# PhysVLM: Enabling Visual Language Models to Understand Robotic Physical Reachability

# PhysVLM:让视觉语言模型理解机器人物理可达性

Weijie Zhou , Manli Tao , Chaoyang Zhao , Haiyun Guo , Honghui Dong , Ming Tang , Jinqiao Wang School of Traffic and Transportation, Beijing Jiaotong University Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences ObjectEye Inc. Guangdong Provincial Key Laboratory of Intellectual Property &Big Data, Guangdong Polytechnic Normal University

周伟杰 ，陶曼丽 ，赵朝阳 ，郭海云 ，董洪辉 ，唐明 ，王金桥 北京交通大学交通运输学院 中国科学院自动化研究所基础模型研究中心 北京ObjectEye公司 广东技术师范大学知识产权与大数据广东省重点实验室

†Corresponding authors: chaoyang.zhao@nlpr.ia.ac.cn, jqwang@nlpr.ia.ac.cn

†通讯作者:chaoyang.zhao@nlpr.ia.ac.cn, jqwang@nlpr.ia.ac.cn

# Abstract

# 摘要

Understanding the environment and a robot’s physical reachability is crucial for task execution. While state-of-the-art vision-language models (VLMs) excel in environmental perception, they often generate inaccurate or impractical responses in embodied visual reasoning tasks due to a lack of understanding of robotic physical reachability. To address this issue, we propose a unified representation of physical reachability across diverse robots, i.e., Space-Physical Reachability Map (S-P Map), and PhysVLM, a vision-language model that integrates this reachability information into visual reasoning. Specifically, the S-P Map abstracts a robot’s physical reachability into a generalized spatial representation, independent of specific robot configurations, allowing the model to focus on reachability features rather than robot-specific parameters. Subsequently, PhysVLM extends traditional VLM architectures by incorporating an additional feature encoder to process the Map, enabling the model to reason about physical reachability without compromising its general vision-language capabilities. To train and evaluate PhysVLM, we constructed a large-scale multi-robot dataset, Phys100K, and a challenging benchmark, EQA-phys, which includes tasks for six different robots in both simulated and real-world environments. Experimental results demonstrate that PhysVLM outperforms existing models, achieving a 14% improvement over GPT-4o on EQA-phys and surpassing advanced embodied VLMs such as RoboMamba and SpatialVLM on the RoboVQA-val and OpenEQA benchmarks. Additionally, the S-P Map shows strong compatibility with various VLMs, and its integration into GPT-4o-mini yields a 7.1% performance improvement.

理解环境和机器人的物理可达性对于任务执行至关重要。尽管最先进的视觉语言模型(VLMs)在环境感知方面表现出色，但由于**缺乏对机器人物理可达性的理解**，它们在具身视觉推理任务中常常生成不准确或不切实际的响应。为了解决这个问题，我们提出了一种**跨多种机器人的物理可达性统一表示**，即**空间物理可达性地图**(S-P Map)，以及**PhysVLM**，一种将这种可达性信息集成到视觉推理中的视觉语言模型。具体来说，S-P Map将机器人的物理可达性抽象为一种通用的空间表示，独立于特定的机器人配置，使模型能够专注于可达性特征而不是机器人特定参数。随后，PhysVLM通过引入一个额外的特征编码器来处理 Map，扩展了传统的VLM架构，使模型能够在推理物理可达性的同时不损害其通用视觉语言能力。为了训练和评估PhysVLM，我们构建了**一个大规模的多机器人数据集Phys100K**，以及一个具有挑战性的基准EQA-phys，该基准包括在模拟和现实环境中为六种不同机器人设计的任务。实验结果表明，PhysVLM优于现有模型，在EQA-phys上比GPT-4o提高了14%，并在RoboVQA-val和OpenEQA基准上超越了RoboMamba和SpatialVLM等先进的具身VLM。此外，S-P Map显示出与各种VLM的强兼容性，将其集成到GPT-4o-mini中带来了7.1%的性能提升。

1. Introduction

1. 引言

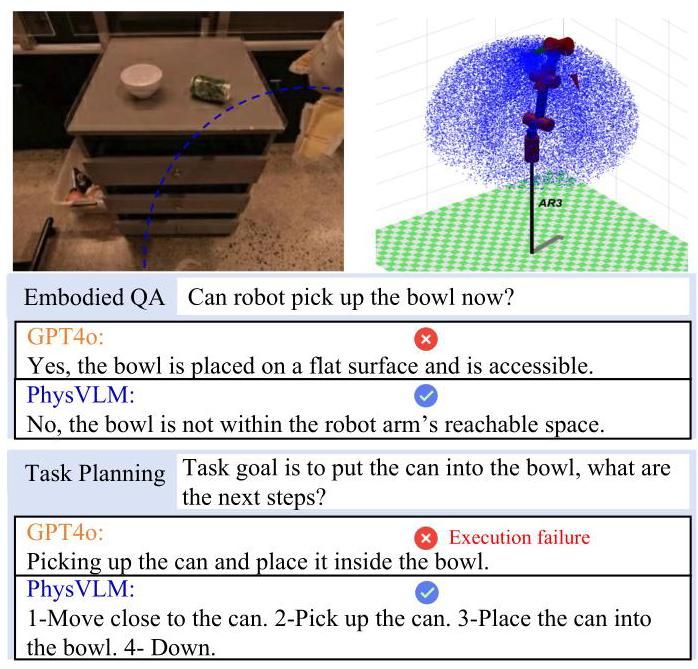


Figure 1. Existing VLM models, such as GPT-4o, may generate inaccurate or impractical responses due to poor comprehension of robotic physical reachability. The proposed PhysVLM integrates vision-language capabilities with an understanding of robotic physical reachability.

图1. 现有的VLM模型，如GPT-4o，由于对机器人物理可达性的理解不足，可能会生成不准确或不切实际的响应。提出的PhysVLM将视觉语言能力与对机器人物理可达性的理解相结合。

Accurate perception of physical reachability is essential for robots to perform tasks effectively. Similar to how humans adjust their actions based on bodily conditions and environmental factors, robots must account for their physical reachability within an environment to ensure efficient and reliable task execution. For instance, in grasping tasks, a robot that fails to assess its reachability may attempt to grasp an object from an unreachable position, leading to task failure or equipment damage [32, 35]. Thus, enhancing a robot’s understanding of physical reachability is crucial for successful task planning and execution in complex environments [31, 41].

准确感知物理可达性对于机器人有效执行任务至关重要。类似于人类根据身体条件和环境因素调整行动，机器人必须考虑其在环境中的物理可达性，以确保高效可靠的任务执行。例如，在抓取任务中，未能评估其可达性的机器人可能会尝试从无法到达的位置抓取物体，导致任务失败或设备损坏[32, 35]。因此，增强机器人对物理可达性的理解对于在复杂环境中成功规划和执行任务至关重要[31, 41]。

Vision-Language Models (VLMs) have shown remarkable progress in environmental understanding [6, 17, 19, 21, 37], and many studies have applied these models to embodied AI to assist robots in perceiving environments and planning tasks [11, 18, 25, 36, 43]. However, while VLMs excel in general environmental perception, they often struggle with tasks that require an understanding of robotic physical reachability (see Figure 1). We identify two key challenges that must be addressed for VLMs to be effective in robotic tasks: (1) how to develop a unified and efficient representation of physical reachability. Robots vary significantly in size, joint types, and other characteristics, making it difficult for VLMs to directly learn these differences; (2) how to enable VLMs to improve their understanding of physical reachability without compromising general vision-language capabilities. Existing VLMs typically combine pre-trained unimodal encoders for vision and language tasks. However, introducing a new modality like physical reachability requires careful architectural and training adjustments to ensure the model can reason about reachability while maintaining its general capabilities.

视觉语言模型(VLMs)在环境理解方面取得了显著进展[6, 17, 19, 21, 37]，许多研究已将这些模型应用于具身AI，以帮助机器人感知环境和规划任务[11, 18, 25, 36, 43]。然而，尽管VLMs在一般环境感知方面表现出色，但在需要理解机器人物理可达性的任务中常常表现不佳(见图1)。我们确定了两个关键挑战，必须解决这些挑战才能使VLMs在机器人任务中有效:(1)如何开发一种**统一且高效的物理可达性表示**。**机器人在尺寸、关节类型等方面差异显著**，使得VLMs难以直接学习这些差异；(2)如何使VLMs在不损害通用视觉语言能力的情况下提高对物理可达性的理解。现有的VLMs通常结合预训练的单模态编码器来处理视觉和语言任务。然而，引入物理可达性等新模态需要仔细的架构和训练调整，以确保模型能够在推理可达性的同时保持其通用能力。

To address these challenges, we propose the Space-Physical Reachability Map (S-P Map), a unified representation that abstracts the physical reachability of diverse robots into a generalized spatial form. The S-P Map is generated by combining robot parameters with egocentric depth images, but crucially, the model learns to focus on the abstracted reachability features (i.e., the gray regions in the S-P Map) rather than the specific robot configurations. This abstraction allows the model to generalize across different robots, as it only needs to reason about which areas are reachable, independent of the robot’s specific characteristics. We introduce PhysVLM, a vision-language model that extends traditional VLM architectures by incorporating an additional feature encoder to process the S-P Map. This design enables PhysVLM to integrate physical reachability information into its reasoning process without compromising its general vision-language capabilities. To train and evaluate PhysVLM, we constructed a large-scale multi-robot dataset, Phys100K, and a challenging benchmark, EQA-phys, which includes tasks for six different robots in both simulated and real-world environments.

为了解决这些挑战，我们提出了空间-物理可达性地图(S-P Map)，这是一种统一的表示方法，将各种机器人的物理可达性抽象为一种广义的空间形式。S-P Map通过将机器人参数与自我中心深度图像结合生成，但关键在于，模型学会关注抽象的可达性特征(即S-P Map中的灰色区域)，而不是具体的机器人配置。这种抽象使得模型能够泛化到不同的机器人，因为它只需要推理哪些区域是可到达的，而不依赖于机器人的具体特性。我们引入了PhysVLM，这是一种视觉-语言模型，它通过引入一个额外的特征编码器来处理S-P Map，从而扩展了传统的VLM架构。这种设计使PhysVLM能够将物理可达性信息整合到其推理过程中，而不影响其一般的视觉-语言能力。为了训练和评估PhysVLM，我们构建了一个大规模的多机器人数据集Phys100K，以及一个具有挑战性的基准EQA-phys，该基准包括在模拟和现实环境中为六种不同机器人设计的任务。

We summarize our contributions as follows:

我们将我们的贡献总结如下:

* We propose a unified and robot-agnostic formulation, the Space-Physical Reachability Map (S-P Map), which abstracts robotic physical reachability in a way that is independent of specific robot configurations, promoting the learning of generalized features.
* 我们提出了一种统一且与机器人无关的公式，即空间-物理可达性地图(S-P Map)，它以独立于特定机器人配置的方式抽象了机器人的物理可达性，促进了广义特征的学习。
* We introduce PhysVLM, a vision-language model that integrates physical reachability with general vision-language capabilities via an additional feature encoder, improving task execution reliability.
* 我们引入了PhysVLM，这是一种视觉-语言模型，它通过额外的特征编码器将物理可达性与一般的视觉-语言能力相结合，提高了任务执行的可靠性。
* We release the EQA-phys benchmark, which includes six robots and question-answer pairs, designed to test the model’s understanding of physical reachability in simulated and real-world environments.
* 我们发布了EQA-phys基准，该基准包括六种机器人和 个问答对，旨在测试模型在模拟和现实环境中对物理可达性的理解。
* Our model achieves a 14% improvement over GPT-4o on the EQA-phys benchmark. In embodied visual reasoning tasks on the RoboVQA-val and OpenEQA benchmarks, it outperforms advanced embodied VLMs such as RoboMamba and SpatialVLM. Furthermore, the S-P Map demonstrates strong compatibility with various VLMs, and integrating it into GPT-4o-mini results in a 7.1% performance improvement.
* 我们的模型在EQA-phys基准上比GPT-4o提高了14%。在RoboVQA-val和OpenEQA基准上的具身视觉推理任务中，它优于RoboMamba和SpatialVLM等先进的具身VLM。此外，S-P Map展示了与各种VLM的强大兼容性，将其集成到GPT-4o-mini中，性能提高了7.1%。

# 2. Related Work

# 2. 相关工作

# 2.1. VLMs in Robotics

# 2.1. 机器人中的VLM

Embodied Question Answering (EQA) tasks require agents to interact with environments to answer questions , 38]. RoboVQA offers a large, diverse dataset for robotic visual question answering. 3D-VLA [44] integrates 3D perception with a generative world model for embodied reasoning, while SpatialVLM [5] enhances VLMs’ spatial understanding using extensive 3D data.

具身问答(EQA)任务要求代理与环境交互以回答问题 ，38]。RoboVQA为机器人视觉问答提供了一个大型、多样化的数据集。3D-VLA [44]将3D感知与生成世界模型相结合，用于具身推理，而SpatialVLM [5]则利用广泛的3D数据增强了VLM的空间理解能力。

Robot task planning involves sequencing subtasks to achieve goals . Code as Policies (CaP) employs OpenAI’s Codex (code-davinci-002) to generate planning code. SayCan [2] combines the Pathways Language Model (PaLM) [7] with robotic affordances to create feasible action plans based on the robot’s capabilities. However, these methods often assume all objects are within the robot’s operational area, ignoring physical reachability and potentially leading to suboptimal or infeasible plans.

机器人任务规划涉及将子任务序列化以实现目标 。代码即策略(CaP) 使用OpenAI的Codex(code-davinci-002)生成规划代码。SayCan [2]将路径语言模型(PaLM)[7]与机器人可供性相结合，根据机器人的能力创建可行的行动计划。然而，这些方法通常假设所有物体都在机器人的操作区域内，忽略了物理可达性，可能导致次优或不可行的计划。

# 2.2. Understanding Physical Reachability

# 2.2. 理解物理可达性

Recent studies use voxel grids with open-vocabulary detection models to assign task-specific attributes, enabling environmental constraint understanding. ReKep [15] generates keypoint proposals and constraints using voxel grids and VLMs, while VoxPoser [14] synthesizes robot trajectories by integrating OWL-ViT [26] and VLMs with voxel-based environment representations. However, these approaches focus on environment modeling without explicitly addressing the robot’s physical reachability.

最近的研究使用体素网格与开放词汇检测模型来分配任务特定属性，从而实现对环境约束的理解。ReKep [15]使用体素网格和VLM生成关键点提案和约束，而VoxPoser [14]通过将OWL-ViT [26]和VLM与基于体素的环境表示相结合，合成了机器人轨迹。然而，这些方法侧重于环境建模，而没有明确解决机器人的物理可达性。

Explicitly representing reachable workspaces remains challenging. Reachability maps [41] model spatial capabilities, and occupancy grids [16] account for obstacles to ensure safe navigation. Additionally, methods like online model predictive control with offline workspace analysis [31] and Reachability Expression-based Motion Planning (REMP) [12] address workspace constraints. Despite these advancements, integrating physical reachability into visual reasoning for complex embodied tasks is still limited, primarily due to the lack of large-scale datasets that include robotic physical parameters in VLM pretraining.

明确表示可达工作空间仍然具有挑战性。可达性地图[41]建模空间能力，而占用网格[16]则考虑障碍物以确保安全导航。此外，像在线模型预测控制与离线工作空间分析[31]和基于可达性表达的运动规划(REMP)[12]等方法解决了工作空间约束。尽管取得了这些进展，将物理可达性整合到复杂具身任务的视觉推理中仍然有限，主要是由于在VLM预训练中缺乏包含机器人物理参数的大规模数据集。

# 3. Method

# 3. 方法

PhysVLM is a large-scale vision-language model designed for visual reasoning that accounts for physical constraints in embodied tasks. As illustrated in Figure 2, PhysVLM integrates instruction text, visual input (RGB image), and an S-P Map abstracts the robotic physical reachability into a unified spatial representation. By combining these inputs, PhysVLM generates responses consistent with both the visual context and the robot’s physical reachability, without being tied to specific robot configurations. The S-P Map is constructed using a unified physical reachability encoding method, which abstracts the physical parameters of various robots alongside their egocentric depth maps into a generalized form. This abstraction allows the model to generalize across different robots, addressing the challenge of learning and reasoning about physical reachability in a robot-agnostic manner.

PhysVLM 是一个为视觉推理设计的大规模视觉语言模型，考虑了具身任务中的物理约束。如图2所示，**PhysVLM 将指令文本、视觉输入(RGB图像)和 S-P 地图(S-P Map)整合在一起**，将机器人物理可达性抽象为统一的空间表示。通过结合这些输入，PhysVLM 生成与视觉上下文和机器人物理可达性一致的响应，而不受特定机器人配置的限制。S-P 地图是使用统一的物理可达性编码方法构建的，该方法将各种机器人的物理参数及其自我中心深度图抽象为通用形式。这种抽象使模型能够推广到不同的机器人，以与机器人无关的方式解决学习和推理物理可达性的挑战。

In this section, we present the core components of PhysVLM. Section 3.1 describes the S-P Map encoding method, Section 3.2 describes the model architecture, and Section 3.3 discusses the training procedures.

在本节中，我们介绍了 PhysVLM 的核心组件。第3.1节描述了 S-P 地图编码方法，第3.2节描述了模型架构，第3.3节讨论了训练过程。

# 3.1. S-P Map Encoding

# 3.1. S-P 地图编码

As illustrated in Figure 2, we model the physical reachability of various robots using a unified approach that abstracts robot-specific parameters into a generalized spatial representation. This abstraction allows the model to focus on the spatial regions that are physically reachable, independent of the specific robot configuration.

如图2所示，我们使用一种统一的方法对各种机器人的物理可达性进行建模，该方法将机器人特定参数抽象为通用的空间表示。这种抽象使模型能够专注于物理上可达的空间区域，而不受特定机器人配置的影响。

where denotes the raw point cloud data from the robot’s RGB-D camera. represents the range of motion for each joint . DH refers to the Denavit-Hartenberg parameters that describe the geometric structure of each joint, and is the extrinsic calibration matrix that transforms coordinates from the camera to the robot’s coordinate system. The function maps these inputs to produce the S-P Map, which abstracts the robot’s physical reachability into a spatial form that is independent of the specific robot configuration.

其中 表示来自机器人 RGB-D 相机的原始点云数据。 表示每个关节 的运动范围。DH 指的是描述每个关节几何结构的 Denavit-Hartenberg 参数， 是将坐标从相机转换到机器人坐标系的外在校准矩阵。函数 将这些输入映射以生成 S-P 地图，该地图将机器人的物理可达性抽象为与特定机器人配置无关的空间形式。

Consider a robot arm with degrees of freedom, where each joint has DH parameters is the joint angle, is the offset along the z-axis, is the link length, and is the twist angle). The homogeneous transformation matrix for each joint is defined as:

考虑一个具有 自由度的机械臂，其中每个关节 具有 DH 参数 是关节角度， 是沿 z 轴的偏移量， 是连杆长度， 是扭转角)。每个关节的齐次变换矩阵定义为:

where is the standard Denavit-Hartenberg transformation function. By multiplying the transformation matrices of all joints, we obtain the transformation matrix from the base frame to the end-effector frame:

其中 是标准的 Denavit-Hartenberg 变换函数。通过将所有关节的变换矩阵相乘，我们得到从基座坐标系到末端执行器坐标系的变换矩阵:

To generate joint configurations, we sample the joint angles from their respective motion ranges , resulting in configurations . By substituting these joint configurations into the forward kinematics equations, we compute the corresponding end-effector positions:

为了生成关节配置，我们从各自的运动范围 中采样关节角度 ，得到配置 。通过将这些关节配置代入正向运动学方程，我们计算相应的末端执行器位置:

where is the origin point in the end-effector frame. We precompute these joint configurations offline, discretize the workspace into a voxel grid , and store it for efficient computation in subsequent steps.

其中 是末端执行器坐标系中的原点。我们离线预计算这些关节配置，将工作空间离散化为体素网格 ，并存储它以在后续步骤中进行高效计算。

Next, as shown in Figure 2, the robot’s raw point cloud is captured from its egocentric RGB-D camera in the camera coordinate system and transformed into the robot’s coordinate system using the extrinsic calibration matrix , resulting in the transformed point cloud :

接下来，如图2所示，机器人的原始点云 从其自我中心 RGB-D 相机在相机坐标系中捕获，并使用外在校准矩阵 转换到机器人坐标系，得到转换后的点云 :

To ensure physical feasibility, we perform a voxel grid lookup to determine whether each point in lies within the precomputed reachable workspace :

为了确保物理可行性，我们执行体素网格查找以确定 中的每个点是否位于预计算的可达工作空间 内:

This step filters the point cloud to include only points within the robot’s reachable workspace, abstracting the robot’s physical reachability into a generalized spatial form.

此步骤过滤点云，仅包含机器人可达工作空间内的点，将机器人的物理可达性抽象为通用的空间形式。

Finally, we transform the valid point cloud back into the camera coordinate system and use the camera’s intrinsic parameters to project these points onto the image plane. We then mark the regions that comply with physical reachability on the original depth map. For areas that are not reachable, we apply a gray mask and outline their boundaries. The resulting S-P Map clearly highlights regions that are beyond the robot’s physical reach, providing a unified and abstracted representation of reachability that is independent of the specific robot configuration. This allows the model to focus on the spatial constraints of the task, without needing to account for the detailed physical parameters of each robot.

最后，我们将有效的点云 转换回相机坐标系，并使用相机的内参将这些点投影到图像平面上。然后，我们在原始深度图上标记符合物理可达性的区域。对于不可达的区域，我们应用灰色遮罩并勾勒其边界。生成的S-P图清晰地突出了机器人物理不可达的区域，提供了一个独立于具体机器人配置的统一且抽象的可达性表示。这使得模型能够专注于任务的空间约束，而无需考虑每个机器人的详细物理参数。

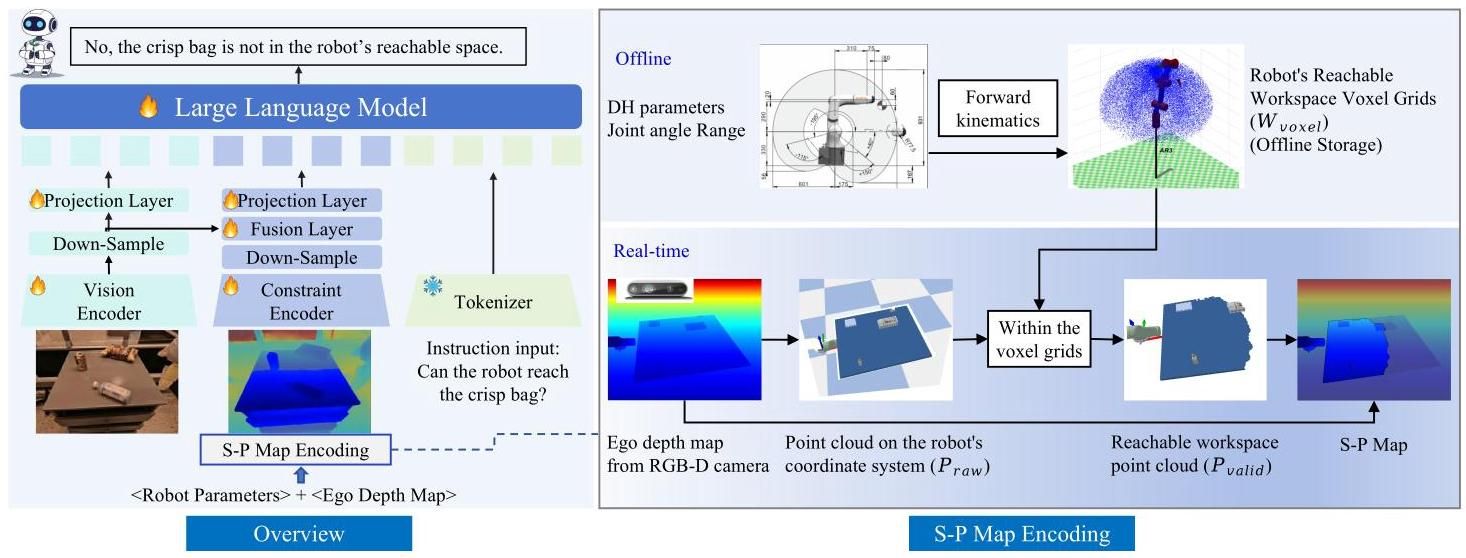


Figure 2. Overview of PhysVLM. Starting with robot parameters and an egocentric depth map, an S-P Map is generated via unified physical reachability encoding. Using the S-P Map, image, and instruction text, PhysVLM generates textual output considering the robot’s physical reachability.

图2. PhysVLM概述。从机器人参数和自我中心深度图开始，通过统一物理可达性编码生成S-P图。使用S-P图、图像和指令文本，PhysVLM生成考虑机器人物理可达性的文本输出。

# 3.2. Model Architecture

# 3.2. 模型架构

To seamlessly integrate robotic physical reachability into PhysVLM while preserving its visual reasoning capabilities, we design a dual-branch architecture: one branch dedicated to vision processing and the other to physical reachability (see Figure 2). These branches operate independently, extracting features from their respective inputs, which are then fused and passed to a unified decoder for final reasoning and response generation.

为了将机器人物理可达性无缝集成到PhysVLM中，同时保留其视觉推理能力，我们设计了一个双分支架构:一个分支专注于视觉处理，另一个分支专注于物理可达性(见图2)。这些分支独立运行，从各自的输入中提取特征，然后融合并传递给统一的解码器进行最终推理和响应生成。

The vision branch leverages a pre-trained Vision Transformer (ViT) [9], specifically the SigLip-400M model [42], to extract high-level visual features from egocentric images. To reduce computational overhead, a Max Pooling layer is applied, followed by a two-layer Multi-Layer Perceptron (MLP) that transforms the visual features into token representations suitable for multimodal fusion.

视觉分支利用预训练的Vision Transformer (ViT) [9]，特别是**SigLip-400M**模型[42]，从自我中心图像中提取高级视觉特征。为了减少计算开销，应用了最大池化层，随后是一个两层的多层感知器(MLP)，将视觉特征转换为适合多模态融合的标记表示。

The physical reachability branch processes the S-P Map, which abstracts the robot’s physical reachability into a generalized spatial form. This branch also utilizes the SigLip- 400M model for feature extraction, followed by Max Pooling and a feature fusion layer. The fusion layer combines the visual and reachability features, and a two-layer MLP further refines these fused features into reachability-specific tokens.

物理可达性分支处理S-P图，该图将机器人的物理可达性抽象为广义的空间形式。该分支也使用SigLip-400M模型进行特征提取，随后是最大池化和特征融合层。融合层结合了视觉和可达性特征，一个两层的MLP进一步将这些融合特征细化为可达性特定的标记。

For the language decoding, we employ the wen-2.5- Instruct-3B model [33, 39] as PhysVLM’s large language model (LLM) decoder, using the Qwen-2.5 tokenizer to process natural language instructions. The decoder integrates multimodal tokens from the vision branch, S-P Maps, and language inputs, generating coherent and contextually relevant textual responses that account for both visual and physical reachability information.

对于语言解码，我们采用 **wen-2.5-Instruct-3B**模型[33, 39]作为PhysVLM的大型语言模型(LLM)解码器，使用Qwen-2.5分词器处理自然语言指令。解码器集成了来自视觉分支、S-P图和语言输入的多模态标记，生成连贯且上下文相关的文本响应，这些响应考虑了视觉和物理可达性信息。

# 3.3. Training

# 3.3. 训练

Training Data Construction. The training data for PhysVLM consists of our Phys100K dataset and general VQA datasets, such as LLaVA-Pretrain, ShareGPT4V, and RoboVQA. Phys100K focuses on question-answering related to physical reachability, aggregating data from RoboVQA (20K samples), ScanNet [8] (10K samples), OpenX-Embodiment [27] (60K samples), and an additional samples from PyBullet.

训练数据构建。PhysVLM的训练数据包括我们的Phys100K数据集和通用VQA数据集，如LLaVA-Pretrain、ShareGPT4V和RoboVQA。Phys100K专注于与物理可达性相关的问题回答，汇总了来自RoboVQA(20K样本)、ScanNet[8](10K样本)、OpenX-Embodiment[27](60K样本)以及来自PyBullet的额外 样本。

Depth maps are essential inputs for generating the S-P Map. For datasets lacking depth maps, we generate them using DepthAnything-v2. Additionally, we employ Grounding DINO [23] and SAM2 [34] to obtain 2D bounding boxes and segmentation results for objects in the images. In PyBullet, we simulate work scenarios using four robotic arms (UR5, FR5, CR5, and FRANKA) to collect RGB images, depth maps, and segmentation results.

深度图是生成S-P图的重要输入。对于缺乏深度图的数据集，我们使用DepthAnything-v2生成它们。此外，我们使用Grounding DINO[23]和SAM2[34]获取图像中物体的2D边界框和分割结果。在PyBullet中，我们使用四个机械臂(UR5、FR5、CR5和FRANKA)模拟工作场景，收集RGB图像、深度图和分割结果。

For the PyBullet data in Phys100K, precise robot configurations can be obtained from the simulator. Therefore, we directly generate the S-P Map using the method described in Section 3.2, and labels indicating whether objects are reachable are obtained through simulated motion. The advantage of the S-P Map is that it abstracts physical reachability into a region-based representation, decoupling the learning process from specific robot configurations. This allows us to generate pseudo-labels for datasets without precise robot parameters. We approximate reachability using the segmentation results, marking regions and the objects within them as "reachable" or "unreachable" based on depth values. Next, we generate question-answer pairs for two main categories:

对于Phys100K中的PyBullet数据，可以从模拟器中获取精确的机器人配置。因此，我们直接使用第3.2节中描述的方法生成S-P图，并通过模拟运动获取物体是否可达的标签。S-P图的优势在于它将物理可达性抽象为基于区域的表示，将学习过程与特定机器人配置解耦。这使得我们能够为没有精确机器人参数的数据集生成伪标签。我们使用分割结果近似可达性，根据深度值将区域及其中的物体标记为“可达”或“不可达”。接下来，我们为两个主要类别生成问答对:

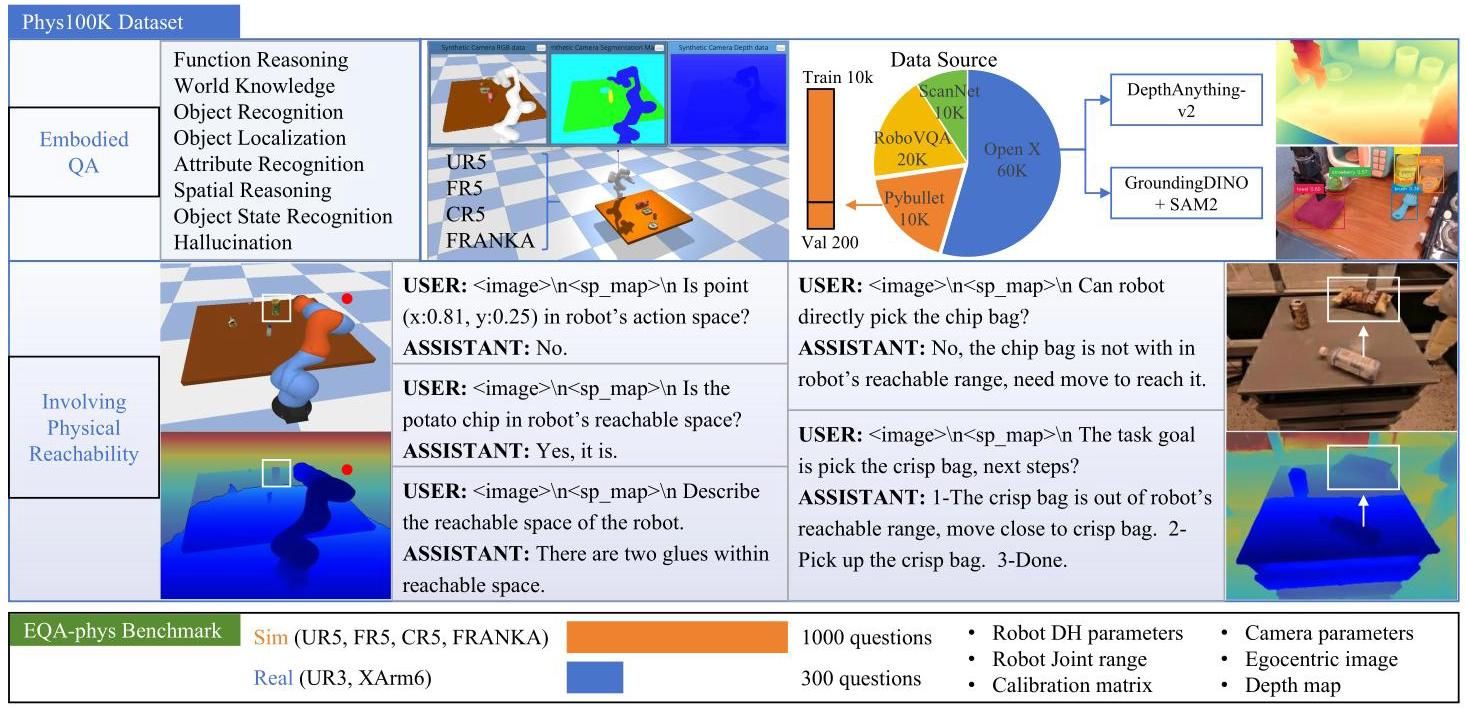


Figure 3. Details of the Phys100K Dataset and EQA-Phys Benchmark.

图3. Phys100K数据集和EQA-Phys基准的详细信息。

* Embodied QA. GPT-4 generates question-answer pairs for ScanNet and RoboVQA, covering categories include Function Reasoning, World Knowledge, Object Recognition, Object Localization, Attribute Recognition, Spatial Reasoning, Object State Recognition, and Hallucination (See Figure 3). Detailed prompts and examples are provided in the appendix to guide the generation process.
* 具身问答。GPT-4为ScanNet和RoboVQA生成问答对，涵盖的类别包括功能推理、世界知识、物体识别、物体定位、属性识别、空间推理、物体状态识别和幻觉(见图3)。附录中提供了详细的提示和示例以指导生成过程。
* Tasks Involving Physical Reachability. We use the "reachable" label with five fixed task templates to generate question-answer pairs, such as "USER:<image> \n<sp\_map> \n Is the [Object] in the robot’s reachable space? ASSISTANT: Yes, it is." Here, [Object] represents the relevant object category, while <image> and <sp\_map> serve as placeholders for image patch tokens and S-P Map patch tokens, respectively. Figure 3 provides examples of question-answer pairs for each object category.
* 涉及物理可达性的任务。我们使用“可达”标签和五个固定任务模板生成问答对，例如“用户:<image> \n<sp\_map> \n [对象] 是否在机器人的可达空间内？助手:是的，它在。” 这里，[对象] 代表相关对象类别，而 <image> 和 <sp\_map> 分别作为图像块标记和 S-P 地图块标记的占位符。图 3 提供了每个对象类别的问答对示例。

Training Pipeline. We adopt a two-stage training process to fully leverage the S-P Map and ensure PhysVLM generalizes across different robots. In the first stage, we align multimodal features using the LLaVA-Pretrain and OpenX-Embodiment datasets from Phys100K. This stage only trains the projection layers, allowing the model to build a foundational understanding of visual inputs and physical reachability, independent of specific robot configurations.

训练流程。我们采用两阶段训练过程，以充分利用 S-P 地图并确保 PhysVLM 在不同机器人之间具有泛化能力。在第一阶段，我们使用 Phys100K 中的 LLaVA-Pretrain 和 OpenX-Embodiment 数据集对齐多模态特征。此阶段仅训练投影层，使模型能够建立对视觉输入和物理可达性的基础理解，独立于特定机器人配置。

In the second stage, we unfreeze all parameters and train the entire model using data from Phys100K, ShareGPT4V, and RoboVQA. This stage enhances PhysVLM’s ability to handle complex visual reasoning tasks with physical reachability constraints, ensuring the model can generalize across diverse environments and robots.

在第二阶段，我们解冻所有参数，并使用 Phys100K、ShareGPT4V 和 RoboVQA 的数据训练整个模型。此阶段增强了 PhysVLM 处理具有物理可达性约束的复杂视觉推理任务的能力，确保模型能够在不同环境和机器人之间泛化。

Implementation Details. PhysVLM is trained for 48 hours using eight A800 GPUs. The training process consists of two stages, each lasting one epoch. The batch size and learning rate are set at 128 and 1e-3 in the first stage, and 64 and 1e-5 in the second stage. The final model is PhysVLM-3B.

实现细节。PhysVLM 使用八块 A800 GPU 训练了 48 小时。训练过程分为两个阶段，每个阶段持续一个 epoch。第一阶段批量大小和学习率分别设置为 128 和 1e-3，第二阶段为 64 和 1e-5。最终模型为 PhysVLM-3B。

# 3.4. EQA-phys Benchmark

# 3.4. EQA-phys 基准

As illustrated in Figure 3, we introduce an embodied QA task focused on physical reachability, termed EQA-phys, which emphasizes QA tasks constrained by physical limitations. This benchmark includes a simulator dataset with 200 samples and 1,000 questions from the PyBullet validation set, as well as a zero-shot evaluation set based on real-world data from UR3 and XArm6 robots in two scenarios. The evaluation set contains 60 samples and 300 questions, all manually annotated by domain experts.

如图 3 所示，我们引入了一个专注于物理可达性的具身问答任务，称为 EQA-phys，该任务强调受物理限制的问答任务。该基准包括一个模拟器数据集，包含来自 PyBullet 验证集的 200 个样本和 1,000 个问题，以及一个基于 UR3 和 XArm6 机器人在两种场景下的真实数据的零样本评估集。评估集包含 60 个样本和 300 个问题，所有数据均由领域专家手动标注。

Table 1. Results of EQA-phys. Comparison of PhysVLM-3B (ours) with API-based VLMs and embodied VLMs.

表 1. EQA-phys 的结果。PhysVLM-3B(我们的)与基于 API 的 VLM 和具身 VLM 的比较。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | REAL-WORLD | | SIMULATOR | | | |  |
|  | | UR3 | XARM6 | UR5 | FR5 | CR5 | FRANKA | ALL |
| API-BASED VLMS | GPT-4O-MINI | 54.3 | 56.0 | 49.4 | 55.4 | 54.6 | 47.1 | 52.8 |
| CLAUDE-3.5 | 56.2 | 60.5 | 54.0 | 58.1 | 55.7 | 54.3 | 56.4 |
| GPT-40 | 56.7 | 61.5 | 55.7 | 58.3 | 57.5 | 52.6 | 57.0 |
| GPT-4o-MINI + S-P MAP | 60.0 | 60.5 |  |  |  |  |  |
| CLAUDE-3.5 + S-P MAP |  |  |  |  |  |  | 60.3 |
|  |  |  |  | 60.7↑1.4 |  |  |  |
| EMBODIED VLMS | SPATIALVLM | 56.3 | 55.1 | 54.6 | 59.1 | 52.0 | 47.5 | 54.1 |
| SPATIALBOT | 51.1 | 50.2 | 50.0 | 48.1 | 53.3 | 54.4 | 51.1 |
| PhysVLM-3B | 64.1 | 63.0 | 71.4 | 75.7 | 74.0 | 78.1 | 71.0 |
|  | | 现实世界 | | 模拟器 | | | |  |
|  | | UR3 | XARM6 | UR5 | FR5 | CR5 | FRANKA | 全部 |
| 基于API的视觉语言模型系统 | GPT-4O-MINI | 54.3 | 56.0 | 49.4 | 55.4 | 54.6 | 47.1 | 52.8 |
| CLAUDE-3.5 | 56.2 | 60.5 | 54.0 | 58.1 | 55.7 | 54.3 | 56.4 |
| GPT-40 | 56.7 | 61.5 | 55.7 | 58.3 | 57.5 | 52.6 | 57.0 |
| GPT-4o-MINI + S-P地图 | 60.0 | 60.5 |  |  |  |  |  |
| CLAUDE-3.5 + S-P地图 |  |  |  |  |  |  | 60.3 |
|  |  |  |  | 60.7↑1.4 |  |  |  |
| 具身视觉语言模型系统 | 空间视觉语言模型 | 56.3 | 55.1 | 54.6 | 59.1 | 52.0 | 47.5 | 54.1 |
| 空间机器人 | 51.1 | 50.2 | 50.0 | 48.1 | 53.3 | 54.4 | 51.1 |
| PhysVLM-3B | 64.1 | 63.0 | 71.4 | 75.7 | 74.0 | 78.1 | 71.0 |

# 4. Experiments

# 4. 实验

# 4.1. Experimental Setting

# 4.1. 实验设置

Tasks. We compare the performance of PhysVLM with other methods across three categories of tasks:

任务。我们在三类任务中比较PhysVLM与其他方法的性能:

* EQA-phys. This benchmark tests the model’s ability to integrate visual reasoning with robotic physical reachability. The real-robot component is used to assess PhysVLM’s zero-shot generalization, highlighting its ability to handle unseen robots and environments.
* EQA-phys。该基准测试模型在视觉推理与机器人物理可达性结合方面的能力。真实机器人组件用于评估PhysVLM的零样本泛化能力，突出其处理未见过的机器人和环境的能力。
* Embodied QA. We evaluate the model’s general visual reasoning ability in embodied tasks using the OpenEQA [25] and RoboVQA-val [29] benchmarks.
* 具身问答。我们使用OpenEQA [25]和RoboVQA-val [29]基准测试模型在具身任务中的一般视觉推理能力。
* Robot Task Planning. For real-world tasks such as "Pick A into B", we assess the model’s ability to understand robotic physical reachability and generate reasonable task plans. As this study does not focus on robot control strategies, we use the natural language planning approach from [29].
* 机器人任务规划。对于“将A放入B”等现实世界任务，我们评估模型理解机器人物理可达性并生成合理任务计划的能力。由于本研究不关注机器人控制策略，我们使用[29]中的自然语言规划方法。

Baselines. We compare our model to several baselines, including API-accessible VLMs such as Claude 3.5 [28], GPT-4o-mini [1], and GPT-4o [1], as well as embodied VLMs like SpatialVLM [5], SpatialBot [4], 3D-VLA [44], and RoboMamba [22]. SpatialVLM and SpatialBot both use the version, which has a similar parameter count to our model. Since the executable versions of 3D-VLA and RoboMamba are unavailable, we compare their reported results on RoboVQA-val.

基线。我们将我们的模型与多个基线进行比较，包括API可访问的VLMs，如Claude 3.5 [28]、GPT-4o-mini [1]和GPT-4o [1]，以及具身VLMs，如SpatialVLM [5]、SpatialBot [4]、3D-VLA [44]和RoboMamba [22]。SpatialVLM和SpatialBot均使用 版本，其参数数量与我们的模型相似。由于3D-VLA和RoboMamba的可执行版本不可用，我们比较了它们在RoboVQA-val上的报告结果。

Evaluation Metrics. For tasks involving physical reachability, we use LLM scoring, following the approach used in existing studies [24, 25]. Assigning 5 points for completely correct responses and 1 point for incorrect responses, we calculate the average score and express it as a percentage. For Embodied QA, we follow the benchmark settings of the corresponding datasets [25, 29]. For task planning, each task type is executed 10 times, and the average success rate serves as the evaluation metric.

评估指标。对于涉及物理可达性的任务，我们使用LLM评分，遵循现有研究[24, 25]的方法。完全正确的回答得5分，错误的回答得1分，我们计算平均分并以百分比表示。对于具身问答，我们遵循相应数据集[25, 29]的基准设置。对于任务规划，每种任务类型执行10次，平均成功率作为评估指标。

# 4.2. Results on EQA-phys

# 4.2. EQA-phys结果

Table 1 shows the results on EQA-phys. Neither API-based nor embodied VLMs can handle the robot’s parameter constraints, resulting in suboptimal outputs with scores around 55%. In contrast, our model successfully completes visual reasoning tasks involving physical reachability, obtaining an average score of . As discussed, these tasks require the model to perform visual reasoning based on an understanding of the robotic physical reachability. The model can only effectively accomplish these tasks by truly understanding robotic physical reachability.

表1显示了EQA-phys的结果。无论是基于API的VLMs还是具身VLMs都无法处理机器人的参数约束，导致输出不理想，得分约为55%。相比之下，我们的模型成功完成了涉及物理可达性的视觉推理任务，获得了 的平均分。正如所讨论的，这些任务要求模型基于对机器人物理可达性的理解进行视觉推理。模型只有真正理解机器人物理可达性，才能有效完成这些任务。

The results in Table 1 demonstrate that prompting API-based VLMs, such as GPT-4o, with the S-P Map (detailed in 4) significantly enhances their performance. This improvement stems from the S-P Map’s ability to abstract physical reachability into a robot-agnostic representation, enabling VLMs to reason about physical constraints that would otherwise be beyond their capabilities. By decoupling reachability from specific robot parameters, the S-P Map facilitates generalization across diverse environments, allowing models to better comprehend physical reachability, even in previously unseen scenarios.

表1中的结果表明，使用S-P Map(详见4)提示基于API的VLMs，如GPT-4o，显著提高了它们的性能。这一改进源于S-P Map将物理可达性抽象为与机器人无关的表示，使VLMs能够推理原本超出其能力的物理约束。通过将可达性与特定机器人参数解耦，S-P Map促进了跨环境的泛化，使模型能够更好地理解物理可达性，即使在以前未见过的场景中也是如此。

Table 1 also demonstrates that PhysVLM-3B achieved scores of over 63% in zero-shot evaluations for UR3 and XArm6 robots, despite operating in new environments with different robot parameters. This performance is attributed to two key factors: (1) the S-P Map abstracts various robot parameters into a unified, transferable representation of physical reachability, and (2) the model’s independent visual and constraint encoding branches allow it to learn generalizable visual features from diverse image-text data, enabling effective reasoning in novel environments.

表1还表明，PhysVLM-3B在UR3和XArm6机器人的零样本评估中得分超过63%，尽管在新环境中操作且机器人参数不同。这一表现归因于两个关键因素:(1) S-P Map将各种机器人参数抽象为统一的、可转移的物理可达性表示，(2) 模型的独立视觉和约束编码分支使其能够从多样化的图像-文本数据中学习可泛化的视觉特征，从而在新环境中进行有效推理。

Table 2. Embodied QA results on the RoboVQA-val set, comparison of PhysVLM (ours) with existing methods. An asterisk (\*) indicates models not pre-trained on the RoboVQA dataset.

表2. RoboVQA-val集上的具身问答结果，PhysVLM(我们的)与现有方法的比较。星号(\*)表示未在RoboVQA数据集上预训练的模型。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | BLEU1 | BLEU2 | BLEU3 | BLEU4 |
| SPATIALVLM\* | 5.1 | 3.0 | 1.9 | 1.2 |
| SPATIALBOT\* | 12.4 | 9.3 | 8.0 | 7.2 |
| 3D-VLA | 48.3 | 38.5 | 31.7 | 26.8 |
| ROBOMAMBA | 54.9 | 44.2 | 39.5 | 36.3 |
| PhysVLM-3B | 65.3 | 62.4 | 50.9 | 43.5 |
|  | BLEU1 | BLEU2 | BLEU3 | BLEU4 |
| 空间视觉语言模型\* | 5.1 | 3.0 | 1.9 | 1.2 |
| 空间机器人\* | 12.4 | 9.3 | 8.0 | 7.2 |
| 三维视觉语言对齐 | 48.3 | 38.5 | 31.7 | 26.8 |
| 机器人曼巴 | 54.9 | 44.2 | 39.5 | 36.3 |
| 物理视觉语言模型-3B | 65.3 | 62.4 | 50.9 | 43.5 |

Table 3. Embodied QA results on the OpenEQA benchmark, comparison of PhysVLM (ours) with existing methods. An asterisk (\*) indicates that only the first 200 samples were tested due to API limitations.

表3. OpenEQA基准测试中的具身问答结果，PhysVLM(我们的方法)与现有方法的比较。星号(\*)表示由于API限制，仅测试了前200个样本。

|  |  |  |  |
| --- | --- | --- | --- |
|  | EM-EQA (SCANNET) | EM-EQA (HM3D) | ALL |
| SPATIALVLM | 42.9 | 44.3 | 43.8 |
| SPATIALBOT | 45.3 | 51.0 | 49.1 |
| GPT4V | 57.4 | 51.3 | 55.3 |
| GPT-4o\* | 68.2 | 65.2 | 66.7 |
| PhysVLM-3B | 60.7 | 51.2 | 57.4 |
|  | EM-EQA(SCANNET) | EM-EQA(HM3D) | 全部 |
| SPATIALVLM | 42.9 | 44.3 | 43.8 |
| SPATIALBOT | 45.3 | 51.0 | 49.1 |
| GPT4V | 57.4 | 51.3 | 55.3 |
| GPT-4o\* | 68.2 | 65.2 | 66.7 |
| PhysVLM-3B | 60.7 | 51.2 | 57.4 |

# 4.3. Results on Embodied QA

# 4.3. 具身问答结果

We demonstrate our model’s effectiveness in handling general embodied visual reasoning tasks. Additionally, we show that incorporating an understanding of physical constraints does not diminish its general visual reasoning capabilities. We compare our model with state-of-the-art embodied VLMs (see Tables 2 and 3). Our model achieves the best performance on the RoboVQA-val benchmark, surpassing other models by in BLEU-4. On the OpenEQA benchmark, our model outperforms existing embodied VLMs and GPT-4V, ranking second, behind only GPT-40.

我们展示了模型在处理通用具身视觉推理任务中的有效性。此外，我们表明，融入对物理约束的理解并不会削弱其通用视觉推理能力。我们将模型与最先进的具身视觉语言模型(VLMs)进行了比较(见表2和表3)。我们的模型在RoboVQA-val基准测试中取得了最佳性能，在BLEU-4指标上超越了其他模型 。在OpenEQA基准测试中，我们的模型表现优于现有的具身VLMs和GPT-4V，排名第二，仅次于GPT-40。

# 4.4. Results on Robot Task Planning

# 4.4. 机器人任务规划结果

Table 4 presents the performance of PhysVLM and baseline models on real-world task planning scenarios. When all objects are within the robot’s physical reach, PhysVLM performs similarly to other models, as directly grabbing or placing the objects succeeds. However, when some objects are outside the physical reach, the model must suggest that the robot move closer before grabbing or placing them. In these cases, our model performs exceptionally well, whereas the task success rates of other models decline significantly. This is attributed to our model’s understanding of robotic physical reachability and its ability to incorporate this understanding in task planning.

表4展示了PhysVLM和基线模型在现实世界任务规划场景中的表现。当所有物体都在机器人的物理可及范围内时，PhysVLM与其他模型表现相似，因为直接抓取或放置物体是成功的。然而，当某些物体超出物理可及范围时，模型必须建议机器人在抓取或放置之前先靠近。在这些情况下，我们的模型表现非常出色，而其他模型的任务成功率显著下降。这归因于我们的模型对机器人物理可及性的理解及其在任务规划中融入这种理解的能力。

Table 4. Task planning results. Comparison of PhysVLM (ours) with other VLMs.

表4. 任务规划结果。PhysVLM(我们的模型)与其他VLMs的比较。

|  |  |  |
| --- | --- | --- |
|  | ALL OBJECTS IN RANGE | PART OBJECTS IN RANGE |
| GPT-4O-MINI | 70.5 | 23.2 |
| CLAUDE-3.5 | 73.6 | 32.1 |
| GPT-40 | 75.9 | 35.8 |
| SPATIALVLM | 64.4 | 21.5 |
| SPATIALBOT | 65.6 | 25.3 |
| PhysVLM-3B | 69.2 | 48.4 |
|  | 范围内所有对象 | 范围内部分对象 |
| GPT-4O-MINI | 70.5 | 23.2 |
| CLAUDE-3.5 | 73.6 | 32.1 |
| GPT-40 | 75.9 | 35.8 |
| SPATIALVLM | 64.4 | 21.5 |
| SPATIALBOT | 65.6 | 25.3 |
| PhysVLM-3B | 69.2 | 48.4 |

Table 5. Ablation study on the S-P Map. We compare the constraint encoder’s performance when using the S-P Map, replacing it with a Depth Map, or providing no input.

表5. S-P图(S-P Map)的消融研究。我们比较了在使用S-P图、用深度图(Depth Map)替换它或不提供任何输入时，约束编码器的性能。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | S-P MAP | DEPTH MAP | EQA-PHYS REAL | EQA-PHYS SIM |
| 1 | ✓ |  | 63.5 | 74.8 |
| 2 |  | ✓ | 58.1 | 62.4 |
| 3 |  |  | 54.2 | 58.8 |
| ID | S-P地图 | 深度图 | EQA-PHYS真实 | EQA-PHYS模拟 |
| 1 | ✓ |  | 63.5 | 74.8 |
| 2 |  | ✓ | 58.1 | 62.4 |
| 3 |  |  | 54.2 | 58.8 |

# 4.5. Ablation Study

# 4.5. 消融研究

In this section, we conduct ablation studies to evaluate the contribution of each component in PhysVLM. We report the average LLM score on tasks involving physical reachability.

在本节中，我们进行消融研究以评估PhysVLM中每个组件的贡献。我们报告了涉及物理可达性任务的平均LLM分数。

The effectiveness of S-P Map. To demonstrate the S-P Map’s contribution, we compare results obtained with and without its input. As shown in Table 5, Experiments 1 and 3 demonstrate that omitting the S-P Map leads to a significant performance decrease for both zero-shot real-world robots and simulators. Specifically, the overall average score drops by in simulation results and 9.3% in real-world robot evaluations. Without the S-P Map input, the model struggles to handle the robotic physical reachability.

S-P图的有效性。为了展示S-P图的贡献，我们比较了有和没有其输入的结果。如表5所示，实验1和3表明，省略S-P图会导致零样本现实世界机器人和模拟器的性能显著下降。具体来说，模拟结果中的总体平均分数下降了 ，现实世界机器人评估中下降了9.3%。没有S-P图输入，模型难以处理机器人物理可达性。

Additionally, Experiments 1 and 2 demonstrate that replacing the S-P Map with a Depth Map significantly degrades the model’s performance on zero-shot tasks. Since the Depth Map does not accurately represent robotic physical reachability, the model cannot rely solely on depth information to understand it.

此外，实验1和2表明，用深度图替换S-P图会显著降低模型在零样本任务上的性能。由于深度图不能准确表示机器人物理可达性，模型不能仅依赖深度信息来理解它。

Effectiveness of an additional feature encoder. To demonstrate the effectiveness of the model architecture, we compare the performance of a feature encoder that shares weights with the visual feature encoder for the S-P Map. In these experiments, we evaluate the average scores on both the EQA-phys and OpenEQA benchmarks. The results in

额外特征编码器的有效性。为了展示模型架构的有效性，我们比较了与S-P图视觉特征编码器共享权重的特征编码器的性能。在这些实验中，我们评估了EQA-phys和OpenEQA基准上的平均分数。结果在

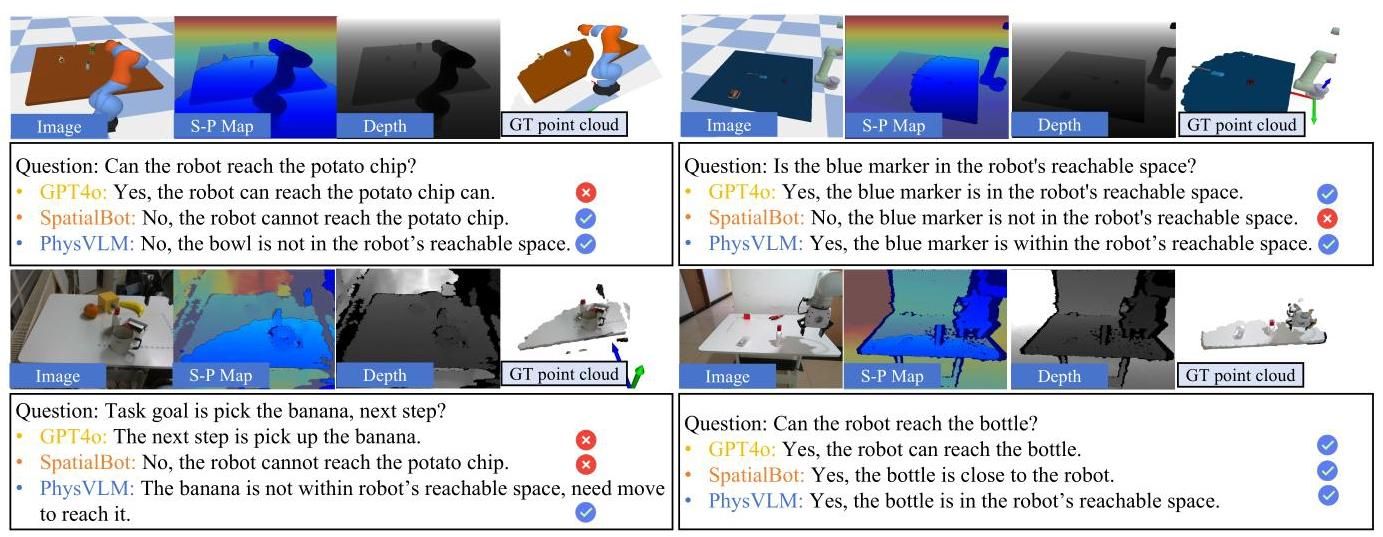


Figure 4. Visual comparison of PhysVLM (ours), GPT-4o, and SpatialBot.

图4. PhysVLM(我们的)、GPT-4o和SpatialBot的视觉比较。

Table 6. Ablation study on the effectiveness of an additional feature encoder. Share indicates shared network and weights with the visual feature encoder.

表6. 额外特征编码器有效性的消融研究。Share表示与视觉特征编码器共享网络和权重。

|  |  |  |
| --- | --- | --- |
|  | EQA-PHYS | OpenEQA |
| INDEPENDENT | 71.0 | 57.4 |
| Share | 68.2 | 56.5 |
|  | EQA-PHYS | OpenEQA |
| 独立 | 71.0 | 57.4 |
| 分享 | 68.2 | 56.5 |

Table 6 show that sharing the feature encoder not only decreases performance on EQA-phys but also impairs general visual reasoning capabilities. This is because S-P Map features differ from images, and the training data contains significantly more image-text pairs than S-P Map data.

表6显示，共享特征编码器不仅降低了EQA-phys的性能，还损害了通用的视觉推理能力。这是因为S-P Map特征与图像不同，且训练数据中包含的图像-文本对远多于S-P Map数据。

Effectiveness of training data. To evaluate the effectiveness of Phys100K, we conduct experiments by selectively removing data from various sources in Phys100K. As shown in Table 7, removing data from PyBullet or other embodied datasets leads to a reduction in overall performance, highlighting the critical role of each data component in the model’s performance.

训练数据的有效性。为了评估Phys100K的有效性，我们通过有选择地移除Phys100K中不同来源的数据进行实验。如表7所示，移除来自PyBullet或其他具身数据集的数据会导致整体性能下降，这凸显了每个数据组件在模型性能中的关键作用。

Table 7. Ablation study on the effectiveness of training data.

表7. 训练数据有效性的消融研究。

|  |  |  |
| --- | --- | --- |
| PART OF PHYS100K | EQA-PHYS REAL | EQA-PHYS SIM |
| ALL | 63.5 | 74.8 |
| W/O PYBULLET | 62.1 | 65.4 |
| W/O OTHER DATASETS | 58.6 | 71.5 |
| 物理100K课程的一部分 | 物理真实环境评估 | 物理模拟环境评估 |
| 全部 | 63.5 | 74.8 |
| 不包括PyBullet | 62.1 | 65.4 |
| 不包括其他数据集 | 58.6 | 71.5 |

# 4.6. Qualitative Results

# 4.6. 定性结果

Figure 4 compares our method with SpatialBot and GPT-4o. SpatialBot uses depth maps and images, while GPT-4o uses standard images. Both struggle with tasks requiring physical reachability, causing visual reasoning errors. In contrast, our approach delivers accurate results. Additionally, incorporating the S-P Map into GPT-4o improves its handling of physical reachability and response accuracy.

图4将我们的方法与SpatialBot和GPT-4o进行了比较。SpatialBot使用深度图和图像，而GPT-4o使用标准图像。两者在处理需要物理可达性的任务时都遇到了困难，导致视觉推理错误。相比之下，我们的方法提供了准确的结果。此外，将S-P Map整合到GPT-4o中，提高了其处理物理可达性和响应准确性的能力。

# 5. Conclusion

# 5. 结论

We introduce PhysVLM, a VLM that incorporates physical reachability into visual reasoning for robotic tasks. The S-P Map provides a unified representation of robotic reachability, facilitating the learning of generalizable features. PhysVLM extends traditional VLMs by adding a physical reachability encoder, enabling the simultaneous processing of visual, reachability, and textual information. Additionally, we present EQA-phys, a benchmark for evaluating embodied QA tasks involving physical reachability. Our experiments show that PhysVLM outperforms existing models, achieving a 14% higher score than GPT-4o on EQA-phys. A limitation is its reduced zero-shot performance on real robots compared to simulations, likely due to the domain gap. Future work will focus on expanding datasets, enhancing real-world performance, and improving the understanding of physical accessibility in vision-language-action models. PhysVLM’s reachability awareness supports safer and more reliable robotic decision-making in industrial and assistive settings, while its unified representation ensures cross-platform adaptability for real-world deployment, bridging crucial gaps between environmental perception and actionable robotic intelligence.

我们介绍了PhysVLM，这是一种将物理可达性融入机器人任务视觉推理的视觉语言模型(VLM)。S-P Map提供了机器人可达性的统一表示，促进了可泛化特征的学习。PhysVLM通过添加物理可达性编码器扩展了传统的VLM，使其能够同时处理视觉、可达性和文本信息。此外，我们提出了EQA-phys，这是一个用于评估涉及物理可达性的具身问答任务的基准。我们的实验表明，PhysVLM优于现有模型，在EQA-phys上的得分比GPT-4o高出14%。一个局限性是其在真实机器人上的零样本性能相比模拟环境有所下降，这可能是由于领域差距所致。未来的工作将集中在扩展数据集、增强现实世界性能以及提高对视觉-语言-动作模型中物理可达性的理解上。PhysVLM的可达性感知支持在工业和辅助环境中更安全、更可靠的机器人决策，而其统一表示确保了跨平台的适应性，为现实世界部署架起了环境感知与可操作机器人智能之间的关键桥梁。

# 6. Acknowledgments

# 6. 致谢

This work was supported by National Key R&D Program of China under Grant No.2022ZD0160601, 2022YFB4300400, and 2018B030322016, National Natural Science Foundation of China under Grants 62176254, 62276260, and U1701266.

本工作得到了国家重点研发计划项目(编号:2022ZD0160601、2022YFB4300400、2018B030322016)和国家自然科学基金(编号:62176254、62276260、U1701266)的支持。

# References

# 参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023. 6

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, 等. Gpt-4技术报告. arXiv预印本 arXiv:2303.08774, 2023. 6

[2] Michael Ahn, Brohan, et al. Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691, 2022. 2

Michael Ahn, Brohan, 等. 按我能做的做，而不是按我说的做:将语言扎根于机器人能力中. arXiv预印本 arXiv:2204.01691, 2022. 2

[3] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Cheb-otar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691, 2022. 2

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Cheb-otar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, 等. 按我能做的做，而不是按我说的做:将语言扎根于机器人能力中. arXiv预印本 arXiv:2204.01691, 2022. 2

[4] Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. arXiv preprint arXiv:2406.13642, 2024. 6

Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, 和 Bo Zhao. Spatialbot:通过视觉语言模型实现精确的空间理解. arXiv预印本 arXiv:2406.13642, 2024. 6

[5] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14455-14465, 2024.2,6

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, 和 Fei Xia. Spatialvlm:赋予视觉语言模型空间推理能力. 在IEEE/CVF计算机视觉与模式识别会议论文集, 页码 14455-14465, 2024.2,6

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24185-24198, 2024. 2

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, 等. Internvl:扩展视觉基础模型并针对通用视觉-语言任务进行对齐. 在IEEE/CVF计算机视觉与模式识别会议论文集, 页码 24185-24198, 2024. 2

[7] Aakanksha Chowdhery, Narang, et al. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240):1-113, 2023. 2

Aakanksha Chowdhery, Narang, 等. Palm:通过路径扩展语言建模. 机器学习研究杂志, 24(240):1-113, 2023. 2

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Hal-ber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated reconstructions of indoor scenes. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5828-5839, 2017. 4

Angela Dai, Angel X Chang, Manolis Savva, Maciej Hal-ber, Thomas Funkhouser, 和 Matthias Nießner. Scannet:室内场景的丰富注释 重建. 在IEEE计算机视觉与模式识别会议论文集, 页码 5828-5839, 2017. 4

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 4

Alexey Dosovitskiy. 一张图片相当于16x16个词:大规模图像识别中的Transformer。arXiv预印本 arXiv:2010.11929, 2020. 4

[10] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378, 2023. 2

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, 等. Palm-e:一个具身多模态语言模型。arXiv预印本 arXiv:2303.03378, 2023. 2

[11] Kiana Ehsani, Tanmay Gupta, Rose Hendrix, Jordi Salvador, Luca Weihs, Kuo-Hao Zeng, Kunal Pratap Singh, Yejin Kim, Winson Han, Alvaro Herrasti, et al. Spoc: Imitating shortest

Kiana Ehsani, Tanmay Gupta, Rose Hendrix, Jordi Salvador, Luca Weihs, Kuo-Hao Zeng, Kunal Pratap Singh, Yejin Kim, Winson Han, Alvaro Herrasti, 等. Spoc:模拟中最短路径的模仿

paths in simulation enables effective navigation and manipu-

在现实世界中实现有效的导航和操作。

lation in the real world. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16238-16250, 2024. 2

在IEEE/CVF计算机视觉与模式识别会议论文集，第16238-16250页，2024. 2

[12] Xiaofeng Gao, Luyao Yuan, Tianmin Shu, Hongjing Lu, and Song-Chun Zhu. Show me what you can do: Capability calibration on reachable workspace for human-robot collaboration. IEEE Robotics and Automation Letters, 7(2):2644- 2651, 2022. 3

Xiaofeng Gao, Luyao Yuan, Tianmin Shu, Hongjing Lu, 和 Song-Chun Zhu. 展示你能做什么:可达工作空间上的人机协作能力校准。IEEE机器人与自动化通讯，7(2):2644-2651, 2022. 3

[13] Yining Hong, Zishuo Zheng, Peihao Chen, Yian Wang, Jun-yan Li, and Chuang Gan. Multiply: A multisensory object-centric embodied large language model in world. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26406-26416, 2024. 2

Yining Hong, Zishuo Zheng, Peihao Chen, Yian Wang, Jun-yan Li, 和 Chuang Gan. Multiply:一个多感官对象中心的具身大语言模型在 世界中。在IEEE/CVF计算机视觉与模式识别会议论文集，第26406-26416页，2024. 2

[14] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. arXiv preprint arXiv:2307.05973, 2023. 2

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, 和 Li Fei-Fei. Voxposer:用于语言模型机器人操作的可组合3D价值图。arXiv预印本 arXiv:2307.05973, 2023. 2

[15] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. arXiv preprint arXiv:2409.01652, 2024. 2

Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, 和 Li Fei-Fei. Rekep:关系关键点约束的时空推理用于机器人操作。arXiv预印本 arXiv:2409.01652, 2024. 2

[16] Lorenzo Jamone, Martim Brandao, Lorenzo Natale, Kenji Hashimoto, Giulio Sandini, and Atsuo Takanishi. Autonomous online generation of a motor representation of the workspace for intelligent whole-body reaching. Robotics and Autonomous Systems, 62(4):556-567, 2014. 2

Lorenzo Jamone, Martim Brandao, Lorenzo Natale, Kenji Hashimoto, Giulio Sandini, 和 Atsuo Takanishi. 自主在线生成工作空间的运动表示用于智能全身到达。机器人与自主系统，62(4):556-567, 2014. 2

[17] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9579-9589, 2024. 2

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, 和 Jiaya Jia. Lisa:通过大语言模型进行推理分割。在IEEE/CVF计算机视觉与模式识别会议论文集，第9579-9589页，2024. 2

[18] Xiaohan Lei, Min Wang, Wengang Zhou, Li Li, and Houqiang Li. Instance-aware exploration-verification-exploitation for instance imagegoal navigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16329-16339, 2024. 2

Xiaohan Lei, Min Wang, Wengang Zhou, Li Li, 和 Houqiang Li. 实例感知的探索-验证-利用用于实例图像目标导航。在IEEE/CVF计算机视觉与模式识别会议论文集，第16329-16339页，2024. 2

[19] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and in large multimodal models. arXiv preprint arXiv:2407.07895, 2024. 2

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, 和 Chunyuan Li. Llava-next-interleave:处理大模态模型中的多图像、视频和 。arXiv预印本 arXiv:2407.07895, 2024. 2

[20] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 9493-9500. IEEE, 2023. 2

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, 和 Andy Zeng. 代码即策略:用于具身控制的语言模型程序。在2023 IEEE国际机器人与自动化会议(ICRA)，第9493-9500页。IEEE, 2023. 2

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024. 2

Haotian Liu, Chunyuan Li, Qingyang Wu, 和 Yong Jae Lee. 视觉指令调优。神经信息处理系统进展，36, 2024. 2

[22] Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Lily Lee, Kaichen Zhou, Pengju An, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Multimodal state space model for efficient robot reasoning and manipulation. arXiv preprint arXiv:2406.04339, 2024. 6

刘嘉明, 刘梦珍, 王振宇, 李莉莉, 周凯晨, 安鹏举, 杨森桥, 张任锐, 郭彦东, 张尚航. Robomamba: 用于高效机器人推理与操作的多模态状态空间模型. arXiv预印本 arXiv:2406.04339, 2024. 6

[23] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with

刘世龙, 曾朝阳, 任天和, 李峰, 张浩, 杨杰, 蒋庆, 李春元, 杨建伟, 苏航, 等. Grounding dino: 将DINO与

grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499, 2023. 4

基于接地的预训练结合用于开放集目标检测. arXiv预印本 arXiv:2303.05499, 2023. 4

[24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In European Conference on Computer Vision, pages 216-233. Springer, 2025. 6

刘源, 段浩东, 张元汉, 李博, 张松阳, 赵王博, 袁一科, 王佳琪, 贺聪慧, 刘子维, 等. Mmbench: 你的多模态模型是全能的吗? 在欧洲计算机视觉会议上, 第216-233页. Springer, 2025. 6

[25] Arjun Majumdar, Ajay, et al. Openeqa: Embodied question answering in the era of foundation models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16488-16498, 2024. 2, 6

阿琼·马宗达, 阿贾伊, 等. Openeqa: 基础模型时代的具身问答. 在IEEE/CVF计算机视觉与模式识别会议论文集上, 第16488-16498页, 2024. 2, 6

[26] Matthias Minderer, Alexey Gritsenko, Austin Stone, et al. Simple open-vocabulary object detection with vision transformers, 2022. 2

马蒂亚斯·明德勒, 阿列克谢·格里岑科, 奥斯汀·斯通, 等. 使用视觉Transformer进行简单的开放词汇目标检测, 2022. 2

[27] Abby O’Neill, Rehman, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 6892-6903. IEEE, 2024. 4

艾比·奥尼尔, 雷曼, 等. Open x-embodiment: 机器人学习数据集和rt-x模型: Open x-embodiment合作0. 在2024年IEEE国际机器人与自动化会议(ICRA)上, 第6892-6903页. IEEE, 2024. 4

[28] Alec Radford, Kim, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748-8763. PMLR, 2021. 6

亚历克·拉德福德, 金, 等. 从自然语言监督中学习可迁移的视觉模型. 在国际机器学习会议上, 第8748-8763页. PMLR, 2021. 6

[29] Pierre Sermanet, Ding, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 645- 652. IEEE, 2024. 6

皮埃尔·塞尔曼内, 丁, 等. Robovqa: 用于机器人的多模态长时推理. 在2024年IEEE国际机器人与自动化会议(ICRA)上, 第645-652页. IEEE, 2024. 6

[30] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10740-10749, 2020. 2

莫希特·施里达尔, 杰西·托马森, 丹尼尔·戈登, 约纳坦·比斯克, 温森·韩, 鲁兹贝·莫塔吉, 卢克·泽特勒莫耶, 迪特尔·福克斯. Alfred: 一个用于解释日常任务中接地指令的基准. 在IEEE/CVF计算机视觉与模式识别会议论文集上, 第10740-10749页, 2020. 2

[31] Chen Song and Darwin Lau. Workspace-based model predictive control for cable-driven robots. IEEE Transactions on Robotics, 38(4):2577-2596, 2022. 2, 3

宋晨, 刘达文. 基于工作空间的电缆驱动机器人模型预测控制. IEEE机器人学汇刊, 38(4):2577-2596, 2022. 2, 3

[32] Giovanni Sutanto, Austin Wang, Yixin Lin, Mustafa Mukadam, Gaurav Sukhatme, Akshara Rai, and Franziska Meier. Encoding physical constraints in differentiable newton-euler algorithm. In Proceedings of the 2nd Conference on Learning for Dynamics and Control, pages 804-813. PMLR, 2020. 2

乔瓦尼·苏坦托, 奥斯汀·王, 林艺新, 穆斯塔法·穆卡达姆, 高拉夫·苏哈特梅, 阿克莎拉·拉伊, 弗兰齐斯卡·迈尔. 在可微牛顿-欧拉算法中编码物理约束. 在第二届学习动力学与控制会议上, 第804-813页. PMLR, 2020. 2

[33] Qwen Team. qwen2.5, 2024. 4

Qwen团队. qwen2.5, 2024. 4

[34] Dominique Thomas, Rodney Rothstein, Nathan Rosenberg, and Yolande Surdin-Kerjan. Sam2 encodes the second methionine s-adenosyl transferase in saccharomyces cerevisiae: physiology and regulation of both enzymes. Molecular and Cellular Biology, 8(12):5132-5139, 1988. 4

多米尼克·托马斯, 罗德尼·罗思坦, 内森·罗森伯格, 约兰德·苏丹-克尔詹. Sam2编码了酿酒酵母中的第二个蛋氨酸S-腺苷转移酶: 两种酶的生理学与调控. 分子与细胞生物学, 8(12):5132-5139, 1988. 4

[35] Yunsheng Tian, Karl DD Willis, Bassel Al Omari, Jieliang Luo, Pingchuan Ma, Yichen Li, Farhad Javid, Edward Gu, Joshua Jacob, Shinjiro Sueda, et al. Asap: Automated sequence planning for complex robotic assembly with physical feasibility. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 4380-4386. IEEE, 2024. 2

田云升, 卡尔·D·D·威利斯, 巴塞尔·阿尔·奥马里, 罗杰亮, 马平川, 李一辰, 法哈德·贾维德, 爱德华·顾, 约书亚·雅各布, 末田信二郎, 等. ASAP: 用于复杂机器人装配的自动化序列规划与物理可行性. 在2024年IEEE国际机器人与自动化会议(ICRA)上, 第4380-4386页. IEEE, 2024. 2

[36] An Dinh Vuong, Minh Nhat Vu, Baoru Huang, Nghia Nguyen, Hieu Le, Thieu Vo, and Anh Nguyen. Language-driven grasp detection. In Proceedings of the IEEE/CVF

范武雄, 武明南, 黄宝如, 阮玉, 黎辉, 武秋, 阮安. 语言驱动的抓取检测. 在IEEE/CVF

Conference on Computer Vision and Pattern Recognition,

计算机视觉与模式识别会议，

pages 17902-17912, 2024. 2

第17902-17912页，2024年。2

[37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024. 2

王鹏，白帅，谭思南，王世杰，范志豪，白金泽，陈克勤，刘雪静，王佳林，葛文斌等。Qwen2-vl:增强视觉语言模型在任何分辨率下的世界感知能力。arXiv预印本arXiv:2409.12191，2024年。2

[38] Yu Wu, Lu Jiang, and Yi Yang. Revisiting embodiedqa: A simple baseline and beyond. IEEE Transactions on Image Processing, 29:3984-3992, 2020. 2

吴宇，江璐，杨毅。重新审视EmbodiedQA:一个简单的基线及超越。IEEE图像处理汇刊，29:3984-3992，2020年。2

[39] An Yang, Baosong Yang, Binyuan Hui, et al. Qwen2 technical report. arXiv preprint arXiv:2407.10671, 2024. 4

杨安，杨宝松，惠宾元等。Qwen2技术报告。arXiv预印本arXiv:2407.10671，2024年。4

[40] Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. Embodied multi-modal agent trained by an from a parallel textworld. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26275- 26285, 2024. 2

杨艺俊，周天怡，李侃学，陶大鹏，李鲁松，沈立，何晓东，蒋静，石玉辉。通过平行文本世界中的 训练的具身多模态代理。IEEE/CVF计算机视觉与模式识别会议论文集，第26275-26285页，2024年。2

[41] Franziska Zacharias, Christoph Borst, and Gerd Hirzinger. Capturing robot workspace structure: representing robot capabilities. In 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 3229-3236. Ieee, 2007. 2

弗朗西斯卡·扎卡里亚斯，克里斯托夫·博斯特，格尔德·希尔辛格。捕捉机器人工作空间结构:表示机器人能力。2007年IEEE/RSJ智能机器人与系统国际会议，第3229-3236页。IEEE，2007年。2

[42] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 4

翟晓华，巴兹尔·穆斯塔法，亚历山大·科列斯尼科夫，卢卡斯·拜尔。语言图像预训练的Sigmoid损失，2023年。4

[43] Ganlong Zhao, Guanbin Li, Weikai Chen, and Yizhou Yu. Over-nav: Elevating iterative vision-and-language navigation with open-vocabulary detection and structured representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16296-16306, 2024. 2

赵甘龙，李冠斌，陈伟凯，余一舟。Over-nav:通过开放词汇检测和结构化表示提升迭代视觉与语言导航。IEEE/CVF计算机视觉与模式识别会议论文集，第16296-16306页，2024年。2

[44] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. arXiv preprint arXiv:2403.09631, 2024. 2, 6

甄浩宇，邱晓文，陈培豪，杨金成，严鑫，杜一伦，洪一宁，甘闯。3d-vla:一个3D视觉-语言-动作生成世界模型。arXiv预印本arXiv:2403.09631，2024年。2, 6