PinchBot：具有引导扩散策略的长时间可变形操控

Alison Bartsch [1] , Arvind Car [1] , Amir Barati Farimani [1]

***Abstract*—陶器制作是一种复杂的艺术形式，需要灵巧、精确和细致的动作，才能将一块黏土慢慢变形成一个有意义且往往实用的三维目标形状。在这项工作中，我们的目标是创建一个仅通过捏合动作即可实现简单陶器目标的机器人系统。这个捏合陶器任务使我们能够探索具有高度多模态和长期性变形操控任务的挑战。为此，我们提出了 PinchBot，一种以目标为条件的扩散策略模型，当其与预训练的三维点云嵌入、任务进度预测和碰撞约束动作投影相结合时，能够成功创建各种简单的陶器目标。有关实验视频和演示数据集的访问，请访问我们的项目网站：https://sites.google.com/andrew.cmu.edu/pinchbot/hom**
。

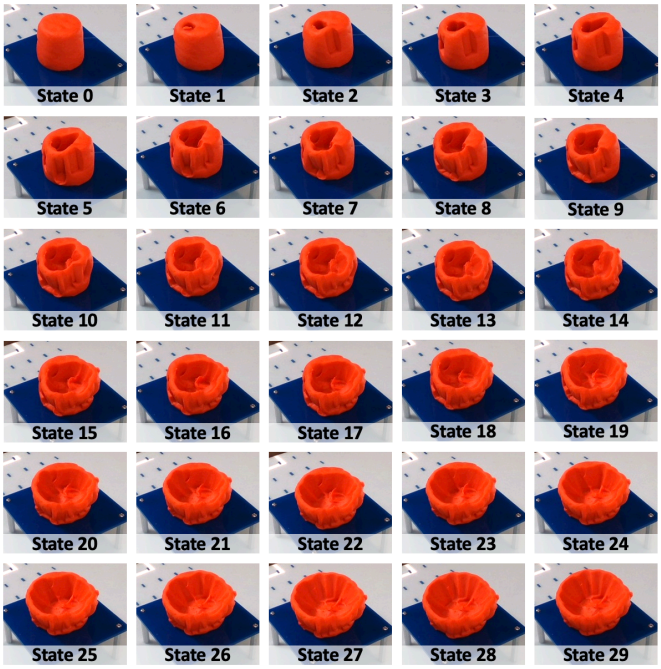

Fig. 1. 我们训练的 PinchBot 策略的完整陶艺雕塑序列。

I. 简介

陶艺创作是一种复杂的艺术形式，需要一系列精确的灵巧且细致的动作，慢慢地将一块黏土从一个语义上无意义的团块变成有意义且常常有用的三维结构。在这项工作中，我们特别关注基于徒手捏制的陶艺，其中艺术家使用手而非工具或陶轮来创作物品。这种基于捏制的陶艺任务是一项非常困难和长周期的可变形操控任务，让我们能够探索和开发能够可靠地与可变形物体互动的机器人系统。由于状态表示、自动遮挡、来自交互的复杂状态变化以及操控任务本身通常具有长周期性质等开放性挑战，可变形物体被认为难以操控。现有的可变形造型工作倾向于关注较短的“长周期”任务，如在黏土中创造字母形状，或者擀面、折叠和切割面团。尽管这些目标很难，但它们通常只需要不到十次的操作即可从初始黏土形状中创造出来。这使得对于规划框架来说任务要简单得多，因为需要优化的轨迹在相对较短的时间范围内。同样地，这简化了策略学习的挑战，因为策略必须学习的动作序列要短得多。在这项工作中，我们认为捏制陶艺任务是真正的长周期任务，将进一步测试我们的系统处理这些复杂长周期目标的能力。此外，捏陶任务并没有明确的正确操作顺序。许多相同操作的组合将产生目标形状，但并非所有这些操作的组合都能生成目标形状。轨迹的高度多模态性加上任务的长时间跨度性质，将允许我们评估和基准测试扩散策略模型变体 [1]–[3] 在完成任务方面的表现。

在这项工作中，我们提出了 PinchBot，一种目标条件的扩散策略变体，它采用了预训练的 3D 点云嵌入模型、任务进度预测和碰撞约束投影，仅通过捏合动作成功创建了各种陶碗。我们提出，这个捏合陶器任务是一个真正的长视野可变形操控任务，可以让我们更好地评估学习框架。本文的主要贡献如下：

- 我们提供了一个公开的数据集，该数据集包含通过示教式教学在 Franka 机器人上收集的人类演示，用于捏制陶器任务。
- 我们提出了一种单一目标条件的策略，该策略能够调整输出的雕刻序列，以成功创建一系列最终的陶艺目标。
- 我们探讨了任务指引、预训练和点云嵌入模型中的选择如何影响这些长时间跨度任务的扩散策略性能。
- 我们展示了如何通过对策略轨迹进行碰撞约束的动作投影来提高策略的性能和一致性。

[1] A. Bartsch, A. Car, and A. B. Farimani are with the Department of Mechanical Engineering at Carnegie Mellon University, Pittsburgh, PA, 15213, USA (e-mail: abartsch@andrew.cmu.edu, acar@andrew.cmu.edu, barati@cmu.edu)

II. 相关工作

基于点云的行为克隆：在操作应用中，先前的研究表明点云提高了行为克隆系统的性能。在 [2]–[5] 中，研究人员展示了基于点云的扩散策略 [1] 的优势。此外，[4] 展示了点云状态观察在简单操作任务中提升了动作分块 Transformer [6] 的效果。在 [2] 中，研究人员结合了大型 Transformer 模型 PointBERT [7] ，该模型在 ShapeNet 重建任务 [8] 上进行了预训练，并用于粘土塑形任务中设定目标。在 [5] 中，研究人员利用了来自 [3] 的 PointNet [9] 模型的轻量化修改版本用于基于轮的陶艺任务。虽然这些方法展示了点云条件扩散策略在粘土操作中的有效性，但在这项工作中，我们旨在探讨点云模型的选择、预训练和目标设定如何影响在更具挑战性的捏制陶艺任务上的表现。

粘土雕塑：已经有许多工作致力于开发机器人系统来实现粘土雕塑的目标。在 [10] 中，研究人员训练了一个动力学模型，以预测基于捏合动作的变形动力学。在后续工作中，他们扩展了这一框架以结合工具 [11] 。还有其他方法开发了利用预训练的点云嵌入进行动力学建模 [12] ，或者处理拓扑状态变化 [13] 。除了动力学建模之外，模仿学习在一组简单形状上显示出潜力 [2] 。最近，研究人员探索

www.xueshuxiangzi.com

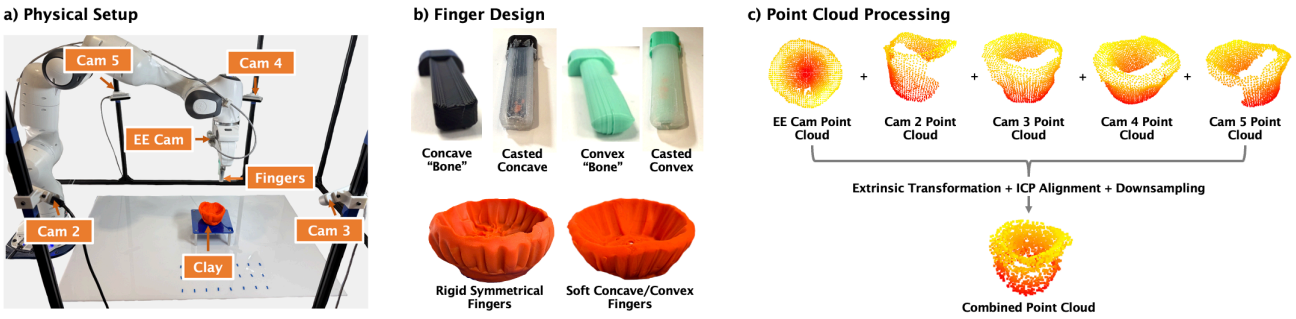

Fig. 2. 硬件设置。a) 我们实验的工作空间，由 4 个 Intel RealSense D415 相机和一个抬高的舞台组成，粘土放置在这个舞台上。b) 用于捏制泥壶任务的凹形和凸形手指设计。两个手指均由 3D 打印的凹形或凸形“骨骼”组成，外包覆着 EcoFlex 制成的软“皮肤”。这种设计选择有助于软化粘土中的凹痕，从而创造出更光滑的表面。c) 点云处理流程，在每个相机画面中，通过色彩阈值分割来隔离粘土，然后将点云投影到机器人的坐标框架中以合并。

了利用大型语言模型进行直接动作规划 [14] ，或用于子目标生成 [15], [16] 。虽然这些方法中的每一个都在探索粘土塑形这一非常具有挑战性的任务，但它们创造的目标形状大约需要 10 次或更少的动作。在这项工作中，我们认为，评估诸如捏陶这样的长时间任务的方法可以更好地突出方法之间的性能差异。在 [5] 中，研究人员成功创建了一个基于轮子的陶艺策略。尽管这是一个更长时间的任务，但由于粘土正在旋转，策略本身的复杂性比捏陶简单得多。0DoF 终端执行器所做的状态变化将对粘土表面在该半径和高度上对称应用。而捏陶任务包含了更多的空间和旋转推理挑战。

III. 捏壶任务

在这项工作中，我们正在探索基于捏合的陶器制作这一具有挑战性的长期任务。在本节中，我们将定义捏合陶器的任务以及动作和轨迹的表示（第 ?? 节），展示硬件设计选择（第 ?? 节），描述完整的 3D 视觉点云处理系统（第 ?? 节），并讨论示范数据收集过程（第 ?? 节）。

我们定义旋钮陶瓷任务为：在场景中放置一个初始为不同尺寸的陶瓷筒，并给出目标陶瓷形状的 3D 点云，执行一系列的捏合操作，直到达到目标陶瓷形状。我们将每个捏合操作定义为末端执行器的位置和方向、指尖之间的最终距离，以及一个终止参数 γ 用于识别每个轨迹中的最终操作（$a_{pinch} = [x, y, z, R_x, R_y, R_z, d_{ee}, \gamma]$ ）。在实际操作中，所有的非最终操作都有 $\gamma = -1$ ，而最终操作有 $\gamma = 1$ 。对于每个捏合陶瓷轨迹，在每次捏合操作之前和之后都会捕捉到一个 3D 点云。因此，轨迹是一个点云状态和 8D 操作对的序列。

为了确保由我们机器人系统制作的捏制陶罐的最终质量，我们在设计指尖时格外注意，以减少机器人捏制成型中常见的严重压痕线。指尖设计的完整可视化展示如图 2 b 所示。我们设计的两个关键方面是：1) 使用凹凸设计，每次挤压时创建一个曲面；2) 使用带有柔软 EcoFlex “皮肤”的 3D 打印“骨架”以减少严重的压痕。“骨架”和“皮肤”分别在每个手指上是凹或凸的形状。非对称手指的设计选择提高了陶器的视觉质量，但确实大幅增加了旋转动作空间，因为夹持器不再被视为对称的。实际上，这意味着我们必须考虑 360° 的完全旋转，而不是像以前的工作中那样将旋转空间收缩到较小的范围内。

我们选择使用 3D 点云作为这个任务的状态表示，因为制陶本质上是一个 3D 任务，并且点云已被证明在扩散策略中能很好地作为 3D 表示。为了在制作的所有阶段完整观察陶器，我们在场景中固定了 4 台相机，并在末端执行器上安装了一台相机以提供顶部视图。这个顶部视图对捕捉碗的内部至关重要，特别是在开口非常小并且从固定的侧视相机被陶器墙体严重遮挡的早期阶段。相机设置的可视化如图 2 a 所示。为了得到最终的粘土点云，每个 5 个相机捕获一个点云。接下来，每个点云被转换为机器人的坐标框架。对于每个相机，这需要外参标定矩阵，而对于末端执行器相机，还需要当前的末端执行器位姿。在每个点云被转换之后，通过位置和基于颜色的阈值隔离出粘土点。最后，通过迭代最近点（ICP）来调整 4 个固定相机视图和末端执行器相机之间的点云对齐，以校正由本体感觉误差引起的变化。然后，最终组合的点云被均匀下采样为固定的 2048 个点。如图 2 c 所示，显示了点云处理的可视化。

在这项工作中，我们通过触觉教学收集了 20 个捏制陶器轨迹的示范数据集。我们将初始黏土圆柱的高度在 5 到 8 厘米之间变化（假设黏土体积不变）。我们将最终陶器碗的直径在 7 到 12 厘米之间变化。由于黏土体积不变，较大的黏土直径也与倾斜的壁相关，而较小的直径则与垂直的壁相关。示范轨迹的长度在 21 到 31 个独立的捏合动作之间。训练前，我们对点云和动作应用关于 z 轴的旋转变换，以 2° 为增量，将数据集大小增加到 3600 个轨迹。每个示范的最终陶器状态点云的可视化如图 3 b 所示。关于数据集，请参见项目网站。

PinchBot 的关键组成部分包括嵌入预训练（第 III-A 节）、目标条件（第 III-B 节）、子目标指导（第 ?? 节）、任务进展指导（第 ?? 节）和碰撞动作投影（第 ?? 节）。该框架的可视化如图 3 所示。

A. 点云嵌入预训练

我们使用预训练作为克服在现实世界演示的小数据集中训练我们的点云嵌入模型的挑战的方法。在本研究中，我们比较了两种点云嵌入模型的性能，这些模型已在文献中被证明对基于点云的扩散策略效果良好，即 PointBERT [7] 和一种来自 [3] 的修改版 PointNet（在本文中被称为 DP3 PointNet）。在 [3] 和 [5] 中，研究者直接使用了 DP3 PointNet 嵌入模型而没有进行任何预训练用于变形塑形任务。然而，在之前的工作中，我们发现对 PointBERT 模型进行预训练对于目标条件点云扩散策略效果良好 [2] 。通过这项工作，我们旨在探索使用每个点云嵌入模型训练的目标条件策略之间的行为差异。我们在 ShapeNet [8] 重建任务上对 PointBERT 和 DP3 PointNet 模型进行预训

www.xueshuxiangzi.com

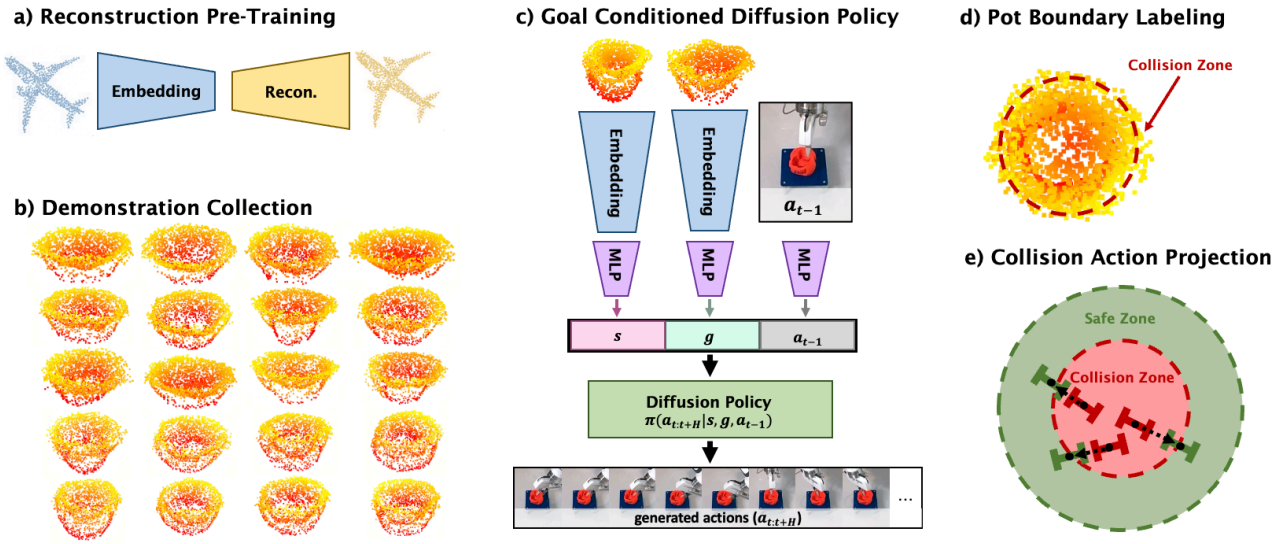

Fig. 3. PinchBot 方法。a) 对重建任务进行预训练嵌入。b) 通过运动示教收集 20 个真实世界中夹捏陶艺的演示。c) 训练目标条件扩散策略变体。d) 陶器实时直径预测用于碰撞投影。e) 将所有不安全的动作位置（即将与现有粘土墙碰撞的动作）投射到安全区内。安全区边界由当前预测的直径决定。

TABLE I
捏制陶艺定量结果。每种嵌入/训练目标变体进行了三次运行。

		8cm Diameter			10cm Diameter			12cm Diameter		
		CD [mm]	EMD [mm]	MSE [mm]	CD [mm]	EMD [mm]	MSE [mm]	CD [mm]	EMD [mm]	MSE [mm]
PointBERT	Binary Pred.	9.1 ± 0.5	7.8 ± 0.6	0.02 ± 0.02	7.1 ± 0.2	6.3 ± 0.5	0.05 ± 0.02	7.9 ± 0.0	7.2 ± 0.2	0.87 ± 0.18
	Cont. Guid.	9.9 ± 0.5	9.2 ± 0.5	0.02 ± 0.01	7.3 ± 0.2	6.5 ± 0.4	0.36 ± 0.06	8.2 ± 0.3	7.2 ± 0.4	1.29 ± 0.68
	Sub-Goal	8.2 ± 0.3	6.9 ± 0.5	0.04 ± 0.04	9.5 ± 0.6	8.7 ± 1.0	0.73 ± 0.07	9.9 ± 0.8	9.5 ± 0.6	3.48 ± 0.62
DP3 PointNet	Binary Pred.	9.2 ± 0.6	8.6 ± 0.7	0.18 ± 0.09	7.2 ± 0.3	6.5 ± 0.4	0.10 ± 0.10	9.9 ± 0.4	9.3 ± 0.7	1.38 ± 0.54
	Cont. Guid.	8.5 ± 0.9	8.6 ± 1.1	0.62 ± 0.48	7.0 ± 0.2	6.5 ± 0.4	0.39 ± 0.13	9.2 ± 0.2	8.7 ± 0.1	1.32 ± 0.11
	Sub-Goal	9.1 ± 1.1	7.9 ± 0.8	0.06 ± 0.05	9.6 ± 0.3	9.1 ± 0.5	1.11 ± 0.31	10.2 ± 0.3	9.9 ± 0.6	3.88 ± 0.85

练，然后在扩散策略目标上微调这些模型。为了微调每个嵌入模型，我们添加了一个 3 层 MLP 结构，将点云嵌入投射到大小为 512 的低维空间中。

B. 目标条件化

在这项工作中，我们旨在训练一个能够适应多种陶器直径目标的单一捏陶策略。目标调节对于一个可适应且可控的策略是必不可少的。此外，目标调节结合终止预测允许策略完全自治，而不需要任何阈值来确定停止条件。没有目标调节的情况下，单个策略必须针对单个目标进行训练，这会简化问题 [5] 。对于目标调节，目标点云与所选的点云模型嵌入在一起，然后该潜在的目标嵌入与状态嵌入和先前的动作连接在一起，以调节动作轨迹的迭代去噪（如图 3 c 所示）。

与其通过连接的潜在状态、目标和前一个动作向量来对去噪过程进行条件化，我们的子目标扩散策略版本是通过连接的潜在状态、N 潜在子目标和前一个动作向量进行条件化的。参数 N 基于动作预测水平和子目标步长来确定，即如果预测水平是 16 个动作且子目标步长是 8，则 $N = 2$ 。通过提供中间子目标，我们正在探索这种中间状态指导是否有助于策略进行长水平动作生成。

为了继续探讨如何更好地指导我们的政策以完成长时间跨度的捏陶任务，我们提出了一个非常简单的策略，让模型预测生成的动作在轨迹中的位置。我们称这个任务为进度引导扩散策略，因为模型正在学习预测它在特定示范中的位置。为此，我们将 γ 从动作修改为预测一个介于 -1 和 1 之间的连续值，以表示沿轨迹的百分比，其中 -1 表示初始状态，1 表示最终状态。

对于我们的任务，单次捏合位置和旋转的轻微误差可能是灾难性的，并导致不可恢复的状态，例如在陶器壁上产生孔洞。在这项工作中，我们开发了一种碰撞检测和动作投影框架，以最大限度地减少有害动作。该过程在图 3 d-e 中进行了可视化。我们通过找到适合投影到圆内的 x,y 平面内的 $> 95\%$ 点的直径，将一个圆拟合到当前状态点云。然后，我们采取从策略生成的动作并识别当前的 x,y 位置是否位于圆内。如果是，这个动作位置会被投影到圆的边缘。这在保持 z、旋转和指尖距离组件的同时修改了动作的 x,y 参数。

通过我们的实验，我们探讨了嵌入模型的选择（第 ?? 节）以及任务进展或子目标指导（第 ?? 节）如何影响策略行为。此外，我们进行了消融研究（第 ?? 节），并使用 T-SNE 分析点云嵌入的潜在空间（第 ?? 节）。对于每种实验变体，我们进行了策略的三次实际运行，并呈现每个评估指标的平均值和标准偏差。每种扩散策略变体的训练均使用了 16 步的动作预测范围，而在测试时策略在重新计划前的执行范围为 4 个动作。子目标扩散策略变体的训练采用了步长为 8 的子目标。

www.xueshuxiangzi.com

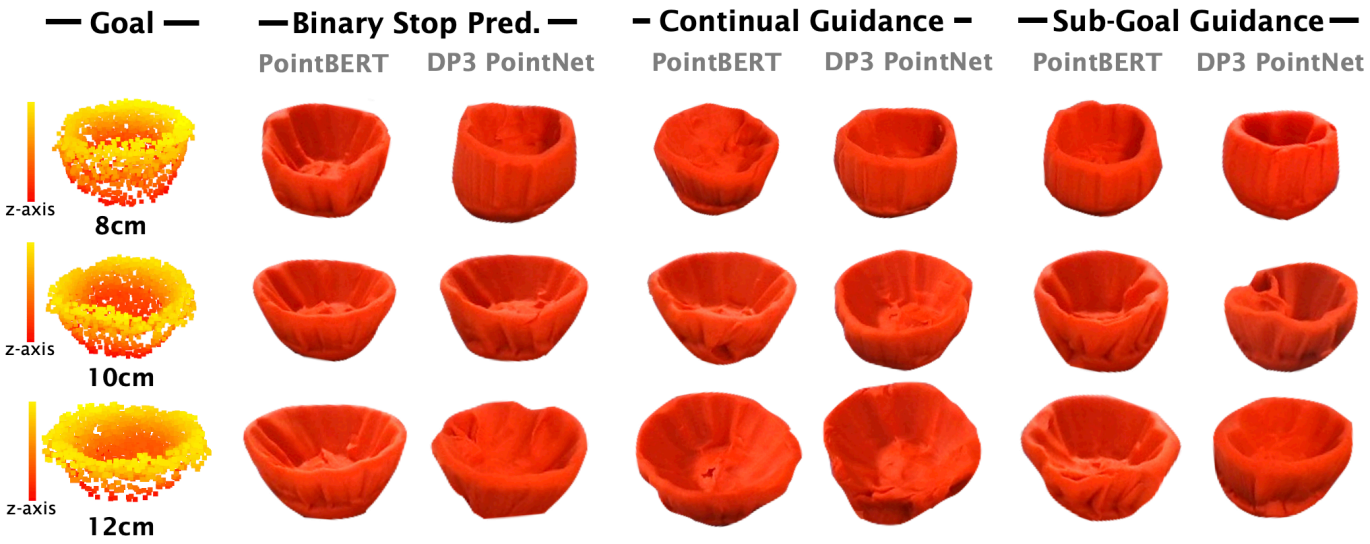

Fig. 4. 最终形状根据点云嵌入和策略变体在 8cm、10cm 和 12cm 碗直径目标上进行调整。

Fig. 5. 最终形状用于消融研究变体，探索预训练、碰撞投影和扩散策略风格的训练如何影响最终方法的性能。

C. 评估指标

为了评估每个策略在创建各种陶器目标方面的表现，我们评估目标陶器直径为 8cm、10cm 和 12cm 时，最终状态和目标点云之间的 Chamfer 距离（CD）[17] 和 Earth Mover's 距离（EMD）[18]。除了这些点对点相似性度量外，我们还提供了目标陶器直径与策略创建的最终陶器直径之间的平均平方误差（MSE）。

通过我们的实验，我们发现 PointBERT 策略在所有定量指标上都优于 DP3 PointNet 变体（如表 I 所示）。此外，PointBERT 策略需要的操作显著更少即可到达最终碗的位置（II）。DP3 PointNet 策略所需的动作数大约是基础演示轨迹的两倍。PointBERT 策略能够更好地判断何时停止以及轨迹中的状态位置。定性分析中，我们发现 DP3 PointNet 策略在何时切换锅的不同区域方面表现困难，并且在区域已经光滑时执行了额外且不必要的夹捏。

我们发现，任务进度指导提高了 DP3 PointNet 策略的行为表现以及对多种碗口直径的适应性（如表格 I 所示）。任务进度指导对 PointBERT 策略的定量性能影响最小（无论是二元预测还是任务进度指导结果，二者都在彼此的标准偏差范围内）。然而，从定性上看，任务进度指导训练确实提高了 PointBERT 策略根据目标点云来区分 R_x 末端执行器旋转变化的能力。动作的 R_x 直接告知陶瓷墙的角度。对于直径为 8 厘米的碗，基础演示中有非常小的 x 轴旋转来创建接近垂直的墙体。通过任务进度指导，PointBERT 策略能够更好地区分目标之间的行为模式，并正确地更广泛地改变 R_x。相比之下，通过二元预测训练出来的 PointBERT 策略能够通过准确的位置变化在每次挤压中更好地匹配基础目标直径和陶瓷结构，并且在旋转上有一些小的变化。当两个策略都能充分满足目标变化时，我们认为任务进度指导产生了一种更具表现力和变化性的策略。另一方面，子目标引导扩散策略在直径为 12 厘米的最大碗时表现不佳。我们推测这可能是因为对于更大直径碗的演示轨迹包含了较小直径碗的中间状态。仅使用子目标条件（即策略未为所有状态提供最终陶瓷目标），会有更多的训练数据点分布用于创建较小直径的碗。

TABLE II
每种方法的平均动作。跨越目标直径。

		# Actions
PointBERT	Binary Pred.	28.4 ± 2.6
	Cont. Guid.	33.3 ± 10.7
	Sub-Goal	34.8 ± 5.3
DP3 PointNet	Binary Pred.	51.3 ± 18.5
	Cont. Guid.	57.4 ± 17.5
	Sub-Goal	35.4 ± 4.2

通过我们的消融研究，我们探讨了预训练、碰撞投影和扩散训练过程对我们方法性能的影响。为此，我们在 10 厘米直径的目标上评估了 PointBERT 和 DP3 PointNet 二进制预测策略，并将其与移除各组件后的变体进行比较（定量结果见表格 III，定性结果在图 5 中展示）。通过我们的消融研究，我们发现预训练、碰撞投影和扩散策略风格的训练本身都对我们提出的方法的成功有所贡献。扩散策略自身对于该方法来说是极其重要的，因为我们的数据高度多模式化。此外，预训练点云嵌入提高了策略正确识别目标已到达的能力。

www.xueshuxiangzi.com

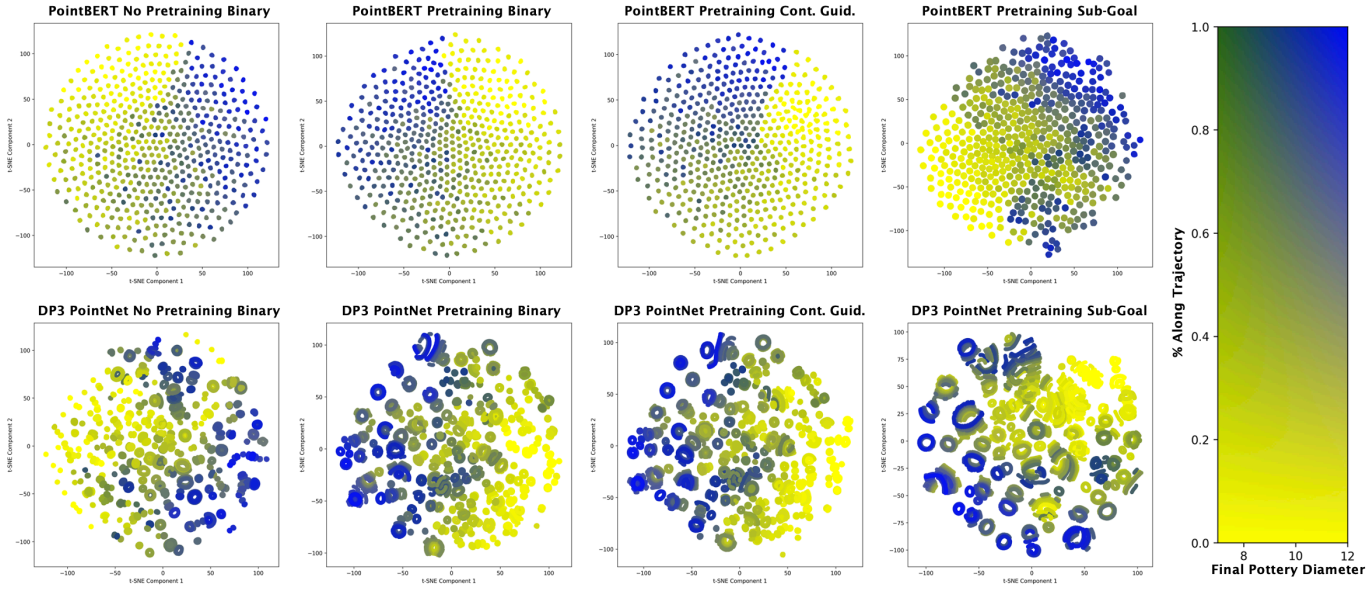

Fig. 6. 不同模型变化下旋转增强演示轨迹嵌入的 T-SNE 图。沿 x 轴的颜色条变化对应于每个轨迹可视化的最终陶器直径，而 y 轴颜色变化则可视化每个相应演示轨迹中的点。

TABLE III
消融研究。10 厘米目标上的二元预测模型。

		CD [mm]	EMD [mm]	MSE [mm]
PointBERT	Pretraining	7.1 ± 0.2	6.3 ± 0.5	0.05 ± 0.02
	No pretraining	7.4 ± 0.2	6.6 ± 0.3	0.78 ± 0.03
	No col. proj.	7.7 ± 0.3	7.9 ± 0.1	1.02 ± 0.25
	Regression	18.9 ± 1.9	17.7 ± 1.7	7.75 ± 2.26
DP3 PointNet	Pretraining	7.2 ± 0.3	6.5 ± 0.4	0.10 ± 0.10
	No pretraining	8.3 ± 0.5	7.7 ± 0.7	0.80 ± 0.91
	No col. proj.	7.5 ± 0.1	7.1 ± 0.2	0.64 ± 0.09
	Regression	20.7 ± 0.4	20.9 ± 0.1	9.86 ± 0.28

为了进一步调查嵌入模型和策略变体之间的性能差异，我们在图 6 中使用 T-SNE 可视化了示范陶器轨迹的潜在嵌入。通过对比两种不同嵌入模型的可视化，我们可以看到在 PointBERT 潜在空间中沿着轨迹的区别比 DP3 PointNet 中更加明显。这证实了 DP3 PointNet 模型难以识别陶器的区域何时已完成和/或目标已达到的结果。此外，与其他 PointBERT 策略变体相比，PointBERT 子目标指导策略的潜在空间在轨迹进展和最终陶器直径上分离效果更差，这与子目标策略的性能不佳相符。最后，我们可以看到，在 DP3 PointNet 变体中，具有连续指导的 DP3 PointNet 策略在最终陶器直径和轨迹进展方面具有最清晰的分离，这与定量结果一致。

在这项工作中，我们介绍了 PinchBot，一种模仿学习框架，利用预训练的点云嵌入、任务进度指导和扩散策略，成功学习一个能够根据目标调整雕塑轨迹的单一目标条件政策，以创造多样的长远制陶目标。尽管我们已经展示了我们的单一政策能够通过少量的实际演示对各种目标进行适度的适应能力，这种行为克隆框架在泛化能力方面仍然存在一定限制。未来的工作可以探索创建一个大得多的数据集，或将动态预测整合到策略学习框架中，以设计出一个更具广泛泛化和适应能力的机器人制陶框架。或者，最近的工作探讨了利用潜在空间强化学习来引导现有的扩散策略 [19] ，这可以直接应用于 PinchBot，以快速将策略适应到更广泛的目标范围。无论这些泛化能力的限制如何，我们提出的框架能够成功学习一种可适应的、目标导向的捏制陶政策，能够制作质量良好的陶碗。这种捏制陶任务的真正长远特性允许我们探索由于嵌入模型和预训练选择而产生的政策行为差异。我们希望在未来，可变形的造型工作将在字母造型目标以外的复杂任务上基准测试，因为当部署在这些更具挑战性的长远任务中时，我们可以更好地识别出方法之间微妙的性能差异。

References

[1] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[2] A. Bartsch, A. Car, C. Avra, and A. B. Farimani, "Sculptdiff: Learning robotic clay sculpting from humans with goal conditioned diffusion policy," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7307–7314.

[3] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," *arXiv preprint arXiv:2403.03954*, 2024.

[4] H. Zhu, Y. Wang, D. Huang, W. Ye, W. Ouyang, and T. He, "Point cloud matters: Rethinking the impact of different observation spaces on robot learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 77 799–77 830, 2024.

[5] U. Yoo, A. Hung, J. Francis, J. Oh, and J. Ichnowski, "Ropotter: Toward robotic pottery and deformable object manipulation with structural priors," in *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*. IEEE, 2024, pp. 843–850.

[6] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[7] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-bert: Pre-training 3d point cloud transformers with masked point modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 19 313–19 322.

www.xueshuxiangzi.com

[8] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[9] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[10] H. Shi, H. Xu, Z. Huang, Y. Li, and J. Wu, "Robocraft: Learning to see, simulate, and shape elasto-plastic objects in 3d with graph networks," *The International Journal of Robotics Research*, vol. 43, no. 4, pp. 533–549, 2024.

[11] H. Shi, H. Xu, S. Clarke, Y. Li, and J. Wu, "Robocook: Long-horizon elasto-plastic object manipulation with diverse tools," *arXiv preprint arXiv:2306.14447*, 2023.

[12] A. Bartsch, C. Avra, and A. B. Farimani, "Sculptbot: Pre-trained models for 3d deformable object manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 12 548–12 555.

[13] D. Bauer, Z. Xu, and S. Song, "Doughnet: A visual predictive model for topological manipulation of deformable objects," in *European Conference on Computer Vision*. Springer, 2024, pp. 92–108.

[14] A. Bartsch and A. B. Farimani, "Llm-craft: Robotic crafting of elasto-plastic objects with large language models," *arXiv preprint arXiv:2406.08648*, 2024.

[15] ——, "Planning and reasoning with 3d deformable objects for hierarchical text-to-3d robotic shaping," *IEEE Robotics and Automation Letters*, pp. 1–8, 2025.

[16] Y. You, B. Shen, C. Deng, H. Geng, S. Wei, H. Wang, and L. Guibas, "Make a donut: Hierarchical emd-space planning for zero-shot deformable manipulation with tools," *IEEE Robotics and Automation Letters*, 2025.

[17] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, pp. 99–121, 2000.

[18] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 605–613.

[19] A. Wagenmaker, M. Nakamoto, Y. Zhang, S. Park, W. Yagoub, A. Nagabandi, A. Gupta, and S. Levine, "Steering your diffusion policy with latent space reinforcement learning," *arXiv preprint arXiv:2506.15799*, 2025.

www.xueshuxiangzi.com