能耗提供了一个合理的数量级估计。通常，给定我们用于RNG的相同晶体管工艺和对模型其他自由参数的一些合理选择，我们可以估计 $E_{cell} \approx 2$ fJ。关于这个模型的详尽推导见附录D。

我们使用一个简单的模型来估算GPU的能耗，该模型低估了实际值。我们计算从训练好的模型生成一个样本所需的总浮点运算次数（FLOPs），然后将其除以制造商给出的FLOP/焦耳规格。进一步的讨论见附录E。

图4：一个可编程的随机比特源

原文翻译

(a) 我们RNG工作特性的实验室测量。输出电压信号处于高电平状态 $(x = 1)$ 的概率可以通过改变输入电压来编程。$\mathrm{P}(x = 1)$ 与输入电压之间的关系可以很好地用一个S型函数来近似。插图显示了不同输入电压下输出电压信号随时间的变化。(b) RNG在无偏点 $(\mathrm{P}(x = 1) = 0.5)$ 时的自相关函数。衰减近似为指数形式，速率为 $\tau_0 \approx 100 ns$。(c) 估计制造变化对RNG性能的影响。图中的每个点代表一个RNG电路的模拟结果，其晶体管参数是根据制造商PDK定义的程序采样的。每种颜色代表一个不同的工艺角，每个工艺角都模拟了约200个RNG的实现。“典型”角代表一个平衡的情况，而其他两个是不对称角，其中两种类型的晶体管（NMOS和PMOS）向相反的方向偏移。慢速NMOS和快速PMOS的情况对我们来说性能最差，这是由于我们设计中的不对称性。

表格3a：RNG工作特性（源自图4a）

控制电压 (任意单位)	输出为1的概率, $\mathrm{P}(x = 1)$
20	~0.0
30	~0.1
40	~0.5 (无偏点)
50	~0.9
60	~1.0

表格3b：RNG的速度与“记忆”（源自图4b）

延迟时间 (单位: $\tau_0 \approx 100$ ns)	归一化自相关
0	1.0
1	~0.37
2	~0.14
3	~0.05

表格3c：RNG对制造变化的鲁棒性（源自图4c）

制造工艺角	平均能耗 (aJ)	平均相关时间 (ns)
快速NMOS, 慢速PMOS	~300	~75
典型 (Typical)	~350	~100
慢速NMOS, 快速PMOS	~500	~120

深度解读

这一部分是论文的“实验验证”核心，它展示了将理论变为现实的关键硬件组件——随机数生成器（RNG）的真实性能，并给出了一个简洁的物理模型来估算整个系统的能耗。

图4：一个强大RNG的三个关键品质 图4的三个子图分别回答了关于这个RNG的三个关键问题，证明了它的实用性。

1. **它是否可控？(图4a)**：答案是肯定的。这个图表显示，通过改变一个“控制电压”，我们可以精确地控制RNG输出“1”的概率。这个关系曲线呈现出完美的S型（Sigmoid），这正是实现玻尔兹曼机更新规则（公式11）所需要的。这证明了RNG可以作为吉布斯采样算法的完美物理基础。
2. **它是否够快？(图4b)**：答案是肯定的。这个图叫做“自相关函数”，它衡量的是RNG在某个时刻的输出与其在一段时间之后的输出有多大关系。你可以看到，这个相关性以大约100纳秒（ns）的时间尺度指数级衰减。这意味着，RNG的“记忆”非常短暂，大约几百纳秒后，它就完全“忘记”了自己之前的状态，可以产生一个全新的、独立的随机比特。这个速度决定了整个计算机的时钟频率，100纳秒的量级意味着它可以达到兆赫兹（MHz）甚至更高的运行速度。
3. **它是否可靠（可制造）？(图4c)**：答案是肯定的。这是工程上最重要的问题。芯片制造不是完美的，同一块晶圆上不同位置的晶体管性能会有微小差异（片内变化），不同批次的晶圆之间也会有系统性差异（工艺角）。图4c通过模拟不同的“工艺角”（比如晶体管偏快或偏慢的极端情况）来测试RNG的鲁棒性。结果显示，即使在最坏的情况下，RNG的能耗和速度虽然有变化，但它依然能够可靠地工作。这给了作者极大的信心，相信这种设计可以被成功地大规模制造出来。

能耗模型：洞察系统效率的来源 公式(12)和(13)提供了一个非常清晰的“能耗预算”模型，让我们能看清整个系统的能量花在了哪里，以及DTM架构的优势所在。 $E = T \cdot K_{mix} \cdot L^2 \cdot E_{cell}$

- E_{cell} **(单细胞能量)**：这是最底层的物理成本，代表芯片上的一个基本计算单元（一个“磁铁”）进行一次翻转决策所消耗的能量。它包括RNG本身的能耗、给RNG提供偏置信号的能耗、时钟同步的能耗以及和邻居通信的能耗。作者估计这个值大约是2飞焦耳（fJ），这是一个极低的数字。
- L^2 **(模型大小)**：代表模型中有多少个这样的基本单元。$L = 70$ 的模型就有 $70 \times 70 = 4900$ 个单元。
- K_{mix} **(混合迭代次数)**：这是**最关键的参数**。它代表为了得到一个有效的样本，每个单元需要进行多少次采样迭代。对于MEBM，由于MET的存在，为了获得好的性能，K_{mix} 可能需要达到数百万甚至更高。而对于DTM，由于每一步都很简单，作者发现 $K_{mix} = 250$ 就足够了。
- T **(去噪步数)**：DTM需要执行T步去噪，而MEBM只需要一步（$T = 1$）。

DTM的胜利之道：现在我们可以清晰地看到DTM是如何赢得这场能效竞赛的。虽然DTM的 T 大于1（比如 $T = 8$），但它在 K_{mix} 这个参数上获得了巨大的优势。假设一个MEBM为了达到同等性能需要 $K_{mix} = 500,000$。

- MEBM能耗 $\propto 1 \cdot 500,000 \cdot L^2 \cdot E_{cell}$
- DTM能耗 $\propto 8 \cdot 250 \cdot L^2 \cdot E_{cell} = 2000 \cdot L^2 \cdot E_{cell}$

通过这个简单的计算，你可以看到，DTM的能耗大约是MEBM的 $2000/500,000 = 1/250$。DTM用增加少量外部步骤（T）的代价，换来了内部核心循环（K_{mix}）效率成百上千倍的提升。这是一种典型的“以架构换效率”的策略，也是整个论文的核心思想。

训练的艺术：通过自适应控制实现稳定

原文翻译 (第6-7页)

IV. 训练DTMs

图1中实验所用的EBMs是通过将标准的蒙特卡洛估计器应用于EBMs的梯度 来训练的，作用于方程(4)，得到：

$$\nabla_\theta \mathrm{L}_{DN}(\theta) = \sum_{t=1}^{T} \mathrm{E}_{Q(x_{t-1},x_t)}[\mathrm{E}_{P_\theta(z_{t-1}\,|\,x_{t-1},x_t)}[\nabla_\theta \mathrm{E}^m_{t-1}] - \mathrm{E}_{P_\theta(x_{t-1},z_{t-1}\,|\,x_t)}[\nabla_\theta \mathrm{E}^m_{t-1}]]$$

(14)

值得注意的是，对t的求和中的每一项都可以独立计算。为了估计方程(14)中的任意一项，首先，从前向过程 $Q(x_{t-1}, x_t)$ 中采样元组 (x_{t-1}, x_t)。然后，对于每个元组，将逆过程EBM适当地钳位到采样值上，并使用吉布斯采样的K次迭代的时间平均来估计内部的期望值。将结果在所有元组上平均，即可得到所需的梯度估计。

需要指出的是，DTCA允许我们的EBMs具有有限且短的混合时间，这使得能够使用足够的采样迭代来获得梯度的几乎无偏估计。由于MEBMs的混合时间很长，在大多数情况下无法获得无偏的梯度估

计 。

一个训练良好的去噪模型通过从噪声中逐步地“拉出”新样本来生成与训练数据相似的样本；一个在Fashion-MNIST数据集上训练的8步去噪模型的输出如图5(a)所示。在最后的时间T，图像是随机比特。随着链的进行，结构开始出现，最终在时间 $t = 0$ 时得到清晰的图像。

DTMs缓解了MEBMs固有的训练不稳定性。MEBMs的参数通常使用一种能产生易于采样的能量景观的策略来初始化 。因此，在训练的早期阶段，从方程(1)中采样是可能的，使用方程(14)产生的梯度估计是无偏的。然而，随着这些梯度的跟进，MEBM被重塑以适应数据分布，并开始变得复杂和多模态。这种诱导的多模态性极大地增加了分布的采样复杂性，导致样本偏离平衡状态。使用非平衡样本计算出的梯度不一定指向有意义的方向，这可能会中止甚至逆转训练过程。

MEBMs中的这种不稳定性导致了不可预测的训练动态，可能对实现细节很敏感。几种不同类型模型的训练动态示例显示在图5(b)中。上图显示了训练期间生成图像的质量，而下图显示了采样器混合程度的度量。图像质量使用FID指标衡量，混合质量使用归一化自相关来衡量：

$$r_{yy}[k] = \frac{\mathrm{E}[(y[j] - \mu)(y[j+k] - \mu)]}{\mathrm{E}[(y[j] - \mu)^2]}$$

(15)

$$\mu = \mathrm{E}(y[j])$$

(16)

其中k是延迟时间；$y[j]$ 是采样链数据在迭代j时的某个低维投影，$y[j] = f(x[j])$; E表示在独立的吉布斯采样链上取的期望值。图5(b)的下图显示了在延迟等于训练期间用于估计梯度的总采样迭代次数时的自相关。通常，如果 r_{yy} 接近1，梯度是使用远离平衡的样本估计的，质量可能很低。如果它接近于零，样本应该接近平衡，并产生高质量的梯度估计。更多讨论见附录H。

非平衡采样的不稳定效应从图5(b)的蓝色曲线中显而易见。在训练开始时，质量和 r_{yy} 都在增加，表明数据的多模态性正在被印刻到模型上。然后 r_{yy} 变得如此之大，以至于梯度的质量开始下降，导致平台期，并最终导致模型质量的恶化。

仅仅是去噪就显著稳定了训练。因为每一层执行的转换更简单，模型必须学习的分布不那么复杂，因此更容易采样。图5(b)中的橙色曲线显示了一个典型去噪模型的训练动态。自相关性...

图5：Fashion-MNIST数据集上的详细结果

原文翻译

(a) 由一个去噪模型生成的图像。在这里，为了获得更好看的图像，将几个二元变量组合起来表示一个灰度像素。灰度级别的噪声是我们嵌入方法的人为产物；见附录G。(b) 一个实验，显示DTMs比MEBMs训练更稳定。用ACP补充DTMs可以完全稳定训练。对于DTMs，显示了所有层中最大的

$r_{yy}[K]$ 值。(c) 扩展EBM复杂性对DTM性能的影响。修改网格大小L以改变潜变量相对于（固定的）数据变量的数量。通常，具有更多连接性和更长允许混合时间的EBM层可以利用更多的潜变量，因此能获得更高的性能。

表格4：训练动态与稳定性对比（源自图5b）

训练轮次 (Epoch)	模型类型	性能 (FID，越低越好)	采样器混合质量 (自相关 $r_{yy}[K]$)
0-20	MEBM	从0.04降至0.02	从0.1升至0.6
20-60	MEBM	在0.02附近停滞后恶化至0.03	持续上升至接近1.0
0-60	DTM	从0.04稳步降至0.015	缓慢上升至约0.4
60-80	DTM	性能开始恶化	持续上升至0.6
0-80	**DTM+ACP**	**从0.04稳步降至0.012**	**始终保持在0.1以下**

表格5：模型扩展对性能的影响（源自图5c）

上图：性能 vs. 潜变量比例 (固定混合时间)

潜变量比例	图连接度	性能 (FID)
~0.1	G_8	~0.025
~0.3	G_8	~0.035 (性能下降)
~0.1	G_{12}	~0.018
~0.3	G_{12}	~0.02 (性能略降)
~0.1	G_{16}	~0.015
~0.3	G_{16}	~0.014 (性能提升)

下图：性能 vs. 混合时间 (固定连接度)

混合迭代次数 (K)	潜变量比例	性能 (FID)
400	~0.1	~0.025

混合迭代次数 (K)	潜变量比例	性能 (FID)
800	~0.1	~0.018
1200	~0.1	~0.015
400	~0.3	~0.022
800	~0.3	~0.016
1200	~0.3	~0.013

深度解读

这一部分揭示了另一个深刻的挑战：**如何有效地训练这些模型**。拥有一个好的模型架构和硬件只是第一步，如果不能稳定地教会它学习，一切都是空谈。

训练的困境：学得越多，错得越离谱 首先，我们来理解为什么训练MEBM如此困难。训练的过程，本质上是根据数据来调整模型的能量景观，即调整公式(10)中的 J_{ij} 和 h_i。这个调整的依据是“梯度”，它告诉模型参数应该朝哪个方向改变才能更好地拟合数据。

然而，计算这个梯度需要从模型当前的分布中进行采样（公式14）。这里就出现了一个恶性循环：

1. **训练初期**：模型很简单，能量景观平缓，采样很容易，梯度计算准确。模型开始学习，性能提升。
2. **学习过程中**：为了拟合复杂的数据，模型的能量景观变得越来越崎岖（MET开始发作）。
3. **采样失效**：采样器（“盲人登山者”）开始被困在某个山谷里，无法充分探索整个能量景观。它带回来的样本是有偏的，不能代表整个分布。
4. **梯度错误**：基于有偏的样本计算出的梯度是错误的，它会给模型一个错误的方向指导。
5. **训练崩溃**：模型根据错误的指导更新自己，结果性能不但不提升，反而开始下降。

图5(b)中的蓝色曲线（MEBM）完美地展示了这个过程。性能（上图）一开始变好，但同时混合质量（下图的自相关 r_{yy}）迅速恶化，当自相关接近1时（意味着采样器完全被困住了），性能就开始崩溃。橙色曲线（DTM）因为每一步任务更简单，这个崩溃的过程被推迟了，但最终还是会发生。

自适应相关惩罚(ACP)：给模型装上“学习监视器” 为了解决这个问题，作者引入了本文的第二个核心算法创新：**自适应相关惩罚 (Adaptive Correlation Penalty, ACP)**。这是一个非常聪明的“闭环控制”思想。

想象一下你是一位老师，正在教一个学生。你不仅要给他出题（损失函数），还要时刻观察他是否“学懵了”（采样混合程度）。

- **自相关 r_{yy}**：这就是你观察学生是否“学懵了”的指标。如果 r_{yy} 很低，说明学生思路清晰，学得很好。如果 r_{yy} 很高，说明他已经钻进牛角尖，快要崩溃了。
- **相关惩罚 λ_t**：这是你的“教学工具”。当学生学懵了，你就把当前的学习任务简化一点（增大惩罚项 λ_t 的权重）。这个惩罚项的作用是“鼓励”模型的能量景观变得更平滑、更容易采样。
- **自适应/闭环控制**：你（ACP算法）会持续监控 r_{yy}。如果 r_{yy} 开始升高并超过一个阈值，你就自动增大 λ_t 来“拉他一把”；如果 r_{yy} 保持在很低的水平，你就可以适当减小 λ_t，让他去挑战更复杂的知识。

图5(b)中的绿色曲线（DTM+ACP）展示了这种方法的惊人效果。在整个训练过程中，自相关（下图）始终被牢牢地压在一个很低的水平。结果是，模型的性能（上图）可以持续、单调地提升，没有任何停滞或崩溃。ACP就像一个经验丰富的教练，确保模型在整个学习过程中始终保持在“最佳学习区”，既不会因为任务太简单而停滞不前，也不会因为任务太难而学不下去。这是一种计算上的“动态平衡”或“稳态”，确保了训练过程的稳定和高效。

模型扩展的艺术：更大不一定更好 图5(c)揭示了设计这些模型时需要考虑的微妙之处，它告诉我们一个重要的工程原则：**简单地堆砌资源并不总是有效**。

- **上图**：在固定的计算预算（混合时间K）下，如果我们只是增加模型的“宽度”（增加潜变量数量），但不增加其“连接性”（信息流动的通路），性能反而可能下降。这就像一个公司，招了很多新员工（潜变量），但没有建立有效的沟通渠道（连接性），结果是人浮于事，效率反而降低了。只有在增加员工的同时，也改善沟通结构，才能提升整体表现。
- **下图**：展示了计算预算的重要性。对于一个给定大小和连接性的模型，给它更多的计算时间（更大的K），它就能学得更好。这说明，模型的潜能需要足够的计算资源才能被激发出来。

总而言之，这一部分不仅展示了如何稳定地训练DTM，还揭示了设计高效模型需要一种系统性的思维，必须在模型大小、内部结构和可用计算资源之间找到一个精妙的平衡点。

原文翻译 (第8页)

和性能比MEBM保持得好得多。

随着训练的进行，DTM最终会变得不稳定，这可以归因于潜变量之间形成了复杂的能量景观。为了解决这个问题，我们修改了训练过程，以惩罚那些混合效果差的模型。我们在损失函数中增加了一项，将优化推向一个易于采样的分布，即：

$$\mathrm{L}_t^{TC} = \mathrm{E}_{Q(x^t)}$$

(17)

其中 $s^{t-1} = (x^{t-1}, z^{t-1})$ 并且 s_i^{t-1} 表示 s^{t-1} 中M个变量的第i个。这一项惩罚了学习到的条件分布与一个具有相同边缘分布的因子化分布之间的距离，是一种总相关惩罚 。

总损失函数是方程(4)和这个总相关惩罚的总和：

$$L = L_{DN} + \sum_{t=1}^{T} \lambda_t L_t^{TC}$$

(18)

参数 λ_t 控制了逆过程中每一步总相关惩罚的相对强度。

我们使用一种**自适应相关惩罚 (Adaptive Correlation Penalty, ACP)** 来设置 λ_t，使其大到足以对每一层都保持采样易处理。在训练期间，我们周期性地测量每个学习到的条件概率在等于梯度估计期间使用的采样迭代次数的延迟下的自相关。如果第j层的自相关接近于零，λ_j 就减小，反之亦然。

我们对相关惩罚强度的闭环控制至关重要，它允许我们在保持训练稳定的同时，最大化EBMs的表达能力。图5(b)中的绿色曲线显示了在这种闭环控制策略下的训练动态示例。模型质量单调增加，并且在整个训练过程中自相关保持很小。在训练本文中用于产生结果的大多数模型时，都采用了这种对相关惩罚的闭环控制，包括图1中所示的模型。

通常，DTMs的性能随着其规模的增加而提高。如图1所示，将DTM的深度从2增加到8，显著提高了生成图像的质量。如图5(c)所示，增加链中EBMs的宽度、度和允许的混合时间通常也能提高性能。

然而，一些微妙之处阻止了这种特定的EBM拓扑结构被无限扩展。图5(c)的上图显示，扩展潜变量的数量（在固定的允许混合时间内）只有在图的连接性也相应扩展时才能提高性能；否则，性能可能会下降。这种依赖性是合理的，因为以这种方式增加潜变量的数量会增加玻尔兹曼机的深度，而这已知会使采样更加困难。超过某一点后，增加模型表达复杂能量景观的能力可能会使其在允许的混合时间 $K \approx 1000$ 的情况下无法学习。

同样的效果也显示在图5(c)的下图中，它表明需要更大的K值来支持更宽的模型，同时保持连接性不变。

总的来说，期望一个硬件高效的EBM拓扑结构能够通过独立扩展来为任意复杂的数据集建模是天真的。例如，没有充分的理由表明，一个从布线角度看很方便的连接模式，也同样适合表示复杂现实世界数据集的相关结构。

V. 结论：扩展热力学机器学习

现代机器学习的核心信条是通过不懈地扩展模型来解决越来越难的问题。利用概率计算机的模型也可以类似地进行扩展，以增强其能力，超越本工作中迄今为止考虑的相对简单的数据集。

然而，我们假设，扩展概率机器学习硬件系统的正确方法不是孤立地进行，而是作为更大的**混合热力学-确定性机器学习 (HTDML)** 系统中的一个组件。这样的混合系统将概率硬件与更传统的机器学习加速器集成在一起。

混合方法是明智的，因为没有先验的理由相信概率计算机应该处理机器学习问题的每一个部分，有时确定性处理器可能是更好的工具。HTDML的目标是设计实用的机器学习系统，在特定任务上以最小的能量达到期望的建模保真度。这种效率将通过跨学科的努力来实现，该努力摒弃了软件/硬件的抽象障碍，以设计尊重物理约束的计算机。

数学上，HTDML的图景可以概括为

$$\Xi_{tot}(S, D, p) = E_{det}(S, D, p) + E_{prob}(S, D, p)$$

(19)

其中 E_{tot} 是某个机器学习系统S评估某个数据集D的模型并达到性能p时消耗的总能量。E_{tot} 分解为一个来自确定性计算机的能量项 E_{det} 和一个来自概率计算机的能量项 E_{prob}。

到目前为止，只有HTDML的极端情况被探索过。

深度解读

这一部分从具体的训练技术转向了对未来的宏大展望，提出了一个非常成熟和务实的愿景：**混合计算**。

对“唯规模论”的反思 在进入结论之前，作者对图5(c)的结果做了一个深刻的总结：**天真地认为一种硬件友好的模型结构可以无限扩展来解决所有问题是错误的**。这是一个非常重要的自省。他们承认，目前为了布线方便而设计的简单网格连接结构，不一定能很好地捕捉真实世界数据（如图像）中复杂的、非局部的关联。这为他们的最终结论——混合系统——埋下了伏笔。它表明，纯粹的概率硬件可能擅长某些事（如高效的随机采样），但在处理其他任务（如从原始像素中提取结构化特征）时可能不是最佳选择。

结论：超越孤立，走向融合 (HTDML) 作者没有像一个狂热的推销员那样宣称他们的概率计算机将取代一切，而是提出了一个更具智慧的观点：**未来的AI系统应该是混合的**。

混合热力学-确定性机器学习 (HTDML) 这个概念的核心思想是：**让专业的人做专业的事**。

- **确定性计算机 (如GPU)**：它们非常擅长执行精确的、确定性的计算，比如卷积、矩阵乘法等。这些操作在处理和编码感官数据（如图像、声音）的早期阶段非常有效。
- **概率计算机 (本文提出的)**：它们非常擅长从复杂的概率分布中进行高效采样，这在生成新数据、进行推理和探索可能性空间时具有无与伦比的能效优势。

HTDML系统就像一个**精英团队**：

- **GPU/传统AI芯片**：扮演“**感知与分析专家**”的角色。它负责接收原始的、混乱的现实世界数据（比如一张彩色照片），并将其处理、编码成一种更抽象、更结构化的“语言”（比如一组二进制代码）。
- **概率计算机**：扮演“**创意与推理引擎**”的角色。它接收由“分析专家”编码好的信息，然后在自己的概率模型空间中进行高效的探索和生成，创造出新的、符合逻辑的、多样化的内容。

公式(19) $E_{tot} = E_{det} + E_{prob}$ 优雅地总结了这个思想。一个任务的总能耗，是确定性部分和概率性部分能耗的总和。到目前为止，整个AI领域几乎只探索了两个极端：

1. **纯确定性系统** ($E_{prob} = 0$)：这就是我们今天所知的几乎所有主流AI系统，它们完全在GPU等确定性硬件上运行。

2. **纯概率性系统** ($E_{det} \approx 0$)：这就是本文前面部分主要展示的，试图用概率计算机完成从头到尾的所有任务。

作者的洞见在于，**最优解几乎肯定位于这两个极端的中间**。通过合理地划分任务，将每个子任务分配给最高效的硬件来执行，整个系统的总能耗可以达到最低。这是一种系统工程的思维方式，它超越了单一技术或算法的局限，从全局最优的角度来设计未来的AI系统。这不仅是一个技术上更优的方案，也是一个商业上更可行的路径，因为它不是要颠覆现有的庞大生态，而是要与之融合，成为其中一个强大而高效的新组件。

原文翻译 (第9页)

图6. 使用神经网络将数据嵌入到DTM中。这里，我们展示了使用一个简单的嵌入模型与DTM结合的结果。DTM被训练来生成CIFAR-10图像，并用一个比传统GAN小约10倍的确定性神经网络达到了性能上的同等水平。

现有的机器学习工作体系有 $E_{tot} = E_{det}$，而本工作中的早期演示则有 $E_{tot} \approx E_{prob}$。像许多工程系统一样，最优解将在中间的某个地方找到，即各个子系统的贡献几乎达到平衡 [65-67]。

极端之间的系统设计代表了完全未被探索的领域。HTDML中的许多基础问题仍有待解决。

例如，需要开发更严谨的方法将数据嵌入到硬件EBMs中，以超越这里考虑的相对简单的数据集。确实，二值化在一般情况下是不可行的，而在概率硬件层面嵌入到更丰富的变量类型（如分类变量）中，效率不是特别高，也不是很有原则性。

解决嵌入问题的一种方法是使用一个小型神经网络将数据映射到概率硬件中。一个展示这一点的初步实验如图6所示。在这里，我们训练了一个小型神经网络，将CIFAR-10数据集 嵌入到一个二元DTM中。

嵌入网络使用自编码器损失进行训练以对数据进行二值化，然后用这些二值化数据来训练DTM。嵌入网络的解码器随后使用GAN目标进行进一步训练，以提高生成图像的质量。这个训练过程在附录I中有更详细的描述。

尽管存在嵌入神经网络的开销，但这个原始的混合模型是高效的。如图所示，传统GAN的生成器需要比我们嵌入网络的解码器大约大10倍，才能达到混合模型的性能。

图6中采用的嵌入过程很可能通过进一步的研究得到显著改进。我们方法的一个主要缺陷是自编码器和DTM没有联合训练，这意味着自编码器学到的嵌入可能不适合信息在DTM中给定其有限连接性的流动方式。在潜空间中使用EBMs的去噪链的问题已经在深度学习文献中被研究过，其中一些工作可能被用来解决这里讨论的嵌入问题 。

本文中使用的模型与即使是基于DTCA的早期概率计算机所能实现的相比，也显得很小。根据我们RNG的大小，可以估计约 10^6 个采样单元可以装入一个 6×6 微米的芯片中（见附录J）。相比之下，图1中显示的最大DTM仅使用了大约50,000个单元。

鉴于我们的模型大小与潜在硬件设备能力之间的差距，一个自然要研究的问题是，这些概率模型如何能在本文考虑的明显方法之外进行扩展。这种扩展可能对应于开发融合多个EBMs以实现逆过程中

每一步的架构。一种可能的方法是构建EBMs的软件定义图形模型，以实现非局部信息路由，这可以缓解一些与固定和局部交互结构相关的问题。

HTDML研究的一个困难是，在GPU上模拟大型硬件EBMs可能是一项具有挑战性的任务。GPU运行这些EBMs的效率远低于概率计算机，而且在处理硬件EBMs时自然出现的稀疏数据结构与常规的张量数据类型不兼容。我们对这些挑战有短期和长期的解决方案。

为了在短期内解决这些挑战，我们开源了一个软件库 ，它能够实现对硬件EBMs的XLA加速 模拟。这个库是用JAX 编写的，并自动化了实现硬件EBM采样的复杂切片操作。我们还提供了额外的代码，包装这个库以实现本文中介绍的具体实验 。

在更长远的未来，使用先进的晶体管工艺 [73-75] 实现大规模的概率计算机，例如本文提出的那种，将显著缓解与HTDML研究相关的挑战，并加速进展的步伐。

表格6：混合DTM-神经网络模型的效率（源自图6）

模型类型	确定性参数数量 (生成器/解码器)	性能 (FID分数，越低越好)
传统 GAN	~50,000	~0.017
传统 GAN	~200,000	~0.014
传统 GAN	~600,000	~0.012
混合 DTM+NN (本文)	**~60,000**	**~0.012**

深度解读

这一页是论文的收尾，它不仅用一个初步的实验支撑了HTDML的愿景，还非常务实地指出了未来的研究方向和挑战，并给出了推动社区发展的具体行动。

HTDML的初步验证 图6和对应的表格是HTDML思想的“概念验证（Proof-of-Concept）”。实验的设置是：

1. **编码器 (Encoder)**：一个小型神经网络，负责将CIFAR-10（一个比MNIST更复杂的彩色图像数据集）的图片“翻译”成DTM能理解的二进制编码。
2. **DTM**：在这些二进制编码上进行训练和生成。
3. **解码器 (Decoder)**：另一个小型神经网络，负责将DTM生成的二进制编码“翻译”回彩色的图像。

实验结果非常惊人。如表格6所示，一个传统的GAN模型需要大约600,000个参数才能达到的性能水平，而这个混合系统的解码器部分只需要大约60,000个参数就能做到。这意味着，通过将核心的、

最困难的“生成”任务交给能效极高的DTM来完成，我们可以**极大地减小**对昂贵的确定性计算资源（即大型神经网络）的依赖。这为HTDML的能效优势提供了强有力的初步证据。

未来的挑战与机遇 作者坦诚地指出了当前研究的局限和未来的方向：

1. **嵌入问题**：如何将真实世界丰富的数据（如彩色图像、声音、语言）高效地“嵌入”到结构相对简单的硬件EBM中，这是一个核心的开放问题。目前的简单二值化方法显然是不够的。
2. **联合训练**：图6的实验中，编码器和DTM是分开训练的，这并非最优。未来的研究需要探索如何将它们联合训练，使得编码器能学习到一种特别适合DTM内部信息流动的“语言”。
3. **扩展模型**：当前的硬件设计潜力巨大。一个 6×6 毫米的芯片就能容纳一百万个计算单元，而目前的实验只用到了五万个。如何设计出能够有效利用这百万单元的、更大更强的概率模型架构，是一个激动人心的研究方向。作者提出了一个可能的方向：“软件定义的图形模型”，即用软件灵活地组织和连接芯片上的多个EBM模块，以实现更强大的信息处理能力。

社区建设：开源的力量 最后，作者指出了一个非常现实的困境：研究一种**尚不存在**的硬件是非常困难的，因为研究者们无法在上面进行实验。这个问题被称为“鸡生蛋还是蛋生鸡”的困境：没有硬件，就没有人为它开发软件；没有软件，就没有理由去投入巨资制造硬件。

作者的解决方案是**开源一个高效的软件模拟库**。这个库可以在现有的GPU上模拟他们提出的概率计算机的行为。这是一个极具战略眼光的举动，旨在：

- **降低研究门槛**：让全世界的研究者都能在自己的电脑上实验和探索DTM和HTDML的思想。
- **构建生态系统**：激发社区的智慧，开发出更多、更好的、为这种新型硬件量身定制的算法和应用。
- **打破“硬件彩票”的诅咒**：通过建立一个活跃的软件和算法社区，为未来真正制造出这种硬件提供强有力的“需求”和“理由”，从而主动地引导硬件发展的方向。

这篇论文以一个宏大的能源危机问题开篇，通过一系列深刻的算法创新和严谨的硬件工程，提出了一个革命性的解决方案，并最终以一个开放、协作的姿态，为整个领域指明了一个激动人心且切实可行的未来方向。它完美地展示了从基础物理、到算法理论、再到硬件工程和社区生态建设的完整科研图景。