基于 Point Transformer 的条件扩散六自由度位姿生成

雷啸天 [1,2] 王菁 [1,2]

1.北方工业大学人工智能与计算机学院，中国·北京 100144

2.大规模流数据集成与分析技术北京市重点实验室，中国·北京 100144

摘要：随着深度学习的发展，基于点云的六自由度抓取感知与生成方法取得了显著进展。传统主要侧重于几何与物理稳定性评价，在面向自然语言指令驱动的抓取任务时，难以在候选生成阶段有效融入语义约束。为此，本文提出一种基于 Point Transformer 条件扩散六自由度抓取生成方法，利用 Point Transformer 对场景点云进行特征提取，在条件扩散去噪网络中引入交叉注意力机制，实现语言指令与点云特征的深层对齐与融合，进而在连续抓取空间中生成任务相关的抓取候选，设计联合一致性评分对候选抓取进行评估与筛选，在保证物理可执行性的同时提升语义匹配性与抓取完成质量。实验结果表明，该方法在语义一致性与抓取成功率等指标上优于对比方法，具备良好的泛化能力与应用潜力。

关键词：六自由度抓取；多模态融合；交叉注意力

Point Transformer–Based Conditional Diffusion for 6–DoF Grasp Pose Generation

Lei Xiaotian[1,2], Wang Jing[1,2]

1.School of Information Seience and Eingineering, North China University of Technology, China Beijing 100144

2.Beijing Key Laboratory of Large-scale Stream Data Integration and Analysis Technology, China Beijing 100144

Abstract: With the rapid development of deep learning, point cloud-based 6-DoF (Degrees of Freedom) grasp perception and generation methods have achieved remarkable progress. Traditional approaches predominantly focus on evaluating geometric and physical stability. However, for natural language instruction-driven grasping tasks, these methods struggle to effectively incorporate semantic constraints during the grasp candidate generation phase.To address this limitation, this paper proposes a Point Transformer-based conditional diffusion method for 6-DoF grasp generation. Specifically, a Point Transformer is utilized to extract features from the scene point cloud, and a cross-attention mechanism is introduced into the conditional diffusion denoising network to achieve deep alignment and fusion between language instructions and point cloud features. This enables the generation of task-relevant grasp candidates within a continuous grasp space. Furthermore, a joint consistency scoring mechanism is designed to evaluate and filter the generated candidates, thereby enhancing semantic matching and overall grasp execution quality while ensuring physical viability.Experimental results demonstrate that the proposed method outperforms baseline approaches in metrics such as semantic consistency and grasp success rate, exhibiting strong generalization capabilities and significant application potential.

Keywords: 6-DOF grasping; Multimodal fusion; Cross-attention

0 引言

机器人抓取[1]旨在为机械臂生成稳定的抓取位姿[2]。现有数据驱动[3]的方法侧重评估物理可行性，难以建模指令中的语义约束，导致抓取结果出现语义失配问题。在语言条件下，其局限性主要凸显于三个方面：（1）表征能力不足：传统点云编码方法对复杂几何及长程依赖关系的建模能力有限，导致抓取候选对位姿敏感的细粒度结构覆盖度不足；（2）条件融合不充分：语言特征多用于后验筛选或简单拼接，无法在抓取候选生成阶段持续发挥引导作用，难以响应偏好特定接触部位的细粒度任务需求；（3）生成范式存在缺陷：六自由抓取解空间具有连续且多峰的特性，传统固定模板采样或单次回归方法，难以在可接受的计算复杂度内，同时覆盖多样物理可行解与语义偏好解。

为此，本文提出一种基于 Point Transformer 的条件扩散六自由度抓取生成方法：首先，采用 Point Transformer 对场景点云进行几何编码，增强细粒度结构与长程依赖的表征能力，提升候选抓取的覆盖质量；其次，在条件扩散去噪网络中引入交叉注意力机制，将语言条件作为上下文信息与点云特征进行精准对齐，使语义约束贯穿候选生成全过程；最后，引入语义相关的一致性评分，与物理可行

性评分共同构建联合评分机制，用于抓取候选的筛选，在保障抓取可行性的同时，提升语言条件任务下的语义匹配性与任务完成率。

1 相关工作

现有六自由度（6-DoF）抓取检测方法经历了从“传统采样”到“端到端回归”，再到“语义融合”的演进，主要分为以下三类：基于采样的方法[4]。通常先在点云或深度图中生成候选抓取，再通过解析指标或学习模型对候选抓取质量进行评估，预测出一个最优的抓取位姿。然而，算法通常需要在高维空间中生成大量候选抓取并逐一打分，严重依赖人工设计的复杂特征提取过程。导致算法整体效率偏低，难以满足复杂场景下的抓取需求。

为此，研究逐渐转向端到端抓取方法[5]。端到端网络通过直接学习感知输入与抓取表示之间的映射关系，提高抓取检测的效率与鲁棒性。

近年来，借助大模型实现跨模态对齐[6]，将人类意图转化为抓取指令。实现了在间接指令下的物体抓取。但多将语义信息作为后验筛选或高层规划条件，与底层抓取生成过程相互割裂，难以形成统一的建模与优化框架。

2 方法

2.1 问题建模与整体框架

本文提出的抓取生成方法：先解耦文本指令 L 得到任务语义L_{sem}与物理语义L_{phy}，给定物体点云$O \in R^{N\times 3}$，构建联合表征并作为条件扩散模型输入，目标是在 6-DoF 空间生成语义合理且物理可执行的抓取位姿 G，最后由联合评分网络筛选与引导。整体流程如图 1 所示。

$$G = \text{Diffusion}(O, L_{\text{sem}}, L_{\text{phy}}) \tag{3}$$

其中 G 采用 6-DoF 抓取规范，对于一个平行夹爪抓取，6D 位姿通常描述的是夹爪参考坐标系相对于世界坐标系或相机坐标系的变换。其完整的抓取位姿 G 通常由相机坐标系下的旋转矩阵和平移向量表示：

$$G = (t, R) \tag{4}$$

$t \in \mathbb{R}^3$表示平移向量，$R \in SO(3)$表示旋转矩阵。

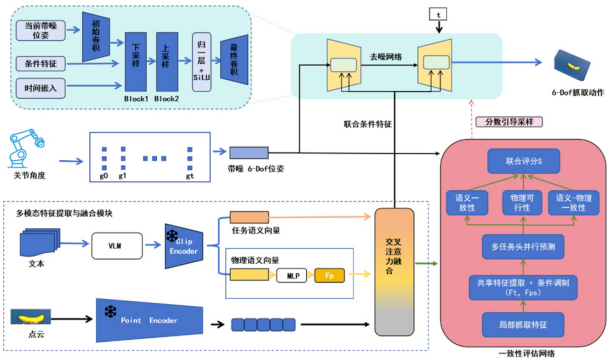

图1 总体流程图

2.2 文本指令的语义解耦

自然语言指令往往包含冗余描述，关键物理约束可能被稀释。为显式引入物理先验，我们利用 VLM 提取任务目标与隐含物理偏好，本文中选用 GPT4o[7] 为抓取生成提供语义条件分解，解耦流程如图 2 所示。

$$\{L_{task}, L_{phy}\} = \text{VLM}(L) \tag{5}$$

给定指令 L，GPT4o 将其解耦为任务语义描述L_{task}与物理语义描述L_{phy}，通过对齐语言语义与视觉物理特征，模型在生成抓取时兼顾语义与物理需求。

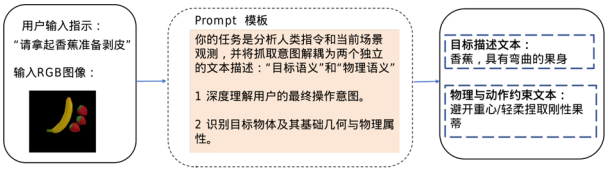

图2 语义与物理约束解耦示意图

2.3 基于点云编码与交叉注意力融合的条件抓取生成

2.3.1 多模态条件融合模块

点云几何编码器：给定输入点云 $P = \{p_i\}_{i=1}^{N}$，$p_i \in \mathbb{R}^3$，其中$p_i \in \mathbb{R}^3$表示空间坐标。几何编码器用于从原始点云中提取具有判别性的几何特征，以表征物体的局部形状与整体结构。本文采用 PointNet[8] 作为点云编码骨干网络，该结构能够在建模局部几何关系的同时捕获长程点间依赖，从而提升复杂几何结构的表达能力。通过该编码器，输入点云被映射为逐点特征表示：

$$F_p = \text{Enc}_{geo}(P), \quad F_p \in \mathbb{R}^{N\times d} \tag{6}$$

其中 d 为特征维度。得到的逐点特征同时编码局部几何信息与全局结构，为后续语义条件融合和抓取生成提供基础几何表示。

语义到物理语义变换模块：在机器人抓取任务中，抓取的物理可行性通常依赖于具体的任务语义。因此，本文设计了一种由语义驱动的物理语义表示方法，首先通过语义到物理语义的映射网络，将文本语义特征F_t投影到物理语义嵌入空间，得到对应的物理语义表示F_p，其定义如下：

$$F_{ps} = \text{MLP}_p(F_t), \quad F_{ps} \in \mathbb{R}^d \tag{7}$$

该表示不直接对应显式物理量，是用于隐式编码语义条件下的物理约束关系，从而在统一的嵌入空间中实现语义信息与物理信息的联合建模。

交叉注意力融合模块：为统一建模语义信息与物理约束，本文通过 Transformer 的交叉注意力机制对文本语义特征、物理语义嵌入与点云几何特征进行融合。

具体而言，在 U-Net 的中间层引入 Cross-Attention

模块，其中点云特征F_p经过线性映射得 到查询向量 Q，文本语义特征F_t与物理语义嵌入F_{ps}分别映射为键 K 和值 V。

注意力计算如下：

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad (8)$$

得到的融合特征用于后续抓取生成。

2.3.2 条件扩散模型模块

扩散模型是一类生成模型，它通过逐步“加噪”和“去噪”的过程来生成数据，为生成多样且符合语义约束的抓取姿态，本文将抓取生成建模为条件扩散过程。给定真实抓取向量 g_0，前向扩散通过逐步加入高斯噪声生成带噪序列 $g_1,...,g_T$:

$$q(g_t \mid g_{t-1})=\mathcal{N}\left(\sqrt{1-\beta_t}\, g_{t-1},\beta_t I\right) \quad (9)$$

任意时刻 t 的噪声样本为：

$$g_t=\sqrt{\bar{\alpha}_t}\, g_0+\sqrt{1-\bar{\alpha}_t}\,\epsilon,\quad \epsilon\sim\mathcal{N}(0,I) \quad (10)$$

在反向生成阶段，模型学习条件概率 $p(g_{t-1}\mid g_t,c)$，其中条件 c 包括点云几何特征F_p文本语义特征 false 以及物理语义嵌入F_{ps}。

在此基础上，本文将抓取生成建模为条件扩散过程。扩散模型以抓取参数 g_t 为生成变量，并在每 t 个时间步接收几何特征F_p、文本语义F_t以及物理语义嵌入F_{ps}作为条件输入，对噪声进行预测：

$$\epsilon_\theta(g_t,t\mid F_p,F_t,F_{ps}) \quad (11)$$

推理阶段从随机噪声逐步去噪生成候选抓取，并通过语义物理一致性评分对采样过程进行引导，从而生成语义合理且物理可行的抓取姿态。

2.4 语义 – 物理一致性评分筛选网络

为了评估候选抓取在语义与物理层面的合理性，本文提出语义 – 物理一致性评分网络，对生成的抓取位姿进行多维度评价与筛选，网络结构如图 3 所示。

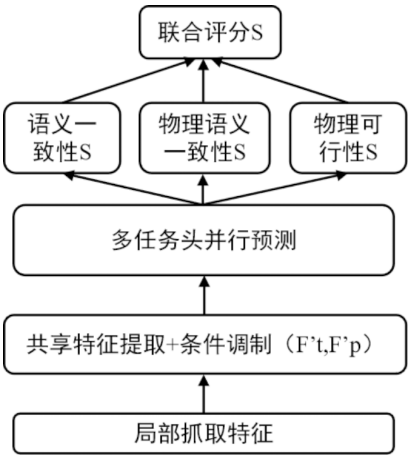

图3 评分网络整体架构

给定候选抓取 g，首先在其局部坐标系中提取抓取邻域点云P_g，并通过局部点云编码器 LE 获得抓取中心特征：

$$f_g=\text{LE}(P_g) \quad (12)$$

为了引入语义条件与物理语义约束，本文采用 FiLM 对局部特征进行条件调制。具体地，由文本语义特征F_t与物理语义嵌入F_{ps}预测调制参数（γ,β）并作用于局部特征：

$$f_{\text{cond}}=\gamma(F_t,F_{ps})\odot f_g+\beta(F_t,F_{ps}) \quad (13)$$

该机制能够根据任务语义和物理属性动态调整抓取区域特征的重要性，使网络能够感知在当前任务约束下的抓取偏好。

在条件调制后的特征基础上，评分网络采用共享主干与多头预测结构，分别输出语义一致性、物理可行性与语义 – 物理一致性评分：S_{sem} ，S_{phy} ，S_{cons} 最终通过加权组合得到统一抓取评分：

$$S=w_{\text{sem}}S_{\text{sem}}+w_{\text{phy}}S_{\text{phy}}+w_{\text{cons}}S_{\text{cons}} \quad (14)$$

该评分用于候选抓取的排序与筛选，并可作为生成阶段的引导信号，从而实现语义合理性与物理可执行性的统一优化。

3 实验

3.1 数据集

实验所使用的物体模型来自公开抓取数据集，基于 GraspNet 数据集并扩展语义任务指令，该数据集用于大量真实场景抓取检测中的物体数据，本文也基于 YCB 数据集扩展了场景表示，包含 66 个物体模型与 36 个语言关键词，关键词类别有：标签类、通用描述类、属性类、功能类。在仿真环境中，多个物体被随机放置于工作台上，物体姿态和位置在每次实验中随机生成，以保证实验的随机性。

3.2 实验设置与评价指标

3.2.1 实验设置

为了验证本文方法在语言条件抓取任务中的有效性，在仿真与真实机器人平台上开展对比实验、消融实验与真实环境抓取实验。

实验在统一的仿真环境与相同场景随机化设置下评估所有方法，每次评测随机生成物体姿态与位置，并在同一场景输入基础上进行抓取决策与执行。实验仿真环境采用 PyBullet 平台构建，实验计算平台采用搭载 NVIDIA GeForce RTX 2080 Ti GPU 的联想 P520 工作站，系统运行于 Ubuntu 20.04 操作系统。

3.2.2 评价指标

本文从尝试级与任务级两个维度，对语言条件驱动的机器人抓取性能进行系统性评估。设评测共包含 M 个语言任务，每个任务独立重复 R 次（本文取 R=15），单次任务最多允许 K 次抓取尝试（本文取 K=8）。针对第 i 个任务、第 r 次重复、第 k 次尝试，定义三类二值变量：抓取成功表征抓取操作的执行结果；物理可行性表征抓取位姿无碰撞且满足力学稳定性；语义一致性是基于 CLIP 特征相似度进行判定，表征抓取结果与指令的匹配程度。尝试级指标定义如下。

抓取成功率（GS）：

$$\mathrm{GS}=\frac{1}{MRK}\sum_{i=1}^{M}\sum_{r=1}^{R}\sum_{k=1}^{K}y_{i,r,k} \tag{15}$$

物理可行率（PF）：

$$\mathrm{PF}=\frac{1}{MRK}\sum_{i=1}^{M}\sum_{r=1}^{R}\sum_{k=1}^{K}f_{i,r,k} \tag{16}$$

语义条件一致性（SC）：

$$\mathrm{SC}=\frac{\sum_{i,r,k}y_{i,r,k}c_{i,r,k}}{\sum_{i,r,k}y_{i,r,k}+\epsilon} \tag{17}$$

任务完成率定义为：若某次任务重复 k 次，抓取同时满足“成功抓取且语义一致”，则该任务完成：

$$z_{i,r}=\mathbb{I}\left(\max_{1\le k\le K}y_{i,r,k}c_{i,r,k}=1\right) \tag{18}$$

$$\mathrm{TC}=\frac{1}{MR}\sum_{i=1}^{M}\sum_{r=1}^{R}z_{i,r} \tag{19}$$

综合评分用于总体对比：

$$\mathrm{OS}=w_{gs}\mathrm{GS}+w_{pf}\mathrm{PF}+w_{sc}\mathrm{SC}+w_{tc}\mathrm{TC} \tag{20}$$

3.3 方法比较

为了验证本文方法的有效性，我们将其与多种代表性抓取方法进行了比较，表 1 给出了不同方法在各项指标上的整体性能对比。

传统抓取方法在物理稳定性方面表现较好，但由于缺乏语言语义建模能力，其语义一致性明显较低，因此限制了在任务中的完成率。

任务条件抓取方法通过引入语义条件能够一定程度提升语义一致性，但由于未显式建模语义与物理之间的关系，其整体性能仍然存在一定局限。

相比之下，本文方法通过引入语义 - 物理一致性建模，并在抓取生成阶段引入一致性评分引导，使模型能够同时兼顾抓取稳定性与语义合理性，因此在语义一致性、抓取成功率以及任务完成率等指标上均取得最优性能。

表1 各项评价指标上的性能对比结果

Method	物理稳定率	语义一致率	抓取成功率	任务完成率	平均分数
GraspNet Baseline	87.3	42.5	82.1	48.6	65.1
GPD	84.6	39.8	79.4	45.2	62.3
Contact-GraspNet	89.5	46.1	85.7	52.8	68.9
Task-Conditioned Grasp	86.8	63.2	83.4	66.5	74.1
Diffusion Grasp (w/o guidance)	90.2	58.4	87.1	63.7	75.3
Ours	93.6	78.9	91.8	84.3	87.2

3.4 不同语义指令类型实验

为了分析不同语义类型对抓取性能的影响，我们将语言指令划分为四类：标签类、通用描述类、形状 / 颜色类和功能类。其中标签类指令直接描述目标物体类别，例如“抓取杯子”；通用描述类指令为通用目标描述，不设定格式限制；形状 / 颜色类指令描述物体外观属性；功能类指令描述物体用途或功能，例如“抓取用于倒水的物体”。

图 4 给出了不同语义指令类型下的任务完成率结果。实验结果表明，当语义表达较为简单时（如 Label 类指令），大多数方法均能够取得较好的抓取成功率。然而随着语义复杂度增加（如 Function 类指令），传统方法的性能显著下降。

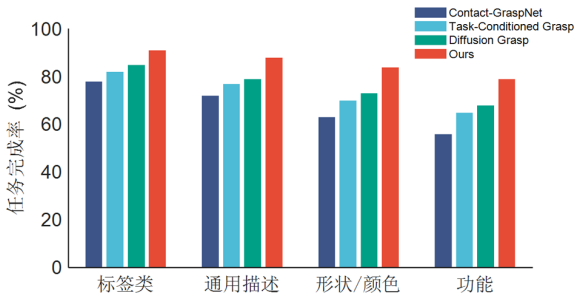

图4 不同语义指令类型下的任务完成率结果

相比之下，本文方法在复杂语义条件下仍能够保持较高的任务完成率，说明语义物理一致性建模能够有效提升模型对复杂任务语义的理解能力。

为了分析模型各模块对整体性能的贡献，我们进行了模块消融实验。实验结果如表 2 所示。

从实验结果可以看到，当仅使用基础扩散抓取模型时，模型能够生成物理可行的抓取姿态，但语义一致性较低，因此任务完成率下降。加入语义嵌入后，模型开始能够利用语义信息进行抓取筛选，从而语义一致性分数得到提升。进一步加入语义到物理映射模块后，模型能够更好地理解语义与抓取物理行为之间的关系。当模型继续引入 FiLM 条件调制模块后，语义信息能够更加有效地作用于抓取局部特征，使模型对抓取区域语义偏好的建模能力得到

表2 不同模块对抓取性能影响的消融实验结果

Variant	Semantic Mapping	FiLM Modulation	Consistency Head	Score Guidance	Semantic Consistency	Grasp Success
Baseline Diffusion	×	×	×	×	52.3	86.7
+ Semantic Embedding	√	×	×	×	61.8	87.2
+ Semantic Physical Mapping	√	×	√	×	70.5	88.3
+ FiLM Modulation	√	√	√	×	74.2	89.5
+ Score Guidance (Full)	√	√	√	√	78.9	91.8

增强。在生成阶段引入一致性评分引导后，模型能够优先选择语义与物理一致性更高的抓取姿态，从而显著提升任务完成率。

4 结语

针对六自由度生成任务中缺乏对语义合理性与物理可行性之间一致性关系的系统建模问题，基于 Point Transformer 条件扩散六自由度抓取生成方法，通过构建语义与物理属性的统一表达，并引入联合评分机制，实现了抓取动作逻辑合理与物理可行的协同优化。实验结果表明，该方法在语义一致性、任务完成率等指标上优于对比方法，具备良好的可行性与应用潜力。

尽管本文方法在实验中表现良好，但是仍存在一定的局限性。在物理因素方面缺乏对三维空间几何深刻理解，没有将更高维的信息统一建模，在后续实验中，将考虑更高维度的灵巧手抓取问题。以更好地应对环境中多样物品的抓取，进一步扩展抓取任务的实用性。

参考文献：

[1] Su á rez-Ruiz F, Zhou X, Pham Q C. Can robots assemble an IKEA chair?[J]. Science Robotics, 2018, 3(17): eaat6385.

[2] Dantam N T, Kingston Z K, Chaudhuri S, et al. Incremental task and motion planning: A constraint-based approach[C]//Robotics: Science and systems. 2016, 12: 00052.

[3] Ten Pas A, Gualtieri M, Saenko K, et al. Grasp pose detection in point clouds[J]. The International Journal of Robotics Research, 2017, 36(13-14): 1455-1473.

[4] Liang H, Ma X, Li S, et al. Pointnetgpd: Detecting grasp configurations from point sets[C]//2019 international conference on robotics and automation (ICRA). IEEE, 2019: 3629-3635.

[5] Fang H S, Wang C, Gou M, et al. Graspnet-1billion: A large-scale benchmark for general object grasping[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11444-11453.

[6] Li M, Zhao Q, Lyu S, et al. Ovgnet: A unified visual-linguistic framework for open-vocabulary robotic grasping[C]//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024: 7507-7513.

[7] Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.

基金项目：国家自然科学基金国际（地区）合作与交流项目（编号：62061136006）资助。

作者简介：雷啸天（2000.10-），男，汉族，陕西省渭南市人，北方工业大学人工智能与计算机学院，硕士在读，研究方向：人工智能，具身智能机器人。