BridgeVLA：基于视觉语言模型的高效 3D 操作学习的输入输出对齐方法

近年来，利用预训练的视觉语言模型 (VLM) 构建视觉语言动作 (VLA) 模型已成为一种高效的机器人操作学习方法。然而，目前只有少数方法将 3D 信号融入 VLM 进行动作预测，而且这些方法未能充分利用 3D 数据固有的空间结构，导致样本效率低下。本文介绍了 BridgeVLA，这是一种新颖的 3D VLA 模型，其 (1) 将 3D 输入投影到多个 2D 图像上，确保输入与 VLM 主干模型对齐；(2) 利用 2D 热图进行动作预测，将输入和输出空间统一在一个一致的 2D 图像空间内。此外，我们还提出了一种可扩展的预训练方法，使 VLM 主干模型能够在下游策略学习之前预测 2D 热图。大量实验表明，该方法能够高效地学习 3D 操作。BridgeVLA 在三个模拟基准测试中均优于最先进的基线方法。在 RLBench 中，它将平均成功率从 81.4% 提升至 88.2%。在 COLOSSEUM 中，它在具有挑战性的泛化设置中表现出显著提升的性能，将平均成功率从 56.7% 提升至 64.0%。在 GemBench 中，它的平均成功率超越了所有对比的基线方法。在真实机器人实验中，BridgeVLA 的平均成功率比最先进的基线方法高出 32%。它在多种分布外的设置（包括视觉干扰和未见指令）中表现出稳健的泛化能力。值得注意的是，它能够在 10 多个任务中实现 96.8% 的成功率，每个任务仅需 3 条轨迹，彰显了其卓越的样本效率。

1 引言

利用预训练的视觉-语言模型 (VLM) [1–4] 开发大型视觉-语言-动作 (VLA) 模型已成为学习可泛化且鲁棒的操作策略的一种有前景的方法 [5–9]。然而，大多数 VLA 模型仅包含 2D 图像输入，并且需要大量的数据收集工作。另一方面，3D 机器人策略在模型设计中利用了 3D 结构先验，并在学习复杂的 3D 机器人操作任务中展现出卓越的样本效率 [10–14]。我们能否开发一个统一的 3D VLA 模型，将 VLA 模型的有效性与 3D 策略的效率相结合？

尽管已经有一些研究探索将 3D 信息集成到 VLM 中以开发 3D VLA 模型 [15, 16]，但这些研究通常将动作转换为不具有空间结构的 token 序列，并使用下一个 token 预测来预测动作。该策略未能像先前高效的 3D 策略 [10–14] 那样利用 3D 结构先验，将观察输入和动作输出对齐到统一的空间，因此导致样本效率低下。开发 3D VLA 模型的另一个重大挑战在于，动作微调中使用的 3D 输入与原始 VLM 预训练中使用的 2D 图像输入之间存在不一致，这导致与原始 VLM 预训练结果存在较大的分布偏移。

图 1 概览。BridgeVLA 是一种新颖的 3D VLA 模型，它将输入和输出对齐到统一的 2D 图像空间中。它使用 2D 热图进行物体定位预训练，并针对 3D 操作的动作预测进行微调。模拟和现实世界的实验结果均表明，它能够高效地学习 3D 操作。

为了应对上述挑战，如图 1 所示，我们提出了 BridgeVLA，这是一种新颖的 3D VLA 模型，它实现了卓越的采样效率和强大的泛化能力。为了确保输入与预训练的 VLM 主干模型对齐，BridgeVLA 将 3D 点云观测值转换为从不同正交投影视图捕获的多个 2D 图像 [13, 14]。为了利用 3D 输入的结构先验，BridgeVLA 经过训练以预测 2D 热图，从而进行平移动作预测。由与投影图像对应的标记生成的 2D 热图与这些图像共享相同的分辨率，从而将输入观测值和输出动作在统一的空间结构内对齐。鉴于原始 VLM 是预训练用于预测标记序列的，这与我们的 VLA 的 2D 热图输出不兼容，我们还引入了一种可扩展的预训练方法，该方法使用以文本输入为条件的热图训练模型来识别地面物体。这种预训练方法使 VLM 能够在后续策略学习微调之前预测热图。总而言之，我们的设计在预训练和微调过程中将输入和输出对齐到共享的二维空间中。

我们在模拟和现实世界中进行了大量的实验，以评估所提出的方法。结果表明，BridgeVLA 能够高效地学习 3D 操作。它在 RLBench [17] 中的表现优于最先进的基线方法，平均成功率从 81.4% 提升至 88.2%。在 COLOSSEUM [18] 中，它在具有挑战性的泛化设置中展现出强劲的性能，成功率从 56.7% 提升至 64.0%。在 GemBench [19] 中，它的平均成功率超越了所有对比的基线方法。在真实机器人实验中，我们评估了七种不同的设置，涵盖了从视觉扰动到操作未知类别的物体等各种场景。BridgeVLA 平均比最先进的方法高出 32%，并且在泛化到多种分布外设置方面表现出色。值得注意的是，BridgeVLA 仅使用 3 条轨迹进行训练，便能在 10 多个任务中实现 96.8% 的成功率，彰显了其卓越的样本效率。总而言之，本文的贡献主要有三方面：

• 我们提出了 BridgeVLA，这是一种新颖的 3D VLA 模型，它能够利用视觉语言模型，通过基于 2D 热图的输入输出对齐，高效地学习 3D 机器人操作。

• 我们提出了一种可扩展的预训练方法，使模型能够通过对象基础 (object grounding) 预测基于文本输入的热图。

• 我们在模拟和实际环境中进行了大量的实验，以全面评估所提出的方法。结果表明，BridgeVLA 在两种环境下均优于最先进的方法，并在实际机器人实验中实现了卓越的样本效率。

2 相关工作

语言条件视觉运动策略。大多数语言条件视觉运动策略使用 Transformer 来处理二维视觉输入，并直接生成用于操作的三维动作 [5–9, 20–25]。在这些工作中，利用预训练的视觉-语言模型 (VLM) 来开发大型视觉-语言-动作 (VLA) 模型因其在学习复杂操作方面的有效性而变得流行 [5–9]。然而，这种基于二维图像的策略通常需要在数据收集上投入大量精力，每个任务通常需要数百条轨迹才能有效学习。另一方面，三维操作策略通过利用三维输入固有的空间结构，在高效学习方面具有巨大的潜力。一些热门的研究将点云数据作为输入 [11, 12, 26–28]。例如，Act3D [12] 提出通过将图像特征提升到观测点云来创建 3D 特征云，并通过对观测空间中的 3D 点进行分类来预测平移动作。另一系列研究利用体素表示观测空间，并预测体素空间内的平移动作，从而将输入观测和输出动作统一在同一空间内 [10, 29]。最近，RVT [13] 和 RVT-2 [14] 提出利用 3D 点云的正交投影将 3D 信号转换为 2D 图像，以避免处理 3D 输入的高计算成本。与上述方法不同，我们的方法旨在将 VLA 模型的有效性和 3D 策略的效率统一在一个统一的框架内，兼具两者的优势。

3D 视觉-语言-动作 (VLA) 模型。虽然 2D VLA 模型已得到广泛研究，但 3D VLA 模型 [15, 28, 30, 31] 的研究仍相对不足。Zhen 等人 [15] 在大型语言模型 (LLM) 之上构建了 3D-VLA，并训练该模型执行 3D 推理、多模态目标生成和机器人规划。Lift3D [30] 提出使用隐式和显式 3D 机器人表征来增强 2D 基础模型（例如 DINOv2 [32]），以学习 3D 操作策略。FP3 [28] 利用转换器融合点云、本体感受状态和语言指令的信息。PointVLA [31] 分别利用 VLM 和点云编码器处理 2D 图像和 3D 点云。两个编码器的嵌入被注入到动作专家中进行动作预测。 SpatialVLA [16] 引入了 Ego3D 位置编码，将 3D 信息注入 2D 图像观测，并引入自适应动作网格，以更易于迁移的方式表示机器人运动。我们的方法与上述方法的不同之处在于，其设计充分利用了 3D 输入的空间结构进行动作预测。此外，它通过将 3D 输入投影到多个 2D 图像中（而非将 3D 信息注入 VLM），弥合了预训练 VLM 的 2D 图像输入与 3D 输入之间的差距。这种设计使其能够同时利用 VLM 主干网络中的广泛知识和 3D 输入中嵌入的空间结构先验。

3 Bridge VLA

3.1 准备工作

BridgeVLA 旨在学习一个多任务 3D 机器人操作策略 π，它将观测 o 和语言指令 l 映射到动作 a：

我们假设可以访问包含 N 条轨迹的专家演示集合 D。每条轨迹包含一条语言指令和一系列观测-动作对τi。观测o是从一个或多个视点捕获的一幅或多幅RGB-D图像。根据前人的研究[10,12,13]，动作a由6自由度末端执行器位姿T∈SE(3)、目标夹持器状态g∈{0,1}和下一个关键帧的碰撞标志c∈{0,1}组成。碰撞标志c指示运动规划器在向目标位姿移动时是否应避免碰撞。关键帧通常捕获轨迹中的重要步骤或瓶颈步骤（详见附录B.1）[33]。 BridgeVLA 通过迭代过程运行：1) 根据当前观测 ot 和指令 l 预测动作 at；2) 使用基于采样的运动规划器 [34–36] 移动到预测的下一个关键帧姿态 Tt；3) 更新观测并重复，直至任务完成或达到最大步长 Hmax。

如图 2 所示，BridgeVLA 采用双阶段训练方案。在预训练阶段，它被训练用于预测目标检测数据集上的二维热图。在微调阶段，点云被投影到多个二维图像中，作为 VLM 主干网络的输入。该模型被训练用于预测二维热图，以估计平移动作和其他动作成分。这种设计在预训练和微调阶段都将输入和输出对齐到共享的二维空间中。

图 2 模型架构。(a) 二维热图预训练：我们在二维物体检测数据集上训练 BridgeVLA。该模型以描述目标物体的图像和语言作为输入，并输出二维热图，该热图突出显示与目标物体对应的感兴趣区域。请注意，此处显示的边界框仅供参考；在输入模型时，图像中并不存在该边界框。(b) 三维动作微调：该模型以三维点云的三幅正交投影图像和语言指令作为输入。它输出三幅二维热图，突出显示末端执行器在下一个关键帧中在所有三个视图中的位置。对于其余的动作组件，它使用多层感知器 (MLP) 处理图像特征标记，以预测下一个关键帧的旋转动作、夹持动作和碰撞标志。

3.2 二维热图预训练

VLM 主干网络最初预训练用于预测不包含空间结构的 token 序列。为了使其具备与下游策略学习相同的预测热图的能力，我们引入了一个预训练阶段，通过热图训练模型识别目标物体。具体来说，我们利用 RoboPoint [37] 的 12 万个目标检测样本作为预训练数据集。对于每幅图像，我们根据所有目标物体的边界框构建真实热图 Hgt。具体来说，对于每个目标，我们构建一个带有空间截断的概率图：

其中 x = (u, v) 表示像素位置，y 表示目标物体边界框的中心，pmin 表示概率阈值。对于所有目标物体，我们通过平均和归一化融合所有目标物体的概率图，得到 Hgt：

其中 Ω 表示像素空间。真实热图示例请参见图 9。

图 9：预训练数据可视化。我们列出了一些预训练数据样本。对于每个样本，左侧显示原始图像；中间显示目标对象的边界框；右侧显示用于训练的地面实况热图。

如图 2 所示，我们将一张图片以及描述目标对象的文本提示输入到 BridgeVLA 的 VLM 主干网络中。在本文中，我们采用 PaliGemma [1] 作为 VLM 主干网络，它由一个 SigLIP 视觉编码器 [38] 和一个 Gemma Transformer 主干网络 [39] 组成。在预训练阶段，PaliGemma 将一张或多张二维图像以及前缀文本（例如，关于图像的问题）作为输入，并输出后缀文本（例如，问题的答案）。该模型使用因果注意力机制来预测后缀文本标记，而对图像标记和前缀文本标记则采用双向注意力机制。这使得图像标记能够融合来自前缀文本的信息。

为了预测热图，我们首先根据输出图像标记的块位置重新排列它们，以重建空间特征网格。然后，凸上采样块 [40] 将网格转换为与输入图像具有相同分辨率的热图。该模型采用交叉熵损失函数进行训练，以预测能够定位图像中所有感兴趣目标位置的热图。需要强调的是，与先前研究 [15, 16] 中使用的传统下一个标记预测方法不同，所提出的预训练策略输出的是具有空间感知的二维热图。此外，该方法具有高度可扩展性，因为它原则上可以利用任何可以构建为热图预测任务的视觉语言数据集，例如关键点检测和语义分割。

3.3 3D 动作微调

在微调过程中，我们首先根据已校准相机捕获的 RGB-D 图像重建场景的点云。为了与 VLM 主干网络的 2D 图像输入对齐，我们从三个视角（顶部、正面和右侧）渲染点云的三幅正交投影图像，并将这些图像用作 VLM 主干网络的输入图像，就像 RVT [13] 和 RVT-2 [14] 中提到的那样。然后，这些图像连同任务指令一起被输入到预先训练好的 VLM 主干网络中，为三个视角分别生成热图。重要的是，我们在 VLM 前向传播过程中不加入任何其他信息（例如机器人状态），以最大限度地减少预训练和微调之间的分布偏差。

对于平移动作，我们对所有三个视角的热图进行反向投影，以估计均匀分布在机器人工作空间中的所有 3D 点网格的得分。