单目三维物体　　用于人机交互的视觉语言模型位置估计

Ari Wahl[1,2], *Dorian Gawlinski*[1,2], *David Przewozny*[1], *Paul Chojecki*[1], *Felix Bießmann*[2], *Sebastian Bosse*[1]

[1] Fraunhofer HHI，德国柏林 [2] 柏林技术高等学校（BHT），德国柏林

arXiv:2603.01224v1 [cs.CV] 1 Mar 2026

摘要

预训练的通用视觉语言模型（VLM）凭借丰富的世界知识和二维物体检测能力，有望提升直观的人机交互 [1]。然而，用于三维坐标检测任务的视觉语言模型并不多见。在这项工作中，我们通过输入腕戴式相机拍摄的单目 RGB 图像、自然语言指令和机器人状态，并返回三维物体位置，研究视觉语言模型的交互能力。我们收集并整理了一个包含超过 100,000 张图像的异构数据集，并使用 QLoRA 和自定义回归头对视觉语言模型进行微调。通过实现条件路由，我们的模型在获得专门的三维位置估计能力的同时，仍能处理通用视觉查询。结果表明，模型具有稳健的预测性能，在测试集上的平均绝对误差中位数为 13 毫米，相较于未经微调的简易基线提升了五倍。在约 25% 的情况下，预测结果处于机器人与物体交互时可接受的范围内。

索引词—单目，3D 物体位置估计，人机交互，视觉语言模型

1. 引言

视觉语言模型已被证明在某些机器人应用场景中有用，例如用于机器人任务的视觉问答或视觉对话 [2]。对于机器人学中的其他任务，视觉语言模型的应用仍是一个活跃的研究课题，后文将对此进行探讨。我们的目标是弥补视觉语言模型在机器人学中的一项关键不足：实现从单目 RGB 图像中估计三维物体位置。自适应特征路由会根据任务动态地将视觉特征导入专用通路：要么回归三维坐标，要么保留视觉语言模型的通用能力。通过使用低秩适配（LoRA）矩阵，我们保持基础模型不变，仅在微调期间训练 LoRA 和回归头参数。随后，我们利用这一特性实现条件路由机制，从而仍可访问原始量化基础模型并使用其全部能力。

本研究得到了柏林市长、柏林参议院总理府研究与技术部门（LZDV）的支持

我们最终实现的模型大小为 37 亿个参数。在测试集上，最终模型的中位数平均绝对误差为 13 毫米，中位数欧几里得距离误差为 27 毫米。

2. 相关工作

面向机器人学的视觉语言模型（VLM）是一个加速发展的领域，其最先进水平（SOTA）不断变化。该领域正在迅速壮大，并采用了广泛的方法，例如直接输出动作的视觉语言动作模型（VLA），或同样被训练用于在二维图像空间中指向物体的视觉语言模型。一些最知名的、将视觉与语言骨干网络用于机器人学的模型属于 RT-X 系列。这是一类直接输出机器人动作的 VLA。RT-1-X（2023年）和 RT-2-X（2023年）模型基于其前身——两种模型架构 RT-1[3] 和 RT-2[4], ，并在 Open X-Embodiment 数据集上训练，该数据集涵盖广泛的机器人场景与任务。它们的泛化能力远优于其前身，但仍不是通用的机器人模型。对于开源的 RT-1-X，已有研究表明，要将其用于未纳入训练集的机器人，还需要额外微调[5], OpenVLA[6],CogACT（2024年）[7] 以及 $\pi 0$（2024年）[8], 同样如此，它们是在 Open X-Embodiment 数据集上训练的其他有前景的开源 VLA。CogACT（2024年）使用视觉语言模型提供认知信息，以通过专门的动作模块引导动作预测。 $\pi 0$ （2024年）模型使用视觉语言模型基础模型和动作专家。LERF[9] 模型也被用于机器人学，其中最知名的是 LERF-TOGO（2023年）[10]和 LifelongLERF（2024年）[11], ，后者通过仅重新渲染场景中发生变化的元素，在实时能力方面取得了进展。近来，通用知识视觉语言模型也越来越多地被训练用于在二维空间中指向物体。例如 RoboPoint（2024年）[12], Molmo（2024年）[13] 和 Moka（2024年）[14], ，后者借助 RGB-D 深度信息将这些点投影到三维。通常，用于单目三维物体检测的视觉语言模型，如 VFMM3D（2024年）[15] 和 MonoDETR（2023年）[16] ，是在自动驾驶数据集（例如 Kitti）上训练的。RT-X 系列、LERF-TOGO、VFMM3D、MonoDETR 和 Molmo 均可使用单目 RGB 相机输入。然而，只有 Open-

VLA、RT-1-X、LERF-TOGO 和 CogACT 是开源的。只有少数模型以仍然完全可获取的通用视觉语言模型作为基础模型，例如 RoboPoint。尽管这份清单并不详尽，但符合我们全部标准并可开箱即用的模型目前尚未开源。

3. 方法

3.1. 系统架构与处理流程

模型在 Slurm 集群上使用 Apptainer 和五块显存为 40 GB 的 NVIDIA Tesla A100 GPU 训练了 4 天。本文讨论的核心任务是：在机器人环境中，使用视觉语言模型（VLM）和安装在机械臂上的单目相机，对物体位置进行开放集检测。该任务意味着检测物体在机器人工作空间中相对于机器人基座的三维坐标。为此，我们利用预训练视觉语言模型的目标检测能力。此外，我们还希望 VLM 能为下游任务提供有用的可供性检测能力。与大多数其他研究不同，本文着重于保留视觉语言模型的通用能力，并仅将三维坐标估计作为一项额外能力。为此，我们采用条件路由机制，将一般问题路由至基础模型，并将与任务相关的问题路由至适配后的模型架构。作为概念验证，我们有意采用一个简单的标识词“question”，仅将特定提示路由至基础模型。输入为一张 RGB图像和一条文本提示，其中包括当前夹爪位置和相对于基座的方向。

3.2. 实验设置

数据匮乏是进一步改进这些领域所面临的主要问题之一，另一个问题则是任务、环境和机器人平台的高度多样性 [2]。为了针对我们的任务微调模型，必须从所用机械臂的工作空间中采集数据。我们使用一台配备 RG2-FT夹爪和安装在腕部的 Logitech Brio网络摄像头的六关节 Doosan A0509 机械臂采集训练数据。采集数据时，夹爪的朝向保持固定，使相机沿夹爪方向朝下，同时夹爪仍处于相机的视野内。这样既简化了整体设置，也能使物体始终处于相机的视野中。

3.3. 数据集

采集的数据集包含约 750 种不同物体，每种物体有 5 段图像序列。采集过程中，机器人将工具中心点 (TCP) 从随机的上方位置移动至物体上方的顶部中心位置；工具中心点定义为夹爪最低处的中点。图像序列和元数据以每秒 2 至 6 帧的速率同步采集。通过采用不同的光照条件，并使机器人接近物体时不仅沿直线移动，还沿曲线和三角形轨迹移动，增强了数据集的鲁棒性。约 60% 的序列包含单个物体，40% 的序列包含多物体场景。数据集中的示例图像见图 1a、1b、1c、1d、1e。

3.4. 训练详情

训练时，数据被划分为 90% 的训练与验证数据和 10% 的测试数据。对每个受测模型均采用五折交叉验证。为防止物体高度信息泄漏，训练、验证和测试采用基于组的数据划分，确保同一物体的所有图像都属于同一组。训练和验证使用 Huber损失，测试使用平均绝对误差（MAE）和欧几里得误差，并使用 MAE 评估误差。MAE 使用 PyTorch 的默认实现，并对坐标进行均值归约。为了直观检查误差范围可能允许抓取或推动任务的样本比例，我们绘制了累积分布函数（CDF）图，如图2所示

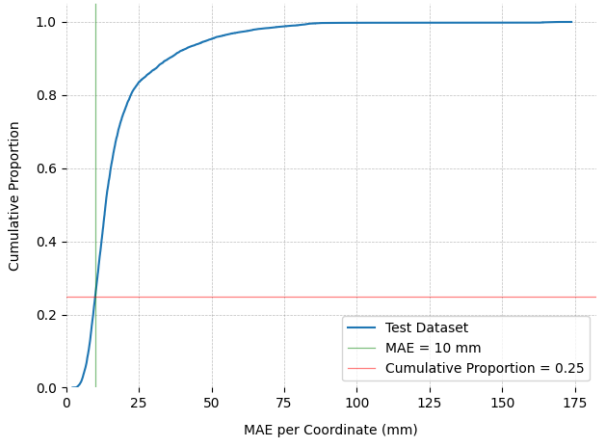


图2：测试集误差结果的累积分布函数（CDF）图

4. 结果

作为基线，我们使用 LLaVA-v1.5 作为特征提取器，并搭配单个线性回归层。在这项任务中，以 7B LLaVA-v1.5 为基础模型的模型表现最佳，性能比基线高出 5 倍。测试的模型包括 7B Mistral、LLaVA-NeXT，以及 0.5B 版本的 LLaVA-Onevision。以 LLaVA-v1.5 为基础模型的最佳模型在训练 10 轮后趋于平稳，Huber 损失约为 20；而基线在验证集上的 Huber 损失约为 100，几乎没有明显改善



(a) 不寻常的设计（冰淇淋模具） (b) 不寻常的光照（半身像玩具） (c) 宽体物体（园艺手套）

(d) 垂直形状的物体（胶棒） (e) 形状不规则的物体（太阳镜）

图 1：不同物体类别的示例图像。

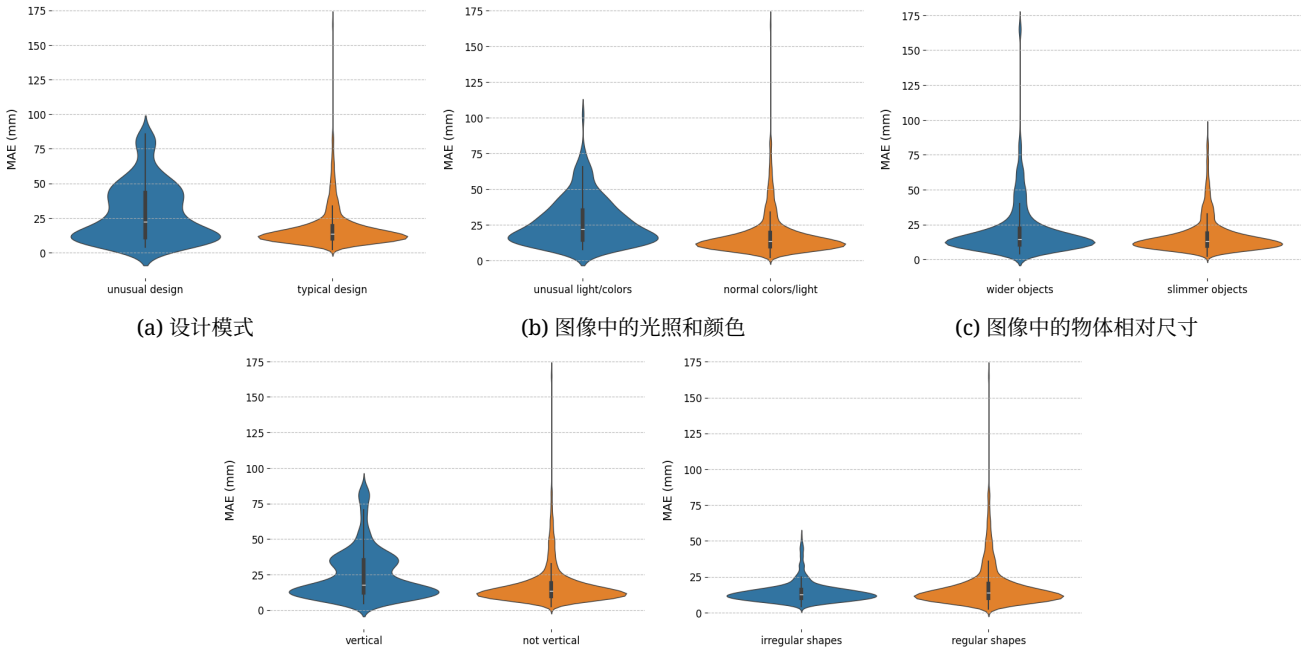
。对于最佳模型，我们进行了详细的误差分析，以探究模型在何处以及为何会出错。虽然模型在 75% 的情况下平均绝对误差低于 20 毫米，但仍存在明显较大的误差：约 5% 的物体误差超过 35 毫米。我们分析了所有离群值数量不少于 25 的组中，各物体组的离群值数量，如表1所示。随后，我们针对误差与物体属性之间的关系提出假设，以了解模型在哪些情况下会出错。第一个假设是，某些垂直形状的物体（如“胶棒”和“汽水瓶”）更难预测，因为从上方无法充分看清它们。第二个假设是，设计独特的物体（如“冰淇淋模具”）更难预测，因为可以推测基础模型会受到传统设计模式的偏向影响，而互联网规模训练数据中这类设计模式占比可能过高。其他假设包括：形状不规则的物体（如“太阳镜”）可能更难预测，因为几何形状有助于找到顶部中心点；较宽物体的位置也更难预测（如“园艺手套”），因为相机越靠近，物体就越可能无法完整出现在视野中。最后，我们推测在这种设定下，z 坐标比其他坐标更难预测，因为仅凭单目 RGB 图像预测物体高度具有挑战性。测试集中与这些假设相关的物体图像示例见图 1a、1b、1c、1d、1e。

组	数量
冰淇淋模具	293
水位	274
人体模型手	234
牙刷套装	225
电源适配器套装	218
药片	217
汽水瓶	207
喷雾罐	206
芥末	179
皂液器	177
开瓶器	144
胶棒	139
3号手术刀柄，12厘米	139
半身像玩具	132
一包筷子	116
玩具烛台	111
旗帜	110
缝纫工具包	87
盘子	84
天鹅绒袋	82
投影仪	71
鸡蛋探测器	68
填充鸡蛋	66
螺旋盖罐	62
玩具剪刀	57
14毫米扳手	53
定时开关	50
TCD解剖剪刀.CVC. ,23CM	43
太阳镜	43
钱包	42
白色蜡笔	40
90度剪刀	34
亚马逊项链	28
公牛雕像	28
降噪耳塞	27
纸杯	25

表1：每个物体组的离群值数量（平均绝对误差 > 40）

5. 讨论

假设检验表明，大多数假设都得到了证实。图 3d、3e、3a、3c、3b 展示了各组分布的小提琴图。与没有垂直形状物体的样本相比，包含垂直形状物体的样本在测试集上的预测误差更高



(a) 设计模式 (b) 图像中的光照和颜色 (c) 图像中的物体相对尺寸

(d) 物体垂直度 (e) 形状规则性

图 3：不同物体和图像属性的误差分布。

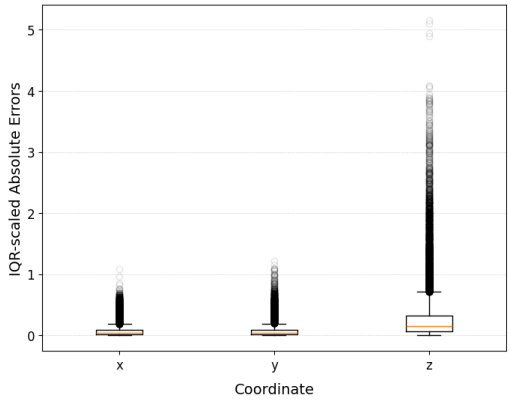


图 4：各坐标四分位距缩放绝对误差的比较

。此外，设计异常的物体、较宽的物体，以及图像中光照或颜色条件异常的物体，其预测误差也更高。曼-惠特尼 U 检验也证实了这一点；检验结果表明，两组之间的差异在显著性水平 0.05 下具有统计显著性，且 p 值已通过 Bonferroni 校正。表示物体高度的 z 坐标预测误差相对高于 x 和 y 坐标。对误差进行四分位距（IQR）缩放后可以观察到这一点，这使得不同范围之间的比较更加公平。x、y 坐标的分布相似，而 z 坐标误差表现出更大的离散程度，表明从单目图像进行深度估计的不确定性更高（见图 4）。在包含各种物体形状、光照条件和物体位置的数据集上进行验证和测试时，模型在未见过的数据上表现出良好的泛化能力。我们最佳模型的测试表现应足以应对大多数情况。欧几里得误差的中位数约为 27 毫米，而经均值归约后的绝对误差中位数约为 13 毫米；我们认为，对于开放集预测任务而言，这是一个成功的结果。从测试集结果的累积分布函数图可以看出，25% 的样本在每个坐标上的平均误差不超过 10 毫米，这意味着该模型有望成功完成抓取或推动等任务。

6. 结论与展望

该模型在开放集预测中对不同物体具有良好的泛化能力。但由于模型仅使用专门采集的数据集进行训练，因此仍会偏向机械臂的工作空间，包括相机模型。未来的目标是通过纳入更多多物体场景，并增加物体接近方式的多样性，来提高数据集的异质性；理想情况下，还应涵盖不同的工作空间和机器人模型。改进不仅可以针对训练数据，也可以针对模型本身。我们计划简化提示输入，并在后期特征融合阶段引入机器人本体感知数据。未来的工作还将采用学习式路由策略。

参考文献

[1] K. Kawaharazuka、Y. Obinata、N. Kanazawa、K. Okada 和 M. Inaba，《预训练视觉语言模型在各种识别行为中的机器人应用》，载于*2023IEEE-RAS 第 22 届国际人形机器人会议（Humanoids）*，2023年，第 1–8 页。

[2] R. Firoozi, J. Tucker, S. Tian 等，“机器人学中的基础模型：应用、挑战与未来”，《国际机器人研究杂志》，2024年9月。

[3]A. Brohan, N. Brown, J. Carbajal 等，“RT-1：面向大规模真实世界控制的机器人变换器”，*arXiv*预印本 *arXiv:2212.06817*，2023年。

[4]A. Brohan, N. Brown, J. Carbajal 等，“RT-2：视觉-语言-动作模型将网络知识迁移至机器人控制”，*arXiv*预印本 *arXiv:2307.15818*，2023年。[5] J. Salzer 和 A. Visser，“将 RT-1-X 基础模型应用于SCARA机器人”，*arXiv*预印本 *arXiv:2409.03299*，2024年。[6] M. J. Kim, K. Pertsch, S. Karamcheti 等，“Open-VLA：开源视觉-语言-动作模型”，*arXiv*预印本 *arXiv:2406.09246*，2024年。[7] Q. Li, Y. Liang, Z. Wang 等，“CogACT：一种基础视觉-语言-动作模型，用于协同机器人操作中的认知与动作”，*arXiv*预印本 *arXiv:2411.19650*，2024年。[8] K. Black, N. Brown, D. Driess 等，“π_0：用于通用机器人控制的视觉-语言-动作流模型”，*arXiv*预印本 *arXiv:2410.24164*，2024年。[9] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa 和 M. Tancik，“LERF：语言嵌入辐射场”，载于*2023 IEEE/CVF*国际计算机视觉会议（*ICCV*），2023年，第19672–19682页。

[10] A. Rashid、S. Sharma、C. M. Kim、J. Kerr、L. Chen、A. Kanazawa 和 K. Goldberg，《用于零样本任务导向抓取的语言嵌入辐射场》，*arXiv*预印本 *arXiv:2309.07970*，2023年。

[11]A. Rashid、C. M. Kim、J. Kerr 等，《Lifelong LERF：使用 FogROS2 进行局部三维语义库存监测》，*arXiv* 预印本 *arXiv:2403.10494*，2024 年。

[12] W. Yuan、J. Duan、V. Blukis 等，“RoboPoint：面向机器人学空间可供性预测的视觉语言模型”，*arXiv*预印本 *arXiv:2406.10721*，2024年。

[13] M. Deitke、C. Clark、S. Lee 等，“Molmo 和 PixMo：面向最先进视觉语言模型的开放权重与开放数据”，*arXiv*预印本 *arXiv:2409.17146*，2024年。[14] F. Liu、K. Fang、P. Abbeel 和 S. Levine，“MOKA：通过基于标记的视觉提示实现开放世界机器人操作”，*arXiv*预印本 *arXiv:2403.03174*，2024年。[15] B. Ding、J. Xie、J. Nie、J. Cao、X. Li 和 Y. Pang，“VFMM3D：借助视觉基础模型释放图像在单目三维物体检测中的潜力”，*arXiv*预印本 *arXiv:2404.09431*，2024年。[16] R. Zhang、H. Qiu、T. Wang、Z. Guo、Z. Cui、Y. Qiao、H. Li 和 P. Gao，“MonoDETR：用于单目三维物体检测的深度引导 Transformer”，载于*2023IEEE/CVF (ICCV)*，2023年，第9121–9132页。