GeoAware-VLA: 隐式几何感知的视觉-语言-动作模型

Ali Abouzeid[1], Malak Mansour[1], Zezhou Sun[1], Dezhen Song[1]

***Abstract*— 视觉-语言-动作（VLA）模型通常难以推广到新的相机视角，这一限制源于它们从 2D 图像中推断出鲁棒 3D 几何结构的困难。我们引入了 GeoAware-VLA，这是一种简单而有效的方法，通过将强大的几何先验知识整合到视觉主干中来增强视点不变性。与训练一个视觉编码器或依赖显式 3D 数据不同，我们将一个冻结、预训练的几何视觉模型用作特征提取器。然后，一个可训练的投影层适应这些富含几何信息的特征以供策略解码器使用，从而减轻了从头开始学习 3D 一致性的负担。通过在 LIBERO 基准子集上的广泛评估，我们展示了 GeoAware-VLA 在零样本推广到新的相机姿态方面取得了显著改进，在模拟中将成功率提高了超过 2 倍。至关重要的是，这些益处转化到了物理世界；我们的模型在现实机器人上表现出明显的性能提升，尤其是在从未见过的相机角度进行评估时。我们的方法证明了它在连续和离散动作空间中的有效性，突显出稳健的几何基础是创建更具通用性的机器人代理的关键组成部分。**

I. 介绍

在非结构化环境中执行多样化操作任务的通用代理的发展是机器人学中的一个核心目标。视觉-语言-动作（VLA）模型是一种流行的范式，它学习将视觉观察和自然语言指令直接映射到机器人动作上。大多数方法在其训练领域内表现出很强的性能 [1]–[3]，但通常难以泛化至这些领域之外，甚至在相机视角发生微小变化时也表现不佳 [4], [5]。这种限制源于从 2D 视觉输入推断一致的 3D 世界模型的困难，这是可靠操作和空间意识的前提。

先前的研究探索了两种主要策略来应对这一挑战。一种方法是将显式的三维表示，如点云，纳入到策略的观测空间 [6]–[8] 中。虽然有效，但这些方法通常需要深度传感器，并且在构建和处理显式三维结构时引入显著的计算开销。另一种方法在于隐性地鼓励模型学习几何上一致的特征而无需在三维中重建场景。这通常是通过在训练过程中利用多视角数据或数据增强来实现产生一种视图不变的策略 [9], [10]。

然而，这些隐式方法也有它们自身的缺点。解缠技术通常依赖于精心策划的多视角数据集进行有效训练，而基于增强的方法从根本上受到生成新视图和所采样视图分布的计算成本限制。这些问题可以通过利用已经从大量数据集中提炼出 3D 理解的强大预训练几何基础模型来解决。例如，视觉几何定位变换器（VGGT）[11] 显式优化以提炼世界丰富的三维表示，其已被训练用于推断关键几何方面，如相机参数、多视角深度、密集点云和点跟踪。

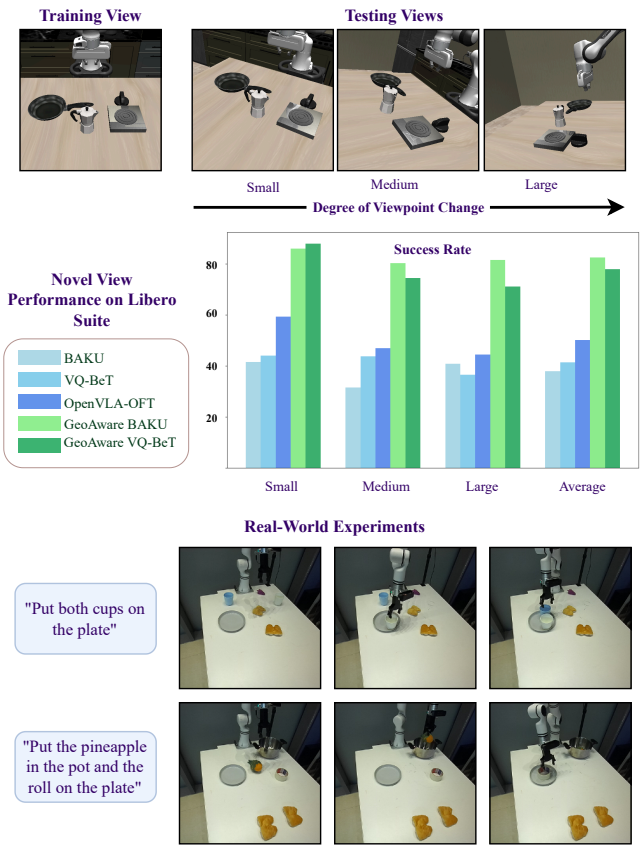

Fig. 1. **（顶部）** LIBERO 数据集中的训练和测试视角的示例。GeoAware-VLA 展现了对新颖视角的零样本泛化能力。**（中期）** GeoAware-VLA 在 LIBERO 数据集上的成功率达到超过最新技术 30%，实现了对新颖视角的优越泛化。**（底部）** 成功进行实际部署过程中的关键中间步骤。

[1]Authors are with the Department of Robotics, Mohamed bin Zayed University of Artificial Intelligence, Masdar City, Abu Dhabi, UAE. Emails:{ali.abouzeid, malak.mansour, chengsong.hu, dezhen.song}@mbzuai.ac.ae

基于这些见解，我们提出了 GeoAware-VLA，这是一种对标准 VLA 架构进行的简单但非常有效的修改，显著增强了其对抗新型摄像机视角变化的鲁棒性。我们的核心假设是，策略在不同视角下泛化的性能从根本上与其视觉编码器的几何精度紧密相关。因此，我们将标准图像编码器替换为强大的 VGGT 主干网络，并将其用作冻结特征提取器。通过利用已经固有地跨不同视角一致的特征，我们的策略免去了从零开始学习 3D 几何结构的任务负担。我们只需要添加一个轻量级、可训练的投影层来将这些强大的特征映射到 BAKU 策略解码器 [1] 的潜在空间中。如实验所示，结合强大的几何先验知识显著提高了零样本泛化到新型摄像机视角的能力，具体结果见图 1。

总之，我们的贡献如下：

- 我们提出了 GeoAware-VLA，这是一种将视觉几何基础模型有效集成到 VLAs 中的方法。
- 地理意识-VLA 在新型视图中实现了成功率翻倍，并在物理机器人上展示了显著的性能提升。
- 我们通过证明我们的方法能够持续提升不同策略解码器的泛化性能来确认其通用性。

II. 相关工作

A. 模仿学习

现代机器人学习植根于模仿学习（IL），早期方法如行为克隆（BC）[12] 遭遇了由于分布变化导致的累积错误问题。这促使开发出更稳健的视觉运动策略，生成性方法如扩散策略 [13] 和基于能量的方法如隐式行为克隆（IBC）[14] 学习复杂的动作分布以增强控制能力。在此基础上，该领域的关注点转向创建大规模、通用的代理。具有影响力的模型如 DeepMind 的 Gato [15] 和 Octo [16] 展示了一个单一的大规模变压器，通过在多样化的多机器人数据集上进行预训练，能够建立强大的且可适应的动力控制基础。这一轨迹最终形成了当前的 VLA 模型范式，这些模型利用网络规模模型的知识来统一感知、语言和控制。这包括一系列具有影响力的著作，如 RT-1 [17] 及其后续作品 [18], [19]，以及像 OpenVLA [2], [20] 和 π- [21], [22] 这样的强大开源模型，这些模型共同推动了泛化和性能的边界。我们的工作基于 BAKU 架构 [1]，这是一种轻量级的 VLA，以其高效的设计选择而脱颖而出，包括多感官观察和动作分块。

B. 机器人学习中的视觉编码器

视觉主干网络，处理原始像素输入，是 VLA 模型的关键组成部分。基础方法通常使用语义编码器如 ResNet [23]，并随后通过添加 FiLM 层来融合来自语言或其他输入的条件信息 [24]。

一个较新的范式涉及像 SigLIP [25] 及其后续版本 SigLIP 2 [26] 这样的模型，这些模型是在大规模的网络图像-文本配对上进行预训练的。这种训练方法创建了与语言自然对齐的表示形式，使其在 VLA 任务中非常有效。然而，这些主干的主要关注点仍然是语义——识别物体是什么，而不是它们精确的三维几何形状。

随着语义编码器在策略学习中的发展，计算机视觉社区中出现了一类强大的几何视觉模型。诸如 DUSt3R [27]、MUST3R [28] 和 VGGT [11] 等架构专门用于从多视角图像进行几何推理，推断相机姿态和密集深度图等属性。尽管这些模型在 3D 重建方面非常有效，但它们并不是作为策略学习的通用基础模型设计的，因此在机器人领域中的应用受到了限制。据我们所知，在此之前，这类几何模型并未被隐式地整合到 VLA 模型中。我们的工作初步探讨了这一方向，研究如何利用这些几何模型的隐式 3D 理解来改善机器人的策略。

C. 视点稳健的机器人策略

为了实现更具视角鲁棒性的策略，已经探索了各种方法。一种方法是通过从模拟器渲染的图像来增强训练数据 [9], [29]。虽然有效，但这引入了仿真到现实转换的挑战。另一种有前景的方法是利用新颖视图合成 [10], [30]。与基于模拟器的方法不同，这些模型可以直接在真实世界的数据上运行，从而避免了仿真到现实的差距。然而，当新视角与原始视角差别很大时，这些模型的表现可能仍然不佳 [10]。

其他作品，如 Act3D [31] 和 3D Diffuser Actor [32]，利用了显式的 3D 表示形式，比如点云或体素网格 [6], [8], [33], [34]。这些方法对视角变化具有鲁棒性，但通常需要可靠的深度数据和相机校准，在大规模数据集中如 [19], [35], [36] 这样的数据可能较为稀缺。与这些方法相比，我们的方法集成了强大的几何先验知识，这减轻了策略从零开始学习 3D 一致性的负担，并避免了对显式深度或校准信息的需求。

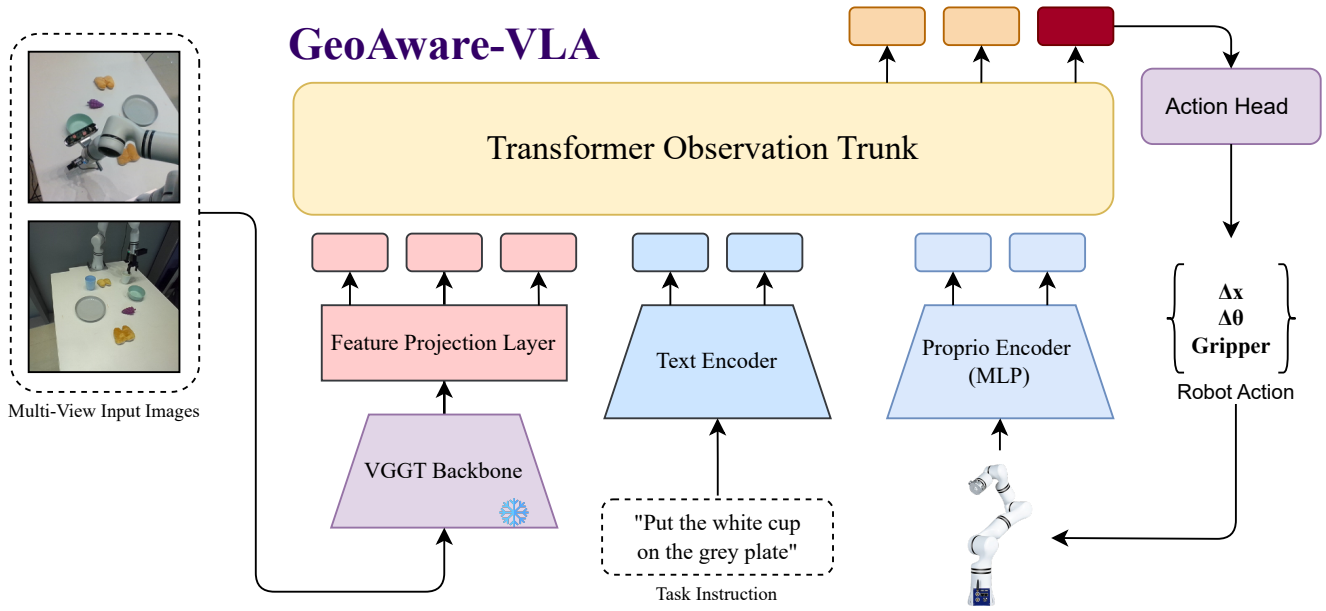

Fig. 2. GeoAware-VLA 的整体架构图。它输入多视角图像，使用 VGGT 提取视图鲁棒特征以生成机器人动作。

III. 问题设置

我们考虑一个多任务模仿学习的标准设置，其中目标是训练一个策略 $\pi(a_t|o_t, l)$，该策略在给定当前观察 o_t 和自然语言指令 l 的情况下预测机器人动作 a_t。观测 o_t 包含一组来自不同摄像机视角的 VRGB 图像，表示为 $\boldsymbol{I_t} = \{I_1, I_2, \ldots, I_V\}$，以及机器人的本体感受状态 $\boldsymbol{p}_t$。这些共同构成了完整的观测 $o_t = (\boldsymbol{I_t}, \boldsymbol{p}_t)$。策略由一个带有可训练权重 θ 的神经网络 π_θ 参数化。每个时间步长 t 的动作 $a_t \in \mathbb{R}^7$ 指定了末端执行器的姿态和夹具的位置。此 7 维的动作向量定义为：

$$a_t = [\Delta x, \Delta y, \Delta z, \Delta\alpha, \Delta\beta, \Delta\gamma, g_t]$$

其中：

- $(\Delta x, \Delta y, \Delta z)$ 是末端执行器相对平移的三维向量。
- $(\Delta\alpha, \Delta\beta, \Delta\gamma)$ 是末端执行器相对旋转的三维轴角表示。
- $g_t \in \{+1, -1\}$ 是表示夹爪状态的标量，其中 $+1$ 表示“打开”，-1 表示“关闭”。

该策略是在一组专家演示数据集 $\mathcal{D} = \{(o_t^{(i)}, l^{(i)}, a_t^{(i)})\}_{i=1}^N$ 上通过行为克隆（BC）训练的，它将模仿学习任务视为一个监督学习问题。目标是找到产生尽可能接近演示数据集中 $\mathcal{D}$ 专家动作 a_t 的策略网络 π_θ 的最优参数 θ^*。这是通过最小化关于网络参数 θ 的损失函数（如均方误差 MSE）来实现的：

$$\mathcal{L}_{BC}(\theta) = \mathbb{E}_{(o_t,l,a_t)\sim\mathcal{D}}\left[||\pi_\theta(o_t, l) - a_t||^2\right]$$

这里，期望 $\mathbb{E}$ 是在专家演示数据集 $\mathcal{D}$ 中的所有样本上取的。通过最小化这个损失，策略网络学会了将观察和语言命令映射到相应的专家级动作。

IV. 方法论

我们提出了 GeoAware-VLA，一种增强标准 VLA 框架的新方法。我们的方法核心是将可训练的视觉编码器替换为一个冻结的、具有几何感知能力的编码器。为了使来自 VGGT 主干网的丰富多层特征能够被策略网络使用，我们引入了一个可训练的视觉投影层。策略架构的其余部分借鉴并建立在 [1] 的元素之上。我们的模型总体架构如图 2 所示。

整体架构可以分为三个阶段：感觉编码、策略解码和动作生成。最终动作 a_t 是视觉观测 o_t^{vis}、本体感受状态 s_t 和语言指令 l_t 的函数：

$$a_t = \pi_\phi(P_\theta(E_{\text{VGGT}}(o_t^{\text{vis}})), E_{\text{lang}}(l_t), E_{\text{proprio}}(s_t))$$

其中 E 表示各自的编码器，P_θ 是我们的可训练视觉投影层，π_ϕ 是包含转换器主体和动作头的策略网络。

A. 感官编码器

每个输入模态都由一个专用编码器处理，生成固定维度的嵌入 D_{repr}，然后再传递给策略的观察主干。

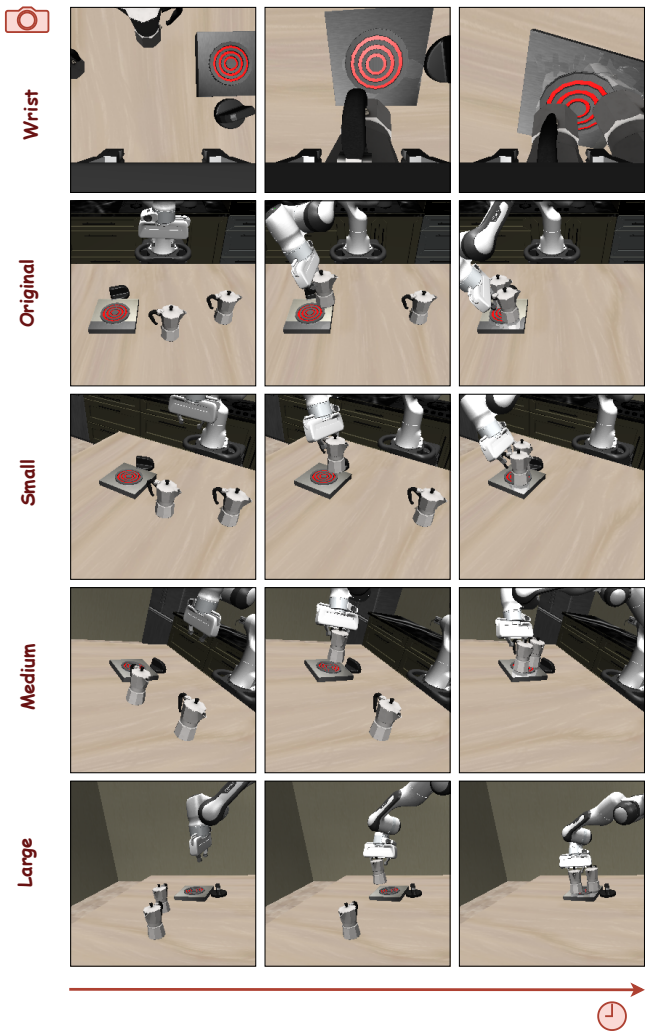

Fig. 3. 我们实验设置中单个片段的可视化表示，展示了不同视角（行）随时间（从左到右的列）的变化。顶部一行显示了手腕摄像头图像，而第二行则显示了俯视视角。模型仅在这两个视角上进行训练。底部三行显示了用于评估模型的三种新颖且未见过的视角。

1) 几何感知视觉编码器: 为了编码视觉观察，我们利用了一个冻结的预训练 VGGT 模型主干，并移除了预测头。VGGT 的一个关键特性是它从多个中间层生成一个特征张量列表，捕获了一种视觉和几何信息的层级结构。我们的可训练视觉投影层设计用于高效地聚合和浓缩这些信息。

令编码器输出一个包含 M 特征张量的列表。我们选择一个均匀间隔的 L 中间层子集进行处理。对于每个摄像机视图，来自单个选定层 l 的特征具有维度 $N_l \times D_{\text{vggt}}$，其中 N_l 是视觉标记的序列长度，D_{vggt} 是 VGGT 的隐藏维度。

视觉投影层处理每组相机视图中从 L 层选择的特征。来自每一层（$\boldsymbol{z}_l \in \mathbb{R}^{N_l \times D_{\text{vggt}}}$）的特征首先通过

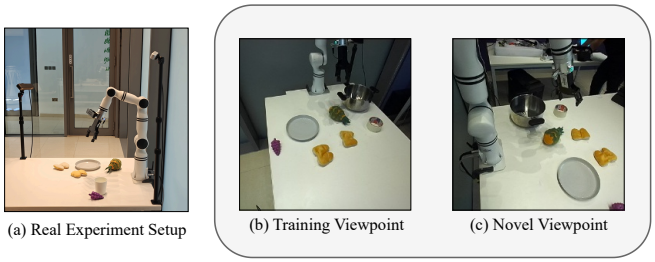

Fig. 4. 真实机器人设置。(a) 硬件示意图。(b) 用于训练所有策略的视角。(c) 用于我们零样本评估实验的视角。

一个具有可训练卷积层和 ReLU 层的 1D 卷积网络，接着是一个自适应平均池化层将特征汇聚成单个向量（$\boldsymbol{f}_l \in \mathbb{R}^{D_{\text{conv}}}$）。这些特定层的向量然后被拼接成一个单一的向量。这个拼接后的向量随后通过最终的多层感知器（MLP）生成视觉嵌入 $\boldsymbol{z}_{\text{vis}} \in \mathbb{R}^{D_{\text{repr}}}$。

单个视图特征的总体变换

$$\boldsymbol{z}_{\text{vis}} = \text{MLP}_\theta\left([\text{Conv1D}_{\theta_1}(\text{Pool}(\boldsymbol{z}_{l_1})); \ldots; \text{Conv1D}_{\theta_L}(\text{Pool}(\boldsymbol{z}_{l_L}))]\right)$$

此视角投影高效地利用轻量级、可训练模块适应强大的、固定的 VGGT 特征。

2) 语言和本体感觉编码器: 我们使用简单的基于 MLP 的投影器处理非视觉模态。

- **语言编码器:** 任务指令字符串 l_t 首先使用预训练的句子转换器 [37] 进行编码，得到一个维度为 $D_{\text{lang\_emb}}$ 的嵌入。然后一个多层感知机将这个嵌入投影到公共表示空间中，生成 $\boldsymbol{z}_{\text{lang}} \in \mathbb{R}^{D_{\text{repr}}}$。
- **本体感觉编码器:** 机器人状态向量 s_t（末端执行器姿态和夹具状态）通过两层 MLP 产生本体感觉嵌入 $\boldsymbol{z}_{\text{proprio}} \in \mathbb{R}^{D_{\text{repr}}}$。

B. 观测主干和策略解码器

我们的策略解码器是一个 GPT 风格的、仅解码器的变压器模型。所有模态（每个摄像机视角的 z_{vis}，z_{lang} 和 z_{proprio}）的编码表示被视为输入标记序列。一个可学习的动作标记被附加到此序列中。该变压器使用因果自注意力掩码处理此序列，输出嵌入对应于动作标记位置的 $h_{\text{action}} \in \mathbb{R}^{D_{\text{hidden}}}$ 作为动作头的输入。

C. 行动头

最终动作是由一个专门的动作头模块根据特征向量 h_{action} 生成的。我们尝试了两种变体来处理不同的动作分布。

- **多层感知器头部：** 一个简单的确定性头部，实现为两层 MLP，直接回归连续动作向量 $a_t \in \mathbb{R}^{D_{\text{act}}}$。该头部适用于单峰动作分布。我们称这种变体为 GeoAware BAKU。
- **VQ-BeT 头部：** 为了潜在地模拟多模态专家行为，我们使用向量量化行为转换器（VQ-BeT）头部。该模块用 VQ-VAE 替换了传统的 k-means 聚类来学习离散的动作代码本。它通过从其代码本中分类适当的动作令牌并回归一个连续的偏移来预测一个动作，使其能够捕捉复杂的多模态动作分布。这个变体是 GeoAware VQ-BeT。

V. 实验设置

A. 模拟基准测试

The LIBERO benchmark [35] 评估多任务和终身机器人操作任务中的知识迁移。它由四个任务套件组成：LIBERO-Spatial，测试在固定任务和物体类型的情况下对新空间布局的泛化；LIBERO-Goal，在保持相同物体和布局的同时引入新任务；LIBERO-Object，在相同的任务和布局下评估新型物体的表现；以及 LIBERO-Long，展示更广泛的物体、布局和背景。每个套件包含 10 个任务，并为每个任务提供 50 个人类演示以进行微调。我们在 Libero 的套件上训练的模型在原始和新颖视点上进行了评估，这些视点均受 [6] 启发，并显示在图 3 中。

B. 真实世界实验

为了在现实世界中评估我们的方法，我们建立了一个桌面操作环境，该环境采用了一只 Realman 65B 机械臂。通过使用 Meta Quest VR 控制器远程操作手臂，我们收集了定制的数据集。场景从两个固定的不同摄像头视角捕获，以提供多视角观察结果给策略。设置如图 4 所示。

1) **把两个杯子放在盘子上** 拿起两个单独的杯子并将它们放在盘子上。这是一个受 Libero Long 任务启发的多物体操作任务。
2) **挑选蓝色的杯子并将其放在盘子上面的碗里** 拿起蓝色的杯子并将其放在一个碗里，这个碗本身位于一个盘子上。此任务测试将物体放入嵌套目标中的能力。
3) **从锅中取出菠萝并放在杯子之间：** 该任务通过要求机器人拾起一个菠萝并将其放置在其他两个物体之间来测试其对空间关系的理解。
4) **将菠萝移到锅里并将卷饼移到盘子里** 执行两个不同的拾取和放置操作：将菠萝移到锅里并将卷饼移到盘子上。这受到来自 Libero Long 的任务的启发。
5) **将蓝色的杯子堆放在白色的杯子上面：** 一个需要高精度稳定垂直对齐的堆叠杯子任务，尤其是在新颖视角下特别困难。

C. 基线方法

我们将我们的**地理感知-VLA** 模型与三个强大的基线进行比较：两个源自 [1] 架构的模型和 OpenVLA-OFT [20]，我们将其包含在内以测试使用更大、预训练模型的性能。

- **巴库：** 具有原始可训练视觉编码器和确定性 MLP 动作头的标准 BAKU 模型。
- **VQ-BeT:** BAKU 模型，但其 MLP 头被 VQ-BeT 动作解码器替换。
- **开放 VLA-光场传输技术：** 我们利用 OpenVLA-OFT 模型作为额外的基线，评估我们的方法在更大更强的预训练视觉语言模型上的性能。

我们实验比较的核心是隔离几何感知视觉编码器的贡献。因此，我们的 GeoAware-VLA 模型与相应基线之间的唯一架构差异在于视觉模块。基线使用基于 Resnet + Film 的标准可训练视觉编码器 BAKU，而我们的模型则使用带有轻量级、可训练投影层的提议冻结 VGGT 编码器。

D. 训练

所有模型均在 PyTorch 中实现，并在一个配备 40GB 显存的 NVIDIA A100 GPU 上进行训练。对于 LIBERO 基准测试中的实验，我们分别为四个任务套件训练了一个单独的策略。每个模型使用 AdamW 优化器进行了 150,000 步的训练。对于我们的真实世界评估，政策在我们的自定义数据集上进行了 50,000 步的训练。在所有实验中，我们使用的批量大小为 64，并向策略提供了来自两个摄像头视角的观察结果。

VI. 结果与讨论

我们评估 GeoAware-VLA 以回答三个关键问题：

TABLE I

成功率（%）在 LIBERO 子集上的表现。标记有*的模型结果取自 [20]。

Model	自由子集								平均	
	空间		对象		目标		长		跨越子集	
	Original	Novel	Original	Novel	Original	Novel	Original	Novel	Original	Novel
Diffusion Policy* [13]	78.3	-	92.5	-	68.3	-	50.5	-	72.4	-
Octo* [16]	78.9	-	85.7	-	84.6	-	51.1	-	75.1	-
OpenVLA-OFT [20]	**98.0**	90.0	98.0	15.7	98.0	81.7	91.0	13.7	96.3	50.2
BAKU [1]	94.0	18.0	**100.0**	49.3	96.0	80.7	87.0	3.7	94.2	37.9
VQ-BeT [38]	94.0	41.3	**100.0**	38.3	98.0	83.0	**93.0**	3.0	96.3	41.4
GeoAware BAKU **(我们的)**	95.0	**94.3**	98.0	98.0	98.0	**90.7**	89.9	47.3	95.2	**82.6**
GeoAware VQ-BeT **(我们的)**	95.0	54.3	**100.0**	**99.0**	**99.0**	85.7	**93.0**	**72.7**	**96.8**	77.9

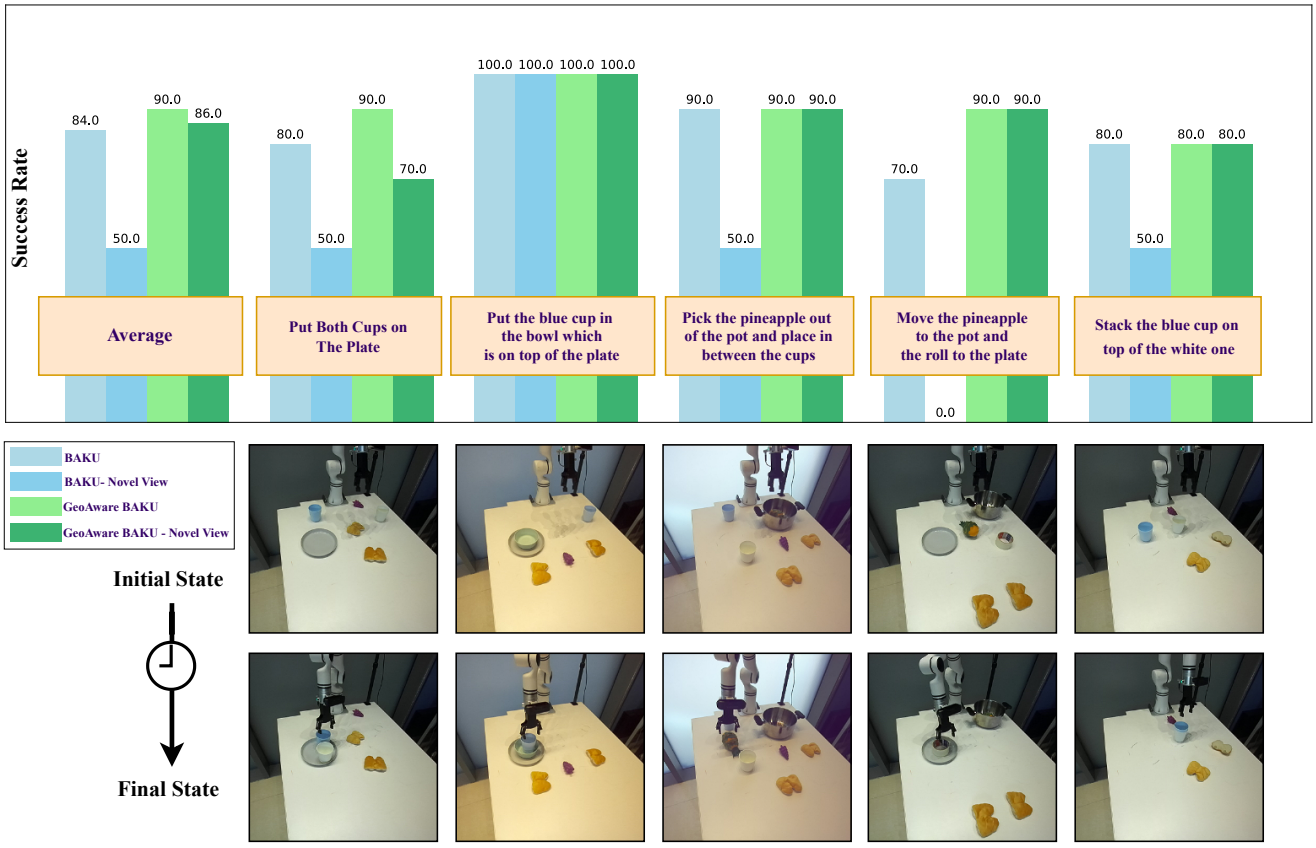

Fig. 5. **（顶部）**对比我们在真实世界中 GeoAware-VLA 的表现与基线模型，突出我们模型在已见和新颖视角上的改进。**(底部)** 评估的真实任务的初始状态和最终状态。

1) 利用预训练的冻结几何编码器进行模仿学习是否是一个可行的策略，即使在训练过的视图上也是如此?
2) 我们的方法在推广到新的相机视角时效果如何，这种推广对视角变化幅度的鲁棒性如何?
3) 实验室中的性能提升能否在实际硬件上实现?
4) 哪些几何主干层对下游策略性能最为关键?

所有结果均报告为每个任务的 10 次运行中的成功率，适用于模拟和现实世界设置。

A. 几何骨架是强大的分布内执行者

首先，我们评估集成冻结几何主干是否影响熟悉任务上的性能。如表 I 所示，我们的 GeoAware 模型不仅在原始训练视图上与基线对照组匹配了高成功率，

TABLE II

新视角下 LIBERO 子集的成功率（%）

Model	Small	Medium	Large	Average
OpenVLA-OFT [20]	59.3	47	44.5	50.2
BAKU [1]	41.5	31.5	40.8	37.9
VQ-BeT [38]	44.0	43.8	36.5	41.4
GeoAware BAKU	86.0	**80.3**	**81.5**	**82.6**
GeoAware VQ-BeT	**88.0**	74.5	71.2	77.9

TABLE III

针对 GEOAWARE VQ-BET 在 LIBERO-LONG 子集上进行的 VGGT 层选择消融研究。展示了已见和新颖视角的成功率（%）。

Model	Seen	Novel
OpenVLA-OFT [20]	**93.0**	13.7
GeoAware VQ-BeT (All Layers)	90.0	**74.3**
GeoAware VQ-BeT (4 Evenly Spaced - Default)	**93.0**	72.7
GeoAware VQ-BeT (Last 4 Layers)	60.0	34.3

而且整体性能也得到了提升，将 VQ-BeT 的原始性能提高到了最高的总体成功率为 96.8%。

这一结果支持我们的假设，即一个轻量级的、可训练的投影层足以将冻结主干中的几何丰富特征适应于策略。它表明依赖于几何先验为策略学习提供了一个有效的基础，而不会损害分布内性能。

B. GeoAware-VLA 实现了在视角变化下的稳健零样本泛化

我们接下来评估模型在零样本设置下对新相机视角的泛化能力。当模型被评估于新的相机姿态时，性能上出现了明显的差异。基线模型的成功率显著下降，表明它们对视角变化敏感（表 I）。

相比之下，我们的 GeoAware 模型表现出更强的鲁棒性，保持了较高的成功率，并且比基线高出两倍多，在整体性能上比下一个基线高出 30%。这种泛化的改进突显了几何先验在实现视点不变性方面的有效性。

为了进一步分析这种鲁棒性，我们检查了在不断增加的视点偏差程度下的性能。表 II 显示，在基线性能较低且不一致的情况下，我们的 GeoAware 模型表现出持续较高的成功率，并且在所有视角下都具有更稳健的退化。

C. 实际性能提升

最后，我们评估在模拟中观察到的性能改进是否可以转移到实际系统。结果显示这些趋势有效地传递到了现实世界。如图 5 所详细展示，我们的 GeoAware BAKU 模型在一个系列的操作任务上实现了可测量的性能提升，甚至在原始训练视角下也是如此。

D. VGGT 层选择的消融研究

为了确定 VGGT 的 24 层主干网络中的哪些层次对策略性能最重要，我们将默认配置（均匀分布的 4 个层次）与两种替代方案进行了比较：使用全部 24 层和仅使用最后 4 层。

如表 III 所示，使用均匀间隔层的默认配置取得了最佳的整体性能。虽然使用所有层在新视图泛化方面提供了小幅提升，但这导致了分布内性能略有下降，并且计算成本更高。值得注意的是，即使是最弱的配置（仅使用最后 4 层），其在新视图上的表现也显著优于下一个最好的基线。尽管整体性能有所下降，但在未见过的视图上仍有 34.3%的成功率，这仍然是 [20] 的两倍多。

VII. 结论与未来工作

在这篇论文中，我们介绍了 GeoAware-VLA，一种显著提升 VLA 模型对新相机视角泛化能力的方法。通过用冻结的预训练几何基础模型（VGGT）和一个轻量级投影层替换标准可训练视觉编码器，我们有效地将强大的 3D 感知先验注入策略中。这一简单的修改缓解了策略从零开始学习几何一致性的难题。

我们在 LIBERO 基准上的实验表明，这种方法非常有效。GeoAware-VLA 不仅在原始训练视角上保持了强劲的性能，还在新型未见过的视图上实现了零样本成功率达到 2 倍以上的显著提升。这些收益在连续和离散动作头中均保持一致，并在一个真实世界的机器人操作平台上得到了成功的验证。

GeoAware-VLA 的成功强调了一个关键见解：视觉主干的几何敏锐度是构建稳健、通用机器人代理的关键因素。我们的工作提供了一种实用且计算效率高的蓝图，利用强大的现成几何模型来增强模仿学习策略。未来的工作可以探索整合其他几何基础模型，将这种方法应用于更广泛的任务和机器人形态，并研究微调 VGGT 主干如何影响性能。

References

[1] S. Haldar, Z. Peng, and L. Pinto, "Baku: An efficient transformer for multi-task policy learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 141 208–141 239, 2024.

[2] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[3] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[4] X. Li, K. Hsu, J. Gu, O. Mees, K. Pertsch, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani, S. Levine, J. Wu, C. Finn, H. Su, Q. Vuong, and T. Xiao, "Evaluating real-world robot manipulation policies in simulation," in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 3705–3728. [Online]. Available: https://proceedings.mlr.press/v270/li25c.html

[5] A. Xie, L. Lee, T. Xiao, and C. Finn, "Decomposing the generalization gap in imitation learning for visual robotic manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 3153–3160.

[6] A. Wilcox, M. Ghanem, M. Moghani, P. Barroso, B. Joffe, and A. Garg, "Adapt3r: Adaptive 3d scene representation for domain transfer in imitation learning," *arXiv preprint arXiv:2503.04877*, 2025.

[7] R. Yang, G. Chen, C. Wen, and Y. Gao, "Fp3: A 3d foundation policy for robotic manipulation," *arXiv preprint arXiv:2503.08950*, 2025.

[8] 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.

[9] J.-C. Pang, N. Tang, K. Li, Y. Tang, X.-Q. Cai, Z.-Y. Zhang, G. Niu, M. Sugiyama, and Y. Yu, "Learning view-invariant world models for visual robotic manipulation," in *The Thirteenth International Conference on Learning Representations*, 2025.

[10] S. Tian, B. Wulfe, K. Sargent, K. Liu, S. Zakharov, V. Guizilini, and J. Wu, "View-invariant policy learning via zero-shot novel view synthesis," *arXiv preprint arXiv:2409.03685*, 2024.

[11] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5294–5306.

[12] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Advances in neural information processing systems*, vol. 1, 1988.

[13] 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.

[14] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson, "Implicit behavioral cloning," in *Conference on robot learning*. PMLR, 2022, pp. 158–168.

[15] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-maron, M. Giménez, Y. Sulsky, J. Kay, J. T. Springenberg *et al.*, "A generalist agent," *Transactions on Machine Learning Research*.

[16] D. Ghosh, H. R. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo *et al.*, "Octo: An open-source generalist robot policy," in *Robotics: Science and Systems*, 2024.

[17] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn *et al.*, "Rt-1: Robotics transformer for real-world control at scale," 2023. [Online]. Available: https://arxiv.org/abs/2212.06817

[18] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[19] Q. Vuong, S. Levine, H. R. Walke, K. Pertsch, A. Singh, R. Doshi, C. Xu, J. Luo, L. Tan, D. Shah *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," in *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023.

[20] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: Optimizing speed and success," *arXiv preprint arXiv:2502.19645*, 2025.

[21] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π0: A vision-language-action flow model for general robot control. corr, abs/2410.24164, 2024. doi: 10.48550," *arXiv preprint ARXIV.2410.24164*.

[22] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, "π0. 5: a vision-language-action model with open-world generalization, 2025," *URL https://arxiv. org/abs/2504.16054*, vol. 1, no. 2, p. 3.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[25] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.

[26] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa *et al.*, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv preprint arXiv:2502.14786*, 2025.

[27] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 697–20 709.

[28] Y. Cabon, L. Stoffl, L. Antsfeld, G. Csurka, B. Chidlovskii, J. Revaud, and V. Leroy, "Must3r: Multi-view network for stereo 3d reconstruction," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1050–1060.

[29] Y. Seo, J. Kim, S. James, K. Lee, J. Shin, and P. Abbeel, "Multi-view masked world models for visual robotic manipulation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 30 613–30 632.

[30] J. Yang, I. Huang, B. Vu, M. Bajracharya, R. Antonova, and J. Bohg, "Mobi-pi: Mobilizing your robot learning policy," *arXiv preprint arXiv:2505.23692*, 2025.

[31] T. Gervet, Z. Xian, N. Gkanatsios, and K. Fragkiadaki, "Act3d: 3d feature field transformers for multi-task robotic manipulation," *arXiv preprint arXiv:2306.17817*, 2023.

[32] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," *arXiv preprint arXiv:2402.10885*, 2024.

[33] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 785–799.

[34] Y. Zhu, Z. Jiang, P. Stone, and Y. Zhu, "Learning generalizable manipulation policies with object-centric 3d representations," in *7th Annual Conference on Robot Learning*.

[35] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 776–44 791, 2023.

[36] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning (CoRL)*, 2023.

[37] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[38] S. Lee, Y. Wang, H. Etukuru, H. J. Kim, N. M. M. Shafiullah, and L. Pinto, "Behavior generation with latent actions," *arXiv preprint arXiv:2403.03181*, 2024.