

SpatialBot: 基于视觉语言模型的精确空间理解

Wenxiao Cai^{1,2,3*}, Iaroslav Ponomarenko⁴, Jianhao Yuan⁵, Xiaoqi Li⁴,
Wankou Yang⁶, Hao Dong⁴, Bo Zhao^{1,3†}

¹ 上海交通大学人工智能学院 ² 斯坦福大学

³ 北京智源人工智能研究院 (BAAI) ⁴ 北京大学 ⁵ 牛津大学 ⁶ 东南大学

摘要

视觉语言模型 (Vision Language Model, VLM) 在 2D 图像理解上已取得优异表现, 但空间理解能力仍然薄弱, 而空间理解是具身智能 (Embodied AI) 的基础。本文提出 SpatialBot, 通过同时利用 RGB 图像与深度图提升空间理解能力。为训练 VLM 的深度感知能力, 本文构建 SpatialQA 与 SpatialQA-E 数据集, 涵盖多种场景与具身任务下的多层次深度相关问题。同时提出 SpatialBench, 从不同层次全面评测 VLM 的空间理解能力。在本文的空间理解基准、通用 VLM 基准以及具身智能任务上的大量实验表明, SpatialBot 带来了显著提升。模型、代码与数据集见 <https://github.com/BAAI-DCAI/SpatialBot>。

1. 引言

近期, 视觉语言模型 (Vision Language Model, VLM) [7, 28, 46, 52, 66, 76] 依托基于视觉编码器的感知与基于语言模型的推理, 在通用 2D 视觉理解与推理上表现出较强能力。然而, 仅从 2D 图像理解空间信息对 VLM 仍然困难, 而这正是实现各类真实世界任务 [5, 22, 36, 43, 58] 的关键, 尤其是操作 (Manipulation) [25, 48, 56, 64] 与导航 [18, 44, 72, 79] 等具身智能 (Embodied AI) 相关任务。

VLM 获得空间理解能力的主要挑战体现在以下方面: 1) 主流 VLM 仅在 RGB 图像上训练, 从未见过深度图, 因此理解深度信息的能力有限; 此外, 其训练任务几乎不需要深度信息即可求解。于是, 直接把深度图输入 VLM 效果很差。2) 缺少专门设计用于训练 VLM 理解深度的数据集。主流 VLM 微调数据集既不提供深度图, 也不包含与深度相关的任务。3) 室内与室外数值深度的尺度不一致, 也是阻碍 VLM 在各类任务中统一处理深度的重要问题。例如, 室内导航与操作等任务需要毫米级精度, 而室外任务无需如此高的精度, 却要求更大的深度范围。

* 本工作完成于 Wenxiao Cai 在上海交通大学访学以及在 BAAI 实习期间。

† 通讯作者: bo.zhao@sjtu.edu.cn。

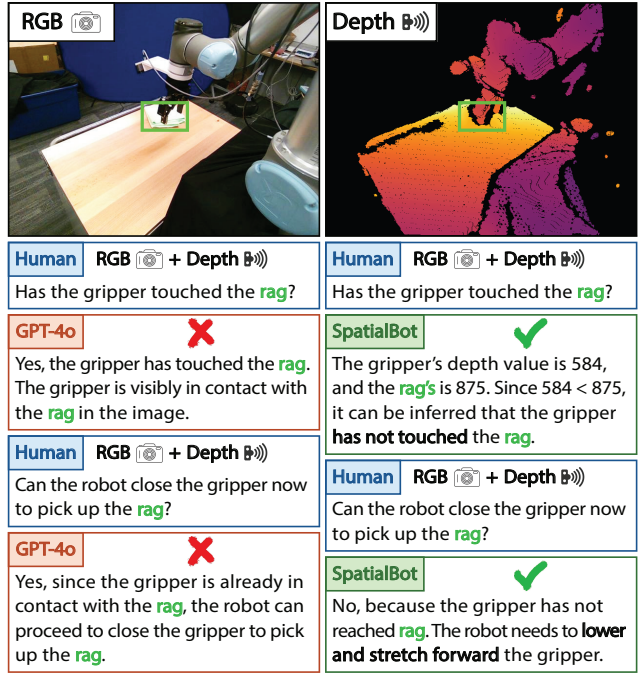


图 1. SpatialBot 的空间理解能力优于 GPT-4o。仅凭单张 RGB 图像, 即便是人也无法判断夹爪 (Gripper) 是否已接触抹布。GPT-4o 即使同时看到 RGB 图像与深度图也判断错误。SpatialBot 先从深度图获取目标物体的深度信息, 进而判断出夹爪尚未接触抹布。

针对这些挑战, 本文提出 SpatialBot, 可通过深度图精确理解空间信息并执行机器人操作。本文设计渐进式训练方案: 先用所提出的包含通用对话任务的 SpatialQA 数据集提升 VLM 的通用空间理解能力, 再借助采集的机器人操作任务数据集 SpatialQA-E, 将该空间理解能力迁移到具身任务上。本文设计了从低层到高层、高度依赖空间理解的多种专用问答任务, 包括低层的深度估计, 中层的目标检测、指代问答 (Referring QA) 与深度比较, 以及需要深度推理的高层任务, 如在通用对话和机器人操作中理解空间关系。为使模型

能够准确获取深度信息，本文设计了 Depth API，允许模型查询单个像素或区域的深度值。

本文用 *SpatialBench* 验证 VLM 的空间理解能力，该基准由人工标注的空间理解与空间推理问答对构成。本文还将 *SpatialBot* 部署到机器人上执行操作任务，例如拿起中间的茶杯并放到最近的砧板上，如图 6 所示。实验结果验证了 *SpatialBot* 能够在三个层次上理解深度。此外，实验还验证了在 *SpatialQA* 上微调 VLM 可以提升其在 MME [23]、MMBench [54] 等通用 VLM 基准上的表现。最后，机器人操作能力表明了 *SpatialBot* 的应用潜力。综上，本文的主要贡献如下：

- 本文提出 *SpatialBot*，在通用视觉识别、空间理解和机器人操作上均取得良好表现。
- 本文构建了大规模 RGB-D 视觉问答数据集 *SpatialQA* 用于训练 *SpatialBot*，以及用于评测 VLM 空间理解性能的 *SpatialBench*。为全面分析深度，设计了三个层次的任务。
- 本文在涉及空间推理的具身任务上微调并部署 *SpatialBot*，并发布聚焦空间关系的机器人操作数据集 *SpatialQA-E*。

2. 相关工作

2.1. VLM 与 RGB 数据集

近年来，视觉语言模型（Vision Language Model, VLM，亦称多模态大语言模型 Multi-modal Large Language Model, MLLM）取得了显著进展 [35]。LLaVA [52] 开创了视觉指令微调（Visual Instruction Tuning），后续工作 [7, 28, 50, 55, 78] 在此基础上引入了规模更大的数据集 [87] 和不同的大语言模型（Large Language Model, LLM）骨干网络 [1, 6, 16, 68]。这些 VLM 主要处理感知 [23]、推理 [54] 和 OCR [45, 81] 相关任务。此外，部分工作在 VLM 之外引入编码器-解码器结构，以完成像素级定位（Pixel-level Grounding）任务 [42, 74, 77, 80, 84, 85]。然而，它们在计数与枚举 [20, 30] 以及空间关系理解 [33] 上的表现欠佳。本文认为，仅凭单目 RGB 图像理解整个空间对 VLM 而言负担过重；引入深度信息可有效增强 VLM 的空间理解能力。

2.2. 通用问答与具身场景中的空间理解

空间理解要求 VLM 的场景理解超越 2D RGB 图像，这在机器人抓取 [19] 等高精度任务中尤为关键。空间理解可借助点云 [19, 73] 或深度图 [33] 实现。已有研究尝试直接从单目 RGB 图像完成深度估计 [47] 与 3D 检测 [17]，但在度量深度估计（Metric Depth Estimation）上精度有限。*SpatialVLM* [12] 和 *SpatialRGPT* [14] 仅从 2D 图像推断空间关系。然而在机器人任务中（例见图 1），来自传感器的深度信息对空间理解不可或缺。近期单目深度估计（Monocular Depth Estimation, MDE）发展迅速。借助大量无监督数据 [8, 75] 与合成数据 [37]，MDE 已能在各类场景下准确估计深度 [60]。因此，本文利用 MDE 为 VLM 所用的 RGB 图像补

充深度信息，从而提升其空间理解能力。尽管单目深度估计模型能力较强，训练大模型直接估计深度并非总是可行。在具身智能（Embodied AI）场景中，所需的精确深度信息必须来自硬件设备，深度估计模型无法达到该精度。此外，让 VLM 从单张 RGB 图像精确理解空间已被证明极为困难 [17, 47]。为将空间理解能力拓展到具身场景，本文提出 *SpatialQA-E*。据我们所知，这是首个聚焦空间关系的操作数据集。*SpatialBot* 采用与 RT [10, 11]、Octo [67]、OpenVLA [39] 等领先视觉-语言-动作模型（Vision-Language-Action Model, VLA）相似的模型结构，并通过在 *SpatialQA-E* 上训练获得操作任务所需的空間知识。

3. SpatialBot

本文利用深度信息引导 VLM 理解空间 [32, 57]，原因在于：相比点云，深度信息更易采集与处理。RGB-D 相机价格低廉，多数机器人都配备此类相机以实时获取 RGB 图像与深度图。此外，得益于单目深度估计（Monocular Depth Estimation, MDE）的强大能力，可以快速且低成本地把大规模 RGB 数据集转换为 RGB-D 数据集。因此，本文引入深度图用于空间理解，并构建包含 RGB-D 图像与深度相关问答对的 *SpatialQA* 数据集。本节详述数据集构建流程：采集 RGB 图像与深度图、用 MDE 从 RGB 图像估计深度图、统一深度图格式、生成用于 VLM 训练的基础 VQA，以及生成深度相关 VQA。该流程易于扩展规模，可从已有 RGB 数据集构建更大的数据集。

3.1. 深度图编码

本文的深度编码旨在保留全部深度信息供 VLM 使用。难点在于室内与室外的一致性：机器人操作 [56] 和室内导航 [9, 27] 等室内场景可能要求毫米级精度，而室外场景的深度值范围很大。已有方法通常采用序数编码（Ordinal Encoding）[24, 75]，但这类编码无法参与基本数学运算。为此，本文用 uint24 或三通道 uint8 存储深度值，以毫米为单位，范围从 1mm 到 131.071m。本文直接保存原始深度值，把后续计算交给 VLM 强大的拟合能力。对于单通道 uint24，直接以毫米为单位，VLM 可从深度图中直接查询所需数值。对于三通道 uint8 图像，本文把数值分布到更宽的范围：三个通道的单位分别为 2^0 、 2^5 和 2^{10} 毫米，各通道分别有 2^5 、 2^5 和 2^7 个可能取值。对于尺寸为 (H, W) 的图像，要把深度值 $d_{H,W}$ （单位毫米）存入三通道 uint8 图像 $I_{H,W}^3$ ，按如下方式编码图像 I ：

$$I_{h,w}^0 = (d_{h,w} / 2^{10}) * 2^1, \quad (1)$$

$$I_{h,w}^1 = (d_{h,w} / 2^5) * 2^3, \quad (2)$$

$$I_{h,w}^2 = (d_{h,w} \% 2^5) * 2^3. \quad (3)$$

第一个通道选取 2^{10} mm 作为单位，是考虑到机器人学中众多桌面抓取任务 [13, 21, 34, 38, 63, 70, 71] 的深度范围，其最大深度通常在 1m 左右。单位取得更大会

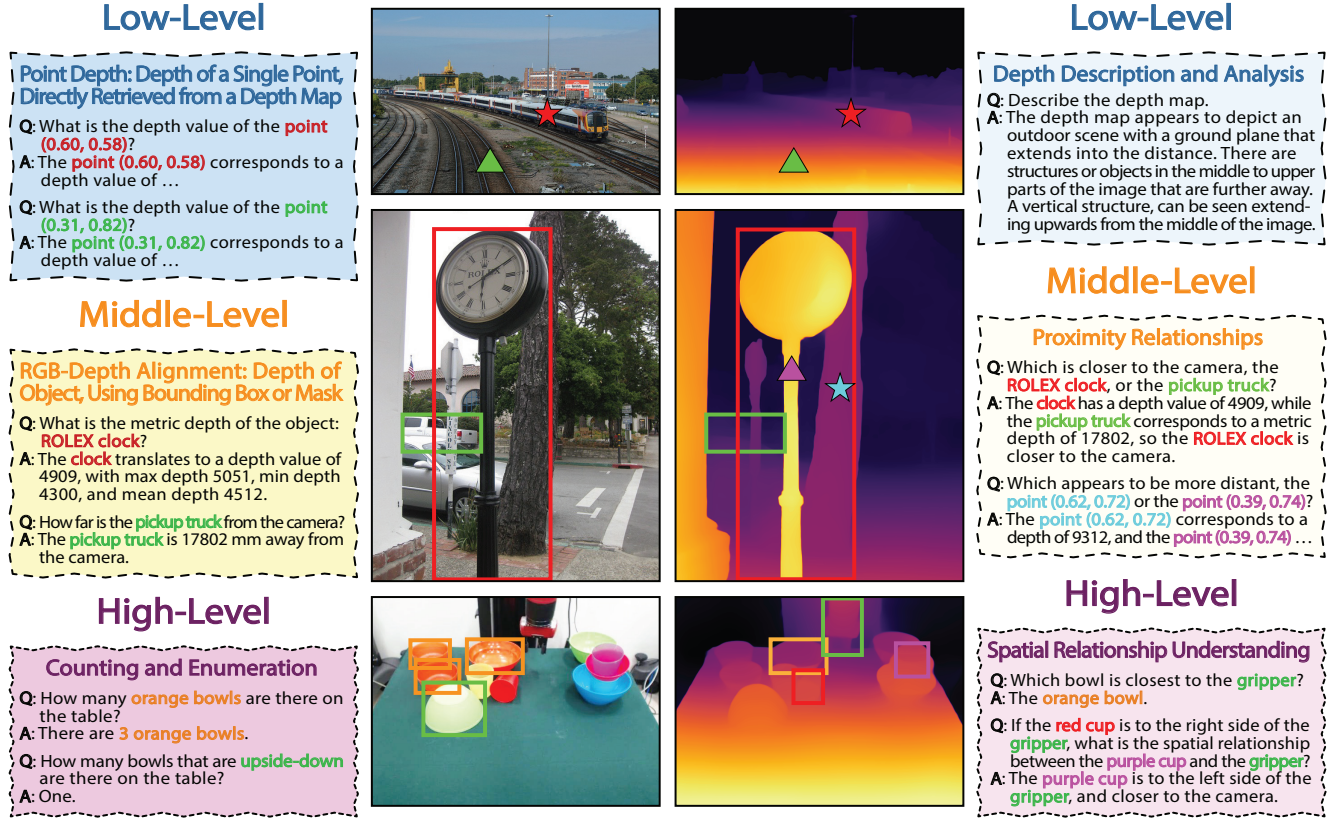


图 2. 所提出的 *SpatialQA* 数据集包含通用 VLM 任务中低层、中层和高层的视觉问答，目标是：(a) 帮助 VLM 理解深度图；(b) 让 VLM 学会对齐 RGB 图像与深度图；(c) 使 VLM 通过同时理解 RGB 图像与深度图更好地完成高层任务，因为深度图提供了清晰的边界信息和空间关系。

导致第一个通道在多数场景下几乎全为零。同理，本文用 2 和 8 作为乘数，以确保三通道深度图之间有更好的区分度。VLM 能够轻松学到本文编码方式与实际深度值之间的关系，实验也验证了这一点。

若原始数据包含传感器估计的深度，则直接使用原始深度值；否则采用 ZoeDepth [8] 模型进行 MDE 估计，因为该模型同时考虑室内与室外场景，并能在这两类场景下准确估计度量深度 (Metric Depth)。注意，本文不使用 MiDaS [60] 等相对深度模型。直接把相对深度 d_r 取倒数当作实际深度 d 是错误的。设图像中的最大、最小深度分别为 d_{max} 和 d_{min} ，转换应遵循：

$$A = \frac{1}{d_{min}} - \frac{1}{d_{max}}, \quad (4)$$

$$B = \frac{1}{d_{max}}, \quad (5)$$

$$d = \frac{1}{(A * d_r + B)}. \quad (6)$$

在问答中询问深度值却以 $\frac{1}{d_r}$ 作为标签 [47]，这种做法是错误的。 $\frac{1}{d_r}$ 虽能反映深度的相对大小，却不保持比例关系（例如 $\frac{1}{d_r} = 0.4$ 并不意味着深度是 0.2 的两

倍）。只有当最大深度为无穷大时， $\frac{1}{d_r}$ 才可视为真实深度乘以一个尺度因子。

3.2. 物体的深度描述

SpatialQA 是 VQA 数据集，本文模型是标准 VLM (图 4)：输入图像与文本，输出文本。为保持通用性，本文不使用独立的图像编码器，因此 *SpatialBot* 无法输出像素级信息。直观上，物体的中心点即可代表其深度；但以杯子为例，内外表面的深度差异显著，单个数值无法准确表示深度。因此，在掩码 (Mask) 可用时，本文用最大、最小、均值和中心点四个深度值描述物体深度。考虑到掩码与深度图都无法做到完全精确，本文取第 95 和第 5 百分位值作为最大与最小深度值。Visual Genome (VG) [41] 中的边界框 (Bounding Box) 很不准确，实验发现用这些边界框提示 SAM [40] 会得到质量不佳的掩码。此时，为避免错误深度误导模型，退而仅用边界框中心点的深度来描述深度。

3.3. 图像来源

RGB-D VLM 数据集应包含帮助 VLM 理解图像的细节问答，涵盖推理、对话、描述与指代 (Referring) [87]。

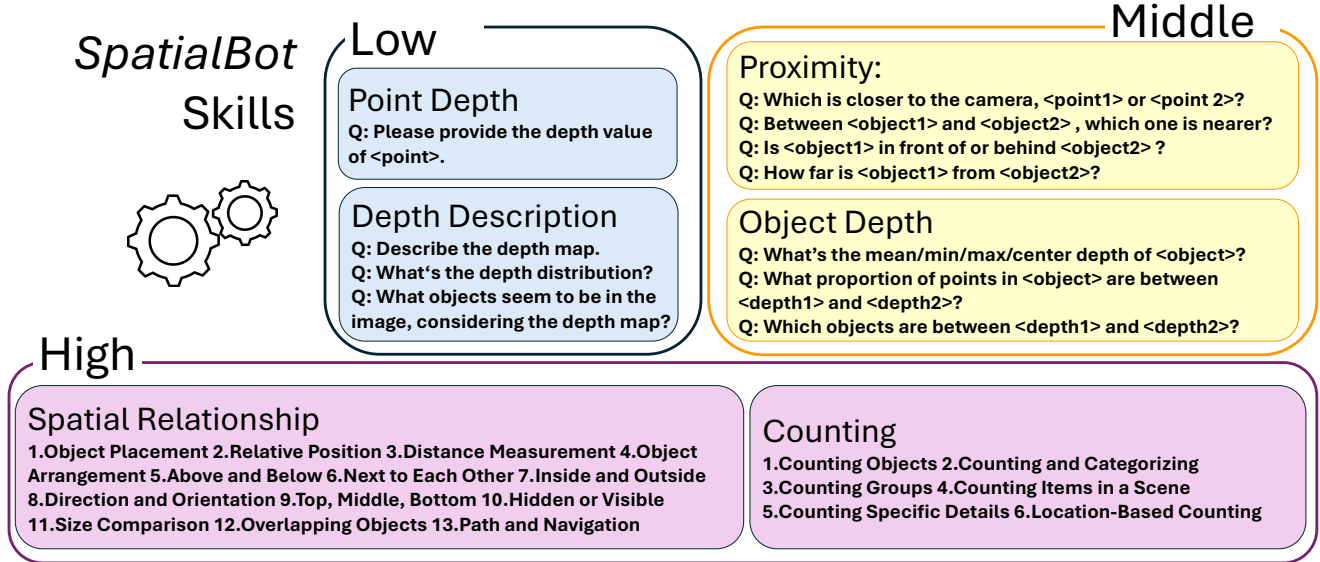


图 3. *SpatialBot* 掌握三个层次的能力: (a) 理解深度图; (b) 对齐 RGB 与深度输入, 并进行远近比较; (c) 利用 RGB-D 完成空间关系理解与计数。

其中需要具体的物体描述, 例如图 4 中, woman (女人) 不是好的描述, 而 the woman in the middle (中间的女人) 或 the woman standing tallest (站得最高的女人) 才是好的描述。已有的图像描述 (Captioning)、定位 (Grounding) 与分割模型 [15, 53, 61, 88] 无法生成如此详细且具体的描述。为此, 本文主要基于已包含详细问答的 VLM 数据 [28] 构建 *SpatialQA*。

SpatialQA 主要包含三类数据来源: COCO [51]、VG [41] 和 Open X-Embodiment (RTX) [56]。因此, 本文以包含 COCO 与 VG 的 Bunny_695k [28] 为基础。Bunny_695k 的图像问答覆盖推理、细节描述、定位等内容, 本文在此基础上补充了深度相关的问答对。本文使用 Bunny_695k 中的边界框, 并以边界框与中心点提示 SAM [40] 获取掩码, 同时确保 SAM 掩码不超出边界框范围, 再选取置信度最高的掩码。RT-X 整合了大量机器人数据集。对于含有传感器深度数据的子集, 直接使用原始深度; 其余子集使用模型估计的度量深度。本文从中选取 7.5k 张图像并人工标注边界框, 查询物体的深度信息; 对剩余图像则只询问图中若干像素的深度。此外, 本文用 GPT-4o 基于 RTX-7.5k 生成对话, 提示 GPT-4o 关注: 机器人在做什么、机器人应如何完成任务、物体数量、物体位置、位置关系以及物体外观。在 *SpatialQA* 的后续版本中, 本文将纳入来源更广泛的图像。

3.4. *SpatialQA* 流程

为帮助 VLM 理解深度输入, 并利用深度信息完成空间关系理解、计数与枚举等高层任务, 本文设计了三步式问答流程。该流程力求有效且易于复现: (a) 流程逐步引导 VLM 学会理解深度、对齐深度与 RGB,

并在高层任务中用深度进行复杂推理; (b) 已有 RGB 数据集可借助该流程轻松转换为 RGB-Depth 数据集。*SpatialQA* 流程见图 5, 待学习的能力见图 3。

低层。为让 VLM 理解深度图并学会从中查询信息, 本文询问点的深度值。VLM 应学会直接从深度输入中读取深度值, 并把点坐标与图像像素对应起来。同时, 由于视觉编码器在预训练中未见过深度图, 本文也期望编码器与投影器学会把深度图与 RGB 图像一并编码。本文还让 *SpatialBot* 在仅给定深度图的情况下描述该深度图, 并推断图中可能有什么。

中层。VLM 已学会编码深度图并从中查询信息后, 接下来应学会使用深度信息。此外, 由于图像与深度同时输入 VLM, 模型还应了解二者之间的关系。首先, 询问远近关系 (Proximity), 即哪个点更近、哪个点更远。其次, 让 VLM 学会用中心点深度、最小深度、最大深度和平均深度描述物体或区域的深度; VLM 应学会在 RGB 图像中定位物体, 再从深度输入中查找深度信息。第三, 询问物体之间的远近关系。

高层。VLM 此时已能理解深度输入、把深度与 RGB 对齐, 并掌握空间世界中的部分远近关系, 本文据此设计任务, 帮助 VLM 在更高层次上应用深度。模型看到深度图后, 物体边界及其周围环境变得更清晰, 因此深度图有助于定位与计数任务。此外, 在 *SpatialQA* 中模型获得了对空间的清晰认知, 有助于判断空间关系与位置关系。

3.5. *SpatialQA-E*

本文提出 *SpatialQA-E*, 把空间理解与空间推理能力扩展到具身任务。本文使用 7 轴 Franka Research 3 机械臂抓取桌面物体, 在移动过程中避开障碍物, 并把物

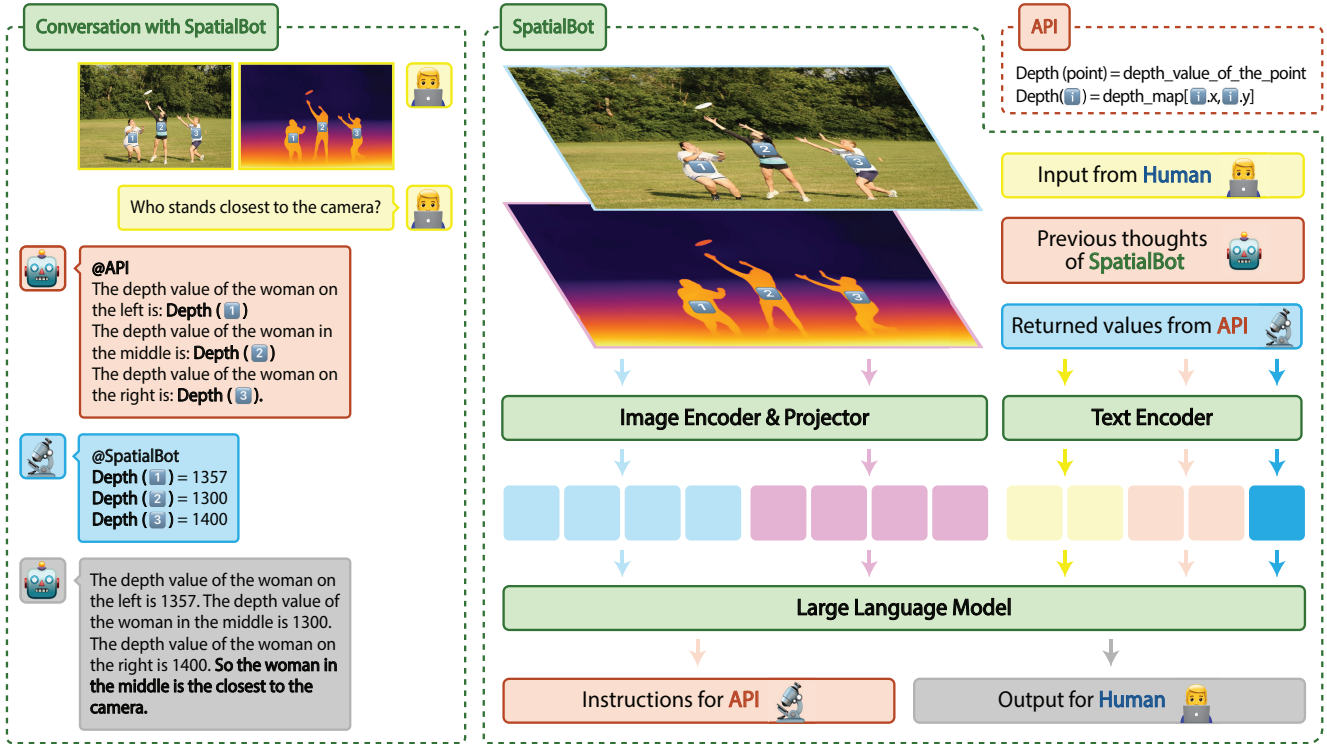


图 4. *SpatialBot* 的架构。输入为一对 RGB 图像与深度图，其中深度图可选。当需要准确的深度信息时，*SpatialBot* 可选择调用 Depth API。

体放到桌上的砧板上。语言指令中包含空间关系，因此模型需要在操作中学习空间推理。*SpatialQA-E* 共包含 2000 条轨迹段 (Episode)。数据集由 4 个步骤组成，见图 6 与图 7：

- 学会抓取与放置茶杯、球、香蕉等物体。
- 找到特定物体与目标位置。数据集包含从机器人视角或人（相机）视角出发、以原级、比较级 (-er) 和最高级 (-est) 表达的空间关系：
 - 位置类：左/ 右/ 中/ 上/ 下，在...上/ 在...内/ 内部/ 外部
 - 尺寸类：高/ 矮/ 大/ 小/ 宽/ 窄/ 大/ 小
 - 错觉类 (Illusion)：本文为物体拍照并打印出来，再把打印件放在桌上。打印件看起来很真实，模型需要通过视觉线索分辨打印件与真实物体，例如深度信息（打印件是平的）和阴影。
- 在移动物体时，机器人需要避开障碍物。

3.6. *SpatialBot* 在具身任务中的应用

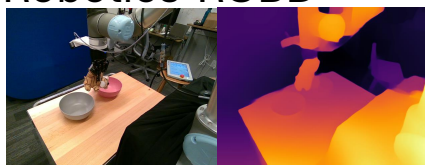
SpatialBot 在 *SpatialQA-E* 上微调后用于具身任务。简言之，它是支持多帧 RGB 或 RGB-D 输入的视觉-语言-动作 (Vision-Language-Action, VLA) 模型。机器人操作任务定义为：在当前时刻 t ，给定历史与当前图像观测 $x_{j=0}^t$ (RGB 或 RGB-D)，模型应学习策

略 $\pi(\cdot | i, x_{j=0}^t)$ 。动作 a_t 从 π 中采样并施加到机器人上。对于配备二指末端执行器 (End Effector) 的机器人，动作空间可表示为 7 自由度 (Degree of Freedom, DoF) 向量 $(\Delta X, \Delta Y, \Delta Z, \Delta R, \Delta P, \Delta Yaw, C)$ ，分别表示位姿 XYZ 的增量变化、旋转 RPYa (滚转、俯仰、偏航) 的增量变化，以及夹爪开合 C 。动作空间中位置与旋转的增量变化被编码为 101 个可能取值，从 0、0.01 直到 1。模型直接输出 7 自由度动作的文本。一段示例对话：'User: What should the robot do to pick up the biggest teacup and move it to the left cutting board? Answer with robot parameters. - SpatialBot: The robot should <0.17, 0.51, 0.44, 0.62, 0.83, 0.07, 1>'。随后把输出解码为机器人控制信号，以控制机器人每一帧的运动。如果模型在微调阶段直接回答机器人参数，本文发现它只能响应机器人相关的问题。为支持多任务训练，本文在机器人回复中加入少量自然语言成分，例如 'The robot should'，然后在机器人数据与通用问答数据（如 *SpatialQA-E* 和 *SpatialQA*）上共同训练模型。本文让模型在机器人任务中预测特殊 token，以保持模型在通用对话中的数值推理能力。本文预测每一帧的增量位姿 (Delta Pose) 而非目标位姿。这一选择把动作空间的每个维度划分为 100 个区间 (Bin)，从而实现更精确的机器人控制。此外，本文发现增量位姿比目标位姿更难学习，表现为损失下降更慢。在部分

Depth
Sensor
Data



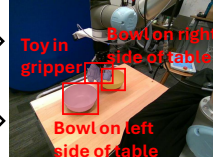
Robotics-RGBD



Annotate

GPT

Bbox & Mask



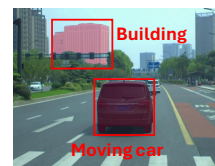
Q: What is the robot doing?
GPT: The robot is using a gripper to hold a green cloth and ...

CV-Depth



Annotate

Bbox & Mask



RGB
Camera
Data

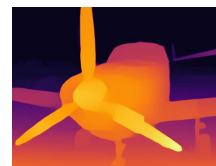


MLLM-General

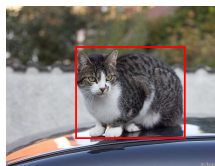


Q: What is in the image?
A: The image shows a small propeller aircraft parked on an airport tarmac. The aircraft has a blue and white color scheme with...

MDE

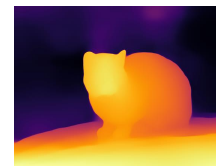


MLLM-Referring

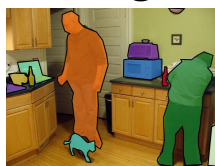


Q: Please provide the bounding box of object: a cat sitting on car
A: [bbox]
Q: What is [bbox]?
A: a cat sitting on car

MDE



CV-Segmentation



GT:
cat: [mask1]
man: [mask2]
man: [mask3]
...

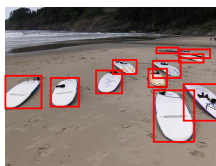
MDE

To QA

Q: What are the people doing?
A: The tall man seems to be looking for something on the table, ...



CV-Detection

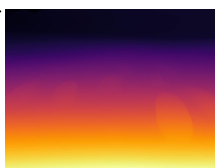


GT:
boat1: [bbox1]
boat2: [bbox2]
...

MDE

SAM

To QA



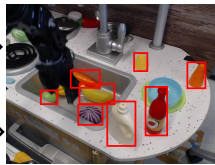
Q: How many boats are there in the image?
A: There are 11 boats in total, and their bounding boxes are [x,y],[x,y],...

Robotics-RGB



MDE

Annotate



To QA

GPT

Q: What can be seen in the image?
A: On the table, there's a ketchup bottle, a carrot...
Q: How should the robot proceed to pick up the carrot?
GPT: The robot should position its gripper above ...

图 5. 图像来源以及 RGB 到 RGB-Depth 数据集的转换流程。构建 *SpatialQA* 数据集需要 RGB 与深度信息、图像的描述或对话、边界框或掩码，以及物体对应的详细描述。

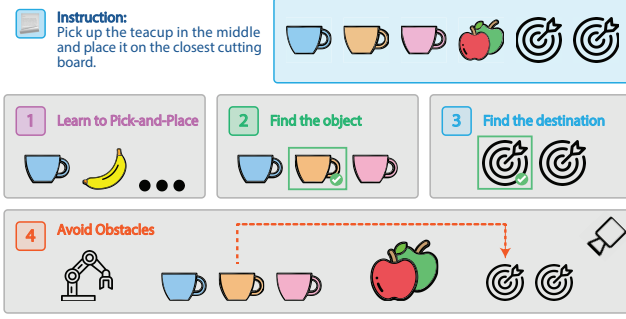


图 6. *SpatialQA-E* 在机器人操作中引入空间关系。

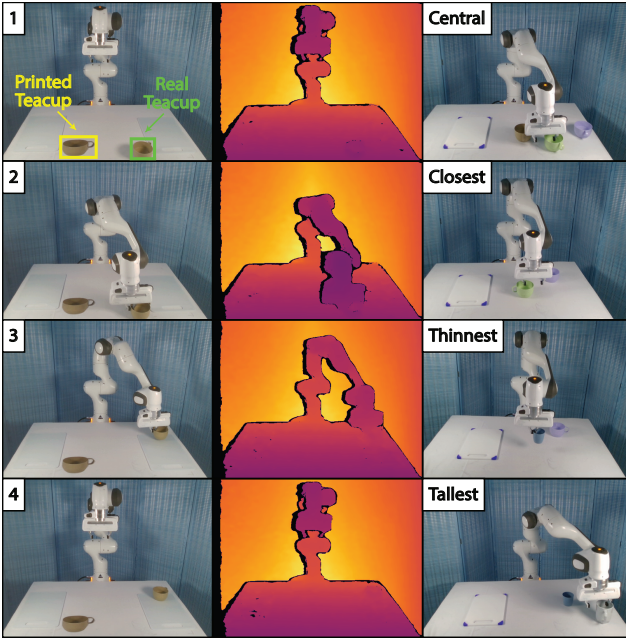


图 7. *SpatialQA-E* 示例。左：抓起真实茶杯并把它放到相机视角下右侧砧板上的 4 个步骤。本文打印了一只茶杯作为干扰项，从深度图中更容易分辨真实茶杯与打印茶杯。右：*SpatialQA-E* 中的 4 个样例设置，其中均指定了空间关系。

数据上增量位姿的损失完全不下降，模型彻底崩溃，无论输入如何都输出相同数值。缓解该问题的直接办法是按指数级增加训练数据量。

3.7. *SpatialBench*

为评估 VLM 在高层任务上的性能，本文标注了 *SpatialBench*。在 120 张图像上提出以下问题：

- [Object 1] 是否接触到/够到了 [Object 2]?
- 若干 [Object] 之间的空间关系是什么?
- 计数与枚举。
- 物体之间的尺寸比较。

所有问题均为是/否或多选题形式。此外，*SpatialBench* 还包含基于 MME [23] 数据集图像的深度与远近问题，

以及本文人工标注的 80 张图像（每张 3 个边界框）。

3.8. *SpatialBot* Depth API 架构

SpatialBot 采用 VLM 结构：图像经图像编码器与多模态投影器（Multi-modal Projector）处理后转换为 token，再与文本 token 一同送入 LLM，最终输出回复。为使模型能准确获取深度信息，本文设计了 **Depth API**。当 *SpatialBot* 的输出包含形如 Depth(point) 的文本时，该 API 会在对应深度图中查询该点的深度值，并把深度值回传给 *SpatialBot*。结合用户问题与 API 返回值，*SpatialBot* 给出最终答案。模型可调用该 API 获取特定点的精确深度值。例如，当 *SpatialBot* 想知道某个物体的深度信息时，先确定该物体的边界框，再用边界框中心点调用 Depth API；若想获取该物体的深度范围，则先观察图中哪些点对应最大与最小深度值，再用这些点的坐标调用 Depth API。不过，为增强模型对深度图本身的理解，训练时本文只允许 *SpatialBot* 在部分数据上调用 API，其余数据必须直接依据深度图回答物体的深度。

4. 实验

首先验证 *SpatialBot* 具备理解深度、从深度图中提取信息并完成高层任务的能力。随后观察到：引入深度图后，模型在 MME [23]、GQA [31] 等通用 VQA 任务上的性能得到提升，说明在 *SpatialQA* 上训练有助于 VLM 在通用任务上取得更好表现。最后，RT-X [56] 上的实验表明，*SpatialBot* 在机器人操作任务中可从深度理解中获益。

实现细节。*SpatialBot* 基于 VLM 系列 Bunny [28] 设计，基座 LLM 采用 Phi-2-3B、Phi-3-4B [1]、QWen-1.5-4B [6] 与 Llama-3-8B [68]。*SpatialBot* 的模型架构见图 4。图像编码器为 SigLIP [83]，图像分辨率 384x384。机器人操作任务中则采用 QWen-1.5-0.5B 与 CLIP [59]，图像分辨率 336x336。模型先在来自 LAION-2B [62] 的两百万图文对（Bunny-pretrain-LAION-2M [28]）上预训练，再在 Bunny_695k [28] 上微调。学习率保持 $2e-4$ ，多模态投影器（Multi-modal Projector）的学习率为 $2e-5$ ；Llama-3-8B 例外，其两个学习率均减半。操作任务中学习率同样减半。多模态投影器在预训练与微调两阶段均可训练，微调时另外加入 LoRA [29] 模块。训练使用 8 张 A100。在 *SpatialQA* 上，Phi-2 [1] 约需 15 小时。

4.1. 空间理解

首先验证 *SpatialBot* 能够从深度图或 Depth API 获取准确的度量深度（Metric Depth）值，并判断远近关系（Proximity）——二者分别对应 *SpatialQA* 中的低层与中层任务。实验使用 *SpatialBench* 提供的边界框（Bounding Box）与度量深度，随后询问图中随机点与物体的深度。向 VLM 给出目标物体名称时，分别考察提供与不提供边界框两种设置。对于真值深度 d_{gt} 与 VLM 估计的深度 d_{est} ，以 $\frac{d_{gt}-d_{est}}{d_{gt}} * 100\%$ 衡量深度精

表 1. *SpatialBench* 上的结果。相同基座 LLM 的模型中，最优结果以 **粗体**标注。LLM-RGB 与 LLM-RGBD 均仅在 RGB 图像上训练，分别以 RGB 与 RGBD 输入测试。深度估计中采用 RGB 输入的 *SpatialBot* 等同于 MDE 任务。

模型	Depth $\uparrow$	Position $\uparrow$	Existence $\uparrow$	Counting $\uparrow$	Reaching $\uparrow$	Size $\uparrow$
GPT-4o-RGB	-	70.6	85.0	84.5	51.7	43.3
GPT-4o-RGBD	-	61.8	90.0	85.2	51.7	40.0
Bunny-Phi2-3B-RGB	70.6	50.0	75.0	89.4	51.7	26.7
<i>SpatialBot</i> -Phi2-3B-RGB	84.1	64.7	80.0	88.0	61.7	28.3
Bunny-Phi2-3B-RGBD	85.8	50.0	75.0	90.4	43.3	28.3
<i>SpatialBot</i> -Phi2-3B-RGBD	>99	61.8	80.0	91.7	55.0	26.7
Bunny-Phi3-4B-RGB	32.3	58.8	75.0	91.0	31.7	16.7
<i>SpatialBot</i> -Phi3-4B-RGB	83.2	64.7	75.0	91.0	40	23.3
Bunny-Phi3-4B-RGBD	63.3	52.9	60.0	85.4	31.7	18.3
<i>SpatialBot</i> -Phi3-4B-RGBD	>99	67.7	70.0	91.7	35.0	21.7
Bunny-QWen1.5-4B-RGB	42.2	50.0	75.0	91.6	26.7	15.0
<i>SpatialBot</i> -QWen1.5-4B-RGB	89.9	52.9	75.0	88.6	46.8	18.3
Bunny-QWen1.5-4B-RGBD	74.6	44.1	70.0	90.7	25.0	15.0
<i>SpatialBot</i> -QWen1.5-4B-RGBD	>99	52.9	60.0	90.5	41.7	26.7
Bunny-Llama3-8B-RGB	58.1	50.0	75.0	91.7	38.3	23.3
<i>SpatialBot</i> -Llama3-8B-RGB	85.6	55.9	80.0	91.2	40.0	20.0
Bunny-Llama3-8B-RGBD	64.0	50.0	75.0	90.4	38.3	25.0
<i>SpatialBot</i> -Llama3-8B-RGBD	>99	53.0	75.0	90.4	45.0	20.0

表 2. 通用 VLM 基准上的结果。相同基座 LLM 的模型中，更优结果以 **粗体**标注。RGB-D 输入仅用于 MME。表中报告 Bunny 以 RGB 训练、并分别以 RGB/RGB-D 测试的结果，用斜杠分隔。*SpatialBot* 在 RGBD 上训练，并在 MME 上以 RGB/RGB-D 测试。

模型	MME ^P $\uparrow$	MME ^C $\uparrow$	MMB ^T $\uparrow$	MMB ^D $\uparrow$	SEED-I $\uparrow$	VQA ^{v2} $\uparrow$	GQA $\uparrow$	POPE $\uparrow$
Bunny-Phi2-3B	1472/1474	286/285	67.90	68.90	69.91	78.98	61.52	86.21
<i>SpatialBot</i> -Phi2-3B	1483/ 1487	310/ 312	70.12	68.56	70.85	79.80	62.28	87.04
Bunny-Phi3-4B	1417/1364	308/319	70.15	70.74	71.04	80.57	61.18	84.60
<i>SpatialBot</i> -Phi3-4B	1431 /1433	337 /329	73.49	73.11	71.64	80.01	62.16	85.47
Bunny-QWen1.5-4B	1340/1364	251/254	69.56	68.56	70.05	80.63	61.55	85.10
<i>SpatialBot</i> -QWen1.5-4B	1378/ 1406	266/ 285	70.91	69.67	70.36	79.69	62.77	86.09
Bunny-Llama3-8B	1574/1542	342/318	73.67	74.15	72.32	80.50	62.18	85.22
<i>SpatialBot</i> -Llama3-8B	1577 /1576	352 /333	75.78	74.83	72.40	80.94	62.90	85.33

度。借助 Depth API 作答的结果见表 1 中的 Depth 与 Proximity 列。此外还考察远近关系判断。深度图理解与高层任务的对话示例见图 12。

4.2. *SpatialBench*

本文在 *SpatialBench* 上对比模型性能，该基准由位置关系、物体存在性、触达 (Reaching) 与尺寸比较四类任务构成。将 GPT-4o 与在 *SpatialQA* 上训练的模型对比可见：在 *SpatialQA* 上训练的 3B、4B 与 8B 模型取得与 GPT-4o 相当的结果。结果报告于表 1。

4.3. 通用 VLM 基准

本文报告以下通用基准上的结果：MME 感知 [23] (MME^P)、MME 认知 (MME^C)、MMBench [54] 测试集与开发集 (MMB^T 与 MMB^D)、SEED Bench Image [45] (SEED(-I))、VQA [26] 的 test-dev 划分 (VQA^{v2})、GQA [31] 以及 POPE [49] (COCO 验证集上三类的平均 F1 分数)。其中多数基准仅凭 RGB 信息即已足够，因此只在 MME^P 与 GQA 上使用 RGB-Depth 输入——这两个基准包含计数、物体存在性与位置类问题，深度信息预期能在此类情形中带来增益。

4.4. SpatialBot 在具身任务中的应用

本文在 *SpatialQA-E* 上微调 *SpatialBot*, 用于真实机器人的操作任务。该模型可视为支持多帧 RGB 或 RGBD 输入的 VLA 模型。基座 LLM 采用 QWen-1.5-0.5B [6], 视觉编码器采用 CLIP [59]。预训练数据集为 Bunny-pretrain-LAION-2M [28], 微调使用 *SpatialQA-E*。模型以历史四帧预测当前帧的末端执行器增量位置, 可本地运行, 也可通过 ssh/sftp 连接在 RTX 4090 GPU 上运行。实验验证了 *SpatialBot* 能够按照空间指令完成操作任务。图 8 给出 *SpatialBot* RGB 与 RGBD 变体的成功率: 借助深度信息, *SpatialBot* 的抓取与放置更为精确。

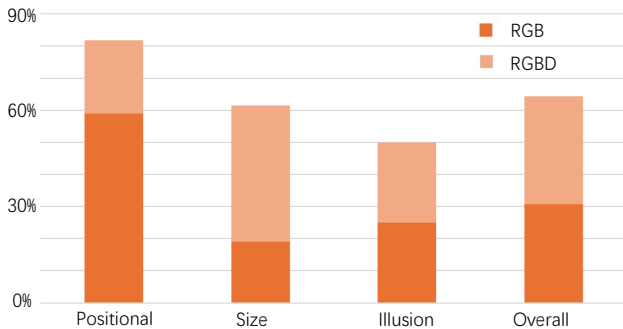


图 8. *SpatialBot* RGB 与 RGBD 变体在抓取与放置 (Pick-and-Place) 任务中的成功率。

5. VLM 的单目深度估计

本文让 VLM 理解深度输入, 从而获得准确的深度信息。读者可能疑惑: 对单个点或物体做单目深度估计 (Monocular Depth Estimation, MDE) 是否困难? 为验证这一点, 本文要求 *SpatialBot* 预测点或物体的深度。结果报告于正文表 1 (*SpatialBot*-RGB 的 Depth 分数)。VLM 似乎尚未具备以纯文本输出方式完成 MDE 的能力。如 [3, 4, 82, 86] 所报告, 生成准确的像素级深度可能需要额外的解码器。本文未采用该做法, 因为这种新结构会限制 *SpatialQA* 的通用性。

6. SpatialBench 元数据

在 *SpatialBench* 中, GPT-4o 就位置关系生成 34 道多选题, 本文以正确选择率作为准确率。GPT-4o 在另外 20 张图像上生成 20 对正向与反向问题。只有模型同时答对同一问题的正向与反向版本, 才算回答正确。在 20 张图像上, 人工标注专家选定一类物体并标注其数量, 再要求模型对该类物体计数。人工标注专家还分别在 20 张图像上设计正向与反向多选题, 分别考察物体 A 是否已到达或接触物体 B, 以及尺寸比较。本文先计算模型的正确选择率; 当模型同时答对一对正向与反向问题时, 给予加分。

7. SpatialQA 中的数据生成

本文在约 50k 张图像上提示 GPT, 为 *SpatialQA* 生成深度图理解、空间理解与机器人场景理解数据。数据来源与数据筛选协议见表 3。三项任务各自使用的 GPT 提示词见表 4。样例图像与生成的问答见图 9、10 和 11。此外, 人工标注专家在 RTX 中为每张图像标注 3 个边界框 (Bounding Box), 并在夹爪 (Gripper) 于图像中清晰可见时对其进行标注。

8. 数据集、模型与基准的使用

SpatialBot 的官方仓库为 <https://github.com/BAAI-DCAI/SpatialBot?tab=readme-ov-file>, 其中提供元数据、代码、脚本、检查点 (Checkpoint)、许可证以及资源链接。

SpatialQA 可在 Hugging Face 下载: <https://huggingface.co/datasets/RussRobin/SpatialQA>。

SpatialQA-E 已在 Hugging Face 提供: <https://huggingface.co/datasets/RussRobin/SpatialQA-E>。

SpatialBench 可通过 <https://huggingface.co/datasets/RussRobin/SpatialBench> 获取。

SpatialBot-3B 基于 Phi-2 [1] 与 SigLip [83], 其检查点可在 <https://huggingface.co/RussRobin/SpatialBot-3B> 下载。

SpatialBot 的预训练检查点可通过 Bunny [28] 模型库获取: <https://github.com/BAAI-DCAI/Bunny>。

9. 结论

本文提出 *SpatialBot*, 一族领先的 VLM, 通过在自建的 *SpatialQA* 与 *SpatialQA-E* 数据集上训练, 实现有效的深度理解, 从而在具身智能中完成精确的机器人操作。*SpatialBot* 能够理解深度输入, 并利用深度信息完成视觉问答与具身任务中的空间理解与空间推理。本文还设计了 *SpatialBench*, 从多个方面评测模型的空间知识表现。在本文基准、通用 VLM 基准以及机器人操作部署上的实验结果验证了 *SpatialBot* 相比对比方法的有效性与优越性。

致谢

本工作受国家自然科学基金 (项目号 62306046) 与国家青年人才支持计划 (项目号 8200800081) 资助。

表 3. SpatialQA 中的 RGB 图像与深度来源。深度图或来自传感器（已包含在原始数据集中），或由 ZoeDepth [8] 做 MDE 得到。

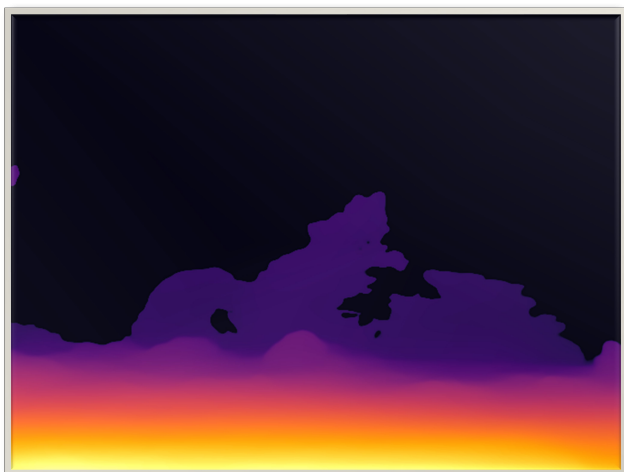
数据	目标	图像数	图像筛选协议与深度图来源
Bunny695k [28]	通用 MLLM 能力	695k	-
VG, COCO	深度图理解	20k	随机选取。提示 GPT 先从深度彩色图做推断，再用 RGB 数据验证推断的正确性。GPT 的推断以问答格式存储。MDE。
KITTI [69]	空间理解	1.75k	从 KITTI 数据集的每个序列中随机选取图像。由于部分场景拍摄于车辆静止或临时停车时（例如红灯处，或前方车辆停车），存在大量重复图像，因此人工过滤掉重复率特别高的图像。由于 KITTI 的深度信息仅包含稀疏点，且 ZoeDepth 已在 KITTI 上微调，本文使用 MDE。
NYU Depth v2 [65]	空间理解	1.5k	采用全部图像与传感器深度图像。
RT-X [56]	机器人场景理解	7.5k	为每张图像标注 3 个边界框 (Bounding Box)，并在夹爪 (Gripper) 可见时对其标注；同时提示 GPT 生成通用问答。若子集提供深度图则直接使用，否则用 MDE 估计深度。
SA-1B [40]	空间理解	15k	从 SA-1B 中随机选取真实世界图像，并提示 GPT-4o 生成关于空间关系的对话。MDE。
2D-3D-S [2]	空间理解	2.9k	从 2d3ds 中随机选取图像，并人工剔除其中物体数不超过 3 个的图像。传感器深度。

表 4. SpatialQA 中使用的 GPT 提示词。

目标	GPT 提示词
深度图理解	Design a conversation between you and a human talking about the depth map. The human asks you to describe the depth map. You should focus on depth value predictions. The colors just represent depth values. Do not directly mention colors on the image in your response, instead, mention the depth distribution they stand for. Looking at the depth map, you should also infer what may be in the image. If something really exists in the rgb image, and can be inferred from the depth map, you can mention they in your response. If possible, pay attention to spatial relationships. When referring to spatial relationships, such as left and right, you should use the real-world left and right, rather than those in the image coordinate system.
空间理解	Design a conversation, consisting of no more than 3 Question-Answer pairs, between you and a person asking about this image. The content within the conversation should be logically connected. You should think of what are spatial relationships of objects in the image. Then generate the conversation according regarding the spatial relationships. Spatial relationships can be about, but not limited to these categories: positional (left/right, below/above, behind/front), distance (further/closer to the camera, further/closer to something), size(big/small, tall/short, wide/thin), reach (has A touched/reached B physically). When describing spatial relationships, always use the real-world orientation as if you are standing in the real scene. e.g. when using right side of object, it should talk about what is on the right side of object in the real world, not on the right side of image'. Only describe the things that you are sure about.
机器人场景理解	Design a conversation, consisting of no more than 3 Question-Answer pairs, between you and a person asking about this image. The content within the conversation should be logically connected. You should first think of robot task: what may the robot want to do with the objects. And then generate the conversation according to robot task. The conversations can include what robot are doing, how should the robot finish robot task, object count, object position, positional relationships, object appearance, etc. Only describe the things that you are sure about. Please note that you are talking to a person about the image and robot. You are not the robot, and you are not talking to the robot.

Q: Can you describe the depth map for me?

A: The depth map shows varying depth values across the scene. The bottom portion of the image represents the foreground with the closest depth values, indicating the presence of grass or a field. Towards the middle of the image, there are structures with intermediate depth values suggesting the presence of multiple trees. The upper section depicts the background with the farthest depth values, likely representing the sky. The trees are positioned such that the one to the left is closer than the rightmost one in the scene.



Q: Can you describe the depth map for me?

A: The depth map indicates variations in distance within the scene. The foreground, from the ground plane closest to us, shows the shallowest depth, hinting at the ground level area immediately in front. Further back, shapes that likely represent standing people and possibly objects like posts or bollards can be inferred, with mid-range depth values. Farther back still, we see deeper values, representing a building facade and a vertically elongated structure, probably a lamp post or pole situated in front and towards the middle. It can be inferred that this scene is of a streetscape with people, a sidewalk, and storefronts.

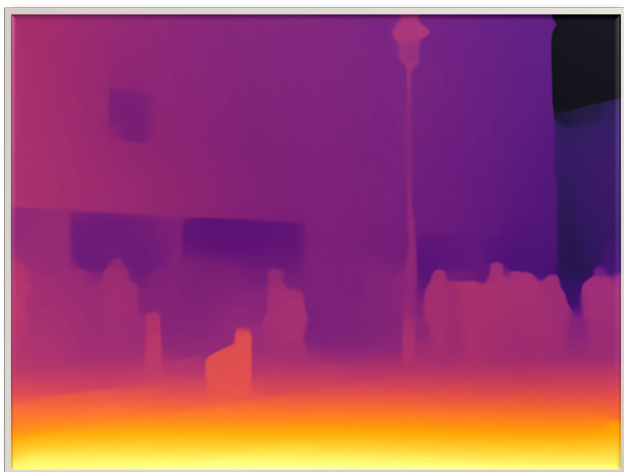


图 9. *SpatialQA* 中深度图理解的样例数据。

Q1: What can you see on the left side of the canal?

A1: On the left side of the canal, there are several wooden boats docked, a large tree, and some buildings, including garages with closed doors.

Q2: What is the woman in the pink coat walking towards?

A2: The woman in the pink coat is walking along the path that runs parallel to the canal, heading towards the background where more buildings and structures are visible.

Q3: Can you describe the scene on the right side of the canal further up?

A3: Further up on the right side of the canal, there are more boats docked and some trees and buildings in the background. There is also another person walking along the path.



Q1: Which penguin seems to be the tallest?

A1: The penguin sitting on top of the igloo seems to be the tallest.

Q2: Which penguin is the smallest in size?

A2: The penguin on the left, with the red hat, seems to be the smallest in size.

Q3: In the real world, which character is bigger, the igloo or the penguin on top of it?

A3: In the real world, the igloo is bigger than the penguin on top of it.

Q1: Has the yellow car reached the motorcycles?

A1: No, the yellow car has not reached the motorcycles. There is a small gap between them.

Q2: Has anyone reached the area with the red stalls?

A2: Yes, there are several people around and inside the area with the red stalls.

Q3: Has the white structure near the motorcycles touched any buildings?

A3: No, the white structure near the motorcycles has not touched any buildings. It is standing independently.



图 10. *SpatialQA* 中空间理解的样例数据。



Q1: What is the robot doing in this image?

A1: The robot appears to be reaching towards two cans placed on a table.

Q2: How many objects are on the table?

A2: There are two cans on the table.

Q3: What is the robot likely trying to do with the cans?

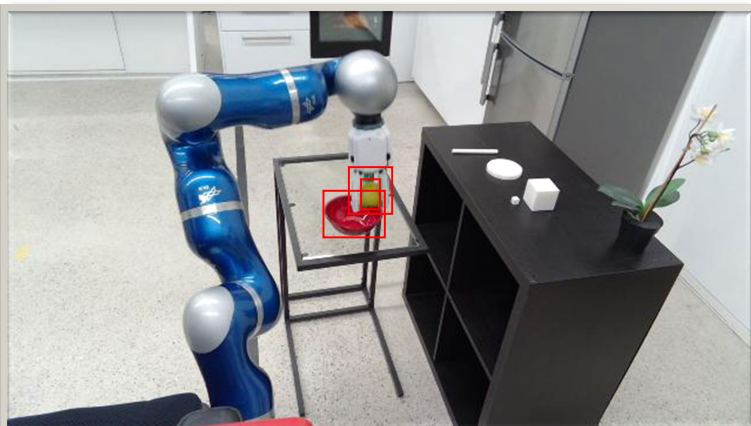
A3: The robot is likely attempting to pick up or move one or both of the cans.

Bounding boxes

Gripper: [0.3,0.117,0.42,0.235],

Can_1: [0.325, 0.147, 0.403, 0.289],

Can_2: [0.232, 0.131, 0.319, 0.3]



Q1: What is the robot doing in this image?

A1: The robot appears to be holding a green object with its gripper arm, placing it on or picking it up from a table.

Q2: What objects are present on the black shelf beside the robot?

A2: There are three objects on the black shelf: a white cylindrical object, a white cube, and a potted plant.

Q3: How should the robot approach picking up the white cube?

A3: The robot should position its gripper arm directly above the white cube, then lower the arm carefully to grip the cube securely before lifting it.

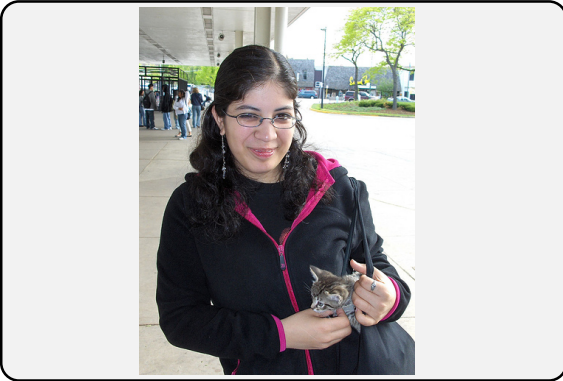
Bounding boxes

Gripper: [0.462, 0.389, 0.514, 0.487],

Bowl: [0.425, 0.443, 0.505, 0.54],

Apple: [0.471, 0.412, 0.499, 0.475]

图 11. *SpatialQA* 中机器人场景的样例数据。



1. What is the depth value of the woman?

SpatialBot-Phi2-3B: 1241 (GT: 1240)

2. What is the depth value of the cat?

SpatialBot-Phi2-3B : 990 (GT: 987)

3. Which is closer to the camera, cat or woman?

SpatialBot-Phi2-3B : cat corresponds to a depth level of 990, and woman corresponds to a depth level of 1241. Since $990 < 1241$, it can be inferred that: cat is closer to the camera, so the answer is: cat. ✓

Bunny-Phi2-3B : The cat. ✓

Bunny-QWen1.5-4B: They are equally close. ✗

LLaVA-34B: The woman. ✗

GPT-4o: The woman. ✗

4. Where is the cat? Answer with the option's letter from the given choices directly.

- (A) The cat is in the woman's bag.
- (B) The cat is in the woman's pocket.
- (C) The cat is out of the woman's bag.
- (D) The cat is not with the woman.

SpatialBot-Phi2-3B: (A) ✓

Bunny-Phi2-3B : (C) ✗

Bunny-QWen1.5-4B : (C) ✗

LLaVA-34B: (C) ✗

GPT-4o: (A) ✓



1. What is the depth value of the sign?

SpatialBot-Phi2-3B: 4137 (GT: 4137)

2. What is the depth value of the white car?

SpatialBot-Phi2-3B : 28482 (GT: 28479)

3. Which is closer to the camera, sign or white car?

SpatialBot-Phi2-3B : The sign corresponds to a depth measurement of 4137, and the white car corresponds to a depth measurement of 28482. Since $4137 < 28482$, it can be inferred that: the sign is closer to the camera. ✓

Bunny-Phi2-3B : The sign. ✓

Bunny-QWen1.5-4B : The sign. ✓

LLaVA-34B: The sign. ✓

GPT-4o: The sign. ✓

4. What is the arrow of sign pointing to? Answer with the option's letter from the given choices directly.

- (A) The car.
- (B) The grass.
- (C) The sky.
- (D) The building.

SpatialBot-Phi2-3B: (B) ✓

Bunny-Phi2-3B : (A) ✗

Bunny-QWen1.5-4B : (D) ✗

LLaVA-34B: (C) ✗

GPT-4o: (A) ✗

图 12. *SpatialBot* 与基线模型的对话示例。先询问物体的深度值，再让模型比较物体之间的深度，最后提出空间关系问题。

参考文献

- [1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 2, 7, 9
- [2] Iro Armeni, Sasha Sax, Amir Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *ArXiv*, abs/1702.01105, 2017. 10
- [3] Dylan Auty and Krystian Mikolajczyk. Learning to prompt clip for monocular depth estimation: Exploring the limits of human language. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2031–2049, 2023. 9
- [4] Dylan Auty and Krystian Mikolajczyk. Language-based depth hints for monocular depth estimation. *ArXiv*, abs/2403.15551, 2024. 9
- [5] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19107–19117, 2021. 1
- [6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2, 7, 9
- [7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023. 1, 2
- [8] S. Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Muller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *ArXiv*, abs/2302.12288, 2023. 2, 3, 10
- [9] Guillaume Le Blanc, Youcef Mezouar, and Philippe Martinet. Indoor navigation of a wheeled mobile robot along visual routes. *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 3354–3359, 2005. 2
- [10] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Anand Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Ho Vuong, F. Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. *ArXiv*, abs/2212.06817, 2022. 2
- [11] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Krzysztof Choromanski, Tianli Ding, Danny Driess, Kumar Avinava Dubey, Chelsea Finn, Peter R. Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil J. Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Sergey Levine, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Pierre Sermanet, Jaspiar Singh, Anika Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Ho Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Ted Xiao, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *ArXiv*, abs/2307.15818, 2023. 2
- [12] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *ArXiv*, abs/2401.12168, 2024. 2
- [13] Lawrence Yunliang Chen, Simeon Adebola, and Ken Goldberg. Berkeley UR5 demonstration dataset. <https://sites.google.com/view/berkeley-ur5/home>. 2
- [14] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. 2024. 2
- [15] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-

- time open-vocabulary object detection. *ArXiv*, abs/2401.17270, 2024. 4
- [16] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023. 2
- [17] Jang Hyun Cho, B. Ivanovic, Yulong Cao, Edward Schmerling, Yue Wang, Xinshuo Weng, Boyi Li, Yurong You, Philipp Krahenbuhl, Yan Wang, and Marco Pavone. Language-image models with 3d understanding. 2024. 2
- [18] Jonathan Crespo, José Carlos Castillo, Óscar Martínez Mozos, and Ramón Barber. Semantic information for robot navigation: A survey. *Applied Sciences*, 2020. 1
- [19] Haoshu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11441–11450, 2020. 2
- [20] Zhen fei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Wanli Ouyang, and Jing Shao. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *ArXiv*, abs/2306.06687, 2023. 2
- [21] Yunhai Feng, Nicklas Hansen, Ziyang Xiong, Chandramouli Rajagopalan, and Xiaolong Wang. Finetuning offline world models in the real world. *ArXiv*, abs/2310.16029, 2023. 2
- [22] Wendy Flores-Fuentes, Gabriel Trujillo-Hernández, Iván Y. Alba-Corpus, Julio César Rodríguez-Quinonez, Jesús E. Mirada-Vega, Daniel Hernández-Balbuena, Fabian Natanael Murrieta-Rico, and Oleg Yu. Sergiyenko. 3d spatial measurement for model reconstruction: A review. *Measurement*, 2022. 1
- [23] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiwu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394, 2023. 2, 7, 8
- [24] Huan Fu, Mingming Gong, Chaohui Wang, K. Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. 2
- [25] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. *arXiv preprint arXiv:2309.02561*, 2023. 1
- [26] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127:398 – 414, 2016. 8
- [27] Faiza Gul, Wan Rahiman, and Syed Sahal Nazli Alhady. A comprehensive study for robot navigation techniques. *Cogent Engineering*, 6, 2019. 2
- [28] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yuezhe Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *ArXiv*, abs/2402.11530, 2024. 1, 2, 4, 7, 9, 10
- [29] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. 7
- [30] Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. Visual hallucinations of multi-modal large language models. *ArXiv*, abs/2402.14683, 2024. 2
- [31] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702, 2019. 7, 8
- [32] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard A. Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew J. Davison, and Andrew William Fitzgibbon. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011. 2
- [33] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. *ArXiv*, abs/2312.14233, 2023. 2
- [34] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi (Jim) Fan. Vima: General robot manipulation with multimodal prompts. *ArXiv*, abs/2210.03094, 2022. 2
- [35] Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Efficient multimodal large language models: A survey. *ArXiv*, abs/2405.10739, 2024. 2
- [36] Zhizhong Kang, Juntao Yang, Zhou Yang, and Sai Cheng. A review of techniques for 3d reconstruction of indoor environments. *ISPRS Int. J. Geo Inf.*, 9:330, 2020. 1
- [37] Bing Wen Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. *ArXiv*, abs/2312.02145, 2023. 2
- [38] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024. 2

- [39] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. [2](#)
- [40] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. [3](#), [4](#), [10](#)
- [41] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32 – 73, 2016. [3](#), [4](#)
- [42] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *ArXiv*, abs/2308.00692, 2023. [2](#)
- [43] Edmund Y. Lam. Computational photography with plenoptic camera and light field capture: tutorial. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 32 11:2021–32, 2015. [1](#)
- [44] Sergey Levine and Dhruv Shah. Learning robotic navigation from experience: principles, methods and recent results. *Philosophical Transactions of the Royal Society B*, 378, 2022. [1](#)
- [45] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *ArXiv*, abs/2307.16125, 2023. [2](#), [8](#)
- [46] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. [1](#)
- [47] Jianing Li, Xi Nan, Ming Lu, Li Du, and Shanghang Zhang. Proximity qa: Unleashing the power of multi-modal large language models for spatial proximity analysis. *ArXiv*, abs/2401.17862, 2024. [2](#), [3](#)
- [48] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18061–18070, 2024. [1](#)
- [49] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023. [8](#)
- [50] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *ArXiv*, abs/2403.18814, 2024. [2](#)
- [51] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. [4](#)
- [52] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023. [1](#), [2](#)
- [53] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499, 2023. [4](#)
- [54] Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *ArXiv*, abs/2307.06281, 2023. [2](#), [8](#)
- [55] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mml: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024. [2](#)
- [56] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. [1](#), [2](#), [4](#), [7](#), [10](#)
- [57] Hailong Pan, Tao Guan, Yawei Luo, Liya Duan, Yuan Tian, Liu Yi, Yizhu Zhao, and Junqing Yu. Dense 3d reconstruction combining depth and rgb information. *Neurocomputing*, 175:644–651, 2016. [2](#)
- [58] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding. In *AAAI Conference on Artificial Intelligence*, 2017. [1](#)
- [59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. [7](#), [9](#)
- [60] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:1623–1637, 2019. [2](#), [3](#)
- [61] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li,

- Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks. *ArXiv*, abs/2401.14159, 2024. 4
- [62] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402, 2022. 7
- [63] Haochen Shi, Huazhe Xu, Samuel Clarke, Yunzhu Li, and Jiajun Wu. Robocook: Long-horizon elasto-plastic object manipulation with diverse tools. *ArXiv*, abs/2306.14447, 2023. 2
- [64] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. *ArXiv*, abs/2109.12098, 2021. 1
- [65] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, 2012. 10
- [66] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multi-modal models. *arXiv preprint arXiv:2312.11805*, 2023. 1
- [67] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. 2
- [68] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. 2, 7
- [69] Jonas Uhrig, N. Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. *2017 International Conference on 3D Vision (3DV)*, pages 11–20, 2017. 10
- [70] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Maximilian Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Ho Vuong, Andre Wang He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, 2023. 2
- [71] Yixuan Wang, Zhuoran Li, Mingtong Zhang, K. Driggs-Campbell, Jiajun Wu, Li Fei-Fei, and Yunzhu Li. D3fields: Dynamic 3d descriptor fields for zero-shot generalizable robotic manipulation. *ArXiv*, abs/2309.16118, 2023. 2
- [72] Steffen Werner, Bernd Krieg-Brückner, Hanspeter A. Mallot, Karin Schweizer, and Christian Freksa. Spatial cognition: The role of landmark, route, and survey knowledge in human and robot navigation. In *GI Jahrestagung*, 1997. 1
- [73] Zhixuan Xu, Chongkai Gao, Zixuan Liu, Gang Yang, Chenrui Tie, Haozhuo Zheng, Haoyu Zhou, Weikun Peng, Debang Wang, Tianyi Chen, Zhouliang Yu, and Lin Shao. Manifoundation model for general-purpose robotic manipulation of contact synthesis with arbitrary objects and robots. 2024. 2
- [74] Jihan Yang, Runyu Ding, Ellis L Brown, Xiaojuan Qi, and Saining Xie. V-irl: Grounding virtual intelligence in real life. *ArXiv*, abs/2402.03310, 2024. 2
- [75] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *ArXiv*, abs/2401.10891, 2024. 2
- [76] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *ArXiv*, abs/2309.17421, 2023. 1
- [77] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *ArXiv*, abs/2310.07704, 2023. 2
- [78] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 2
- [79] Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*, 2024. 1
- [80] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. *ArXiv*, abs/2312.10032, 2023. 2
- [81] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhao Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *ArXiv*, abs/2311.16502, 2023. 2
- [82] Ziyao Zeng, Daniel Wang, Fengyu Yang, Hyungseob Park, Yangchao Wu, Stefano Soatto, Byung-Woo Hong, Dong Lao, and Alex Wong. Wordepth: Variational language prior for monocular depth estimation. *ArXiv*, abs/2404.03635, 2024. 9
- [83] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *2023 IEEE/CVF International Confer-*

- ence on Computer Vision (ICCV), pages 11941–11952, 2023. 7, 9
- [84] Ao Zhang, Wei Ji, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *ArXiv*, abs/2311.04498, 2023. 2
- [85] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chun yue Li, and Jianwei Yang. Llava-grounding: Grounded visual chat with large multi-modal models. *ArXiv*, abs/2312.02949, 2023. 2
- [86] Renrui Zhang, Ziyao Zeng, and Ziyu Guo. Can language understand depth? *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 9
- [87] Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 2, 3
- [88] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *ArXiv*, abs/2304.06718, 2023. 4