# Spatial-Temporal Graph Diffusion Policy with Kinematic Modeling for Bimanual Robotic Manipulation

基于运动学建模的双手机器人操作时空图扩散策略

Qi Hao Xiang Deng Rui Shao Yinchuan Jianye

齐 郝 邓翔 邵锐 尹川 建业

Longxiang Michael Yu Wang Liqiang Nie

龙翔 Michael Yu Wang 聂立强

Harbin Institute of Technology (Shenzhen) Greate Bay University

哈尔滨工业大学(深圳) 大湾区大学

Huawei Noah’s Ark Lab Shandong Computer Science Center

华为诺亚方舟实验室 山东省计算中心

lvqi@stu.hit.edu.cn, dengxiang@hit.edu.cn

lvqi@stu.hit.edu.cn, dengxiang@hit.edu.cn

# Abstract

# 摘要

Despite the significant success of imitation learning in robotic manipulation, its application to bimanual tasks remains highly challenging. Existing approaches mainly learn a policy to predict a distant next-best end-effector pose (NBP) and then compute the corresponding joint rotation angles for motion using inverse kinematics. However, they suffer from two important issues: (1) rarely considering the physical robotic structure, which may cause self-collisions or interferences, and (2) overlooking the kinematics constraint, which may result in the predicted poses not conforming to the actual limitations of the robot joints. In this paper, we propose Kinematics enhanced Spatial-TemporAl gRaph Diffuser (KStar Diffuser). Specifically, (1) to incorporate the physical robot structure information into action prediction, KStar Diffuser maintains a dynamic spatial-temporal graph according to the physical bimanual joint motions at continuous timesteps. This dynamic graph serves as the robot-structure condition for denoising the actions; (2) to make the NBP learning objective consistent with kinematics, we introduce the differentiable kinematics to provide the reference for optimizing KStar Diffuser. This module regularizes the policy to predict more reliable and kinematics-aware next end-effector poses. Experimental results show that our method effectively leverages the physical structural information and generates kinematics-aware actions in both simulation and real-world.

尽管模仿学习在机器人操作中取得了显著成功，但其在双手任务中的应用仍然极具挑战性。现有方法主要通过学习策略来预测下一个最佳末端执行器姿态(NBP)，然后通过逆运动学计算相应的关节旋转角度以进行运动。然而，这些方法存在两个重要问题:(1)很少考虑物理机器人结构，可能导致自碰撞或干扰；(2)忽视运动学约束，可能导致预测的姿态不符合机器人关节的实际限制。在本文中，我们提出了基于运动学增强的时空图扩散器(KStar Diffuser)。具体而言，(1)为了将物理机器人结构信息纳入动作预测，KStar Diffuser根据连续时间步长的物理双手关节运动维护一个动态时空图。该动态图作为机器人结构条件用于去噪动作；(2)为了使NBP学习目标与运动学一致，我们引入了可微运动学，为优化KStar Diffuser提供参考。该模块正则化策略，以预测更可靠且具有运动学意识的下一末端执行器姿态。实验结果表明，我们的方法有效利用了物理结构信息，并在仿真和实际环境中生成了具有运动学意识的动作。

# 1. Introduction

# 1. 引言

Bimanual manipulation represents a fundamental capability for robotic systems to perform complex tasks requiring two-arm coordination. While imitation learning has demonstrated remarkable success in single-arm manipulation , its extension to bimanual scenarios faces unique challenges as robots need to coordinate dual-arm movements while conforming to physical constraints. These challenges significantly impact the reliability and feasibility of predicted actions in real-world applications.

双手操作 代表了机器人系统执行需要双臂协调的复杂任务的基本能力。尽管模仿学习在单臂操作 中取得了显著成功，但其扩展到双手场景面临独特挑战，因为机器人需要协调双臂运动，同时遵守物理约束。这些挑战显著影响了预测动作在实际应用中的可靠性和可行性。

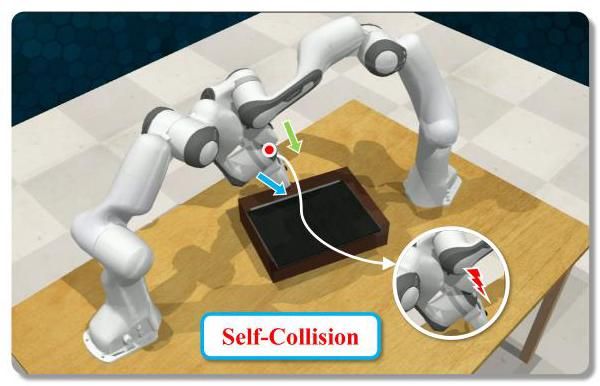


Figure 1. The self-collision problem in bimanual manipulation tasks due to overlooking the robotic structure.

图1. 由于忽视机器人结构而导致的双手操作任务中的自碰撞问题。

A straightforward approach would be to learn the entire motion trajectory by predicting joint positions at consecutive timesteps. However, this poses significant challenges as trajectories are typically long and difficult to learn. Therefore, mainstream approaches [15, 17, 35, 49] typically adopt a two-stage pipeline: first predicting the next-best end-effector pose (NBP), then computing joint rotations through inverse kinematics [21, 26]. However, the dominant NBP-based approaches, though simplify the learning process, often lead to unreliable motion generation. Concretely, the predicted poses may violate physical structure constraints , causing potential arm collisions, or exceeding joint limits due to inadequate consideration of kinematic feasibility, as Figure 1 illustrated.

一种直接的方法 是通过预测连续时间步长的关节位置来学习整个运动轨迹。然而，这带来了重大挑战，因为轨迹通常较长且难以学习。因此，主流方法[15, 17, 35, 49]通常采用两阶段流程:首先预测下一个最佳末端执行器姿态(NBP)，然后通过逆运动学[21, 26]计算关节旋转。然而，基于NBP的主导方法虽然简化了学习过程，但往往导致不可靠的运动生成。具体而言，预测的姿态可能违反物理结构约束 ，导致潜在的手臂碰撞，或由于对运动学可行性的考虑不足而超出关节限制，如图1所示。

In real-world bimanual manipulation tasks, existing approaches exhibit significant limitations in motion execution despite their promising performance in trajectory prediction. Empirical observations indicate that while individual end-effector poses are feasible in isolation, their concurrent execution frequently results in inter-arm collisions [48] and kinematically infeasible configurations [35, 39]. The fundamental challenge lies in the optimization of end-effector poses exclusively in Cartesian space, which introduces a substantial discrepancy between motion prediction and physical constraints. Existing approaches incorporate proprioception information in a simplistic manner, (e.g., gripper rotation angles [17, 49] and end-effector poses [53]), but disregarding the robot’s intrinsic structural properties e.g., the kinematic chain and joint configurations during the prediction phase. This oversimplified representation fails to capture the crucial spatial correlations between robotic arms, joint links, and their potential interactions. Consequently, the predicted dual-arm trajectories may satisfy task objectives in Cartesian space while violating crucial spatial constraints during execution.

在实际的双臂操作任务中，现有方法 在轨迹预测方面表现出色，但在运动执行方面存在显著局限性。实证观察表明，尽管单个末端执行器的姿态在孤立情况下是可行的，但它们的同时执行经常导致双臂碰撞[48]和运动学上不可行的配置[35, 39]。根本挑战在于仅在笛卡尔空间中优化末端执行器姿态，这导致了运动预测与物理约束之间的显著差异。现有方法 以简单的方式整合了本体感知信息(例如，夹爪旋转角度[17, 49]和末端执行器姿态[53])，但在预测阶段忽略了机器人的内在结构特性，例如运动链和关节配置。这种过度简化的表示未能捕捉到机械臂、关节连杆及其潜在相互作用之间的关键空间关联。因此，预测的双臂轨迹可能在笛卡尔空间中满足任务目标，但在执行过程中违反了关键的空间约束。

[[1]](#footnote-28)

Moreover, the conventional paradigm of treating inverse kinematics as a post-processing procedure introduces additional complexities. The disjunction between kinematic constraints and the pose prediction learning objective often yields solutions that exhibit apparent smoothness in Cartesian space while manifesting discontinuous or mechanically infeasible trajectories in the joint space. This limitation becomes particularly pronounced in configurations approaching kinematic singularities or joint limits, where the mapping between Cartesian and joint spaces becomes ill-conditioned. These inherent limitations necessitate the integration of both structural and kinematic constraints into the policy learning framework.

此外，将逆运动学视为后处理过程的传统范式 引入了额外的复杂性。运动学约束与姿态预测学习目标之间的脱节往往导致在笛卡尔空间中表现出明显平滑性，而在关节空间中表现出不连续或机械上不可行的轨迹。这种局限性在接近运动学奇点或关节极限的配置中尤为明显，其中笛卡尔空间与关节空间之间的映射变得病态。这些固有的局限性要求将结构和运动学约束整合到策略学习框架中。

To address the above issues, we propose Kinematics enhanced Spatial-TemporAl gRaph Diffuser (KStar Diffuser), a novel framework that explicitly incorporates both robot structures and kinematics into the bimanual motion generation process. Our key insight is that the physical structure and kinematic properties of the robot should guide the learning process of pose prediction, rather than be treated as independent post-processing constraints. Specifically, (1) to incorporate structural awareness, we construct a dynamic spatial-temporal graph from the robot’s URDF specifications, where nodes represent joint properties and edges capture both spatial relations and temporal dependencies. This graph structure is encoded via Graph Convolutional Network (GCN) [25] to provide explicit physical constraints for the diffusion process; (2) for kinematic feasibility, we regularize the NBP learning objective by incorporating joint-space prediction, where the predicted joint positions are mapped to reference end-effector poses through differentiable forward kinematics. These kinematically-feasible poses then serve as conditions to guide the diffusion process, ensuring that the generated motions satisfy both structural and kinematic constraints. Our main contributions are as follows:

为了解决上述问题，我们提出了运动学增强的时空图扩散器(KStar Diffuser)，这是一种新颖的框架，明确地将机器人结构和运动学整合到双臂运动生成过程中。我们的关键见解是，机器人的物理结构和运动学特性应指导姿态预测的学习过程，而不是被视为独立的后处理约束。具体来说，(1)为了整合结构意识，我们从机器人的URDF规范中构建了一个动态时空图，其中节点表示关节特性，边捕捉空间关系和时间依赖性。该图结构通过图卷积网络(GCN)[25]进行编码，为扩散过程提供明确的物理约束；(2)为了运动学可行性，我们通过引入关节空间预测来正则化NBP学习目标，其中预测的关节位置通过可微的前向运动学映射到参考末端执行器姿态。这些运动学上可行的姿态随后作为条件来指导扩散过程，确保生成的运动满足结构和运动学约束。我们的主要贡献如下:

* Different from existing approaches that optimize end-effector poses solely in Cartesian space, we propose a novel spatial-temporal robot graph that explicitly models the robot physical configuration to guide the generative action denoising procedure.
* 与现有方法仅在笛卡尔空间中优化末端执行器姿态不同，我们提出了一种新颖的时空机器人图，明确建模机器人物理配置以指导生成动作去噪过程。
* We introduce a kinematics regularizer that augments the NBP learning objective by introducing joint-space supervision. This regularizer leverages forward kinematics to provide kinematically-feasible reference poses, effectively guiding the diffusion process to conform to kinematic constraints.
* 我们引入了一种运动学正则器，通过引入关节空间监督来增强NBP学习目标。该正则器利用前向运动学提供运动学上可行的参考姿态，有效指导扩散过程符合运动学约束。
* Extensive experiments show that our proposed KStar Diffuser is superior in both simulation and real-world scenarios, surpassing baselines more than 10% in success rate.
* 大量实验表明，我们提出的KStar Diffuser在模拟和实际场景中均表现出色，成功率超过基线10%以上。

# 2. Related Work

# 2. 相关工作

Diffusion Models in Bimanual Robotic Controls. Diffusion models such as Denoising Diffusion Probabilistic Models (DDPM) [19] have achieved great success in the fields of image generation and video generation [41, 55, 66]. Thus, recent work [8, 45, 52, 57, 67] has been devoted to applying its powerful generation capabilities to action generation for robotic manipulation tasks in a imitation learning mode. With the 2D images [9, 30- or point cloud observations , the policy is trained to output the action sequence of joint positions or end-effector poses by iterative denoising process.

双臂机器人控制中的扩散模型。诸如去噪扩散概率模型(DDPM)[19]等扩散模型在图像生成 和视频生成[41, 55, 66]领域取得了巨大成功。因此，最近的工作[8, 45, 52, 57, 67]致力于将其强大的生成能力应用于模仿学习模式下的机器人操作任务的动作生成。通过2D图像[9, 30- 或 点云观测 ，策略被训练为通过迭代去噪过程输出关节位置或末端执行器姿态的动作序列。

Bimanual manipulation tasks, which resemble humanlike actions more closely, have garnered increasing attention [10, 14, 17, 33, 65]. However, existing research primarily extends single-arm manipulation methods to dual-arm tasks , without considering about the distinct challenges specific to bimanual settings. The relative independence of the two arms makes it essential to consider self-collision avoidance when predicting actions. In practice, the robot’s structure significantly influences its motion since the joint types determine movement directions and joint angles constrain the range of motion. Although some studies integrate proprioception in robotic control, they often use simply low-dimensional vectors to represent body information, lacking depth in structural exploration and overlooking critical spatial relationships between dual arms. Therefore, we study the physical structure information of the dual-arm robot system, and improve both the accuracy and adaptability of robotic movements.

双手操作任务，因其更接近人类动作，已引起越来越多的关注[10, 14, 17, 33, 65]。然而，现有研究主要将单臂操作方法扩展到双臂任务 ，而未考虑双手操作特有的挑战。双臂的相对独立性使得在预测动作时，必须考虑自碰撞避免 。实际上，机器人的结构对其运动有显著影响，因为关节类型决定了运动方向，而关节角度限制了运动范围。尽管一些研究 在机器人控制中整合了本体感知，但它们通常仅使用低维向量来表示身体信息，缺乏对结构的深入探索，并忽视了双臂之间的关键空间关系。因此，我们研究了双臂机器人系统的物理结构信息，并提高了机器人运动的准确性和适应性。

Robotic Kinematics Modeling. Robotic kinematics modeling is a classic problem in the robotic control where inverse kinematics (IK) presents a fundamental challenge, i.e., deriving the joint configuration given the end-effector pose. The inherent non-uniqueness of IK solutions as a significant obstacle for learning algorithms , often produce inaccurate models by averaging over nonconvex feasible sets. Recent advancements integrate neural solutions to enhance the adaptability and precision of kinematic models. Bócsi et al. [5] leveraged support vector machines to parameterize quadratic programs, aligning solutions with IK in specific workspace regions. Current data-driven imitation learning approaches primarily rely on probabilistic modeling of action distributions to predict subsequent actions, but lack the reliability guarantees from the aspect of kinematics. In this paper, we construct a spatial-temporal graph to learn the robot physical representation, and provide a kinematics-awared latent end-effect embedding for diffusion policy as guidance. In this way, o "ur approach enhances the robot’s ability to perform tasks with superious precision and adaptability.

机器人运动学建模。机器人运动学建模 是机器人控制 中的一个经典问题，其中逆运动学(IK)提出了一个基本挑战，即给定末端执行器姿态推导关节配置。IK解的内在非唯一性对学习算法 构成了重大障碍，通常通过对非凸可行集进行平均来产生不准确的模型。最近的进展整合了神经解决方案 以增强运动学模型的适应性和精度。Bócsi等人[5]利用支持向量机对二次程序进行参数化，使解与特定工作空间区域中的IK对齐。当前数据驱动的模仿学习方法 主要依赖于动作分布的概率建模来预测后续动作，但缺乏从运动学角度的可靠性保证。在本文中，我们构建了一个时空图来学习机器人物理表示，并为扩散策略提供了一个运动学感知的潜在末端效应嵌入作为指导。通过这种方式，我们的方法增强了机器人执行任务的精确性和适应性。

# 3. Method

# 3. 方法

# 3.1. Preliminary

# 3.1. 预备知识

Diffusion Policy. Chi et al. [9] propose the diffusion policy which represents a robot’s visuomotor policy as a conditional denoising diffusion process. It learns a model distribution conditioned on observation to approximate joint distribution . The whole procedure consists of a forward process and a reverse process.

扩散策略。Chi等人[9]提出了扩散策略，将机器人的视觉运动策略表示为条件去噪扩散过程。它学习一个以观察 为条件的模型分布 来近似联合分布 。整个过程包括前向过程和反向过程。

(1) Forward Process: For the Markov chain with Gaussian transitions parameterized, the policy learns a fixed inference procedure :

(1) 前向过程:对于参数化为高斯转移的马尔可夫链，策略 学习一个固定的推理过程 :

Thus, can be expressed as a linear combination of and a noise variable :

因此， 可以表示为 和噪声变量 的线性组合:

The training loss is obtained:

得到训练损失:

where is the random noise sampling at iteration.

其中 是在 次迭代时采样的随机噪声。

(2) Reverse Process: Starting from a sample , the reverse steps are:

(2) 反向过程:从样本 开始，反向步骤如下:

By the iterative denoising process, the policy generates as its next action. The continuous predicted action sequence forms the complete action trajectory.

通过迭代去噪过程，策略生成 作为其下一个动作。连续的预测动作序列形成了完整的动作轨迹。

Robotic Kinematics. Robotics kinematics describes the relationship between the robot joints and its end-effector. It can be divided into forward kinematics and inverse kinematics for controlling robot movements.

机器人运动学。机器人运动学描述了机器人关节与其末端执行器之间的关系。它可以分为正向运动学和逆运动学，用于控制机器人运动。

(1) Forward Kinematics (FK): Given a specific configuration of joint angles, i.e., , it aims to construct a function , computing the end-effector position and orientation that are represented as a pose matrix .

(1) 正向运动学(Forward Kinematics, FK):给定一组特定的关节角度配置，即 ，其目标是构建一个函数 ，计算表示为位姿矩阵 的末端执行器位置和方向。

where is the homogeneous transformation matrix for joint , and means the task space. denotes the configuration space, while the configuration refers to the specific arrangement of the robotic joints in a given pose.

其中 是关节 的齐次变换矩阵， 表示任务空间。 表示配置空间，而配置指的是机器人关节在给定位姿中的特定排列。

(2) Inverse Kinematics (IK): In contrast, IK aims to find the joint configurations which achieve a desired end-effector pose. It is defined as a mapping from the task space to the set of possible configurations that satisfy the given pose matrix :

(2) 逆向运动学(Inverse Kinematics, IK):与之相反，IK旨在找到实现期望末端执行器位姿的关节配置。它被定义为从任务空间 到满足给定位姿矩阵 的可能配置集 的映射 :

It is noted that this mapping is often non-unique, especially for redundant manipulators with more than .

需要注意的是，这种映射通常不是唯一的，特别是对于具有超过 个自由度的冗余机械臂。

# 3.2. Task Definition

# 3.2. 任务定义

Given the dataset with language instructions and the RGB-D observations , we aim to learn a policy which predicts actions . Here, is composed of a trajectory and gripper opening or closing action , where denotes the trajectory length and represents the number of robot joints. In the bimanual manipulation, is usually to be 12 or 14 as each robot arm has 6 or 7-DoF. Referring to prior works , it is inefficient to train the policy on all trajectory points. Thus, a keyframe discovery method is used to extract a set of keyframe indices and the prediction actions are the key set of end-effector poses .

给定包含语言指令 和RGB-D观测 的数据集 ，我们的目标是学习一个策略 ，该策略预测动作 。这里， 由轨迹 和夹爪开合动作 组成，其中 表示轨迹长度， 表示机器人关节的数量。在双手操作中， 通常为12或14，因为每个机械臂具有6或7个自由度。参考先前的工作 ，在所有轨迹点上训练策略是低效的。因此，使用关键帧发现方法提取一组 关键帧索引 ，预测动作是关键末端执行器位姿集 。

# 3.3. KStar Diffuser

# 3.3. KStar扩散器

# 3.3.1. Overview

# 3.3.1. 概述

Mainstream methods train a policy to predict actions, but take little consideration about the mechanical robot structure which determines its motion. We thus propose a spatial-temporal graph to model both the static physical structure and dynamic history movement information. In addition, to reduce kinematically infeasible predictions for the end-effector pose, we introduce a differentiable kinematics module that provides kinematics-aware references to the policy network. The overview of our proposed Kinematics enhanced Spatial-TemporAl gRaph Diffuser (KStar Diffuser) is shown in Figure B.

主流方法 训练策略来预测动作，但很少考虑决定其运动的机械机器人结构。因此，我们提出了一个时空图来建模静态物理结构和动态历史运动信息。此外，为了减少对末端执行器位姿的运动学不可行预测，我们引入了一个可微运动学模块，为策略网络提供运动学感知参考。我们提出的运动学增强时空图扩散器(KStar Diffuser)的概述如图B所示。

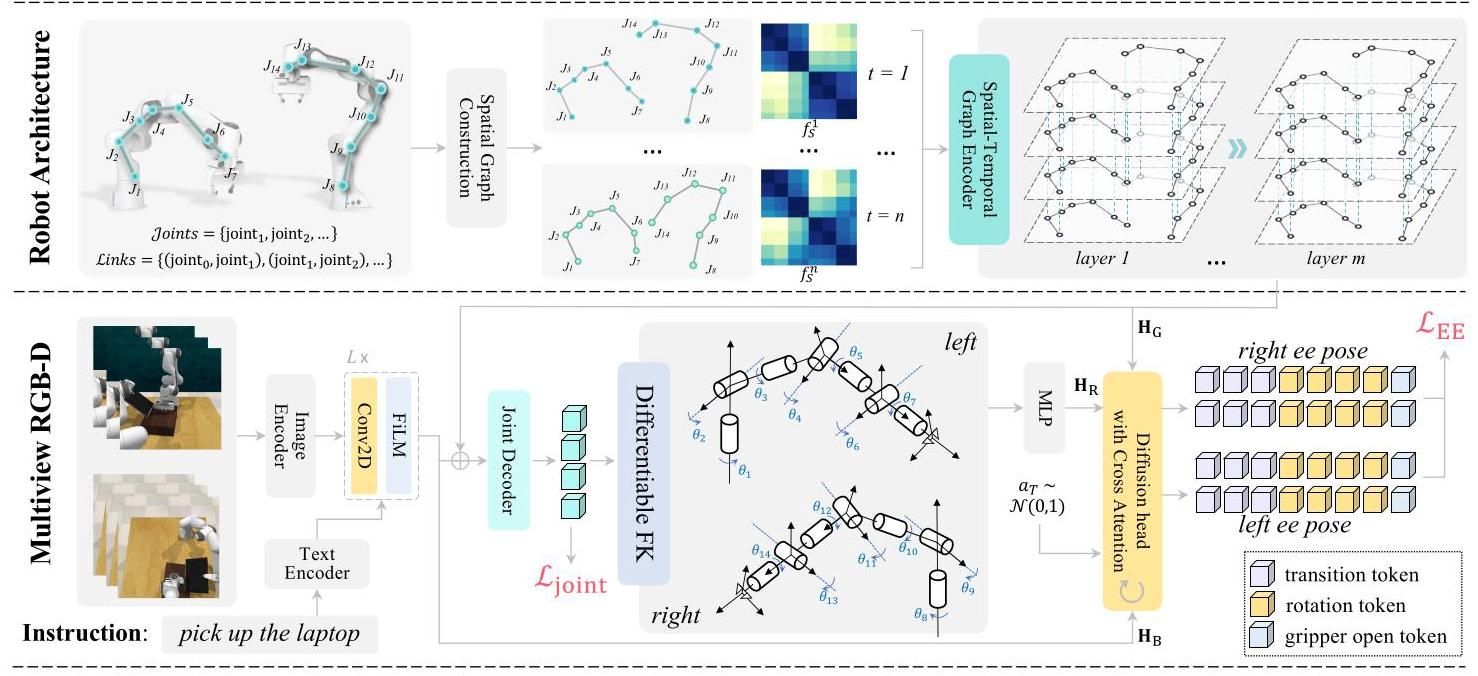


Figure 2. Overview of KStar Diffuser. The top part presents the spatial-temporal graph which is constructed according to the robot architecture. The bottom part shows our backbone and the proposed kinematics regularizer. For the backbone, it extracts the multimodal information which consists of multiview RGB-D observations and language instruction, and then generates bimanual 6D end-effector poses. The kinematics regularizer enhances pose learning by incorporating joint-level predictions, which are mapped to reference end-effector poses through differentiable forward kinematics (FK).

图2. KStar扩散器概述。顶部展示了根据机器人架构构建的时空图。底部展示了我们的骨干网络和提出的运动学正则化器。对于骨干网络，它提取了由多视角RGB-D观测和语言指令组成的多模态信息，然后生成双手6D末端执行器姿态。运动学正则化器通过结合关节级预测来增强姿态学习，这些预测通过可微分正向运动学(FK)映射到参考末端执行器姿态。

# 3.3.2. Backbone

# 3.3.2. 骨干网络

Given the language instruction and multiview RGB-D observation images , we first adopt the Transformer-based encoders to extract their features and , respectively. Then the features are fused with the -layers FiLM block [42] to obtain the hidden states . Each layer is combined with an up-sampling 2D convolution layer:

给定语言指令 和多视角RGB-D观测图像 ，我们首先采用基于Transformer的编码器分别提取它们的特征 和 。然后，这些特征与 层FiLM块[42]融合，以获得隐藏状态 。每一层都与一个上采样2D卷积层结合:

where and the is initialized with .

其中 和 用 初始化。

Our backbone uses the last hidden states as the condition to guide a diffusion head to denoise and generate the bimanual end-effector pose. It is noted that we attach the historical observation images to provide more information for capturing the motion tendency. Following Chi et al. [9], we let the policy predict next actions during training to alleviate multimodal problems. Here, we set both and to 2 . The action prediction is as follows:

我们的骨干网络 使用最后的隐藏状态 作为条件，指导扩散头去噪并生成双手末端执行器姿态。值得注意的是，我们附加了 历史观测图像，以提供更多信息来捕捉运动趋势。根据Chi等人[9]的方法，我们让策略在训练期间预测接下来的 个动作，以缓解多模态问题。这里，我们将 和 都设置为2。动作预测如下:

The learning objective is:

学习目标为:

# 3.3.3. Spatial-Temporal Robot Graph

# 3.3.3. 时空机器人图

The physical architecture impacts the motion of the whole robot, determining whether it can complete the task. Meanwhile, the historical spatial information is also important to future movements. Thus, we propose a spatial-temporal graph method to model the robot architecture at each step and the robot motion at continuous timesteps, representing the static spatial information and dynamic motion features.

物理架构影响整个机器人的运动，决定其是否能完成任务。同时，历史空间信息对未来运动也很重要。因此，我们提出了一种时空图方法，用于在每一步建模机器人架构和连续时间步的机器人运动，表示静态空间信息和动态运动特征。

Spatial Structure Graph Construction. To represent the robot structure, we first parse the Unified Robot Description Format (URDF) file which is usually used to describe the static physical structure of robot such as joint types, joint limits, and link lengths. Then, we define a dual-arm system as an undirected graph based on the joint and link configuration. Here, and represents the node set of joints and edge set of links respectively, where denotes the number of joint. We use to denote the value of -th node feature. It consists of three attributes as follows:

空间结构图构建。为了表示机器人结构，我们首先解析统一机器人描述格式(URDF)文件，该文件通常用于描述机器人的静态物理结构，如关节类型、关节限制和连杆长度。然后，我们基于关节和连杆配置将双臂系统定义为一个无向图 。这里， 和 分别表示关节的节点集和连杆的边集，其中 表示关节数量。我们使用 表示 个节点特征的值。它由以下三个属性组成:

* Joint Coordinate: we use a vector in Cartesian coordinate system, , to denote the absolute coordinates of the -th joint. The vector is normalized according to the workspace boundary for stable convergence of model training.
* 关节坐标:我们使用笛卡尔坐标系中的向量 来表示第 个关节的绝对坐标。该向量根据工作空间边界进行归一化，以确保模型训练的稳定收敛。
* Joint Distance: To measure the spatial relationship between the node and the other nodes , we compute the Euclidean Distance between and :
* 关节距离:为了测量节点 与其他节点 之间的空间关系，我们计算 和 之间的欧几里得距离:

where and means the Euclidean norm.

其中 和 表示欧几里得范数。

* Body Label: To discriminate the source of node , we use a one-hot vector as one of its features. It can also help the policy capture the motion modes of different robot arms such as the symmetry.
* 身体标签:为了区分节点 的来源，我们使用一个独热向量 作为其特征之一。它还可以帮助策略捕捉不同机器人手臂的运动模式，例如对称性。

We concatenate , and to form the completed feature , where demotes the dimension of node features.

我们将 和 连接起来，形成完整的特征 ，其中 表示节点特征的维度。

Spatial-Temporal Graph Learning. Given the same instruction and observation, different historical robot poses result in different predictions. Thus, we build the spatial structure graph with the temporal motion information. Specifically, we build a spatial-temporal graph by combining from historical timesteps, where denotes the number of history steps. In , the node set contains from historical static spatial graph. Additionally, we add edges linking the same joint node joint at different timesteps, in order to establish the correlation of joint motions at continuous times. It can be formulated as follows:

时空图学习。在相同的指令和观察下，不同的历史机器人姿态会导致不同的预测。因此，我们结合时间运动信息构建空间结构图。具体来说，我们通过结合历史时间步的 构建时空图 ，其中 表示历史步数。在 中，节点集 包含来自历史静态空间图的 。此外，我们添加边 连接不同时间步的相同关节节点 ，以建立连续时间关节运动的相关性。它可以表示为:

In this way, we obtain the whole spatial-temporal graph. A Graph Convolutional Network (GCN) is then adopted to propagate and aggregate node features across the graph. The GCN layer updates each node feature by aggregating the features of its neighboring nodes, thereby capturing the relational and structural information of the robotic arms. We use the node feature of the last encoder layer as the representation of the robot structure to condition the denois-ing process.

通过这种方式，我们获得了整个时空图。然后采用图卷积网络(GCN)在图中传播和聚合节点特征。GCN层通过聚合其相邻节点的特征来更新每个节点特征 ，从而捕捉机器人手臂的关系和结构信息。我们使用最后一个编码器层的节点特征 作为机器人结构的表示，以条件化去噪过程。

# 3.3.4. Kinematics Regularizer

# 3.3.4. 运动学正则化器

To control the end-effector effectively, the generated pose trajectory must be processed by an Inverse Kinematics (IK) solver, which calculates the joint configurations to achieve the specified poses. However, because the predicted trajectory is generated without considering the robot kinematic constraints, it often falls outside the IK solver’s feasible range, resulting in high failure rates during execution. To address this limitation, we propose a kinematics regularizer into the end-effector pose learning objective. This regularizer aligns the predicted poses with the robot kinematic constraints, ensuring that the generated trajectory remains within the solvable space of the IK solver, thus enhancing then reliability of trajectory execution.

为了有效控制末端执行器，生成的姿态轨迹必须通过逆运动学(IK)求解器进行处理，该求解器计算关节配置以实现指定的姿态。然而，由于预测的轨迹是在不考虑机器人运动学约束的情况下生成的，它通常超出IK求解器的可行范围，导致执行期间的高失败率。为了解决这一限制，我们提出在末端执行器姿态学习目标中加入运动学正则化器。该正则化器将预测的姿态与机器人运动学约束对齐，确保生成的轨迹保持在IK求解器的可解空间内，从而提高轨迹执行的可靠性。

Differentiable Kinematics. Given a joint configuration , the corresponding end-effector pose can be computed using forward kinematics, represented as a mapping . This mapping from joint space to end-effector space is differentiable, namely Differentiable Forward Kinematics (DFK), enabling the use of gradients to optimize our control policy. Leveraging DFK, our policy learns to predict the next joint configuration , from which we compute an intermediate end-effector pose . By using as a reference, we guide a denoising process to generate a precise and executable end-effector pose.

可微运动学。给定一个关节配置 ，可以使用前向运动学计算相应的末端执行器姿态 ，表示为映射 。这种从关节空间到末端执行器空间的映射是可微的，即可微前向运动学(DFK)，使得我们可以利用梯度来优化控制策略。利用DFK，我们的策略学会预测下一个关节配置 ，从中我们计算出一个中间末端执行器姿态 。通过使用 作为参考，我们引导去噪过程生成一个精确且可执行的末端执行器姿态。

Specifically, we combine structure features with the last hidden state , projecting to the joint space and using the DFK to obtain reference as follows:

具体来说，我们将结构特征 与最后一个隐藏状态 结合，投影到关节空间并使用DFK获得参考 如下:

To ensure the consistency between predicted and actual joint angles, we minimize the joint loss:

为了确保预测的关节角度与实际关节角度的一致性，我们最小化关节损失:

Conditioning Diffusion Process on Kinematics. To enforce kinematic consistency, we condition the diffusion process on the reference representation , an auxiliary input encoding kinematic constraints. This allows the predicted pose trajectory to stay within feasible space. Given the diffusion steps from Eq.(9) and Eq.(10), we have:

在运动学上条件化扩散过程。为了强制运动学一致性，我们在参考表示 上条件化扩散过程，这是一个编码运动学约束的辅助输入。这使得预测的姿态轨迹保持在可行空间内。给定方程(9)和方程(10)中的扩散步骤，我们有:

(18)

Incorporating DFK into the diffusion process allows gradients from the pose loss to propagate back through the kinematic function, ensuring that each denoising step maintains compliance with joint constraints, thereby optimizing the end-effector’s control accuracy and robustness.

将DFK纳入扩散过程使得姿态损失的梯度可以通过运动学函数反向传播，确保每个去噪步骤都符合关节约束，从而优化末端执行器的控制精度和鲁棒性。

# 3.4. Training and Inference

# 3.4. 训练与推理

Training. We use the conditional action generation mode to train KStar Diffuser, which is formulated as a conditional denoising diffusion. The loss function is defined as the Mean Square Error (MSE) as follows:

训练。我们使用条件动作生成模式来训练KStar Diffuser，这被表述为条件去噪扩散。损失函数定义为均方误差(MSE)如下:

where is obtained by the forward diffusion process and is the combination of is the trade-off coefficient.

其中 通过前向扩散过程获得， 是 的组合， 是权衡系数。

Inference. Sampling from a Gaussian noise , the policy performs iterations to gradually denoise a random noise into the noise-free action :

推理。从高斯噪声 中采样，策略 执行 次迭代，逐步将随机噪声 去噪为无噪声动作 :

We use the predicted as the final action to control robot execution.

我们使用预测的 作为最终动作来控制机器人执行。

# 4. Experiment

# 4. 实验

# 4.1. Dataset and Evaluation Settings

# 4.1. 数据集与评估设置

Dataset. Bimanual manipulation tasks demand high levels of coordination, synchronization, and symmetry awareness between the two robotic arms, making them inherently more challenging than single-arm tasks. To assess the capabilities of KStar Diffuser in these areas, we conducted comprehensive experiments using the RLBench2 benchmark [17], an extended version of RLBench tailored for bimanual manipulation and comprising tasks closely resembling real-world scenarios. Please refer to Appendix A for more details about RLBench2 and real-world tasks.

数据集。双手操作任务要求两个机械臂之间高度协调、同步和对称意识，这使得它们比单臂任务更具挑战性。为了评估KStar Diffuser在这些领域的能力，我们使用RLBench2基准[17]进行了全面实验，RLBench2是RLBench的扩展版本，专为双手操作设计，包含与现实场景非常相似的任务。有关RLBench2和现实世界任务的更多详细信息，请参阅附录A。

Evaluation Settings. To evaluate the policy performance, we employ success rate as the primary metric. Although the policy generates multiple sequential actions during execution, we primarily focus on the final goal achievement rather than intermediate steps. Each task is associated with a specific success criterion defined by its target state. To comprehensively assess the policy’s capability, we conduct experiments with varying numbers of demonstrations (20 and 100) during training. Figure 3 illustrates our experimental setup, including both the simulation environment and the Cobot Agilex ALOHA robot. Detailed descriptions of simulation tasks and real-world experimental settings are provided in the Appendix B.

评估设置。为了评估策略性能，我们采用成功率作为主要指标。尽管策略在执行过程中生成多个连续动作，但我们主要关注最终目标达成情况，而不是中间步骤。每个任务都与其目标状态定义的特定成功标准相关联。为了全面评估策略的能力，我们在训练过程中使用不同数量的演示(20和100)进行实验。图3展示了我们的实验设置，包括仿真环境和Cobot Agilex ALOHA机器人。仿真任务和现实世界实验设置的详细描述见附录B。

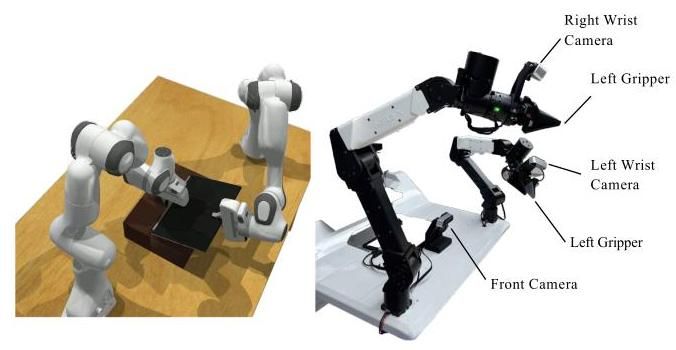


Figure 3. The left: the simulation environment of pick\_laptop task. The right: the ALOHA device used in the real-world tasks.

图3. 左:pick\_laptop任务的仿真环境。右:现实世界任务中使用的ALOHA设备。

# 4.2. Baselines

# 4.2. 基线

We systematically evaluate KStar Diffuser against state-of-the-art methods in two major categories:

我们系统地评估了KStar Diffuser在两大类中最先进方法中的表现:

Transformer-based methods: (1) Action Chunking with Transformers (ACT) [65] employs a Conditional VAE architecture, consisting of an encoder-decoder framework for joint angle sequence prediction; (2) Robotic View Transformer Leader Following (RVT-LF) [17] leverages RVT [15] as its backbone, incorporating a multi-view transformer for cross-view information aggregation and image re-rendering, coupled with a leader following mechanism for action prediction; (3) Perceiver-Actor Leader Following (PerAct-LF) [17] adopts the leader following paradigm based on PerAct [49], utilizing a perceiver Transformer to encode both instructions and voxel observations for optimal voxel action generation; (4) PerAct2 [17] enhances PerAct by implementing a unified feature space for dual-arm actions and employing combined self-attention for synchronized bimanual action prediction.

基于Transformer的方法:(1) 使用Transformer进行动作分块(ACT)[65]采用了条件变分自编码器(Conditional VAE)架构，包含一个用于联合角度序列预测的编码器-解码器框架；(2) 机器人视角Transformer领导者跟随(RVT-LF)[17]以RVT [15]为骨干，结合多视角Transformer进行跨视角信息聚合和图像重新渲染，并采用领导者跟随机制进行动作预测；(3) 感知器-执行器领导者跟随(PerAct-LF)[17]基于PerAct [49]采用领导者跟随范式，利用感知器Transformer对指令和体素观测进行编码，以生成最佳体素动作；(4) PerAct2 [17]通过实现双臂动作的统一特征空间并采用组合自注意力机制进行同步双手动作预测，增强了PerAct。

Diffusion-based methods: (1) Joint-based Diffusion Policy (DP-J) [9] adopts a diffusion model to robotic manipulation within an imitation learning framework, focusing on joint angle prediction; (2) EndEffector-based Diffusion Policy (DP-EE) is a DP variant which we reimplement Diffusion Policy to predict end-effector poses instead of joint angles, offering an alternative control paradigm; (3) 3D Diffusion Policy (DP3) [61] enhances 3D perception by incorporating point clouds for joint angle prediction.

基于扩散的方法:(1) 基于关节的扩散策略(DP-J)[9]在模仿学习框架内采用扩散模型进行机器人操作，专注于关节角度预测；(2) 基于末端执行器的扩散策略(DP-EE)是DP的变体，我们重新实现了扩散策略以预测末端执行器姿态而非关节角度，提供了一种替代控制范式；(3) 3D扩散策略(DP3)[61]通过结合点云进行关节角度预测，增强了3D感知。

# 4.3. Comparison Results with SOTA Methods

# 4.3. 与SOTA方法的比较结果

Experimental Results on RLBench2. As shown in Table 1, the KStar Diffuser significantly outperforms other state-of-the-art baselines, achieving more than higher overall performance with both 20 and 100 training demonstrations. We found that:

RLBench2上的实验结果。如表1所示，KStar Diffuser显著优于其他最先进的基线方法，在20次和100次训练演示中均实现了超过 的整体性能提升。我们发现:

(1) Similar to learning a single-arm policy, the process of learning a bimanual policy can adapt quickly and achieve a high success rate, given a relatively consistent distribution of task trajectories. In push\_box, where the objective is for both arms to push a box toward a specified target along fixed trajectories, our KStar Diffuser and other baseline models perform well. However, as task complexity increases, success rates decrease. For example, in the lift\_ball task, both arms must lift a large ball simultaneously to complete the task. Any asynchrony in movement can cause instability, leading to the ball slipping and ultimately resulting in task failure. Our KStar Diffuser achieves its robust performance on such bimanual tasks by explicitly modeling the spatial and motion relationships between the two arms, surpassing other methods more than 6%.

(1) 类似于学习单臂策略，学习双手策略的过程可以快速适应并实现高成功率，前提是任务轨迹分布相对一致。在push\_box任务中，目标是让双臂沿着固定轨迹将箱子推向指定目标，我们的KStar Diffuser和其他基线模型表现良好。然而，随着任务复杂度的增加，成功率下降。例如，在lift\_ball任务中，双臂必须同时抬起一个大球才能完成任务。任何动作的不同步都可能导致不稳定，使球滑落，最终导致任务失败。我们的KStar Diffuser通过显式建模双臂之间的空间和运动关系，在此类双手任务中表现出色，比其他方法高出6%以上。

(2) Distinct from single-arm systems, bimanual robotic systems possess the capability for collaborative manipulation. Methods which are directly adapted from single-arm to bimanual manipulation exhibit high failure rates in tasks, e.g., pick\_laptop, as they lack consideration of the spatial and motion relationships between the arms. Specifically, as shown in Figure 4, this task involves picking up a notebook lying flat on a cabinet surface. Given that the notebook rests fully against the tabletop, direct grasping by the robotic arm is not possible. Instead, an effective strategy is to control one arm to push the notebook outward from the cabinet by a short distance, allowing the other arm to pick it up. KStar Diffuser achieves a success rate approximately higher than other methods, demonstrating its ability to capture the coordinated motion patterns required for collaborative object manipulation between two arms.

(2) 与单臂系统不同，双手机器人系统具有协作操作的能力。直接从单臂操作适应到双手操作的方法在任务中表现出高失败率，例如pick\_laptop，因为它们缺乏对双臂之间空间和运动关系的考虑。具体来说，如图4所示，该任务涉及从柜子表面拿起一个平放的笔记本。由于笔记本完全紧贴桌面，机械臂无法直接抓取。相反，有效的策略是控制一只手臂将笔记本从柜子向外推一小段距离，让另一只手臂将其拿起。KStar Diffuser的成功率比其他方法高出约 ，展示了其捕捉双臂协作操作所需协调运动模式的能力。

Table 1. The experimental result on simulated tasks. We train the policy with the setting of different training demonstrations, i.e. , to test its capability comprehensively with 100 times each task. The best results are in bold. Each result is reported with three seeds on average.

表1. 模拟任务的实验结果。我们使用不同训练演示的设置(即 )训练策略，以全面测试其能力，每个任务进行100次测试。最佳结果以粗体显示。每个结果均基于三个种子的平均值报告。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Method | Backbone | Push Box | Lift Ball | Handover Item (easy) | Pick Laptop | Sweep Dustpan | Overall |
| 20 demos | | | | | | | |
| ACT [65] | Transformer |  |  |  |  |  |  |
| RVT-LF [17] | Transformer |  |  |  |  |  |  |
| PerAct-LF [17] | Transformer |  |  |  |  |  |  |
| PerAct2 [17] | Transformer |  |  |  |  |  |  |
| DP-J [9] | Diffusion |  |  |  |  |  |  |
| DP-EE [9] | Diffusion |  |  |  |  |  |  |
| DP3 [61] | Diffusion |  |  |  |  |  |  |
| KStar Diffuser (Ours) | Diffusion |  |  |  |  |  |  |
| 100 demos | | | | | | | |
| ACT [65] | Transformer |  |  |  |  |  |  |
| RVT-LF [17] | Transformer |  |  |  |  |  |  |
| PerAct-LF [17] | Transformer |  |  |  |  |  |  |
| PerAct2 [17] | Transformer |  |  |  |  |  |  |
| DP-J [9] | Diffusion |  |  |  |  |  |  |
| DP-EE [9] | Diffusion |  |  |  |  |  |  |
| DP3 [61] | Diffusion |  |  |  |  |  |  |
| KStar Diffuser (Ours) | Diffusion |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 方法 | 骨干网络 | 推箱子 | 举球 | 交接物品(简单) | 拿起笔记本电脑 | 扫簸箕 | 总体 |
| 20个演示 | | | | | | | |
| ACT [65] | Transformer |  |  |  |  |  |  |
| RVT-LF [17] | Transformer |  |  |  |  |  |  |
| PerAct-LF [17] | Transformer |  |  |  |  |  |  |
| PerAct2 [17] | Transformer |  |  |  |  |  |  |
| DP-J [9] | 扩散 |  |  |  |  |  |  |
| DP-EE [9] | 扩散 |  |  |  |  |  |  |
| DP3 [61] | 扩散 |  |  |  |  |  |  |
| KStar扩散器(我们的) | 扩散 |  |  |  |  |  |  |
| 100个演示 | | | | | | | |
| ACT [65] | Transformer |  |  |  |  |  |  |
| RVT-LF [17] | Transformer |  |  |  |  |  |  |
| PerAct-LF [17] | Transformer |  |  |  |  |  |  |
| PerAct2 [17] | Transformer |  |  |  |  |  |  |
| DP-J [9] | 扩散 |  |  |  |  |  |  |
| DP-EE [9] | 扩散 |  |  |  |  |  |  |
| DP3 [61] | 扩散 |  |  |  |  |  |  |
| KStar扩散器(我们的) | 扩散 |  |  |  |  |  |  |

Real-world Experimental Results. To comprehensively evaluate the policy’s effectiveness, we build 2 tasks in real-world based on the simulation benchmark. The performances on real-world tasks are illustrated in Table 2.

实际实验结果。为了全面评估策略的有效性，我们基于仿真基准构建了2个实际任务。实际任务中的性能如表2所示。

Similar to the simulation results, we observe that policies which do not consider about bimanual scenarios, i.e., ACT, DP, and DP3, demonstrate limited capability across all bimanual tasks, achieving around 20% success ratio on average. Although PerAct2 is designed for bimanual tasks via mapping bimanual actions into a shared learning space, it fails to capture the spatial structure of the bimanual system, leading to ineffective arm coordination during execution. Furthermore, we also found PerAct2 faces significant inverse kinematics issues with its predicted end-effector poses, including joint configuration conflicts and unreachable positions which are shown in Figure 4. It is likely due to PerAct2’s limited ability to capture the complex spatial constraints and kinematic relationships within the bimanual robotic system. In contrast, KStar Diffuser achieves superior bimanual coordination, surpassing other methods by over , as it successfully captures the motion patterns between dual arms and predicts feasible end-effector poses.

与仿真结果类似，我们观察到不考虑双手场景的策略，即ACT、DP和DP3，在所有双手任务中表现出有限的能力，平均成功率约为20%。尽管PerAct2通过将双手动作映射到共享学习空间来设计用于双手任务，但它未能捕捉到双手系统的空间结构，导致执行过程中手臂协调无效。此外，我们还发现PerAct2在其预测的末端执行器姿态方面面临显著的逆运动学问题，包括关节配置冲突和不可达位置，如图4所示。这可能是由于PerAct2在捕捉双手机器人系统内复杂的空间约束和运动学关系方面的能力有限。相比之下，KStar Diffuser实现了卓越的双手协调，超过其他方法超过 ，因为它成功捕捉了双臂之间的运动模式并预测了可行的末端执行器姿态。

Table 2. The result of real-world tasks. We train all policies with 100 demostrations and test 15 times. The best result are in bold.

表2. 实际任务的结果。我们使用100次演示训练所有策略，并测试15次。最佳结果以粗体显示。

|  |  |  |  |
| --- | --- | --- | --- |
| Methods | Lift Plate | Handover | Overall |
| ACT [65] |  |  |  |
| DP [9] |  |  |  |
| DP3 [61] |  |  |  |
| PerAct2 [17] |  |  |  |
| KStar Diffuser (Ours) |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| 方法 | 提升板 | 交接 | 总体 |
| ACT [65] |  |  |  |
| DP [9] |  |  |  |
| DP3 [61] |  |  |  |
| PerAct2 [17] |  |  |  |
| KStar扩散器(我们的) |  |  |  |

Table 3. Ablation study of model components.

表3. 模型组件的消融研究。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ST Graph | KR | Handover-S | Handover-R | Overall |
| ✓ | ✓ |  |  |  |
| ✓ | ✘ |  |  |  |
| ✘ | ✘ |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ST图 | KR | 切换-S | 切换-R | 总体 |
| ✓ | ✓ |  |  |  |
| ✓ | ✘ |  |  |  |
| ✘ | ✘ |  |  |  |

# 4.4. Ablation Studies

# 4.4. 消融研究

Effects of Model Components. To systematically evaluate the contribution of each component in KStar Diffuser, we conduct ablation experiments on the handover\_item task in both simulated and real-world environments. We design a progressive ablation process by first removing the Differential Forward Kinematics module while retaining the Spatial-Temporal Graph (ST Graph), and then completely disabling both ST Graph and Kinematics Regularizer (KR).

模型组件的影响。为了系统评估KStar Diffuser中每个组件的贡献，我们在模拟和真实环境中对handover\_item任务进行了消融实验。我们设计了一个逐步消融的过程，首先移除差分正向运动学模块，同时保留时空图(ST Graph)，然后完全禁用ST图和运动学正则化器(KR)。

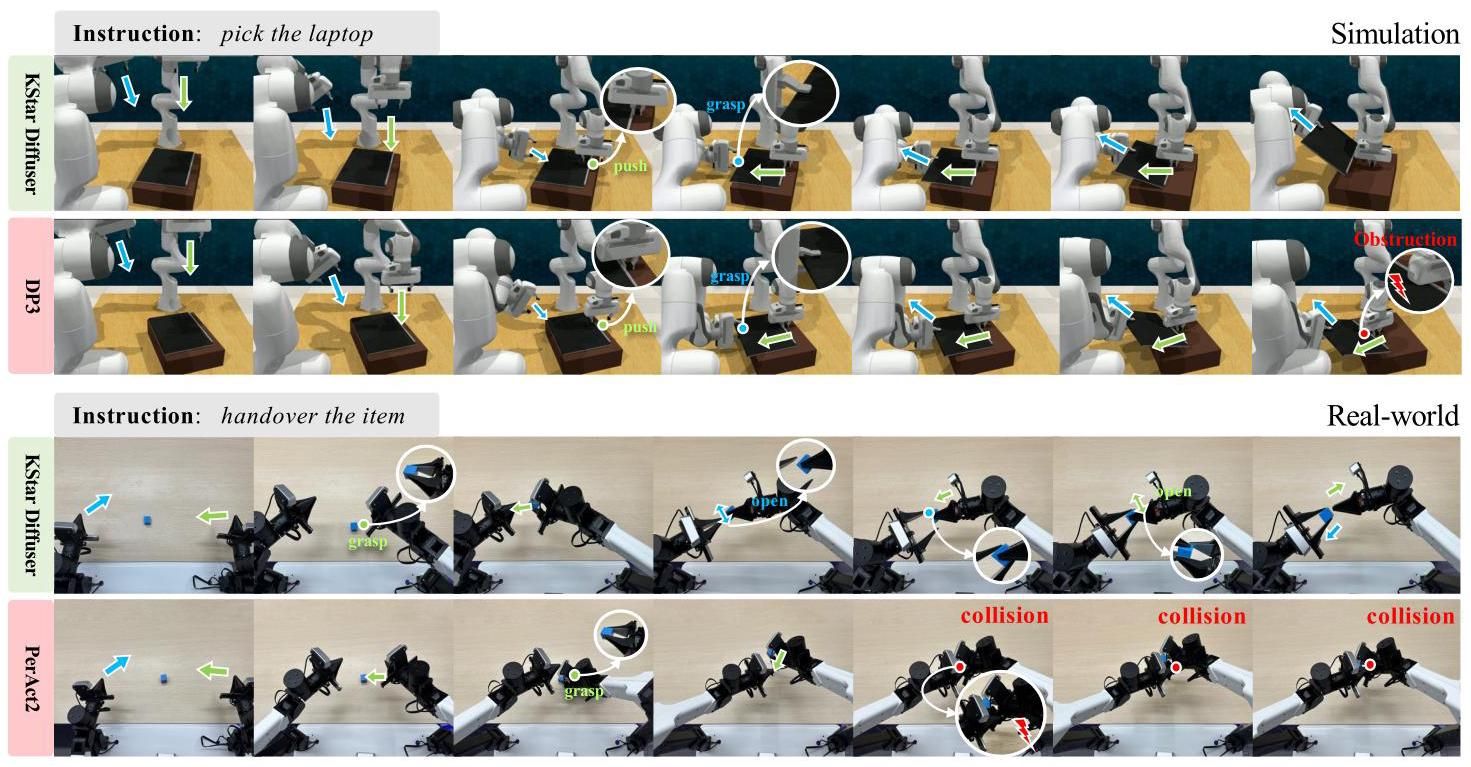


Figure 4. The visualization of bimanual manipulation on simulated RLBench2 and real-world tasks. The blue annotations represent the motion of the robot’s left arm, while the green annotations indicate the motion of the right arm.

图4. 在模拟RLBench2和真实任务中的双手操作可视化。蓝色注释表示机器人左臂的运动，绿色注释表示右臂的运动。

The experimental results in Table 3 demonstrate the crucial role of each component. The removal of KR leads to a significant decrease in success rate, particularly pronounced in real-world scenarios. This performance degradation can be attributed to the fundamental differences between simulated and real environments. While simulated environments maintain consistent and noise-free inputs, real-world scenarios introduce various perturbations, e.g., sensor noise and light reflection, making the policy more susceptible to kinematic constraint violations without the regularization effect of KR. Further ablation by removing both ST Graph and KR results in a substantial performance drop across all experimental settings. This observation illustrates two key aspects: First, the ST Graph effectively captures the spatial-temporal dependencies among joints, which is essential for coordinating the relative positioning and interactions between robotic arms. Second, the graph structure’s explicit encoding of the robotic physical architecture enhances the policy’s robustness against unexpected perturbations by maintaining spatial and temporal coherence. We conduct extensive ablation studies of action chunking size, history lengths and trade-off coefficient, please refer to Appendix for more details.

表3中的实验结果表明了每个组件的关键作用。移除KR导致成功率显著下降，尤其是在真实场景中。这种性能下降可以归因于模拟环境和真实环境之间的根本差异。虽然模拟环境保持了稳定且无噪声的输入，但真实场景引入了各种扰动，例如传感器噪声和光反射，使得策略在没有KR正则化效果的情况下更容易违反运动学约束。进一步移除ST图和KR会导致所有实验设置中的性能大幅下降。这一观察结果说明了两个关键方面:首先，ST图有效地捕捉了关节之间的时空依赖关系，这对于协调机械臂之间的相对位置和交互至关重要。其次，图结构对机器人物理架构的显式编码通过保持时空一致性增强了策略对意外扰动的鲁棒性。我们对动作块大小、历史长度和权衡系数进行了广泛的消融研究，更多细节请参见附录 。

# 4.5. Qualitative Analysis

# 4.5. 定性分析

We further present qualitative analysis in Figure 4. We compare the performance of KStar Diffuser with DP3 and PerAct2 in executing bimanual manipulation tasks within both simulated and real-world environments, respectively.

我们在图4中进一步展示了定性分析。我们分别比较了KStar Diffuser与DP3和PerAct2在模拟和真实环境中执行双手操作任务的性能。

In the simulation task, since the laptop rests flat on the cabinet, direct lifting is not feasible. One robotic arm initiates a forward push, creating space, while the other arm concurrently grasps and elevates the laptop. KStar Diffuser effectively models this dual-arm coordination, generating a precise trajectory of synchronized actions. Conversely, DP3, adapted from a single-arm policy to a dual-arm configuration, fails to achieve effective coordination. Concretely, after executing the push motion, the right arm does not halt, obstructing the left arm’s lifting process.

在模拟任务中，由于笔记本电脑平放在柜子上，直接抬起不可行。一个机械臂开始向前推动，创造空间，而另一个机械臂同时抓住并抬起笔记本电脑。KStar Diffuser有效地模拟了这种双臂协调，生成了精确的同步动作轨迹。相反，DP3从单臂策略调整为双臂配置，未能实现有效协调。具体来说，在执行推动动作后，右臂没有停止，阻碍了左臂的抬起过程。

In the real-world task, KStar Diffuser generates an executable item transfer trajectory between the left and right arms, with no collisions throughout the task, reflecting its strong environmental adaptability and collision avoidance capabilities. Conversely, PerAct2 encounters collisions during the handover process (marked in red), indicating less effective handling of dynamic real-world variables and a lack of kinematic awareness on the robot movements. More qualitative analysis can be found in Appendix D.

在真实任务中，KStar Diffuser生成了左右臂之间可执行的物品转移轨迹，整个任务过程中没有发生碰撞，反映了其强大的环境适应性和碰撞避免能力。相反，PerAct2在交接过程中遇到了碰撞(用红色标记)，表明其对动态真实变量的处理效果较差，并且缺乏对机器人运动的运动学意识。更多定性分析请参见附录D。

# 5. Conclusion

# 5. 结论

In this paper, we have proposed a novel Kinematics enhanced Spatial-TemporAl gRaph Diffuser (KStar Diffuser) that explicitly incorporates both robot structures and kinematics into the bimanual motion generation process. It consists of a spatial-temporal robot graph that explicitly models the robot physical configuration to guide the generative action denoising procedure, and a kinematics regularizer that augments the NBP learning objective by introducing joint-space supervision. Extensive experiments demonstrate that KStar Diffuser outperforms the baselines by a large margin on both simulation and real-world tasks.

在本文中，我们提出了一种新颖的运动学增强时空图扩散器(KStar Diffuser)，它将机器人结构和运动学显式地纳入双手运动生成过程。它由一个时空机器人图组成，该图显式地建模机器人物理配置以指导生成动作去噪过程，以及一个运动学正则化器，通过引入关节空间监督来增强NBP学习目标。大量实验表明，KStar Diffuser在模拟和真实任务中均大幅优于基线方法。

Limitations and Future Directions. While we explored robot structure impacts through GNN modeling and kinematic constraints, the core control logic of end-effector pose prediction and inverse kinematics remains. In the future, we aim to leverage neural networks to directly model joint movements, aligning the robot motion space with the Cartesian space of the human world.

局限性与未来方向。虽然我们通过GNN建模和运动学约束探索了机器人结构的影响，但末端执行器姿态预测和逆运动学的核心控制逻辑仍然存在。未来，我们旨在利用神经网络直接建模关节运动，使机器人运动空间与人类世界的笛卡尔空间对齐。

# References

# 参考文献

[1] Lucia Angelini, Manuela Uliano, Angela Mazzeo, Mattia

Lucia Angelini, Manuela Uliano, Angela Mazzeo, Mattia

Penzotti, and Marco Controzzi. Self-collision avoidance in bimanual teleoperation using collisionik: algorithm revision and usability experiment. In 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids), pages 112-118, 2022. 2

Penzotti, and Marco Controzzi. 使用CollisionIK进行双手遥操作中的自碰撞避免:算法修订和可用性实验。在2022年IEEE-RAS第21届国际人形机器人会议(Humanoids)中，第112-118页，2022年。2

[2] Andreas Aristidou, Joan Lasenby, Yiorgos Chrysanthou, and Ariel Shamir. Inverse kinematics techniques in computer graphics: A survey. In Computer graphics forum, pages 35- 58. Wiley Online Library, 2018. 2

Andreas Aristidou, Joan Lasenby, Yiorgos Chrysanthou, and Ariel Shamir. 计算机图形学中的逆运动学技术:综述。在计算机图形学论坛中，第35-58页。Wiley在线图书馆，2018年。2

[3] Weibang Bai, Ningshan Zhang, Baoru Huang, Ziwei Wang, Francesco Cursi, Ya-Yen Tsai, Bo Xiao, and Eric M Yeat-man. Dual-arm coordinated manipulation for object twisting with human intelligence. In 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 902-908. IEEE, 2021. 2

白伟邦, 张宁山, 黄宝如, 王子维, Francesco Cursi, 蔡雅妍, 肖波, 和 Eric M Yeatman. 基于人类智能的双臂协调操作物体扭转. 在 2021 年 IEEE 系统、人与控制论国际会议 (SMC) 上, 第 902-908 页. IEEE, 2021. 2

[4] Aaron M Bestick, Samuel A Burden, Giorgia Willits, Nikhil Naikal, S Shankar Sastry, and Ruzena Bajcsy. Personalized kinematics for human-robot collaborative manipulation. In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 1037-1044. IEEE, 2015. 2

Aaron M Bestick, Samuel A Burden, Giorgia Willits, Nikhil Naikal, S Shankar Sastry, 和 Ruzena Bajcsy. 人机协作操作的个性化运动学. 在 2015 年 IEEE/RSJ 智能机器人与系统国际会议 (IROS) 上, 第 1037-1044 页. IEEE, 2015. 2

[5] Botond Bócsi, Duy Nguyen-Tuong, Lehel Csató, Bernhard Schoelkopf, and Jan Peters. Learning inverse kinematics with structured prediction. In 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 698- 703. IEEE, 2011. 3

Botond Bócsi, Duy Nguyen-Tuong, Lehel Csató, Bernhard Schoelkopf, 和 Jan Peters. 基于结构化预测的逆运动学学习. 在 2011 年 IEEE/RSJ 智能机器人与系统国际会议上, 第 698-703 页. IEEE, 2011. 3

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakr-ishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817, 2022. 1

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, 等. Rt-1: 用于大规模现实世界控制的机器人变压器. arXiv 预印本 arXiv:2212.06817, 2022. 1

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding,

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding,

Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818, 2023. 1

Danny Driess, Avinava Dubey, Chelsea Finn, 等. Rt-2: 将网络知识转移到机器人控制的视觉-语言-动作模型. arXiv 预印本 arXiv:2307.15818, 2023. 1

[8] Shizhe Chen, Ricardo Garcia, Cordelia Schmid, and Ivan Laptev. Polarnet: point clouds for language-guided robotic manipulation. arXiv preprint arXiv:2309.15596, 2023. 2

陈世哲, Ricardo Garcia, Cordelia Schmid, 和 Ivan Laptev. Polarnet: 用于语言引导机器人操作的点云. arXiv 预印本 arXiv:2309.15596, 2023. 2

[9] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research, page 02783649241273668, 2023. 1, 2, 3, 4, 6, 7, 14

程驰, 徐振佳, 冯思远, Eric Cousineau, 杜一伦, Benjamin Burchfiel, Russ Tedrake, 和 宋书然. 扩散策略: 通过动作扩散的视觉运动策略学习. 国际机器人研究杂志, 第 02783649241273668 页, 2023. 1, 2, 3, 4, 6, 7, 14

[10] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shu-ran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. arXiv preprint arXiv:2402.10329, 2024. 1, 2, 3

程驰, 徐振佳, 潘楚尔, Eric Cousineau, Benjamin Burchfiel, 冯思远, Russ Tedrake, 和 宋书然. 通用操作界面: 无需野外机器人的野外机器人教学. arXiv 预印本 arXiv:2402.10329, 2024. 1, 2, 3

[11] Akos Csiszar, Jan Eilers, and Alexander Verl. On solving the inverse kinematics problem using neural networks. In 2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP), pages 1-6, 2017. 3

Akos Csiszar, Jan Eilers, 和 Alexander Verl. 使用神经网络解决逆运动学问题. 在 2017 年第 24 届机电一体化与机器视觉实践国际会议 (M2VIP) 上, 第 1-6 页, 2017. 3

[12] Aaron D’Souza, Sethu Vijayakumar, and Stefan Schaal. Learning inverse kinematics. In Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180), pages 298-303. IEEE, 2001. 3

Aaron D’Souza, Sethu Vijayakumar, 和 Stefan Schaal. 学习逆运动学. 在 2001 年 IEEE/RSJ 智能机器人与系统国际会议论文集上. 扩展机器人在下一个千年的社会角色 (Cat. No. 01CH37180), 第 298-303 页. IEEE, 2001. 3

[13] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. arXiv preprint arXiv:2401.02117, 2024.1,3

傅子鹏, Tony Z Zhao, 和 Chelsea Finn. Mobile Aloha: 通过低成本全身遥操作学习双手移动操作. arXiv 预印本 arXiv:2401.02117, 2024.1,3

[14] Jianfeng Gao, Xiaoshu Jin, Franziska Krebs, Noémie Jaquier, and Tamim Asfour. Bi-kvil: Keypoints-based visual imitation learning of bimanual manipulation tasks. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 16850-16857. IEEE, 2024. 2

高剑峰, 金晓舒, Franziska Krebs, Noémie Jaquier, 和 Tamim Asfour. Bi-kvil: 基于关键点的双手操作任务视觉模仿学习. 在 2024 年 IEEE 国际机器人与自动化会议 (ICRA) 上, 第 16850-16857 页. IEEE, 2024. 2

[15] 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, pages 694- 710. PMLR, 2023. 1, 2, 3, 6

Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, 和 Dieter Fox. Rvt: 用于3D物体操作的机器人视觉变换器。在机器人学习会议上，第694-710页。PMLR, 2023. 1, 2, 3, 6

[16] Reinhard Grassmann, Vincent Modes, and Jessica Burgner-Kahrs. Learning the forward and inverse kinematics of a 6-dof concentric tube continuum robot in se (3). In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 5125-5132. IEEE, 2018. 2

Reinhard Grassmann, Vincent Modes, 和 Jessica Burgner-Kahrs. 学习6自由度同心管连续体机器人在SE(3)中的正向和逆向运动学。在2018年IEEE/RSJ智能机器人与系统国际会议(IROS)上，第5125-5132页。IEEE, 2018. 2

[17] Markus Grotz, Mohit Shridhar, Tamim Asfour, and Dieter Fox. Peract2: A perceiver actor framework for bimanual manipulation tasks. arXiv preprint arXiv:2407.00278, 2024. 1,2,3,6,7,13

Markus Grotz, Mohit Shridhar, Tamim Asfour, 和 Dieter Fox. Peract2: 用于双手操作任务的感知者-执行者框架。arXiv预印本 arXiv:2407.00278, 2024. 1,2,3,6,7,13

[18] Abdullah Aamir Hayat, Ratan OM Sadanand, and Subir K Saha. Robot manipulation through inverse kinematics. In Proceedings of the 2015 conference on advances in robotics, pages 1-6, 2015. 2

Abdullah Aamir Hayat, Ratan OM Sadanand, 和 Subir K Saha. 通过逆向运动学进行机器人操作。在2015年机器人进展会议上，第1-6页, 2015. 2

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840-6851, 2020. 2

Jonathan Ho, Ajay Jain, 和 Pieter Abbeel. 去噪扩散概率模型。神经信息处理系统进展, 33:6840-6851, 2020. 2

[20] Binghao Huang, Yixuan Wang, Xinyi Yang, Yiyue Luo, and Yunzhu Li. 3d-vitac: Learning fine-grained manipulation

Binghao Huang, Yixuan Wang, Xinyi Yang, Yiyue Luo, 和 Yunzhu Li. 3d-vitac: 通过视觉-触觉感知学习精细操作

with visuo-tactile sensing. arXiv preprint arXiv:2410.24091, 2024. 2

arXiv预印本 arXiv:2410.24091, 2024. 2

[21] Kenneth H Hunt. Structural kinematics of in-parallel-actuated robot-arms. 1983. 1

Kenneth H Hunt. 并联驱动机械臂的结构运动学。1983. 1

[22] Michael I Jordan and David E Rumelhart. Forward models: Supervised learning with a distal teacher. In Backpropaga-tion, pages 189-236. Psychology Press, 2013. 3

Michael I Jordan 和 David E Rumelhart. 正向模型:使用远端教师进行监督学习。在反向传播中，第189-236页。Psychology Press, 2013. 3

[23] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragki-adaki. 3d diffuser actor: Policy diffusion with scene representations. arXiv preprint arXiv:2402.10885, 2024. 2, 3

Tsung-Wei Ke, Nikolaos Gkanatsios, 和 Katerina Fragkiadaki. 3d扩散执行者:使用 场景表示进行策略扩散。arXiv预印本 arXiv:2402.10885, 2024. 2, 3

[24] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246, 2024. 2

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, 等. Openvla: 一个开源的视觉-语言-动作模型。arXiv预印本 arXiv:2406.09246, 2024. 2

[25] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016. 2

Thomas N Kipf 和 Max Welling. 使用图卷积网络进行半监督分类。arXiv预印本 arXiv:1609.02907, 2016. 2

[26] CS George Lee. Robot arm kinematics, dynamics, and control. Computer, 15(12):62-80, 1982. 1

CS George Lee. 机械臂运动学、动力学和控制。计算机, 15(12):62-80, 1982. 1

[27] Maolin Lei, Ting Wang, Chen Yao, Huan Liu, Zhi Wang, and Yongsheng Deng. Real-time kinematics-based self-collision avoidance algorithm for dual-arm robots. Applied Sciences, 10(17):5893, 2020. 1

Maolin Lei, Ting Wang, Chen Yao, Huan Liu, Zhi Wang, 和 Yongsheng Deng. 基于实时运动学的双机械臂自碰撞避免算法。应用科学, 10(17):5893, 2020. 1

[28] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for human pose and shape estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3383-3393, 2021. 3

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, 和 Cewu Lu. Hybrik: 一种用于 人体姿态和形状估计的混合解析-神经逆向运动学解决方案。在IEEE/CVF计算机视觉与模式识别会议(CVPR)上，第3383-3393页, 2021. 3

[29] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for human pose and shape estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3383-3393, 2021. 3

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, 和 Cewu Lu. Hybrik: 一种用于 人体姿态和形状估计的混合解析-神经逆向运动学解决方案。在IEEE/CVF计算机视觉与模式识别会议上，第3383-3393页, 2021. 3

[30] Xiaojie Li, Shaowei He, Jianlong Wu, Yue Yu, Liqiang Nie, and Min Zhang. Mask again: Masked knowledge distillation for masked video modeling. In Proceedings of the ACM International Conference on Multimedia, page 2221-2232. ACM, 2023. 2

李晓杰, 何少伟, 吴建龙, 余越, 聂立强, 张敏. 再次掩码:用于掩码视频建模的掩码知识蒸馏. 在ACM国际多媒体会议论文集, 第2221-2232页. ACM, 2023. 2

[31] Xiaojie Li, Jianlong Wu, Shaowei He, Shuo Kang, Yue Yu, Liqiang Nie, and Min Zhang. Fine-grained key-value memory enhanced predictor for video representation learning. In Proceedings of the ACM International Conference on Multimedia, page 2264-2274. ACM, 2023.

李晓杰, 吴建龙, 何少伟, 康硕, 余越, 聂立强, 张敏. 细粒度键值记忆增强预测器用于视频表示学习. 在ACM国际多媒体会议论文集, 第2264-2274页. ACM, 2023.

[32] Xiaojie Li, Yibo Yang, Xiangtai Li, Jianlong Wu, Yue Yu, Bernard Ghanem, and Min Zhang. Genview: Enhancing view quality with pretrained generative model for self-supervised learning. In Proceedings of the European Conference on Computer Vision. Springer, 2024. 2

李晓杰, 杨一博, 李祥泰, 吴建龙, 余越, Bernard Ghanem, 张敏. Genview:使用预训练生成模型增强视图质量用于自监督学习. 在欧洲计算机视觉会议论文集. Springer, 2024. 2

[33] I Liu, Chun Arthur, Sicheng He, Daniel Seita, and Gau-rav Sukhatme. Voxact-b: Voxel-based acting and stabilizing policy for bimanual manipulation. arXiv preprint arXiv:2407.04152, 2024. 2

刘I, Arthur Chun, 何思成, Daniel Seita, Gaurav Sukhatme. Voxact-b:基于体素的双臂操作和稳定策略. arXiv预印本 arXiv:2407.04152, 2024. 2

[34] Qi Lv, Xiang Deng, Gongwei Chen, Michael Yu Wang, and Liqiang Nie. Decision mamba: A multi-grained state space model with self-evolution regularization for offline rl. In

吕琪, 邓翔, 陈功伟, Michael Yu Wang, 聂立强. Decision Mamba:具有自进化正则化的多粒度状态空间模型用于离线强化学习. 在

Advances in Neural Information Processing Systems, pages

神经信息处理系统进展, 第

22827-22849. Curran Associates, Inc., 2024. 1

22827-22849页. Curran Associates, Inc., 2024. 1

[35] Xiao Ma, Sumit Patidar, Iain Haughton, and Stephen James. Hierarchical diffusion policy for kinematics-aware multitask robotic manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18081-18090, 2024. 1, 2

马骁, Sumit Patidar, Iain Haughton, Stephen James. 用于运动学感知多任务机器人操作的分层扩散策略. 在IEEE/CVF计算机视觉与模式识别会议论文集, 第18081-18090页, 2024. 1, 2

[36] Matthew T Mason. Mechanics of robotic manipulation. MIT press, 2001. 2

Matthew T Mason. 机器人操作的力学. MIT出版社, 2001. 2

[37] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. Advances in Neural Information Processing Systems, 36, 2024. 1

穆尧, 张青龙, 胡孟康, 王文海, 丁明宇, 金俊, 王斌, 戴继峰, 乔宇, 罗平. EmbodiedGPT:通过具身思维链进行视觉语言预训练. 神经信息处理系统进展, 36, 2024. 1

[38] Yoshihiko Nakamura, Kiyoshi Nagai, and Tsuneo Yoshikawa. Dynamics and stability in coordination of multiple robotic mechanisms. The International Journal of Robotics Research, 8(2):44-61, 1989. 1

中村吉彦, 永井清, 吉川恒夫. 多机器人机制协调中的动力学与稳定性. 国际机器人研究杂志, 8(2):44-61, 1989. 1

[39] Valerio Ortenzi, Naresh Marturi, Michael Mistry, Jeffrey Kuo, and Rustam Stolkin. Vision-based framework to estimate robot configuration and kinematic constraints. IEEE/ASME Transactions on Mechatronics, 23(5):2402- 2412, 2018. 2

Valerio Ortenzi, Naresh Marturi, Michael Mistry, Jeffrey Kuo, Rustam Stolkin. 基于视觉的框架用于估计机器人配置和运动学约束. IEEE/ASME机电一体化汇刊, 23(5):2402-2412, 2018. 2

[40] Eric Paljug, Xiaoping Yun, and Vijay Kumar. Control of rolling contacts in multi-arm manipulation. IEEE Transactions on Robotics and Automation, 10(4):441-452, 1994. 1

Eric Paljug, 云晓平, Vijay Kumar. 多臂操作中滚动接触的控制. IEEE机器人与自动化汇刊, 10(4):441-452, 1994. 1

[41] William Peebles and Saining Xie. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4195-4205, 2023. 2

William Peebles, 谢赛宁. 使用Transformer的可扩展扩散模型. 在IEEE/CVF国际计算机视觉会议论文集, 第4195-4205页, 2023. 2

[42] Ethan Perez, Florian Strub, Harm De Vries, Vincent Du-moulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI conference on artificial intelligence, 2018. 4

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, Aaron Courville. Film:使用通用条件层进行视觉推理. 在AAAI人工智能会议论文集, 2018. 4

[43] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952, 2023. 2

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, Robin Rombach. SDXL:改进潜在扩散模型用于高分辨率图像合成. arXiv预印本 arXiv:2307.01952, 2023. 2

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1 (2):3, 2022. 2

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, 和 Mark Chen. 基于CLIP潜在特征的分层文本条件图像生成. arXiv预印本 arXiv:2204.06125, 1 (2):3, 2022. 2

[45] Allen Z Ren, Justin Lidard, Lars L Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. arXiv preprint arXiv:2409.00588, 2024. 2

Allen Z Ren, Justin Lidard, Lars L Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, 和 Max Simchowitz. 扩散策略优化. arXiv预印本 arXiv:2409.00588, 2024. 2

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684-10695, 2022. 2

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, 和 Björn Ommer. 基于潜在扩散模型的高分辨率图像合成. 在IEEE/CVF计算机视觉与模式识别会议论文集, 页码 10684-10695, 2022. 2

[47] Nilanjan Sarkar, Xiaoping Yun, and Vijay Kumar. Dynamic control of 3-d rolling contacts in two-arm manipulation. IEEE Transactions on Robotics and Automation, 13(3): 364-376, 1997. 1

Nilanjan Sarkar, Xiaoping Yun, 和 Vijay Kumar. 双臂操作中三维滚动接触的动态控制. IEEE机器人与自动化汇刊, 13(3): 364-376, 1997. 1

[48] Youshik Shin and Zeungnam Bien. Collision-free trajectory planning for two robot arms. Robotica, 7(3):205-212, 1989. 1, 2

Youshik Shin 和 Zeungnam Bien. 双臂机器人的无碰撞轨迹规划. 机器人学, 7(3):205-212, 1989. 1, 2

[49] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In Conference on Robot Learning, pages 785-799. PMLR, 2023.1,2,3,6,13

Mohit Shridhar, Lucas Manuelli, 和 Dieter Fox. Perceiver-Actor: 用于机器人操作的多任务Transformer. 在机器人学习会议, 页码 785-799. PMLR, 2023.1,2,3,6,13

[50] Christian Smith, Yiannis Karayiannidis, Lazaros Nalpan-tidis, Xavi Gratal, Peng Qi, Dimos V Dimarogonas, and Danica Kragic. Dual arm manipulation-a survey. Robotics and Autonomous systems, 60(10):1340-1353, 2012. 1

Christian Smith, Yiannis Karayiannidis, Lazaros Nalpan-tidis, Xavi Gratal, Peng Qi, Dimos V Dimarogonas, 和 Danica Kragic. 双臂操作综述. 机器人与自主系统, 60(10):1340-1353, 2012. 1

[51] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213, 2024. 2

Octo模型团队, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, 等. Octo: 一个开源的通用机器人策略. arXiv预印本 arXiv:2405.12213, 2024. 2

[52] Vitalis Vosylius, Younggyo Seo, Jafar Uruç, and Stephen James. Render and diffuse: Aligning image and action spaces for diffusion-based behaviour cloning. arXiv preprint arXiv:2405.18196, 2024. 2

Vitalis Vosylius, Younggyo Seo, Jafar Uruç, 和 Stephen James. 渲染与扩散: 对齐图像与动作空间以实现基于扩散的行为克隆. arXiv预印本 arXiv:2405.18196, 2024. 2

[53] Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. arXiv preprint arXiv:2409.20537, 2024. 2

Lirui Wang, Xinlei Chen, Jialiang Zhao, 和 Kaiming He. 通过异构预训练Transformer扩展本体感知视觉学习. arXiv预印本 arXiv:2409.20537, 2024. 2

[54] Yixiao Wang, Yifei Zhang, Mingxiao Huo, Ran Tian, Xi-ang Zhang, Yichen Xie, Chenfeng Xu, Pengliang Ji, Wei Zhan, Mingyu Ding, et al. Sparse diffusion policy: A sparse, reusable, and flexible policy for robot learning. arXiv preprint arXiv:2407.01531, 2024. 2

Yixiao Wang, Yifei Zhang, Mingxiao Huo, Ran Tian, Xiang Zhang, Yichen Xie, Chenfeng Xu, Pengliang Ji, Wei Zhan, Mingyu Ding, 等. 稀疏扩散策略: 一种稀疏、可重用且灵活的机器人学习策略. arXiv预印本 arXiv:2407.01531, 2024. 2

[55] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in Neural Information Processing Systems, 36, 2024. 2

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, 和 Jun Zhu. ProlificDreamer: 通过变分分数蒸馏实现高保真且多样化的文本到3D生成. 神经信息处理系统进展, 36, 2024. 2

[56] Chris Welman. Inverse kinematics and geometric constraints for articulated figure manipulation. 1993. 2

Chris Welman. 逆向运动学与关节图形操作的几何约束. 1993. 2

[57] Kun Wu, Yichen Zhu, Jinming Li, Junjie Wen, Ning Liu, Zhiyuan Xu, Qinru Qiu, and Jian Tang. Discrete policy: Learning disentangled action space for multi-task robotic manipulation. arXiv preprint arXiv:2409.18707, 2024. 2

Kun Wu, Yichen Zhu, Jinming Li, Junjie Wen, Ning Liu, Zhiyuan Xu, Qinru Qiu, 和 Jian Tang. 离散策略: 学习解耦动作空间以实现多任务机器人操作. arXiv预印本 arXiv:2409.18707, 2024. 2

[58] Wenke Xia, Dong Wang, Xincheng Pang, Zhigang Wang, Bin Zhao, Di Hu, and Xuelong Li. Kinematic-aware prompting for generalizable articulated object manipulation with llms. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 2073-2080. IEEE, 2024. 2

Wenke Xia, Dong Wang, Xincheng Pang, Zhigang Wang, Bin Zhao, Di Hu, 和 Xuelong Li. 基于运动学感知的提示用于LLMs的可推广关节对象操作. 在2024年IEEE国际机器人与自动化会议 (ICRA), 页码 2073-2080. IEEE, 2024. 2

[59] Andrea Maria Zanchettin, Paolo Rocco, Luca Bascetta, Ioan-nis Symeonidis, and Steffen Peldschus. Kinematic analysis and synthesis of the human arm motion during a manipulation task. In 2011 IEEE international conference on robotics and automation, pages 2692-2697. IEEE, 2011. 2

Andrea Maria Zanchettin, Paolo Rocco, Luca Bascetta, Ioan-nis Symeonidis, 和 Steffen Peldschus。在操作任务中人体手臂运动的运动学分析与合成。在2011年IEEE国际机器人与自动化会议上，第2692-2697页。IEEE, 2011. 2

[60] Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, Xue Bin Peng, and Jiajun Wu. Generalizable humanoid manipulation with improved 3d diffusion policies. arXiv preprint arXiv:2410.10803, 2024. 1, 2

Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, Xue Bin Peng, 和 Jiajun Wu。具有改进3D扩散策略的可泛化人形操作。arXiv预印本 arXiv:2410.10803, 2024. 1, 2

[61] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy. arXiv preprint arXiv:2403.03954, 2024. 1, 2, 6, 7

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, 和 Huazhe Xu。3D扩散策略。arXiv预印本 arXiv:2403.03954, 2024. 1, 2, 6, 7

[62] Haoyu Zhang, Meng Liu, Yuhong Li, Ming Yan, Zan Gao, Xiaojun Chang, and Liqiang Nie. Attribute-guided collaborative learning for partial person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(12):14144-14160, 2023. 2

Haoyu Zhang, Meng Liu, Yuhong Li, Ming Yan, Zan Gao, Xiaojun Chang, 和 Liqiang Nie。属性引导的协作学习用于部分行人重识别。IEEE模式分析与机器智能汇刊，45(12):14144-14160, 2023. 2

[63] Haoyu Zhang, Meng Liu, Zixin Liu, Xuemeng Song, Yaowei Wang, and Liqiang Nie. Multi-factor adaptive vision selection for egocentric video question answering. In Proceedings of the 41st International Conference on Machine Learning, pages 59310-59328. PMLR, 2024. 2

Haoyu Zhang, Meng Liu, Zixin Liu, Xuemeng Song, Yaowei Wang, 和 Liqiang Nie。多因素自适应视觉选择用于第一人称视频问答。在第41届国际机器学习会议论文集上，第59310-59328页。PMLR, 2024. 2

[64] Junjie Zhang, Chenjia Bai, Haoran He, Wenke Xia, Zhigang Wang, Bin Zhao, Xiu Li, and Xuelong Li. Sam-e: Leveraging visual foundation model with sequence imitation for embodied manipulation. arXiv preprint arXiv:2405.19586, 2024.2 ,3

Junjie Zhang, Chenjia Bai, Haoran He, Wenke Xia, Zhigang Wang, Bin Zhao, Xiu Li, 和 Xuelong Li。Sam-e:利用视觉基础模型与序列模仿进行具身操作。arXiv预印本 arXiv:2405.19586, 2024.2 ,3

[65] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. arXiv preprint arXiv:2304.13705, 2023. 1,2,3,6,7,14

Tony Z Zhao, Vikash Kumar, Sergey Levine, 和 Chelsea Finn。学习低成本硬件的细粒度双手操作。arXiv预印本 arXiv:2304.13705, 2023. 1,2,3,6,7,14

[66] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, march 2024. URL https://github.com/hpcaitech/Open-Sora, 1(3):4. 2

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, 和 Yang You。Open-sora:为所有人民主化高效视频制作，2024年3月。URL https://github.com/hpcaitech/Open-Sora, 1(3):4. 2

[67] Minjie Zhu, Yichen Zhu, Jinming Li, Junjie Wen, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, et al. Scaling diffusion policy in transformer to 1 billion parameters for robotic manipulation. arXiv preprint arXiv:2409.14411, 2024. 2

Minjie Zhu, Yichen Zhu, Jinming Li, Junjie Wen, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, 等。将扩散策略扩展到10亿参数用于机器人操作。arXiv预印本 arXiv:2409.14411, 2024. 2

# Spatial-Temporal Graph Diffusion Policy with Kinematic Modeling for Bimanual Robotic Manipulation

# 基于运动学建模的时空图扩散策略用于双手机器人操作

Supplementary Material

补充材料

We provide a more comprehensive tasks descriptions in Section A, encompassing both simulated environments and real-world scenarios. In Section B, we elaborate on the implementation details, including code base of baseline methods, and hyperparameter configurations for the backbone, graph encoder modules and optimization process. Notably, to thoroughly validate efficacy of KStar Diffuser, we conduct extensive ablation studies analyzing the impact of demonstration quantity, action chunking size, observation history length, and the control learning objective coefficient . The result is presented in Section C. Finally, we present more qualitative analysis in Section D.

我们在A部分提供了更全面的任务描述，包括模拟环境和现实场景。在B部分，我们详细阐述了实现细节，包括基线方法的代码库，以及骨干网络、图编码器模块和优化过程的超参数配置。值得注意的是，为了彻底验证KStar Diffuser的有效性，我们进行了广泛的消融研究，分析了演示数量、动作分块大小、观察历史长度和控制学习目标系数 的影响。结果在C部分展示。最后，我们在D部分提供了更多的定性分析。

# A. Task Descriptions

# A. 任务描述

We conduct extensive experiments on both simulated tasks and real-world tasks to evaluate the effectiveness of our proposed KStar Diffuser. Specifically, we selected five tasks from RLBench2, ranging from basic symmetrical tasks to advanced coordination-requiring tasks, including push\_box, lift\_ball, handover\_item\_easy, sweep\_dustpan, and pick\_laptop. For real-world evaluations, we created a similar setup with two bimanual tasks: lift\_plate and handover\_item\_easy. We present the details about both simulated and real-world tasks in Table A. For each simulated task, we evaluate the model 100 times, whereas for each real-world task, we conduct 15 evaluations.

我们在模拟任务和现实任务上进行了广泛的实验，以评估我们提出的KStar Diffuser的有效性。具体来说，我们从RLBench2中选择了五个任务，从基本的对称任务到需要高级协调的任务，包括push\_box、lift\_ball、handover\_item\_easy、sweep\_dustpan和pick\_laptop。对于现实世界的评估，我们创建了一个类似的设置，包含两个双手任务:lift\_plate和handover\_item\_easy。我们在表A中展示了模拟和现实任务的详细信息。对于每个模拟任务，我们评估模型100次，而对于每个现实任务，我们进行了15次评估。

Table A. Tasks Details.

表A. 任务详情。

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Duration | #Keyframes | Instruction |
| Simulated Tasks | | | |
| push\_box | 4.33s | 2.1 | "Push the box to the red area." |
| lift\_ball | 4.40s | 4.0 | "Lift the ball." |
| handover\_item\_easy | 7.17s | 7.5 | "Handover the item." |
| sweep\_dustpan | 4.93s | 7.3 | "Sweep the dust to the pan.’ |
| pick\_laptop | 3.97s | 7.2 | "Pick up the notebook." |
| Real-world Tasks | | | |
| lift\_plate | 6.37s | 3.4 | "Lift the plate." |
| handover\_item\_easy | 9.52s | 8.6 | "Handover the item." |

|  |  |  |  |
| --- | --- | --- | --- |
| 任务 | 持续时间 | 关键帧数量 | 指令 |
| 模拟任务 | | | |
| 推箱子 | 4.33s | 2.1 | "将箱子推到红色区域。" |
| 举球 | 4.40s | 4.0 | "举起球。" |
| 简单物品交接 | 7.17s | 7.5 | "交接物品。" |
| 扫灰尘 | 4.93s | 7.3 | "将灰尘扫到簸箕里。" |
| 拿笔记本电脑 | 3.97s | 7.2 | "拿起笔记本。" |
| 现实世界任务 | | | |
| 举盘子 | 6.37s | 3.4 | "举起盘子。" |
| 简单物品交接 | 9.52s | 8.6 | "交接物品。" |

# A.1. Simulated Tasks

# A.1. 模拟任务

push\_box. As illustrated in Figure A(a), the task requires the robot to utilize both arms to push a heavy box, weighing , and transport it to a designated target area through a fix trajectory. The completion of the task is defined as successfully moving the box to the specified location. The scenario involves two key elements: a large box and a target area, with the primary challenge being the considerable weight of the box, which exceeds the capacity of a single arm to manage effectively. Notably, this task necessitates the use of both arms simultaneously, as a single robot is incapable of accomplishing it independently.

推箱子。如图A(a)所示，该任务要求机器人使用双臂推动一个重达 的箱子，并通过固定轨迹将其运输到指定的目标区域。任务的完成定义为成功将箱子移动到指定位置。该场景涉及两个关键元素:一个大箱子和一个目标区域，主要挑战在于箱子的重量相当大，超过了单臂有效处理的能力。值得注意的是，该任务需要同时使用双臂，因为单个机器人无法独立完成。

lift\_ball. As depicted in Figure A(b), the task entails the robot using both arms to lift a large ball, achieving a minimum height of 0.95 meters to meet the success criteria. The object in this task is the large ball, presenting a significant coordination challenge. Due to the ball’s size and the inability of the gripper to securely grasp it, the operation relies on coordinated non-prehensile manipulation, demanding precise synchronization of the arms during the lifting process. This task cannot be accomplished by a single robot because of the object’s dimensions.

举球。如图A(b)所示，该任务要求机器人使用双臂举起一个大球，达到至少0.95米的高度以满足成功标准。该任务中的物体是一个大球，提出了显著的协调挑战。由于球的尺寸较大且夹爪无法牢固抓握，操作依赖于协调的非抓取式操纵，要求在举升过程中双臂的精确同步。由于物体的尺寸，该任务无法由单个机器人完成。

handover\_item\_easy. The task requires the robot to handover a red item by utilizing one arm to securely grasp and lift the item to a height of while ensuring the other arm remains idle and unengaged, as shown in Figure . The object involved is a single red block, and the key challenge lies in coordinating the handover process effectively. Successful completion is determined when the item is accurately identified, grasped, and positioned at the required height with no actions performed by the idle arm.

简单物品交接。该任务要求机器人使用一只手臂牢固抓握并举起一个红色物品到 的高度，同时确保另一只手臂保持空闲且未参与，如图 所示。涉及的物体是一个红色方块，主要挑战在于有效协调交接过程。当物品被准确识别、抓握并定位到所需高度且空闲手臂未执行任何动作时，任务成功完成。

sweep\_dustpan. It is shown in Figure A(d) that the task involves the robot using a broom to sweep dust into a dust pan, requiring precise coordination of the sweeping motion to effectively collect the dust. The objects involved include a broom, a dust pan, supporting objects, and the dust itself. Successful completion is defined as all the dust being gathered inside the dust pan. The primary challenge lies in the accuracy and control of the sweeping motion to ensure that the dust is properly directed into the pan.

扫簸箕。如图A(d)所示，该任务要求机器人使用扫帚将灰尘扫入簸箕，需要精确协调扫动动作以有效收集灰尘。涉及的物体包括扫帚、簸箕、支撑物和灰尘本身。任务成功完成定义为所有灰尘都被收集到簸箕内。主要挑战在于扫动动作的准确性和控制，以确保灰尘被正确引导到簸箕中。

pick\_laptop. As illustrated in Figure A(e), the task requires the robot to pick up a notebook placed on top of a block by first manipulating it into a position suitable for grasping. This involves performing non-prehensile actions, such as pushing or sliding, to adjust the notebook’s orientation before securely grasping and lifting it off the block. The objects involved are a notebook and a block. Successful completion is defined as the robot lifting the notebook off the block. Although the task can be performed with a single robotic arm, precise coordination is essential for effective manipulation.

拾取笔记本电脑。如图A(e)所示，该任务要求机器人拾取放置在方块顶部的笔记本电脑，首先将其调整到适合抓握的位置。这涉及执行非抓取式动作，如推动或滑动，以调整笔记本电脑的方向，然后牢固抓握并将其从方块上举起。涉及的物体是笔记本电脑和方块。任务成功完成定义为机器人将笔记本电脑从方块上举起。尽管该任务可以由单只机械臂完成，但精确的协调对于有效操纵至关重要。

# A.2. Real-world Tasks

# A.2. 现实世界任务

lift\_plate. As shown in Figure A(f), the task involves the robot using both arms to lift a plate, maintaining an elevated position for over 3 seconds to meet the success criteria. The target object is a plate requiring coordinated manipulation. Due to the plate’s width and the need for stable control, the operation demands precise bimanual manipulation, requiring synchronized lifting motions from both arms. This task cannot be accomplished by a single robot arm given the plate’s dimensions and stability requirements.

举盘子。如图A(f)所示，该任务要求机器人使用双臂举起一个盘子，保持举升位置超过3秒以满足成功标准。目标物体是一个需要协调操纵的盘子。由于盘子的宽度和稳定控制的需求，操作需要精确的双臂操纵，要求双臂同步举升动作。鉴于盘子的尺寸和稳定性要求，该任务无法由单只机械臂完成。

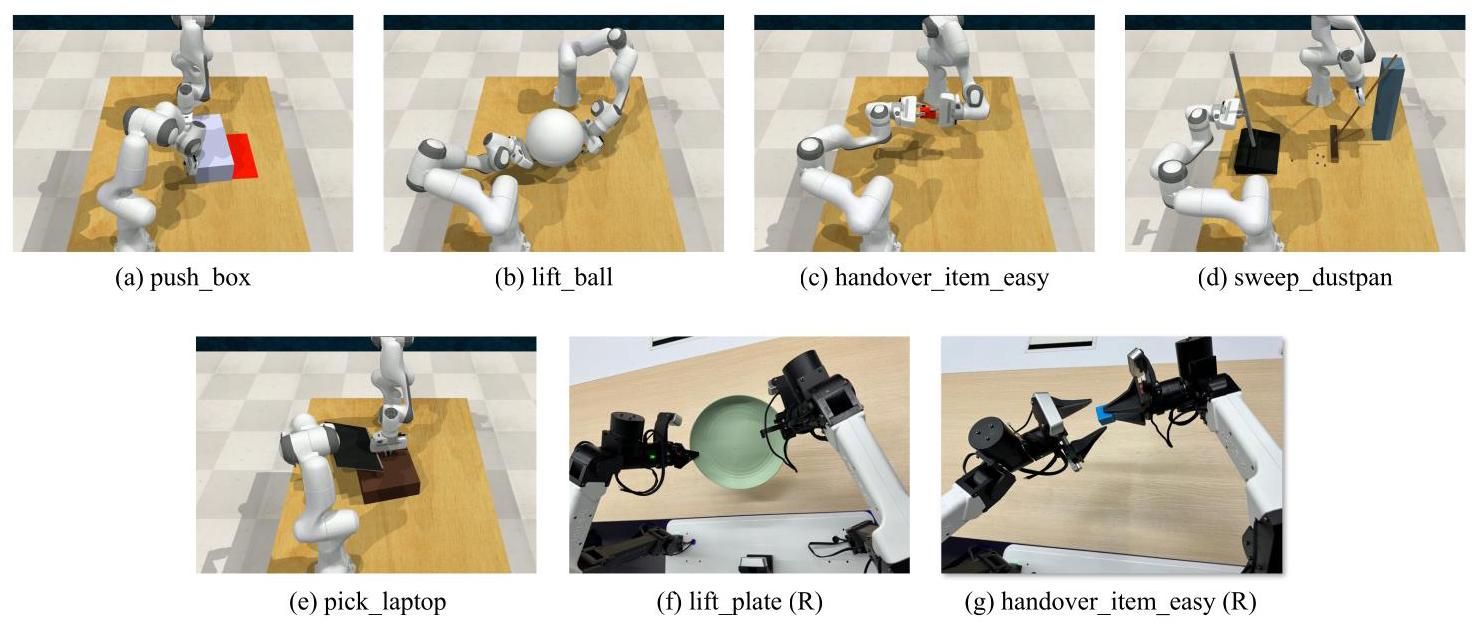


Figure A. The visualization of simulated tasks and real-world tasks. The task with "(R)" means the real-world tasks.

图A. 模拟任务和现实世界任务的可视化。带有“(R)”的任务表示现实世界任务。

handover\_item\_easy. Similar to the corresponding simulated task, this task requires the robot to handover a blue item by utilizing one arm to securely grasp and lift the item as shown in Figure A(g). However, the robot do not need to lift the item to a height of , but instead considering the security, while ensuring the other arm remains idle and unengaged, The object involved is a blue block, and the key challenge is as the same as that of in simulation. Successful completion is determined when the item is accurately identified, grasped, and positioned at the required height with no actions performed by the idle arm.

简单物品交接。与相应的模拟任务类似，该任务要求机器人使用一只手臂牢固抓握并举起一个蓝色物品，如图A(g)所示。然而，考虑到安全性，机器人不需要将物品举到 的高度，而是 ，同时确保另一只手臂保持空闲且未参与。涉及的物体是一个蓝色方块，主要挑战与模拟中的相同。当物品被准确识别、抓握并定位到所需高度且空闲手臂未执行任何动作时，任务成功完成。

# B. Implementation Details

# B. 实现细节

# B.1. Details of KStar Diffuser

# B.1. KStar Diffuser的细节

# B.1.1. Backbone

# B.1.1. 骨干网络

Vision Branch. Following the existing work [17, 49], we employ multiview RGB-D observation as visual input, where the resolution of RGB images is and the depth data is processed into point clouds using the camera’s intrinsic and extrinsic parameters. The camera views contain front, overhead, right\_over\_shoulder, left\_over\_shoulder, right\_wrist, left\_wrist. We uniformly use a Vision Transformer which is trained from scratch as the vision encoder.

视觉分支。遵循现有工作[17, 49]，我们采用多视角RGB-D观测作为视觉输入，其中RGB图像的分辨率为 ，深度数据使用相机的内外参数处理为点云 。相机视角包括正面、俯视、右肩上方、左肩上方、右手腕、左手腕。我们统一使用从头开始训练的视觉Transformer作为视觉编码器。

Language Branch. For the language instruction, we encode the instruction with CLIP’s language encoder . Specifically, the input sentence is preprocessed by the CLIP’s tokenizer and then encoded to a sequence of dimensions . We use its hidden state of " [CLS]" as the textual feature. It is worth noting that we extracted all text features in advance. Therefore, throughout the training phase, the weights of the text encoder do not participate in the gradient calculation of backpropagation.

语言分支。对于语言指令，我们使用CLIP的语言编码器 对指令进行编码。具体来说，输入句子通过CLIP的分词器进行预处理，然后编码为维度序列 。我们使用其"[CLS]"的隐藏状态作为文本特征。值得注意的是，我们提前提取了所有文本特征。因此，在整个训练阶段，文本编码器的权重不参与反向传播的梯度计算。

Fusion Module. The fusion module begins with upsam-pling the visual features, followed by feature-wise modulation (FiLM) to integrate textual features, projecting them into a high-dimensional semantic space. This hierarchical fusion process is performed iteratively times, where we empirically set to 3 .

融合模块。融合模块首先对视觉特征进行上采样，然后通过特征调制(FiLM)整合文本特征，将其投影到高维语义空间。这个分层融合过程迭代执行 次，我们经验性地将 设置为3。

The relevant hyperparameters of the backbone are presented in Table B.

主干网络的相关超参数如表B所示。

Table B. The hyperparameters of Backbone.

表B. 主干网络的超参数。

|  |  |  |
| --- | --- | --- |
|  | Vision | Text |
| patch size | 16 |  |
| hidden size | 64 | 512 |
| #layers | 6 | 12 |
| #heads | 8 | 8 |
| intermediate size | 64 | 2048 |
| dropout ratio | 0.1 | 0.0 |
| activation | irelu | quickgelu |
| trainable | ✓ | ✘ |

|  |  |  |
| --- | --- | --- |
|  | 视觉 | 文本 |
| 补丁大小 | 16 |  |
| 隐藏大小 | 64 | 512 |
| 层数 | 6 | 12 |
| 头数 | 8 | 8 |
| 中间大小 | 64 | 2048 |
| 丢弃率 | 0.1 | 0.0 |
| 激活 | irelu | quickgelu |
| 可训练 | ✓ | ✘ |

[[2]](#footnote-91)

# B.1.2. Spatial-Temporal Graph

# B.1.2. 时空图

We construct the spatial graph based on the URDF file of robotic arms. In the simulation setup, we use two Franka Panda arms, each with 7 joints, resulting in a total of 14 joints. In the real-world setup, we employ a bimanual ALOHA device, which has 12 joints, with 6 joints per arm. Therefore, the number of nodes and edges in the spatial graph is set to 14 and 12 for the simulation, respectively, and 12 and 10 for the real-world setup.

我们基于机械臂的URDF文件构建了空间图。在仿真设置中，我们使用了两台Franka Panda机械臂，每台有7个关节，总共有14个关节。在现实世界的设置中，我们使用了双手机器人ALOHA设备，它有12个关节，每只手臂有6个关节。因此，空间图中的节点和边的数量在仿真中分别设置为14和12，在现实世界中分别设置为12和10。

For the dynamic spatio-temporal graph, we combine the spatial graphs of three consecutive timesteps and add inter-timestep edges connecting the same joint node across different timesteps. In simulated tasks, this results in a graph with 42 nodes and 33 edges. In real-world tasks, the graph has 36 nodes and 28 edges.

对于动态时空图，我们将三个连续时间步的空间图结合起来，并添加了连接不同时间步中相同关节节点的跨时间步边。在仿真任务中，这导致了一个有42个节点和33条边的图。在现实世界任务中，该图有36个节点和28条边。

We use the Graph Convolutional Graph (GCN) encoder to obtain the spatial-temporal graph representation. The detailed hyperparameters are presented in Table C.

我们使用图卷积网络(GCN)编码器来获取时空图的表示。详细的超参数在表C中给出。

Table C. The hyperparameters of GCN Encoder.

表C. GCN编码器的超参数。

|  |  |  |
| --- | --- | --- |
|  | Simulation | Real-world |
| #nodes | 42 | 36 |
| #edges | 36 | 28 |
| node dimension | 19 | 19 |
| hidden size | 128 | 128 |
| intermediate size | 128 | 128 |
| #layers | 4 | 4 |

|  |  |  |
| --- | --- | --- |
|  | 模拟 | 现实世界 |
| 节点数量 | 42 | 36 |
| 边数量 | 36 | 28 |
| 节点维度 | 19 | 19 |
| 隐藏层大小 | 128 | 128 |
| 中间层大小 | 128 | 128 |
| 层数 | 4 | 4 |

# B.1.3. Optimization Details

# B.1.3. 优化细节

During training, we follow the setup of Diffusion Policy, using the DDPM scheduler to forward and denoise, where the step of forward process and reverse process is set to 100 and 1, respectively. Table D shows the training hyperparamters. Table D. The training hyperparameters.

在训练过程中，我们遵循扩散策略(Diffusion Policy)的设置，使用DDPM调度器进行前向和去噪操作，其中前向过程和反向过程的步数分别设置为100和1。表D展示了训练超参数。表D. 训练超参数。

|  |  |  |
| --- | --- | --- |
|  |  | Values |
|  | batch size | 64 |
|  | learning rate |  |
|  | warmup step |  |
|  | weight decay |  |
|  | lr scheduler | cosine |
|  | training step | 150k |
|  | Optimizer | AdamW |

|  |  |  |
| --- | --- | --- |
|  |  | 值 |
|  | 批量大小 | 64 |
|  | 学习率 |  |
|  | 预热步数 |  |
|  | 权重衰减 |  |
|  | 学习率调度器 | 余弦 |
|  | 训练步数 | 150k |
|  | 优化器 | AdamW |

# B.2. Code Base

# B.2. 代码库

The code bases employed for our evaluations are detailed as follows:

我们评估所使用的代码库详细如下:

* ACT: https://github.com/markusgrotz/peract\_bimanual
* ACT: https://github.com/markusgrotz/peract\_bimanual
* RVT-LF: https://github.com/markusgrotz/peract\_bimanual
* RVT-LF: https://github.com/markusgrotz/peract\_bimanual
* PerAct-LF: https://github.com/markusgrotz/peract\_bimanual
* PerAct-LF: https://github.com/markusgrotz/peract\_bimanual
* PerAct2: https://github.com/markusgrotz/peract\_bimanual
* PerAct2: https://github.com/markusgrotz/peract\_bimanual
* DP-J: https://github.com/real-stanford/diffusion\_policy
* DP-J: https://github.com/real-stanford/diffusion\_policy
* DP3: https://github.com/YanjieZe/3D-Diffusion-Policy
* DP3: https://github.com/YanjieZe/3D-Diffusion-Policy

# C. Extensive Ablation Studies

# C. 广泛的消融研究

To evaluate KStar Diffuser more comprehensively, we conduct following extensive experiments.

为了更全面地评估KStar Diffuser，我们进行了以下广泛的实验。

(1) The Effects of Demonstration Quantity. Given the critical role of demonstration quantity in imitation learning, we conduct an ablation study by training the KStar diffuser with 50 demonstrations, with results reported in Table E. Additionally, we provide a comparative analysis of policy performance across varying numbers of demonstrations (20, 50, and 100), as shown in Figure 5. The results demonstrate a clear positive correlation between demonstration quantity and policy performance. With 20 demonstrations, the policy achieves basic task completion capabilities. Increasing to 50 demonstrations yields significant performance improvements, with success rates rising by around across the task suite. The upward trend continues as we scale to 100 demonstrations, indicating that the policy benefits from larger demonstration sets. These findings suggest that expanding the demonstration dataset consistently enhances the policy’s ability to learn and generalize manipulation tasks.

(1) 演示数量的影响。鉴于演示数量在模仿学习中的关键作用，我们通过使用50个演示训练KStar diffuser进行了消融研究，结果如表E所示。此外，我们还提供了不同演示数量(20、50和100)下策略性能的比较分析，如图5所示。结果表明，演示数量与策略性能之间存在明显的正相关关系。使用20个演示时，策略具备基本的任务完成能力。增加到50个演示时，性能显著提升，任务套件的成功率提高了约 。当我们扩展到100个演示时，上升趋势继续，表明策略受益于更大的演示集。这些发现表明，扩展演示数据集持续增强了策略学习和泛化操作任务的能力。

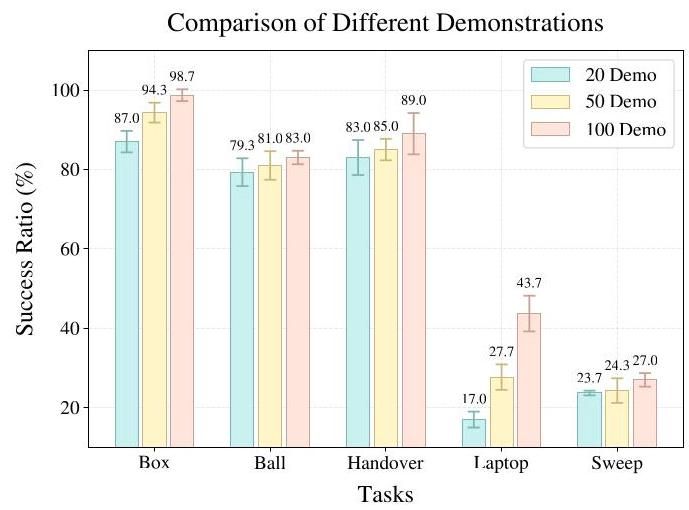


Figure B. The comparison of different number of demonstrations.

图B. 不同演示数量的比较。

(2) The Effects of Action Chunking Size. As mentioned in previous work , action chunking prediction serves as an effective approach to address the multimodality challenges inherent in robotic manipulation tasks. To empirically verify this mechanism within our framework, we conducted an evaluation of the KStar Diffuser under various action chunking size configurations. Our experimental setup explores three distinct action chunking strategies: 1) singlestep prediction where only the next optimal pose is generated (action chunk ),2) two-step prediction of consecutive optimal poses (action chunk ), and 3 ) five-step prediction of sequential optimal poses (action chunk = 5).

(2) 动作分块大小的影响。如先前工作 所述，动作分块预测是解决机器人操作任务中固有的多模态挑战的有效方法。为了在我们的框架内实证验证这一机制，我们对KStar Diffuser在各种动作分块大小配置下进行了评估。我们的实验设置探索了三种不同的动作分块策略:1)单步预测，仅生成下一个最优姿态(动作分块 )；2)两步预测，生成连续的最优姿态(动作分块 )；3)五步预测，生成连续的最优姿态(动作分块=5)。

Table E. The experimental result on simulated tasks. We train the policy with the setting of different training demonstrations, i.e. [20,100], to test its capability comprehensively. The best results are in bold. Each result is reported with three seeds on average.

表E. 模拟任务的实验结果。我们使用不同训练演示数量(即[20,100])的设置训练策略，以全面测试其能力。最佳结果以粗体显示。每个结果均基于三个种子的平均值报告。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Push Box | Lift Ball | Handover Item (easy) | Pick Laptop | Sweep Dustpan | Overall |
| Demonstration Quantity | | | | | | |
| KStar Diffuser (num demos=20) |  |  |  |  |  |  |
| KStar Diffuser (num demos=50) |  |  |  |  |  |  |
| KStar Diffuser (num demos ) |  |  |  |  |  |  |
| Action Chunking Length | | | | | | |
| KStar Diffuser (chunking ) |  |  |  |  |  |  |
| KStar Diffuser (chunking=2) |  |  |  |  |  |  |
| KStar Diffuser (chunking=5) |  |  |  |  |  |  |
| Historical Observation Length | | | | | | |
| KStar Diffuser (history=0) |  |  |  |  |  |  |
| KStar Diffuser (history=1) |  |  |  |  |  |  |
| KStar Diffuser (history=2) |  |  |  |  |  |  |
| Coefficient | | | | | | |
| KStar Diffuser |  |  |  |  |  |  |
| KStar Diffuser |  |  |  |  |  |  |
| KStar Diffuser |  |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 推箱子 | 举球 | 交接物品(简单) | 拿笔记本电脑 | 扫簸箕 | 总体 |
| 演示数量 | | | | | | |
| KStar扩散器(演示次数=20) |  |  |  |  |  |  |
| KStar扩散器(演示次数=50) |  |  |  |  |  |  |
| KStar扩散器(演示次数 ) |  |  |  |  |  |  |
| 动作分块长度 | | | | | | |
| KStar扩散器(分块 ) |  |  |  |  |  |  |
| KStar扩散器(分块=2) |  |  |  |  |  |  |
| KStar扩散器(分块=5) |  |  |  |  |  |  |
| 历史观察长度 | | | | | | |
| KStar扩散器(历史=0) |  |  |  |  |  |  |
| KStar扩散器(历史=1) |  |  |  |  |  |  |
| KStar扩散器(历史=2) |  |  |  |  |  |  |
| 系数 | | | | | | |
| KStar扩散器 |  |  |  |  |  |  |
| KStar扩散器 |  |  |  |  |  |  |
| KStar扩散器 |  |  |  |  |  |  |

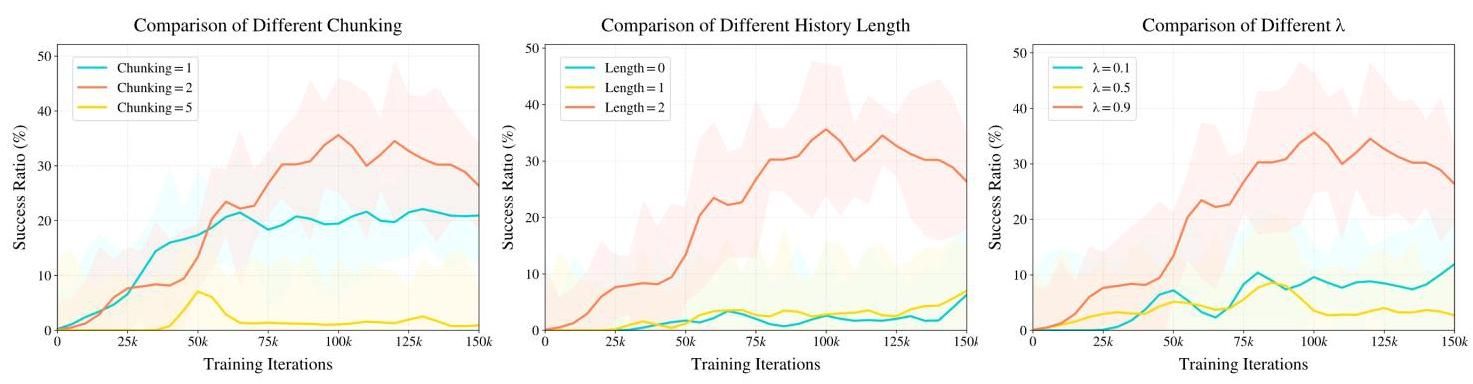


Figure C. The left: The result of different action chunking size. The middle: The result of history lengths. The right: The result of different coefficient .

图C。左侧:不同动作分块大小的结果。中间:历史长度的结果。右侧:不同系数 的结果。

The results presented in Table E reveal an interesting trade-off between prediction horizon and model performance. While single-step prediction (action chunk = 1) provides basic capabilities, it usually meets the multimodal problem. The two-step prediction strategy (action chunk ) emerges as the optimal configuration, demonstrating superior success rates and motion quality across all tasks. Notably, attempting to predict longer sequences (action chunk = 5) leads to decreased performance, with success rates dropping by around compared to the two-step configuration. This performance degradation suggests that when predicting next best pose, extended prediction horizons introduce excessive complexity into the learning problem, making it challenging for the policy to capture and generate accurate action sequences. As shown in Figure C, we present the variation in the success ratio for the handover\_item\_easy task, under different action chunking size configurations as the number of training steps increases.

表E中的结果揭示了预测范围与模型性能之间的有趣权衡。虽然单步预测(动作分块=1)提供了基本能力，但它通常会遇到多模态问题。两步预测策略(动作分块 )成为最佳配置，在所有任务中表现出更高的成功率和运动质量。值得注意的是，尝试预测更长的序列(动作分块=5)会导致性能下降，与两步配置相比，成功率下降了约 。这种性能下降表明，在预测下一个最佳姿势时，延长的预测范围会给学习问题引入过多的复杂性，使得策略难以捕捉和生成准确的动作序列。如图C所示，我们展示了在不同动作分块大小配置下，随着训练步骤的增加，handover\_item\_easy任务的成功率变化。

(3) The Effects of Historical Observation Length. During our experiments, we found that the historical information plays an important role in action prediction quality. We conduct an ablation study across three history lengths: 0 (current observation only), 1, and 2 steps. The results are shown in Table E. With no historical information (0- step), the policy fails to learn effective policies, as it cannot capture the temporal dependencies crucial for manipulation tasks. Adding one historical step enables basic learning capabilities, with the model achieving preliminary success in simpler tasks. Further extending to two historical steps yields optimal performance, showing an approximately improvement over the single-step configuration and demonstrating enhanced stability across all tasks. While longer history lengths might provide more temporal context, they risk introducing redundant information that could potentially obscure relevant features, as observed in our preliminary experiments. These findings suggest that while temporal context is essential for understanding the current state and predicting actions, an appropriate history window, e.g., 2 steps, provides an optimal balance between capturing necessary temporal dependencies and maintaining computational efficiency. As shown in Figure , we present the variation in the success ratio for the handover\_item\_easy task, under different history length configurations as the number of training steps increases.

(3) 历史观察长度的影响。在我们的实验中，我们发现历史信息在动作预测质量中起着重要作用。我们对三种历史长度进行了消融研究:0(仅当前观察)、1步和2步。结果如表E所示。在没有历史信息(0步)的情况下，策略无法学习有效的策略，因为它无法捕捉到对操作任务至关重要的时间依赖性。添加一个历史步骤使模型具备了基本的学习能力，在较简单的任务中取得了初步成功。进一步扩展到两个历史步骤时，性能达到最佳，与单步配置相比，性能提高了约 ，并在所有任务中表现出更高的稳定性。虽然更长的历史长度可能会提供更多的时间上下文，但它们可能会引入冗余信息，从而可能掩盖相关特征，正如我们在初步实验中观察到的那样。这些发现表明，虽然时间上下文对于理解当前状态和预测动作至关重要，但适当的历史窗口(例如2步)在捕捉必要的时间依赖性和保持计算效率之间提供了最佳平衡。如图 所示，我们展示了在不同历史长度配置下，随着训练步骤的增加，handover\_item\_easy任务的成功率变化。

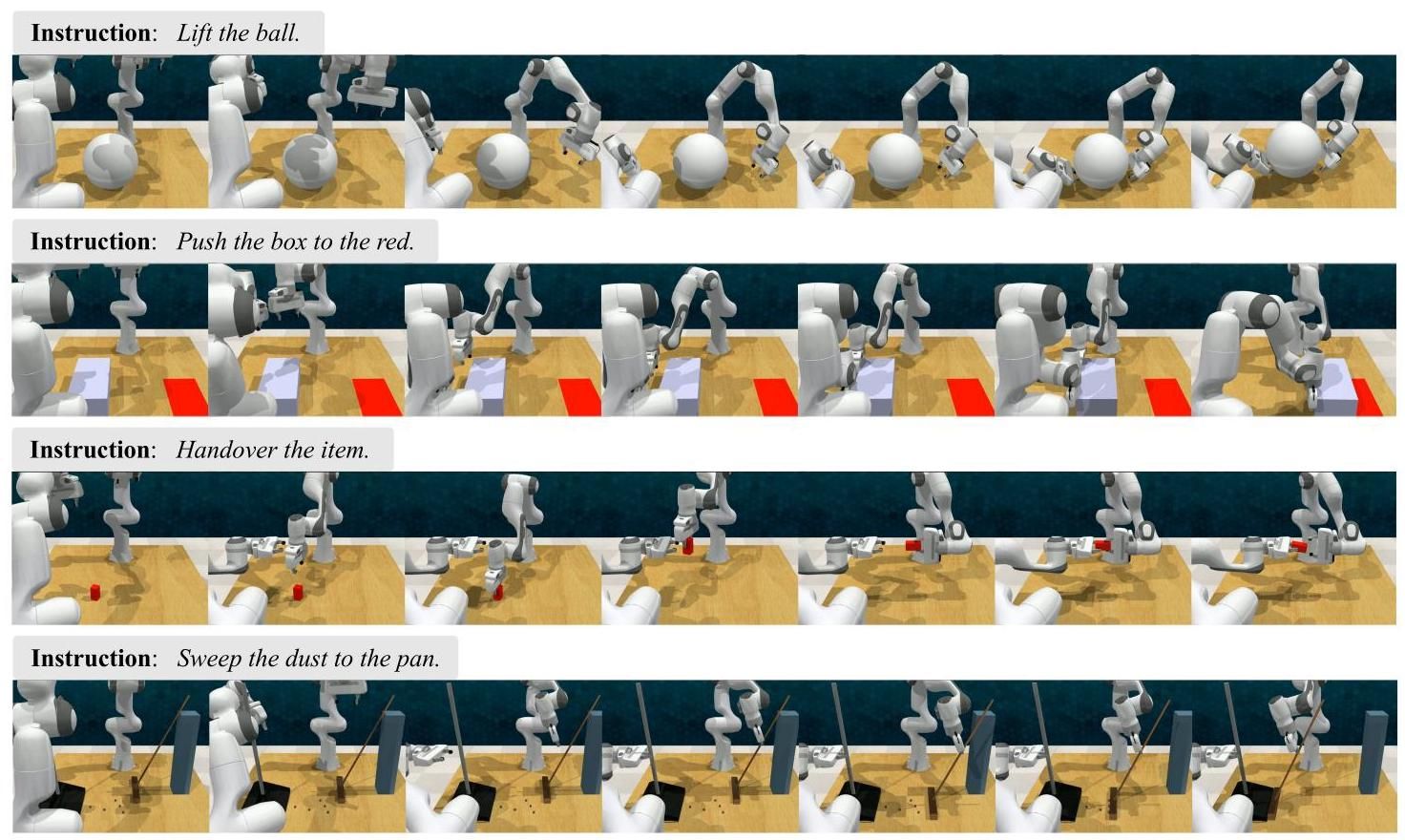


Figure D. The visualization of simulated tasks, including push\_box, lift\_ball, handover\_item\_easy, sweep\_dustpan.

图D。模拟任务的可视化，包括push\_box、lift\_ball、handover\_item\_easy、sweep\_dustpan。

(4) The Effects of Coefficient . In policy learning, a kinematic regularization term is incorporated into the next best pose learning objective to balance task effectiveness and motion constrains. The regularization strength is controlled by the coefficient , which determines the extent of kinematic constraints imposed on the learning process. The choice of significantly influences policy performance. Larger values of (approaching 1.0) correspond to weaker kinematic constraints, enabling more flexible motion patterns. Conversely, smaller values of (e.g.,0.1) impose stricter kinematic constraints, yielding more conservative policies that emphasize motion smoothness over task efficiency. Empirical results, as shown in Table E, demonstrate that policy performance reaches its optimum at . For smaller values of , the learned policies exhibit overly cautious behavior, manifesting in trajectories that prioritize smoothness at the expense of task efficiency and optimal path planning. Conversely, as approaches 1.0, the diminished kinematic constraints result in less regulated motion patterns, potentially compromising trajectory naturalness and precision. These findings indicate that achieves a better trade-off between preserving natural motion characteristics and ensuring efficient task execution. This configuration effectively minimizes the adverse effects of both excessive motion constraints and insufficient regulation, ultimately yielding superior performance across our task suite. As shown in Figure C, we present the variation in the success ratio for the handover\_item\_easy task, under different coefficient configurations as the number of training steps increases.

(4) 系数 的影响。在策略学习中，将运动学正则化项纳入下一个最佳姿势学习目标中，以平衡任务有效性和运动约束。正则化强度由系数 控制，该系数决定了在学习过程中施加的运动学约束的程度。 的选择显著影响策略性能。较大的 值(接近1.0)对应于较弱的运动学约束，使得运动模式更加灵活。相反，较小的 值(例如0.1)施加了更严格的运动学约束，产生了更保守的策略，强调运动平滑性而非任务效率。如表E所示的实证结果表明，策略性能在 时达到最佳。对于较小的 值，学习到的策略表现出过于谨慎的行为，表现为优先考虑平滑性而牺牲任务效率和最优路径规划的轨迹。相反，当 接近1.0时，减弱的运动学约束导致运动模式缺乏调节，可能会影响轨迹的自然性和精确性。这些发现表明， 在保持自然运动特性和确保高效任务执行之间实现了更好的权衡。这种配置有效地最小化了过度运动约束和不足调节的负面影响，最终在我们的任务套件中产生了卓越的性能。如图C所示，我们展示了在不同系数 配置下，随着训练步骤的增加，handover\_item\_easy任务的成功率变化。

# D. Qualitative Analysis

# D. 定性分析

We show more qualitative result in Figure D. Through the novel approach of encoding robotic arm structural information as graph representations and explicitly incorporating kinematic constraints, our model demonstrates exceptional performance in motion symmetry, synchronization, and coordination across dual-arm manipulation tasks. In the lift\_ball experiment, the model achieves precise bilateral symmetry in spatial positioning through learned structured representations. The dual arms maintain stable symmetric configurations while preventing object instability through synchronized force application patterns, highlighting the efficacy of our structure-aware control paradigm. Furthermore, in the push\_box task, the model exhibits remarkable geometric symmetry in motion planning and execution. By leveraging embedded kinematic information, the robotic arms consistently maintain equidistant positioning relative to the target object’s center of mass while executing synchronized trajectories along parallel paths. This precise symmetrical control not only ensures operational stability during object manipulation but also establishes a robust framework for dual-arm cooperative control in complex manipulation scenarios.

我们在图D中展示了更多的定性结果。通过将机械臂结构信息编码为图表示并显式结合运动学约束的新颖方法，我们的模型在双臂操作任务中的运动对称性、同步性和协调性方面表现出色。在lift\_ball实验中，模型通过学习到的结构化表示实现了空间定位的精确双边对称性。双臂通过同步的施力模式保持稳定的对称配置，同时防止物体不稳定，凸显了我们结构感知控制范式的有效性。此外，在push\_box任务中，模型在运动规划和执行中表现出显著的几何对称性。通过利用嵌入的运动学信息，机械臂在执行沿平行路径的同步轨迹时，始终相对于目标物体的质心保持等距定位。这种精确的对称控制不仅确保了物体操作过程中的操作稳定性，还为复杂操作场景中的双臂协同控制建立了一个稳健的框架。

1. \*This work was done when and were intern at Huawei Noah’s Ark Lab. Corresponding author

   \*这项工作是在 和 在华为诺亚方舟实验室实习期间完成的。 通讯作者 [↑](#footnote-ref-28)
2. https://huggingface.co/openai/clip-vit-base-patch32

   https://huggingface.co/openai/clip-vit-base-patch32 [↑](#footnote-ref-91)