GRAVMAD: GROUNDED SPATIAL VALUE MAPS GUIDED ACTION DIFFUSION FOR GENERALIZED 3D MANIPULATION

GRAVMAD: 基于接地空间价值图的引导动作扩散用于广义三维操作

Yangtao Chen* [1] , Zixuan Chen* [1] , Junhui Yin [1] , Jing Huo [†1] , Pinzhuo Tian [3] , Jieqi Shi [2] , Yang Gao [1,2]

陈阳涛 * [1] ， 陈子轩 * [1] ， 尹俊辉 [1] ， 霍晶 [†1] ， 田品卓 [3] ， 石杰琦 [2] ， 高扬 [1,2]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China yangtaochen@smali.nju.edu.cn,{che huojing, gaoy}@nju.edu.cn, yinjunhui@smail.nju.edu.cn

[1] 新型软件技术国家重点实验室，南京大学，中国 yangtaochen@smali.nju.edu.cn,{chenzx, huojing, gaoy}@nju.edu.cn, yinjunhui@smail.nju.edu.cn

[2] School of Intelligence Science and Technology, Nanjing University (Suzhou Campus), Suzhou, China jayceesjq@gmail.com

[2] 智能科学与技术学院，南京大学 (苏州校区)，苏州，中国 jayceesjq@gmail.com

[3] School of Computer Engineering and Science, Shanghai University, Shanghai, China pinzhuo@shu.edu.cn

[3] 计算机工程与科学学院，上海大学，上海，中国 pinzhuo@shu.edu.cn

Abstract

摘要

Robots' ability to follow language instructions and execute diverse 3D manipulation tasks is vital in robot learning. Traditional imitation learning-based methods perform well on seen tasks but struggle with novel, unseen ones due to variability. Recent approaches leverage large foundation models to assist in understanding novel tasks, thereby mitigating this issue. However, these methods lack a task-specific learning process, which is essential for an accurate understanding of 3D environments, often leading to execution failures. In this paper, we introduce GravMAD, a sub-goal-driven, language-conditioned action diffusion framework that combines the strengths of imitation learning and foundation models. Our approach breaks tasks into sub-goals based on language instructions, allowing auxiliary guidance during both training and inference. During training, we introduce Sub-goal Keypose Discovery to identify key sub-goals from demonstrations. Inference differs from training, as there are no demonstrations available, so we use pre-trained foundation models to bridge the gap and identify sub-goals for the current task. In both phases, GravMaps are generated from sub-goals, providing GravMAD with more flexible 3D spatial guidance compared to fixed 3D positions. Empirical evaluations on RLBench show that GravMAD significantly outperforms state-of-the-art methods, with a 28.63% improvement on novel tasks and a 13.36% gain on tasks encountered during training. Evaluations on real-world robotic tasks further show that GravMAD can reason about real-world tasks,

associate them with relevant visual information, and generalize to novel tasks. These results demonstrate Grav-MAD's strong multi-task learning and generalization in 3D manipulation. Video demonstrations are available at: https://gravmad.github.io

机器人根据语言指令执行多样化三维操作任务的能力在机器人学习中至关重要。传统基于模仿学习的方法在已见任务上表现良好，但由于任务多样性，对新颖未见任务的适应能力较差。近期方法利用大型基础模型辅助理解新任务，从而缓解该问题。然而，这些方法缺乏针对具体任务的学习过程，而这对于准确理解三维环境至关重要，常导致执行失败。本文提出 Grav-MAD，一种基于子目标驱动、语言条件的动作扩散框架，结合了模仿学习与基础模型的优势。我们的方法根据语言指令将任务拆分为子目标，在训练和推理阶段均提供辅助引导。训练时引入子目标关键姿态发现 (Sub-goal Keypose Discovery) 以从示范中识别关键子目标。推理阶段无示范可用，故利用预训练基础模型弥合差距，识别当前任务的子目标。在两个阶段，均从子目标生成 GravMaps，为 GravMAD 提供比固定三维位置更灵活的三维空间引导。RLBench 上的实证评估表明，GravMAD 在新颖任务上显著优于最先进方法，训练任务上亦有提升。真实机器人任务评估进一步显示，GravMAD 能推理现实任务，关联相关视觉信息，并泛化至新任务。结果证明 Grav-MAD 在三维操作中的强大多任务学习与泛化能力。视频演示见:https://gravmad.github.io

1 INTRODUCTION

1 引言

One of the ultimate goals of general-purpose robot manipulation learning is to enable robots to perform a wide range of tasks in real-world 3D environments based on natural language instructions (Hu et al. 2023a). To achieve this, robots must understand task language instructions and align them with the spatial properties of relevant objects in the scene. Additionally, robots must effectively generalize across different tasks and environments; otherwise, their practical application will be limited (Zhou et al. 2023). For example, if a robot has learned the policy for the task "Take the chicken off the grill", it should also be able to perform the task "Put the chicken on the grill" Without this generalizability, its utility will be greatly reduced. Recent research in robot learning for 3D manipulation tasks has focused on two mainstream approaches: imitation learning-based methods and pre-trained foundation model-based methods. Imitation learning-based methods learn

通用机器人操作学习的终极目标之一是使机器人能够基于自然语言指令，在真实三维环境中执行广泛任务 (Hu 等，2023a)。为此，机器人必须理解任务语言指令，并将其与场景中相关物体的空间属性对齐。此外，机器人必须在不同任务和环境间有效泛化，否则其实用性将受限 (Zhou 等，2023)。例如，若机器人已学会“将鸡肉从烤架上取下”的策略，它也应能执行“将鸡肉放到烤架上”的任务。缺乏这种泛化能力，其应用价值将大打折扣。近期三维操作任务的机器人学习研究聚焦于两大主流方法: 基于模仿学习的方法和基于预训练基础模型的方法。基于模仿学习的方法学习

*These authors contributed equally.

* 这些作者贡献相同。

† Corresponding author.

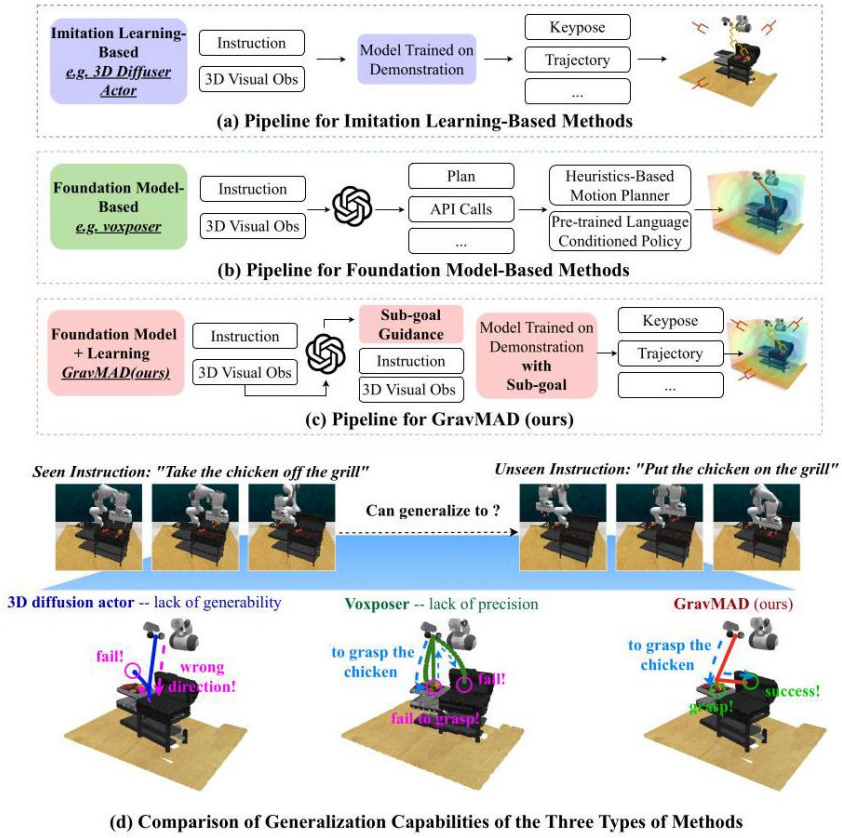


Figure 1: Comparison of Pipelines. (a) Imitation learning-based methods learn end-to-end policies that map language and 3D observations to actions for precise manipulation. (b) Foundation models-based methods use LLMs/VLMs to process inputs, generate plans, and execute actions with predefined primitives for task generalization. (c)(d) GravMAD combines both, using sub-goal guidance to leverage the language understanding of foundation models and the policy learning of imitation learning for precise and generalized manipulation.

图 1: 流程对比。(a) 基于模仿学习的方法学习端到端策略，将语言和三维观测映射到动作，实现精确操作。(b) 基于基础模型的方法利用大型语言模型 (LLMs)/视觉语言模型 (VLMs) 处理输入，生成计划，并通过预定义原语执行动作，实现任务泛化。(c)(d) GravMAD 结合两者，利用子目标引导，发挥基础模型的语言理解能力和模仿学习的策略学习优势，实现精确且泛化的操作。

end-to-end policies from expert demonstrations in attempt to address 3D manipulation tasks (Walke et al. 2023; Padalkar et al., 2024; Argall et al., 2009; Chen et al., 2024a). By designing various learning frameworks, such as incorporating different 3D representations (Shridhar et al. 2023; Chen et al., 2023a; Goyal et al., 2023), policy representations (Ze et al., 2024; Ke et al., 2024; Yan et al., 2024), and multi-stage architectures (Gervet et al. 2023, Goyal et al. 2024), imitation learning-based policies can map perceptual information and language instructions to actions that complete complex 3D manipulation tasks. However, these policies often overfit to specific tasks (Xie et al. 2024; Zhang et al. 2024), leading to significant performance degradation or even failure when applied to tasks that differ from those encountered during training (Brohan et al., 2023a; Zitkovich et al., 2023).

† 通讯作者。

通过专家示范学习端到端策略以解决三维操作任务 (Walke 等，2023；Padalkar 等，2024；Argall 等，2009；Chen 等，2024a)。通过设计多种学习框架，如引入不同的三维表示 (Shridhar 等，2023；Chen 等，2023a；Goyal 等，2023)、策略表示 (Ze 等，2024；Ke 等，2024；Yan 等，2024) 和多阶段架构 (Gervet 等，2023；Goyal 等，2024)，基于模仿学习的策略能够将感知信息和语言指令映射为完成复杂三维操作任务的动作。然而，这些策略常常对特定任务过拟合 (Xie 等，2024；Zhang 等，2024)，导致在应用于训练中未遇到的任务时性能显著下降甚至失败 (Brohan 等，2023a；Zitkovich 等，2023)。

Another line of cutting-edge research seeks to leverage foundation models trained on internet-scale data (OpenAI, 2023; Yang et al., 2023b) to enhance policy generalization across a variety of tasks (Brohan et al., 2023b; Hu et al., 2023b; Huang et al., 2023). Unlike traditional imitation learning-based methods, approaches using pre-trained foundation models typically decouple perception, reasoning, and control during manipulation (Sharan et al. 2024). However, this decoupling often leads to a limited understanding of scenes and manipulation tasks (Huang et al. 2024), allowing robots to conceptually grasp tasks but failing to accurately complete tasks in 3D environments, resulting in failures. This underscores a key challenge: both imitation learning-based and foundation model-based approaches struggle to balance precision and generalization when adapting to novel 3D manipulation tasks. Such a challenge raises a crucial question: Can the strengths of both approaches be combined to achieve precise yet generalized 3D manipulation?

另一条前沿研究路线旨在利用基于互联网规模数据训练的基础模型 (OpenAI, 2023；Yang 等, 2023b) 提升策略在多任务间的泛化能力 (Brohan 等，2023b；Hu 等，2023b；Huang 等，2023)。与传统的基于模仿学习的方法不同，使用预训练基础模型的方法通常在操作过程中将感知、推理和控制解耦 (Sharan 等，2024)。然而，这种解耦往往导致对场景和操作任务的理解有限 (Huang 等，2024)，使机器人能够概念性地把握任务，但无法准确完成三维环境中的任务，导致失败。这凸显了一个关键挑战: 基于模仿学习和基础模型的方法在适应新颖三维操作任务时都难以在精确性和泛化性之间取得平衡。由此引发一个重要问题: 能否结合两者优势，实现既精确又具泛化能力的三维操作?

To this end, inspired by the approach of introducing task sub-goals to achieve efficient execution in robotic manipulation (Black et al., 2024; Kang et al., 2023; Xian et al., 2023; Ma et al., 2024), we propose discovering key sub-goals for 3D manipulation tasks as a bridge between foundation models and learned policies, leading to the development of Grounded Spatial Value Maps-guided Action Diffusion (GravMAD), a novel sub-goals-driven, language-conditioned action diffusion framework. Specifically, a new data distillation method called Sub-goal Keypose Discovery is introduced during the training phase. This method identifies the key sub-goals required for each sub-task stage from the demonstrations. In the inference phase, pre-trained foundation models are leveraged to interpret the robot's 3D visual observations and task language instructions, directly identifying task sub-goals. Once the task sub-goals are obtained, the voxel value maps introduced in Voxposer (Huang et al. 2023) are used to generate the corresponding Grounded Spatial Value Maps (GravMaps). These maps reflect both the cost associated with each sub-goal and the ideal gripper openness. The closer to the sub-goal, the lower (cooler) the cost; the farther away, the higher (warmer) the cost, while also indicating the gripper's state within the sub-goal range. Thus, they serve as intuitive tools for grounding language instructions into 3D robotic workspaces. Finally, the generated GravMaps are integrated with the policy diffusion architecture proposed in 3D diffuser actor (Ke et al. 2024), forming the GravMAD framework. This enables the robot to utilize 3D visual observations, task language instructions, and GravMaps guidance to denoise random noise into precise end-effector poses. As shown in Fig. 1, GravMAD effectively combines the precise manipulation capabilities of imitation learning-based methods with the reasoning and generalization abilities of foundation model-based approaches. We extensively evaluate GravMAD on RLBench (James et al. 2020), a representative benchmark for instruction-following 3D manipulation tasks. The

results show that GravMAD not only performs well on tasks encountered during training but also significantly outperforms state-of-the-art baseline methods in terms of generalization to novel tasks. Additionally, we validate these findings through 10 real-world robotic manipulation tasks.

为此，受引入任务子目标以实现机器人操作高效执行方法的启发 (Black 等，2024；Kang 等，2023；Xian 等，2023；Ma 等，2024)，我们提出发现三维操作任务关键子目标，作为基础模型与学习策略之间的桥梁，进而开发了基于子目标驱动、语言条件的动作扩散新框架——基于落地空间价值图的动作扩散 (Grounded Spatial Value Maps-guided Action Diffusion，GravMAD)。具体而言，在训练阶段引入了一种名为子目标关键姿态发现 (Sub-goal Keypose Discovery) 的新型数据蒸馏方法，该方法从示范中识别每个子任务阶段所需的关键子目标。在推理阶段，利用预训练基础模型解读机器人的三维视觉观测和任务语言指令，直接识别任务子目标。获得子目标后，采用 Voxposer(Huang 等，2023) 中引入的体素价值图生成对应的落地空间价值图 (GravMaps)。这些图反映了每个子目标相关的代价及理想的夹爪开合状态。越接近子目标，代价越低 (颜色越冷)；越远离，代价越高 (颜色越暖)，同时指示夹爪在子目标范围内的状态。因此，它们作为将语言指令落地到三维机器人工作空间的直观工具。最后，将生成的 GravMaps 与 3D diffuser actor(Ke 等，2024) 提出的策略扩散架构结合，形成 GravMAD 框架，使机器人能够利用三维视觉观测、任务语言指令和 GravMaps 引导，将随机噪声去噪为精确的末端执行器姿态。如图 1 所示，GravMAD 有效结合了基于模仿学习方法的精确操作能力与基于基础模型方法的推理和泛化能力。我们在 RLBench(James 等，2020) 这一代表性指令驱动三维操作任务基准上对 GravMAD 进行了广泛评估。结果表明，GravMAD 不仅在训练任务上表现优异，在新颖任务的泛化能力上也显著优于最先进的基线方法。此外，我们通过 10 个真实机器人操作任务验证了这些发现。

In summary, our contributions are: 1) We propose leveraging key sub-goals in 3D manipulation tasks to bridge the gap between foundation models and learned policies. In the training phase, we introduce a data distillation method, Sub-goal Keypose Discovery, to identify task sub-goals. In the inference phase, foundation models are used for this purpose. 2) We generate GravMaps from these sub-goals, translating task language instructions into 3D spatial sub-goals and reflecting spatial relationships in the environment. 3) We propose a new action diffusion framework, GravMAD, guided by GravMaps. It is sub-goal-driven and language-conditioned, combining the precision of imitation learning with the generalization capabilities of foundation models. 4) The simulation experiments are conducted on 20 tasks in RLBench, comprising two types: 12 base tasks directly selected from RLBench, and 8 novel tasks created by modifying scene configurations or task instructions. GravMAD achieves at least 13.36% higher success rates than state-of-the-art baselines on the 12 base tasks encountered during training, and surpasses them by 28.63% on the 8 novel tasks, highlighting its strong generalization capabilities. Experiments on 10 real-world robotic tasks further validate GravMAD's effectiveness.

总结来说，我们的贡献包括:1) 我们提出利用 3D 操作任务中的关键子目标，弥合基础模型与学习策略之间的差距。在训练阶段，我们引入了一种数据蒸馏方法——子目标关键姿态发现 (Sub-goal Keypose Discovery)，以识别任务子目标。在推理阶段，则使用基础模型完成此任务。2) 我们从这些子目标生成 GravMaps，将任务语言指令转化为 3D 空间子目标，并反映环境中的空间关系。3) 我们提出了一种新的动作扩散框架 GravMAD，以 GravMaps 为引导。该框架以子目标驱动并基于语言条件，结合了模仿学习的精确性与基础模型的泛化能力。4) 在 RLBench 的 20 个任务上进行了仿真实验，任务分为两类:12 个直接选自 RLBench 的基础任务和 8 个通过修改场景配置或任务指令创建的新颖任务。GravMAD 在训练中遇到的 12 个基础任务上成功率至少比最先进基线高出 13.36%，在 8 个新颖任务上则超出 28.63%，彰显了其强大的泛化能力。对 10 个真实机器人任务的实验进一步验证了 GravMAD 的有效性。

2 RELATED WORKS

2 相关工作

Learning 3D Manipulation Policies from Demonstrations. Recent works have employed various perception methods to learn 3D manipulation policies from demonstrations to tackle the complexity of reasoning in 3D space. These methods include using 2D images (Chen et al. 2024b; Zitkovich et al., 2023; Jang et al., 2022), voxels (Shridhar et al., 2023; James et al., 2022), point clouds (Chen et al., 2023a; Yuan et al., 2023), multi-view virtual images (Chen et al., 2023b; Goyal et al., 2024), and feature fields (Gervet et al. 2023). To support policy learning, some studies (Ke et al. 2024; Xian et al., 2023; Yan et al., 2024; Ze et al., 2024) have integrated 3D scene representations with diffusion models (Ho et al., 2020). These approaches attempt to handle the multi-modality of actions, in contrast to behavior cloning methods that train deterministic policies. By leveraging 3D representation learning, these policies can accurately complete tasks by accounting for the spatial properties of objects, such as orientation and position. This is especially effective for tasks that closely resemble those encountered during training (Ze et al. 2023). However, these policies often lack the language understanding and generalization abilities of foundation models. Our method builds upon the diffusion architecture (Ke et al., 2024), enhancing its ability to utilize demonstration data through imitation learning, while integrating foundation models to improve generalization, combining the strengths of both approaches.

从示范中学习 3D 操作策略。近期工作采用多种感知方法，从示范中学习 3D 操作策略，以应对 3D 空间推理的复杂性。这些方法包括使用二维图像 (Chen et al. 2024b；Zitkovich et al., 2023；Jang et al., 2022)、体素 (Shridhar et al., 2023；James et al., 2022)、点云 (Chen et al., 2023a；Yuan et al., 2023)、多视角虚拟图像 (Chen et al., 2023b；Goyal et al., 2024) 和特征场 (Gervet et al. 2023)。为支持策略学习，一些研究 (Ke et al. 2024；Xian et al., 2023；Yan et al., 2024；Ze et al., 2024) 将 3D 场景表示与扩散模型 (Ho et al., 2020) 结合。这些方法试图处理动作的多模态性，有别于训练确定性策略的行为克隆方法。通过利用 3D 表示学习，这些策略能够准确完成任务，考虑物体的空间属性，如朝向和位置。这对与训练时任务高度相似的任务尤为有效 (Ze et al. 2023)。然而，这些策略通常缺乏基础模型的语言理解和泛化能力。我们的方法基于扩散架构 (Ke et al., 2024)，通过模仿学习增强其利用示范数据的能力，同时整合基础模型以提升泛化，结合两者优势。

Foundation Models for 3D Manipulation. Recent foundation models trained on internet-scale data have shown strong zero-shot and few-shot generalization, offering new opportunities for complex 3D manipulation tasks (Hu et al., 2023a; Zhou et al., 2023). While some approaches fine-tune

用于 3D 操作的基础模型。近期基于互联网规模数据训练的基础模型展现了强大的零样本和少样本泛化能力，为复杂 3D 操作任务提供了新机遇 (Hu et al., 2023a；Zhou et al., 2023)。尽管部分方法进行了微调

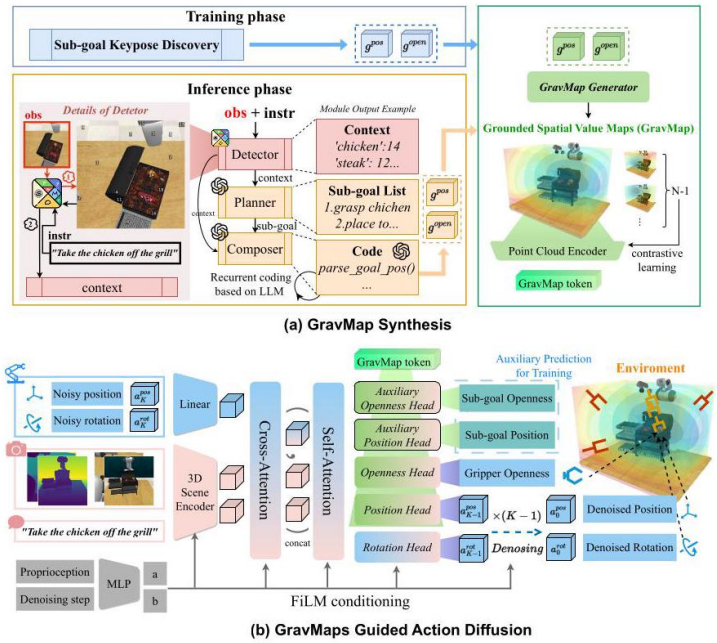


Figure 2: GravMAD Overview. (a) GravMap Synthesis: During training, we use Sub-goal Keypose Discovery to obtain sub-goals g^{pos} and g^{open}. During inference, the Detector, Planner, and Composer pipeline interprets visual observations and language instructions to derive g^{pos} and g^{open}, which are processed into a GravMap and encoded as a GravMap token. (b) GravMaps Guided Action Diffusion: The policy network perceives the scene and denoises noisy actions guided by the GravMap token. After K denoising steps, the clean actions are executed by the robot.

图 2:GravMAD 概览。(a)GravMap 合成: 训练期间，我们使用子目标关键姿态发现获得子目标 g^{pos} 和 g^{open}。推理时，检测器、规划器和合成器流水线解读视觉观测和语言指令，推导出 g^{pos} 和 g^{open}，将其处理成 GravMap 并编码为 GravMap 标记。(b)GravMaps 引导的动作扩散: 策略网络感知场景，并在 GravMap 标记引导下对噪声动作进行去噪。经过 K 步去噪后，机器人执行干净的动作。

vision-language models with embodied data (Driess et al. 2023; Li et al. 2024), this increases computational costs due to the large data requirements. Alternatively, foundational vision models can generate visual representations for 3D manipulation tasks (Zhang et al. 2024; 2023), but they often lack the reasoning capabilities needed for complex tasks. To address these challenges, some studies leverages large language models (LLMs) as high-level planners (Brohan et al. 2023b; Hu et al., 2023b; Huang et al., 2022), generating language-based plans executed by lower-level policies. Others utilize LLMs' code-writing abilities to control robots via API calls or to create value maps for planning robot trajectories (Liang et al., 2023; Huang et al., 2023). However, these methods often sacrifice precision due to a rough understanding of complex 3D scenes. Recent works have combined the reasoning capabilities of foundation models with fine-grained control in 3D manipulation to overcome this limitation (Huang et al. 2024; Sharan et al. 2024). For example, Huang et al. (2024) uses pre-trained vision-language models (VLMs) to provide spatial constraints and a nonlinear solver to generate precise grasp poses. Our method combines the learning power of diffusion architectures with the generalization of VLMs. VLMs generate spatial value maps that guide action diffusion, enabling precise control and multi-task generalization in 3D manipulation tasks.

结合具身数据的视觉-语言模型 (Driess 等，2023；Li 等，2024)，由于大量数据需求，计算成本增加。另一种方法是基础视觉模型可为三维操作任务生成视觉表示 (Zhang 等，2024；2023)，但它们通常缺乏完成复杂任务所需的推理能力。为解决这些挑战，一些研究利用大型语言模型 (LLMs) 作为高层规划者 (Brohan 等，2023b；Hu 等，2023b；Huang 等，2022)，生成基于语言的计划，由低层策略执行。另一些则利用 LLMs 的代码编写能力，通过 API 调用控制机器人或创建价值地图以规划机器人轨迹 (Liang 等，2023；Huang 等，2023)。然而，这些方法因对复杂三维场景理解粗糙，常牺牲精度。近期工作结合了基础模型的推理能力与三维操作中的细粒度控制以克服此限制 (Huang 等，2024；Sharan 等，2024)。例如，Huang 等 (2024) 使用预训练视觉-语言模型 (VLMs) 提供空间约束，并用非线性求解器生成精确抓取姿态。我们的方法结合了扩散架构的学习能力与 VLMs 的泛化能力。VLMs 生成空间价值地图，引导动作扩散，实现三维操作任务中的精确控制和多任务泛化。

3 METHOD

3 方法

In this section, we introduce GravMAD, a multi-task, sub-goal-driven, language-conditioned diffusion framework for 3D manipulation, as shown in Fig. 2 We divide GravMAD's design into three parts: Section 3.1 defines the problem setting, Section 3.2 explains the definition and generation of GravMaps, and Section 3.3 details how GravMaps guide action diffusion in 3D manipulation.

本节介绍 GravMAD，一种多任务、子目标驱动、语言条件的三维操作扩散框架，如图 2 所示。我们将 GravMAD 的设计分为三部分: 第 3.1 节定义问题设置，第 3.2 节解释 GravMaps 的定义与生成，第 3.3 节详述 GravMaps 如何指导三维操作中的动作扩散。

3.1 Problem Formulation

3.1 问题表述

We consider a problem setting where expert demonstrations consist of a robot trajectory $(o_1, a_1, o_2, a_2, \ldots)$ and a natural language instruction $\ell \in \mathcal{L}$ that describes the task goal. Each

我们考虑的问题设置中，专家示范包含机器人轨迹 $(o_1, a_1, o_2, a_2, \ldots)$ 和描述任务目标的自然语言指令 $\ell \in \mathcal{L}$。每个

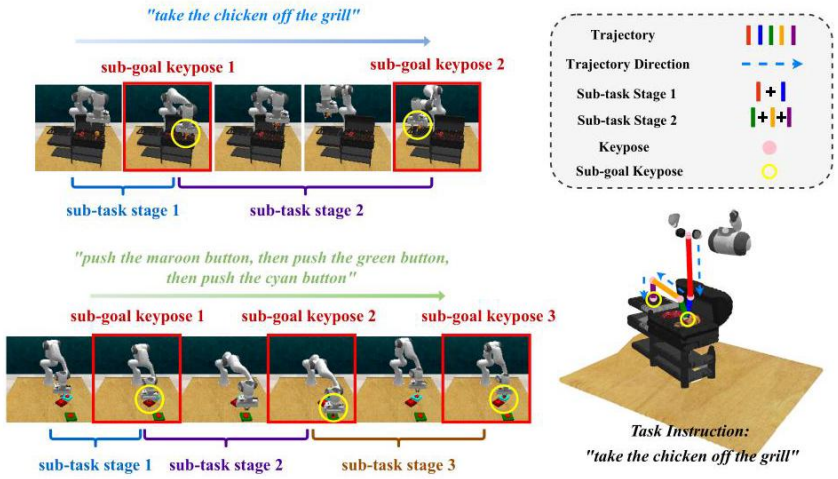


Figure 3: Visualization of sub-goal keyposes and sub-task stages. The left sub-figure shows image-based sub-goal keyposes and sub-task stages for ”take the chicken off the grill” and ”push the_____ button” tasks. The right shows the sub-goal key poses and sub-task stages in the trajectory for the ”take the chicken off the grill” task.

图 3: 子目标关键姿态和子任务阶段的可视化。左图展示了“将鸡肉从烤架上取下”和“按下 _____ 按钮”任务的基于图像的子目标关键姿态和子任务阶段。右图展示了“将鸡肉从烤架上取下”任务轨迹中的子目标关键姿态和子任务阶段。

observation $o_t \in \mathcal{O}$ includes RGB-D images from one or more viewpoints. Each action $a_t \in \mathcal{A}$ contains the 3D position of the robot's end-effector $a^{\text{pos}} \in \mathbb{R}^3$, a 6D rotation $a^{\text{rot}} \in \mathbb{R}^6$, and a binary gripper state $a^{\text{open}} \in \{0, 1\}$. To address potential discontinuities from quaternion constraints and ensure smooth optimization, we utilize the 6D rotation representation (Ke et al. 2024). In this setting, we assume that a robotic task is composed of multiple sub-tasks, with each sub-task completed when the robot reaches a sub-goal $g_t \in \mathcal{G}$, which specifies the 3D position $g^{\text{pos}} \in \mathbb{R}^3$, the gripper openness $g^{\text{open}} \in \{0, 1\}$, and the 6D rotation $g^{\text{rot}} \in \mathbb{R}^6$. Based on this, we construct a new dataset $\mathcal{D} = \{\zeta_1, \zeta_2, \ldots\}$ from expert demonstrations. Each demonstration ζ consists of trajectories with sub-goals $\{(o_1, g_1, a_1), (o_2, g_2, a_2), \ldots\}$ and the corresponding language instruction ℓ. Our goal is to learn a policy $\pi : (\mathcal{O}, \mathcal{L}, \mathcal{G}) \mapsto \mathcal{A}$, which maps observations o_t, sub-goals g_t, and instructions ℓ to actions a_t. To facilitate sub-task segmentation and efficiently learn the policy, we frame the robot's 3D manipulation learning problem as a keypose prediction problem following prior works (James & Davison, 2022; James et al., 2022; Goyal et al., 2023; Shridhar et al., 2023). Our model progressively predicts the next keypose based on current observations and uses a sampling-based motion planner (Klemm et al. 2015) to plan the trajectory between two keyposes. In the existing keypose discovery method (James & Davison, 2022), a pose is identified as a keypose when the robot's joint velocities are near zero, and the gripper state remains unchanged. Our work filters the task's sub-goals based on these keyposes to facilitate sub-task segmentation and ensure efficient completion of the overall task.

观测 $o_t \in \mathcal{O}$ 包括来自一个或多个视角的 RGB-D 图像。每个动作 $a_t \in \mathcal{A}$ 包含机器人末端执行器的位置 3D $a^{\text{pos}} \in \mathbb{R}^3$、旋转 6D $a^{\text{rot}} \in \mathbb{R}^6$ 以及二元夹爪状态 $a^{\text{open}} \in \{0, 1\}$。为了解决四元数约束可能导致的不连续性并确保优化的平滑性，我们采用了旋转表示方法 (Ke 等，2024)。在此设置中，我们假设机器人任务由多个子任务组成，每个子任务在机器人达到子目标 $g_t \in \mathcal{G}$ 时完成，该子目标指定了位置 3D $g^{\text{pos}} \in \mathbb{R}^3$、夹爪开合度 $g^{\text{open}} \in \{0, 1\}$ 和旋转 6D $g^{\text{rot}} \in \mathbb{R}^6$。基于此，我们从专家示范中构建了一个新数据集 $\mathcal{D} = \{\zeta_1, \zeta_2, \ldots\}$。每个示范 ζ 由带有子目标 $\{(o_1, g_1, a_1), (o_2, g_2, a_2), \ldots\}$ 的轨迹及相应的语言指令 ℓ 组成。我们的目标是学习一个策略 $\pi : (\mathcal{O}, \mathcal{L}, \mathcal{G}) \mapsto \mathcal{A}$，将观测 o_t、子目标 g_t 和指令 ℓ 映射到动作 a_t。为了促进子任务分割并高效学习策略，我们将机器人的三维操作学习问题框定为关键姿态预测问题，遵循先前工作 (James & Davison, 2022；James 等，2022；Goyal 等，2023；Shridhar 等，2023)。我们的模型基于当前观测逐步预测下一个关键姿态，并使用基于采样的运动规划器 (Klemm 等，2015) 规划两个关键姿态之间的轨迹。在现有的关键姿态发现方法 (James & Davison, 2022) 中，当机器人关节速度接近零且夹爪状态保持不变时，该姿态被识别为关键姿态。我们的工作基于这些关键姿态过滤任务的子目标，以促进子任务分割并确保整体任务的高效完成。

3.2 GRAVMAP: GROUNDED SPATIAL VALUE MAPS

3.2 GRAVMAP: 基于环境的空间价值图

To tackle generalization challenges in 3D manipulation tasks, we introduce the spatial value maps (GravMap), an adaptation of the voxel value maps proposed by Huang et al., denoted as m . GravMaps are adaptively synthesized based on task variations, translating language instructions into 3D spatial sub-goals and reflecting the spatial relationships within the environment. This provides precise guidance for robotic action diffusion. Each GravMap m contains two voxel maps: (1) a spatial cost map m_c , with lower values near the sub-goal and higher costs further away, and (2) a gripper openness map m_o , indicating where the gripper should open or close. As shown in Fig. 2(a), GravMaps are generated differently for training and inference. In training, they are identified from expert demonstrations using the sub-goal keypose discovery method. During inference, pre-trained models generate them from language instructions and observed images.

为应对三维操作任务中的泛化挑战，我们引入了空间价值图 (GravMap)，这是对 Huang 等人提出的体素价值图的改进，记作 m。GravMap 根据任务变化自适应合成，将语言指令转化为三维空间子目标，并反映环境中的空间关系，为机器人动作扩散提供精确指导。每个 GravMap m 包含两个体素图:(1) 空间代价图 m_c，子目标附近值较低，远离子目标代价较高；(2) 夹爪开合图 m_o，指示夹爪应开合的位置。如图 2(a) 所示，GravMap 在训练和推理阶段的生成方式不同。训练时，利用子目标关键姿态发现方法从专家示范中识别生成；推理时，预训练模型根据语言指令和观测图像生成。

GravMap Synthesis with Sub-goal keypose Discovery during Training. We define each sub-task stage in 3D manipulation as: (1) the process where the robotic end-effector transitions from not touching an object to making contact, or (2) the interaction between the end-effector or tool and a new object, where a series of operations are performed before disengaging. To efficiently segment these sub-task stages and find sub-goals, we build upon the existing keypose discovery method (James & Davison, 2022) and propose a novel data distillation method called sub-goal keypose discovery.

训练过程中通过子目标关键姿态发现进行 GravMap 合成。我们将 3D 操作中的每个子任务阶段定义为:(1) 机器人末端执行器从未接触物体到接触物体的过程，或 (2) 末端执行器或工具与新物体之间的交互，在断开前执行一系列操作。为了高效地分割这些子任务阶段并找到子目标，我们基于现有的关键姿态发现方法 (James & Davison, 2022)，提出了一种新颖的数据蒸馏方法，称为子目标关键姿态发现。

The sub-goal keypose discovery process iterates over each keypose $K_p^i \in \{K_p\}_1^{N_k}$, where N_k is the number of keyposes in a task. For each keypose, the corresponding observation-action pair $\left(o_{K_p^i}, a_{K_p^i}\right)$ is passed to the function S_K , which outputs a Boolean value to determine whether the given keypose should be discovered as a sub-goal keypose. The decision is made based on whether the keypose satisfies the discovery constraint: $S_K\left(\left(o_{K_p^i}, a_{K_p^i}\right)\right) = \begin{cases} 1, & \text{if discovery constraints are met} \\ 0, & \text{otherwise} \end{cases}$. The function S_K can incorporate multiple constraints. In our paper, we define two constraints for S_K , depending on the type of manipulation task, as shown in Fig. 3 (1) For grasping tasks, such as "take the chicken off the grill", sub-goal keyposes are discovered based on the following constraints: a change in the gripper's open/close state and a significant change in touch force. (2) For contact-based tasks, such as "push the "push the "button", sub-goal keyposes are discovered solely based on significant changes in touch force. For more details on sub-goal keypose discovery, please refer to Appendix A. 2

子目标关键姿态发现过程遍历每个关键姿态 $K_p^i \in \{K_p\}_1^{N_k}$ ，其中 N_k 为任务中的关键姿态数量。对于每个关键姿态, 传入对应的观察-动作对 $\left(o_{K_p^i}, a_{K_p^i}\right)$ 至函数 S_K ，该函数输出布尔值以判断该关键姿态是否应被发现为子目标关键姿态。决策基于关键姿态是否满足发现约束: $S_K\left(\left(o_{K_p^i}, a_{K_p^i}\right)\right) = \begin{cases} 1, & \text{if discovery constraints are met} \\ 0, & \text{otherwise} \end{cases}$ 。函数 S_K 可包含多个约束。本文中，我们根据操作任务类型为 S_K 定义了两种约束，如图 3 所示:(1) 对于抓取任务，如“从烤架上取下鸡肉”，子目标关键姿态基于夹爪开合状态变化和触觉力显著变化发现；(2) 对于基于接触的任务，如“按按钮”，子目标关键姿态仅基于触觉力显著变化发现。更多子目标关键姿态发现细节请参见附录 A.2。

After discovering the sub-goal keyposes, the sub-task stages can be quickly segmented, and the corresponding sub-goals can be identified. The end-effector position g^{pos} and gripper openness g^{open} at these sub-goals are then input to the Gravmap generator to generate the GravMaps m for training The process of the GravMap generator is illustrated in Algorithm 1 in Appendix A.1, adapted from Huang et al. (2023).

发现子目标关键姿态后，可快速分割子任务阶段并识别对应子目标。然后将这些子目标处的末端执行器位置 g^{pos} 和夹爪开合度 g^{open} 输入 GravMap 生成器，生成用于训练的 GravMaps m 。GravMap 生成器的过程见附录 A.1 中的算法 1，改编自 Huang 等人 (2023)。

GravMap Synthesis with Foundation Model during Inference. During the inference phase, we use pre-trained foundation models to synthesize GravMaps. First, to enable the robot to tie the task-related words with their manifestation in the 3D environment, we introduce a Set-of-Mark (SoM) (Yang et al. 2023a)-based Detector. This Detector uses Semantic-SAM (Li et al. 2023) to perform semantic segmentation on the observed RGB images and assigns numerical tags to the segmented regions Next, the Detector uses GPT-4o to select task-relevant objects and their corresponding tags from the labeled images as contextual information $\mathcal{C}$. Based on the task instructions ℓ and the context $\mathcal{C}$ provided by the Detector, we apply the LLM-based Planner proposed by Huang et al. to infer a series of text-based sub-goals. Then, an LLM-based Composer (Huang et al. 2023) recursively generates code to parse each sub-goal. During execution, the code uses the context $\mathcal{C}$ to obtain the end-effector positions g^{pos} and

gripper openness states g^{open} corresponding to each sub-goal. Finally, g^{pos} and g^{open} are fed into the GravMap generator shown in Algorithm 1, skipping the data augmentation process to generate the GravMaps. Details of this process can be found in Appendix A.3.2

推理阶段基于基础模型进行 GravMap 合成。推理时，我们使用预训练基础模型合成 GravMaps。首先，为使机器人将任务相关词汇与 3D 环境中的表现关联，我们引入基于 Set-of-Mark(SoM)(Yang 等，2023a) 的检测器。该检测器利用 Semantic-SAM(Li 等，2023) 对观察到的 RGB 图像进行语义分割，并为分割区域分配数值标签。接着，检测器使用 GPT-4o 从标注图像中选择任务相关物体及其对应标签作为上下文信息 $\mathcal{C}$ 。基于任务指令 ℓ 和检测器提供的上下文 $\mathcal{C}$ ，我们应用 Huang 等人提出的基于大语言模型 (LLM) 的规划器推断一系列文本子目标。随后，基于 LLM 的编排器 (Huang 等，2023) 递归生成代码以解析每个子目标。执行过程中，代码利用上下文 $\mathcal{C}$ 获取对应每个子目标的末端执行器位置 g^{pos} 和夹爪开合状态 g^{open} 。最后，将 g^{pos} 和 g^{open} 输入算法 1 所示的 GravMap 生成器，跳过数据增强过程生成 GravMaps。该过程详情见附录 A.3.2。

We synthesize the GravMaps via sub-goal keypose discovery during training or foundation models during inference. GravMaps m are then downsampled using farthest point sampling (FPS) and encoded into token t_m with the DP3 (Ze et al. 2024) encoder, a lightweight MLP network.

我们通过训练期间的子目标关键姿态发现或推理期间的基础模型合成 GravMaps(重力图)。然后，GravMaps m 使用最远点采样 (FPS) 进行下采样，并通过 DP3(Ze 等，2024) 编码器——一个轻量级的多层感知机 (MLP) 网络——编码成 token t_m 。

3.3 GrayMaps Guided Action Diffusion

3.3 GravMaps 引导的动作扩散

After obtaining the GravMaps, they can be used to guide the action diffusion process, as shown in Fig. 2(b). Before the diffusion process begins, the robot should first perceive the 3D environment.

获得 GravMaps 后，可以用它们来引导动作扩散过程，如图 2(b) 所示。在扩散过程开始之前，机器人应首先感知三维环境。

3D Scene Perception. Building on previous works (Gervet et al., 2023; Ke et al., 2024), we use a 3D scene encoder to transform language instructions and multi-view RGB-D images into scene tokens, enhancing the robot's 3D scene perception. RGB images are encoded using a pre-trained CLIP ResNet50 backbone (Radford et al. 2021) and a feature pyramid network. These features are lifted into 3D feature clouds using 3D positions derived from depth images and camera intrinsics. Simultaneously, the CLIP language encoder converts task instructions into language tokens. These tokens interact with the 3D feature cloud to generate scene tokens (t_s) , enabling the robot to capture 3D environmental information.

三维场景感知。基于先前的工作 (Gervet 等，2023；Ke 等，2024)，我们使用三维场景编码器将语言指令和多视角 RGB-D 图像转换为场景 token，增强机器人的三维场景感知。RGB 图像通过预训练的 CLIP ResNet50 骨干网络 (Radford 等，2021) 和特征金字塔网络进行编码。这些特征利用深度图像和相机内参得到的三维位置提升为三维特征点云。同时，CLIP 语言编码器将任务指令转换为语言 token。这些 token 与 3D 特征点云交互生成场景 token (t_s)，使机器人能够捕捉三维环境信息。

GravMaps Guided Action Diffusion. GravMAD builds upon the 3D trajectory diffusion architecture introduced by 3D Diffuser Actor (Ke et al. 2024) and further integrates GravMap tokens t_m to guide the action diffusion process. Specifically, GravMAD models policy learning as the reconstruction of the robot's end-effector pose using diffusion probabilistic models (DDPMs) (Ho et al. 2020). The end-effector pose is represented as $e = (a^{\text{pos}}, a^{\text{rot}})$. Starting with Gaussian noise $e_K = \left(a_K^{\text{pos}}, a_K^{\text{rot}}\right)$, the denoising networks $\epsilon_\theta^{\text{pos}}$ and $\epsilon_\theta^{\text{rot}}$ perform K iterative steps to progressively reconstruct the clean

GravMaps 引导的动作扩散。GravMAD 基于 3D Diffuser Actor(Ke 等，2024) 提出的三维轨迹扩散架构，进一步整合了 GravMap token t_m 以引导动作扩散过程。具体而言，GravMAD 将策略学习建模为使用扩散概率模型 (DDPMs)(Ho 等，2020) 重构机器人的末端执行器姿态。末端执行器姿态表示为 $e = (a^{\text{pos}}, a^{\text{rot}})$。从高斯噪声 $e_K = \left(a_K^{\text{pos}}, a_K^{\text{rot}}\right)$ 开始，去噪网络 $\epsilon_\theta^{\text{pos}}$ 和 $\epsilon_\theta^{\text{rot}}$ 执行 K 次迭代步骤，逐步重建干净的

pose $e_0 = \left(a_0^{\text{pos}}, a_0^{\text{rot}}\right)$:

姿态 $e_0 = \left(a_0^{\text{pos}}, a_0^{\text{rot}}\right)$:

$$a_{k-1}^{\text{pos}} = \alpha\left(a_k^{\text{pos}} - \gamma\epsilon_\theta^{\text{pos}}\left(e_k, k, p, t_s, t_m\right) + \mathcal{N}\left(0, \sigma^2 I\right)\right), \tag{1}$$

$$a_{k-1}^{\text{rot}} = \alpha\left(a_k^{\text{rot}} - \gamma\epsilon_\theta^{\text{rot}}\left(e_k, k, p, t_s\right) + \mathcal{N}\left(0, \sigma^2 I\right)\right),$$

where α, γ, and σ are functions of the iteration step k, determined by the noise schedule. $\mathcal{N}\left(0, \sigma^2 I\right)$ is Gaussian noise. Here, p represents proprioceptive information (a short action history). The denoising networks use 3D relative position attention layers (Gervet et al., 2023, Xian et al., 2023, Ke et al. 2024), with FiLM (Perez et al. 2018) conditioning applied to each layer based on proprioception p and denoising step k. As shown in Fig. 2(b), after passing through linear layers, a_K^{pos} and a_K^{rot} are concatenated and attend to the 3D scene tokens t_s via cross-attention. A self-attention layer then refines this representation to produce end-effector contextual features. These features are processed by five prediction heads: the Position Head, Rotation Head, Openness Head, Auxiliary Openness Head, and Auxiliary Position Head. In all but the rotation head, contextual features undergo cross-attention with GravMap tokens, followed by an MLP to predict the target values. See Appendix A. 4 for details.

其中 α, γ 和 σ 是迭代步数 k 的函数，由噪声调度决定。$\mathcal{N}\left(0, \sigma^2 I\right)$ 是高斯噪声。这里，p 表示本体感受信息 (短期动作历史)。去噪网络采用三维相对位置注意力层 (Gervet 等，2023；Xian 等，2023；Ke 等，2024)，并基于本体感受 p 和去噪步数 k 对每层应用 FiLM(Perez 等，2018) 条件调节。如图 2(b) 所示，经过线性层后，a_K^{pos} 和 a_K^{rot} 被拼接并通过交叉注意力关注三维场景 token t_s。随后，自注意力层对该表示进行细化，生成末端执行器的上下文特征。这些特征由五个预测头处理: 位置头、旋转头、开合度头、辅助开合度头和辅助位置头。除旋转头外，所有上下文特征均与 GravMap token 进行交叉注意力，然后通过 MLP 预测目标值。详情见附录 A.4。

The first two prediction heads predict the noise added to the original pose using the $L1$ norm, with the losses defined as:

前两个预测头使用 $L1$ 范数预测加到原始姿态上的噪声，损失定义为:

$$\mathcal{L}_{\text{pos}} = \left\| \epsilon_k^{\text{pos}} - \epsilon_\theta^{\text{pos}} \left(e_k, k, p, t_s, t_m\right) \right\|, \tag{2}$$

$$\mathcal{L}_{\text{rot}} = \left\| \epsilon_k^{\text{rot}} - \epsilon_\theta^{\text{rot}} \left(e_k, k, p, t_s\right) \right\|,$$

where iteration k is randomly selected, and ϵ_k^{pos} and ϵ_k^{rot} are randomly sampled as the ground truth noise.

其中迭代步数 k 随机选择，ϵ_k^{pos} 和 ϵ_k^{rot} 随机采样作为真实噪声。

The third prediction head is used to predict the gripper's open/close state, and we use binary cross-entropy (BCE) loss for supervision:

第三个预测头用于预测夹爪的开合状态，我们使用二元交叉熵 (BCE) 损失进行监督:

$$\mathcal{L}_{\text{open}} = \text{BCE}\left(f_\theta^{\text{open}} \left(e_k, k, p, t_s, t_m\right), a^{\text{open}}\right) \tag{3}$$

The last two prediction heads enable GravMAD to better focus on the ideal end-effector pose at sub-goals, with the loss functions defined as follows:

最后两个预测头使 GravMAD 能够更好地聚焦于子目标处理想的末端执行器姿态，损失函数定义如下:

$$\mathcal{L}_{\text{aux\_pos}} = \left\| g^{\text{pos}} - f_\theta^{\text{aux\_pos}} \left(e_k, k, p, t_s, t_m\right)\right\|, \tag{4}$$

$$\mathcal{L}_{\text{aux\_open}} = \text{BCE}\left(f_\theta^{\text{aux\_open}} \left(e_k, k, p, t_s, t_m\right), g^{\text{open}}\right),$$

where f_θ represents the pose prediction network in GravMAD, while g^{pos} and g^{open} denote the ground truth sub-goal positions and gripper openness, respectively.

其中 f_θ 表示 GravMAD 中的姿态预测网络，g^{pos} 和 g^{open} 分别表示真实的子目标位置和夹爪开合度。

In addition to the losses related to robot actions mentioned above, a contrastive learning loss is applied to enhance feature representations from GravMaps. Positive pairs are features from the same GravMap, while negative pairs come from different GravMaps. In each forward pass, one GravMap is extracted from the dataset, and $N-1$ different GravMaps are randomly generated. The loss maximizes similarity between positive pairs and minimizes it between negative pairs:

除了上述与机器人动作相关的损失外，还应用了对比学习损失以增强 GravMaps 的特征表示。正样本对来自同一 GravMap，负样本对则来自不同的 GravMap。在每次前向传播中，从数据集中提取一个 GravMap，并随机生成 $N-1$ 个不同的 GravMap。该损失最大化正样本对的相似度，最小化负样本对的相似度:

$$\mathcal{L}_{\text{con}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(f_{g_i}\cdot f_{g_i}^{+}/T\right)}{\sum_{j=1}^{N}\exp\left(f_{g_i}\cdot f_{g_j}/T\right)}, \tag{5}$$

where T is the temperature parameter, f_{g_i} represents the feature of the i -th sample, and $f_{g_i^+}$ represents the positive feature of the i -th sample.

其中 T 是温度参数，f_{g_i} 表示第 i 个样本的特征，$f_{g_i^+}$ 表示第 i 个样本的正样本特征。

At this stage, the training objective of GravMAD can be formulated by combining the losses from Eq. 2, 3, 4, and 5 as follows:

在此阶段，GravMAD 的训练目标可通过结合公式 2、3、4 和 5 中的损失函数表述如下:

$$\mathcal{L}_{\text{GravMAD}} = \mathcal{L}_{\text{open}} + \omega_1\cdot\mathcal{L}_{\text{pos}} + \omega_2\cdot\mathcal{L}_{\text{rot}} + \omega_3\cdot\mathcal{L}_{\text{aux\_pos}} + \mathcal{L}_{\text{aux\_open}} + \omega_4\cdot\mathcal{L}_{\text{con}}, \tag{6}$$

where $\omega_1, \omega_2, \omega_3, \omega_4$ are adjustable hyperparameters. For more detailed implementation of GravMap and GravMAD, please refer to Appendix A.

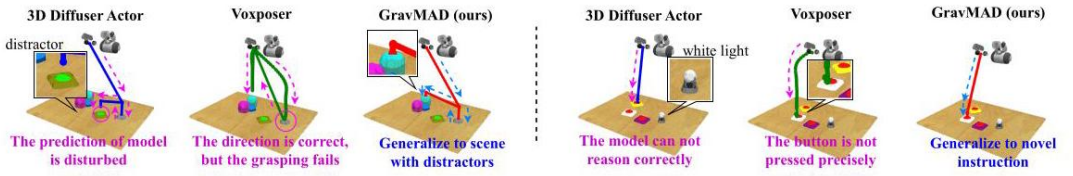
其中 $\omega_1, \omega_2, \omega_3, \omega_4$ 为可调节的超参数。有关 GravMap 和 GravMAD 的更详细实现，请参见附录 A。

4 EXPERIMENTS

4 实验

We aim to answer the following questions: (i) Can GravMAD achieve superior generalization in novel 3D manipulation tasks compared to SOTA models? (See Sec. 4.2) (ii) Is GravMAD's performance competitive on the 3D manipulation tasks encountered during training? (See Sec. 4.3) (iii) What key design elements contribute significantly to GravMAD's overall performance? (See Sec. 4.4)

我们旨在回答以下问题:(i) 与最先进模型相比，GravMAD 能否在新颖的 3D 操作任务中实现更优的泛化能力？(见第 4.2 节)(ii)GravMAD 在训练过程中遇到的 3D 操作任务上的表现是否具有竞争力？(见第 4.3 节)(iii) 哪些关键设计元素对 GravMAD 的整体性能贡献显著？(见第 4.4 节)



Models	Avg. Success ↑	Avg. Rank ↓	Close Drawer	Close Jar Banana	Close Jar Distractor	Condition Block	Meat On Grill	Open Drawer Small	Stack cups blocks	Push Buttons Light
Voxposer Huang et al. 2023	34.29	2.6	96.00 ±4.00	17.33 ±19.73	22.67 ±10.07	25.00 ±23.26	38.67 ±12.22	6.67 ±2.31	0.00 ±0.00	68.00 ±18.33
Act3D Gervet et al. 2023	17.83	3.5	66.67±9.24	29.33±9.24	41.33±4.62	0.00±0.00	1.33±2.31	2.67±4.62	0.00±0.00	1.33±2.31
3D Diffuser Actor Ke et al. 2024	29.38	2.9	81.33±6.11	48.00±4.00	42.67±4.62	27.00±10.15	0.00±0.00	2.67±4.62	2.67±2.31	30.67±12.86
GravMAD (VLM)	**62.92**	1.0	**97.33**±2.31	**84.00**±00.00	**86.67**±2.31	**74.00**±11.14	**45.33**±4.62	**21.33**±12.86	**18.67**±2.31	**76.00**±8.00
Performance gain	**+28.63**	-	**+1.33**	**+36.00**	**+44.00**	**+47.00**	**+6.66**	**+14.66**	**+16.00**	**+8.00**

Table 1: Generalization to 8 novel RLBench tasks. Evaluations on 8 novel tasks are conducted using 3 seeds, with 25 test episodes per task, utilizing the final checkpoints from training on 12 base tasks. Performance gains are compared to the best-performing baselines, indicated by underlines.

表 1: 对 8 个新颖 RLBench 任务的泛化能力。对 8 个新任务的评估使用 3 个随机种子，每个任务测试 25 个回合，利用在 12 个基础任务上训练得到的最终检查点。性能提升与表现最好的基线模型比较，提升部分用下划线标出。

4.1 ENVIRONMENTAL SETUP

4.1 环境设置

To thoroughly investigate these questions, we conduct our experiments on a representative instruction-following 3D manipulation benchmark, RLBench (James et al. 2020). Simulation experiments are conducted on two types of tasks to provide a comprehensive evaluation of GravMAD. 1) Base tasks, To evaluate GravMAD's performance across 3D manipulation tasks encountered during training, we select 12 base tasks from RLBench's 100 language-conditioned tasks, each featuring 2 to 60 variations in instructions, such as handling objects of different colors or quantities. For each base task, we collect 20 demonstrations for training and evaluate the final checkpoints using 3 random seeds over 25 episodes. Detailed descriptions of these tasks are provided in Appendix B.1 2) Novel tasks. To further test GravMAD's generalization capabilities, we modify the scene configurations or task instructions of several base tasks to create 8 novel tasks across 3 novelty categories as illustrated in fig. 10. These modifications introduce significant challenges for the robot regarding instruction comprehension, environmental perception, and policy generalization, as described in Appendix B.2 For each novel task, we evaluate the final checkpoints trained on the 12 base tasks. We use 3 random seeds over 25 episodes for each novel task. For all tasks, we use a front-view 256 X 256 RGB-D camera and a Franka Panda robot with parallel grippers. Additionally, we further validate GravMAD on 10 real-world robotic tasks, with details provided in Appendix D.6

为全面探讨这些问题，我们在具有代表性的指令驱动 3D 操作基准 RLBench(James 等，2020) 上进行实验。仿真实验涵盖两类任务，以全面评估 GravMAD。1) 基础任务。为评估 GravMAD 在训练中遇到的 3D 操作任务上的表现，我们从 RLBench 的 100 个语言条件任务中选取 12 个基础任务，每个任务包含 2 至 60 种指令变体，如处理不同颜色或数量的物体。每个基础任务收集 20 个示范用于训练，最终检查点在 3 个随机种子下进行 25 个回合的评估。任务详细描述见附录 B.1。2) 新颖任务。为进一步测试 GravMAD 的泛化能力，我们通过修改若干基础任务的场景配置或任务指令，创建了 8 个跨 3 个新颖类别的新任务，如图 10 所示。这些修改在指令理解、环境感知和策略泛化方面对机器人提出了显著挑战，详见附录 B.2。每个新任务均使用在 12 个基础任务上训练的最终检查点进行评估，采用 3 个随机种子，每个任务 25 个回合。所有任务均使用前视 256×256 RGB-D 相机和配备平行夹爪的 Franka Panda 机器人。此外，我们还在 10 个真实机器人任务上进一步验证了 GravMAD，详情见附录 D.6。

Baselines. We compare GravMAD against various baselines, covering both foundation model-based and imitation learning-based methods. For the foundation model-based approach, we use VoxPoser (Huang et al. 2023) as the baseline. VoxPoser leverages GPT-4 to generate code for constructing value maps, which are then used by a heuristic-based motion planner to synthesize robotic arm trajectories. We reproduce this baseline in our tasks using prompt templates from Huang et al. and our SoM-based Detector, with five camera viewpoints in RLBench. For the imitation learning-based baselines, we select: (1) 3D Diffuser Actor (Ke et al. 2024), which combines 3D scene representations with a diffusion policy for robotic manipulation tasks. To highlight instruction-following tasks, we use the enhanced language-conditioned version provided by Ke et al.; and (2) Act3D (Gervet et al. 2023), which uses a 3D feature field within a policy transformer to represent the robot's workspace. Differences between GravMAD and these baselines are detailed in Appendix A.5

基线方法。我们将 GravMAD 与多种基线方法进行比较，涵盖基于基础模型和基于模仿学习的方法。对于基于基础模型的方法，我们采用 VoxPoser(Huang 等，2023) 作为基线。VoxPoser 利用 GPT-4 生成构建价值地图的代码，随后由基于启发式的运动规划器合成机器人臂轨迹。我们在 RLBench 中使用 Huang 等人的提示模板和基于 SoM 的检测器，结合五个摄像机视角，复现了该基线在我们的任务中的表现。对于基于模仿学习的基线，我们选择:(1)3D Diffuser Actor(Ke 等，2024)，该方法结合了三维场景表示与扩散策略，用于机器人操作任务。为突出指令执行任务，我们采用 Ke 等提供的增强语言条件版本；(2)Act3D(Gervet 等，2023)，其在策略变换器中使用三维特征场表示机器人的工作空间。GravMAD 与这些基线的差异详见附录 A.5。

Training and Evaluation Details. GravMAD runs in a multi-task setting during both the training and testing phases. All models complete 600k training iterations on an NVIDIA RTX4090 GPU, with the final checkpoint selected using three random seeds for evaluation. During testing, except for the novel task "push buttons light", which must be completed in 3 time steps, all other tasks must be completed in 25 time steps; otherwise, they are considered failures. Evaluation metrics include the average success rate and rank. The success rate measures the proportion of tasks completed according to language instructions. Meanwhile, the average rank calculates the average of the rankings of each model in all tasks, reflecting the overall performance of the model in the tasks. Two settings are used to generate context $\mathcal{C}$ during testing: Manual and VLM. In the manual setting, we manually provide the Detector with the precise 3D coordinates of task-related objects in the simulation to generate accurate context. In the VLM setting, we use a Detector implemented with SoM and GPT-4o to locate task-related objects and generate context.

训练与评估细节。GravMAD 在训练和测试阶段均以多任务设置运行。所有模型均在 NVIDIA RTX4090 GPU 上完成 600k 次训练迭代，最终检查点通过三个随机种子进行评估时选取。测试期间，除新颖任务“按按钮点亮”必须在 3 个时间步内完成外，所有其他任务必须在 25 个时间步内完成，否则视为失败。评估指标包括平均成功率和排名。成功率衡量根据语言指令完成任务的比例；平均排名则计算模型在所有任务中的排名平均值，反映模型整体任务表现。测试时生成上下文 $\mathcal{C}$ 采用两种设置: 手动和 VLM。在手动设置中，我们手动向检测器提供仿真中任务相关物体的精确三维坐标以生成准确上下文；在 VLM 设置中，使用基于 SoM 和 GPT-4o 实现的检测器定位任务相关物体并生成上下文。

Models	Avg. Success ↑	Avg. Rank ↓	Close Jar	Open Drawer	Meat off Grill	Slide Block	Put in Drawer
Voxposer (Huang et al., 2023)	15.11	4.5	$12.00_{\pm 10.58}$	$10.67_{\pm 8.33}$	$45.33_{\pm 24.44}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$
Act3D (Gervet et al.. 2023	34.11	4.3	$61.33_{\pm 4.62}$	$41.33_{\pm 4.62}$	$60.0_{\pm 6.92}$	$78.67_{\pm 2.31}$	$49.33_{\pm 10.07}$
3D Diffuser Actor (Ke et al., 2024)	55.81	2.3	$66.67_{+2.31}$	$\underline{88.00}_{\pm 6.93}$	$88.00_{+4.00}$	$\underline{84.00}_{\pm 0.00}$	$\underline{94.67}_{\pm 2.31}$
GravMAD (Manual)	69.17	1.3	$\mathbf{100.00}_{\pm 0.00}$	$76.67_{\pm 4.62}$	$\mathbf{89.33}_{+2.31}$	$\mathbf{93.33}_{\pm 2.31}$	$78.67_{\pm 6.11}$
Performance gain	+13.36	-	+33.33	-13.33	+1.33	+9.33	-16.00
GravMAD (VLM)	56.72	2.1	$\mathbf{100.00}_{\pm 0.00}$	$58.67_{\pm 2.31}$	$70.67_{\pm 2.31}$	$80.00_{\pm 0.00}$	$61.33_{+9.24}$
Performance gain	+0.91	-	+33.33	-29.33	-17.33	-4.00	-33.34
Models	Push Buttons	Stack Blocks	Place Cups	Place Wine	Screw Bulb	Insert Peg	Stack Cups
Voxposer (Huang et al. 2023	$80.00_{\pm 13.86}$	$16.00_{+12.00}$	$6.67_{+8.33}$	$5.33_{+2.31}$	$4.00_{+4.00}$	$0.00_{+0.00}$	$1.33_{+2.31}$
Act3D (Gervet et al.. 2023	$66.67_{\pm 2.31}$	$0.00_{+0.00}$	$0.00_{+0.00}$	$45.33_{\pm 2.31}$	$6.67_{+2.31}$	$0.00_{+0.00}$	$0.00_{+0.00}$
3D Diffuser Actor (Ke et al., 2024)	$94.67_{\pm 2.31}$	$13.67_{\pm 2.89}$	$5.33_{\pm 6.11}$	$82.67_{\pm 2.31}$	$29.33_{\pm 2.31}$	$2.67_{\pm 4.62}$	$20.00_{\pm 0.00}$
GravMAD (Manual)	$\mathbf{98.67}_{\pm 2.31}$	$\mathbf{56.67}_{\pm 4.62}$	$5.33_{\pm 2.31}$	$77.33_{\pm 4.62}$	$\mathbf{66.67}_{\pm 6.11}$	$\mathbf{32.00}_{\pm 6.93}$	$\mathbf{57.33}_{+2.31}$
Performance gain	+4.00	+40.67	-1.34	-5.34	+37.34	+29.33	+37.33
GravMAD (VLM)	$\mathbf{97.33}_{\pm 2.31}$	$\mathbf{51.33}_{\pm 6.11}$	$5.33_{\pm 4.62}$	$33.33_{\pm 4.62}$	$\mathbf{54.67}_{\pm 6.11}$	$\mathbf{18.67}_{\pm 4.62}$	$\mathbf{49.33}_{\pm 2.31}$
Performance gain	+2.66	+35.33	-1.34	-49.34	+25.34	$+\mathbf{16.00}$	+29.33

模型	平均成功率 ↑	平均排名 ↓	关闭罐子	打开抽屉	从烤架取肉	滑动积木	放入抽屉
Voxposer (Huang 等, 2023)	15.11	4.5	$12.00_{\pm 10.58}$	$10.67_{\pm 8.33}$	$45.33_{\pm 24.44}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$
Act3D (Gervet 等, 2023)	34.11	4.3	$61.33_{\pm 4.62}$	$41.33_{\pm 4.62}$	$60.0_{\pm 6.92}$	$78.67_{\pm 2.31}$	$49.33_{\pm 10.07}$
3D Diffuser Actor (Ke 等, 2024)	55.81	2.3	$66.67_{+2.31}$	$\underline{88.00}_{\pm 6.93}$	$88.00_{+4.00}$	$\underline{84.00}_{\pm 0.00}$	$\underline{94.67}_{\pm 2.31}$
GravMAD(手动)	69.17	1.3	$\mathbf{100.00}_{\pm 0.00}$	$76.67_{\pm 4.62}$	$\mathbf{89.33}_{+2.31}$	$\mathbf{93.33}_{\pm 2.31}$	$78.67_{\pm 6.11}$
性能提升	+13.36	-	+33.33	-13.33	+1.33	+9.33	-16.00
GravMAD(视觉语言模型)	56.72	2.1	$\mathbf{100.00}_{\pm 0.00}$	$58.67_{\pm 2.31}$	$70.67_{\pm 2.31}$	$80.00_{\pm 0.00}$	$61.33_{+9.24}$
性能提升	+0.91	-	+33.33	-29.33	-17.33	-4.00	-33.34
模型	按按钮	堆叠积木	放置杯子	放置酒杯	旋紧灯泡	插入销钉	堆叠杯子
Voxposer (Huang 等, 2023)	$80.00_{\pm 13.86}$	$16.00_{+12.00}$	$6.67_{+8.33}$	$5.33_{+2.31}$	$4.00_{+4.00}$	$0.00_{+0.00}$	$1.33_{+2.31}$
Act3D (Gervet 等, 2023)	$66.67_{\pm 2.31}$	$0.00_{+0.00}$	$0.00_{+0.00}$	$45.33_{\pm 2.31}$	$6.67_{+2.31}$	$0.00_{+0.00}$	$0.00_{+0.00}$
3D Diffuser Actor (Ke 等, 2024)	$94.67_{\pm 2.31}$	$13.67_{\pm 2.89}$	$5.33_{\pm 6.11}$	$82.67_{\pm 2.31}$	$29.33_{\pm 2.31}$	$2.67_{\pm 4.62}$	$20.00_{\pm 0.00}$
GravMAD(手动)	$\mathbf{98.67}_{\pm 2.31}$	$\mathbf{56.67}_{\pm 4.62}$	$5.33_{\pm 2.31}$	$77.33_{\pm 4.62}$	$\mathbf{66.67}_{\pm 6.11}$	$\mathbf{32.00}_{\pm 6.93}$	$\mathbf{57.33}_{+2.31}$
性能提升	+4.00	+40.67	-1.34	-5.34	+37.34	+29.33	+37.33
GravMAD(视觉语言模型)	$\mathbf{97.33}_{\pm 2.31}$	$\mathbf{51.33}_{\pm 6.11}$	$5.33_{\pm 4.62}$	$33.33_{\pm 4.62}$	$\mathbf{54.67}_{\pm 6.11}$	$\mathbf{18.67}_{\pm 4.62}$	$\mathbf{49.33}_{\pm 2.31}$
性能提升	+2.66	+35.33	-1.34	-49.34	+25.34	$+\mathbf{16.00}$	+29.33

Table 2: Multi-task test results on 12 base tasks. All models are trained on 12 base tasks with 20 demonstrations each. Final checkpoints are evaluated across 3 seeds with 25 test episodes per task Performance gains are compared to the best-performing baselines.

表 2:12 个基础任务上的多任务测试结果。所有模型均在 12 个基础任务上训练，每个任务有 20 个示范。最终检查点在 3 个随机种子下评估，每个任务 25 个测试回合。性能提升相较于表现最佳的基线模型进行比较。

4.2 GENERALIZATION PERFORMANCE OF GRAVMAD TO NOVEL TASKS

4.2 GravMAD 对新任务的泛化性能

In Table 1, we present the generalization performance of models trained on 12 base tasks when tested on 8 novel tasks, along with visualized trajectories from two of these tasks. The results show that changes in task scenarios and instructions negatively impact the test performance of all pre-trained models to some extent. However, GravMAD exhibits superior generalization across all 8 novel tasks compared to the baseline models. In terms of average success rate, GravMAD outperforms VoxPoser, Act3D, and 3D Diffuser Actor by 28.63%, 45.09%, and 33.54%, respectively. VoxPoser leverages large models to achieve a certain level of performance on novel tasks, but its heuristic motion planner fails to grasp object properties and task interaction conditions, leading to poor results on tasks requiring fine manipulation, as shown in the trajectory visualizations. Similarly, 3D Diffuser Actor and Act3D struggle to transfer skills from training to novel tasks, primarily due to overfitting to training-specific tasks, which hampers generalization. In contrast, GravMAD uses VLM-generated GravMaps to guide action diffusion, enabling effective object interaction and strong performance on novel tasks. These results clearly demonstrate GravMAD's superior generalization.

在表 1 中，我们展示了在 12 个基础任务上训练的模型在 8 个新任务上的泛化性能，并附上了其中两个任务的轨迹可视化。结果表明，任务场景和指令的变化在一定程度上对所有预训练模型的测试表现产生负面影响。然而，GravMAD 在所有 8 个新任务上均表现出优越的泛化能力，优于基线模型。就平均成功率而言，GravMAD 分别比 VoxPoser、Act3D 和 3D Diffuser Actor 高出 28.63%、45.09% 和 33.54%。VoxPoser 利用大型模型在新任务上达到一定性能，但其启发式运动规划器未能准确把握物体属性和任务交互条件，导致在需要精细操作的任务中表现不佳，如轨迹可视化所示。同样，3D Diffuser Actor 和 Act3D 在技能从训练任务向新任务迁移时表现不佳，主要因过拟合训练特定任务，限制了泛化能力。相比之下，GravMAD 利用视觉语言模型 (VLM) 生成的 GravMaps 引导动作扩散，实现了有效的物体交互和新任务上的强劲表现。这些结果清晰地证明了 GravMAD 的卓越泛化能力。

4.3 TEST PERFORMANCE OF GRAVMAD ON BASE TASKS

4.3 GravMAD 在基础任务上的测试表现

Table 2 compares the performance of all models on 12 base tasks. GravMAD (Manual) outperforms Act3D and Voxposer across all tasks and exceeds the best baseline, 3D Diffuser Actor, in 9 out of 12 tasks, with an average success rate improvement of 13.36%. Despite the Detector's coarse SoM positioning affecting GravMAD (VLM)'s performance, it still outperforms Act3D and Voxposer on all tasks, with a **0.91**% higher average success rate than 3D Diffuser Actor. These results clearly show that GravMAD remains highly competitive even on previously seen tasks. As long as task-related object positions are accurate, the generated GravMap effectively reflects sub-goals and guides action diffusion, enabling precise execution by GravMAD. GravMAD (Manual) underperforms 3D Diffuser Actor in the "open drawer", "put in drawer", and "place wine" tasks due to slight deviations between the manually provided object positions and the sub-goals. In high-precision tasks, even small deviations can impact performance. For example, in the "open drawer" task, the robot needs to grasp the center of the small handle for optimal performance. After manually adjusting the sub-goal to better align with the handle, performance improved.

GravMAD (VLM) also struggles in tasks like ”Place Wine” due to inaccuracies in the object positions provided by the Detector, especially when Semantic SAM fails to provide precise locations or the camera doesn’t capture the full scene. For further analysis of failure cases, please refer to Appendix B.3

表 2 比较了所有模型在 12 个基础任务上的表现。GravMAD(手动) 在所有任务中均优于 Act3D 和 VoxPoser,并在 12 个任务中有 9 个超过了表现最佳的基线 3D Diffuser Actor,平均成功率提升 13.36%。尽管检测器粗略的 SoM 定位影响了 GravMAD(VLM) 的表现，但其仍在所有任务中优于 Act3D 和 VoxPoser，平均成功率比 3D Diffuser Actor 高 **0.91**% 。这些结果清楚表明，即使在已见任务上，GravMAD 依然具有很强的竞争力。只要任务相关的物体位置准确，生成的 GravMap 能有效反映子目标并引导动作扩散，使 GravMAD 能够精准执行。GravMAD(手动) 在“打开抽屉”、“放入抽屉”和“放置酒瓶”任务中表现不及 3D Diffuser Actor，原因是手动提供的物体位置与子目标存在细微偏差。在高精度任务中，即使是小偏差也会影响表现。例如，在“打开抽屉”任务中，机器人需抓住小把手的中心以达到最佳效果。手动调整子目标以更好地对齐把手后，表现有所提升。GravMAD(VLM) 在“放置酒瓶”等任务中也表现不佳，主要因检测器提供的物体位置不准确，尤其是当 Semantic SAM 未能提供精确位置或摄像头未捕捉完整场景时。有关失败案例的进一步分析，请参见附录 B.3。

4.4 ABLATIONS

4.4 消融实验

Extensive ablation studies are conducted to analyze the role of each key design element in GravMAD, with the results shown in Fig 4. The following findings are revealed: 1) Impact of replacing GravMaps with specific sub-goal position and openness: Replacing GravMaps with sub-goals g^{pos} and g^{open} (w/o GravMap) results in a significant performance drop. Without GravMaps, the policy

进行了大量消融研究以分析 GravMAD 中各关键设计元素的作用，结果见图 4。发现如下:1) 用具体的子目标位置和开启度替代 GravMaps 的影响: 用子目标 g^{pos} 和 g^{open} (无 GravMap) 替代 GravMaps 导致性能显著下降。没有 GravMaps，策略

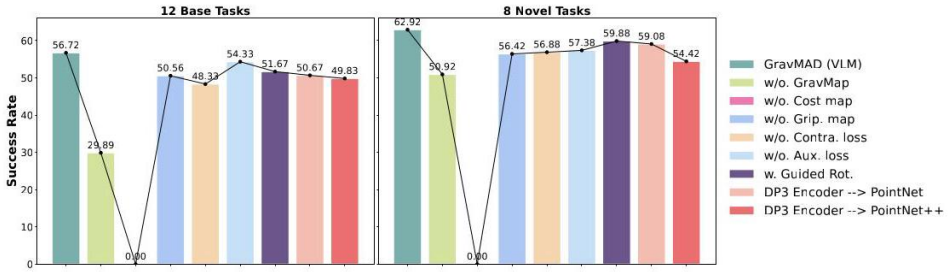


Figure 4: Ablation Studies. We evaluate the impact of key design elements by reporting the average success rates across 12 base tasks and 8 novel tasks. In the results, ” $\rightarrow$ ” denotes replacement, ”w/o” indicates ”without”, and ”w.” signifies ”with”.

图 4: 消融研究。我们通过报告 12 个基础任务和 8 个新任务的平均成功率来评估关键设计元素的影响。结果中，“$\rightarrow$ ”表示替代，“w/o”表示“无”，“w.”表示“有”。

lacks regional context, becoming overly sensitive to precise positions and unable to generalize to slight spatial variations. 2) Importance of both cost map and gripper map in GravMaps: The combination of the cost map and gripper map within GravMaps is essential for guiding the model's attention to sub-goal locations and ensuring effective gripper usage. The absence of the gripper map causes a moderate decline in performance (w/o. Grip. map). In contrast, omitting the cost map causes zero-gradient issues during training, leading to incorrect predictions and task failure. This occurs because the encoder cannot process such input. Additional experiments for this ablation, detailed in Appendix D.5, highlight the cost map's impact on performance. (w/o. Cost map) 3) Significance of contrastive learning loss and auxiliary losses: Removing the contrastive learning loss $\mathcal{L}_{\text{con}}$ results in highly similar features from the point cloud encoder, diminishing their effectiveness in action denoising and leading to a decline in model performance (w/o. Contra. loss). Similarly, the absence of auxiliary losses $\mathcal{L}_{\text{aux\_pos}}$ and $\mathcal{L}_{\text{aux\_open}}$ weakens the model's focus on sub-goals, leading to a noticeable drop in performance (w/o Aux. loss). 4) Effect of GravMap tokens on guiding rotation actions: Conditioning rotation actions with GravMap tokens in the action diffusion process results in a performance drop, likely due to the inherent nature of rotation actions, which makes them difficult to be guided explicitly through value maps (w. Guided Rot.). 5) Impact of different point cloud encoders on GravMap performance. Replacing the DP3 encoder in GravMAD with PointNet(Qi et al.,2017a)(DP3 Encoder → PointNet) or PointNet++(Qi et al.,2017b)(DP3 Encoder → PointNet++) leads to a performance decline. We suspect that lightweight encoders help prevent overfitting to training data details, enhancing GravMAD's generalization ability across different tasks or unseen data.

缺乏区域上下文，导致对精确位置过于敏感，无法对轻微的空间变化进行泛化。2)GravMaps 中代价图和夹爪图的重要性: 代价图与夹爪图的结合对于引导模型关注子目标位置并确保夹爪的有效使用至关重要。缺少夹爪图会导致性能适度下降 (无夹爪图)。相比之下，省略代价图会在训练过程中引发零梯度问题，导致错误预测和任务失败，这是因为编码器无法处理此类输入。附录 D.5 中详细的消融实验进一步强调了代价图对性能的影响 (无代价图)。3) 对比学习损失和辅助损失的重要性: 去除对比学习损失 $\mathcal{L}_{\text{con}}$ 会使点云编码器提取的特征高度相似，降低其在动作去噪中的有效性，导致模型性能下降 (无对比损失)。同样，缺少辅助损失 $\mathcal{L}_{\text{aux\_pos}}$ 和 $\mathcal{L}_{\text{aux\_open}}$ 会削弱模型对子目标的关注，导致性能明显下降 (无辅助损失)。4)GravMap 标记对引导旋转动作的影响: 在动作扩散过程中用 GravMap 标记条件化旋转动作会导致性能下降，这可能是由于旋转动作的固有特性，使其难以通过价值图明确引导 (有引导旋转)。5) 不同点云编码器对 GravMap 性能的影响。用 PointNet(Qi et al.,2017a)(DP3 编码器 → PointNet) 或 PointNet++(Qi et al.,2017b)(DP3 编码器 → PointNet++) 替换 GravMAD 中的 DP3 编码器会导致性能下降。我们推测轻量级编码器有助于防止对训练数据细节的过拟合，从而增强 GravMAD 在不同任务或未见数据上的泛化能力。

5 CONCLUSION AND DISCUSSION

5 结论与讨论

In this paper, we introduce GravMAD, a novel action diffusion framework that facilitates generalized 3D manipulation using sub-goals. GravMAD grounds language instructions into spatial subgoals within the 3D workspace through grounded spatial value maps (GravMaps). During training, these GravMaps are generated from demonstrations by Sub-goal Keyposes Discovery. In the inference phase, GravMaps are constructed by leveraging foundational models to directly predict sub-goals Consequently, GravMAD seamlessly integrates the precision of imitation learning with the strong generalization capabilities of foundational models, leading to superior performance across a variety of manipulation tasks. Extensive experiments on the RLBench benchmark and real-robot tasks show that

GravMAD achieves competitive performance on training tasks. It also generalizes well to novel tasks, demonstrating its potential for practical use across diverse 3D environments. Despite its promising results, GravMAD has some limitations. First, its effectiveness is highly dependent on prompt engineering, which can be challenging for inexperienced users. Additionally, visual-language models (VLMs) have limited detection capabilities and are sensitive to changes in camera perspective, affecting performance, and preventing optimal efficiency and accuracy. Future work will address these issues to enhance the performance of the model, expand its applicability, and validate its use on more complex and long-horizon real-robot tasks.

本文提出了 GravMAD，一种利用子目标实现泛化 3D 操作的新型动作扩散框架。GravMAD 通过基于空间的价值图 (GravMaps) 将语言指令映射到 3D 工作空间中的空间子目标。在训练阶段，这些 GravMaps 由子目标关键姿态发现方法从示范中生成。在推理阶段，GravMaps 通过利用基础模型直接预测子目标来构建。因此，GravMAD 无缝结合了模仿学习的精确性与基础模型的强大泛化能力，在多种操作任务中表现优异。在 RLBench 基准和真实机器人任务上的大量实验表明，GravMAD 在训练任务上取得了竞争性表现，并且能够很好地泛化到新任务，展示了其在多样化 3D 环境中实际应用的潜力。尽管结果令人鼓舞，GravMAD 仍存在一些局限。首先，其效果高度依赖于提示工程，这对缺乏经验的用户来说具有挑战性。此外，视觉语言模型 (VLMs) 检测能力有限且对摄像机视角变化敏感，影响性能，阻碍了效率和准确性的最优化。未来工作将针对这些问题进行改进，以提升模型性能，拓展其适用范围，并验证其在更复杂和长时域真实机器人任务中的应用。

6 ACKNOWLEDGMENTS

6 致谢

This work was supported in part by the National Natural Science Foundation of China under Grant 62276128, Grant 62192783, Grant 62206166; in part by the Jiangsu Science and Technology Major Project BG2024031; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20243051; in part by the Shanghai Sailing Program under Grant No.23YF1413000; in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX24_0263; in part by the Fundamental Research Funds for the Central Universities (14380128); in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

本工作部分由中国国家自然科学基金资助，项目编号 62276128、62192783、62206166；部分由江苏省科技重大专项 BG2024031 资助；部分由江苏省自然科学基金资助，项目编号 BK20243051；部分由上海市扬帆计划资助，项目编号 23YF1413000；部分由江苏省研究生科研与实践创新计划资助，项目编号 KYCX24_0263；部分由中央高校基本科研业务费资助 (14380128)；部分由新型软件技术与产业化协同创新中心支持。

REFERENCES

参考文献

Brenna D. Argall, Sonia Chernova, Manuela M. Veloso, and Brett Browning. A survey of robot learning from demonstration. Robotics Auton. Syst., 57(5):469-483, 2009.

Brenna D. Argall, Sonia Chernova, Manuela M. Veloso, and Brett Browning. 机器人示范学习综述。Robotics Auton. Syst., 57(5):469-483, 2009.

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Rich Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In International Conference on Learning Representations (ICLR), 2024.

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Rich Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. 利用预训练图像编辑扩散模型实现零样本机器人操作。发表于国际学习表征会议 (ICLR)，2024 年。

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 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 Sontakke, Austin Stone, Clayton Tan, Huong T. Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: robotics transformer for real-world control at scale. In Proceedings of Robotics: Science and Systems (RSS), 2023a.

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 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 Sontakke, Austin Stone, Clayton Tan, Huong T. Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, 和 Brianna Zitkovich. RT-1: 面向大规模现实世界控制的机器人变换器 (robotics transformer)。发表于《机器人科学与系统会议》(RSS)，2023a。

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In Conference on robot learning (CoRL), pp. 287-318. PMLR, 2023b.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian 等. 按我所能行事，而非我所言: 将语言基础与机器人可供性 (affordances) 相结合。发表于机器人学习会议 (CoRL)，第 287-318 页。PMLR，2023b。

Shizhe Chen, Ricardo Garcia Pinel, Cordelia Schmid, and Ivan Laptev. Polarnet: 3d point clouds for language-guided robotic manipulation. In Conference on Robot Learning (CoRL), pp. 1761-1781. PMLR, 2023a.

Shizhe Chen, Ricardo Garcia Pinel, Cordelia Schmid, 和 Ivan Laptev. Polarnet: 用于语言引导机器人操作的三维点云。发表于机器人学习会议 (CoRL)，第 1761-1781 页。PMLR，2023a。

Zixuan Chen, Wenbin Li, Yang Gao, and Yiyu Chen. Tild: Third-person imitation learning by estimating domain cognitive differences of visual demonstrations. In AAMAS, pp. 2421-2423, 2023b.

Zixuan Chen, Wenbin Li, Yang Gao, 和 Yiyu Chen. Tild: 通过估计视觉示范的领域认知差异进行第三人称模仿学习。发表于多智能体系统国际会议 (AAMAS)，第 2421-2423 页，2023b。

Zixuan Chen, Ze Ji, Jing Huo, and Yang Gao. Scar: Refining skill chaining for long-horizon robotic manipulation via dual regularization. In NeurIPS, pp. 111679-111714, 2024a.

Zixuan Chen, Ze Ji, Jing Huo, 和 Yang Gao. Scar: 通过双重正则化优化长时域机器人操作的技能链。发表于神经信息处理系统大会 (NeurIPS)，第 111679-111714 页，2024a。

Zixuan Chen, Ze Ji, Shuyang Liu, Jing Huo, Yiyu Chen, and Yang Gao. Cognizing and imitating robotic skills via a dual cognition-action architecture. In AAMAS, pp. 2204-2206, 2024b.

Zixuan Chen, Ze Ji, Shuyang Liu, Jing Huo, Yiyu Chen, 和 Yang Gao. 通过双重认知-动作架构认知与模仿机器人技能。发表于多智能体系统国际会议 (AAMAS)，第 2204-2206 页，2024b。

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), International Conference on Machine Learning (ICML), volume 202, pp. 8469-8488. PMLR, 2023.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, 和 Pete Florence. Palm-e: 一种具身多模态语言模型。发表于国际机器学习大会 (ICML)，由 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, 和 Jonathan Scarlett 编辑，卷 202，第 8469-8488 页。PMLR，2023。

Ricardo Garcia, Shizhe Chen, and Cordelia Schmid. Towards generalizable vision-language robotic manipulation: A benchmark and llm-guided 3d policy. arXiv preprint arXiv:2410.01345, 2024.

Ricardo Garcia, Shizhe Chen, 和 Cordelia Schmid. 面向通用视觉-语言机器人操作的基准与大语言模型 (LLM) 引导的三维策略。arXiv 预印本 arXiv:2410.01345，2024。

Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. In Conference on Robot Learning (CoRL), pp. 3949-3965. PMLR, 2023.

Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, 和 Katerina Fragkiadaki. Act3d: 用于多任务机器人操作的三维特征场变换器。发表于机器人学习会议 (CoRL)，第 3949-3965 页。PMLR，2023。

Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer

for 3d object manipulation. In Conference on Robot Learning (CoRL), pp. 694-710. PMLR, 2023.

Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, 和 Dieter Fox. Rvt: 用于三维物体操作的机器人视角变换器。发表于机器人学习会议 (CoRL)，第 694-710 页。PMLR，2023。

Ankit Goyal, Valts Blukis, Jie Xu, Yijie Guo, Yu-Wei Chao, and Dieter Fox. Rvt2: Learning precise manipulation from few demonstrations. Proceedings of Robotics: Science and Systems (RSS), 2024.

Ankit Goyal, Valts Blukis, Jie Xu, Yijie Guo, Yu-Wei Chao, 和 Dieter Fox. Rvt2: 从少量示范中学习精确操作。机器人学: 科学与系统会议 (RSS)，2024 年。

Jianye Hao, Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Zhaopeng Meng, Peng Liu, and Zhen Wang. Exploration in deep reinforcement learning: From single-agent to multiagent domain. IEEE Transactions on Neural Networks and Learning Systems, 35(7):8762-8782, 2024.

Jianye Hao, Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Zhaopeng Meng, Peng Liu, 和 Zhen Wang. 深度强化学习中的探索: 从单智能体到多智能体领域。IEEE 神经网络与学习系统汇刊，35(7):8762-8782, 2024 年。

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In NeurIPS, 2020.

Jonathan Ho, Ajay Jain, 和 Pieter Abbeel. 去噪扩散概率模型。NeurIPS 会议，2020 年。

Yafei Hu, Quanting Xie, Vidhi Jain, Jonathan Francis, Jay Patrikar, Nikhil Keetha, Seungchan Kim, Yaqi Xie, Tianyi Zhang, Zhibo Zhao, et al. Toward general-purpose robots via foundation models: A survey and meta-analysis. arXiv preprint arXiv:2312.08782, 2023a.

Yafei Hu, Quanting Xie, Vidhi Jain, Jonathan Francis, Jay Patrikar, Nikhil Keetha, Seungchan Kim, Yaqi Xie, Tianyi Zhang, Zhibo Zhao 等. 通过基础模型实现通用机器人: 综述与元分析。arXiv 预印本 arXiv:2312.08782, 2023a。

Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. arXiv preprint arXiv:2311.17842, 2023b.

Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, 和 Yang Gao. 三思而后行: 揭示 GPT-4V 在机器人视觉语言规划中的强大能力。arXiv 预印本 arXiv:2311.17842, 2023b。

Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models. arXiv preprint arXiv:2403.08248, 2024.

Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, 和 Yang Gao. COPA: 通过部件空间约束与基础模型实现通用机器人操作。arXiv 预印本 arXiv:2403.08248, 2024 年。

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In International Conference on Machine Learning (ICML), pp. 9118-9147. PMLR, 2022.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, 和 Igor Mordatch. 语言模型作为零样本规划器: 为具身智能体提取可执行知识。国际机器学习大会 (ICML)，第 9118-9147 页。PMLR, 2022 年。

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. In Conference on Robot Learning (CoRL), pp. 540-562. PMLR, 2023.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, 和 Li Fei-Fei. Voxposer: 用于机器人操作的可组合三维价值图与语言模型。机器人学习会议 (CoRL)，第 540-562 页。PMLR, 2023 年。

Stephen James and Andrew J Davison. Q-attention: Enabling efficient learning for vision-based robotic manipulation. IEEE Robotics and Automation Letters, 7(2):1612-1619, 2022.

Stephen James 和 Andrew J Davison. Q-attention: 实现基于视觉的机器人操作的高效学习。IEEE 机器人与自动化快报，7(2):1612-1619, 2022 年。

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. IEEE Robotics and Automation Letters, 5(2):3019- 3026, 2020.

Stephen James, Zicong Ma, David Rovick Arrojo, 和 Andrew J Davison. RLBench: 机器人学习基准与学习环境。IEEE 机器人与自动化快报，5(2):3019-3026, 2020 年。

Stephen James, Kentaro Wada, Tristan Laidlow, and Andrew J Davison. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13739-13748, 2022.

Stephen James, Kentaro Wada, Tristan Laidlow, 和 Andrew J Davison. 粗到细的 Q-attention: 通过离散化实现视觉机器人操作的高效学习。IEEE/CVF 计算机视觉与模式识别会议 (CVPR) 论文集，第 13739-13748 页，2022 年。

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In Conference on Robot Learning (CoRL), pp. 991-1002. PMLR, 2022.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, 和 Chelsea Finn. BC-Z: 通过机器人模仿学习实现零样本任务泛化。机器人学习会议 (CoRL)，第 991-1002 页。PMLR, 2022 年。

Xuhui Kang, Wenqian Ye, and Yen-Ling Kuo. Imagined subgoals for hierarchical goal-conditioned policies. In CoRL 2023 Workshop on Learning Effective Abstractions for Planning (LEAP), 2023.

Xuhui Kang, Wenqian Ye, 和 Yen-Ling Kuo. 为分层目标条件策略设计的想象子目标。CoRL 2023 学习有效规划抽象研讨会 (LEAP)，2023 年。

Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. In Conference on Robot Learning (CoRL). PMLR, 2024.

Tsung-Wei Ke, Nikolaos Gkanatsios, 和 Katerina Fragkiadaki. 3D 扩散演员: 基于三维场景表示的策略扩散。机器人学习会议 (CoRL)。PMLR, 2024 年。

Sebastian Klemm, Jan Oberländer, Andreas Hermann, Arne Roennau, Thomas Schamm, J Marius Zollner, and Rüdiger Dillmann. Rrt*-connect: Faster, asymptotically optimal motion planning. In 2015 IEEE international conference on robotics and biomimetics (ROBIO), pp. 1670-1677. IEEE, 2015.

Sebastian Klemm, Jan Oberländer, Andreas Hermann, Arne Roennau, Thomas Schamm, J Marius Zollner, 和 Rüdiger Dillmann. RRT*-Connect: 更快的渐近最优运动规划。2015 年 IEEE 机器人与仿生国际会议 (ROBIO)，第 1670-1677 页。IEEE, 2015 年。

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767, 2023.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, 和 Jianfeng Gao. Semantic-SAM: 任意粒度的分割与识别。arXiv 预印本 arXiv:2307.04767, 2023 年。

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models as effective robot imitators. In International Conference on Learning Representations (ICLR), 2024.

李星航，刘明焕，张汉博，于存军，徐杰，吴洪涛，张志廉，景雅，张伟南，刘华平，李航，孔涛。视觉-语言基础模型作为高效的机器人模仿者。发表于国际学习表征会议 (ICLR)，2024 年。

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 IEEE International Conference on Robotics and Automation (ICRA), pp. 9493-9500. IEEE, 2023.

Jacky Liang，黄文龙，夏飞，徐鹏，Karol Hausman，Brian Ichter，Pete Florence，Andy Zeng。代码即策略: 用于具身控制的语言模型程序。发表于 IEEE 国际机器人与自动化会议 (ICRA)，第 9493-9500 页。IEEE，2023 年。

Xiao Ma, Sumit Patidar, Iain Haughton, and Stephen James. Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18081-18090, 2024.

马晓，Sumit Patidar，Iain Haughton，Stephen James。分层扩散策略用于运动学感知的多任务机器人操作。发表于 IEEE/CVF 计算机视觉与模式识别会议 (CVPR)，第 18081-18090 页，2024 年。

OpenAI. GPT-4 technical report. CoRR, abs/2303.08774, 2023.

OpenAI。GPT-4 技术报告。CoRR，abs/2303.08774，2023 年。

Abhishek Padalkar, Ajinkya Jain, Alex Bewley, Alexander Herzog, Alex Irpan, Alexander Khazatsky, Anant Raj, Anikait Singh, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim,

Bernhard Schölkopf, Brian Ichter, Cewu Lu, Charles Xu, Chelsea Finn, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Chuer Pan, Chuyuan Fu, Coline Devin, Danny Driess, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Federico Ceola, Fei Xia, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Giulio Schiavi, Gregory Kahn, Hao Su, Haoshu Fang, Haochen Shi, Heni Ben Amor, Henrik I. Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Igor Mordatch, Ilija Radosavovic, and et al. Open x-embodiment: Robotic learning datasets and RT-X models. In IEEE International Conference on Robotics and Automation (ICRA), pp. 6892-6903. IEEE, 2024.

Abhishek Padalkar，Ajinkya Jain，Alex Bewley，Alexander Herzog，Alex Irpan，Alexander Khazatsky，Anant Raj，Anikait Singh，Anthony Brohan，Antonin Raffin，Ayzaan Wahid，Ben Burgess-Limerick，Beomjoon Kim，Bernhard Schölkopf，Brian Ichter，卢策武，Charles Xu，Chelsea Finn，Chenfeng Xu，程驰，黄晨光，Christine Chan，潘楚尔，傅楚源，Coline Devin，Danny Driess，Deepak Pathak，Dhruv Shah，Dieter Büchler，Dmitry Kalashnikov，Dorsa Sadigh，Edward Johns，Federico Ceola，夏飞，Freek Stulp，周高岳，Gaurav S. Sukhatme，Gautam Salhotra，严戈，Giulio Schiavi，Gregory Kahn，苏浩，方浩书，石浩辰，Heni Ben Amor，Henrik I. Christensen，古田裕树，Homer Walke，方洪杰，Igor Mordatch，Ilija Radosavovic，等。开放 x-具身: 机器人学习数据集与 RT-X 模型。发表于 IEEE 国际机器人与自动化会议 (ICRA)，第 6892-6903 页。IEEE，2024 年。

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI conference on artificial intelligence (AAAI), 2018.

Ethan Perez，Florian Strub，Harm De Vries，Vincent Dumoulin，Aaron Courville。FILM: 具有通用条件层的视觉推理。发表于 AAAI 人工智能会议，2018 年。

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 652-660, 2017a.

Charles R Qi，苏浩，莫凯春，Leonidas J Guibas。PointNet: 基于点集的 3D 分类与分割深度学习方法。发表于 IEEE/CVF 计算机视觉与模式识别会议 (CVPR)，第 652-660 页，2017 年 a。

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In NeurIPS, pp. 5099-5108, 2017b.

Charles Ruizhongtai Qi，李毅，苏浩，Leonidas J. Guibas。PointNet++: 度量空间中点集的深层次特征学习。发表于 NeurIPS，第 5099-5108 页，2017 年 b。

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (ICML), pp. 8748-8763. PMLR, 2021.

Alec Radford，Jong Wook Kim，Chris Hallacy，Aditya Ramesh，Gabriel Goh，Sandhini Agarwal，Girish Sastry，Amanda Askell，Pamela Mishkin，Jack Clark，等。基于自然语言监督的可迁移视觉模型学习。发表于国际机器学习会议 (ICML)，第 8748-8763 页。PMLR，2021 年。

SP Sharan, Ruihan Zhao, Zhangyang Wang, Sandeep P Chinchali, et al. Plan diffuser: Grounding llm plan-

ners with diffusion models for robotic manipulation. In Bridging the Gap between Cognitive Science and Robot Learning in the Real World: Progresses and New Directions, 2024.

SP Sharan，赵瑞涵，王章扬，Sandeep P Chinchali，等。Plan Diffuser: 利用扩散模型为机器人操作赋能大型语言模型规划器。发表于认知科学与机器人学习现实世界桥梁: 进展与新方向，2024 年。

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In Conference on Robot Learning (CoRL), pp. 785-799. PMLR, 2023.

Mohit Shridhar，Lucas Manuelli，Dieter Fox。Perceiver-Actor: 用于机器人操作的多任务变换器。发表于机器人学习会议 (CoRL)，第 785-799 页。PMLR，2023 年。

Mohit Shridhar, Yat Long Lo, and Stephen James. Generative image as action models. In 8th Annual Conference on Robot Learning (CoRL), 2024.

Mohit Shridhar，Yat Long Lo，Stephen James。生成式图像作为动作模型。发表于第八届机器人学习年会 (CoRL)，2024 年。

Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata V2: A dataset for robot learning at scale. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), Conference on Robot Learning (CoRL), volume 229, pp. 1723-1736. PMLR, 2023.

Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, 和 Sergey Levine。Bridgedata V2: 大规模机器人学习数据集。收录于 Jie Tan, Marc Toussaint, 和 Kourosh Darvish(编)，机器人学习会议 (CoRL)，第 229 卷，第 1723-1736 页。PMLR，2023 年。

Zhou Xian, Nikolaos Gkanatsios, Theophile Gervet, Tsung-Wei Ke, and Katerina Fragkiadaki. Chaineddiffuser: Unifying trajectory diffusion and keypose prediction for robotic manipulation. In Conference on Robot Learning (CoRL), pp. 2323-2339. PMLR, 2023.

Zhou Xian, Nikolaos Gkanatsios, Theophile Gervet, Tsung-Wei Ke, 和 Katerina Fragkiadaki。Chaineddiffuser: 统一轨迹扩散与关键姿态预测的机器人操作方法。收录于机器人学习会议 (CoRL)，第 2323-2339 页。PMLR，2023 年。

Annie Xie, Lisa Lee, Ted Xiao, and Chelsea Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation. In IEEE International Conference on Robotics and Automation (ICRA), pp. 3153-3160. IEEE, 2024.

Annie Xie, Lisa Lee, Ted Xiao, 和 Chelsea Finn。分解视觉机器人操作模仿学习中的泛化差距。收录于 IEEE 国际机器人与自动化会议 (ICRA)，第 3153-3160 页。IEEE，2024 年。

Ge Yan, Yueh-Hua Wu, and Xiaolong Wang. Dnact: Diffusion guided multi-task 3d policy learning. arXiv preprint arXiv:2403.04115, 2024.

Ge Yan, Yueh-Hua Wu, 和 Xiaolong Wang。Dnact: 基于扩散引导的多任务三维策略学习。arXiv 预印本 arXiv:2403.04115，2024 年。

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. CoRR, abs/2310.11441, 2023a.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, 和 Jianfeng Gao。集合标记提示释放 GPT-4V 中卓越的视觉定位能力。CoRR，abs/2310.11441，2023a。

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). CoRR, abs/2309.17421, 2023b.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, 和 Lijuan Wang。Lmms 的曙光: 基于 gpt-4v(ision) 的初步探索。CoRR，abs/2309.17421，2023b。

Yida Yin, Zekai Wang, Yuvan Sharma, Dantong Niu, Trevor Darrell, and Roei Herzig. In-context learning enables robot action prediction in llms. arXiv preprint arXiv:2410.12782, 2024.

Yida Yin, Zekai Wang, Yuvan Sharma, Dantong Niu, Trevor Darrell, 和 Roei Herzig。上下文学习使大型语言模型 (LLMs) 能够预测机器人动作。arXiv 预印本 arXiv:2410.12782，2024 年。

Wentao Yuan, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. M2t2: Multi-task masked transformer for object-centric pick and place. In Conference on Robot Learning (CoRL), pp. 3619-3630. PMLR, 2023.

Wentao Yuan, Adithyavairavan Murali, Arsalan Mousavian, 和 Dieter Fox。M2t2: 面向物体中心的多任务掩码变换器用于抓取与放置。收录于机器人学习会议 (CoRL)，第 3619-3630 页。PMLR，2023 年。

Yanjie Ze, Ge Yan, Yueh-Hua Wu, Annabella Macaluso, Yuying Ge, Jianglong Ye, Nicklas Hansen, Li Erran Li, and Xiaolong Wang. Gnfactor: Multi-task real robot learning with generalizable neural feature fields. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), Conference on Robot Learning (CoRL), volume 229, pp. 284-301. PMLR, 2023.

Yanjie Ze, Ge Yan, Yueh-Hua Wu, Annabella Macaluso, Yuying Ge, Jianglong Ye, Nicklas Hansen, Li Erran Li, 和 Xiaolong Wang。Gnfactor: 具备泛化能力的多任务真实机器人学习神经特征场。收录于 Jie Tan, Marc Toussaint, 和 Kourosh Darvish(编)，机器人学习会议 (CoRL)，第 229 卷，第 284-301 页。PMLR，2023 年。

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In Proceedings of Robotics: Science and Systems (RSS), 2024.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, 和 Huazhe Xu。三维扩散策略: 通过简单三维表示实现泛化视觉运动策略学习。收录于机器人科学与系统会议 (RSS)，2024 年。

Junjie Zhang, Chenjia Bai, Haoran He, Zhigang Wang, Bin Zhao, Xiu Li, and Xuelong Li. Sam-e: Leveraging visual foundation model with sequence imitation for embodied manipulation. In International Conference on

Machine Learning (ICML), 2024.

Junjie Zhang, Chenjia Bai, Haoran He, Zhigang Wang, Bin Zhao, Xiu Li, 和 Xuelong Li。Sam-e: 结合视觉基础模型与序列模仿的具身操作方法。收录于国际机器学习大会 (ICML)，2024 年。

Tong Zhang, Yingdong Hu, Hanchen Cui, Hang Zhao, and Yang Gao. A universal semantic-geometric representation for robotic manipulation. In Conference on Robot Learning (CoRL), pp. 3342-3363. PMLR, 2023.

Tong Zhang, Yingdong Hu, Hanchen Cui, Hang Zhao, 和 Yang Gao。机器人操作的通用语义-几何表示。收录于机器人学习会议 (CoRL)，第 3342-3363 页。PMLR，2023 年。

Hongkuan Zhou, Xiangtong Yao, Yuan Meng, Siming Sun, Zhenshan BIng, Kai Huang, and Alois Knoll. Language-conditioned learning for robotic manipulation: A survey. arXiv preprint arXiv:2312.10807, 2023.

Hongkuan Zhou, Xiangtong Yao, Yuan Meng, Siming Sun, Zhenshan Bing, Kai Huang, 和 Alois Knoll。基于语言条件的机器人操作学习综述。arXiv 预印本 arXiv:2312.10807，2023 年。

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong T. Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: vision-language-action models transfer web knowledge to robotic control. In Conference on Robot Learning (CoRL), volume 229, pp. 2165-2183. PMLR, 2023.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong T. Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: 视觉-语言-动作模型将网络知识转移到机器人控制中。发表于机器人学习会议 (CoRL)，第 229 卷，第 2165-2183 页。PMLR，2023 年。

A ADDITIONAL IMPLEMENTATION DETAILS

A 附加实现细节

A.1 GRAVMAP GENERATION PROCESS

A.1 GravMap 生成过程

Algorithm 1: GravMap Generation Process

算法 1:GravMap 生成过程

Input: End-effector position g^{pos} , initial gripper openness $g^{\text{open}}_{\text{init}}$, target gripper openness g^{open} ,

输入: 末端执行器位置 g^{pos} ， 初始夹爪开合度 $g^{\text{open}}_{\text{init}}$ ， 目标夹爪开合度 g^{open} ，

map size S_m , offset range β_o , radius β_r , downsample ratio β_d , number of sampled points

地图大小 S_m ， 偏移范围 β_o ， 半径 β_r ， 下采样比例 β_d ， 采样点数量

N_p , inference mode flag inference

N_p ， 推理模式标志 inference

Output: GravMap m

输出:GravMap m

begin

开始

// Initialize cost map m_c , gripper map m_g , and avoidance map m_a with size S_m^3

// 初始化代价地图 m_c ， 夹爪地图 m_g ， 避障地图 m_a ， 大小为 S_m^3

Initialize m_c, m_q , and m_a with shape $S_m \times S_m \times S_m$, setting $m_c\,(u, v, w) = 1$,

初始化 m_c, m_q 和 m_a ， 形状为 $S_m \times S_m \times S_m$ ， 设置为 $m_c\,(u, v, w) = 1$,

$m_g\,(u, v, w) = g^{\text{open}}_{\text{init}}$, and $m_a\,(u, v, w) = 0$ for all voxels(u, v, w);

为所有体素 (u, v, w) 初始化 $m_g\,(u, v, w) = g^{\text{open}}_{\text{init}}$ 和 $m_a\,(u, v, w) = 0$;

// Convert g^{pos} from world coordinates(x, y, z)to voxel coordinates(i, j, k)

// 将 g^{pos} 从世界坐标 (x, y, z) 转换为体素坐标 (i, j, k)

Extract $(x, y, z) \leftarrow g^{\text{pos}}$; Convert(x, y, z)to voxel coordinates(i, j, k);

提取 $(x, y, z) \leftarrow g^{\text{pos}}$ ； 将 (x, y, z) 转换为体素坐标 (i, j, k)；

// Determine voxel coordinates (i', j', k') based on mode

// 根据模式确定体素坐标 (i', j', k')

if not inference then

如果不是推理模式，则

// Apply random offsets $\delta_i, \delta_j, \delta_k$ to(i, j, k)for data augmentation

// 对 (i, j, k) 应用随机偏移 $\delta_i, \delta_j, \delta_k$ 以进行数据增强

Sample $\delta_i, \delta_j, \delta_k \sim$ Uniform $(-\beta_o, \beta_o)$;

采样 $\delta_i, \delta_j, \delta_k \sim$ Uniform $(-\beta_o, \beta_o)$;

Update voxel coordinates: $(i', j', k') = (i + \delta_i, j + \delta_j, k + \delta_k)$;

更新体素坐标: $(i', j', k') = (i + \delta_i, j + \delta_j, k + \delta_k)$;

else

else

// Use original voxel coordinates in inference mode

// 在推理模式下使用原始体素坐标

$(i', j', k') = (i, j, k)$;

// Compute Euclidean distance from (i', j', k') for all(u, v, w)

// 计算所有 (u, v, w) 点到 (i', j', k') 的欧氏距离

For each voxel(u, v, w), compute $D(u, v, w) = \sqrt{(u - i')^2 + (v - j')^2 + (w - k')^2}$

对每个体素 (u, v, w)， 计算 $D(u, v, w) = \sqrt{(u - i')^2 + (v - j')^2 + (w - k')^2}$

// Construct avoidance map m_a

// 构建避让地图 m_a

Set $m_a(u, v, w) = 1$ for all occupied voxels in the scene;

为场景中所有占用体素设置 $m_a(u, v, w) = 1$;

// Update m_a by excluding voxels near the target (i', j', k')

// 通过排除靠近目标 (i', j', k') 的体素来更新 m_a

Set $m_a(u, v, w) = 0$ for voxels where $D(u, v, w) < 0.15 \cdot S_m$;

为满足 $D(u, v, w) < 0.15 \cdot S_m$ 的体素设置 $m_a(u, v, w) = 0$;

Smooth m_a with Gaussian filter $(\sigma = 10)$;

用高斯滤波器 $(\sigma = 10)$ 平滑 m_a ;

// Compute and normalize the cost map m_c

// 计算并归一化代价地图 m_c

Set $m_c(u, v, w) = \frac{D(u,v,w)}{\max D}$ for all voxels(u, v, w);

为所有体素 (u, v, w) 设置 $m_c(u, v, w) = \frac{D(u,v,w)}{\max D}$;

// Combine m_c and m_a into the final cost map

// 将 m_c 和 m_a 合并为最终代价图

Update $m_c(u, v, w) = 2 \cdot m_c(u, v, w) + m_a(u, v, w)$ for all voxels(u, v, w);

更新所有体素 (u, v, w) 的 $m_c(u, v, w) = 2 \cdot m_c(u, v, w) + m_a(u, v, w)$;

Normalize m_c to the range $[0, 1]$;

将 m_c 归一化到范围 $[0, 1]$;

// Set m_q within radius β_r of (i', j', k')

// 在半径 β_r 内设置 m_q ， 以 (i', j', k') 为中心

Set $m_g(u, v, w) = g^{\text{open}}$ for voxels where $D(u, v, w) \leq \beta_r$;

为满足 $D(u, v, w) \leq \beta_r$ 条件的体素设置 $m_g(u, v, w) = g^{\text{open}}$;

// Downsample both m_c and m_g

// 对 m_c 和 m_g 进行下采样

Downsample m_c and m_g by β_d ;

将 m_c 和 m_g 按 β_d 比例下采样;

Select N_p points $\{v_p\}$ from the downsampled m_c and m_g using Farthest Point Sampling;

使用最远点采样从下采样后的 m_c 和 m_g 中选择 N_p 个点 $\{v_p\}$;

// Construct GravMap m using sampled points

// 使用采样点构建 GravMap m

Form $m = \{(v_p, m_c(v_p), m_g(v_p))\}_{p=1}^{N_p}$;

形成 $m = \{(v_p, m_c(v_p), m_g(v_p))\}_{p=1}^{N_p}$;

return m ;

返回 m ;

A.2 HEURISTICS FOR SUB-GOAL KEYPOSE DISCOVERY

A.2 子目标关键姿态发现的启发式方法

Building on keypose discovery (James & Davison, 2022), we propose the Sub-goal Keypose Discovery method to identify sub-goal keyposes from demonstrations, focusing on changes in the gripper's state and touch forces. This is particularly relevant for object manipulation tasks, where the robot's interactions with objects can be segmented into discrete sub-goals.

基于关键姿态发现 (James & Davison, 2022)，我们提出了子目标关键姿态发现方法，用于从示范中识别子目标关键姿态，重点关注夹持器状态和触觉力的变化。这对于物体操作任务尤为重要，因为机器人与物体的交互可以分割为离散的子目标。

The implementation of the Sub-goal Keypose Discovery algorithm starts with a set of pre-computed keyposes, which are frames selected from the demonstration sequence through an initial keypose discovery process. We introduce two functions: touch_change, shown in Algorithm 2, and gripper_change, shown in Algorithm 3, to evaluate whether a keypose qualifies as a sub-goal.

子目标关键姿态发现算法的实现始于一组预先计算的关键姿态, 这些关键姿态是通过初始关键姿态发现过程从示范序列中选取的帧。我们引入了两个函数: 触觉变化 (touch_change)，见算法 2，以及夹持器状态变化 (gripper_change)，见算法 3，用于评估关键姿态是否符合子目标条件。

The first function checks for significant changes in the gripper's touch forces, while the second evaluates changes in the gripper's open/close state. The pseudocode in Algorithm 4 outlines the heuristic steps for identifying sub-goal keyposes.

第一个函数检测夹持器触觉力的显著变化，第二个函数评估夹持器开闭状态的变化。算法 4 中的伪代码概述了识别子目标关键姿态的启发式步骤。

One current limitation of the Sub-goal Keypose Discovery method is its inability to effectively handle tasks involving tool use, which we plan to address in future research.

子目标关键姿势发现方法目前的一个限制是无法有效处理涉及工具使用的任务，我们计划在未来的研究中解决这一问题。

Algorithm 2: touch_change Function

算法 2:touch_change 函数

Input: Demonstration sequence demo, Keypose index k , Threshold touch_threshold, Tolerance

输入: 演示序列 demo，关键姿势索引 k ，阈值 touch_threshold，容差

delta

delta

Output: Boolean indicating significant touch force change

输出: 布尔值，指示触摸力是否发生显著变化

begin

开始

Set start to max $(0, k$ - touch_threshold) ;

将 start 设为 max $(0, k$ - touch_threshold)；

for each index j from start to k-1 do

对每个索引 j 从 start 到 k-1 执行

if Touch forces at j differ from Touch forces at k within tolerance delta then

如果 j 处的触摸力与 k 处的触摸力在容差 delta 内不同，则

return True;

返回 True；

return False;

返回 False;

Algorithm 3: gripper_change Function

算法 3:gripper_change 函数

Input: Demonstration sequence demo, Keypose index k , Threshold gripper_threshold

输入: 演示序列 demo，关键姿势索引 k ，阈值 gripper_threshold

Output: Boolean indicating gripper state change

输出: 布尔值，指示夹持器状态是否变化

begin

开始

Set start to max $(0, k$ - gripper_threshold) ;

将 start 设为 max $(0, k$ - gripper_threshold) ;

for each index j from start to $k-1$ do

对于从 start 到 $k-1$ 的每个索引 j 执行

if Gripper state at j differs from Gripper state at k then

如果索引 j 处的夹爪状态与索引 k 处的夹爪状态不同，则

return True;

返回 True;

return False;

返回 False;

Algorithm 4: Heuristics for Sub-goal Keypose Discovery

算法 4: 子目标关键姿态发现的启发式方法

Input: Demonstration sequence demo, Task type task_str, Threshold parameters

输入: 示范序列 demo，任务类型 task_str，阈值参数

touch_threshold, gripper_threshold, delta

touch_threshold，gripper_threshold，delta

Output: List of sub-goal keyposes sub_goal_keyposes

输出: 子目标关键姿态列表 sub_goal_keyposes

begin

开始

Initialize sub_goal_keyposes as an empty list;

初始化 sub_goal_keyposes 为空列表；

Identify keyposes from demo using keypose discovery method;

使用关键姿态发现方法从 demo 中识别关键姿态；

for each keypose k in keyposes do

对于关键姿态列表中的每个关键姿态 k 执行

if task_str is a task involving touch without grasping then

如果 task_str 是涉及触摸但不抓取的任务，则

if touch_change(demo, k , touch_threshold, delta) then

如果 touch_change(demo, k , touch_threshold, delta) 为真，则

Append k to sub_goal_keyposes;

将 k 添加到 sub_goal_keyposes；

else

否则

if gripper_change(demo, k , gripper_threshold) or touch_change(demo, k ,

如果 gripper_change(demo, k , gripper_threshold) 或 touch_change(demo, k ,

touch_threshold, delta) then

touch_threshold, delta) 则

Append k to sub_goal_keyposes;

将 k 添加到 sub_goal_keyposes;

Append the last keypose to sub_goal_keyposes;

将最后一个关键姿态添加到 sub_goal_keyposes;

return sub_goal_keyposes;

返回 sub_goal_keyposes;

Algorithm 5: GravMap Generation

算法 5:GravMap 生成

Input: Instruction ℓ , Observed RGB Image $\mathcal{O}$

输入: 指令 ℓ ， 观测到的 RGB 图像 $\mathcal{O}$

Prompt : Prompt for Detector $\mathcal{P}_{\text{det}}$, Prompt for Planner $\mathcal{P}_{\text{plan}}$, Prompt for Composer $\mathcal{P}_{\text{com}}$ ，

提示: 检测器提示 $\mathcal{P}_{\text{det}}$ ， 规划器提示 $\mathcal{P}_{\text{plan}}$ ， 组合器提示 $\mathcal{P}_{\text{com}}$ ，

Few-shot task specified prompt $\mathcal{P}_{\text{task}} = \left\{\mathcal{P}'_{\text{det}}, \mathcal{P}'_{\text{plan}}, \mathcal{P}'_{\text{com}}\right\}$, Cost Map Prompt $\mathcal{P}_{\text{cost}}$ ，

少样本任务指定提示 $\mathcal{P}_{\text{task}} = \left\{\mathcal{P}'_{\text{det}}, \mathcal{P}'_{\text{plan}}, \mathcal{P}'_{\text{com}}\right\}$ ， 代价地图提示 $\mathcal{P}_{\text{cost}}$ ，

Gripper Map Prompt $\mathcal{P}_{\text{gripper}}$

夹爪地图提示 $\mathcal{P}_{\text{gripper}}$

Output: GravMap m

输出:GravMap m

begin

开始

$\mathcal{O}' \leftarrow$ Semantic-SAM(O); // Label objects with numerical tags

$\mathcal{O}' \leftarrow$ Semantic-SAM(O); // 用数字标签标注对象

$\mathcal{C} \leftarrow$ Detector $(\ell, \mathcal{O}', \mathcal{P}_{\text{det}}, \mathcal{P}'_{\text{det}})$; // Select relevant objects and get

$\mathcal{C} \leftarrow$ Detector $(\ell, \mathcal{O}', \mathcal{P}_{\text{det}}, \mathcal{P}'_{\text{det}})$; // 选择相关对象并获取

corresponding 3D positions as context

对应的三维位置作为上下文

$ST \leftarrow$ Planner $\left(\ell, \mathcal{C}, \mathcal{P}_{\text{plan}}, \mathcal{P}'_{\text{plan}}\right)$; // Infer sub-tasks $ST = (st_1, st_2, \ldots, st_i)$

$ST \leftarrow$ 规划器 $\left(\ell, \mathcal{C}, \mathcal{P}_{\text{plan}}, \mathcal{P}'_{\text{plan}}\right)$; // 推断子任务 $ST = (st_1, st_2, \ldots, st_i)$

Function calls, parameters $\leftarrow$ Composer $(ST, \mathcal{C}, \mathcal{P}_{\text{com}}, \mathcal{P}'_{\text{com}})$; // Generate API

函数调用，参数 $\leftarrow$ 组合器 $(ST, \mathcal{C}, \mathcal{P}_{\text{com}}, \mathcal{P}'_{\text{com}})$; // 生成 API

calls and their parameters for generating g^{pos} and g^{open}

调用及其参数用于生成 g^{pos} 和 g^{open}

$g^{\text{pos}} \leftarrow$ get_cost_map(Function calls, parameters, $\mathcal{P}_{\text{cost}}$);

$g^{\text{pos}} \leftarrow$ get_cost_map(函数调用，参数，$\mathcal{P}_{\text{cost}}$);

$g^{\text{open}} \leftarrow$ get_gripper_map(Function calls, parameters, $\mathcal{P}_{\text{gripper}}$);

$g^{\text{open}} \leftarrow$ get_gripper_map(函数调用，参数，$\mathcal{P}_{\text{gripper}}$);

$m \leftarrow$ GravMap generator $(\text{cat}(g^{\text{pos}}, g^{\text{open}}))$;

$m \leftarrow$ GravMap 生成器 $(\text{cat}(g^{\text{pos}}, g^{\text{open}}))$;

return m ;

返回 m ;

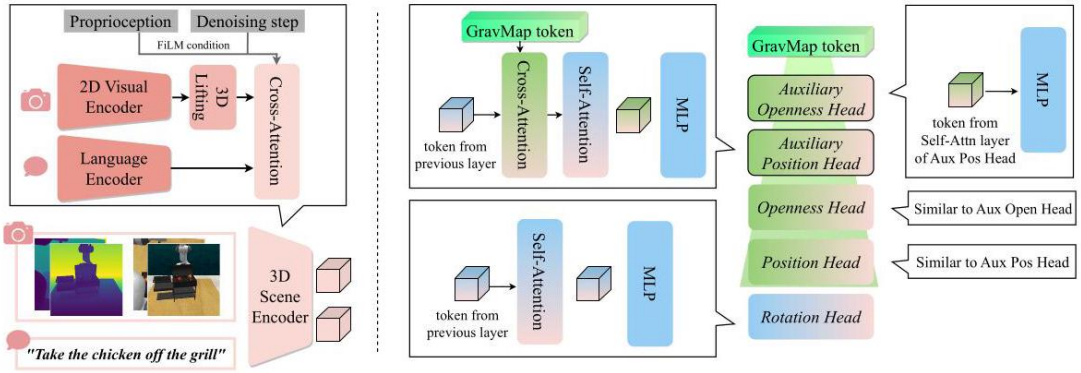


Figure 5: Detailed description of the modules in GravMAD, including the 3D Scene Encoder and the prediction heads

图 5:GravMAD 模块的详细描述，包括 3D 场景编码器和预测头

A.3 DETAILS OF GRAVMAP SYNTHESIS

A.3 GravMap 合成细节

A.3.1 TRAINING PHASE

A.3.1 训练阶段

To facilitate GravMap synthesis, we assign a goal action to each keypose by linking it to the action performed at the nearest future sub-goal. This association enables us to determine the relevant cost and gripper state for different regions of the GravMap. In the first map, $m_c \in \mathbb{R}^{w \times h \times d}$, the cost is lower near the positions of the robotic end-effector at these sub-goal keyposes and higher as the distance increases. In the second map, $m_o \in \mathbb{R}^{w \times h \times d}$, areas near the end-effector’s position at the sub-goal keyposes reflect the gripper state at the sub-goal, while other areas reflect the gripper state at the current frame.

为了便于 GravMap 的合成，我们通过将每个关键姿势与最近未来子目标执行的动作关联，赋予其一个目标动作。此关联使我们能够确定 GravMap 不同区域的相关代价和夹爪状态。在第一个地图 $m_c \in \mathbb{R}^{w \times h \times d}$ 中，靠近这些子目标关键姿势处机器人末端执行器位置的代价较低，距离越远代价越高。在第二个地图 $m_o \in \mathbb{R}^{w \times h \times d}$ 中，靠近子目标关键姿势末端执行器位置的区域反映子目标的夹爪状态，而其他区域则反映当前帧的夹爪状态。

A.3.2 INFERENCE PHASE

A.3.2 推理阶段

In this section, we introduce the complete pipeline for GravMap generation, as outlined in Algorithm 5 This includes Algorithm 1, which details the process of generating a GravMap from a language instruction ℓ and an

observed RGB image $\mathcal{O}$. The GravMap generation pipeline integrates VLMs to interpret instructions, ground them in the visual context, and translate them into coarse 3D voxel representations, i.e., the GravMap.

本节介绍了 GravMap 生成的完整流程，如算法 5 所示。包括算法 1，详细说明了如何从语言指令 ℓ 和观察到的 RGB 图像 $\mathcal{O}$ 生成 GravMap。GravMap 生成流程整合了视觉语言模型 (VLMs) 以理解指令，将其与视觉上下文结合，并转换为粗略的 3D 体素表示，即 GravMap。

Our pipeline consists of the following three components:

我们的流程包含以下三个部分:

- Detector. Starting with an instruction ℓ and an observed RGB image $\mathcal{O}$, the RGB image is passed through the Semantic-SAM segmentation model, which labels each object with a numerical tag, producing a labeled image $\mathcal{O}'$. The GPT4o-based Detector uses the prompts $\mathcal{P}_{\text{det}}$ and $\mathcal{P}'_{\text{det}}$ (adapted from Huang et al. (2024)) to select relevant objects and obtain their 3D positions. The output is a set of selected objects, or context $\mathcal{C}$, which includes the objects' identities and their spatial coordinates in the 3D environment. In the VLM setting, the Detector accesses initial RGB images from four views: wrist, left shoulder, right shoulder, and front camera. In the manual setting, precise 3D object attributes are provided from the simulation.

  - 检测器。以指令 ℓ 和观察到的 RGB 图像 $\mathcal{O}$ 为起点，RGB 图像通过 Semantic-SAM 分割模型，为每个物体标注数字标签，生成标注图像 $\mathcal{O}'$。基于 GPT4o 的检测器使用提示 $\mathcal{P}_{\text{det}}$ 和 $\mathcal{P}'_{\text{det}}$ (改编自 Huang 等人 (2024)) 选择相关物体并获取其 3D 位置。输出是一组选定物体或上下文 $\mathcal{C}$，包括物体身份及其在 3D 环境中的空间坐标。在 VLM 设置中，检测器访问来自手腕、左肩、右肩和前置摄像头的四个视角的初始 RGB 图像。在手动设置中，精确的 3D 物体属性由仿真提供。

- Planner. The GPT4o-based Planner takes the instruction ℓ, context $\mathcal{C}$, and planner-specific prompts $\mathcal{P}_{\text{plan}}$ and $\mathcal{P}'_{\text{plan}}$ (adapted from Huang et al. (2023)) to infer a sequence of sub-tasks $(st_1, st_2, \ldots, st_i)$. Each sub-task describes an action or interaction needed to fulfill the instruction ℓ. Progress is tracked based on the robot's gripper state (open/closed) and whether it is holding an object. The current sub-task is then passed to the Composer for further processing.

  - 规划器。基于 GPT4o 的规划器接收指令 ℓ、上下文 $\mathcal{C}$ 以及规划器专用提示 $\mathcal{P}_{\text{plan}}$ 和 $\mathcal{P}'_{\text{plan}}$ (改编自 Huang 等人 (2023))，推断一系列子任务 $(st_1, st_2, \ldots, st_i)$。每个子任务描述完成指令 ℓ 所需的动作或交互。进度基于机器人夹爪状态 (开/闭) 及是否持有物体进行跟踪。当前子任务随后传递给组合器进行进一步处理。

- Composer. Following Huang et al. (2023), the GPT4o-based Composer parses each inferred sub-task st_i using corresponding prompts $\mathcal{P}_{\text{com}}$ and $\mathcal{P}'_{\text{com}}$. The Composer generates the sub-goal position g^{pos} and sub-goal openness g^{open} by recursively generating code. This includes calls to get_cost_map and get_gripper_map, which are triggered by cost map prompt $\mathcal{P}_{\text{cost}}$ and gripper map prompt $\mathcal{P}_{\text{gripper}}$. For example, for a sub-task like "push close the topmost drawer," the Composer might generate: get_cost_map('a point 30 cm into the topmost drawer handle') and get_gripper_map('close everywhere'). Natural language parameters are parsed by GPT to generate code that assigns values to g^{pos} and g^{open}. The final GravMap generator in Algorithm 1 then processes g^{pos} and g^{open} to generate the GravMap m.

- 组合器。遵循 Huang 等人 (2023) 的做法，基于 GPT4o 的组合器使用相应的提示词解析每个推断出的子任务 st_i，$\mathcal{P}_{\text{com}}$ 和 $\mathcal{P}'_{\text{com}}$。组合器通过递归生成代码来生成子目标位置 g^{pos} 和子目标开合度 g^{open}。这包括调用 get_cost_map 和 get_gripper_map，这些调用由代价地图提示词 $\mathcal{P}_{\text{cost}}$ 和夹爪地图提示词 $\mathcal{P}_{\text{gripper}}$ 触发。例如，对于“推近最上层抽屉”这样的子任务，组合器可能生成:get_cost_map(’ 一个点 30 cm 进入最上层抽屉把手’) 和 get_gripper_map(’ 全程关闭’)。自然语言参数由 GPT 解析，生成赋值给 g^{pos} 和 g^{open} 的代码。算法 1 中的最终 GravMap 生成器随后处理 g^{pos} 和 g^{open} 以生成 GravMap m。

The prompts mentioned above can be found on the website: https://gravmad.github.io

上述提示词可在网站 https://gravmad.github.io 上找到

A.4 DETAIL OF MODEL ARCHITECTURE AND HYPER-PARAMETERS FOR GRAVMAD

A.4 GravMAD 模型架构及超参数细节

The detailed hyperparameters of GravMAD are listed in Table 3. Additionally, Fig. 5 provides a detailed overview of GravMAD’s modules, including the 3D Scene Encoder and the prediction heads.

GravMAD 的详细超参数列于表 3。此外，图 5 提供了 GravMAD 模块的详细概览，包括 3D 场景编码器和预测头。

(a) The 3D Scene Encoder processes visual and language information separately, merging them via a cross-attention mechanism, with proprioception integrated through FiLM. This allows the model to understand tasks like ”Take the chicken off the grill” in a 3D environment. First, the visual input is processed by a 2D Visual Encoder, transforming image data into feature representations. These 2D features are then passed through a 3D lifting module, converting them into 3D representations using depth information. Simultaneously, the language input, such as the instruction ”Take the chicken off the grill”, is encoded into language tokens by the Language Encoder. Finally, the 3D visual features and language tokens are combined through cross-attention, producing 3D Scene tokens.

(a) 3D 场景编码器分别处理视觉和语言信息，通过交叉注意力机制融合，体感信息通过 FiLM 集成。这使模型能够理解如“把鸡从烤架上拿下来”这样的 3D 环境任务。首先，视觉输入由 2D 视觉编码器处理，将图像数据转换为特征表示。随后，这些 2D 特征通过 3D 提升模块，利用深度信息转换为 3D 表示。与此同时，语言输入如指令“把鸡从烤架上拿下来”由语言编码器编码成语言标记。最后，3D 视觉特征与语言标记通过交叉注意力结合，生成 3D 场景标记。

(b) Each prediction head consists of Attention layers and an MLP. The Auxiliary Position Head receives tokens from the previous layer, which first go through cross-attention with GravMap tokens, followed by self-attention to refine the features. The tokens are then passed through an MLP to output the sub-goal end-effector position. Similarly, the Auxiliary Openness Head takes tokens from the self-attention layer of the Auxiliary Position Head and uses an MLP to predict the sub-goal gripper openness. The Position Head follows the same process as the

Auxiliary Position Head, while the Openness Head mirrors the Auxiliary Openness Head. The Rotation Head processes tokens with self-attention and an MLP to predict rotation error.

(b) 每个预测头由注意力层和多层感知机 (MLP) 组成。辅助位置头接收来自上一层的标记，先与 GravMap 标记进行交叉注意力，再通过自注意力细化特征。然后标记通过 MLP 输出子目标末端执行器位置。类似地，辅助开合度头接收辅助位置头自注意力层的标记，使用 MLP 预测子目标夹爪开合度。位置头的流程与辅助位置头相同，开合度头则对应辅助开合度头。旋转头通过自注意力和 MLP 处理标记，预测旋转误差。

A.5 Comparison between GravMAD and other baseline models

A.5 GravMAD 与其他基线模型的比较

We compare GravMAD with Voxposer (Huang et al., 2023) and 3D Diffuser Actor (Ke et al., 2024) in Fig. 6

我们在图 6 中比较了 GravMAD 与 Voxposer(Huang 等，2023) 及 3D Diffuser Actor(Ke 等，2024)。

(a) Voxposer. We describe our reproduction of Voxposer on RLBench. Voxposer uses our SOM-driven Detector to process the input observation and instruction, generating context information. The

(a) Voxposer。我们描述了在 RLBench 上复现 Voxposer 的过程。Voxposer 使用我们基于 SOM 的检测器处理输入的观察和指令，生成上下文信息。该

	Values
Sub − goalKeyposeDiscovery	
touch_threshold	2
Tolerance: delta	0.005
gripper_threshold	4
GravMap	
map_size: S_m	100
offset_range: β_o	3
radius: β_r	3
downsample ratio: β_d	4
number of sampled points: N_p	1024
Model	
image_size	256
token_dim	120
diffusion_timestep	100
noise_scheduler: position	scaled_linear
noise_scheduler: rotation	squaredcos
action_space	absolute pose
Train	
batch_size	8
optimizer	Adam
train_iters	600K
learning_rate	$1e^{-4}$
weight_decay	$5e^{-4}$
loss weight: ω_1	30
loss weight: ω_2	10
loss weight: ω_3	30
loss weight: ω_4	10
Evaluation	
maximal step except push_button_light	25
maximal step of push_button_light	3

	数值
Sub − goalKeyposeDiscovery	
触摸阈值	2
容差: 增量	0.005
夹持阈值	4
重力图 (GravMap)	
地图大小: S_m	100
偏移范围: β_o	3
半径: β_r	3
降采样比例: β_d	4
采样点数量: N_p	1024
模型	
图像尺寸	256
标记维度	120
扩散时间步 (diffusion_timestep)	100
噪声调度器: 位置	缩放线性
噪声调度器: 旋转	平方余弦 (squaredcos)
动作空间	绝对姿态
训练	
批量大小	8
优化器	Adam 优化器
训练迭代次数	600K
学习率	$1e^{-4}$
权重衰减	$5e^{-4}$
损失权重: ω_1	30
损失权重: ω_2	10
损失权重: ω_3	30
损失权重: ω_4	10
评估	
除按键灯 (push_button_light) 外的最大步数	25
按键灯 (push_button_light) 的最大步数	3

Table 3: Hyper-parameters for GravMAD, including Sub-goal Keypose Discovery, GravMap, model configuration, training, and evaluation.

表 3:GravMAD 的超参数，包括子目标关键姿势发现、GravMap、模型配置、训练和评估。

Planner then receives this context and outputs a sub-goal, representing an intermediate step necessary for the overall motion plan. The Composer processes this sub-goal, producing three maps: Cost Map, Rotation Map, and Gripper Map. These maps guide the robot's movement toward the target in the environment. Note that Voxposer's testing process involves a different number of steps compared to 3D Diffuser Actor and GravMAD, completing only after all LLM inferences are executed.

规划器接收该上下文并输出一个子目标，表示整体运动规划所需的中间步骤。组合器处理该子目标，生成三张地图：代价地图、旋转地图和夹爪地图。这些地图引导机器人在环境中朝目标移动。注意，Voxposer 的测试过程涉及的步骤数与 3D Diffuser Actor 和 GravMAD 不同，只有在所有大型语言模型 (LLM) 推理完成后才结束。

(b) 3D Diffuser Actor. We use the language-enhanced version of 3D Diffuser Actor as a baseline. 3D Diffuser employs a 3D Scene Encoder to transform visual and language inputs into 3D Scene tokens, providing an understanding of the 3D environment. An MLP encodes noisy estimates of position and rotation into corresponding tokens, which are then fed, along with the 3D Scene tokens, into a denoising network for action diffusion. This network, conditioned on proprioception and the denoising step, includes attention layers, Openness Head, Position Head, and Rotation Head. During diffusion, noisy position/rotation tokens attend to 3D Scene tokens, and cross-attention with instruction tokens enhances language understanding. These instruction tokens are also used in the prediction processes of the Openness, Position, and Rotation heads.

(b)3D Diffuser Actor。我们使用语言增强版的 3D Diffuser Actor 作为基线。3D Diffuser 采用 3D 场景编码器将视觉和语言输入转换为 3D 场景标记，提供对 3D 环境的理解。一个多层感知机 (MLP) 将位置和旋转的噪声估计编码为相应的标记，然后与 3D 场景标记一起输入去噪网络进行动作扩散。该网络以本体感受和去噪步骤为条件，包含注意力层、开放度头、位置头和旋转头。在扩散过程中，带噪声的位置/旋转标记关注 3D 场景标记，并通过与指令标记的交叉注意力增强语言理解。这些指令标记也用于开放度、位置和旋转头的预测过程。

(c) GravMAD (ours). GravMAD shares components with Voxposer, such as the Detector, Planner, and Composer, but incorporates task-specific prompt engineering. Unlike Voxposer, which uses maps for planning, GravMAD encodes these maps into tokens using a point cloud encoder, which are then employed in the action diffusion process. Compared to 3D Diffuser Actor, the key difference is that GravMAD uses GravMap tokens instead of language tokens, improving generalization. Additionally, GravMAD introduces two auxiliary tasks to predict subgoals, enhancing representation learning.

(c)GravMAD(本方法)。GravMAD 与 Voxposer 共享检测器、规划器和组合器等组件，但引入了任务特定的提示工程。与使用地图进行规划的 Voxposer 不同，GravMAD 通过点云编码器将这些地图编码为标记，随后用于动作扩散过程。与 3D Diffuser Actor 相比，关键区别在于 GravMAD 使用 GravMap 标记替代语言标记，提升了泛化能力。此外，GravMAD 引入了两个辅助任务以预测子目标，增强表示学习。

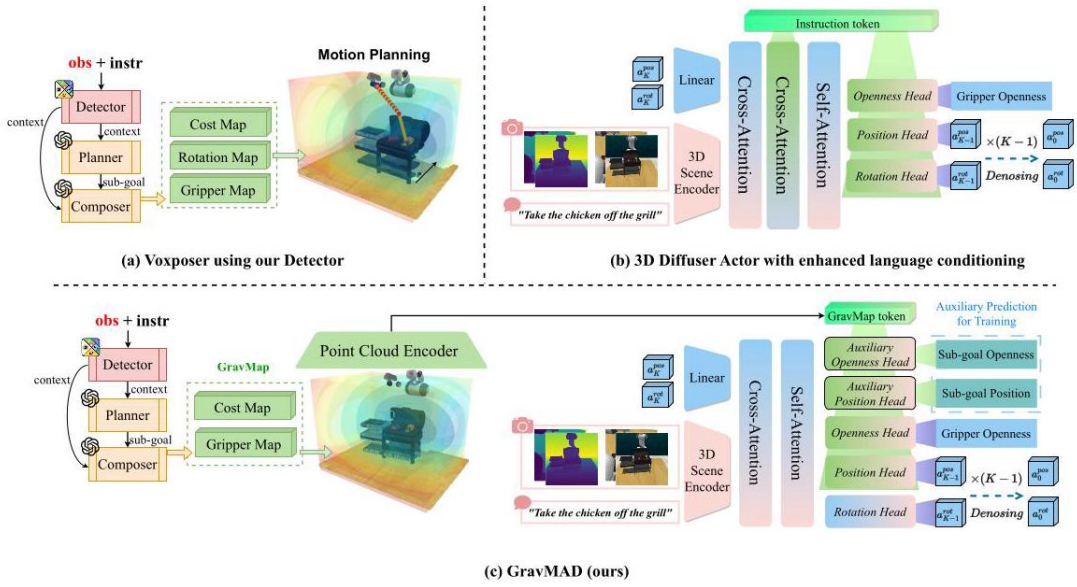


Figure 6: Comparison of GravMAD with Voxposer and 3D Diffuser Actor. Unlike Voxposer, which uses planning, GravMAD leverages GravMaps for learning. Compared to 3D Diffuser Actor, GravMAD employs GravMap tokens to guide the action diffusion process and introduces auxiliary position and openness heads to improve representation learning.

图 6:GravMAD 与 Voxposer 和 3D Diffuser Actor 的比较。不同于使用规划的 Voxposer，GravMAD 利用 GravMap 进行学习。相比 3D Diffuser Actor，GravMAD 采用 GravMap 标记引导动作扩散过程，并引入辅助位置头和开放度头以提升表示学习。

B ADDITIONAL EXPERIMENTAL DETAILS

B 附加实验细节

B.1 BASE TASK

B.1 基础任务

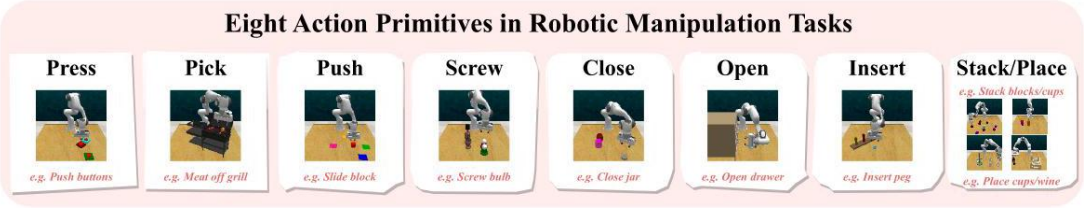


Figure 7: Eight action primitives in robotic manipulation tasks.

图 7: 机器人操作任务中的八个动作原语。

For the selection of base tasks, our primary criterion is to ensure they comprehensively cover the fundamental action primitives in robotic manipulation tasks. Therefore, we follow Garcia et al. (2024) and further summarize the eight essential action primitives required for robotic manipulation, as shown in Fig. 7. In line with this criterion, we select 12 base tasks from RLBench (James et al. 2020), as illustrated in Fig. 8. These 12 tasks also include short-term tasks (close jar, open drawer; meat off grill, slide block, push buttons, place wine), long-horizon tasks (put item in drawer, stack blocks, stack cups), and tasks that require high-precision manipulation (screw bulb, insert peg, place cups). Each base task contains 2 to 60 variants in the instructions, covering differences in color, placement, category, and count. In addition to instruction variations, the objects, distractors, and their positions and scenes are randomly initialized in the environment. The templates representing task goals in the instructions are also modified while maintaining their semantic meaning. A summary of the 12 tasks is provided in Table 4.

在基础任务的选择上，我们的主要标准是确保它们全面涵盖机器人操作任务中的基本动作原语。因此，我们遵循 Garcia 等人 (2024) 的工作，进一步总结了机器人操作所需的八个基本动作原语，如图 7 所示。基于此标准，我们从 RLBench(James 等人，2020) 中选取了 12 个基础任务，如图 8 所示。这 12 个任务包括短期任务 (关闭罐子、打开抽屉；从烤架取肉、滑动积木、按按钮、放置酒瓶)、长时任务 (将物品放入抽屉、堆叠积木、堆叠杯子) 以及需要高精度操作的任务 (拧灯泡、插销钉、放置杯子)。每个基础任务包含 2 到 60 个指令变体，涵盖颜色、位置、类别和数量的差异。除了指令变体外，环境中的物体、干扰物及其位置和场景均随机初始化。指令中表示任务目标的模板也会被修改，但保持其语义不变。12 个任务的总结见表 4。

We provide a detailed description of each task below and explain modifications from RLBench origin code-base.

下面我们详细描述每个任务，并说明相较于 RLBench 原始代码库的修改。

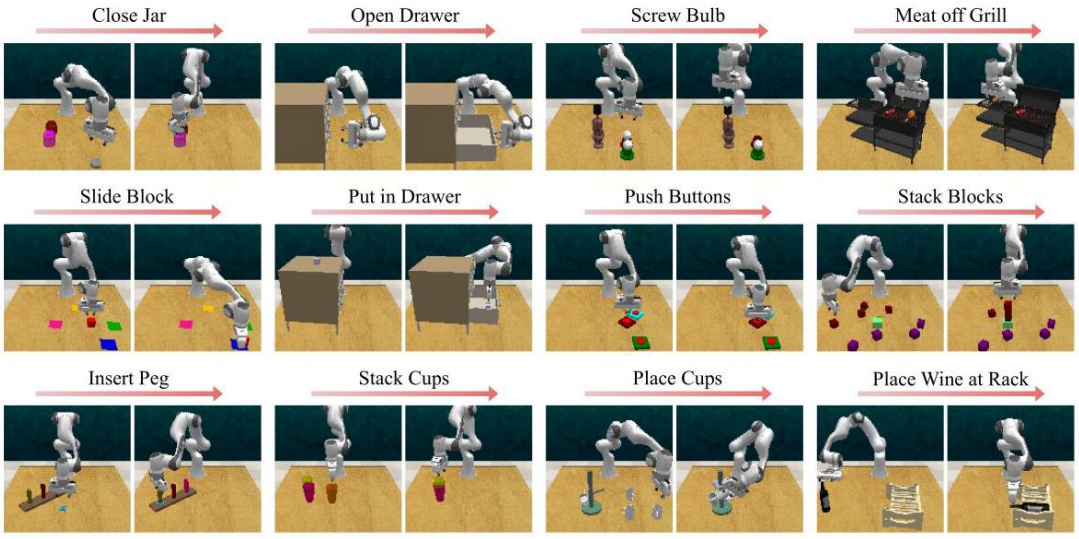


Figure 8: Visualization of 12 base tasks.

图 8:12 个基础任务的可视化。

Table 4: The 12 base tasks selected from RLBench (James et al., 2020)

表 4: 从 RLBench(James 等人，2020) 中选取的 12 个基础任务

Task	Variation Type	#of Variations	Avg. Keyposes	Language Template
close jar	color	20	6.0	”close the —jar”
open drawer	placement	3	3.0	”open the —drawer”
screw bulb	color	20	7.0	”screw in the —light bulb”
meat off grill	category	2	5.0	”take the —off the grill”
slide block	color	4	4.7	”slide the block to —target”
put in drawer	placement	3	12.0	”put the item in the - drawer”
push buttons	color	50	3.8	”push the - button, [then the - button]”
stack blocks	color, count	60	14.6	”stack —blocks”
insert peg	color	20	5.0	”put the ring on the - spoke”
stack cups	color	20	10.0	”stack the other cups on top of the - cup”
place cups	count	3	11.5	”place —cups on the cup holder”
place wine	count	3	5.0	”stack the wine bottle to the - of the rack”

任务	变体类型	变体数量	平均关键姿势	语言模板
关闭罐子	颜色	20	6.0	“关闭—罐子”
打开抽屉	放置	3	3.0	“打开—抽屉”
拧灯泡	颜色	20	7.0	“拧入—灯泡”
从烤架取肉	类别	2	5.0	“从烤架上取下—”
滑动积木	颜色	4	4.7	“将积木滑动到—目标”
放入抽屉	放置	3	12.0	“将物品放入—抽屉”
按按钮	颜色	50	3.8	“按—按钮，[然后按—按钮]”
堆积积木	颜色，数量	60	14.6	“堆叠—积木”
插入销钉	颜色	20	5.0	“将环放在—辐条上”
堆叠杯子	颜色	20	10.0	“将其他杯子堆叠在—杯子上”
放置杯子	计数	3	11.5	“将—个杯子放在杯架上”
放置酒瓶	计数	3	5.0	“将酒瓶堆放在架子的—侧”

B.1.1 CLOSE JAR

B.1.1 关闭罐子

Task: Close the jar by placing the lid on the jar.

任务: 通过将盖子盖在罐子上来关闭罐子。

filename: close_jar.py

文件名:close_jar.py

Modified: The modified success condition registers a single DetectedCondition to check if the jar lid is correctly placed on the jar using a proximity sensor, discarding the previous condition of checking if nothing is grasped by the gripper.

修改: 修改后的成功条件注册了一个单一的 DetectedCondition(检测条件), 通过接近传感器检查罐盖是否正确放置在罐子上，舍弃了之前检查夹爪是否空闲的条件。

Success Metric: The jar lid is successfully placed on the jar as detected by the proximity sensor.

成功指标: 通过接近传感器检测到罐盖成功放置在罐子上。

B.1.2 OPEN DRAWER

B.1.2 打开抽屉

Task: Open the drawer by gripping the handle and pulling it open.

任务: 通过抓住把手并拉开来打开抽屉。

filename: open_drawer.py

文件名:open_drawer.py

Modified: The cam_over_shoulder_left camera's position and orientation were modified to better observe the drawer. The camera was repositioned to $[0.2, 0.90, 1.10]$ and reoriented to [0.5 *math.pi,0,0] .

修改: 调整了 cam_over_shoulder_left 摄像头的位置和朝向，以更好地观察抽屉。摄像头被重新定位到 $[0.2, 0.90, 1.10]$ 并重新定向到 [0.5 *math.pi,0,0] 。

Success Metric: The drawer is successfully opened to the desired position as detected by the joint condition on the drawer's joint.

成功指标: 通过抽屉关节的关节条件检测到抽屉成功打开到预期位置。

B.1.3 SCREW BULB

B.1.3 拧灯泡

Task: Screw in the light bulb by picking it up from the holder and placing it into the lamp. filename: light_bulb_in.py

任务: 从灯座上取下灯泡并将其拧入灯具。文件名:light_bulb_in.py

Modified: No.

修改: 无。

Success Metric: The light bulb is successfully screwed into the lamp and detected by the proximity sensor.

成功指标: 灯泡成功拧入灯具，并通过接近传感器检测到。

B.1.4 MEAT OFF GRILL

B.1.4 从烤架上取肉

Task: Take the specified meat off the grill and place it next to the grill.

任务: 将指定的肉从烤架上取下并放置在烤架旁边。

filename: meat_off_grill.py

文件名:meat_off_grill.py

Modified: The cam_over_shoulder_right camera's position and orientation were modified to better observe the drawer. The camera was repositioned to $[0.20, -0.36, 1.85]$ and reoriented to [-0.85*math.pi, 0, math.pi].

修改:cam_over_shoulder_right 摄像头的位置和方向被调整，以更好地观察抽屉。摄像头被重新定位到 $[0.20, -0.36, 1.85]$ ，并重新定向为 [-0.85*math.pi, 0, math.pi]。

Success Metric: The specified meat is successfully removed from the grill and detected by the proximity sensor.

成功标准: 指定的肉类成功从烤架上移除，并被接近传感器检测到。

B.1.5 SLIDE BLOCK

B.1.5 滑动方块

Task: Slide the block to the target of a specified color.

任务: 将方块滑动到指定颜色的目标处。

filename: slide_block_to_color_target.py

文件名:slide_block_to_color_target.py

Modified: No.

修改: 无。

Success Metric: The block is successfully detected on top of the target color as indicated by the proximity sensor.

成功标准: 方块被接近传感器检测到成功放置在目标颜色上方。

B.1.6 PUT IN DRAWER

B.1.6 放入抽屉

Task: Put the item in the specified drawer.

任务: 将物品放入指定的抽屉中。

filename: put_item_in_drawer.py

文件名:put_item_in_drawer.py

Modified: The cam_over_shoulder_left camera's position and orientation were modified to better observe the drawer. The camera was repositioned to $[0.2, 0.90, 1.15]$ and reoriented to [0.5 *math.pi,0,0] .

修改:cam_over_shoulder_left 摄像头的位置和方向被调整，以更好地观察抽屉。摄像头被重新定位到 $[0.2, 0.90, 1.15]$ ，并重新定向为 [0.5 *math.pi,0,0] 。

Success Metric: The item is successfully placed in the drawer as detected by the proximity sensor.

成功标准: 物品被成功放入抽屉中，并被接近传感器检测到。

B.1.7 PUSH BUTTONS

B.1.7 按钮按压

Task: Press the buttons of the specified color in order

任务: 按顺序按下指定颜色的按钮

filename: push_buttons.py

文件名:push_buttons.py

Modified: No.

修改: 否。

Success Metric: The buttons are successfully pushed in order.

成功标准: 按钮按顺序成功按下。

B.1.8 STACK BLOCKS

B.1.8 堆叠积木

Task: Stack a specified number of blocks of the same color in a vertical stack.

任务: 将指定数量的同色积木垂直堆叠。

filename: stack_blocks.py

文件名:stack_blocks.py

Modified: No.

修改: 否。

Success Metric: The blocks are successfully stacked according to the specified color and number.

成功标准: 积木按指定颜色和数量成功堆叠。

B.1.9 INSERT PEG

B.1.9 插入插销

Task: Insert a square ring onto the spoke with the specified color.

任务: 将方形环插入指定颜色的辐条上。

filename: insert_onto_square_peg.py

文件名:insert_onto_square_peg.py

Modified: No.

修改: 否。

Success Metric: The square ring is successfully placed onto the correctly colored spoke.

成功标准: 方形环成功放置在正确颜色的辐条上。

B.1.10 STACK CUPS

B.1.10 堆叠杯子

Task: Stack two cups on top of the cup with the specified color.

任务: 将两个杯子堆叠在指定颜色的杯子上。

filename: stack_cups.py

文件名:stack_cups.py

Modified: No.

修改: 否。

Success Metric: The cups are successfully stacked with the correct cup as the base.

成功指标: 杯子成功堆叠，且以正确的杯子作为底座。

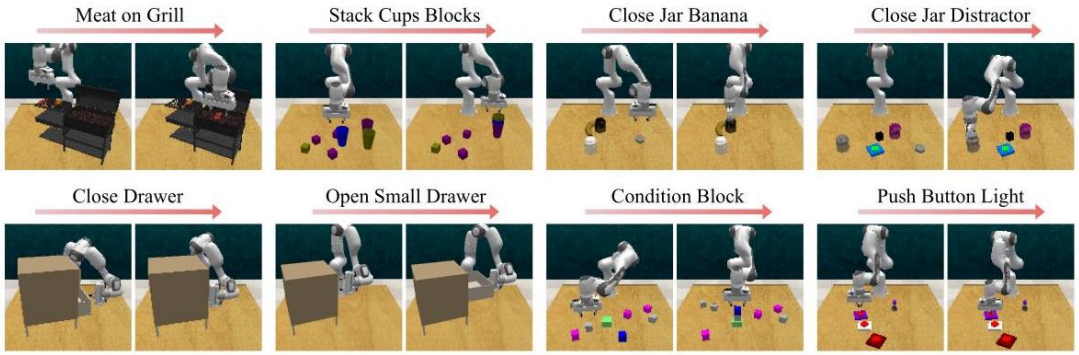


Figure 9: Visualization of 8 novel tasks.

图 9:8 个新颖任务的可视化。

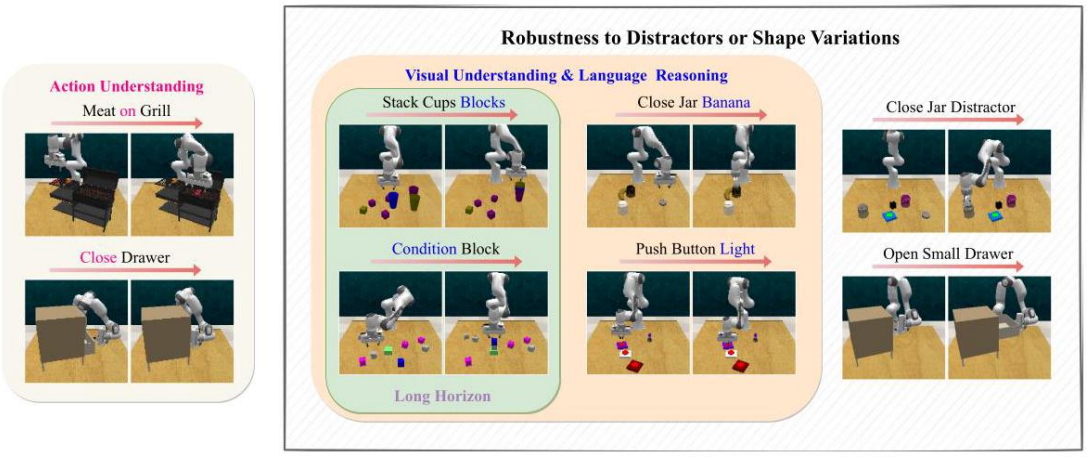


Figure 10: Three Novelty Categories for the Novel Tasks.

图 10: 新颖任务的三种新颖性类别。

B.1.11 PLACE CUPS

B.1.11 放置杯子

Task: Place a specified number of cups onto a cup holder.

任务: 将指定数量的杯子放置到杯架上。

filename: place_cups.py

文件名:place_cups.py

Modified: No.

是否修改: 否。

Success Metric: The cups are successfully placed onto the holder according to the task instructions.

成功指标: 杯子根据任务指令成功放置到杯架上。

B.1.12 Place Wine at Rack Location

B.1.12 在酒架位置放置酒瓶

Task: Place the wine bottle onto the specified location on the wine rack.

任务: 将酒瓶放置到酒架指定位置。

filename: place_wine_at_rack_location.py

文件名:place_wine_at_rack_location.py

Modified: No.

是否修改: 否。

Success Metric: The wine bottle is successfully placed at the correct rack location and released from the gripper.

成功指标: 酒瓶成功放置在正确的酒架位置并从夹持器释放。

B.2 NOVEL TASK

B.2 新颖任务

As shown in Fig. 9, we create 8 novel tasks that differ from the original training tasks to test policy generalization. These tasks feature scenes and objects similar to those in the training tasks. We further define the novelty

categories of the 8 novel tasks in our experiments to better explain the generalization improvements brought by GravMAD. As shown in Fig. 10, the designed novel tasks introduce three types of challenges to the model: Action Understanding (meat on grill, close drawer), Visual Understanding & Language Reasoning (stack cups blocks, push buttons light, close jar banana, condition block)-including two long-horizon tasks (stack cups blocks and condition block), and Robustness to Distractors or Shape Variations (stack cups blocks, push buttons light, close jar banana, close jar distractor, open small drawer, condition block).

如图 9 所示，我们创建了 8 个与原始训练任务不同的新颖任务，用以测试策略的泛化能力。这些任务的场景和物体与训练任务相似。我们进一步定义了实验中 8 个新颖任务的新颖性类别，以更好地解释 GravMAD 带来的泛化提升。如图 10 所示，设计的新颖任务为模型引入了三类挑战: 动作理解 (烤架上的肉，关闭抽屉)、视觉理解与语言推理 (堆叠杯子积木，按按钮开灯，关闭香蕉罐，条件积木)——其中包括两个长时序任务 (堆叠杯子积木和条件积木)，以及对干扰物或形状变化的鲁棒性 (堆叠杯子积木，按按钮开灯，关闭香蕉罐，关闭罐子干扰物，打开小抽屉，条件积木)。

Table 5: The 8 novel tasks changed based on base tasks.

表 5: 基于基础任务修改的 8 个新颖任务。

Task	Variation Type	#of Variations	Avg. Keyposes	Language Template
close drawer	placement	3	2.0	”close the —drawer”
close jar banana	placement	2	6.0	”close the jar closer to the banana”
close jar distractors	color	20	6.0	”close the —jar”
condition block	color, count	72	11.0	”build a tall tower out of - - cubes, and add a black block if it exists”
meat on grill	category	2	5.0	”put the —on the grill”
open small drawer	placement	3	3.0	”open the —drawer”
stack cups blocks	color	20	10.0	”Identify the most common color in the block pile, and stack the other cups on the cup that matches that color”
push button light	color	20	2.0	”push the button with the same color as the light”

任务	变体类型	变体数量	平均关键帧	语言模板
关闭抽屉	放置	3	2.0	“关闭—抽屉”
关闭香蕉罐	放置	2	6.0	“关闭靠近香蕉的罐子”
关闭罐子干扰项	颜色	20	6.0	“关闭—罐子”
条件块	颜色，数量	72	11.0	“用–个积木搭建一座高塔，如果存在黑色积木则添加”
烤架上的肉	类别	2	5.0	“把—放到烤架上”
打开小抽屉	放置	3	3.0	“打开—抽屉”
堆叠杯子积木	颜色	20	10.0	“识别积木堆中最常见的颜色，将其他杯子堆叠在与该颜色相匹配的杯子上”
按按钮灯	颜色	20	2.0	“按与灯颜色相同的按钮”

Specifically, Action Understanding refers to tasks involving changes in interaction actions with objects; Visual Understanding & Language Reasoning involve introducing entirely new operational rules or conditions compared to known tasks; and Robustness to Distractors or Shape Variations includes tasks that require interaction based on fixed object attributes (such as color, size, distance, or distractors). A summary of the seven tasks is provided in Table 5 . We provide a detailed description of each novel task below and explain the modifications from the base tasks.

具体来说，动作理解 (Action Understanding) 指涉及与物体交互动作变化的任务；视觉理解与语言推理 (Visual Understanding & Language Reasoning) 涉及引入与已知任务完全不同的操作规则或条件；对干扰物或形状变化的鲁棒性 (Robustness to Distractors or Shape Variations) 包括基于固定物体属性 (如颜色、大小、距离或干扰物) 进行交互的任务。七个任务的总结见表 5。以下我们详细描述每个新任务，并解释相较于基础任务的修改。

B.2.1 MEAT ON GRILL

B.2.1 烤架上的肉

Task: Place either a chicken or a steak on the grill depending on the variation.

任务: 根据不同变体，将鸡肉或牛排放置在烤架上。

filename: meat_on_grill.py

文件名:meat_on_grill.py

Base task: meat off grill.

基础任务: 将肉从烤架上取下。

Modified: The task requires placing meat onto the grill, whereas the base task involves removing it. The cam_over_shoulder_right camera's position and orientation were modified to better observe the drawer. The camera was repositioned to $[0.20, -0.36, 1.85]$ and reoriented to $[-0.85*$ math.pi, 0, math.pi].

修改: 该任务要求将肉放到烤架上，而基础任务是将肉取下。cam_over_shoulder_right 摄像头的位置和方向被调整，以更好地观察抽屉。摄像头被重新定位到 $[0.20, -0.36, 1.85]$，并重新定向为 $[-0.85*$ [math.pi, 0, math.pi]。

Success Metric: The selected meat (chicken or steak) is successfully placed on the grill and released from the gripper.

成功标准: 所选肉类 (鸡肉或牛排) 成功放置在烤架上，并从夹持器中释放。

B.2.2 STACK CUPS BLOCKS

B.2.2 堆叠杯子和积木

Task: Identify the most common color in the block pile, and stack the other cups on the cup that matches that color.

任务: 识别积木堆中最常见的颜色，并将其他杯子堆叠在与该颜色匹配的杯子上。

filename: stack_cups_blocks.py

文件名:stack_cups_blocks.py

Base task: Stack cups.

基础任务: 堆叠杯子。

Modified: The task involves identifying the cup that matches the most common color among the distractor blocks, then stacking the other two cups on top. The base task is simply stacking the cups without considering block colors.

修改: 任务要求识别与干扰积木中最常见颜色匹配的杯子，然后将另外两个杯子堆叠在其上。基础任务仅仅是堆叠杯子，不考虑积木颜色。

Success Metric: Success is measured when the correct cup is stacked with the other cups based on the color identification and all cups are within the target area defined by the proximity sensor.

成功标准: 当基于颜色识别正确堆叠杯子，且所有杯子均位于接近传感器定义的目标区域内时，任务成功。

B.2.3 CLOSE JAR BANANA

B.2.3 关闭靠近香蕉的罐子

Task: Close the jar that is closer to the banana by screwing on its lid.

任务: 通过旋紧盖子关闭靠近香蕉的罐子。

filename: close_jar_banana.py

文件名:close_jar_banana.py

Base task: close jar.

基础任务: 关闭罐子。

Modified: The task involves identifying the jar closer to the banana and screwing its lid on, while the base task only requires closing a jar without proximity consideration.

修改后: 任务包括识别靠近香蕉的罐子并旋紧其盖子，而基础任务仅要求关闭罐子，不考虑距离。

Success Metric: The lid is successfully placed on the jar closest to the banana, confirmed by the proximity sensor.

成功标准: 盖子成功盖在最靠近香蕉的罐子上，由接近传感器确认。

B.2.4 CLOSE JAR DISTRACTOR

B.2.4 关闭罐子干扰任务

Task: Close the jar by screwing on the lid, while distractor objects are present.

任务: 在存在干扰物的情况下，通过旋紧盖子关闭罐子。

filename: close_jar_distractor.py

文件名:close_jar_distractor.py

Base task: close jar.

基础任务: 关闭罐子。

Modified: The task includes distractor objects, such as a button and block, which are colored and placed near the jars. These objects have been encountered during training, adding complexity compared to the base task.

修改后: 任务中包含干扰物，如按钮和积木，这些物体有颜色并放置在罐子附近。训练时已遇到这些物体，增加了任务复杂度。

Success Metric: The jar lid is successfully placed on the target jar, confirmed by the proximity sensor.

成功标准: 罐子盖成功盖在目标罐子上，由接近传感器确认。

B.2.5CLOSE DRAWER

B.2.5 关闭抽屉

Task: Close one of the drawers (bottom, middle, or top) by sliding it shut.

任务: 通过滑动关闭一个抽屉 (底部、中部或顶部)。

filename: close_drawer.py

文件名:close_drawer.py

Base task: open drawer.

基础任务: 打开抽屉。

Modified: The task involves closing the drawer instead of opening it.

修改后: 任务改为关闭抽屉而非打开。

Success Metric: The selected drawer is closed successfully, confirmed by the joint position of the drawer.

成功标准: 所选抽屉成功关闭，由抽屉关节位置确认。

B.2.6 OPEN DRAWER SMALL

B.2.6 打开小抽屉

Task: Open one of the smaller drawers (bottom, middle, or top) by sliding it open.

任务: 通过滑动打开较小的抽屉之一 (底部、中部或顶部)。

filename: open_drawer_small.py

文件名:open_drawer_small.py

Base task: open drawer.

基础任务: 打开抽屉。

Modified: The task involves opening a smaller drawer compared to the base task, with adjusted camera settings for better visibility.

修改说明: 任务涉及打开比基础任务更小的抽屉，并调整摄像头设置以获得更好的可视性。

Success Metric: The selected drawer is opened successfully, verified by the joint position of the drawer.

成功标准: 所选抽屉成功打开，通过抽屉的关节位置进行验证。

B.2.7 CONDITION BLOCK

B.2.7 条件积木

Task: Stack a specified number of blocks and, if the black block is present, add it to the stack. filename: condition_block.py

任务: 堆叠指定数量的积木，如果存在黑色积木，则将其加入堆叠。文件名:condition_block.py

Base task: stack blocks.

基础任务: 堆积木。

Modified: The task involves stacking a specified number of blocks, with an additional requirement to include the black block if it is present.

修改说明: 任务涉及堆叠指定数量的积木，额外要求如果存在黑色积木，则必须将其包含在堆叠中。

Success Metric: The correct number of target blocks are stacked, and if the black block is present, it is also correctly added to the stack.

成功标准: 正确数量的目标积木被堆叠，如果存在黑色积木，也被正确加入堆叠。

B.2.8 Push Button Light

B.2.8 按钮灯

Task: Push the button that matches the color of a light bulb on the first attempt.

任务: 首次尝试按下与灯泡颜色匹配的按钮。

filename: push_buttons_light.py

文件名:push_buttons_light.py

Base task: push button.

基础任务: 按按钮。

Modified: The task involves pressing a single button that matches the color of a light bulb. The button must be pressed correctly on the first attempt; repeated attempts are not allowed.

修改说明: 任务涉及按下与灯泡颜色匹配的单个按钮，必须在首次尝试时正确按下，不允许重复尝试。

Success Metric: The correct button matching the light bulb's color is pressed on the first attempt.

成功标准: 首次尝试正确按下与灯泡颜色匹配的按钮。

B.3 FAILURE CASES OF GRAVMAD

B.3 GravMAD 的失败案例分析

In this section, we analyze why GravMAD underperforms compared to the baseline model 3D Diffuser Actor on certain base tasks, particularly in the "Place Wine" task and drawer-related tasks.

本节中，我们分析了 GravMAD 在某些基础任务上表现不如基线模型 3D Diffuser Actor 的原因，特别是在“放置酒瓶”任务和与抽屉相关的任务中。

As discussed in the main paper, GravMaps represent spatial relationships in 3D space, but this introduces a challenge: areas close to the sub-goal often share the same cost value, as seen in the value map on the right side of Fig. 11 (a). This uniform cost value can mislead the robot into assuming it should complete the sub-goal within that area. For tasks requiring precise actions, such as the "Open Drawer" task, GravMaps' coarse guidance may lead to suboptimal performance compared to 3D Diffuser Actor. In the left schematic of Fig. 11(a), the robot must grasp the center of a small handle to achieve optimal performance in the "Open Drawer" task. This high precision

demand on the end-effector results in a lower success rate for GravMAD. This limitation extends to the ”Put in Drawer” task, which depends on the successful completion of ”Open Drawer”. Similarly, in the ”Place Wine” task, insufficient predictive accuracy causes the robot to misalign the bottle with the correct slot by one unit, leading to failure.

如主论文所述，GravMaps 表示三维空间中的空间关系，但这带来了一个挑战: 靠近子目标的区域通常具有相同的代价值，如图 11(a) 右侧的价值图所示。这种统一的代价值可能误导机器人认为应在该区域内完成子目标。对于需要精确操作的任务，如“打开抽屉”任务，GravMaps 的粗略引导可能导致性能不如 3D Diffuser Actor。在图 11(a) 左侧的示意图中，机器人必须抓住小把手的中心才能在“打开抽屉”任务中达到最佳表现。对末端执行器的高精度要求导致 GravMAD 的成功率较低。这一限制同样影响“放入抽屉”任务，该任务依赖于“打开抽屉”的成功完成。类似地，在“放置酒瓶”任务中，预测精度不足导致机器人将酒瓶与正确槽位错开一个单位，导致失败。

In the VLM setting, sub-goal accuracy often suffers, as shown in Fig. 11(b), further reducing model performance. These inaccuracies typically arise from two factors: (1) SAM may fail to accurately

在 VLM(视觉语言模型) 设置中，子目标的准确性常常受到影响，如图 11(b) 所示，进一步降低了模型性能。这些不准确通常源于两个因素:(1)SAM 可能无法准确

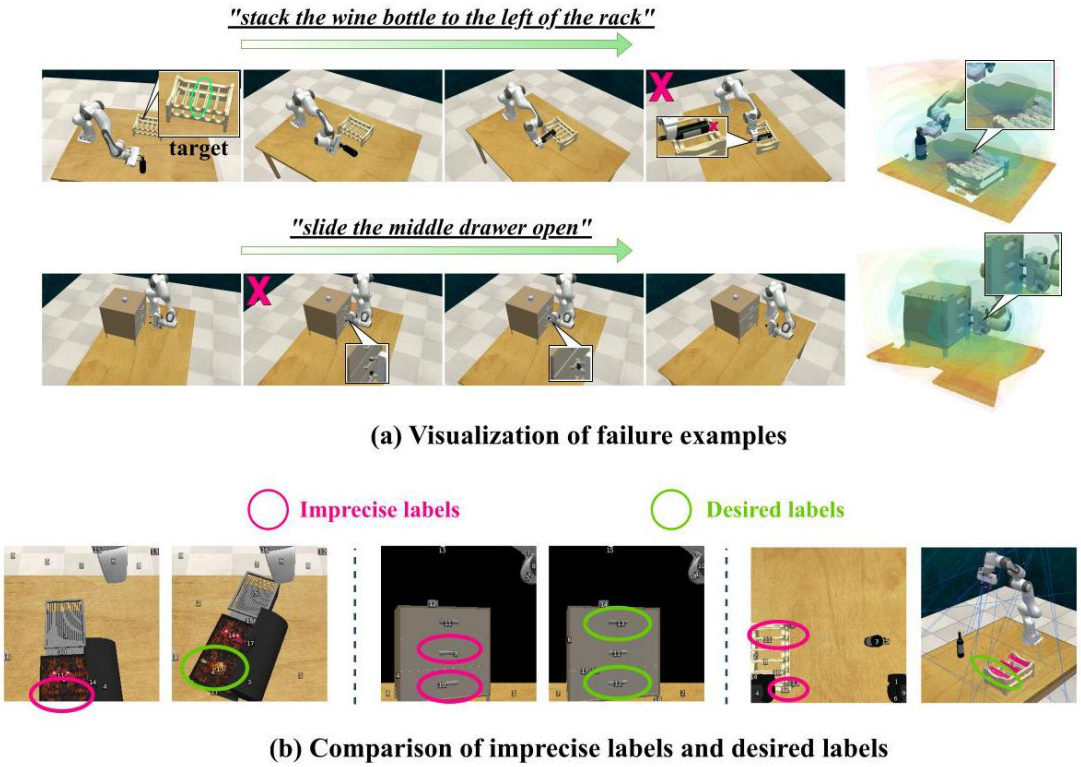


Figure 11: Failure cause analysis, including (a) visualization of failure examples; (b) comparison of imprecise labels and expected labels.

图 11: 失败原因分析，包括 (a) 失败示例的可视化；(b) 不精确标签与预期标签的比较。

identify ideal areas, leading to imprecise contextual information from the Detector module for tasks like ”Place Wine”, ”Open Drawer”, and ”Put in Drawer”; (2) the camera’s positioning may not capture the full scene, leaving some task-relevant objects out of view, as seen in tasks like ”Meat off Grill”. To overcome these VLM limitations, potential solutions include: (1) integrating multi-view information into the Detector for a more comprehensive scene observation; and (2) using a more granular segmentation model to provide GPT-4 with a wider range of labels, improving the quality of the context generated by the Detector.

识别理想区域，导致检测器模块提供的上下文信息不精确，影响“放置酒瓶”、“打开抽屉”和“放入抽屉”等任务；(2) 摄像头位置可能无法捕捉完整场景，部分任务相关物体未被视野覆盖，如“烤肉取下”任务所示。为克服这些 VLM 限制，潜在解决方案包括:(1) 将多视角信息整合进检测器，实现更全面的场景观察；(2) 采用更细粒度的分割模型，为 GPT-4 提供更丰富的标签，提升检测器生成上下文的质量。

C DISCUSSION

C 讨论

C.1 The Relationship and Differences Between GravMap and Voxposer

C.1 GravMap 与 Voxposer 的关系与区别

The GravMap in GravMAD and the value maps in Voxposer (Huang et al. 2023) share the following connections and differences:

GravMAD 中的 GravMap 与 Voxposer(Huang et al. 2023) 中的价值图存在以下联系与差异:

- Number of value maps involved: Voxposer utilizes multiple value maps, including the cost map, rotation map, gripper openness map, and velocity map. In our method, we only combine the cost map and gripper map, and their numerical values remain identical at this stage.

  - 价值图数量:Voxposer 使用多种价值图，包括代价图、旋转图、夹爪开合图和速度图。而我们的方法仅结合了代价图和夹爪图，且此阶段它们的数值保持一致。

- Structure and processing: We further downsample the cost map and gripper openness map, transforming them into a point cloud structure containing position information and gripper states (x, y, z, m_c, m_g) , which we term GravMap. This sparse data structure not only efficiently represents sub-goals but also allows feature extraction using a point cloud encoder.

  - 结构与处理: 我们进一步下采样代价图和夹爪开合图，将其转化为包含位置信息和夹爪状态的点云结构 (x, y, z, m_c, m_g) ，称为 GravMap。这种稀疏数据结构不仅高效表示子目标，还允许使用点云编码器进行特征提取。

C.2 THE REASON FOR NOT USING THE ROTATION MAP FROM VOXPOSER

C.2 不使用 Voxposer 旋转图的原因

GravMap does not currently use the rotation map from Voxposer because incorporating the rotation map could introduce significant distributional shifts between the guidance provided during the training and inference phases. During training, precise rotation guidance can be derived from expert trajectories. However, during inference, off-the-shelf foundation models often struggle to accurately interpret rotation information from visual and linguistic inputs, making it challenging to provide precise rotation guidance. To address this issue, future research will explore integrating rotation information from expert trajectories with object poses to generate few-shot prompts for off-the-shelf foundation models (Yin et al., 2024). This approach aims to enable LLMs to produce effective rotation guidance while reducing distributional shifts relative to the training data.

GravMap 目前未使用 Voxposer 的旋转图，因为引入旋转图可能导致训练与推理阶段指导信息的分布显著偏移。训练时，可以从专家轨迹中获得精确的旋转指导；但推理时，现成的基础模型往往难以准确解读视觉和语言输入中的旋转信息，难以提供精确的旋转指导。为解决此问题，未来研究将探索结合专家轨迹中的旋转信息与物体姿态，生成针对现成基础模型的少样本提示 (Yin et al., 2024)。该方法旨在使大型语言模型 (LLMs) 生成有效的旋转指导，同时减少与训练数据的分布偏移。

C.3 Further Details on Sub-goal Keypose Discovery

C.3 关于子目标关键姿态发现的更多细节

C.3.1 WHY SUB-GOALS ARE EXTRACTED DIFFERENTLY DURING TRAINING AND INFERENCE

C.3.1 为什么训练和推理阶段子目标提取方式不同

During the training phase of GravMAD, we use Sub-goal Keypose Discovery to extract sub-goals and generate GravMaps based on them. In contrast, during the inference phase, sub-goals are inferred by foundation models to generate GravMaps. The reasons for adopting different methods to generate GravMaps during the training and inference phases are as follows:

在 GravMAD 的训练阶段，我们使用子目标关键姿态发现方法提取子目标并基于其生成 GravMaps。相比之下，推理阶段的子目标由基础模型推断以生成 GravMaps。训练和推理阶段采用不同方法生成 GravMaps 的原因如下:

- Efficiency and reliability during training: Using Sub-goal Keypose Discovery to extract sub-goals during training is both simple and efficient. If foundation models were directly used to generate GravMaps as guidance during training, while they can indeed produce GravMaps, the results are generally coarser, less precise, and slower compared to expert trajectories. For example, due to limitations such as camera resolution

or angles, foundation models may fail to fully observe the scene in some cases, leading to inaccurate sub-goal positions (failure cases are discussed in Appendix B.3). Under such circumstances, the quality of the training data cannot be guaranteed. Additionally, using foundation models to process large-scale data is practically infeasible due to their slow processing speed.

- 训练过程中的效率和可靠性: 在训练过程中使用子目标关键姿态发现 (Sub-goal Keypose Discovery) 来提取子目标既简单又高效。如果直接使用基础模型 (foundation models) 生成 GravMaps 作为训练指导，虽然确实可以生成 GravMaps，但结果通常较为粗糙、精度较低且速度较慢，相较于专家轨迹而言。例如，由于摄像头分辨率或角度等限制，基础模型在某些情况下可能无法完全观察场景，导致子目标位置不准确 (失败案例详见附录 B.3)。在这种情况下，训练数据的质量无法得到保证。此外，由于基础模型处理速度较慢，使用其处理大规模数据在实际中不可行。

- Simplifying the problem by avoiding semantic reasoning: Extracting sub-goals from expert trajectories focuses solely on analyzing the robot's actions, thereby avoiding the complexity of semantic understanding and reasoning. Our key insight is that in task trajectories, certain actions in expert trajectories inherently carry semantic information (i.e., sub-goals, which may involve direct interactions with objects). These actions often exhibit distinctive features, such as the opening and closing of the gripper. The Keypose Discovery method (James & Davison, 2022) has already performed an initial filtering of these key actions, narrowing the scope for sub-goal selection. Based on this, we can quickly identify sub-goals through heuristic methods, which are also effective for long-horizon tasks.

- 通过避免语义推理简化问题: 从专家轨迹中提取子目标仅关注分析机器人动作，从而避免了语义理解和推理的复杂性。我们的关键见解是，在任务轨迹中，专家轨迹中的某些动作本身就携带语义信息 (即子目标，可能涉及与物体的直接交互)。这些动作通常表现出显著特征，如夹爪的开合。关键姿态发现方法 (James & Davison, 2022) 已经对这些关键动作进行了初步筛选，缩小了子目标选择的范围。在此基础上，我们可以通过启发式方法快速识别子目标，这对长时序任务同样有效。

It is worth noting that using different sub-goal generation methods during the training and inference phases may lead to a distributional shift. This occurs because the sub-goals generated by foundation models during inference are often less precise compared to those derived from expert trajectories, resulting in a discrepancy between the distributions of the training and inference phases. To address this issue, we apply data augmentation to the precise sub-goals generated from expert trajectories during the training phase. Specifically, as described in Line 279 of Algorithm 1, we introduce random offsets to the sub-goals generated during training (this processing is not applied to sub-goals generated during inference) and then generate GravMaps based on these perturbed sub-goals. This approach effectively reduces the risk of distributional shift to a certain extent.

值得注意的是，在训练和推理阶段使用不同的子目标生成方法可能导致分布偏移。这是因为推理阶段由基础模型生成的子目标通常不如专家轨迹生成的精确，导致训练和推理阶段的分布存在差异。为了解决这一问题，我们在训练阶段对专家轨迹生成的精确子目标进行了数据增强。具体来说，如算法 1 第 279 行所述，我们对训练阶段生成的子目标引入随机偏移 (推理阶段生成的子目标不进行此处理)，然后基于这些扰动后的子目标生成 GravMaps。这种方法在一定程度上有效降低了分布偏移的风险。

C.3.2 Why use Sub-Goal Keypose Discovery to filter keyposes

C.3.2 为什么使用子目标关键姿态发现来筛选关键姿态

The Sub-goal Keypose Discovery method is essential for GravMAD because the original keyposes include both sub-goal keyposes and the intermediate steps required to achieve these sub-goals. These intermediate steps may involve precise alignment of the robotic arm with objects. However, foundation models often struggle to generate these intermediate steps, and even if they can, the results may exhibit significant distributional shifts compared to the guidance provided during the training phase. Additionally, generating only sub-goals reduces the complexity and difficulty of task reasoning for the foundation model while also simplifying the prompt engineering.

子目标关键姿态发现方法对于 GravMAD 至关重要，因为原始关键姿态包含了子目标关键姿态以及实现这些子目标所需的中间步骤。这些中间步骤可能涉及机器人手臂与物体的精确对齐。然而，基础模型通常难以生成这些中间步骤，即使能够生成，结果也可能与训练阶段提供的指导存在显著的分布偏移。此外，仅生成子目标可以降低基础模型的任务推理复杂度和难度，同时简化提示工程。

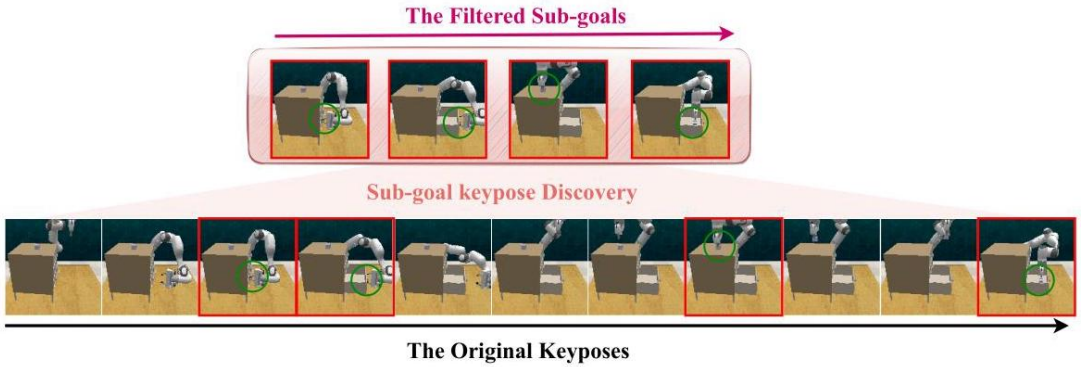


Figure 12: A comparison between the original keyposes and the filtered keyposes in the long-horizon task put item in drawer.

图 12: 长时序任务“将物品放入抽屉”中原始关键姿态与筛选后关键姿态的对比。

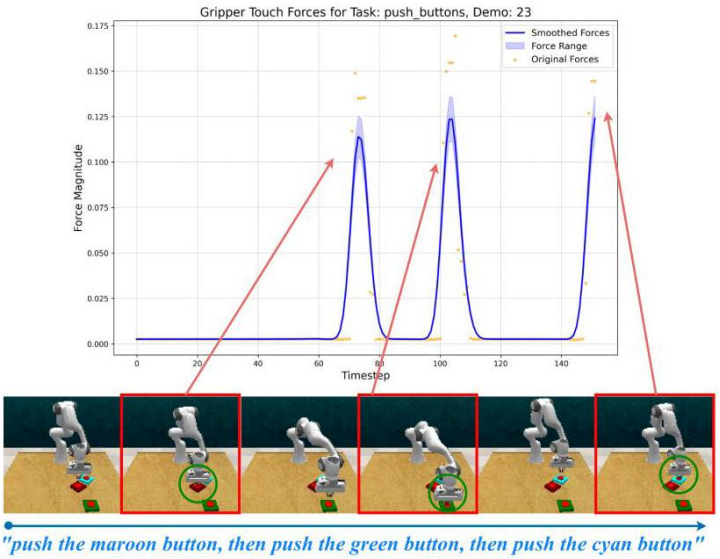


Figure 13: Visualization of sub-goal keypose discovery determining significant changes in gripper_torch_force during the push button task.

图 13: 子目标关键姿态发现在“按按钮”任务中确定夹爪触碰力 (gripper_touch_force) 显著变化的可视化。

As shown in Fig. 12, for the long-horizon task put item in drawer, if only traditional keypose discovery methods are used, the extracted sub-goal stages would include 11 stages. In contrast, when using our Sub-goal Keypose Discovery, the filtered sub-goals are reduced to just 4 stages, perfectly aligning with the most critical phases of the task. This significantly reduces model inference time and improves task execution efficiency.

如图 12 所示，对于长时序任务“将物品放入抽屉”，如果仅使用传统的关键姿态发现方法，提取的子目标阶段将包括 11 个阶段。相比之下，使用我们的子目标关键姿态发现，筛选后的子目标仅剩 4 个阶段，完美对应任务中最关键的阶段。这显著减少了模型推理时间，提高了任务执行效率。

C.3.3 Criteria for ”Significant Changes” in Sub-goal Keypose Discovery

C.3.3 子目标关键姿态发现中“显著变化”的判定标准

To clearly explain the specific criteria for ”significant changes” in our Sub-goal Keypose Discovery method, we visualized the changes in gripper_touch_force using the push buttons task as an example. As shown in Fig. 13, when the button is pressed, the gripper_touch_force value increases from nearly 0 to $0.1 \sim 0.15$. As the robotic arm lifts, the gripper_touch_force returns to 0 . By analyzing these force changes, we can intuitively identify the sub-goal frames.

为了清晰说明我们子目标关键姿态发现方法中“显著变化”的具体判定标准，我们以“按按钮”任务为例，展示了夹爪触碰力 (gripper_touch_force) 的变化。如图 13 所示，当按钮被按下时，夹爪触碰力值从接近 0 上升到 $0.1 \sim 0.15$ 。当机器人手臂抬起时，夹爪触碰力回落至 0。通过分析这些力的变化，我们可以直观地识别子目标帧。

	Voxposer	Act3D	3D Diffuser Actor	GravMAD (ours)
Avg. Inference Time for Keypose Prediction (secs)	/	0.04	1.78	1.81
Avg. Task Completion Time (secs)	448.01	14.08	49.45	97.04
Avg. Inference Time per Sub-task Stage (secs)	90.47	/	/	40.64

	Voxposer	Act3D	三维扩散器角色 (3D Diffuser Actor)	GravMAD(本方法)
关键姿势预测平均推理时间 (秒)	/	0.04	1.78	1.81
任务完成平均时间 (秒)	448.01	14.08	49.45	97.04
每个子任务阶段平均推理时间 (秒)	90.47	/	/	40.64

Table 6: Comparison of Inference Times.

表 6: 推理时间比较。

Models	Avg. Success ↑	Avg. Rank ↓	Close Jar	Open Drawer	Meat off Grill	Slide Block	Put in Drawer
Voxposer	15.11	5.88	12.00	10.67	45.33	0.00	0.00
Voxposer (Manual)	22.06	4.92	13.33	18.67	69.33	0.00	0.00
ChainedDiffuser (Oracle)	29.72	4.42	82.67	0.00	52.00	2.67	0.00
Act3D	34.11	5.38	61.33	41.33	60.00	78.67	49.33
3D Diffuser Actor	55.81	3.00	66.67	88.00	88.00	84.00	94.67
GravMAD (Manual)	69.17	1.63	100.00	76.67	89.33	93.33	78.67
GravMAD (VLM)	56.72	2.79	100.00	58.67	70.67	80.00	61.33
Models	Push Buttons	Stack Blocks	Place Cups	Place Wine	Screw Bulb	Insert Peg	Stack Cups
Voxposer	80.00	16.00	6.67	5.33	4.00	0.00	1.33
Voxposer (Manual)	86.67	36.67	13.33	10.67	6.67	0.00	9.33
ChainedDiffuser (Oracle)	62.67	15.00	22.33	48.67	25.33	4.00	41.33
Act3D	66.67	0.00	0.00	45.33	6.67	0.00	0.00
3D Diffuser Actor	94.67	13.67	5.33	82.67	29.33	2.67	20.00
GravMAD (Manual)	98.67	56.67	5.33	77.33	66.67	32.00	57.33
GravMAD (VLM)	97.33	51.33	5.33	33.33	54.67	18.67	49.33

模型	平均成功率 ↑	平均排名 ↓	关闭罐子	打开抽屉	从烤架取肉	滑动积木	放入抽屉
Voxposer	15.11	5.88	12.00	10.67	45.33	0.00	0.00
Voxposer(手动)	22.06	4.92	13.33	18.67	69.33	0.00	0.00
ChainedDiffuser(Oracle)	29.72	4.42	82.67	0.00	52.00	2.67	0.00
Act3D	34.11	5.38	61.33	41.33	60.00	78.67	49.33
3D Diffuser Actor	55.81	3.00	66.67	88.00	88.00	84.00	94.67
GravMAD(手动)	69.17	1.63	100.00	76.67	89.33	93.33	78.67
GravMAD(VLM)	56.72	2.79	100.00	58.67	70.67	80.00	61.33
模型	按按钮	堆叠积木	放置杯子	放置酒杯	拧灯泡	插入销钉	堆叠杯子
Voxposer	80.00	16.00	6.67	5.33	4.00	0.00	1.33
Voxposer(手动)	86.67	36.67	13.33	10.67	6.67	0.00	9.33
ChainedDiffuser(Oracle)	62.67	15.00	22.33	48.67	25.33	4.00	41.33
Act3D	66.67	0.00	0.00	45.33	6.67	0.00	0.00
3D Diffuser Actor	94.67	13.67	5.33	82.67	29.33	2.67	20.00
GravMAD(手动)	98.67	56.67	5.33	77.33	66.67	32.00	57.33
GravMAD(VLM)	97.33	51.33	5.33	33.33	54.67	18.67	49.33

Table 7: Additional Multi-task test results on 12 base tasks.

表 7:12 个基础任务的额外多任务测试结果。

Models	Avg. Success	Avg. Rank	Close Drawer	Close Jar Banana	Close Jar Distractor	Condition Block	Meat On Grill	Open Drawer Small	Stack cups blocks	Push Buttons Light
Voxposer (Huang et al. 2023)	34.29	3.25	96.00	17.33	22.67	25.00	38.67	6.67	0.00	68.00
ChainedDiffuser (Oracle) (Xian et al., 2023)	43.22	2.75	84.33	82.67	85.00	48.00	29.00	0.00	41.33	30.00
Act3D [Gervet et al., 2023]	17.83	4.25	66.67	29.33	41.33	0.00	1.33	2.67	0.00	1.33
3D Diffuser Actor (Ke et al., 2024)	29.38	3.375	81.33	48.00	42.67	27.00	0.00	2.67	2.67	30.67
GravMAD (VLM)	62.92	1.125	97.33	84.00	86.67	74.00	45.33	21.33	18.67	76.00

模型	平均成功率	平均排名	关闭抽屉	关闭香蕉罐	关闭干扰罐	条件区块	烤架上的肉	打开小抽屉	堆叠杯子积木	按按钮灯
Voxposer (Huang 等, 2023)	34.29	3.25	96.00	17.33	22.67	25.00	38.67	6.67	0.00	68.00
ChainedDiffuser (Oracle) (Xian 等, 2023)	43.22	2.75	84.33	82.67	85.00	48.00	29.00	0.00	41.33	30.00
Act3D [Gervet 等, 2023]	17.83	4.25	66.67	29.33	41.33	0.00	1.33	2.67	0.00	1.33
3D Diffuser Actor (Ke 等, 2024)	29.38	3.375	81.33	48.00	42.67	27.00	0.00	2.67	2.67	30.67
GravMAD (视觉语言模型)	62.92	1.125	97.33	84.00	86.67	74.00	45.33	21.33	18.67	76.00

Table 8: Additional generalization results on 8 novel tasks.

表 8:8 个新任务的额外泛化结果。

D ADDITIONAL EXPERIMENTAL RESULTS

D 额外实验结果

D.1 INFERENCE TIME

D.1 推理时间

We test the inference time of all models under the setting of 8 novel tasks using a single NVIDIA 4090 GPU. The results, shown in Table 6 (in seconds), indicate the following: models like Act3D and 3D Diffuser Actor, which do not rely on foundation model inference, have shorter inference times but lower success rates. In contrast, Voxposer spends a significant amount of time synthesizing trajectories. Our GravMAD requires more time than Act3D and 3D Diffuser Actor because it waits for the foundation model to process information and infer sub-goals for sub-tasks.

我们在使用单个 NVIDIA 4090 GPU 的 8 个新任务设置下测试了所有模型的推理时间。结果如表 6 所示 (单位: 秒), 表明: 像 Act3D 和 3D Diffuser Actor 这类不依赖基础模型推理的模型推理时间较短, 但成功率较低。相比之下，Voxposer 在合成轨迹上花费了大量时间。我们的 GravMAD 推理时间比 Act3D 和 3D Diffuser Actor 更长，因为它需要等待基础模型处理信息并推断子任务的子目标。

D.2 ADDITIONAL BASELINE EXPERIMENTS

D.2 额外基线实验

We introduce two additional baseline methods for performance comparison: Voxposer (Manual) and Chained Diffuser (Xian et al. 2023) (Oracle). Voxposer (Manual) means that we manually provide ground truth object pose information to Voxposer instead of relying on the inference results of the foundation model. In Chained Diffuser (Oracle), we provide the ideal position for each keypose, with the connections between keyposes generated using the local trajectory diffuser module from Chained Diffuser. The performance comparisons of these two baseline methods on 12 base tasks and 8 novel tasks are shown in Table 7 and Table 8, respectively.

我们引入了两个额外的基线方法进行性能对比:Voxposer(手动) 和 Chained Diffuser(Xian 等,2023)(Oracle)。Voxposer(手动) 指的是我们手动向 Voxposer 提供真实的物体位姿信息，而非依赖基础模型的推理结果。在 Chained Diffuser(Oracle) 中，我们为每个关键位姿提供理想位置，关键位姿之间的连接由 Chained Diffuser 的局部轨迹扩散模块生成。这两个基线方法在 12 个基础任务和 8 个新任务上的性能对比分别见表 7 和表 8。

From the experimental results, we observe the following:

从实验结果中，我们观察到以下几点:

- In the base task setting, Voxposer (Manual) shows a slight performance improvement when provided with ground truth object information but still falls short compared to our GravMAD (Manual).

- 在基础任务设置中，当提供真实物体信息时，Voxposer(手动) 表现出轻微的性能提升，但仍不及我们的 GravMAD(手动)。

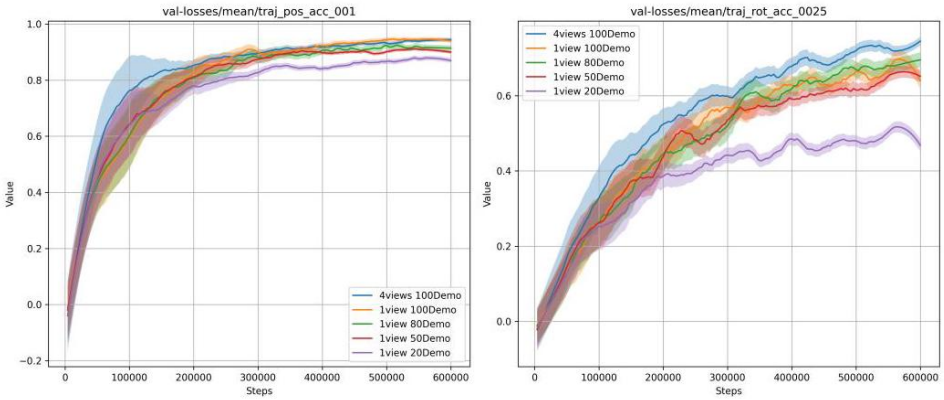


Figure 14: Comparison of validation curves under varying viewpoints and data sizes.

图 14: 不同视角和数据量下的验证曲线对比。

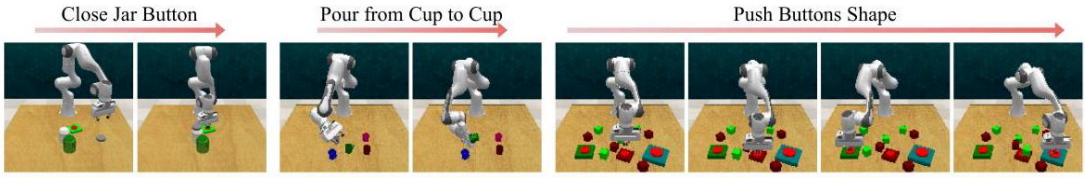


Figure 15: Visualization of additional novel tasks.

图 15: 额外新任务的可视化。

- For Chained Diffuser (Oracle), the keyposes come from ideal waypoints predefined in simulation, and the model effectively connects these keyposes, achieving a high success rate. However, in real-world scenarios, manually providing each keypose is impractical. Even with precise keyposes, Chained Diffuser (Oracle) still performs worse than our GravMAD (VLM).

- 对于 Chained Diffuser(Oracle)，关键位姿来自仿真中预定义的理想路径点，模型有效连接这些关键位姿，取得了较高的成功率。然而，在现实场景中，手动提供每个关键位姿是不切实际的。即使关键位姿精确，Chained Diffuser(Oracle) 的表现仍不及我们的 GravMAD(VLM)。

D.3 Scalability of GravMAD

D.3 GravMAD 的可扩展性

To evaluate the scalability of our proposed method with respect to data volume, we conduct training comparisons using five different demonstration dataset sizes and visualize the corresponding validation curves. The experimental results are presented in Fig. 14, with the validation curves reflecting two key metrics:

为了评估我们方法在数据量方面的可扩展性，我们使用五种不同示范数据集规模进行训练对比，并可视化相应的验证曲线。实验结果见图 14，验证曲线反映了两个关键指标:

1) The proportion of predicted positions in the validation set with an error less than 0.01 (left subplot in Fig. 14).

1) 验证集中预测位置误差小于 0.01 的比例 (图 14 左子图)。

2) The proportion of predicted rotations in the validation set with an error less than 0.025 (right subplot in Fig. 14).

2) 验证集中预测旋转误差小于 0.025 的比例 (图 14 右子图)。

The results in Fig. 14 clearly demonstrate that the model's performance improves as the number of expert demonstrations and the number of viewpoints increase. The key observations are as follows:

图 14 的结果清晰表明，随着专家示范数量和视角数量的增加，模型性能提升。主要观察点如下:

- With only 20 expert demonstrations, the model exhibits low overall performance, particularly in predicting rotation angles.
  - 仅使用 20 个专家示范时，模型整体表现较低，尤其是在预测旋转角度方面。
- Models trained with four viewpoints achieve significantly better performance, but this improvement comes at the cost of increased training time.
  - 使用四个视角训练的模型表现显著提升，但这种提升是以增加训练时间为代价的。
- As the number of expert demonstrations grows, the marginal improvement in model performance diminishes. This could be attributed to the model's parameter size not scaling proportionally with the increase in data volume.
  - 随着专家示范数量的增加，模型性能的边际提升逐渐减小。这可能是因为模型参数规模未能与数据量的增长成比例扩展。

These results highlight the benefits of larger datasets for enhancing model performance. However, they also underscore the need for further optimization in model architecture and resource allocation to effectively harness the potential of large-scale data. Without such improvements, the diminishing returns observed with increasing data may limit scalability in practical applications.

这些结果凸显了更大数据集在提升模型性能方面的优势，但同时也强调了在模型架构和资源分配上进一步优化的必要性，以有效利用大规模数据的潜力。若无此类改进，随着数据量增加而出现的收益递减可能限制实际应用中的可扩展性。

Task	Variation Type	#of Variations	Avg. Keyposes	Language Template
push button shape	color	20	2.0	”Press the buttons in order of their size, from smallest to largest”
button close jar	color	20	8.0	”after close the - jar, push the button”
pour from cup to cup	color	20	6.0	”pour liquid from the - cup to the - cup”

任务	变体类型	变体数量	平均关键姿势	语言模板
按钮形状	颜色	20	2.0	“按按钮的大小顺序，从最小到最大依次按下”
关闭罐子的按钮	颜色	20	8.0	“关闭罐子后，按下按钮”
从杯子倒到杯子	颜色	20	6.0	“将液体从 - 杯倒入 - 杯”

Table 9: Description of Additional Novel Tasks.

表 9: 额外新颖任务的描述。

Additional Novel Task			Voxposer Act3D 3D Diffuser Actor GravMAD (Ours)	
Push Buttons Shape (Difficult Task)	0	0	0	62.66
Button Close Jar (Combination of Skills)	0	0	0	0
Pour From Cup to Cup (Completely New)	0	0	0	0

额外的新任务			Voxposer Act3D 三维扩散器角色 GravMAD(我们的)	
按按钮形状 (困难任务)	0	0	0	62.66
关闭罐子按钮 (技能组合)	0	0	0	0
从杯子倒到杯子 (全新任务)	0	0	0	0

Table 10: Generalization Performance Comparison on Additional Novel Tasks.

表 10: 额外新任务上的泛化性能比较。

D.4 ADDITIONAL NOVEL TASKS

D.4 额外新任务

We evaluate the performance of baseline methods and GravMAD on three additional novel tasks, with detailed descriptions provided in Table 9 and Fig. 15 These tasks include a highly challenging one (Push Buttons Shape), a task that requires integrating skills learned during training (Button Close Jar), and a task involving entirely new objects compared to the training set (Pour From Cup to Cup).

我们在三个额外的新任务上评估了基线方法和 GravMAD 的性能，详细描述见表 9 和图 15。这些任务包括一个极具挑战性的任务 (按键形状)，一个需要整合训练中学到技能的任务 (按钮关闭罐子)，以及一个涉及与训练集完全不同新物体的任务 (从杯子倒到杯子)。

The results are presented in Table 10 . The ”Push Buttons Shape” task evaluates the model’s ability to handle long-horizon planning, language reasoning, and robustness to visual perturbations. Under these conditions, all baseline methods fail to complete the task, whereas GravMAD performs well, showcasing its potential for generalization. For the ”Button Close Jar” task, the results indicate that GravMAD still struggles with long-horizon tasks requiring the integration of multiple skills. In the entirely new task ”Pour From Cup to Cup”, GravMAD successfully identifies task-relevant objects but fails to complete the task due to incorrect actions. This failure is likely caused by a significant mismatch between the training data and the test environment.

结果见表 10。“按键形状”任务评估模型处理长时序规划、语言推理及对视觉扰动的鲁棒性能力。在这些条件下，所有基线方法均未能完成任务，而 GravMAD 表现良好，展示了其泛化潜力。对于“按钮关闭罐子”任务，结果表明 GravMAD 在需要整合多项技能的长时序任务上仍存在困难。在完全新颖的“从杯子倒到杯子”任务中，GravMAD 成功识别了任务相关物体，但因动作错误未能完成任务。这一失败很可能是训练数据与测试环境之间存在显著不匹配所致。

D.5 ADDITIONAL ABLATION STUDY

D.5 额外消融研究

To investigate the impact of the cost map on model performance, we perform more detailed experiments on the ”w/o Cost map” ablation setting. In this ablation study, due to the inherent limitations of the encoder, the GravMap containing only the gripper map cannot be effectively processed. For instance, when the sub-goal requires the robotic arm to perform a ”close everywhere” operation, m_q becomes a zero structure. Such an m_g cannot be properly parsed by the DP3 Encoder, resulting in gradient vanishing during the training process. To address this issue, we modify the gripper map in the ”w/o Cost map” setting by changing the closed state representation from 0 to -1, enabling the encoder to correctly process this data structure. The experimental results are shown in Fig. 16. The results show that removing the cost map causes a significant performance drop compared to the original model: a decrease of 11.97% on 12 base tasks and 21.04% on 8 novel tasks. These findings clearly highlight the critical role of the cost map in ensuring the performance of the GravMAD model.

为探究代价图对模型性能的影响，我们在“无代价图”消融设置下进行了更详细的实验。在该消融研究中，由于编码器的固有限制，仅包含夹爪图的 GravMap 无法被有效处理。例如，当子目标要求机械臂执行“全闭合”操作时，m_q 变成了零结构。这样的 m_g 无法被 DP3 编码器正确解析，导致训练过程中梯度消失。为解决此问题，我们在“无代价图”设置中将夹爪图的闭合状态表示由 0 改为-1，使编码器能够正确处理该数据结构。实验结果见图 16。结果显示，移除代价图相比原模型导致性能显著下降:12 个基础任务下降 11.97%，8 个新任务下降 21.04%。这些发现清晰地凸显了代价图在保障 GravMAD 模型性能中的关键作用。

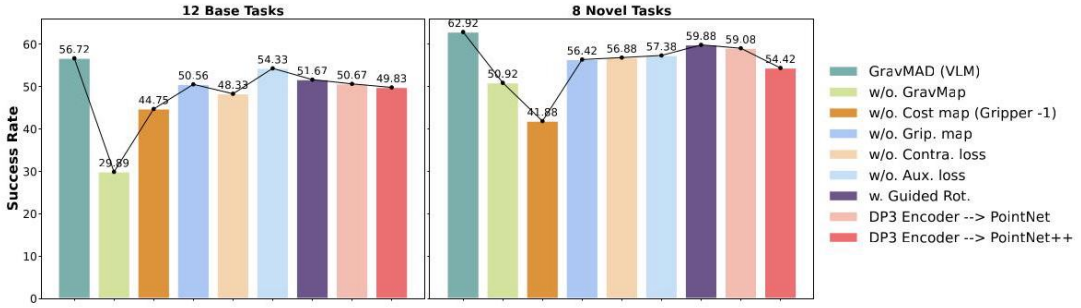
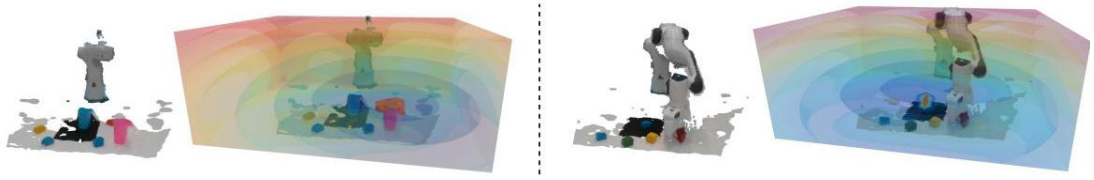


Figure 16: Additional Ablation Studies. We represent the gripper closure in the gripper map under ”w/o. Cost map” as -1 instead of 0, enabling the encoder to correctly process this data structure.

图 16: 额外消融研究。我们在“无代价图”设置下将夹爪图中的夹爪闭合状态表示为-1 而非 0，使编码器能够正确处理该数据结构。

D.6 REAL WORLD EVALUATION

D.6 真实环境评估



Real-world Task	Open Drawer	Toy in Drawer	Mouse on Pad	Stack Cup	Stack Block Same
GravMAD (%)	80	90	100	60	50
Real-world Task	Place Cup	Stack Block	Stack Cup Blocks	Wired Mouse on Pad	Colored Toy in Drawer
GravMAD (%)	10	40	40	100	70

现实任务	打开抽屉	抽屉里的玩具	鼠标在鼠标垫上	叠杯子	叠放相同积木
GravMAD(%)	80	90	100	60	50
现实任务	放置杯子	叠积木	叠杯子积木	有线鼠标在鼠标垫上	抽屉里的彩色玩具
GravMAD(%)	10	40	40	100	70

Table 11: Real-robot Results. Success rates of GravMAD on 10 real-world tasks. These tasks include both manipulation and placement challenges. Above the table are the point clouds and GravMaps for Stack Cup Blocks and Stack Block, respectively.

表 11: 真实机器人结果。GravMAD 在 10 个真实世界任务中的成功率。这些任务包括操作和放置挑战。表格上方分别是堆叠杯子积木和堆叠积木的点云和 GravMaps。

We use a Franka Emika robot to validate Grav-MAD's multi-task generalization ability across 10 real-world tasks. Each task involves variations in placement, and some tasks include color variations. Compared to the base tasks, the novel tasks introduce new objects and new instructions. The base tasks include:

我们使用 Franka Emika 机器人验证 Grav-MAD 在 10 个真实世界任务中的多任务泛化能力。每个任务涉及放置的变化，有些任务还包括颜色变化。与基础任务相比，新任务引入了新物体和新指令。基础任务包括:

Figure 17: Real-Robot Setup with RealSense D435i and Franka Panda.

图 17: 配备 RealSense D435i 和 Franka Panda 的真实机器人设置。

- Open Drawer (task description: open top drawer)
  - 打开抽屉 (任务描述: 打开顶层抽屉)
- Place Cup (task description: put the yellow toy in the top drawer)
  - 放置杯子 (任务描述: 将黄色玩具放入顶层抽屉)
- Mouse on Pad (task description: put the wireless mouse on pad)
  - 鼠标放垫上 (任务描述: 将无线鼠标放在鼠标垫上)
- Stack Cup (task description: stack color1 cup on top of color2 cup)
  - 堆叠杯子 (任务描述: 将 color1 杯子堆叠在 color2 杯子上)
- Stack Block Same (task description: stack blocks with the same color)
  - 堆叠同色积木 (任务描述: 堆叠颜色相同的积木)
- Place Cup (task description: place one cup on the cup holder)
  - 放置杯子 (任务描述: 将一个杯子放在杯架上)

The novel tasks involve:

新任务包括:

- Stack Block (task description: stack color1 block on top of color2 block)
  - 堆叠积木 (任务描述: 将 color1 积木堆叠在 color2 积木上)
- Stack Cup Blocks (task description: identify the most common color in the block pile, and stack the other cups on the cup that matches that color)
  - 堆叠杯子积木 (任务描述: 识别积木堆中最常见的颜色，并将其他杯子堆叠在与该颜色匹配的杯子上)
- Wired Mouse on Pad (task description: put the wired mouse on pad)
  - 有线鼠标放垫上 (任务描述: 将有线鼠标放在鼠标垫上)

- Colored Toy in Drawer (task description: put the Black and white toy in the top drawer)

- 彩色玩具放抽屉 (任务描述: 将黑白玩具放入顶层抽屉)

We position a RealSense D435i camera in front of the robot to capture images, which are downsampled from the original resolution of 1280×720 to 256×256 , as shown in fig. 17 During training, we collect

我们将 RealSense D435i 摄像头置于机器人前方采集图像，图像从原始分辨率 1280×720 下采样至 256×256 ，如图 17 所示。训练期间，我们收集

20 demonstrations for each base task to train the model. During inference, similar to the simulation setup, GravMAD predicts the next keypose, and we use the BiRRT planner provided by MoveIt! ROS to guide the robot to reach the predicted keypose. For evaluation, we run 10 episodes for each task and report the success rate.

每个基础任务 20 次示范以训练模型。推理时，与仿真设置类似，GravMAD 预测下一个关键姿态，我们使用 MoveIt! ROS 提供的 BiRRT 规划器引导机器人达到预测的关键姿态。评估时，每个任务运行 10 次，报告成功率。

The inference performance of GravMAD on 6 base tasks and 4 novel tasks is shown in Table 11 These results demonstrate that GravMAD can effectively reason about 3D manipulation tasks in real-world robotic scenarios, leveraging associated visual information and generalizing to novel tasks. The video demonstrations are available at: https://gravmad.github.io

GravMAD 在 6 个基础任务和 4 个新颖任务上的推理性能如表 11 所示。这些结果表明，GravMAD 能够有效地推理现实机器人场景中的三维操作任务，利用相关视觉信息并推广到新任务。视频演示可见于:https://gravmad.github.io

E LIMITATIONS AND POTENTIAL SOLUTIONS

限制与潜在解决方案

Despite GravMAD demonstrating strong generalization capabilities across the 3 categories and 8 novel tasks showcased, it still has certain limitations. The following section discusses some of the limitations not covered in the main text and their potential solutions:

尽管 GravMAD 在展示的 3 个类别和 8 个新颖任务中表现出强大的泛化能力，但仍存在一定的局限性。以下部分讨论了主文中未涉及的一些限制及其潜在解决方案:

- Limitations of heuristic Sub-goal Keypose Discovery: The current method relies on predefined heuristic rules, which may struggle to adapt to tasks with more complex or ambiguous sub-goal structures. Future research could explore more adaptive or learning-based strategies, such as incorporating diffusion models (Black et al. 2024) or generative models (Shridhar et al. 2024) to generate sub-goals, to further enhance the robustness and flexibility of the method.

- 启发式子目标关键姿态发现的局限性: 当前方法依赖预定义的启发式规则, 可能难以适应具有更复杂或模糊子目标结构的任务。未来研究可探索更具适应性或基于学习的策略，如引入扩散模型 (Black et al. 2024) 或生成模型 (Shridhar et al. 2024) 来生成子目标，以进一步增强方法的鲁棒性和灵活性。

- Dependence on Detector accuracy and inference time: The Detector's accuracy during the inference phase has a significant impact on the results, and its relatively long inference time remains a bottleneck. Future work could integrate observations from multiple viewpoints to provide a more comprehensive scene understanding and improve detection accuracy. Alternatively, more granular segmentation models could be leveraged to provide richer labels for foundation models, thereby improving the quality of the context generated by the Detector.

- 对检测器准确性和推理时间的依赖: 检测器在推理阶段的准确性对结果影响显著, 其相对较长的推理时间仍是瓶颈。未来工作可整合多视角观测以提供更全面的场景理解并提升检测准确率。或者，可利用更细粒度的分割模型为基础模型提供更丰富的标签，从而提升检测器生成上下文的质量。

- Limited guidance for end-effector orientation: The current GravMap framework does not effectively guide the robot's end-effector orientation, limiting its applicability to tasks requiring precise orientation control. A potential improvement involves combining rotation information from expert trajectories with object poses to generate few-shot prompts for off-the-shelf foundation models (Yin et al. 2024). By leveraging such few-shot prompts, foundation models could produce more precise and effective rotation guidance.

- 末端执行器姿态指导有限: 当前 GravMap 框架未能有效指导机器人末端执行器的姿态，限制了其在需要精确姿态控制任务中的适用性。潜在改进方案是结合专家轨迹中的旋转信息与物体姿态，为现成基础模型生成少样本提示 (Yin et al. 2024)。通过利用此类少样本提示，基础模型可产生更精确有效的旋转指导。

- Challenges in generalization: While GravMAD performs exceptionally well on tasks similar to those seen during training, its generalization ability is still limited for tasks with significant differences from the training set, such as entirely unseen tasks or challenging tasks requiring a combination of multiple learned skills. Expanding GravMAD's capability to flexibly integrate multiple learned skills will be a key direction for future research. One feasible direction is to combine exploration-based learning with reinforcement learning (Hao et al., 2024).

- 泛化挑战: 虽然 GravMAD 在与训练任务相似的任务上表现优异，但对于与训练集差异显著的任务 (如全新任务或需多种已学技能组合的复杂任务), 其泛化能力仍有限。扩展 GravMAD 灵活整合多种已学技能的能力将是未来研究的关键方向。一种可行方向是结合基于探索的学习与强化学习 (Hao et al., 2024)。

- Dependence on GravMap for Sub-goal Representation: The GravMap framework relies on point cloud structures for sub-goal representation, which, while effective, may add unnecessary complexity in scenarios where simpler representations, such as a single point or relative coordinates, could suffice. The competitive performance of the "w/o GravMap" variant on novel tasks suggests that alternative representations could simplify the model without compromising performance. Defining sub-goals as relative coordinates with respect to

the gripper's current position, leveraging proprioceptive information, is a promising direction. This approach could possibly introduce more data variation, enhance adaptability to spatial changes, handle imprecise subgoals, and naturally encode directional information. Future research could explore this direction further to achieve a balance between simplicity and performance, potentially enhancing the generalization capability of the model while reducing reliance on GravMap.

- 对 GravMap 子目标表示的依赖:GravMap 框架依赖点云结构进行子目标表示,虽有效,但在某些场景下可能增加不必要的复杂度, 而简单表示如单点或相对坐标即可满足需求。“无 GravMap”变体在新任务上的竞争性表现表明，替代表示可简化模型且不损失性能。将子目标定义为相对于夹持器当前位置的相对坐标，利用本体感知信息，是一个有前景的方向。该方法可能引入更多数据变异性，增强对空间变化的适应性，处理不精确子目标，并自然编码方向信息。未来研究可进一步探索此方向，以实现简洁性与性能的平衡，潜在提升模型泛化能力并减少对 GravMap 的依赖。

By addressing these limitations, we anticipate that GravMAD will demonstrate stronger adaptability and practical value in more diverse tasks.

通过解决这些限制，我们预期 GravMAD 将在更多样化任务中展现更强的适应性和实用价值。