arXiv:2603.02139v1 [cs.RO] 2 Mar 2026

重新思考相机选择：鱼眼相机特性在机器人操作中的实证研究

Han Xue[1*] NanMin[2*] Xiaotong Liu[3,4*] Wendi Chen[1,4] Yuan Fang[1,4] JunLv[5] Cewu Lu[1,4,5,‡] ChuanWen[1‡]上海交通大学 [2]东南大学 [3]中国科学技术大学[4]上海创新研究院 [5]NoematrixLtd.

[*] 同等贡献。‡通讯作者。

摘要

鱼眼相机因其极宽的视场角（*FoV*）而在机器人操作中被迅速采用，其速度已远超人们对它们在策略学习中下游影响的系统性认识。本文提出了首个全面的实证研究以弥合这一差距，严格分析腕部安装的鱼眼相机在模仿学习中的特性。通过在仿真与真实世界中的大量实验，我们探究了三个关键研究问题：空间定位、场景泛化与硬件泛化。研究结果表明：（*1*）宽视场显著增强空间定位，但这一收益在很大程度上取决于环境的视觉复杂度。（*2*）鱼眼训练策略虽易在简单场景中过拟合，但在以足够的环境多样性进行训练时，可实现更优的场景泛化。（*3*）尽管朴素的跨相机迁移会导致失败，我们将根本原因确定为尺度过拟合，并证明借助简单的随机尺度增强（*RSA*）策略可以提升硬件泛化性能。总体而言，我们的发现为机器人学习中鱼眼数据集的大规模采集与有效使用提供了具体、可操作的指导。更多结果与视频请见 https://robo-fisheye.github.io/。

1. 引言

有效的视觉感知是实现稳健机器人操作的基石。尽管标准针孔相机一直是默认选择，但机器人学领域中越来越多的工作（例如，UMI [5]，RDT2 [47]，GEN-0 [46]，$\pi_{0.5}$ [59]）开始采用鱼眼相机，这类相机通常安装在机器人的腕部。这类相机的视场极大，通常超过 180°，能够捕捉到比传统相机宽广得多的视野。这种广角能力已在自动驾驶 [9, 34] 和同步定位与建图（SLAM） [20,54],等领域证明了其巨大价值，在这些领域中，全面感知场景至关重要。

这一新兴趋势表明，未来将依赖大规模的鱼眼数据集，用于训练模仿学习策略和视觉-语言-动作（VLA）模型 [2, 13,46, 47, 59]。然而，这种应用的发展速度正迅速超越我们对其系统性的理解。具体而言，使用腕部安装的鱼眼相机进行机器人模仿学习的实际优势和潜在挑战，在很大程度上仍未得到探索。本文通过对鱼眼相机特性对策略性能的影响开展全面系统的分析，弥合这一关键空白。

为提供可落地的洞见，我们围绕鱼眼镜头的两个决定性特征及其对策略学习的直接影响来组织分析：

- **(1) 宽视场的优势：** 其主要优势在于巨大的视场。这自然促使我们研究其对策略能力的下游影响——具体而言，这种更加丰富的上下文如何提升策略对世界的理解。
- **(2) 畸变带来的挑战：** 这种宽视场是通过严重的径向畸变实现的，而针孔模型中并不存在这种特性。这带来了独特的挑战，尤其是在具有不同内参的不同硬件之间实现模型泛化。

基于这一框架，我们提出了三个关键研究问题。前两个旨在探究宽视场优势的影响，第三个则直面硬件特有畸变带来的挑战：

1. **空间定位：** 广角视场在多大程度上增强策略的空间推理与定位能力？ 2. **场景泛化：** 鱼眼相机是否能提升策略面对新颖或干扰性背景时的鲁棒性与泛化能力？ 3.**硬件泛化：** 策略能多好地

在一台鱼眼相机上训练后，迁移到具有不同内参的新的、未见过的鱼眼镜头?

为严谨地回答这些问题，我们在仿真和真实世界中开展了大量实验。在仿真实验中，我们在 MuJoCo [48]物理引擎中实现了逼真的鱼眼相机模型。这使我们能够针对两个广受认可的模仿学习基准测试 Robomimic [28] 和 MimicGen [29], 中的多项任务，采用相机和背景设置的多种组合开展大规模实验。在真实世界验证中，我们在三项操作任务中使用了多种不同的实体摄像头和背景，并设计了严格的评估协议，以验证仿真实验中的发现。

我们的全面研究揭示了若干重要发现，并为如何在真实世界中利用鱼眼相机扩大机器人数据采集规模提供了可行指导：

1. **鱼眼相机能够提升空间定位能力，但这一效果取决于场景内部的复杂度**。我们发现，更广的视场能够捕捉更多环境特征点，并提升策略的空间定位能力。然而，在缺少鲜明视觉特征的环境中（例如，纯色背景），这一优势会减弱。**指导建议**：优先在视觉复杂且特征丰富的环境中采集数据。

2. **鱼眼相机需要场景间多样性以防止过拟合**。虽然鱼眼训练策略更容易在简单场景中过拟合，但如果在背景（例如纹理、光照和干扰物）方面以足够的多样性进行训练，其场景泛化能力会超过标准相机。

   指导：在数据采集过程中最大化环境多样性，以释放泛化潜力。

3. **鱼眼镜头的跨相机泛化是一项显著挑战，但该挑战可部分应对**。我们观察到，将策略简单迁移到新的鱼眼镜头会导致性能急剧下降。然而，我们证明所提出的随机尺度增强（RSA）能够有效缓解这种性能下降。这表明该挑战并非不可克服，以数据为中心的方法是有前景的方向。

   指导：在训练期间使用强尺度导向的数据增强，以提升跨相机迁移能力。

综上所述，我们的发现与指导为机器人学界提供了严谨的基础，使其能够自信且有效地大规模采用鱼眼相机，以训练下一代通用策略。

2. 相关工作

面向机器人操作的模仿学习。 模仿学习是一种强大的范式，用于教会机器人掌握复杂的操作技能，这些技能来自专家演示。其核心在于行为克隆（BC），它将策略学习表述为将观测映射到动作的监督问题 [3, 32, 41]。近年来，该领域经历了向基于扩散的策略的范式转变 [4, 5, 12, 14, 18, 22, 57], ，这类策略擅长对复杂的多模态动作分布进行建模。尽管有这些进展，模仿学习仍然面临诸如分布偏移 [17, 37] 和部分可观测性 [6, 52]等挑战。缓解这些问题的一个关键策略是收集丰富且高质量的演示数据 [2, 18, 46, 59], ，涵盖多样的行为轨迹，以及带有丰富上下文信息和大感受野的传感器观测。在这项工作中，我们表明，鱼眼相机凭借其广角视场角（FoV），与标准针孔相机相比，能大幅增强策略可获得的上下文信息。我们进一步对其影响进行了系统分析，展示了由鱼眼相机独特特性所带来的性能提升，以及更好的场景泛化与硬件泛化。

机器人鱼眼相机。鱼眼相机凭借其宽广的视场，广泛应用于自动驾驶 [9, 25, 34, 53], 同步定位与建图 [20, 51, 54],监控 [24, 55], 和空中机器人 [1, 15]等领域。近年来，鱼眼相机的应用已扩展至机器人操作，以拓宽工作空间观测范围 [5, 18, 46, 50]。其中，UMI [5] 采用鱼眼相机进行便携式数据采集，提供了宝贵的应用层面见解。然而，目前尚缺乏系统性分析来分离特定光学属性对策略学习的定量影响。这一缺口也反映在当前的基准测试领域：现有鱼眼数据集 [42, 43, 56]缺少机器人操作任务，而常用的机器人基准测试 [19, 21, 28, 31] 则未纳入鱼眼图像流。我们的工作旨在填补这一空白，通过模拟和真实世界实验，对鱼眼相机在机器人操作中的应用开展系统性研究，并最终提出一套关于如何有效部署鱼眼相机的实用指南。

鱼眼仿真。 鱼眼相机广泛应用于真实世界的机器人操作，但标准仿真器缺乏鱼眼渲染功能，阻碍了基准测试，也限制了对基于鱼眼相机的策略的研究。因此，为了支持真实环境与仿真环境中的一致实验，开展可靠的鱼眼仿真至关重要。目前有几种方法可以从针孔渲染图合成鱼眼图像：3D高斯泼溅 [23] 能够实现高保真鱼眼图像合成 [27, 35, 44], ，扩散模型也展现出潜力 [10, 49], ，但这两种方法对于交互式仿真而言仍然过于缓慢或需要过多资源。相比之下，经典投影模型 [26] 以及 OmniCV-Lib [39] 等工具包通过两阶段渲染管线提供了稳定高效的替代方案：首先生成全景中间视图，然后将其重新投影到

鱼眼领域。为满足稳定性、准确性和效率方面的要求，我们设计了一种受 OmniCV-Lib 启发的两阶段管线 [39] 用于鱼眼仿真。

3. 研究设计

我们设计了一个涵盖仿真和真实世界的综合实验框架，以研究鱼眼相机特性。本节将详细介绍问题表述（第 3.1 节）、为开展这项研究而开发的核心方法组件（第 3.2 节）、我们的模仿学习框架（第 3.3 节）、仿真（第 3.4 节）和真实世界（第 3.5 节）的实验设置，以及我们的评估协议（第 3.6 节）。

3.1. 问题表述

我们的研究旨在系统地探究鱼眼镜头对模仿学习的影响。我们的分析围绕三个关键问题展开，这些问题源于鱼眼相机的核心特性（例如宽视场和畸变）：

1. **（RQ1）空间定位：** 鱼眼相机的宽视场在多大程度上增强了策略的空间定位能力？ 2. **（研究问题2）场景泛化：** 鱼眼相机能否提升策略应对新颖或干扰性背景的鲁棒性和泛化能力？ 3. **（RQ3）硬件泛化：** 在一台鱼眼相机上训练的策略，迁移到具有不同内在参数的、未见过的鱼眼镜头上时，表现如何?

为严谨地回答这些问题，我们定义了一组用于分析的核心因素，如图1所示。这些因素作为实验中的自变量，用于分离并测量各自的具体影响：

- **相机模型：** 这是我们的主要比较维度。在后续针对不同分析因素开展的所有实验中，我们都将鱼眼相机的性能特征与作为对照组的标准针孔相机进行直接比较。我们具体关注腕部安装相机，这是近期大型野外数据采集项目（例如 RDT2 [47], GEN-0 [46]）采用的一种配置；较短的观测距离会自然地放大两种相机模型之间的差异。

- **场景复杂度：** 为回答 RQ1（空间定位）这一问题，我们针对背景设置建立了两个对照组：特征贫乏（例如，纯色）和特征丰富（例如，纹理丰富）。通过比较这两种极端情况下的性能，我们评估背景纹理可用性对模型空间定位能力的影响。
- **场景多样性：** 为解决研究问题2（场景泛化），我们建立了一套扩展方法，在训练过程中将背景数量从单一背景扩展到 N 不同场景。随后，通过测量策略在完全零样本迁移中的表现，严格检验这种多样性提升的有效性——具体而言，将其部署到未包含在训练分布中的完全未见场景中进行测试。
- **相机参数：** 为应对RQ3（硬件泛化），我们通过评估该策略在以零样本方式部署于未见相机内参（即不同视场和畸变配置）时的性能，来衡量其跨硬件泛化能力。

3.2. 实现鱼眼中心模仿学习

为探究我们的研究问题，尤其是RQ1与RQ3，我们首先必须解决两个重大技术缺口：标准仿真器中缺乏逼真的鱼眼相机支持，以及跨相机泛化方面的挑战。我们引入两个组件来解决这些问题。

MuJoCo中的鱼眼相机仿真。据我们所知，现有的机器人学仿真基准均未为鱼眼相机提供原生支持。因此，我们在MuJoCo [48]物理引擎中实现了鱼眼相机模型，改编自Robomimic [28] 与MimicGen [29]所使用的流程。我们采用两阶段投影流程来仿真鱼眼相机，如图2所示。该过程首先在我们的仿真环境中捕获完整的360度球形视图。为此，我们放置六个虚拟相机，每个相机沿一个基本方向（前、后、左、右、上、下）定向，并将所得的六幅图像组装成单个立方体贴图。在流程的第一阶段，将该立方体贴图的六个面投影并拼接在一起，形成称为等距圆柱投影图像的中间全景表示。这种常见格式有效地将球面视图“展开”到二维平面上。在第二阶段，通过应用特定的投影模型，将该等距圆柱投影图像变换为最终的鱼眼视图，从而能够仿真各种镜头特性。整个过程利用受OmniCV-Lib等库启发的全向摄像头功能实现 [39]。

随机尺度增强（RSA）。为解决 RQ3 并缓解跨相机泛化中观察到的性能骤降，我们提出了一种简单却非常有效的数据增强技术，称为随机尺度增强（RSA），如图所示。3。我们假设，将策略迁移到未见过的鱼眼镜头时，面临的主要挑战是不同镜头内参导致的物体尺度显著变化。我们的 RSA 方法通过迫使策略适应尺度变化，直接应对这一问题。我们不再使用固定尺度裁剪（例如，0.95），而是为每张训练图像从一个较宽的均匀分布中采样一个随机尺度因子 s（例如，$U(0.7, 1.3)$）。随后，将图像按该尺度进行中心裁剪，并调整至标准网络输入尺寸。关键在于，如果 $s > 1.0$，这一操作便会实现“缩小”效果：将源图像缩小，并用黑色填充周围画布。

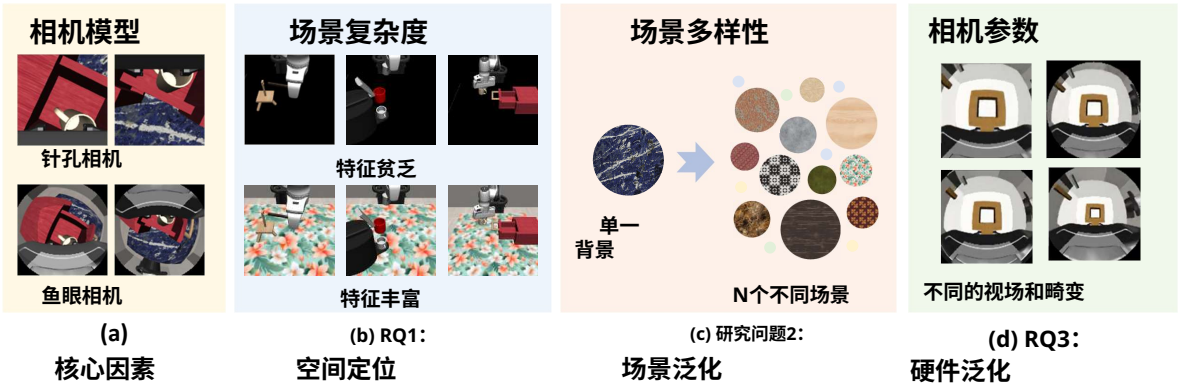


图1. 为解答我们的研究问题（RQs）所分析的四个因素概览。我们研究：**（a）相机模型**（鱼眼与针孔）作为主要比较；**（b）场景复杂度**（贫乏与丰富），用于空间定位（RQ1）；**（c）场景多样性**（1个与N个场景），用于场景泛化（研究问题2）；以及**（d）相机参数**（变化的内参），用于硬件泛化（RQ3）。

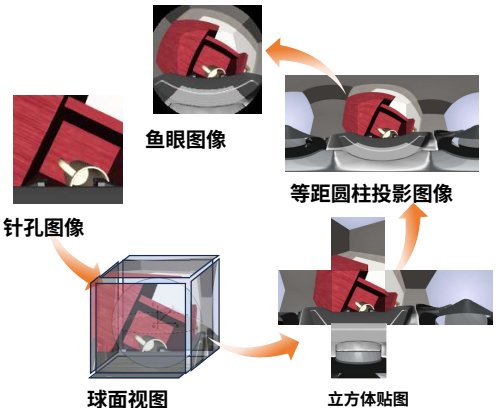


图2 MuJoCo中鱼眼相机仿真的实现流程 [48]。

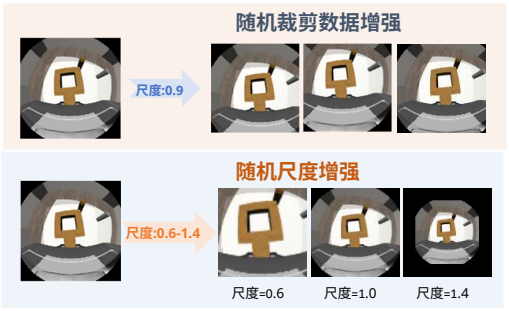


图3. 随机裁剪增强（固定尺度）与随机尺度增强（RSA）在跨相机泛化中的对比。

我们认为，这种简单而有效的增强策略可以防止网络对物体的绝对像素尺度过拟合。相反，它促使策略学习相对空间关系，例如目标物体相对于机器人末端执行器的尺度，这是一种在不同相机系统之间更具泛化性的线索。

3.3. 模仿学习框架

我们基于标准的最先进视觉模仿学习框架构建系统。以下将按照从核心算法到输入和输出的逻辑顺序，详细介绍我们的设计选择。

核心算法。 我们采用扩散策略 [4]框架对收集的大量视觉数据进行建模。之所以选择该框架，是因为它在真实世界操作任务中表现出色，并已成为基于视觉的机器人模仿学习中强大且广受认可的基线。按照标准做法，我们采用 U-Net [36]架构作为噪声预测网络，并使用 DDIM [45] 调度器进行高效推理。

策略输入。 我们将策略设计为无状态，完全依赖视觉数据。这是直接检验我们的研究问题（RQs）的关键决策。

- **视觉编码器：** 我们针对各领域的需求指定了不同的编码器：– **仿真：** 对于仿真中的受控大规模实验，我们采用未经预训练的标准 ResNet-18 [16] 。它是广泛采用、鲁棒且计算高效的基线，确保我们的结果能够与现有仿真基准测试进行比较。– **真实世界：** 在真实世界验证中，域偏移和视觉保真度是更重要的考量，因此我们利用 CLIP [33] 视觉Transformer（ViT [8]）提取的特征。我们采用其强大的预训练视觉特征，以增强模型应对物理环境中固有的新纹理和光照条件的鲁棒性。

- **不使用本体感知输入：** 这项分析中的一个核心设计决策是，在策略训练期间省略本体感知状态（例如末端执行器位姿和关节状态）。

这会形成一种无状态策略，迫使模型完全依赖视觉输入进行空间定位。正如先前关于空间泛化的研究所示 [58], , 移除本体感知可防止策略对简单的状态向量过拟合，并直接分离和检验视觉编码器的能力，尤其是鱼眼相机的视场优势——这对于需要依靠强视觉线索进行定位的任务尤为重要。

策略输出（动作空间）。在仿真实验中，我们遵循 Robomimic [28], 的默认设置，采用增量动作（连续帧之间的相对变换）作为动作空间。在真实机器人实验中，我们采用 UMI [5], 的默认设置，使用相对动作（相对于动作块第一帧的变换）。先前研究 [58] 表明，与绝对动作相比，相对动作空间能提供更好的空间泛化能力，尤其是在缺少本体感觉输入的情况下。

3.4. 仿真实验设置

•基准测试： 两个基于 MuJoCo 的 [48] 基准测试——Robomimic [28] 和 MimicGen [29], ——均经过改造，以支持鱼眼相机渲染。

• 摄像头配置： 我们为实验定义了两种摄像头配置，如图2（b）所示：**– 针孔摄像机：** 一台或两台腕部安装的针孔相机（90° 视场），不设置第三视角相机。**– 鱼眼相机：** 一台或两台腕部安装的鱼眼相机（235° 视场），不设置第三视角相机。与先前类似 UMI 的研究一致，我们的设置特意排除了第三视角相机（例如RDT2 [47], GEN-0 [46]）。这样做是为了避免引入混杂变量，以免干扰对鱼眼相机影响的独立分析。

•任务： 我们从 Robomimc [28] 和 MimicGen [29] 中选取了六项具有挑战性的任务，如图 S1所示。这些任务用于评估不同维度的能力，包括：**高精度操作**(工具悬挂、穿线）、**空间泛化** (正方形、组装），以及**长时程任务执行** (咖啡、马克杯清理）。

3.5. 真实世界实验设置

• 硬件平台： 我们使用一台 Flexiv Rizon 4 [11]机械臂和 DH AG-160-95 夹爪 [7]作为硬件平台，如图 5所示。数据采集通过 Meta Quest 3 [30]进行遥操作。

• 相机配置： 我们使用两种相机配置（图 5），不使用第三视角相机，与仿真设置一致：

– 针孔摄像机： 一台腕部安装的针孔相机（60°视场），不使用第三视角相机。**– 鱼眼相机：** 一台腕部安装的鱼眼相机（180°视场），不使用第三视角相机。

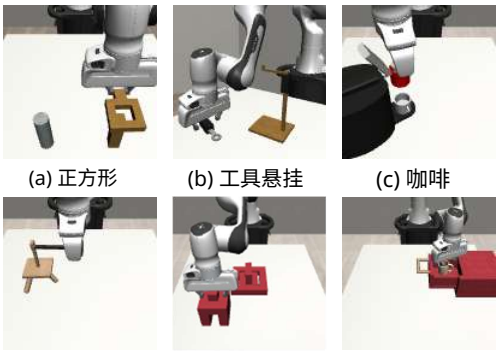


图 4. 仿真实验中的六项任务。

• 任务： 我们设计了三个任务，如图5所示，用于测试各种机器人操作技能，包括**空间泛化**(拿起杯子）、**可变形物体操作**(折叠毛巾）以及**高精度旋转操作**(悬挂中国结）。

3.6. 评估协议

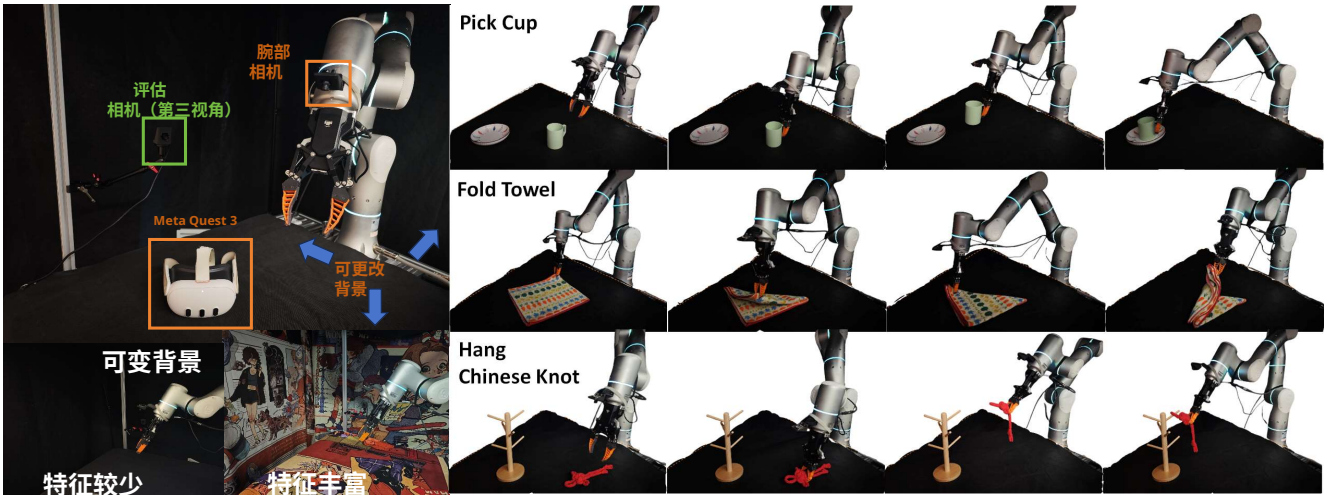
我们在仿真和真实世界中进行了严格评估，以衡量策略性能和泛化能力。我们的协议旨在确保不同相机系统之间的结果可靠且可公平比较。

• 仿真实验在基于仿真的分析中，我们遵循 Robomimic 和 MimicGen 基准测试所确立的标准评估流程。性能采用**成功率 (SR)**进行量化。对于每次实验运行，我们选取训练期间性能最高的策略检查点，然后执行该策略 50 条执行轨迹进行评估。

• 真实世界实验我们的真实世界评估协议旨在提供细致且稳健的性能衡量，同时确保公平性。在复杂的真实世界操作任务中，二元成功率往往过于稀疏，难以捕捉策略行为的细微差异。遵循先前工作[18], , 我们定义了一种归一化多阶段评分指标。每项任务被分解为若干关键阶段（通常为 2–3 个）。策略每成功完成一个阶段即可获得一分，最终报告**归一化得分**:

$$\text{归一化得分} = \frac{\text{获得的总分}}{\text{阶段总数}}$$

对于每种实验设置，该指标均取 $N = 20$ 次试验的平均值，比简单的通过/失败指标更精细地反映策略能力。为确保对所有实验条件进行公平、直接的比较，我们对实验设置进行了细致控制。每次运行前，我们都会将机器人位姿及所有相关物体位姿重置为预先设定的初始状态。



(a). 真实世界硬件设置 **(b). 真实世界操作任务**

图5。(a) 真实世界实验设置，其中包含用于研究场景复杂度（RQ1）和场景泛化（RQ2）的**可变背景**。(b) 真实世界实验中的三项任务：拿起杯子、折叠毛巾和悬挂中国结。

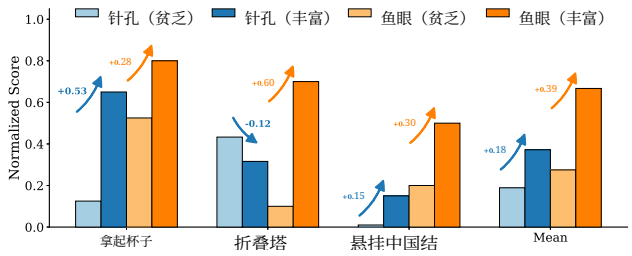


图6. 在三个任务中，不同场景复杂度（特征贫乏 与特征丰富）下鱼眼相机/针孔相机的真实世界性能（RQ1）。

4. 实验分析

4.1. (RQ1) 空间定位

在本节中，我们研究 RQ1：鱼眼相机更广的视场能否帮助策略定位？

由于我们依赖仅腕部视角视觉，策略必须利用背景线索进行定位。我们认为，鱼眼相机的优势源于其更宽的视场能够捕获更多静态背景特征，这些特征可作为稳定的视觉锚点。因此，我们提出以下可检验的假设：

假设：鱼眼相机更宽的视场能够整合更密集的背景特征，从而实现更优的策略定位。因此，策略性能将高度依赖于训练场景的视觉丰富度。

为验证这一点，我们开展了由两部分组成的实验。首先，我们测量了“特征贫乏”（例如，统一色彩）与“特征丰富”（例如，图 5(a) 中复杂背景）背景下的任务性能。图 6和表 1中的结果证实，丰富的背景至关重要。更重要的是，图 6 显示，**在真实世界中，鱼眼相机带来的性能提升更为显著**（平均提升 +0.39，而针孔相机为 +0.18），仿真中的提升则较小。我们认为，这得益于更强大的 CLIP [33]编码器以及真实世界中复杂的纹理，而鱼眼相机的视场能够充分利用这些纹理。

其次，为了证明这种提升源于更优的定位能力，我们探究编码器隐含的空间感知能力。我们在训练好的视觉编码器上微调一个轻量级 MLP 头，使其根据图像预测机器人的本体感知状态（位置/方向）；误差越低，表明空间感知能力越强。表2 中的结果十分明确：使用鱼眼相机训练的编码器始终能得到更低的本体感知误差，而**特征丰富环境中的鱼眼相机**性能大幅领先（例如，平移误差为 1.73 cm），这证实它学到了最准确的空间表征。

综合来看，这些发现证实，鱼眼相机的定位效果取决于环境。这为数据采集提供了重要指导：**为了最大限度地提升策略性能，应在视觉复杂且特征丰富的环境中采集数据，以充分发挥鱼眼相机的空间定位能力**。

4.2.（研究问题2）场景泛化

本节探讨研究问题2：鱼眼相机如何影响对新背景的泛化能力？

在机器人操作中，腕部安装相机的运动会自然引发背景变化，这可以视为一种用于场景泛化的**隐式数据增强** 形式。我们认为，鱼眼相机的宽视场会显著增强这一效果：它能够以更宽的视场捕捉更大范围的背景变化，并引入

表1. 不同场景复杂度下鱼眼/针孔相机的仿真性能（特征贫乏与特征丰富）（RQ1）。特征丰富背景下的性能以括号中的绝对差值表示，该差值是相对于同一相机在特征贫乏背景基线下的结果计算得出。

实验因素		仿真任务成功率						平均值
相机	特征	正方形	工具悬挂	咖啡	螺纹组装	马克杯清洁	–	
针孔（单目）	Poor	0.40	0.52	0.36	0.04	0.14	0.40	0.31
针孔（单目）	Rich	0.48 (+0.08)	0.56 (+0.04)	0.34 (-0.02)	0.18 (+0.14)	0.12 (-0.02)	0.38 (-0.02)	0.34 (+0.03)
鱼眼（单目）	Poor	0.68	0.80	**0.80**	0.30	0.24	0.58	0.57
鱼眼（单目）	Rich	**0.74 (+0.06)**	**0.84 (+0.04)**	0.76 (-0.04)	**0.56 (+0.26)**	**0.48 (+0.24)**	**0.60 (+0.02)**	**0.66 (+0.09)**
双针孔	Poor	0.50	0.44	0.26	0.22	0.44	0.40	0.38
双针孔	Rich	0.70 (+0.20)	0.34 (-0.10)	0.36 (+0.10)	0.38 (+0.16)	0.34 (-0.10)	0.56 (+0.16)	0.45 (+0.07)
鱼眼（双重）	Poor	0.86	0.84	0.74	**0.68**	**0.56**	0.66	0.72
鱼眼（双重）	Rich	**0.88 (+0.02)**	**0.88 (+0.04)**	**0.86 (+0.12)**	0.66（-0.02）	0.44（-0.12）	**0.80 (+0.14)**	**0.75 (+0.03)**

表2。使用本体感觉预测作为代理任务，对视觉编码器的空间感知进行定量探测（RQ1）。我们在预训练视觉编码器上微调一个轻量级MLP头，通过在三个真实世界任务中预测机器人的本体感知状态，评估所学习空间表征的质量。

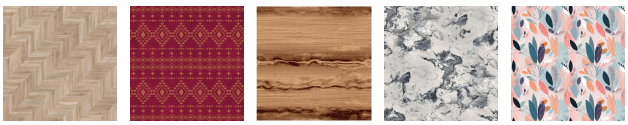
Task	相机特征	平移误差(cm) ↓	旋转误差(°) ↓
拿起杯子	针孔 Poor	12.309	15.345
	针孔 Rich	5.367	7.612
	鱼眼（低）	3.369	3.677
	鱼眼 Rich	**2.362**	**3.394**
折叠毛巾	针孔（低）	4.204	6.829
	针孔（丰富）	5.329	6.434
	鱼眼（低）	3.837	3.398
	鱼眼（丰富）	**2.908**	**2.952**
悬挂中国结	针孔（低）	14.168	12.464
	针孔 Rich	7.683	9.377
	鱼眼（低）	8.804	7.256
	鱼眼（丰富）	**5.143**	**4.887**

在机器人运动过程中，鱼眼畸变会带来更强的数据增强。这引出了我们的核心假设：

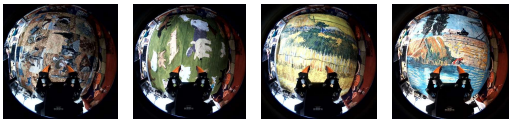
假设：鱼眼训练策略能够更有效地利用场景多样性来提升泛化能力；随着独特训练场景数量的增加，其性能扩展曲线更加陡峭。

为验证这一假设，我们设计了实验，明确测量**场景多样性的缩放效应**。我们改变独特训练场景的数量，N，同时保持数据总量不变，并评估策略在不同未见场景中的零样本性能，这些场景涵盖仿真和真实世界（见图 7）。我们系统地测试 $N = \{1, 8, 16, 32\}$ 在仿真中的表现，以及 $N = \{1, 2, 4, 6, 8\}$ 在真实世界中的表现。

图 8所示结果有力地支持了我们的假设。我们观察到，与传统相机相比，鱼眼相机表现出显著更大的扩展潜力。值得注意的是，在真实世界设置中，鱼眼策略在未见环境中的零样本成功率



(a) 仿真中的 5 个未见测试场景



(b) 真实世界中的 4 个未见测试场景

图 7. 用于评估的未见场景：(a) 仿真；(b) 真实世界实验（研究问题2）。

在仅用八个多样化场景训练后便迅速超过 95%。相比之下，仿真中的扩展曲线较为平缓。我们认为，这种差异主要源于两点：1）视觉编码器不同（仿真中使用未预训练的 ResNet-18 [16]，而真实世界中使用预训练的 CLIP [33]）；2）仿真背景图像的视觉复杂度相对较低（见图 5及补充文件）。

总体而言，这些发现证实，鱼眼相机更宽的视场可作为一种强大的隐式数据增强方式，使策略能够更好地利用场景多样性，实现稳健的跨场景泛化。这为大规模数据收集提供了重要指导：**最大限度地提高场景多样性，对于充分发挥鱼眼相机的泛化能力至关重要**。

4.3. (RQ3) 硬件泛化

本节探讨 RQ3：策略部署到新的鱼眼镜头上时，能否保持性能？

这一问题具有重要的实际意义。首先，随着基于鱼眼相机的大规模数据集（例如用于预训练视觉语言动作模型）日益普及 [18, 46, 47], ，策略能够在定制的机器人系统上进行微调和部署至关重要，而这些系统不可避免地会使用内参不同的镜头。其次，随着硬件

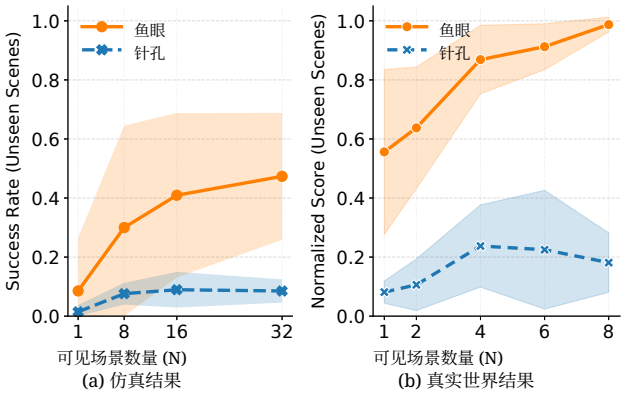


图 8. 策略性能随训练场景数量的增加而提升，分别体现在 (a) 咖啡任务的仿真实验和 (b) 拿起杯子 任务的真实世界实验中（研究问题 2）。

不断升级，策略必须始终向后兼容旧数据。然而，鱼眼镜头显著且多样的畸变特征，使这种跨相机迁移成为一项不容小觑的挑战。

对于腕部安装相机，我们观察到策略严重依赖于空间关系（例如物体距离），这些关系来自图像中物体与夹爪的**绝对尺度**。当引入新镜头时，这些绝对尺度会发生变化，导致策略误判场景（例如将物体感知得比实际更近或更远），并造成灾难性失败（见图 9）。由此引出我们的关键假设：

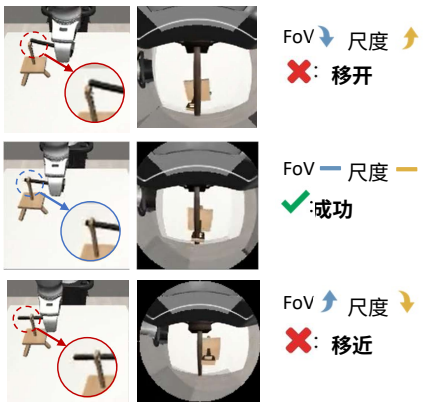


图 9. 跨硬件泛化的失败案例（RQ3）。策略倾向于过拟合绝对尺度。

假设：跨镜头迁移的主要挑战在于策略过拟合于物体绝对尺度。通过数据增强迫使策略转而关注相对尺度关系（例如物体相对于夹爪的尺度），可以缓解这一问题。

如第3.2节所述，我们提出了**随机尺度数据增强（RSA）**，这是应对这一问题的一种简单而有效的策略（见图3）。RSA 迫使网络在多种不同的相对尺度下观察目标物体和夹爪手指，从而打破其对绝对像素大小的依赖。

为了验证这一假设，我们在仿真中进行了大量实验，因为在真实世界中评估众多硬件配置成本高昂（真实世界验证见补充材料）。我们在单一相机配置（“已见参数”）上训练策略，并在五种具有不同畸变和视场参数的未见配置上评估其零样本迁移性能（更多详情见补充材料）。

图 10 中的结果一目了然。采用常规增强训练的基线策略在部署到未见镜头上时，性能大幅下降（例如“参数 3”、“参数 4”）。相比之下，采用我们的 RSA 训练的策略在所有配置下都保持较高的成功率，展现出鲁棒泛化能力。这有力地支持了我们的假设：学习相对尺度是实现跨相机鲁棒性的关键。

这项分析为鱼眼数据提供了我们的最终指导合集：**采用 RSA 等强力的基于尺度的数据增强，确保策略对硬件变化具有鲁棒性，并能够利用来自不同镜头来源的数据。**

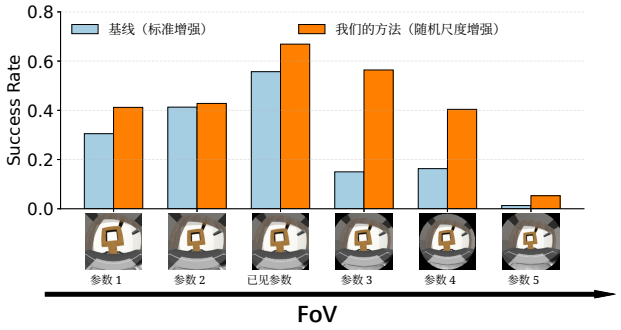


图10. 在仿真中的六项任务上取平均后，不同未见过的相机参数下的策略性能（RQ3）。

5. 结论

我们首次系统分析了用于模仿学习的腕部安装的鱼眼相机，考察了空间定位、场景泛化与硬件泛化。主要发现如下：（1）宽视场的定位收益关键取决于环境特征丰富度。（2）鱼眼相机能够更好地利用数据多样性，从而实现更优的场景泛化。（3）由尺度过拟合导致的跨相机迁移失败，可通过我们提出的随机尺度增强得到有效解决。这些发现为机器人学界提供了具体、可操作的指导原则，并为使用鱼眼相机开展大规模数据收集以及训练稳健的通用型机器人策略奠定了实证基础。

参考文献

[1] 使用双鱼眼相机的自主空中机器人 - gao - 2020年 - 《野外机器人学杂志》 - Wiley在线图书馆。https://onlinelibrary.wiley.com/doi/10.1002/rob.21946。2[2] Kevin Black、Noah Brown、Danny Driess、Adnan Esmail、Michael Equi、Chelsea Finn、Niccolo Fusai、Lachy Groom、Karol Hausman、Brian Ichter等人。 π_0：面向通用机器人控制的视觉-语言-动作流模型。arXiv预印本 arXiv:2410.24164，2024年。1, 2, 21[3] Sylvain Calinon、Florent D'halluin、Eric L. Sauser、Darwin G. Caldwell和Aude G. Billard。通过模仿学习并复现手势。IEEE《机器人与自动化杂志》，17(2):44–54，2010年。2[4] Cheng Chi、Zhenjia Xu、Siyuan Feng、Eric Cousineau、Yilun Du、Benjamin Burchfiel、Russ Tedrake和Shuran Song。扩散策略：通过动作扩散进行视觉运动策略学习。《国际机器人研究杂志》，2024年。2, 4, 12, 21[5] Cheng Chi、Zhenjia Xu、Chuer Pan、Eric Cousineau、Benjamin Burchfiel、Siyuan Feng、Russ Tedrake和Shuran Song。通用操作接口：无需真实环境机器人的真实环境机器人教学。载于《机器人学：科学与系统》（RSS）会议论文集，2024年。1, 2, 5, 15[6] Pim de Haan、Dinesh Jayaraman和Sergey Levine。模仿学习中的因果混淆。载于《神经信息处理系统进展》，第11693–11704页，2019年。2[7] 大寰机器人科技有限公司。AG系列连杆式自适应电动夹爪。https://en.dh-robotics.com/product/ag，2025年。5[8] Alexey Dosovitskiy。一张图像相当于16x16个词：用于大规模图像识别的Transformer。arXiv预印本 arXiv:2010.11929，2020年。4[9] Ciar´an Eising、Jonathan Horgan和Senthil Yogamani。使用环视鱼眼相机的低速车辆自动化近场感知。IEEE智能交通系统汇刊，23 13993，2022年 9 13976 1, 2 ( ): – 。[10] I-Sheng Fang、Yue-Hua Han和Jun-Cheng Chen。将相机设置作为词元：在潜在扩散模型上建模摄影。载于SIGGRAPH Asia 2024会议论文，第1–11页，美国纽约州纽约市，2024年。美国计算机协会。2[11] Flexiv。Rizon——具备力控制的自适应七轴机械臂。https://www.flexiv.com/products/rizon，2025年。5[12] Zipeng Fu、Tony Z. Zhao和Chelsea Finn。Mobile ALOHA：利用低成本全身遥操作学习双臂移动操作。载于第八届机器人学习会议论文集，第4066–4083页。PMLR，2025年。2[13] Dibya Ghosh、Homer Rich Walke、Karl Pertsch、Kevin Black、Oier Mees、Sudeep Dasari、Joey Hejna、Tobias Kreiman、Charles Xu、Jianlan Luo等人。Octo：一种开源通用型机器人策略。载于《机器人学：科学与系统》，2024年。1[14] Huy Ha、Pete Florence和Shuran Song。扩展规模与蒸馏压缩：语言引导的机器人技能习得。载于《弥合认知科学与真实世界机器人学习之间的鸿沟：进展与新方向》，2024年。2[15] Jan Hausberg、Ryoichi Ishikawa、Menandro Roxas 和 Takeshi Oishi。利用激光雷达和鱼眼相机，通过直接与间接观测实现无人机—地面车辆相对定位，2020年。2[16] 何恺明、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于 IEEE计算机视觉与模式识别会议论文集，第770–778页，2016年。4，7[17] Jonathan Ho 和 Stefano Ermon。生成对抗模仿学习。载于《神经信息处理系统进展》第29卷，第4565–4573页。Curran Associates, Inc.，2016年。2[18] Yingdong Hu、Fanqi Lin、Pingyue Sheng、Chuan Wen、Jiacheng You 和 Gao Yang。机器人操作模仿学习中的数据缩放定律。载于第一届 X-Embodiment 机器人学习研讨会，2024年。2，5，7，12，16[19] Stephen James、Zicong Ma、David Rovick Arrojo 和 Andrew J. Davison。RLBench：机器人学习基准与学习环境。《IEEE机器人与自动化快报》，5(2):3019–3026，2020年。2[20] Shunping Ji、Zijie Qin、Jie Shan 和 Meng Lu。基于多鱼眼相机阵列的全景同步定位与建图。《国际摄影测量与遥感杂志》，159:169–183，2020年。1，2[21] Zhenyu Jiang、Yuqi Xie、Kevin Lin、Zhenjia Xu、Weikang Wan、Ajay Mandlekar、Linxi Jim Fan 和 Yuke Zhu。DexMimicGen：通过模仿学习自动生成双臂灵巧操作数据。载于2025年IEEE国际机器人与自动化会议（ICRA），第16923–16930页，2025年。2[22] Tsung-Wei Ke、Nikolaos Gkanatsios 和 Katerina Fragkiadaki。3D扩散器演员：基于3D场景表示的策略扩散。载于《机器人学习第八届会议论文集》，第1949–1974页。PMLR，2025年。2[23] Bernhard Kerbl、Georgios Kopanas、Thomas Leimk¨uhler 和 George Drettakis。用于实时辐射场渲染的3D高斯泼溅。《ACM图形学汇刊》，42(4)，2023年。2[24] Janusz Konrad、Mertcan Cokbas、M. Ozan Tezcan 和 Prakash Ishwar。用于室内监控的俯视鱼眼相机：挑战与最新进展。《成像前沿》，3，2024年。2[25] Varun Ravi Kumar、Senthil Yogamani、Markus Bach、Christian Witt、Stefan Milz 和 Patrick M¨ader。UnRectDepthNet：采用通用框架处理常见相机畸变模型的自监督单目深度估计。载于2020年IEEE/RSJ智能机器人与系统国际会议（IROS），第8177–8183页，2020年。2[26] Wenhui Li、You Qu、Ying Wang 和 Jialun Liu。鱼眼图像的相机特定仿真方法。载于2019年第14届IEEE智能系统与知识工程国际会议（ISKE），第955–961页，2019年。2[27] Yi-Ruei Liu、You-Zhe Xie、Yu-Hsiang Hsu、I.-Sheng Fang、Yu-Lun Liu 和 Jun-Cheng Chen。每一种相机效果，每时每刻，同时实现：用于生成基于物理的相机效果数据的4D高斯光线追踪，

每时每刻，同时实现：用于生成基于物理的相机效果数据的4D高斯光线追踪，2025年。2[28] Ajay Mandlekar、Danfei Xu、Josiah Wong、Soroush Nasiriany、Chen Wang、Rohun Kulkarni、Li Fei-Fei、Silvio Savarese、Yuke Zhu 和 Roberto Mart´ın-Mart´ın。离线人类示范中哪些因素对机器人操作学习至关重要。载于机器人学习会议，第1678–1690页。PMLR，2022年。2，3，5[29] Ajay Mandlekar、Soroush Nasiriany、Bowen Wen、Iretiayo Akinola、Yashraj Narang、Linxi Fan、Yuke Zhu 和 Dieter Fox。MimicGen：利用人类示范实现可扩展机器人学习的数据生成系统。载于《迈向通用机器人：可扩展技能习得的学习范式 @ CoRL2023》，2023年。2，3，5[30] Meta。Meta Quest 3。https://www.meta.com/quest/quest-3，2025年。5[31] Tongzhou Mu、Zhan Ling、Fanbo Xiang、Derek Yang、Xuanlin Li、Stone Tao、Zhiao Huang、Zhiwei Jia 和 Hao Su。ManiSkill：具有大规模示范数据的泛化操作技能基准。arXiv电子预印本，文章编号 arXiv:2107.14483，2021年。2[32] Dean A. Pomerleau。ALVINN：基于神经网络的自主陆地车辆。载于第二届神经信息处理系统国际会议论文集，第305–313页，美国马萨诸塞州剑桥，1988年。麻省理工学院出版社。2[33]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议，第8748–8763页。PMLR，2021年。4，6，7，14[34]Varun Ravi Kumar、Senthil Yogamani、Hazem Rashed、Ganesh Sitsu、Christian Witt、Isabelle Leang、Stefan Milz 和 Patrick M¨ader。OmniDet：基于环视摄像机的自动驾驶多任务视觉感知网络。《IEEE机器人与自动化快报》，6(2):2830–2837，2021年。1，2[35] Yuan Ren、Guile Wu、Runhao Li、Zheyuan Yang、Yibo Liu、Xingxin Chen、Tongtong Cao 和 Bingbing Liu。UniGaussian：通过统一高斯表示，利用多种相机模型重建驾驶场景。CoRR，abs/2411.15355，2024年。2[36] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。U-Net：用于生物医学图像分割的卷积网络。载于国际医学图像计算与计算机辅助介入会议，第234–241页。施普林格，2015年。4[37] St´ephane Ross、Geoffrey J. Gordon 和 J. Andrew Bagnell。将模仿学习和结构化预测归约为无悔在线学习。《机器学习研究期刊》，15:627–635，2011年。2[38] Ethan Rublee、Vincent Rabaud、Kurt Konolige 和 Gary Bradski。ORB：SIFT 或 SURF 的一种高效替代方案。载于2011年国际计算机视觉会议，第2564–2571页，2011年。18，19[39] Kaustubh Sadekar。OmniCV-Lib：用于全向（360度）相机的计算机视觉库。https://github.com/kaustubh- sadekar/OmniCV- Lib，2020年。2, 3[40]Vaibhav Saxena、Matthew Bronars、Nadun Ranawaka Arachchige、Kuancheng Wang、Woo Chul Shin、Soroush Nasiriany、Ajay Mandlekar 和 Danfei Xu。《机器人操作大规模数据集学习中的关键因素》。arXiv预印本 arXiv:2506.13536，2025年。16[41] Stefan Schaal。模仿学习是通往人形机器人的途径吗？《认知科学趋势》，3：233–242，1999。2[42] Tobias Scheck、Roman Seidel 和 Gangolf Hirtz。从 theodore 中学习：用于深度迁移学习的合成全向俯视室内数据集。发表于 IEEE/CVF冬季计算机视觉应用会议（WACV）论文集，2020年。2[43] Ahmed Rida Sekkat、Yohan Dupuis、Varun Ravi Kumar、Hazem Rashed、Senthil Yogamani、Pascal Vasseur 和 Paul Honeine。Synwoodscape：用于自动驾驶的合成环视鱼眼相机数据集。《IEEE机器人与自动化快报》，7(3)：8502–8509，2022年。2[44]Changha Shin、Woong Oh Cho 和 Seon Joo Kim。Seam360gs：基于真实全向图像的无缝360度高斯溅射。发表于 IEEE/CVF 计算机视觉国际会议（ICCV）论文集，第 28970–28979页，2025年。2[45] Jiaming Song、Chenlin Meng 和 Stefano Ermon。去噪扩散隐式模型。发表于国际学习表征会议。4[46] Generalist AI团队。GEN-0：通过物理交互扩展规模的具身基础模型。Generalist AI博客，2025年。https://generalistai.com/blog/nov-04-2025-GEN-0。1, 2, 3, 5, 7, 15[47] RDT团队。RDT2：通过扩展规模的UMI数据实现零样本跨具身泛化，2025年。1, 3, 5, 7[48] Emanuel Todorov、Tom Erez 和 Yuval Tassa。MuJoCo：用于基于模型控制的物理引擎。发表于2012年IEEE/RSJ智能机器人与系统国际会议论文集，第5026–5033页。IEEE，2012年。2, 3, 4, 5[49] Andrey Voynov、Amir Hertz、Moab Arar、Shlomi Fruchter 和 Daniel Cohen-Or。曲线扩散：一种具有光学几何控制的生成模型。发表于《计算机视觉——ECCV 2024：第18届欧洲计算机视觉会议，米兰，意大利，2024年9月29日至10月4日，论文集》，第LXXVII部分，第149–164页，柏林、海德堡，2024年。施普林格出版社。2[50] Chen Wang、Haochen Shi、Weizhuo Wang、Ruohan Zhang、Li Fei-Fei 和 Karen Liu。DexCap：用于灵巧操作的可扩展便携式动作捕捉数据采集系统。发表于RSS 2024研讨会：机器人数据生成，2024年。2[51] Yahui Wang、Shaojun Cai、Shi-Jie Li、Yun Liu、Yangyan Guo、Tao Li 和 Ming-Ming Cheng。CubemapSLAM：分段针孔单目鱼眼SLAM系统。arXiv电子印本，文章编号 arXiv:1811.12633，2018年。2[52] Chuan Wen、Jierui Lin、Trevor Darrell、Dinesh Jayaraman 和 Gao Yang。利用观测历史进行行为克隆时对抗模仿者智能体。《神经信息处理系统进展》，33：2564–2575，2020年。2[53] Zizhang Wu、Wenkai Zhang、Jizheng Wang、Man Wang、Yuanzhu Gan、Xinchao Gou、Muqing Fang 和 Jing Song。

解耦与向量化：一种基于环视鱼眼相机的自动驾驶三维视觉感知方法。发表于2021年IEEE/RSJ智能机器人与系统国际会议（IROS）论文集，第5576 5582页，2021年 2 – 。[54] Weijian Xie、Guanyi Chu、Quanhao Qian、Yihao Yu、Shangjin Zhai、Danpeng Chen、Nan Wang、Hujun Bao 和 Guofeng Zhangv。面向背靠背鱼眼相机的全向稠密同步定位与建图。发表于2024年IEEE机器人与自动化国际会议（ICRA）论文集，第1653 1660页，2024年 1, 2 – 。[55] Tao Yang、Zhi Li、Fangbing Zhang、Bolin Xie、Jing Li 和 Linfeng Liu。基于无结构相机阵列的全景无人机监控与回收系统。《IEEE Access》，7：25763–25778，2019年。2[56] Senthil Yogamani、Ciaran Hughes、Jonathan Horgan、Ganesh Sistu、Sumanth Chennupati、Michal Uricar、Stefan Milz、Martin Simon、Karl Amende、Christian Witt、Hazem Rashed、Sanjaya Nayak、Saquib Mansoor、Padraig Varley、Xavier Per-rotton、Derek Odea 和 Patrick Perez。Woodscape：用于自动驾驶的多任务、多相机鱼眼数据集。发表于2019年IEEE/CVF计算机视觉国际会议（ICCV）论文集，第9307–9317页，韩国首尔，2019年。IEEE。2[57] Yanjie Ze、Gu Zhang、Kangning Zhang、Chenyuan Hu、Muhan Wang 和 Huazhe Xu。三维扩散策略：通过简单的三维表征实现可泛化的视觉运动策略学习。发表于第2届灵巧操作研讨会：设计、感知与控制（RSS），2024年。2[58] Juntu Zhao、Wenbo Lu、Di Zhang、Yufeng Liu、Yushen Liang、Tianluo Zhang、Yifeng Cao、Junyuan Xie、Yingdong Hu、Shengjie Wang 等。视觉运动策略是否需要本体感觉状态？arXiv预印本 arXiv:2509 18644，2025年 5 。 。[59] Zhongyi Zhou、Yichen Zhu、Junjie Wen、Chaomin Shen 和 Yi Xu。利用预训练知识进行开放世界具身推理的视觉-语言-动作模型。arXiv预印本 arXiv:2505.21906，2025年。1, 2

重新思考相机选择：关于机器人操作中鱼眼相机特性的实证研究的补充材料

A. 概述

本补充材料提供了全面的实现细节、扩展分析和额外的真实世界验证，以支持主论文中呈现的研究结果。材料组织如下：

• **第 B（实验设置细节）**重点介绍实验设置细节，涵盖每个仿真任务的可视化、双相机设置以及实验场景的比较。

• **第D（实现细节）**提供训练所使用的具体超参数，以及真实世界实验中归一化得分的定义。

• **第D（RQ1：空间定位）**详细介绍探测模型，并呈现关于本体感知的额外消融研究。

•**第 E（研究问题2：场景泛化）**详细介绍了实验中使用的场景数据集，并提供了仿真和真实世界任务的细粒度曲线。

• **第 F（RQ3：硬件泛化）**说明了仿真中的相机参数，并展示了仿真和真实世界中跨相机实验的定量结果。

B. 实验设置细节

B.1. 仿真任务

图中展示了所有仿真任务。S1。这些任务涵盖了多种行为，包括抓放、精确插入（例如穿线和正方形任务），以及需要将多种行为串联起来完成的长时程任务（例如工具悬挂和马克杯清理）。每项任务的详细配置，包括轨迹数量和数据来源，见表 S1。

B.2. 仿真中的双相机设置

为了扩大视场以实现更全面的场景感知，我们在仿真环境中添加了一个与原腕部摄像头相对的新相机。原腕部摄像头的位置参数为 pos="0.05 0 0"，新增相机则放置在其相对一侧，位置参数为 pos="−0.05 0 0"。二者呈对称分布，从而有效扩大了视场。不同相机的可视化结果如图S2

表 S1. **仿真任务概览**。PH：熟练人类数据集；D0：默认重置分布；D1：扩展重置分布。

Task	轨迹数量	数据来源	数据类型
正方形	200	Robomimic	PH
工具悬挂	200	Robomimic	PH
咖啡	500	MimicGen	D1
穿线	500	MimicGen	D0
组装	500	MimicGen	D0
马克杯清理	500	MimicGen	D0

C. 实现细节

C.1. 训练超参数

我们的策略实现基于 Diffusion Policy 框架 [4]。遵循 [18]，中确立的协议，我们严格控制超参数，以便公平比较鱼眼相机和针孔相机。

仿真和真实世界实验的具体超参数详见表S2。

表 S2. **详细超参数**。 我们报告仿真（Sim）和真实世界（Real）实验所用的具体设置。超参数值与 [4] 和 [18]保持一致。

配置	仿真	真实世界
模型架构		
视觉骨干网络	ResNet-18（无预训练）	CLIP ViT-B/16
池化方法	空间 Softmax	空间 Softmax
去噪网络	条件 U-Net1D	条件 U-Net1D
动作空间	相对动作	相对动作
动作预测时域（T_{act}）	8	8
观测时域（T_{obs}）	2	2
预测时域（T_{pred}）	16	16
输入数据		
图像分辨率	128×128	224×224
图像预处理	随机裁剪	随机裁剪
本体感觉输入	**无**（无状态）	**无**（无状态）
优化		
优化器	AdamW	AdamW
权重衰减	1×10^{-6}	1×10^{-6}
学习率调度	余弦衰减	余弦衰减
学习率（UNet）	1×10^{-4}	1×10^{-4}
学习率（编码器）	1×10^{-4}	1×10^{-5}
批次大小	16	64
训练轮数	2000	500
EMA 衰减	0.75	0.9999

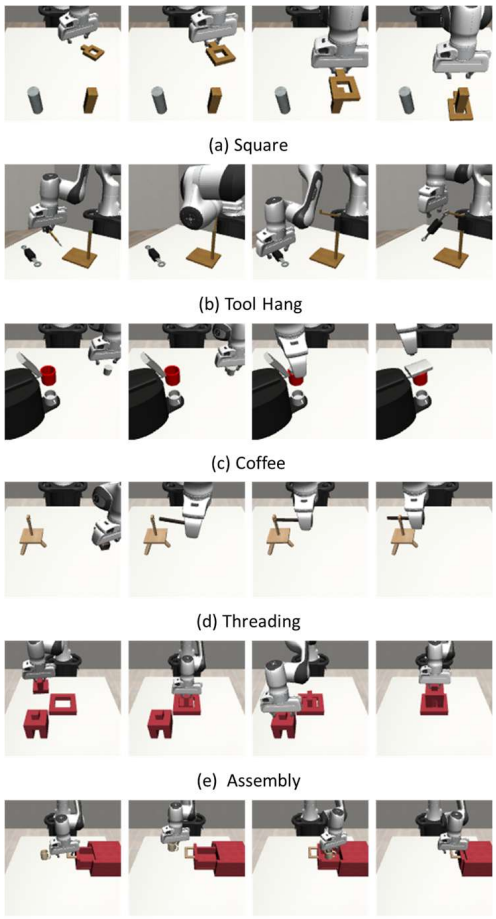


图 S1. **仿真任务示例**。每行展示一个不同任务的连续快照：(a) 正方形，抓取黄色方块并将其插入黄色目标槽。(b) 工具悬挂，插入针形挂钩并悬挂工具。(c) 咖啡，将咖啡胶囊放入机器并合上盖子。(d) 穿线，抓取针并将其插入针孔。(e) 组装，依次插入两个不规则方块。(f) 马克杯清理，打开抽屉，将马克杯放入其中，然后关上抽屉。

C.2. 真实世界任务得分指标定义

如正文所述，我们在真实世界评估中采用归一化多阶段评分指标。与二元成功率相比，该指标能够更细致地评估策略能力。各项任务的详细评分标准如下：

• **拿起杯子**：目标是拿起杯子，并将其直立放在杯垫上。– 阶段 *1* （ *0.00* 分）：未能抓取杯子，或抓取了杯子但未能将其放到杯垫上（例如，杯子掉落或未放到目标位置）。– 阶段*2*（*0.50* 分）：成功抓取杯子并将其放到杯垫上，但杯子倾倒了（未保持直立

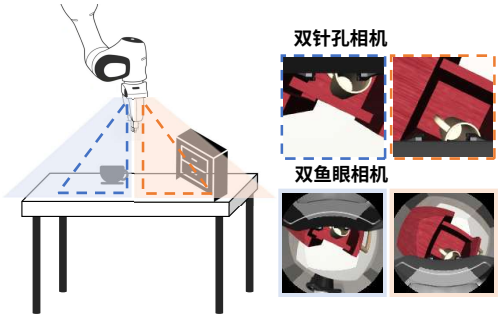


图 S2。**不同相机的可视化**。橙色区域表示针孔腕部相机的视场，蓝色区域表示新添加的对称排列相机的视场。上方显示针孔腕部相机拍摄的图像，下方显示双鱼眼相机拍摄的图像，蓝色和橙色边框分别区分与其各自视场对应的成像效果。

）。– 阶段*3*（*1.00* 分）：成功抓取杯子并将其放到杯垫上，且保持直立。

• **折叠毛巾**：目标是连续折叠毛巾两次。机器人必须抓取一个角，将其折向对角线，然后抓取另一个角，完成第二次折叠。

– 阶段*1*（*0.25* 分）：成功抓取第一个角并将其抬离桌面。– 阶段*2*（*0.50* 分）：成功松开夹爪并完成第一次折叠。– 阶段*3*（*0.75* 分）：成功定位并抓取第二个角（第一次折叠后），并将其抬离桌面。– 阶段*4*（*1.00* 分）：成功松开夹爪并完成第二次折叠。

• **悬挂中国结**：目标是将中国结挂到支架上的指定挂钩上。由于大多数基线方法都能成功完成初始抓取阶段，因此我们不将其纳入评分标准。我们只关注将中国结牢固挂好所需的精准放置。

– 阶段*1*（*0.00*分）：未能将中国结挂到挂钩上（例如，中途掉落或未挂中挂钩）。– 阶段*2*（*1.00*分）：成功将中国结移至目标位置，并将其牢固挂到挂钩上。

对于每种评估设置，我们进行20次试验，并报告**累计得分**（所有试验得分的总和）。

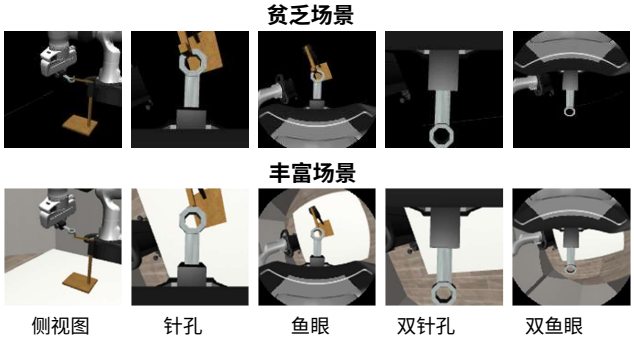
D. RQ1（空间定位）的补充实验

本节首先展示 RQ1 仿真实验中的场景（第D.1节）。随后，我们详细介绍用于探测视觉编码器空间感知能力的模型细

节（第D.2节）。此外，为进一步验证我们关于空间定位的发现，我们评估了策略在使用本体感觉输入时的表现（第 D.3 节和第D.4节），以及使用第三视角输入时的表现（第D.5节）。

D.1. 仿真中的实验场景

图S3展示了RQ1实验中工具悬挂任务的一个场景示例，比较了包括第三视角相机、针孔相机和鱼眼相机在内的不同相机在两种不同环境设置下的可视化效果：贫乏场景（单一暗场景）和丰富场景（包含多样元素的场景）。



图S3. **RQ1中的场景对比**。贫乏（单一暗色）场景与丰富（多样元素）场景中侧视、针孔、鱼眼、双针孔和双鱼眼相机的可视化。

D.2. 实现细节：本体感知预测任务

方法。 在论文正文中，我们证明了基于鱼眼相机的策略在真实世界任务中的本体感觉预测误差更低。此处，我们详细介绍这项探测任务的实验设置。为了明确量化视觉表征所捕获的空间信息，我们在本体感觉预测任务上对视觉编码器进行微调。具体而言，我们从已学习的操作策略中提取视觉编码器（CLIP ViT [33]），并连接一个轻量级多层感知机（MLP）头。该 MLP 由两个全连接层组成，包含 256 个隐藏单元，并采用 ReLU 激活函数。我们使用均方误差（MSE）损失对整个网络（编码器 +MLP 头）进行微调，以回归机器人的末端执行器位姿（3D 位置和四元数方向）。为了严格评估所学习空间表征的泛化能力，我们没有简单地划分训练数据，而是针对每种任务设置专门采集了一组**由 30 条额外轨迹组成的测试集**。这些轨迹是在与训练数据完全相同的环境设置下采集的，但在训练阶段严格保留为未见数据。

D.3. 消融实验：带本体感知的仿真策略性能

动机。 如第 3.3 节所述，为了严格隔离并评估不同相机模型（RQ1）的视觉空间定位能力，我们有意从策略输入中排除了本体感知状态（例如末端执行器位姿和关节位置）。然而，在实际的机器人应用中，通常可以获取本体感知信息。本节中，我们首先在仿真中进行消融实验，以研究加入本体感知状态如何影响使用针孔相机和鱼眼相机训练的策略。本节中的所有实验均在“特征丰富”背景设置下进行，以确保比较公平。

仿真实验设置。本节所有实验均在“特征丰富”的背景设置下进行，以确保公平比较。我们评估了六项仿真任务的表现。为简洁起见，在表S3中，我们使用以下缩写：**Tool** （工具悬挂）、**Sqr**（正方形）、**Cof** （咖啡）、**Thrd**（穿线）、**Asm** （组装）和 **Mug** （马克杯清理）。我们比较两种腕部相机配置：

- **单个：**机器人配备一台腕部安装相机。
- **双个：**机器人配备两台腕部安装相机（提供不同视角）以减少遮挡。

结果与分析。定量结果汇总于表S3中。**平均值**列清楚地展现了两种相机设置对本体感知依赖程度的差异：

1.**针孔相机敏感性：**移除本体感知后，基于针孔相机的策略性能急剧下降。对于单个针孔相机配置，平均成功率从0.62骤降至0.34（下降**45%**）。对于双针孔相机配置，平均成功率从0.64骤降至0.45（下降**30%**）。这证实了在缺少机器人状态的明确指导时，窄视场难以保持稳定定位。

2. **鱼眼鲁棒性：**与之形成鲜明对比的是，基于鱼眼相机的策略表现出显著的鲁棒性。单鱼眼配置的均值成功率仍然很高（仅从 0.75 略微下降至 **0.66**），而双鱼眼设置的成功率下降幅度微乎其微（从 0.81 降至 **0.75**）。这一实证结果证明，鱼眼相机的广角视野能够有效地隐式编码机器人与环境之间的空间关系，因此显式状态输入在很大程度上是多余的。

D.4. 消融实验：具有本体感知能力时的策略性能

真实世界实验设置。 为了在真实世界中验证我们的仿真发现，我们使用正文所述的真实世界设置开展了相同的消融研究

表 S3. **仿真中本体感知输入的消融实验**。 我们比较了有（有状态）和无（无状态）本体感知输入时的成功率。(·)中的数值表示相对于基线有状态性能的下降（或提升）。针孔相机的相对性能大幅下降（例如，平均下降 41%），而鱼眼相机则保持了稳定的性能（例如，平均仅下降 6%）。

实验因素		仿真任务成功率						平均值
相机	状态	正方形	工具悬挂	咖啡	螺纹装配	马克杯清洁	_	
针孔（单目）	带状态	0.82	0.68	0.38	0.72	0.46	0.66	0.62
	无状态	0.48（-41%）	0.56（-18%）	0.34（-11%）	0.18（-75%）	0.12（-74%）	0.38（-42%）	0.34（-45%）
鱼眼（单目）	有状态	0.86	0.88	0.88	0.72	0.58	0.60	0.75
	无状态	**0.74**（-14%）	**0.84** (-5%)	**0.76** (-14%)	**0.56** (-22%)	**0.48** (-17%)	**0.60** (0%)	**0.66** (-12%)
双针孔	使用状态	0.92	0.54	0.38	0.76	0.58	0.66	0.64
	不使用状态	0.70 (-24%)	0.34 (-37%)	0.36 (-5%)	0.38 (-50%)	0.34 (-41%)	0.56 (-15%)	0.45 (-30%)
鱼眼（双重）	带状态	0.94	0.88	0.88	0.78	0.56	0.80	0.81
	不含状态	**0.88** (-6%)	**0.88** (0%)	**0.86** (-2%)	**0.66** (-15%)	**0.44** (-21%)	**0.80** (0%)	**0.75** (-7%)

论文。我们采用“特征丰富”（可更换且纹理丰富的背景）设置，以最大化视觉特征提取的潜力。我们比较了策略在有无本体感知情况下，于三项真实世界任务中的归一化得分：**拿起杯子**、**折叠毛巾**和**悬挂中国结**。

表 S4。**真实世界本体感知消融**。我们报告了特征丰富设置下的归一化得分。括号中的数值（·）表示相对于基线有状态性能的下降幅度。值得注意的是，不使用本体感知的鱼眼相机策略（0.67）表现优于即使使用本体感知的针孔相机策略（0.52）。

实验因素		真实世界归一化分数			平均值
相机	状态	拿起杯子	折叠毛巾	悬挂 打结	
针孔	带状态	0.75	0.37	0.45 0.52	
	无状态	0.65(-13%)	0.32(-14%)	0.15(-67%) 0.37(-29%)	
鱼眼	有状态	0.98	0.92	0.70	0.87
	无状态	**0.80**(-18%)	**0.70**(-24%)	**0.50**(-29%) **0.67**(-23%)	

结果。 表 S4 中展示的真实世界结果表明，与仿真相比，鱼眼相机的优势更加显著：

1. **卓越的空间定位：** 最关键的指标是不使用本体感知（不含状态）时的表现，它体现了相机纯粹依靠视觉进行定位的能力。在这一情形下，鱼眼相机的表现显著优于针孔相机。例如，在具有挑战性的可变形物体任务（折叠毛巾）中，鱼眼相机策略在不使用状态信息时得分为**0.70**，而针孔相机策略的得分仅为**0.32**。平均而言，仅使用视觉时，鱼眼相机的平均得分为**0.67**，几乎是针孔相机平均得分**0.37**的两倍。
2. **降低状态依赖：** 虽然加入本体感知信息后，两种相机的表现都有所提升（这可能是由于真实世界固有的噪声和动态性所致），但针孔相机对这类信息的依赖要强得多。在悬挂中国结 任务中，针孔相机策略依赖状态信息，才能将原本不及格的 0.15 分提升至 0.45 分。相比之下，鱼眼相机策略仅凭视觉就已达到 0.50 的较高基线。

消融总结。 这些真实世界实验与仿真结果一致，证实了我们的假设：鱼眼相机的宽视场能够捕捉足够的全局背景信息，即使没有机器人状态信息，也能实现高精度操作。总体而言，这些结果进一步支持了我们在 RQ1 中的结论：鱼眼相机天生具备更出色的空间定位能力，显著降低了对精确机器人状态输入的依赖。

D.5. 消融研究：第三视角相机集成的影响

虽然主要研究通过排除额外传感器，单独考察了腕部安装相机的影响，但在真实世界的机器人部署中，通常会采用多模态配置，例如将腕部相机与本体感觉或第三人称视角结合起来。为探究鱼眼相机在这些更复杂感知范式中的表现，我们在仿真中评估了“腕部 + 第三人称”配置。尽管近期最先进的框架（如 UMI[5] 和 GEN-0[46]）主要依赖腕部相机，但我们的探索为更广泛的部署场景提供了有益见解。

表 S5. 仿真中的第三视角相机消融 on.

配置（双摄像机 + 3rd）	Sqr	Tool	Cof	Thrd	Asm	马克杯	平均值
针孔基线	0.94	0.78	0.78	0.80	0.56	0.66	0.75
鱼眼（**我们的方法**）	**0.96**	**0.84**	**0.82**	**0.82**	**0.66**	**0.72**	**0.80**
改进	*+0.02*	*+0.06*	*+0.04*	*+0.02*	*+0.10*	*+0.06*	***+0.05***

如表S5所总结，鱼眼相机始终比针孔相机带来 5% 的均值性能提升

即使提供了第三人称视角，基线性能依然如此。值得注意的是，在高精度的“组装”（Assembly，简称“Asm”）任务中，鱼眼配置实现了 10% 的改进，这表明即使存在全局视角，鱼眼相机的宽视场优势也并非冗余。相反，鱼眼镜头能够提供关键的局部上下文信息，例如精确的夹爪-物体相对位姿；固定的全局相机可能会因遮挡或分辨率有限而难以捕捉这些信息。这种稳定的性能提升证明，我们研究中发现的鱼眼相机优势具有稳健性，并且能够推广到更全面的传感器套件中，进一步证明了在未来的通用型机器人策略中采用鱼眼相机的合理性。

E. RQ2（场景泛化）的补充实验

E.1. 环境多样性可视化

动机。 在主论文中，我们证明了鱼眼相机的宽视场能够显著增强空间定位能力，尤其是在特征丰富环境中（RQ1）；同时，这种能力会随着场景多样性的增加而提升（RQ2）。为验证我们的实验设置能够严格评估泛化能力，而非简单的记忆，我们同时提供了背景数据集的定性可视化和定量分布分析。

可视化设置。 图S4 展示了本研究所采用的多样化纹理。在构建包含不同场景数量的数据集时（N），场景泛化实验的一个关键方面，是将环境多样性带来的收益与数据规模增加带来的收益区分开来。因此，不同于以往通过扩大数据量 [18], 的工作，我们采用固定总数据量协议。

• **仿真环境：** 这些纹理来源于 MimicLab [40]资产库，用于替换原始场景材质。纹理类型从几何图案到自然材质不等，引入了对编码器构成挑战的高频视觉特征。训练使用了 32 种纹理，测试则使用了另外 5 种此前未见过的背景。

我们在所有实验中均采用咖啡任务，并将数据预算固定为**500 条轨迹**。随着不同场景的数量（N）增加，我们轮换使用不同的场景渲染，以确保轨迹在各环境中均匀分布。例如，在 $N = 32$设置中，我们采用均衡组合：20 个场景各贡献 16 条轨迹，另有 12 个场景各贡献 15 条轨迹，从而保持总数为 500 条轨迹。

• **真实世界场景：**我们利用一组图案布料（例如抽象艺术、网格）设计了一套可变背景系统，以引入多样化的视觉外观。总共有8个不同的背景场景被构建用于训练。为评估零样本泛化能力，我们采用了另外一组此前未见的4个背景场景进行测试。

“拿起杯子”任务在真实环境中进行了测试。所有实验的训练预算均固定为**200 条轨迹**。随着不同场景数量（N）的增加，我们将轨迹预算均匀分配到各场景中。例如，在 $N = 1$ 设置中，我们使用来自单个场景的 200 条轨迹；在 $N = 8$ 设置中，我们使用来自 8 个场景的轨迹，每个场景 25 条。这样可以确保任何性能提升都仅归因于视觉数据多样性的增加，而非演示数据数量的增加。

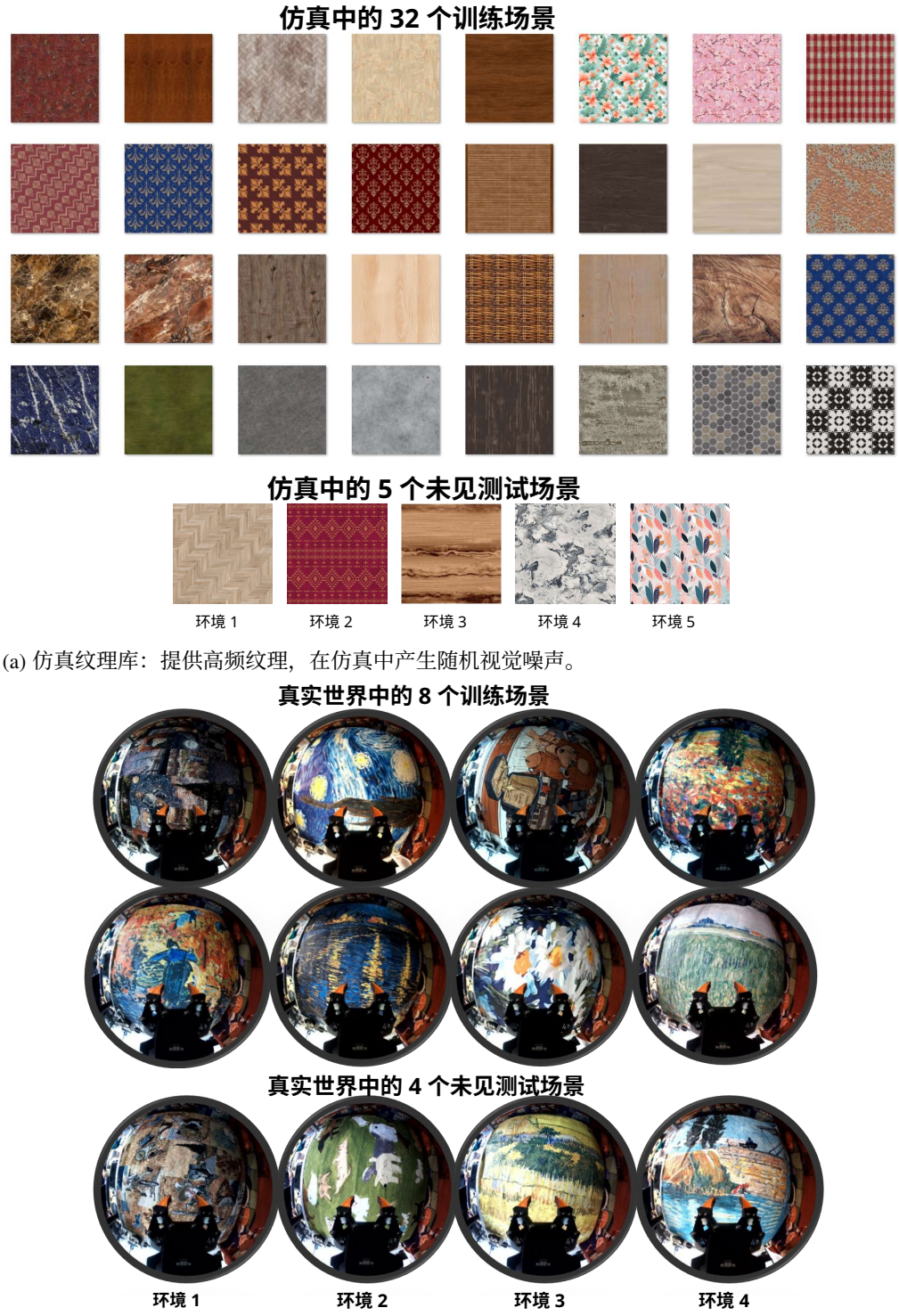
数据分布分析。 为确保我们的训练集/测试集划分在统计上严谨，并覆盖可能环境的语义空间，我们采用了数据驱动的选择策略，而非随意手动选择。具体而言，我们使用预训练 CLIP 视觉编码器的 CLS 标记，提取所有候选场景的全局语义特征。随后，我们对这些嵌入进行 K 均值聚类，以识别不同的视觉簇。

• **仿真**（$K = 8$）**：** 如图S5(a)所示，仿真纹理聚为 8 个不同的簇。
• **真实世界**（$K = 4$）**：** 如图S5(b)所示，真实世界场景聚为 8 个不同的簇。
为了构建均衡的评估协议，我们从每个簇中恰好抽取一个代表性场景，作为**留出测试集**（以红框标示），其余场景则构成训练集。这种方法确保测试集不会偏向任何特定的纹理类型，并能严格检验策略在整个视觉分布范围内的泛化能力。我们在仿真实验中采用设置 $N \in \{1, 8, 16, 32\}$，在真实世界实验中采用设置 $N \in \{1, 2, 4, 6, 8\}$，从而确保环境覆盖范围逐步扩大。

E.2. 逐场景泛化分析

研究动机。 在论文正文中，我们证明了增加训练场景的多样性（N）可以显著提升未见环境中的平均零样本泛化性能。然而，汇总平均指标可能会掩盖不同测试场景之间的难度差异。例如，策略可能在视觉上较为简单的背景中表现出色，却在更复杂的背景中失利，从而导致平均值虚高。为考察泛化的一致性，我们进行了细致的分项分析，绘制了每个个体 未见测试场景的性能扩展曲线。

分析设置。 我们将图 8 中的汇总结果拆分为每个留出环境的具体性能轨迹：

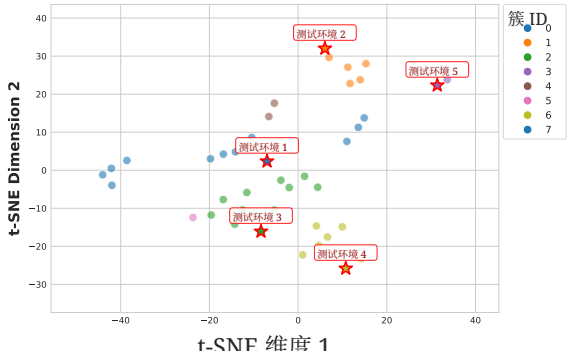


(a) 仿真纹理库：提供高频纹理，在仿真中产生随机视觉噪声。

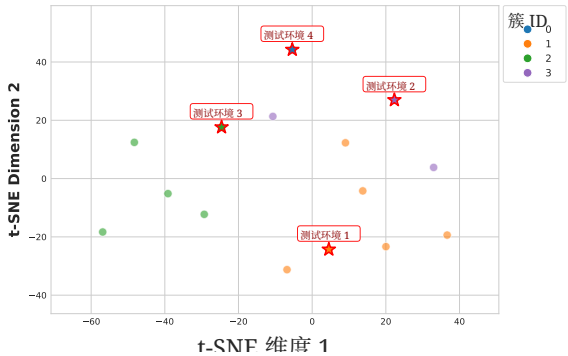
(b) 真实世界背景：图案布料提供多样的视觉特征，用于真实世界实验

R S.

图S4. **场景多样性的综合可视化**。我们通过确保仿真和真实世界两个领域都具有较高的场景复杂度，明确弥合两者之间的视觉差距。（**a**）在仿真中，我们将 32 种不同的纹理（木材、大理石、织物、瓷砖）随机应用于背景表面。（**b**）在真实世界中，我们使用可变背景，并采用视觉上各异的纹理布料，严格测试空间定位和泛化能力。

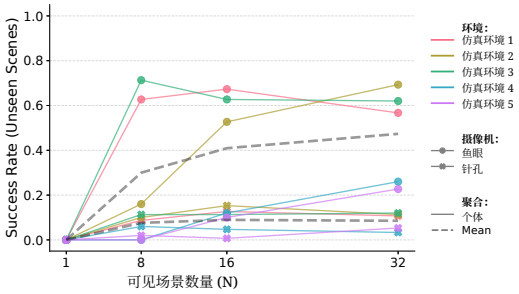


（a）仿真场景分布（$K = 8$）

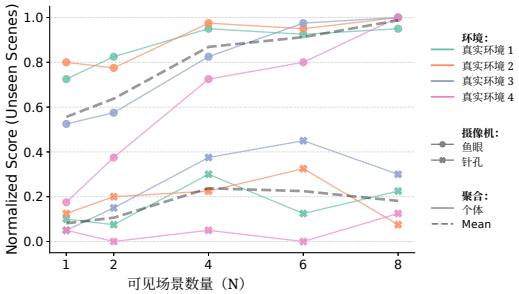


（b）真实世界场景分布（$K = 4$）

图 S5。**可视化泛化实验中的场景分布**。我们通过使用 K 均值对场景图像的 CLIP 嵌入进行聚类，来可视化场景分布。**红色星标**突出显示**留出测试场景**。



(a) **仿真任务（5 个场景）**



(b) **真实世界任务（4 个场景）**

图 S6. **各场景的泛化性能**。我们将每个个体留出测试场景的性能扩展曲线可视化，以考察泛化一致性。**虚线**表示汇总后的均值性能，实线则表示具体的测试环境。

• **仿真细分（5 条曲线）：** 我们报告了 5 个未见仿真测试场景中每个场景的成功率随训练多样性增加而变化的情况（如图 S6所示）（$N \in \{1, 8, 16, 32\}$）。

• **真实世界场景细分（4 条曲线）：** 我们报告了 4 个未见过的真实世界场景中各场景的归一化分数扩展情况（如图S6所示）——包括具有挑战性的“《星夜》”和“《向日葵》”图案——并展示其随训练多样性增加而发生的变化（$N \in \{1, 2, 4, 6, 8\}$）。

这些逐场景可视化旨在验证鱼眼相机的宽视场是否能带来普遍的泛化优势，使其在不同视觉分布下均保持稳健，而非仅适用于特定纹理类型。

E.3. 环境复杂度与编码器选择分析

为了解决环境变化对策略性能的影响，我们对场景代表性进行了定量分析，并论证了不同领域中的视觉编**场景代**

表性：通过将成功率映射到具体的特征密度，我们评估鱼眼相机的优势是否存在于典型的真实世界环境中，而不仅限于极端情况（表 S6）。**编码器选择：**特征密度为我们在不同领域使用不同视觉编码器提供了技术依据。它揭示了仿真与真实世界环境之间巨大的复杂度差距，因此有必要提升模型的表征能力——从稀疏仿真环境中的 ResNet-18 转向特征丰富的真实世界中的 CLIP——以有效处理捕获的上下文（表 S7）。

场景代表性与纹理敏感性：如表S6所示，我们在视觉纹理程度各异的真实环境中评估了这些策略。即使在“典型”设置下，例如特征密度处于中位数的标准木桌（2111.43 ± 341.87），鱼眼策略的表现也始终显著优于针孔基线（$+0.4250$）。虽然在高纹理环境中性能提升最为显著（$+0.8062$），但在低纹理场景中的稳健成功表明，我们的发现也适用于日常实际部署场景。

视觉编码器选择：视觉编码器的选择

表 S6. 不同真实世界环境中的性能。

典型测试场景 （真实世界变化）	**特征密度** （ORB[38] 点/帧）	**鱼眼评分** 针孔**鱼眼改进**	
无纹理（贫乏）	*1299.35 ± 176.81*（低）	0.1250 **0.5250**	*+0.4000*
木桌（典型）	*2111.43 ± 341.87*（中位数）	0.1250 **0.5500**	*+0.4250*
高纹理（丰富）	3574.99 ± 102.54（高）	0.1813 **0.9875**	*+0.8062*

（仿真使用 ResNet-18，真实世界使用 CLIP）源于不同领域之间视觉复杂度的巨大差异。如表S7所示，真实世界场景的特征密度几乎是仿真环境的 13 倍（3574。99 对比 268。77）。轻量级 ResNet-18 足以处理视觉信息稀疏的仿真数据，而鱼眼镜头在真实世界中捕获的高密度信息，则需要 CLIP 强大的表征能力。关键的是，消融实验结果表明，无论使用哪种编码器，鱼眼配置在两个领域中始终优于针孔基线，这证明广角视场角（FoV）的优势独立于具体的神经网络架构。

表 S7. 视觉编码器消融实验。

领域 **特征密度**（ORB[38]点/帧）	**仿真（成功率）** 268.77 ± 30.27		**真实世界（平均得分）** 3574.99 ± 102.54	
编码器	ResNet-18	**CLIP**	ResNet-18	**CLIP**
针孔	0.4467	0.5333	0.4250	0.7000
鱼眼	**0.7533**	**0.77**	**0.7000**	**0.8875**

F. RQ3（硬件泛化）的补充实验

本节分五个小节，全面分析跨相机泛化。首先介绍实验方案（第F.1节）。为揭示跨摄像头失效的根本原因，我们在仿真环境（第F.2节）和真实场景（第F.3节）中进行了尺度敏感性分析。基于这些分析所得的见解，我们展示了随机尺度增强在仿真环境中缓解尺度过拟合的效果（第F.4节）。最后，我们在多种实体摄像头上进行了大量零样本真实场景验证（第F.5节），证实了所提方法在实际应用中的稳健性。

F.1. 跨相机评估的仿真实验方案

摄像头建模。 为了严格评估零样本硬件泛化，我们设定了用于训练的“已见”摄像头配置，并准备了一组多样化的“未见”配置用于测试。在获取等距圆柱投影图像后，我们可以通过应用具有不同畸变参数的投影模型来模拟各种鱼眼效果。如表S8所示，我们采用配置了不同参数的鱼眼模型，模拟更换镜头时遇到的几何域偏移（例如，从广角鱼眼镜头切换到视角更窄的鱼眼镜头）。图S7 展示了使用不同参数进行模拟的可视化效果。

所用的具体模型如下：

- **扩展统一摄像机模型（EUCM）：** 该模型是统一摄像机模型（UCM）的扩展。通过引入两个参数，该模型提高了对大视场镜头的精度。其参数如下：**– f:** 焦距，用于控制图像的缩放级别。**– a:** 形状参数，取值范围为（0, 1], ，用于控制投影曲线的形状。**– b:** 畸变参数，用于调节非线性畸变的程度。

- **双球摄像机模型（DS）：** 该模型通过组合两个球体来描述光路，从而有效模拟大视场相机的投影几何。其参数为：

– f: 焦距，控制图像的缩放级别。**– a :** 混合参数，控制两个球体模型之间的混合比例。**– xi :** 畸变参数，表示两个球体中心之间的偏移。

表 S8. **RQ3 的仿真摄像头参数。** 我们定义了一种用于训练的已见配置，以及不同的配置用于零样本评估。焦距和畸变的变化模拟了显著的几何偏移。

配置名称	**方法**	**焦距**	**畸变**	**尺度**
已见参数	EUCM	45	a: 0.4 b: 2.0	0.9
参数 1	EUCM	60	a: 0.5 b: 2.0	1.0
参数 2	DS	50	a: 0.5 xi: 0.1	1.0
参数 3	EUCM	45	a: 0.4 b: 2。0	1.0
参数 4	EUCM	45	a: 0。4 b: 2。5	1.0
参数 5	EUCM	35	a: 0。4 b: 1.2	1.0

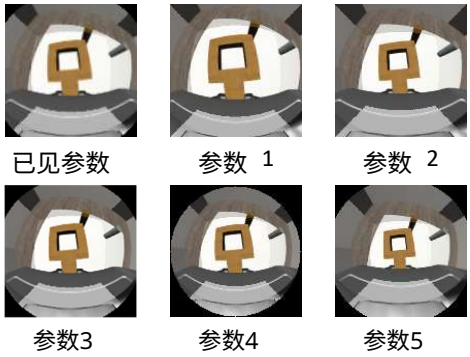


图 S7. **使用具有不同畸变参数的投影模型模拟鱼眼效果**。每个子图展示了应用一组不同参数后的视觉效果：“已见参数”代表一组基线参数，而“参数1”至“参数5”展示了畸变强度、视场及其他投影特性的变化，呈现不同参数配置如何改变鱼眼渲染效果。

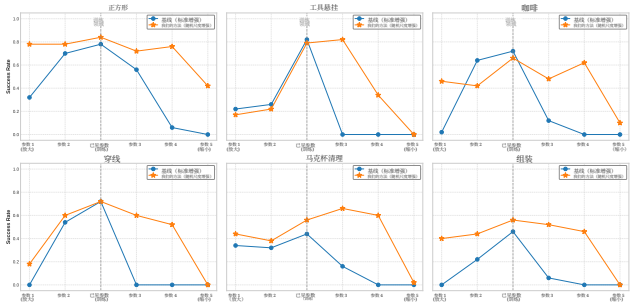


图 S8。**仿真中每个任务的跨相机泛化性能**。我们评估了在单一“已见参数”配置上训练的策略在多种未见相机设置下的零样本相机迁移性能。**基线（蓝色）**在相机参数偏离时表现出急剧的性能下降

远离训练域时，表明其对物体绝对尺度存在严重过拟合。相比之下，我们的 **RSA 方法（橙色）**展现出宽广的泛化平台期，在各种未见过的相机参数下都能保持稳健的性能。

F.2. 机制分析：尺度过拟合与尺度不变性

假设。 在正文中，我们假设跨相机迁移的主要失效模式是**“尺度过拟合。”** 由于不同的镜头内参会将物体投影为不同尺寸（例如，更宽的视场会使物体看起来更小），标准策略会过拟合训练集中物体的绝对像素尺度。当镜头发生变化时，这一绝对尺度先验便会失效，从而导致失败。

定量分析。为验证这一点，我们将仿真中全部六项任务的性能可视化，如图 S8所示。

• **基线失败（“峰形”曲线）：** 使用标准增强训练的策略（蓝线）仅在“已见参数”和“参数2”上表现良好，二者的视觉尺度相近。在引入显著尺度变化的配置（例如参数 4 和参数 5）下，性能急剧下降，这证实该策略高度依赖训练镜头提供的特定几何线索。

• **RSA 成功（“平台”曲线）：** 相比之下，使用我们的**随机尺度增强（RSA）**（橙色线）训练的策略在广泛的参数范围内保持了较高的成功率。RSA 有效地迫使网络学习**尺度不变特征**（例如相对空间关系），而不是记忆绝对像素大小。

F.3. 真实场景跨相机验证

为了在仿真中验证我们的假设，我们在真实机器人上进行了尺度敏感性分析。由于实际更换传感器模块以精确控制内参并不可行，我们通过应用不同的中心裁剪缩放因子来模拟几何域偏移

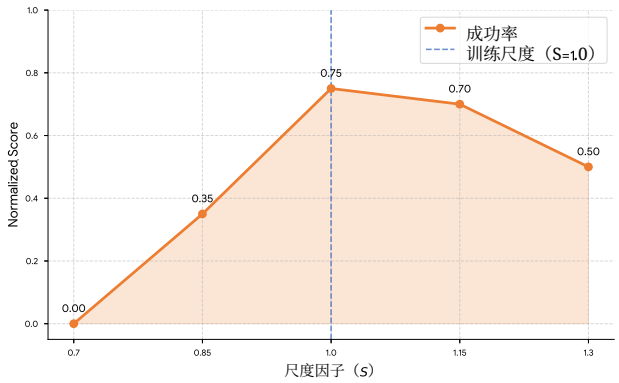


图 S9. **真实世界尺度敏感性分析**。为探究跨相机迁移失败的根本原因，我们对鱼眼输入应用不同尺度因子（S）的中心裁剪，以模拟几何域偏移。结果呈现出典型的**“倒 V 形”性能下降**：策略在接近训练尺度（$S = 1.0$）时表现稳健，但随着尺度显著偏离，性能会急剧恶化（例如，在 $S = 0.7$ 时成功率为 0%）。这证实了标准策略对物体绝对尺度高度敏感，也验证了采用我们的随机尺度数据增强（RSA）策略的必要性。

（S），在推理期间对输入图像进行处理，从而有效模拟焦距和视场的变化。这一操作有效模拟了改变镜头参数所产生的“放大”（视场更窄）和“缩小”（视场更广）效果。我们评估了策略在 $S \in [0.7, 1.3]$ 尺度范围内的表现，其中 $S = 1.0$ 表示训练分布。

定量分析：对尺度的敏感性。如图所示。S9 （尺度敏感性分析），基线策略呈现出一种典型的**“倒V形”性能曲线**。



图 S10. **尺度变化导致失效的定性可视化**。我们对几何尺度偏移（通过中心裁剪模拟）下基线策略的具体失效模式进行了可视化。**（左）**放大时（裁剪=0.7），物体看起来更大，误导策略将其判断为更近；因此，机器人尝试在更浅的深度处抓取（未达到目标）。**（右）**缩小时（裁剪=1.3），物体看起来更小，导致策略将其判断为更远；因此，机器人在更深的深度处抓取（超过目标或与目标发生碰撞）。**（中）**只有在训练尺度下（裁剪=1.0）才能准确操作。这些行为明确证实，跨摄像头失效源于策略对物体绝对像素尺度的过拟合。**更多详情请参阅补充文件中的视频**。

虽然策略在接近训练尺度（$S = 1.0$，得分 0.75）时仍能保持稳健的性能，但随着尺度偏离，该策略的性能会急剧下降。值得注意的是，“放大”操作（$S = 0.7$）会使成功率骤降至**0.0**，而“缩小”（$S = 1.3$）则会使性能降至**0.5**。这种大幅下降证实，标准鱼眼策略对物体的绝对尺度高度敏感，也印证了我们的仿真假设：尺度过拟合是跨相机迁移的主要瓶颈。

定性分析：深度—尺度歧义。为了解失效机制，我们对图 S10 中的具体轨迹执行行为进行了可视化。结果揭示了视觉尺度与深度估计误差之间的明显关联：

- **深度低估（放大）：**在 $S = 0.7$时，物体在图像帧中看起来明显更大。策略将这一视觉线索误解为物体比实际更近，因此尝试在**更浅的深度**处抓取（未能到达目标）。
- **深度高估（缩小）：**相反，在$S = 1.3$时，物体看起来更小。策略认为物体更远，因此会在**更深的深度**处抓取（通常会与目标碰撞或抓取过头）。

这些发现提供了实证证据：在缺乏尺度不变训练（例如 RSA）的情况下，基于鱼眼相机的策略在空间推理时严重依赖绝对像素大小，因此难以直接跨相机迁移。

F.4. 随机尺度增强的有效性

基于第F.3节所定义的尺度偏移协议，我们评估了随机尺度增强（RSA）在提升策略鲁棒性方面的有效性。我们的结果表明，RSA能够持续带来泛化能力提升，可作为抵御几何域偏移的手段；此类偏移在典型情况下会导致标准模仿学习策略失效。我们评估的详细结果列于表S9，其中量化了不同策略在变化的尺度因子下的性能。

•泛化能力持续提升：采用 RSA 训练的策略在所有评估尺度下，性能均稳步优于标准增强（Standard Augmentation）基线。例如，当尺度因子为 $S = 0.85$时，采用 RSA 训练的扩散策略得分为 0.950，而标准增强的得分为 0.350。

•减少尺度导致的性能下降：标准增强对尺度变化表现出敏感性，尤其是在放大情况下（$S = 0.70$），此时 Dif- fusion 策略的性能降至零。相比之下，RSA 在相同条件下仍保持 0.725 的得分，从而减轻物体放大导致的失败。

• 与大规模模型的集成：与 $\pi_{0.5}$ 架构集成后，RSA 有助于提升尺度不变性。值得注意的是，即使在显著缩小（S =1.30）的情况下，配备 RSA 的 $\pi_{0.5}$ 仍保持 1.000 的得分，而标准增强的得分则降至 0.150。

这些结果表明，RSA 促使视觉编码器优先关注相对空间关系——例如目标物体相对于夹爪的大小——而非绝对像素尺度。这种优先级有助于实现更稳健的跨硬件部署，因为它降低了镜头特定几何变化的影响。

表 S9。在模拟尺度变化下“抓取与放置”任务的归一化得分。我们比较不同参数（S）下的**标准增强**（标准增强）和**RSA**（随机尺度增强）。

策略模型	增强策略	$S = 0.70$ (放大)	参数 1	$S = 1.0$ (已见)	$S = 1.15$	$S = 1.30$ (缩小)
扩散策略 [4]	标准增强	0.000	0.350	0.750	0.750	0.500
	RSA（我们的方法）	**0.725**	**0.950**	**1.000**	**0.750**	**0.650**
$\pi_{0.5}$ [2]	标准增强	0.375	0.875	1.000	0.600	0.150
	RSA（我们的方法）	**0.900**	**1.000**	**1.000**	**0.975**	**1.000**

F.5. 零样本跨相机验证 n

为了评估 RSA 的实际效用，我们在真实环境中对不同的实体鱼眼镜头进行了零样本迁移实验。如表 S10所示，在标准 180° 镜头上训练的策略被直接部署到配备窄角（150°）和广角（220°）镜头的硬件上，由此产生了明显的几何和尺度变化。为确保研究结果能够代表现代通用策略，我们使用最先进的 $\pi_{0.5}$ 架构进行了这项验证。

• **应对硬件引起的尺度偏移：** 标准增强（基线）策略在遇到硬件变化时，性能明显下降。具体而言，切换到广角镜头（$\sim 0.8\times$尺度变化）时，基线得分降至 0.0025。
• **真实世界泛化：**相比之下，RSA 能够缓解不同物理镜头所引起的这些硬件偏移。对于窄角镜头（$\sim 1.2\times$ 尺度偏移），RSA 将分数从 0.5000 提升至 0.9500。即便在更具挑战性的广角镜头场景中，RSA 也将性能恢复至 0.6000。
• **实际部署意义：**这些结果表明，RSA 是一种超越基于仿真的启发式方法、适用于真实世界机器人学的有效策略。通过将策略与绝对像素尺度解耦，RSA 有助于在多样化的相机系统中复用现有数据集。

表 S10. 使用 RSA 的真实世界跨相机泛化。

测试相机	视场角	引起的尺度偏移	基线（标准增强）	RSA（我们的方法）
已见相机	180°	1.0×（已见）	1.0000	**1.0000**
窄角镜头	150°	∼1.2× (放大)	0.5000	**0.9500**
广角镜头	220°	∼0.8×(缩小)	0.0025	**0.6000**

总而言之，这些发现表明，RSA 能有效弥合不同硬件配置之间的差距。通过促进尺度不变特征的学习，RSA 为在各种机器人平台上部署基于视觉的策略提供了一条切实可行的途径，无需针对特定硬件进行微调。