家庭服务场景 3DGS 重建与人形机器人自主作业仿真

闫奕菲[1]，朱林青[1]，叶林奇[1]，邢琳[2]，冯瑞[3]
（1. 上海大学未来技术学院，上海市 200444；2. 上海电器科学研究院，上海市 200063；3. 复旦大学计算与智能创新学院，上海市 200438）

摘要：针对具身智能任务人形机器人测试成本高、真实场景难复现、传统仿真环境真实感差的问题，为实现机器人自主家庭服务，*采用三维高斯溅射（3D Gaussian Splatting, 3DGS）技术构建具有较高视觉真实感的家庭仿真场景，在 Unity 仿真环境中部署宇树 G1 人形机器人模型并完成感知、导航与抓取的全流程仿真适配，搭建端到端基于关键词匹配的语音指令驱动自主作业系统，实现自然语言指令直接驱动机器人完成找物、导航、识别和抓取全流程家庭任务。*实验对 Bottle、Tomato、Cup 和 Apple 四类物体进行抓取-递送任务测试，共完成 260 次有效实验，记录成功率和各阶段耗时数据。结果表明，所构建场景能够较好保持真实家庭环境的纹理、光照和空间布局特征，并支撑完整机器人作业流程验证；四类物体的抓取与递送表现存在差异，结合实验现象推测，该差异可能与物体尺寸和几何形状有关。该仿真平台可用于家庭服务具身智能算法验证与任务仿真测试，为后续迁移提供仿真基础。

关键词：3D 高斯溅射；具身智能；Unity 仿真；人形机器人；语音指令解析；自主抓取
中图分类号：TP242.6；TP391.41　　文献标志码：A　　文章编号：xxxx-xxxx (2026) xx

3DGS Reconstruction and Autonomous Operation Simulation of Humanoid Robots in Home Service Scenarios

Yifei Yan[1]，*Linqing Zhu*[1]，Linqi Ye[1]，*Lin Xing*[2]，*Rui Feng*[3]
(1.School of Future Technology, Shanghai University, Shanghai 200444, China; 2. Shanghai Electrical Apparatus Research Institute, Shanghai 200063, China; 3. School of Computing and Intelligent Innovation, Fudan University, Shanghai 200438, China)

Abstract: To address the high testing cost of humanoid robot evaluation in scenes, and the limited realism of traditional simulation environments, *this study constructs a home simulation scene with high visual realism using 3D Gaussian Splatting (3DGS) for autonomous home service by humanoid robots. The Unitree G1 humanoid robot is deployed in the Unity simulation environment and adapted for the full process of perception, navigation and grasping. An end-to-end keyword-based voice command-driven autonomous operation system is developed, enabling natural language commands to directly drive the robot to perform household tasks including object search, navigation, recognition and grasping.* Pick-and-delivery experiments are conducted on four object categories, namely Bottle, Tomato, Cup, and Apple, with 260 valid trials in total, recording success rates and phase-level execution times. The results show that the constructed scene can effectively preserve the texture, illumination, and spatial layout characteristics of the real home environment, and can support complete robot task execution. Differences in grasping and delivery performance are observed among the four object categories, which may be

收稿日期：2026-0605　　修回日期：2026-07-21
基金项目：上海市科委项目“具身本体通用运动框架与多模态融合感知关键技术”（项目编号：24511103304）
第一作者：闫奕菲(2004-)，女，汉族，本科，本科，研究方向为机器人工程。
通讯作者：叶林奇(1992-)，男，汉族，副教授，博士，研究方向为具身智能。
论文视频附件链接：https://linqi-ye.github.io/video/3dgs.mp4

associated with object size and geometry.The proposed simulation platform can support embodied intelligence algorithm validation and task-level simulation testing in home service scenarios, providing a simulation foundation for future transfer.

Keywords: 3D Gaussian Splatting; embodied intelligence; Unity simulation; humanoid robot; voice command parsing; autonomous grasping

0 引言

人形机器人进入家庭服务场景是具身智能领域的长期研究目标。与结构化工业环境不同，家庭场景具有非结构化布局、动态变化和物体种类繁多等特点，对机器人的感知精度、导航鲁棒性和操作灵活性提出了极高要求[1]。目前，家庭服务机器人的研发测试面临两重核心挑战。

其一，真实家庭场景测试成本高、耗时长且难以标准化复现。不同住宅户型、家具布局和光照条件使得在真实环境中进行大规模、可重复的机器人实验几乎不可行。因此，构建高保真仿真测试平台成为支撑家庭机器人算法研发与验证的重要技术路径。其二，传统机器人仿真环境多依赖手工几何建模与纹理贴图构建场景，存在纹理重复单一、光影生硬、细节缺失等问题，导致渲染效果与照片级真实感之间存在显著差距。Bono 等[2]指出，现有仿真器偏重几何精度而渲染真实感不足，严重影响视觉感知类算法在仿真环境中的验证可信度。

三维高斯溅射（3D Gaussian Splatting, 3DGS）是近年来兴起的新型场景表示与渲染技术。Kerbl 等[3]于 2023 年在 SIGGRAPH 首次提出该技术，其核心思想是通过一组带可学习参数的各向异性三维高斯球体显式表示场景的几何与外观信息，并借助基于瓦片的高效可微光栅化算法实现实时照片级新视角合成。与传统的神经辐射场（NeRF）隐式表示相比[4]，3DGS 在渲染速度与重建质量之间取得了显著优势，可在消费级 GPU 上实现 1080p 分辨率下不低于 30 FPS 的实时渲染。Zhu 等[5]在 3DGS 机器人应用综述中系统梳理了该技术在场景理解、操作交互、导航规划等方面的研究进展，指出 3DGS 的显式几何表达和实时渲染特性使其具有构建机器人仿真环境的天然优势。在 3DGS 仿真与机器人操作的交叉领域，近期涌现了 PhysGaussian[6]、GaussianGrasper[7]、ManiGaussian[8]等一系列探索性工作，在物理仿真融合、开放词汇抓取规划、多任务操作等方向取得了重要进展。在仿真平台层面，RoboGSim[9]首次将 3DGS 与物理引擎结合，实现了照片级场景重建到数字孪生的闭合链路，验证了 3DGS 构建机器人仿真环境的可行性。然而，上述工作多聚焦于单一操作任务，将 3DGS 应用于构建完整的具身智能仿真测试平台，并实现从场景重建到任务执行的完整链路集成，相关系统级工作仍较为少见。

在移动机器人导航与避障方面，Tang 等[10]系统综述了全局路径规划与局部避障算法，为本文导航模块的全局-局部混合策略设计提供了参考。在数字孪生与仿真平台构建方面，Wang 等[11]基于 Unity3D 构建了果园数字孪生虚拟交互系统，验证了 Unity3D 在农业场景中的有效性，与本文采用 Unity3D 构建家庭服务仿真平台的方法论一致。

在交互层面，随着大语言模型和云端语音服务的快速发展，通过自然语言指令直接驱动机器人执行任务成为新的交互范式[12]。当前中文云端语音服务在近场普通话场景下已具备高识别准确率和低响应延迟，使得实时语音驱动机器人作业成为可能。现有语音机器人系统如 SayCan[13]、Robi Butler[14]等多聚焦于轮式移动平台或机械臂方案，将自然语音交互与人形机器人全身自主作业进行系统级集成与验证的工作尚不多见。

针对上述问题，本文设计并实现了一套基于 3DGS 具有较高视觉真实感的仿真场景的人形机器人语音指令驱动自主作业系统。本文的主要贡献包

括：（1）利用 3DGS 技术重建具有真实家庭外观特征的仿真场景，改善传统手工建模场景中纹理、光照和生活化细节表达不足的问题，为家庭服务任务提供更接近真实外观的视觉环境；（2）在上述仿真场景中部署宇树 G1 人形机器人模型，完成感知、导航、识别与抓取的全流程仿真适配；（3）搭建端到端基于关键词匹配的语音指令驱动自主作业系统，实现自然语言指令驱动机器人自主完成找物、导航、识别和抓取全套家庭任务，形成从场景重建、机器人部署到任务执行的完整链路。实验以 Bottle、Tomato、Cup 和 Apple 四类常见家庭物体为测试对象，通过对成功率、各阶段耗时等多维度指标的记录与分析，验证系统在仿真环境中的作业性能与可靠性。

1 基于 3DGS 的家庭高保真仿真场景构建

1.1 家庭场景数据采集

本文以真实家庭室内空间作为 3DGS 重建对象, 采集范围覆盖客厅、餐桌、厨房台面、卫生间和室内通道等区域。这些区域分别对应语音指令、目标搜索、递水、拿取水果和自主导航等任务环节, 因而数据采集不只记录房间整体布局, 还重点保留桌面物体、杯具、水果、柜体边缘、通道宽度和机器人可停靠观察点等细节。通过在采集阶段引入任务需求, 后续重建场景能够同时服务于视觉真实感展示和机器人作业测试。

采集时采用围绕式多视角拍摄方式, 相机沿房间边界、桌面周围和机器人可能移动路径连续取景, 保证相邻图像之间具有足够重叠区域。对于水杯、水瓶、水果等后续抓取目标, 额外补充近距离视角和不同高度视角, 以覆盖机器人头部相机、站立观察视角和手臂接近目标时的局部视角。采集过程中尽量保持曝光稳定和运动连续, 避免快速移动造成模糊, 同时减少人员走动、物体临时移动和大面积遮挡对重建一致性的影响。

图像采集设备为小米 14，原始拍摄模式为 4K、60 frame/s。当前用于 3DGS 重建的导出视频采用 HEVC 编码，分辨率为 1920×1080，帧率为 60 frame/s，时长约 205.78 s，共包含 12434 帧可选图像。本文采用 Postshot 完成训练图像筛选、相机位姿估计和 3DGS 训练。Postshot 采用 Use Best Images 策略，从原视频中选择 412 张图像作为训练输入；相机位姿采用 Compute From Images 模式根据训练图像自动估计，并启用相机位姿与点云中心化。

除 Postshot 选择的 412 张训练图像外，本文从原视频中人工筛选 205 张候选图像，用于检查客厅、餐桌、厨房台面、卫生间和卧室等区域的场景覆盖情况，并从中选择 5 张代表性关键帧用于论文展示和探索性图像评价。人工筛选的 205 张候选图像不作为 Postshot 训练图像数量统计。

原始图像采集完成后, 对图像序列进行筛选和整理, 剔除明显模糊、曝光异常、重复度过高或重叠区域不足的图像, 保留能够反映空间结构和纹理细节的有效视角。与此同时, 记录机器人初始位置、典型目标位置、可通行区域、桌面操作区域和障碍物边界等辅助信息。这些信息一方面用于 3DGS 重建前的数据组织, 另一方面为后续导航区域划分、碰撞体构建和抓取目标标注提供依据。

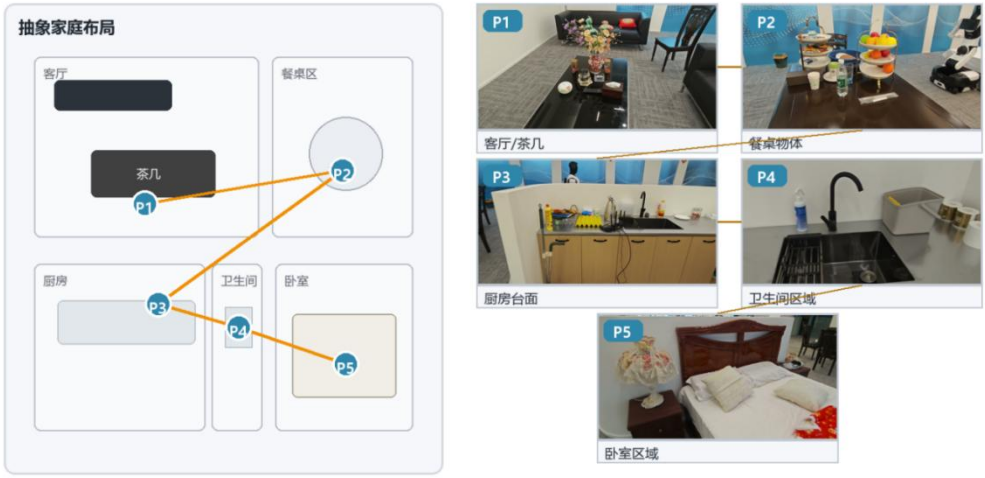

图 1　场景图像采集路径示意图
Fig. 1　Scene Image Acquisition Path Schematic

1.2 3DGS 场景重建、纹理与光影优化

3DGS 家庭场景重建流程如图 2 所示。

图 2　3DGS 家庭场景重建流程
Fig. 2　Workflow of 3DGS Household Scene Reconstruction

本文采用 Postshot 完成家庭场景 3DGS 重建。训练配置选择 Splat3，输入图像长边限制为 1920 像素，Max Splat Count 设置为 3000，启用 Anti-Aliasing，最大训练步数设置为 30 kSteps。Postshot 首先根据筛选后的多视角图像估计相机位姿并初始化场景结构，随后通过比较渲染图像与输入图像，迭代优化三维高斯基元的位置、尺度、旋转、不透明度和颜色参数。与人工网格建模和贴图不同，3DGS 直接从真实图像中学习场景的颜色、纹理和光照信息，更适合保留家庭环境中的复杂材质、局部遮挡和生活化细节。

家庭场景中的透明、反光和细长结构容易影响重建质量，例如玻璃杯、水瓶、金属边缘和水果高光区域可能出现局部模糊、漂浮点或边界断裂。针对这些问题，本文在训练前剔除明显模糊和曝光异常的图像，并提高桌面目标、机器人观察点和狭窄通道附近的视角覆盖密度。训练完成后，将 3DGS 资源导入 Unity 并完成尺度与坐标对齐。3DGS 资源主要承担真实家庭外观渲染，机器人导航、物理碰撞和抓取接触则由轻量化碰撞体、NavMesh 及独立刚体模型实现。

真实采集图像与对应的 3DGS 渲染结果如图 3 所示。重建结果能够保持主要家具布局、物体颜色和室内光照趋势，但在反光表面、细长结构和局部边界处仍存在一定模糊和伪影。

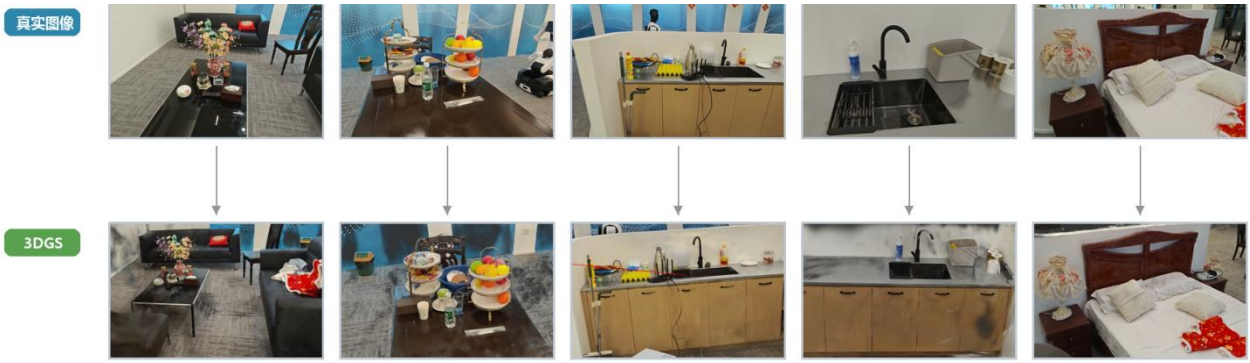

图 3　真实采集图像与 3DGS 渲染结果对比
Fig. 3　Comparison between Real Captured Images and 3DGS Rendering Results

为初步评价当前展示图的视觉一致性，本文从人工筛选的 205 张候选图像中选取客厅、餐桌、厨房、卫生间和卧室各 1 张代表性实拍图，并导出对应区域的近似视角 3DGS 渲染图。计算前将图像中心裁剪并统一缩放至 512×512 像素。PSNR 根据 RGB 均方误差计算，SSIM 采用 11×11 高斯窗口，LPIPS 采用 AlexNet 特征网络。五类场景的评价结果依次为：客厅 9.87 dB、0.2477、0.6919；餐桌 10.45 dB、0.3531、0.7162；厨房 12.53 dB、0.4841、0.7634；卫生间 12.85 dB、0.5717、0.7430；卧室 12.52 dB、0.5227、0.6074，其中三项数值分别对应 PSNR、SSIM 和 LPIPS。五组平均值分别为 11.64 dB、0.4359 和 0.7044。厨房、卫生间和卧室的 SSIM 相对较高，客厅和餐桌相对较低，这可能与视角偏差、目标遮挡、反光表面及光照变化有关；较高的 LPIPS 也表明实拍图与渲染图在感知特征层面仍存在差异。

需要说明的是，评价图像来自与训练数据相同的原始视频，无法完全排除其与 Postshot 选择的 412 张训练图像重合；渲染图也采用人工调整的近似视角，而非严格相同的相机内外参。因此，该结果不属于标准训练—留出划分下的同位姿评价，仅用于辅助分析当前展示图的视觉一致性，不作为照片级重建精度或方法性能对比的充分证据。

1.3 仿真场景适配与轻量化处理

3DGS 重建的原始场景仅支持新视角图像渲染，不具备碰撞检测、路径规划等物理交互能力。本文对场景进行了以下适配：

（1）碰撞与导航网格生成：3DGS 重建结果主要用于真实场景视觉渲染，并不直接提供刚体碰撞和路径规划所需的物理结构。为支持机器人导航、避障和抓取交互，本文采用“静态场景简化、可交互物体精细化”的分层碰撞建模方式构建物理交互层。对于墙体、柜体、桌面边界和大部分规则静态障碍物，采用 BoxCollider 进行近似；对于圆形桌子等近似轴对称结构，采用 CylinderCollider 进行近似。该方式能够降低碰撞检测和 NavMesh 烘焙复杂度，同时保留机器人导航避障所需的主要空间约束。对于水瓶、番茄、杯子和苹果等任务相关可交互物体，则采用独立刚体模型和较精细的网格碰撞体，以保证抓取、夹持、运输和释放过程中的接触判断稳定性。

在完成碰撞体配置后，基于 Unity NavMesh 系统烘焙可行走区域。由于本文场景为家庭室内平地环境，地面坡度变化较小，NavigationAgent 参数设置为 radius 0.25 m、height 1.3 m、step height 0.1 m、max slope 10°。其中 NavMeshAgent 主要用于生成导航参考路径，并不直接等同于 G1 机器人完整身体包络；机器人实际运动由 PID 路径跟踪控制、行走策略和 Unity 物理碰撞体共同约束。烘焙完成后，本文检查机器人初始点、桌面抓取点和用户投递点之间的路径连通性，并通过后续 260 次取物递送实验验证该分层碰撞建模与 NavMesh 适配方式能够支撑导航、抓取、运输和投递任务。

（2）语义区域标注：以统一的命名规范管理场景中的功能区域（如桌面、用户站立点），每个区域包含位置、朝向、对齐方式及安全停靠距离等属性。导航模块通过区域名称即可解析目标坐标，使语音模块输出的地点关键词可直接映射至空间位置，实现从语言到空间的端到端贯通。

（3）物体随机化：为支持可重复实验，在每轮任务开始前在一定区域内随机放置目标物体，保证物体间最小间距，并同步调整对应的机器人抓取点，确保抓取点始终位于机器人可达工作空间内。

2 人形机器人家庭作业系统设计

2.1 系统整体架构

本文系统采用“交互层—调度层—执行层”三层架构，如图 4 所示。

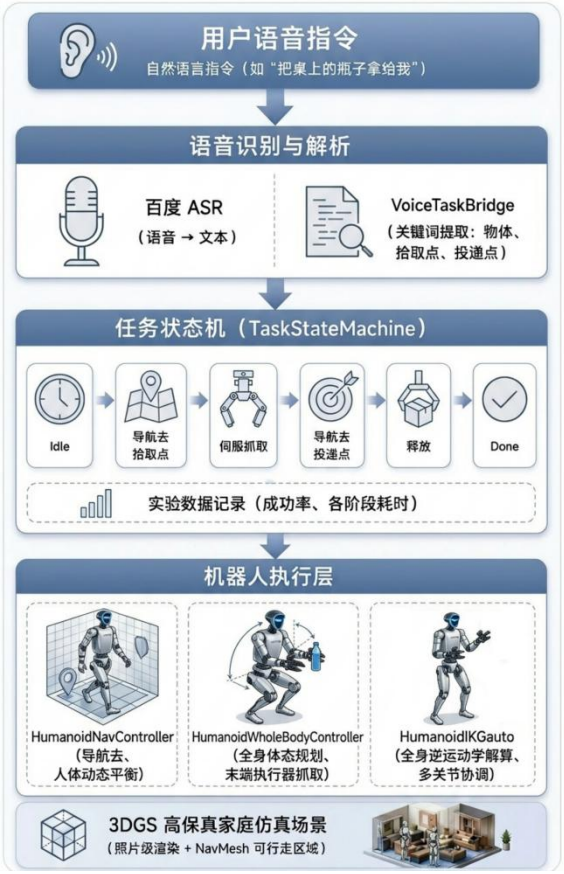

图 4 系统总体架构图

Fig. 4 Overall system architecture

交互层负责语音到任务指令的转化：云端 ASR 完成语音识别，本地 VoiceTaskBridge 解析出目标物体、拾取地点与投递地点，形成任务三元组。调度层以有限状态机为核心，接收任务指令后按预设状态序列依次调度下层模块，管理状态转换时机与异常处理，并实时记录实验数据。执行层由导航控制、抓取伺服、IK 位姿控制三个模块组成，分别负责空间移动、物体抓取与全身关节协调。三层间数据流单向传递，松耦合接口便于各层独立调试与替换。

2.2 人形机器人仿真模型导入与参数配置

宇树 G1 人形机器人 URDF 模型导入 Unity 后，自动构建包含关节链的 GameObject 层级树。系统遍历所有铰接体组件，按关节类型分别归入臂体执行关节和手部夹爪关节。

步态控制：双足行走采用前馈-反馈架构，前馈部分由正弦相位振荡器生成的周期性步态信号驱动腿部关节，反馈部分通过 ML-Agents 框架训练的强化学习策略实现。策略以 38 维观测向量（身体姿态、关节位置与速度、速度指令、步态相位等）为输入，输出 12 维连续动作用于下肢关节目标修正量，经平滑处理后与前馈叠加后由 ArticulationBody 关节驱动器执行，导航模块设定的线速度与角速度指令自动调节步幅与步频。

步态策略的训练参数如下：训练在 Unity 2022.3 PhysX 物理引擎中进行，物理仿真步长 0.01 s，对应控制频率 100Hz，32 个机器人并行训练。观测空间为 38 维向量，包含体坐标系重力方向（3 维）、角速度（3 维）、线速度（3 维）、12 个下肢关节位置（12 维）和速度（12 维）、期望指令$(vr,\omega r,cr)$（3 维）以及步伐相位sin/cos编码（2 维）。动作空间为 12 维连续向量，对应下肢关节目标修正量，输出经 EMA 平滑（系数 0.9）后驱动关节。

训练采用 PPO 算法，网络结构为 3 层隐藏层×512 单元，输入归一化，最大训练步数 20M，batch_size 2048，buffer_size 20480，学习率 3×10^{-4}，PPO clip epsilon 0.2，entropy beta 0.005，GAE lambda 0.95，discount gamma 0.995。每步奖励由以下各项之和构成：存活奖励(1.0)、横滚惩罚（-0.1$|\phi|$）、俯仰惩罚（-0.1$|\theta|$）、角速度跟踪（1-2$|\dot{\psi}-\omega r|$）、前向速度跟踪（1-2$|v_z-vr|$）、侧向抑制（-$|v_x|$），episode 在横滚角或俯仰角超过 30°时终止。

为验证 PPO 步态策略训练过程的稳定性，本文进一步记录训练过程中的累计奖励、episode length、policy loss 和 policy entropy，结果如图 5 所示。训练初期累计奖励和 episode length 快速上升，约 1000 万步后进入相对稳定区间；policy loss 保持在较低范围内波动，策略熵随训练逐步下降并趋于平稳，说明训练得到的步态策略已具备较稳定的仿真行走控制能力。

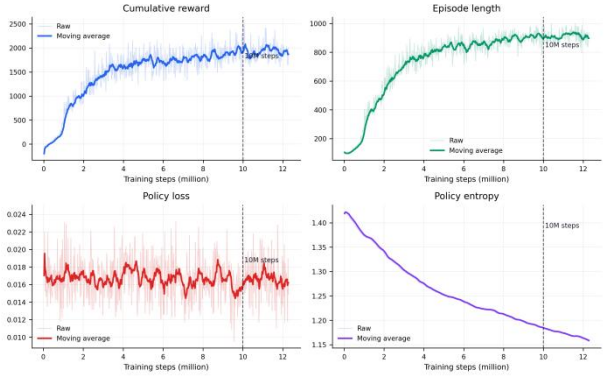

图 5　G1 步态策略 PPO 训练曲线

Fig. 5　PPO training curves of the G1 locomotion policy

IK 位姿控制：双臂末端位姿由逆运动学系统独立控制，通过设定手部相对于身体坐标系的目标偏移量实现空间定位。身体高度由蹲起参数统一调节，站立、下蹲、深蹲等姿态随任务阶段平滑切换。

夹爪配置：双手装配平移式两指夹爪，闭合目标值按物体类型独立预设以适配不同尺寸，开合通过目标关节值的平滑过渡实现，避免阶跃指令引发的抓取冲击。

2.3 语音指令解析模块

人形机器人在家庭服务场景中面向非专业用户，自然语言是最直观的交互方式。本文设计的语音指令解析模块采用“云端识别+本地解析”的两级架构，将用户的中文口语指令转化为机器人可执行的任务规约。该模块定位为基于关键词匹配的轻量级语音指令接口，通过预定义的物体与地点关键词词典实现意图解析，不具备连续对话和指代消解能力。

2.3.1 模块架构

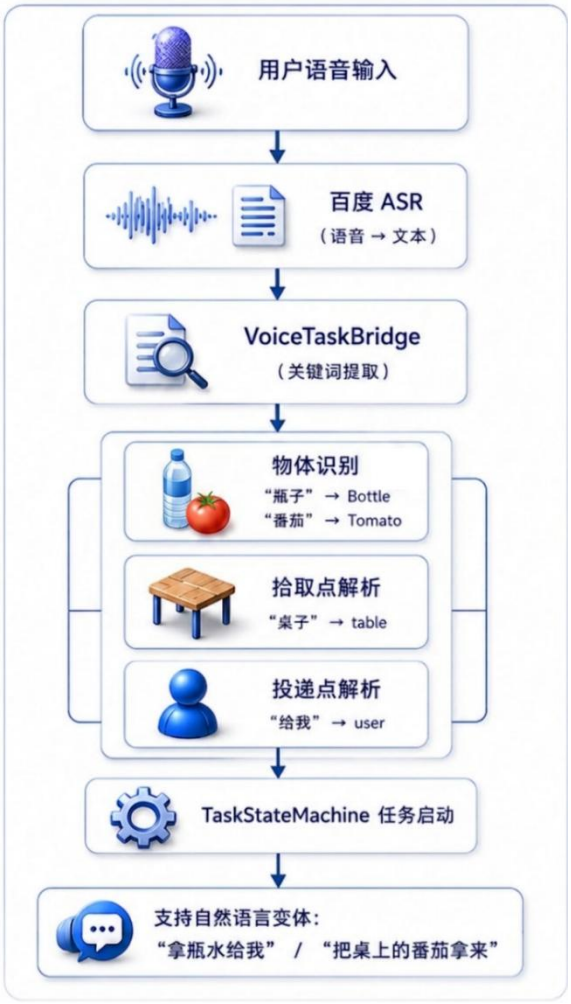

图 6　语音指令解析模块架构

Fig. 6　Architecture of the voice command parsing module

模块整体架构如图 6 所示。音频采集与语音识别由云端 ASR（Automatic Speech Recognition）服务承担，识别结果以文本形式返回 Unity 场景；本地语音任务桥接器负责轮询识别结果、解析任务意图并触发任务执行。该架构将计算密集的声学建模与语言模型推理置于云端，本地仅需轻量级文本匹配，兼顾了识别准确率与响应实时性。

2.3.2 语音识别输入

ASR 组件以独立 GameObject 形式挂载于 Unity 场景中，持续采集麦克风输入并周期性地将识别结果写入可读文本字段。考虑到家庭环境中可能存在背景噪声、无效拾音等干扰，桥接器在每帧轮询时执行三重过滤：

（1）滤除空字符串及与上一帧相同的重复结

果，防止同一条识别结果被多次触发；

（2）滤除 ASR 服务返回的中间状态提示文本（如“录音中”、“识别中”等），仅保留最终识别结果；

（3）当机器人正在执行任务（即有限状态机处于非 Idle 状态）时，拒绝接受新指令，并向用户反馈“任务进行中，请稍候”，避免并发冲突导致的不可预期行为。

2.3.3 自然语言指令解析

在获得用户的中文自然语言文本后，桥接器需从中提取三项关键信息：目标物体、拾取地点、投递地点。考虑到家庭服务场景中用户指令的句式相对固定（如“帮我把桌上的水瓶拿给我”、“去桌子上给我拿个番茄”），本文采用基于关键词字典与介词模式匹配的规则解析方法，而非训练大规模语言模型，以降低部署复杂度并保证实时性。

2.4 机器人自主导航、目标识别、抓取流程

前文已构建 3DGS 高保真仿真场景并部署 G1 机器人模型，在上述仿真环境中实现机器人自主完成找物-导航-抓取-递送全套任务还需要一个将语音指令转化为连续动作执行的控制系统。本文设计了以有限状态机为顶层调度框架、以 NavMesh 路径规划与 PID 路径跟踪控制律实现导航、以基于位置的闭环抓取伺服实现精准抓取、以逆运动学位姿控制协调全身动作的分层控制架构。

2.4.1 有限状态机总体设计

系统核心调度采用有限状态机，将一次完整的取物-递送任务抽象为 6 个顺序状态：空闲（Idle）、前往抓取点（GoingToPickup）、抓取中（Picking）、前往投递点（GoingToDeliver）、投递中（Delivering）、完成（Done），如图 7 所示。

图 7 取物递送任务流程图

Fig. 7 Pick-and-Delivery Task Flowchart

正常任务流程沿主链路单向推进：Idle 状态接收语音模块下发的任务指令后初始化物体专属姿态参数并转入 GoingToPickup 状态；导航至目标抓取点附近后触发到达回调，转入 Picking 状态；抓取成功后转入 GoingToDeliver 状态；导航至投递点后转入 Delivering 状态执行物体释放；释放完成后经高度校验确认物体已在桌面上方，转入 Done 状态并最终回至 Idle，等待下一次任务指令。

2.4.2 自主导航模块

机器人导航采用“NavMesh 全局路径规划 + PID 路径跟踪控制”的混合策略。Unity NavMesh 基于与 3DGS 视觉场景对齐的轻量化几何代理和

碰撞体烘焙可行走区域，并承担全局路径规划职责；实际行走速度指令由独立 PID 路径跟踪控制器生成，以避免使用 NavMeshAgent 直接驱动机器人模型而导致步态失稳。

底层运动控制采用带安全约束的比例控制律：角速度由航向偏差比例生成并限幅，线速度由目标距离比例生成并限幅。系统施加三项安全约束：

（1）航向偏差超过设定阈值时线速度强制归零，优先原地转身；

（2）线速度随航向偏差增大而线性衰减，转弯时自动减速；

（3）角速度与线速度互锁抑制，防止高线速与高角速并存导致摔倒。

导航模块内置卡死检测与脱困机制：当指令速度大于阈值而位移与航向变化持续低于阈值时判定为卡死，触发三段式脱困——先后退、再原地旋转、再前进，随后重新规划路径。终端接近采用“预对齐—直线逼近”两步法：在目标前方安全距离处停步、转向对准目标方向，再以低速直线逼近，直至满足到达判定条件。为了简化抓取流程，抓取位置引入横向补偿偏移，使目标物体位于右手正前方工作空间内。

针对家庭场景中的动态物体如行人等，导航模块在全局路径跟踪基础上增加局部行人检测与让行机制。当行人进入机器人导航区域时，机器人优先沿当前朝向后退；后方障碍达到安全距离或发生接触时停止并等待；当行人离开机器人前方区域后，系统重新计算至原目标的 NavMesh 路径并恢复任务。

2.4.3 基于位置的闭环抓取伺服模块

到达抓取点后，机器人需完成对目标物体的精准抓取。由于机器人导航停止位置、预设抓取姿态和逆运动学求解均可能产生末端残余位置误差，仅依赖预设姿态的开环抓取容易出现夹爪与目标未充分对齐的问题。本文设计了基于夹爪中心位置实时反馈的位置闭环抓取伺服策略：以右夹爪中心世界坐标为反馈量、目标物体世界坐标为参考量，通过增量式 IK 偏移迭代修正夹爪位置，使夹爪中心逐步逼近物体质心。

抓取过程分为三个阶段。阶段 0 初始化——张开夹爪，加载物体对应的预抓取姿态，记录右手 IK 基准位置并后缩防碰撞。阶段 1：安全对齐——以物体前方固定安全间隙为目标，在身体局部坐标系下伺服 X、Y、Z 三轴偏移量，使手爪在保持安全距离的前提下与物体水平对准。阶段 2：质心精准对齐——三轴同步向零误差收敛，伺服增益和步进限幅逐级提升，使夹爪中心与物体质心精确重合。阶段 3 闭合验证——夹爪闭合过程中 XZ 方向伺服持续修正，闭合稳定后通过距离检测与碰撞检测双重判据判定抓取成败，成功则保持夹持并进入运输阶段，失败则上提机械臂并重新尝试抓取。

3 仿真实验与结果分析

3.1 实验环境与参数

实验平台运行于配备 Intel Core i7 处理器、32 GB 内存及 NVIDIA GeForce RTX 4060 Laptop GPU 的计算机上，操作系统为 Ubuntu 22.04。仿真环境基于 Unity 2022.3 构建，3DGS 场景在 Unity 中采用高斯基元投影与混合方式进行实时前向渲染。在 Ultra 画质、1920×1080 分辨率及垂直同步开启的条件下，对 3DGS 场景进行了约 23 s 的初步运行性能测试。去除启动阶段异常采样后，运行帧率主要分布在 52.32～63.55 FPS，平均约 56.8 FPS，对应平均帧时间约 17.6 ms。

实验对象为四类常见家庭物品：Bottle（圆柱形水瓶）、Tomato（不规则形状番茄）、Cup（圆台体水杯）和 Apple（不规则球形苹果），如图 8 所示。Bottle 和 Tomato 各自测试 80 条数据，Cup 和 Apple 各自测试 50 条数据。

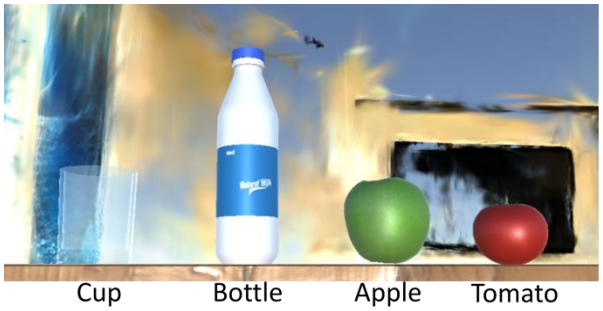

图 8 四种待抓取物体
Fig. 8 Four types of objects to be grasped

每轮开始前随机采样重置物体位置，物体随机化的具体参数如下：抓取区域为 70 cm×15 cm 的桌面矩形区域，物体位置在 XZ 平面内服从均匀分布，最小物体间距约束为 0.3 m（通过碰撞检测保证，最多重试 50 次）。物体起始高度保持预设原始值，起始姿态保持原始旋转，仅随机化 XZ 位置。随机种子采用 Unity 默认设置，未固定种子，因此每次实验产生不同的物体布局序列。机器人站立位置由物体位置通过仿射映射计算得出，保持物体在桌面上的相对拓扑关系；每次实验中物体布局不同，机器人站位也随之变化，但站位的随机性完全来源于物体布局，不引入额外随机因素。任务执行结果统计如表 1 所示。

表 1 任务执行结果统计
Table 1 Task execution results

指标	Bottle	Tomato	Cup	Apple
总次数	80	80	50	50
完全成功	77	65	28	25
抓取失败	0	1	6	18
运输掉落	1	11	16	6
机器人摔倒	2	3	0	1

3.2 典型家庭任务实验：语音控制抓取水瓶水果

每次任务的标准流程为：机器人通过语音指令从当前站立点出发，经 NavMesh 路径规划与 PID 路径控制导航至桌边抓取点，通过三阶段闭环伺服完成物体抓取，携带物体导航至用户投递点，释放物体，如图 9 所示。

动态行人条件下的任务过程验证：为考察系统在行人走动干扰下的任务连续性，在 Unity 交互层中加入沿预设路径移动的虚拟行人。实验过程中，当行人进入机器人导航区域后，机器人能够触发后退让行，并在后方空间不足时停止等待；行人通过后，机器人重新规划路径并恢复原任务，最终完成水瓶抓取与递送。实验表明，系统能够在动态行人条件下完成障碍检测、让行和任务恢复完整流程。

图 9 任务过程截图
Fig. 9 Task Screenshots

实验记录到三类失败模式：抓取失败（三阶段伺服超过最大重试次数仍未满足抓取判据）、运输掉落（导航至投递点途中物体 Y 坐标降至 0.1m 以下）、机器人摔倒（身体倾角超过 30°）。

3.3 实验结果分析

成功率分析：抓取成功率方面，Bottle 为 100%，Tomato 为 98.8%，Cup 为 88%，Apple 为 64%。递送成功率方面，Bottle 为 96.2%，Tomato 为 82.3%，Cup 为 63.6%，Apple 为 78.1%。递送成功率定义为成功抓取后完整完成运输与投递的比率，其差异直接反映了运输阶段的可靠性。为了量化上述差异的统计可靠性，表 2 给出了成功率及 95%置信区间（Wilson 分数区间），表 3 给出了差异显著性检验结果。图 10 以柱状图形式对比了 Bottle、Tomato、Cup 和 Apple 在抓取和递送两个维度上的成功率差异。四类物体的尺寸与夹爪最大开合尺寸（官方参数最大尺寸为 9cm）对比如表 4 所示。

在α=0.01 水平上，Bottle 的递送成功率显著优于 Tomato、Cup 和 Apple。Tomato 与 Cup 的差异

在α=0.05 水平显著，但在α=0.01 水平不显著。Tomato 与 Apple、Cup 与 Apple 的差异均不显著。效应量分析表明，Bottle 与 Cup 的差异最大（Cohen's h=0.905，大效应），其他对比多为小到中等效应。

表 2　成功率及 95%置信区间
Table 2　Success rates with 95% confidence intervals

指标	Bottle	Tomato	Cup	Apple
抓取成功率	100%	98.8%	88%	64%
95% CI	[95.4%，100%]	[93.3%，99.8%]	[76.2%，94.4%]	[50.1%，75.9%]
递送成功率	96.2%	82.3%	63.6%	78.1%
95% CI	[89.5%，98.7%]	[72.4%，89.1%]	[48.9%，76.2%]	[61.2%，89.0%]

表 3　四类物体递送成功率差异检验
Table 3　Significance test of delivery success rate differences among four objects

检验项目	Bottle vs Tomato	Bottle vs Cup	Bottle vs Apple	Tomato vs Cup	Tomato vs Apple	Cup vs Apple
比例差	14.0%	32.6%	18.1%	18.6%	4.2%	-14.5%
z-statistic	2.851	4.824	3.039	2.308	0.506	-1.357
p-value	0.0044	<0.001	0.0024	0.0210	0.6128	0.1746
Cohen's h	0.479	0.905	0.584	0.426	0.104	-0.321

表 4　四类物体尺寸与夹爪比例
Table 4　Object dimensions and gripper opening ratios for four object categories

物体	Bottle	Tomato	Cup	Apple
直径（cm）	6.99	7.80	8.03	8.77
高度（cm）	24.11	8.26	9.62	8.60
占夹爪比例（%）	77.7	86.7	89.2	97.4
形状特征	规则圆柱	不规则球体	圆台	不规则球体

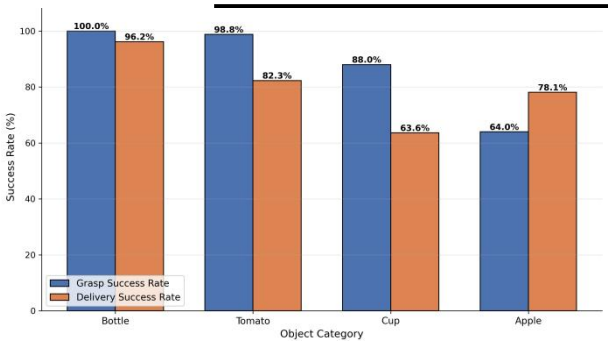

图 10　不同物体抓取与递送成功率
Fig. 10　Grasp and delivery success rates by object category

失败模式分析：图 11 以环形图展示了四类物体各自的结果分类分布。主要失败模式包括抓取失败、运输掉落和机器人摔倒。抓取失败共出现 25 次，其中 Apple 为 18 次、Cup 为 6 次、Tomato 为 1 次、Bottle 为 0 次。结合表 4 可以看出，Apple 和 Cup 的直径分别占夹爪最大开合尺寸的 97.4%和 89.2%，其可用抓取空间相对较小；Bottle 的直径占比最低，为 77.7%，在本实验中未出现抓取失败。上述结果表明，物体尺寸与夹爪开合范围可能影响抓取容差。

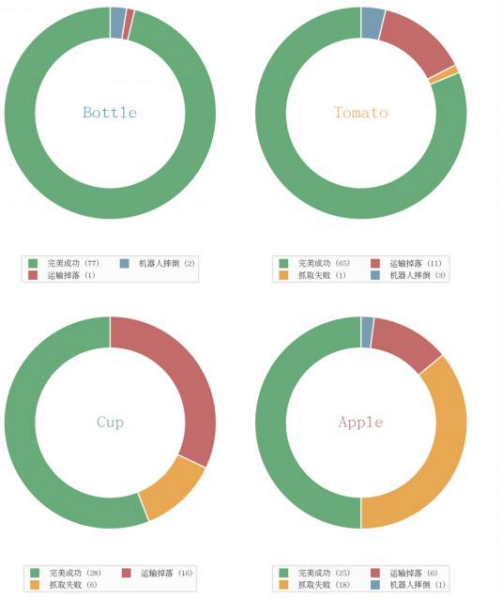

图 11 实验结果分类分布

Fig. 11 Classification distribution of experimental results

运输掉落共出现 34 次，其中 Cup 为 16 次、Tomato 为 11 次、Apple 为 6 次、Bottle 为 1 次。结合物体尺寸和几何外形推测，不同物体的夹爪接触形式和抓取容差可能存在差异，并可能影响运输阶段的稳定性。但本文未直接记录夹持力、摩擦系数、接触状态和行走加速度，因此上述分析仅用于解释实验现象，不能作为动力学因果证明。

机器人摔倒共出现 6 次，其中 Bottle 为 2 次、Tomato 为 3 次、Apple 为 1 次。根据实验记录，摔倒均发生在导航转向或障碍物接近阶段。

耗时分析：由于四类物体的耗时分布基本一致，图 12 以堆叠柱状图形式仅展示了 Bottle 和 Tomato 两类成功任务各阶段的平均耗时分解。Bottle 成功任务平均总耗时约 47.4 s，Tomato 约 49.8 s，两者总体接近。四阶段耗时占比呈现一致的模式，导航去程与导航回程为任务总耗时的主要组成部分。导航耗时占比高的原因在于 G1 人形机器人受限于双足步态稳定性，采用较低的安全行走速度，单个导航行程通常需 20s 左右。

表 5 给出了成功任务各阶段耗时的完整描述性统计，失败任务由于抓取失败、运输掉落或机器人摔倒等不同阶段提前终止，后续阶段耗时不存在，若直接纳入阶段平均耗时会人为缩短总耗时并造成统计偏差。因此，失败任务不纳入表 5 的阶段耗时统计，而仅用表 1 和图 11 中的失败模式分类分析。对于失败任务，系统记录其失败类型和失败发生阶段，用于评估系统可靠性与主要失效原因。

表 5 成功任务各阶段耗时统计

Table 5 Phase-level time statistics of successful tasks

任务	阶段	均值±*SD*(s)	中位数(s)	Q1(s)	Q3(s)	范围(s)
Bottle (n=77)	导航去	21.3±2.4	21.4	20.5	22.9	[11.2, 26.6]
	抓取	3.8±0.2	3.8	3.6	4.0	[3.4, 4.3]
	导航回	18.9±1.3	19.3	18.7	19.7	[15.5, 20.6]
	放下	3.5±0.0	3.5	3.5	3.5	[3.5, 3.5]
	总耗时	47.4±2.5	47.6	46.1	49.2	[37.9, 54.0]
Tomato (n=65)	导航去	22.0±1.4	22.5	21.2	23.0	[18.7, 24.3]
	抓取	4.4±0.2	4.3	4.2	4.4	[4.1, 5.4]
	导航回	20.0±2.7	18.6	17.9	21.6	[17.2, 28.7]
	放下	3.5±0.0	3.5	3.5	3.5	[3.5, 3.5]
	总耗时	49.8±2.8	48.8	48.2	51.5	[45.5, 59.1]

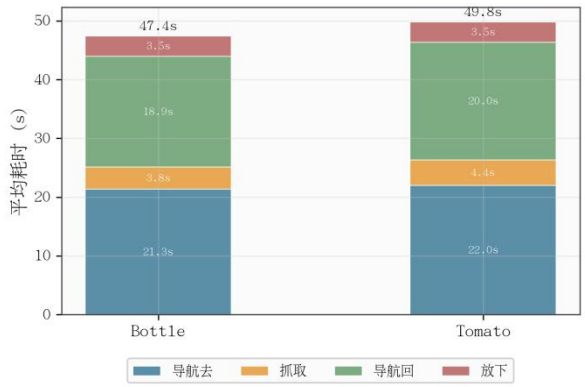

图 12 成功任务各阶段耗时分解

Fig. 12 Phase-level time breakdown of successful tasks

耗时稳定性分析：图 13 同样以散点图形式仅展示了 Bottle 和 Tomato 全部 160 次实验的总耗时分布。成功任务（实心标记）的总耗时主要集中在 40~60 s 区间，两类物体分布范围相近。失败任务（空心标记）的耗时因失败阶段不同而呈现较大离散度：抓取失败耗时最短（任务提前终止），运输掉落居中。图中虚线标注了各类物体成功任务的平均耗时，Bottle 与 Tomato 的平均耗时未呈现显著差异。

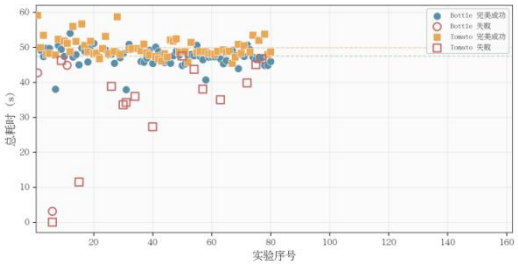

图 13 各次实验任务耗时分布
Fig. 13 Per-trial task completion time distribution

典型任务剖面：图 14 选取了 Bottle 一次中位数耗时的成功任务，以水平甘特图的形式直观展示了从任务启动到完成的完整时间剖面。各阶段以不同颜色区分，时长标注于色块中央。该剖面清晰反映了系统的时序逻辑：导航至抓取点（约 21 s）→伺服抓取（约 4 s）→携带物体导航至投递点（约 20 s）→ 释放物体（约 3.5s），整个链路无等待间隙，状态转换紧凑。

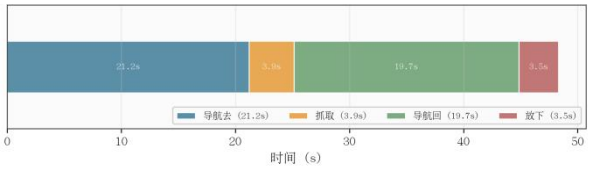

图 14 典型成功任务执行时间线
Fig. 14 Execution timeline of a typical successful task

3.4 与其他仿真平台的对比分析

为直观比较传统手工建模贴图场景与 3DGS 场景的视觉表现，本文选取 Unity 手工建模贴图家庭场景和 Postshot 重建的 3DGS 家庭场景进行并列展示，如图 15 所示。手工场景由规则几何模型、人工材质和常规贴图构建，3DGS 场景则由真实家庭视频重建获得。由于两类图片并非来自同一场景的严格同位姿渲染，图 15 用于代表性视觉外观比较，不作为像素级定量基线。

为进一步说明本文仿真平台与其他机器人仿真平台的差异，本文将所构建的 3DGS+Unity 平台与 Isaac Sim、Gazebo 进行对比，如表 6 所示。需要说明的是，本文平台并非用于替代 Isaac Sim 或 Gazebo 等高精度物理仿真器，而是面向真实家庭场景外观复现与具身智能任务流程验证。Isaac Sim 和 Gazebo 更适合机器人动力学、传感器建模、控制算法和 ROS 生态验证；而本文平台的优势在于可由真实家庭图像序列重建具有真实纹理、光照和空间细节的仿真场景，从而支持导航、抓取、递送和语音交互等家庭服务任务的端到端测试。

此外，3DGS 表征主要用于真实场景视觉重建，并非 Isaac Sim 或 Gazebo 的原生物理场景格式。若将 3DGS 场景直接迁移到上述平台，需要额外完成网格化、碰撞体重建、材质转换和物理属性设定，这一过程可能引入新的误差。因此，本文不将 Isaac Sim 或 Gazebo 作为同一 3DGS 场景下的直接实验基线，而是从平台功能、场景构建方式、视觉真实感、物理仿真能力和任务验证适用性等方面进行结构化比较。

实验结果表明，在添加碰撞网格、导航网格和可交互物体模型后，本文平台能够支持机器人完成“导航-抓取-运输-投递”的完整任务链路。260 次实验中的任务成功率、耗时统计、失败类型分析和语音交互鲁棒性结果说明，3DGS+Unity 平台可支撑本文家庭服务任务的端到端仿真验证，并在真实场景视觉复现方面形成对传统手工建模仿真场景的补充。

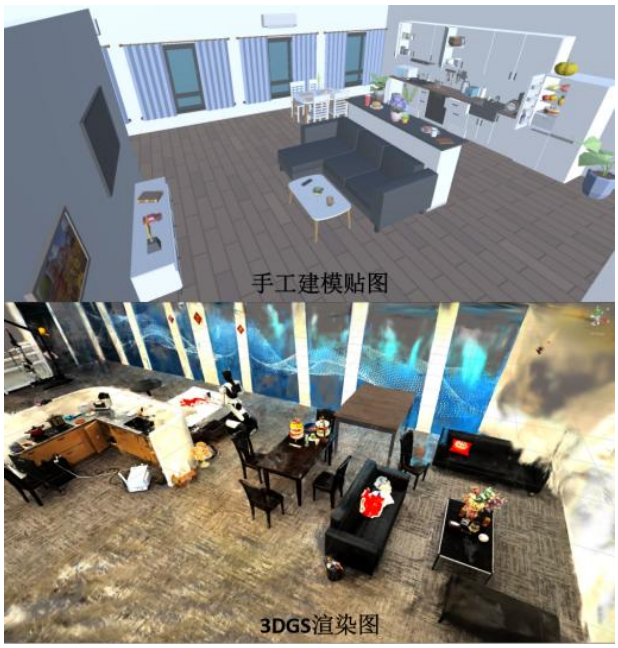

图 15 手工建模贴图家庭场景与 3DGS 家庭场景的视觉对比
Fig. 15 Visual comparison between the manually modeled and

textured home scene and the 3DGS-reconstructed home scene

从图 15 可以看出，手工建模贴图场景具有清晰且规则的几何边界，但家具材质、表面纹理和光照分布相对理想化；3DGS 场景直接来源于真实家庭图像，能够保留更加复杂的生活化物体分布、非规则纹理、局部遮挡和自然光照过渡。因此，该对比能够说明 3DGS 在特定家庭视觉外观复现方面的优势，但不能单独证明其提高了导航、抓取或递送成功率。

表 6　3DGS+Unity 平台与 Isaac Sim、Gazebo 的对比分析
Table 6　Comparison of the 3DGS+Unity platform with Isaac Sim and Gazebo

对比维度	本文 3DGS+Unity 平台	Isaac Sim	Gazebo
场景来源	真实家庭场景图像序列重建，保留真实纹理和光照	USD、Mesh 和材质建模，依赖人工或资产库构建	SDF/URDF、Mesh 和常规材质建模
视觉真实感	较高，接近真实采集场景外观	较高，但取决于模型、材质和渲染设置	中等，偏工程仿真和控制验证
物理仿真能力	依赖 Unity 碰撞体、Navmesh 和刚体动力学，适合任务流程验证	较强，适合机器人动力学、接触和传感器仿真	较强，适合机器人控制、导航和 ROS 生态验证
家庭场景复现成本	较低，可由实拍图像快速重建真实外观	中等到较高，需要模型、材质和 USD 场景搭建	中等到较高，需要 SDF/URDF、网格和材质配置
本文任务适用性	适合验证真实外观场景下的导航、抓取、递送和语音交互流程	适合高精度机器人仿真和传感器闭环验证	适合传统机器人控制、导航和 ROS 集成验证
平台定位	真实外观场景下的补充性任务仿真测试平台	高保真机器人仿真平台	传统机器人仿真平台

3.5 语音指令解析可靠性测试

为评估语音指令解析模块在实际使用中的可靠性，由 2 名说话人在安静和背景噪声两种环境下完成 120 次语音输入，其中每种环境 60 次；每种环境包含 36 条有效指令（含标准句式和口语变体）和 24 条干扰指令（含词典外物体、无关指令）。其中有效指令用于测试意图识别准确率（即系统能否从语音指令中正确解析出〈物体, 拾取点, 投递点〉三元组），干扰指令用于测试词典外拒绝率（OOD rejection rate）和误触发率。分别在安静室内环境和有背景噪声两种条件下朗读全部指令，测试结果如表 7 所示。

表 7　语音指令解析测试结果
Table 7　Voice command parsing test results

指标	环境	有效指令	干扰指令
意图识别准确率	安静	91.7%	—
	噪声	88.9%	—
OOD 拒绝率	安静	—	100%
	噪声	—	100%
误触发率	安静	—	0%
	噪声	—	0%

结果表明：有效指令的意图识别平均准确率为 90.3%，主要错误来源于云端 ASR 的完全识别失败（2 次）和词误识别（如“番茄”识别为“分针”、“瓶子”识别为“被子”、“瓶水”识别为“勺子”）；干扰指令的 OOD 拒绝率为 100%，误触发率为 0%，表明系统对词典外指令具有较强的拒绝能力，不会因无关语音导致误操作。当前测试仅覆盖近场普通话场景，噪声条件为短视频背景音，更复杂的噪声环境（多人交谈、电视播报等）下的表现以及更大规模的用户研究将在未来工作中进一步评估。

4　结论与展望

本文面向家庭人形机器人低成本、可复现的任务验证需求，在具有真实家庭外观特征的 3DGS 仿真场景中部署宇树 G1 人形机器人模型并完成感知、导航与抓取的全流程仿真适配，搭建了“云端语音识别+本地关键词解析”的语音指令解析模块和“状态机调度+导航控制+伺服抓取+IK 规划”的分层作业控制系统。实验以 Bottle、Tomato、Cup 和 Apple 四类物体为测试对象，共完成 260 次有效

实验，其中，Bottle 抓取成功率为 100%，递送成功率为 96.2%；Tomato 抓取成功率为 98.8%，递送成功率为 82.3%；Cup 抓取成功率为 88%，递送成功率为 63.6%；Apple 抓取成功率为 64%，递送成功率为 78.1%。不同物体之间的抓取成功率和递送成功率存在差异，结合实验现象推测，该差异可能与物体尺寸、可抓取姿态、夹爪接触形式及运输过程中的运动扰动有关。

实验结果表明，所构建的 3DGS 场景能够保持真实家庭环境的主要空间布局、纹理和光照趋势，可支撑完整的机器人自主作业链路验证，为家庭人形机器人算法研发提供了可重复运行的仿真测试平台。

本文工作存在以下局限性，需在后续研究中加以改进。

（1）感知闭环的局限：当前抓取伺服回路的反馈信号直接来源于仿真引擎的世界坐标，未经过相机成像、目标检测或位姿估计等视觉感知环节，本质上是基于位置的闭环抓取伺服。在未来 Sim2Real 迁移中，需将 RGB-D 相机的位姿估计结果（及其不确定性）接入控制回路，并在仿真中注入感知噪声以验证闭环鲁棒性。

（2）语音模块的局限：当前语音指令解析模块基于关键词匹配实现，不支持连续对话、指代消解和多意图理解，关键词词典为硬编码，对词典外词汇直接返回失败。云端 ASR 的识别准确率受环境噪声和口语变体影响，尚未覆盖远场、多说话人、强噪声和方言等复杂条件。未来将引入大语言模型增强的自然语言理解模块，并在真实部署条件下评估端到端可靠性。

（3）物体泛化验证仍需扩展：本文在 Bottle、Tomato、Apple 和 Cup 的基础上进行了抓取-递送测试，初步验证了系统对不同几何特性物体的适应能力。然而，当前验证的物体类别仍然有限，对于更大尺寸、更重、表面更光滑或形状更不规则的物体，系统的泛化能力仍需在更大规模的基准测试中进一步评估。

（4）动态场景验证的局限：本文在静态 3DGS 背景的 Unity 交互层中补充了单个虚拟行人的定性避让验证，但尚未覆盖多人活动、复杂运动轨迹、动态光照变化及长时间统计测试。3DGS 视觉表征本身仍为静态表示，后续可进一步引入 4D Gaussian Splatting 或可变形 3DGS，以增强动态场景表达能力。

（5）本文物理交互层采用轻量化碰撞建模方式，静态场景碰撞体并不追求与 3DGS 视觉外观逐点一致，尚未对碰撞网格误差进行独立几何量化评估。本文主要通过路径连通性检查和任务级实验验证其在家庭服务具身智能任务中的可用性。

（6）实验可复现性与动力学记录的局限：本文采用 Unity 默认随机状态，未保存固定随机种子及任务布局序列，严格复现相同实验序列仍存在限制。同时，实验未直接记录夹持力、摩擦系数、夹爪—物体接触状态和机器人行走加速度，因此关于运输掉落原因的分析仅为基于几何特征和失败现象的合理推测。

未来工作将围绕以下方向展开：

（1）引入自适应指尖或柔性夹持机构提升对多样化物体的运输可靠性；

（2）在现有单动态行人验证基础上，进一步扩展多人活动、复杂运动轨迹和动态光照条件下的系统测试；

（3）探索仿真环境中的策略训练向真实机器人的虚实迁移；

（4）为进一步提高后续迁移可行性，后续实验将考虑引入域随机化策略：在视觉域随机调整光照强度、色温和相机曝光参数；在物理域随机化物体质量、表面摩擦系数和夹爪闭合力；在感知域为目标位置注入高斯噪声并引入随机感知延迟，以模拟真实视觉位姿估计的不确定性。上述策略将用于评估导航、抓取和递送流程对视觉误差、物理参数误差和执行延迟的鲁棒性。

参考文献：

[1] LI H Z, DING X L. Adaptive and intelligent robot task planning for home service: A review[J]. Engineering Applications of Artificial Intelligence, 2023, 117: 105618.

[2] Bono G, Poirier H, Antsfeld L, et al. Learning to navigate efficiently and precisely in real environments[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 17837-17846.

[3] Kerbl B, Kopanas G, Leimkühler T, et al. 3d gaussian splatting for real-time radiance field rendering[J]. ACM Trans. Graph., 2023, 42(4): 139:1-139:14.

[4] Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.

[5] Zhu S, Wang G, Kong X, et al. 3d gaussian splatting in robotics: A survey[J]. arXiv preprint arXiv:2410.12262, 2024.

[6] Xie T, Zong Z, Qiu Y, et al. Physgaussian: Physics-integrated 3d gaussians for generative dynamics[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 4389-4398.

[7] Zheng Y, Chen X, Zheng Y, et al. Gaussiangrasper: 3d language gaussian splatting for open-vocabulary robotic grasping[J]. IEEE Robotics and Automation Letters, 2024, 9(9): 7827-7834.

[8] Lu G, Zhang S, Wang Z, et al. Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 349-366.

[9] Li X , Li J , Zhang Z ,et al.RoboGSim: A Real2Sim2Real Robotic Gaussian Splatting Simulator[J]. 2024.

[10] Tang Y, Qi S, Zhu L, et al. Obstacle avoidance motion in mobile robotics[J]. Journal of System Simulation, 2024, 36(1): 1-26.

[11] 王红军，林俊强，邹湘军，张坡，周铭轩，邹伟锐，唐昀超，罗陆锋。基于数字孪生的果园虚拟交互系统构建 [J]. 系统仿真学报，2024,36 (6):1493-1508. Wang H J, Lin J Q, Zou X J, et al. Construction of a Virtual Interactive System for Orchards Based on Digital Twin[J]. Journal of System Simulation, 2024, 36(6): 1493-1508.

[12] Jeong H, Lee H, Kim C, et al. A survey of robot intelligence with large language models[J]. Applied sciences, 2024, 14(19): 8868.

[13] Ahn M, Brohan A, Brown N, et al. Do as i can, not as i say: Grounding language in robotic affordances[J]. arXiv preprint arXiv:2204.01691, 2022.

[14] Xiao A, Janaka N, Hu T, et al. Robi Butler: Multimodal Remote Interaction with a Household Robot Assistant[J]. arXiv preprint arXiv:2409.20548, 2024.