arXiv:2303.03221v1 [cs.HC] 6 Mar 2023

Stargazer：一种基于细微的讲解者提示捕捉教程视频的交互式摄像机器人

Jiannan Li 多伦多大学 多伦多，加拿大

Bryan Wang 多伦多大学 多伦多，加拿大

Angela Yang 多伦多大学 多伦多，加拿大

Mauricio Sousa 多伦多大学 多伦多，加拿大

Paula Akemi Aoyaui 多伦多大学 多伦多，加拿大

Ravin Balakrishnan 多伦多大学 多伦多，加拿大

Tovi Grossman 多伦多大学 加拿大多伦多

Karthik Mahadevan 多伦多大学 多伦多，加拿大

Nicole Yu 多伦多大学 多伦多，加拿大

Anthony Tang 新加坡管理大学 新加坡

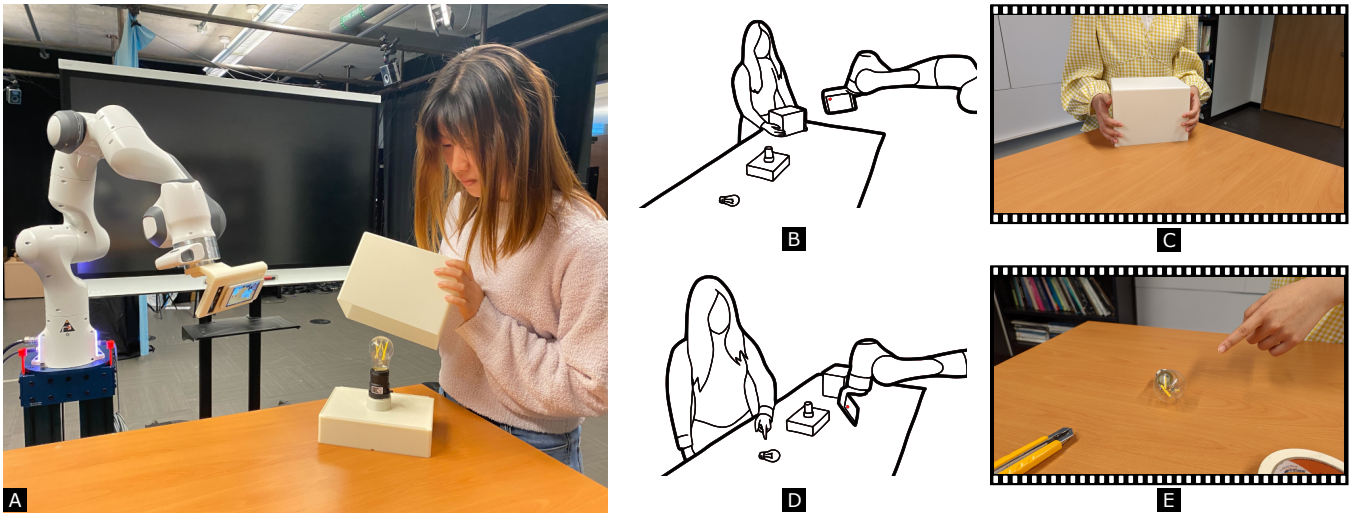

图1：（A）Stargazer 是一种利用摄像机器人创建教程视频的新方法。（B 和 C）Stargazer 的摄像机会自动跟随感兴趣区域，例如手持灯罩的讲师的手部。（D 和 E）Stargazer 还能够检测讲师的信号，例如指向手势，并通过聚焦于灯泡来响应。

摘要

实时和预先录制的视频教程是教授烹饪或电子产品原型制作等身体技能的有效方式。由专门的摄影师跟随讲师的活动可以提高制作质量。然而，不具备无法获得摄影师协助的人，往往只能在固定摄像机的限制下工作。我们介绍 Stargazer：一种创新方法，利用摄像机机器人辅助创作教程内容。该机器人会根据讲师动作自主跟踪感兴趣区域，以拍摄动态镜头。讲师可以通过手势和语音等细微提示调整 Stargazer 的摄像机行为，从而将摄像机控制指令流畅地融入教学活动。我们对六位讲师进行了用户研究，他们各自教授一项不同技能。研究结果显示，参与者能够使用 Stargazer 创作动态教程视频，呈现多种拍摄对象、摄像机取景和摄像机角度组合。

特此授权出于个人或课堂使用目的，在不收取费用的情况下制作本作品的全部或部分数字或纸质副本，前提是不得为营利或商业利益制作或分发副本，且副本必须保留本声明以及第一页上的完整引文。对于由非作者拥有的本作品组成部分，其版权必须得到尊重。注明出处后可进行摘要。除此之外的复制、再版、发布到服务器或向列表重新分发，均须事先获得具体许可和/或支付费用。请通过 permissions@acm.org 联系我们申请许可。*CHI '23，2023年4月23日至28日*，德国汉堡© 2023 版权所有归作者/权利所有者所有。出版权已授权给 ACM。ACM
ISBN 978-1-4503-9421-5/23/04...$15.00
https://doi.org/10.1145/3544548.3580896

CCS 概念
• **以人为中心的计算 → 交互系统与工具**；• **计算机系统组织 → 机器人学**。

关键词
摄像机、机器人、教学视频

ACM 参考格式：
Jiannan Li、Mauricio Sousa、Karthik Mahadevan、Bryan Wang、Paula Akemi Aoyaui、Nicole Yu、Angela Yang、Ravin Balakrishnan、Anthony Tang 和 Tovi Grossman。2023。《Stargazer：一种基于细微的讲解者提示拍摄教程视频的交互式摄像机器人》。载于《*2023*年*CHI*计算系统中的人为因素会议论文集》（*CHI '23*），*2023*年*4*月*23*日至*28*日，德国汉堡。ACM，纽约，美国，16页。https://doi.org/10.1145/3544548.3580896

1 引言
人们会通过现场或预先录制的教程视频学习新的身体技能 [37], ，从修理损坏的键盘到学习数字制造 [42]。在大多数教授身体技能的教程视频中，讲师会逐步演示如何完成任务 [14]。由于这些步骤可能涉及在不同地点开展的活动，且所需的细节程度各异，单个固定摄像机往往无法以足够清晰的画面记录每个步骤 [39]。因此，需要频繁调整摄像机参数，包括视点、角度和缩放级别。

专业视频制作（如烹饪和家居装修节目）会配备多名专职摄像机操作员，他们会根据指导者的动作主动重新调整摄像机的位置和参数。然而，大多数讲师无法获得这类资源，因此只能依赖一台或多台预先配置好的固定摄像机。虽然可以在录制过程中重新配置固定摄像机，但讲师需要停下演示（例如切菜）来操作摄像机。这会打断演示，增加讲师的工作量，也需要进行更多后期处理，以合并使用不同摄像机配置拍摄的片段。

电影制作人和研究人员都曾探索以操控摄像机的机器人取代人工操作员的构想 [1, 27]。近期的摄像机器人（主要是无人机）能够自主追踪移动主体 [6, 24, 41]。然而，对于正在被拍摄的用户而言，一边演示某个实体操作过程等活动，一边控制摄像机器人的行为仍是一项挑战。传统的机器人控制界面采用操纵杆 [52], 、手势 [49], 和语音 [16]，而这些方式都需要专门的输入动作，会打断教学实施。如果不将这些中断剪辑掉，它们可能会分散观众的注意力，妨碍学习 [9], ；但后期处理会增加讲师的工作量。近期的用户界面研究探索了通过演讲者在演示过程中使用的手势和语音来触发屏幕视觉效果 [22, 32, 48] 。本研究采用的方法是：(1) 确定教程视频使用的镜头类型；(2) 利用讲师在演示过程中本就用于面向观众 进行讲解的交际信号，以不造成干扰的方式控制摄像机操作。例如，讲师可能会指向物体的某个部分以突出强调，用语音引导观众的注意力，或挥手作自我介绍。

我们介绍 Stargazer，这是一种新颖的方法，利用摄像机器人拍摄桌面规模实体任务的动态教程视频（图 1A）。讲师可以通过细微信号控制摄像机的高层行为，例如当前的感兴趣区域和摄像机取景，并实现实时控制。Stargazer 的摄像机可自主追踪并拍摄感兴趣区域（图 1B-E）。请注意，我们的目标并非打造一台能够理解人类交际行为各种细微之处的全自动机器人；相反，我们旨在探索一套控制指令，使其能够支持制作信息丰富的教程视频所需的摄像工作，并自然融入常规教学活动中。

通过分析三个热门类别下的 50 个教程视频，我们确定了机器人应支持的三种常见镜头类型和三种摄像机参数控制方式。讲师通过手势和语音来控制镜头类型和摄像机参数，而这些手势和语音也是他们引导观众注意力时常用的方式。Stargazer 能够定位每种镜头类型中的关注拍摄对象，并平稳地跟踪他们。

我们邀请了六名参与者，其中包括两名专业电影制作人，使用 Stargazer 为其擅长的实体任务拍摄教程视频。所有参与者都能够在无需任何专用控制装置的情况下，仅凭 Stargazer 提供的线索来制作视频，并且对所制作视频的质量感到满意。根据参与者对所制作视频发表的出声思维评论以及研究后访谈数据，我们确定了使用机器人进行协作式视频拍摄的机遇与挑战。

2 相关工作
我们的研究建立在教程视频、摄像机器人以及用于控制机器人的人类线索等既有工作的基础上。我们的交互设计受近期表演驱动的交互式演示领域进展的启发。下面我们将回顾相关工作。

2.1 教程视频
教程视频在在线视频和直播网站上十分常见，为学习实体任务提供了便利的方式 [14, 17, 37] 。人机交互（HCI）、认知科学和教育技术领域的研究探讨了教程视频的各种属性，例如视角 [26], 、节奏 [58], 以及讲师身份 [23], ，可能会如何影响学习。用户界面研究尤其考察了用于浏览 [10], 、导航 [11, 61], 、标注 [29], 和编辑 [12] 教程视频的新型界面。我们的工作建立在这些关于教程视频的既有研究之上，但旨在促进此类视频的拍摄，而不是其后期处理或观看。

2.2 交互式演示
研究人员探索了演示者的表演（包括语音、 [32], 手势 [22, 48], 和绘制草图 [43, 55],）如何驱动实时屏幕图形，从而使现场或录制的演示更具信息量和吸引力。这些交互式演示界面尤其精妙的一点在于，它们允许演示者将图形效果控制融入内容呈现中。RealityTalk [32] 利用口头演示中的关键词来触发视觉增强效果。Saquib 等人 [48] 使演示者能够通过手势直接操控图形元素和效果

以及姿势的变化。Hall等人 [22] 也研究了手势的使用，重点关注对数据可视化的修改。我们与 Stargazer 的用户交互设计，受到了先前研究的启发：这些研究通过利用手势和语音，将效果控制流畅地融入演示者表演中。Stargazer 不同于控制由固定摄像机捕捉的数字制品，而是专注于通过讲师动作控制动态摄像机。

2.3 人机交互中的人类线索

先前研究探讨了人类团队在协作物理任务中使用的协调机制。这些任务需要共同注意——引导合作伙伴注意某个物体或感兴趣的区域。人机交互研究人员提出了多种促进共同注意的方法。一种直接的方法是让用户标注机器人的传感器输入（例如摄像头流） [51, 60], , 但这通常需要额外的设备。跟踪人类的注视可以实现预测性机器人规划和动作执行 [2, 4, 5, 25, 40],, 也可以在检测到用户犹豫时提示机器人接管 [47]。手势是人类和机器人引导注意力的另一种方式 [49], , 包括在协作完成物理任务时进行指向、展示和展示物体 [20]。人类甚至可以通过语言引导机器人的注意力 [56],, 这种方式已用于免手操作机器人编程 [16] , 以及在抓取放置任务中指挥机器人 [36]。我们注意到，先前研究通常使用显式通信机制（即指向、指令等）或更隐式的通信机制（即眼神）来操控机器人。Stargazer 的方法同时采用显式和隐式交流行为。不过，我们将这些行为视为演讲者与观众 之间表演性对话的一部分，而非演讲者与机器人之间的对话。

2.4 摄像机器人

远程协助和远程临场领域的研究大多关注由远程用户通过图形用户界面手动操作的摄像机器人（例如， [21, 31, 57, 59]）。近期的远程协助摄像机器人，例如 RemoteCoDe [46], , 能够根据远程协助者的面部朝向进行隐式控制。另一类研究探索了能够对被拍摄者发出的信号做出反应的摄像机器人。自主摄像无人机可以跟踪拍摄对象，并根据拍摄对象的位置 [41], 、姿势 [24], 和动作 [6], 调整飞行计划，这些调整通常依据电影摄影原则 [27] , 或基于学习到的偏好 [6]。Rakita等人 [44] 则设计了一种方法，让机械臂动态拍摄另一只正在接受遥操作的机械臂，以辅助遥操作；在此研究中，拍摄对象并非人类。尽管这些机器人能够响应人类信号，但它们通常不允许被拍摄者主动发起显著的机器人行为变化，例如更换拍摄对象或调整摄像机取景。Stargazer 旨在弥补这一不足，为讲师提供更精细的实时摄像机行为控制，以拍摄教程视频。

3 教程视频中的摄像机使用特征

Stargazer 的行为与交互设计参考了对现有教程视频内容及其视觉语言的研究。我们研究了来自三个类别的 50 个身体技能指导视频，旨在了解：(1) 这些视频中常见的镜头类型有哪些？(2) 创作者在教程视频中如何运用取景、角度和运动？

3.1 视频语料库

我们的视频语料库包含 50 部教学视频，其中 15 部涉及计算机与电子产品，18 部涉及食品与娱乐，17 部涉及爱好与手工艺。视频平均时长为 5 分 22 秒。截至 2022 年 7 月，视频平均观看次数为 2,176,470 次（最少：20,824 次；最多：37,067,768 次）。我们基于 WikiHow [1] 上的实体任务构建了语料库。WikiHow 是一个知名数据库，收录了数千篇用户撰写的教程文章，涵盖各种日常任务，构建方法参考了先前研究中使用的策略 [38]。

为确定桌面尺度实体任务的类别，我们参考了网站上的“热门类别”，并发现三个类别：“计算机与电子产品”、“爱好与手工艺”以及“食品与娱乐”。随后，我们从每个类别中随机抽取文章，收集了 50 篇教程文章，并使用文章标题在 YouTube[2] 上搜索视频，搜索时使用隐身模式。接着，我们依据以下标准筛选视频：1）回应从 WikiHow 收集的教程问题；2）以英语录制；3）时长不超过 20 分钟；4）累计观看次数超过两万次。我们将观看次数作为代理指标，以确保抽样视频具有合理的质量。近期一项针对教学视频和信息类视频观众的调查发现，受访者中偏好观看时长超过 20 分钟视频的人不到 10% [30]。因此，我们将研究重点放在时长不超过 20 分钟的视频上，以便团队在资源限制下分析足够多样化的数据集。

3.2 分析

我们以镜头层级对抽样视频进行编码（共 884 个镜头，平均每个视频 17.68 个镜头），重点关注每个镜头所捕捉的内容，以及这些内容如何通过独特的视觉语言（即角度和取景）进行传达。我们采用自下而上的方式制定了一个包含两个主要类别的编码方案：镜头主体和摄影机设置（即角度、取景和运动）。镜头主体涵盖画面中展示的内容（例如主持人、物体）。我们的摄影机设置编码方案参考了常见的电影摄影实践 [7]。该类别涉及视觉语言，包括摄像机角度（例如高角度、过肩镜头）、摄像机取景（例如中景、特写）和摄像机运动（例如平移镜头、横向移动、环绕运动）。两名研究人员通过共同分析五个视频来制定并测试编码方案，随后分别处理其余视频。

3.3 镜头中的拍摄对象

我们观察到，教程视频镜头主要聚焦于少数几类拍摄对象——讲师、讲师的动作，以及用于演示任务的物体。讲师的动作主要表现为用手操作工具和其他人工制品。教程视频镜头通常同时呈现讲师的手部和被操作的物体（手部+物体，占视频总时长的 53.51%，见图2 A 和 D）。第二类

[1] https://www.wikihow.com/Main-Page [2] https://www.youtube.com/

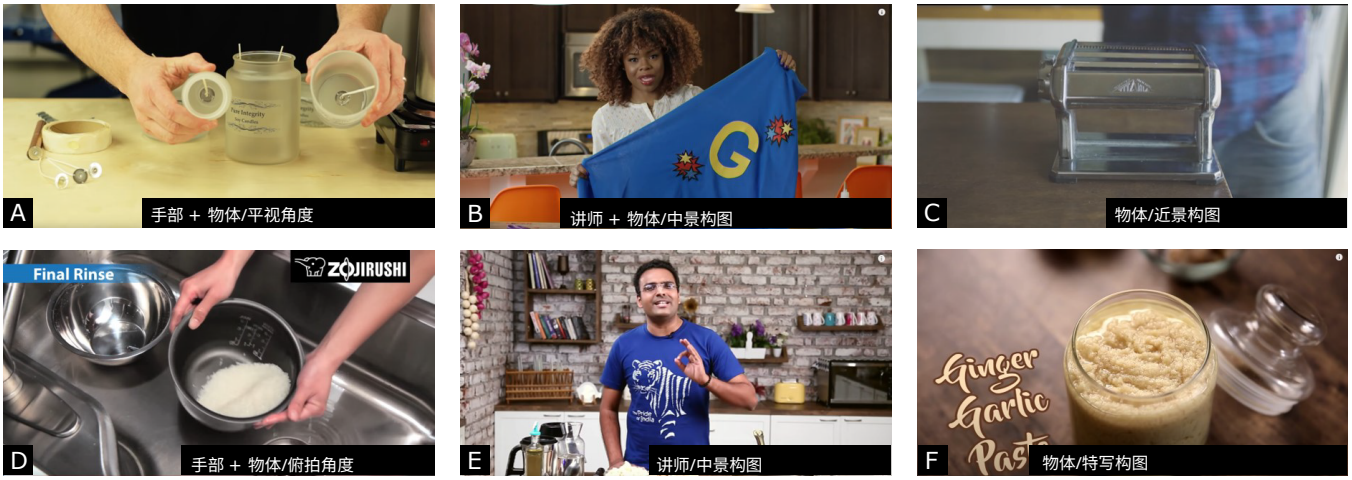

图2：视频分析中的示例帧。（A）以平视角度拍摄手部+物体的镜头（B）采用中景构图拍摄讲师+物体的镜头（C）采用特写构图拍摄物体的镜头（D）以高机位角度拍摄手部+物体的镜头（E）采用中景构图拍摄讲师的镜头（F）采用大特写构图拍摄物体的镜头

镜头则更多聚焦于讲师向观众进行口头"讲解"，因此主要呈现讲师讲话并向观众展示物体的画面（讲师，占视频总时长的10.18%；以及讲师+物体，占视频总时长的 11.88%；见图2 B 和 E）。最后，在分析的视频总时长中，有 17.65% 包含仅呈现物体的镜头（物体；见图2 C 和 F）。这些镜头通常展示任务中使用的材料或工具，或任务完成后的成果。

3.4 视觉语言

尽管电影制作中的视觉语言涵盖诸多元素，我们的分析聚焦于与摄像机定位相关的要素，即摄像机取景、角度和运动。

3.4.1 摄像机取景。我们发现，展示每种类型拍摄对象的镜头主要采用两种取景方式。展示手部+物体或单独展示物体的镜头主要采用特写构图（手部+物体占79.92%，单独展示物体占64.74%，见图2A、C、D），其次是大特写构图（手部+物体占15.86%，单独展示物体占28.21%，见图2F）。聚焦于讲师的镜头更多采用中景构图（讲师占66.67%，讲师+物体占87.62%，见图2B和E），而非特写构图（讲师占33.33%，讲师+物体占10.48%）。

3.4.2 摄像机角度。在标注摄像机角度时，我们采用了常见的摄像机角度类型，并根据教程视频的情境进行了调整。遵循电影摄影惯例，我们将向下拍摄拍摄对象（手部／讲师／物体）的镜头标记为俯拍角度，将向上拍摄拍摄对象的镜头标记为仰拍角度。俯拍镜头也归入俯拍角度类别。与拍摄对象处于同一高度或略高于拍摄对象的镜头，则标记为平视角度。我们发现，大多数聚焦于讲师的镜头都采用平视角度（讲师为98.89%，讲师+物体为82.86%；图2B和E）。相比之下，聚焦于手部+物体和物体的镜头中，俯拍角度（22.42%）、主观视角（34.02%）和平视角度（29.57%）都很常见。

图2A和D展示了聚焦于手部+物体的平视角度和俯拍镜头示例。

3.4.3 摄像机运动。我们抽样的视频中，大多数镜头（70.59%）都是用固定摄像机拍摄的。约15%的镜头存在抖动、突兀的摄像机运动，这对视觉质量的损害最大。只有少数镜头（8.13%）采用了稳定的摄像机运动，例如平移镜头（绕摄像机自身的垂直轴旋转）、横向移动（沿与镜头方向垂直的水平轴移动摄像机）、环绕运动（围绕拍摄对象沿弧线移动）和推轨运动（朝拍摄对象靠近或远离）[7]。这可能是因为大多数讲师无法聘请专职摄影师。

3.4.4 总结。我们对教程视频中摄像机使用情况的探索，让我们了解了 Stargazer 需要支持的镜头类型范围——镜头聚焦的对象（拍摄对象），以及制作人员需要调整或使用的各项参数。接下来，我们将介绍 Stargazer 的交互设计，该设计采用了这套镜头类型术语，并讨论 Stargazer 如何让制作人员在拍摄过程中无需中断讲解或预先设定摄像机路径，即可在不同镜头类型之间切换。

4 STARGAZER 交互设计

Stargazer 可帮助指导者使用安装在高度灵活的七自由度机械臂上的单摄像机，制作"一镜到底"的教程视频。系统会跟踪拍摄对象，并根据指导者的动作和语音动态调整摄像机取景和角度。例如，在教程过程中，指导者可以让 Stargazer 依次将镜头对准要用到的每件工具——逐一指向这些工具，让 Stargazer 平移镜头展示每一件。最后，指导者可以对观众说："现在，如果你从上方看我如何将 A 放入 B 中"，Stargazer 就会

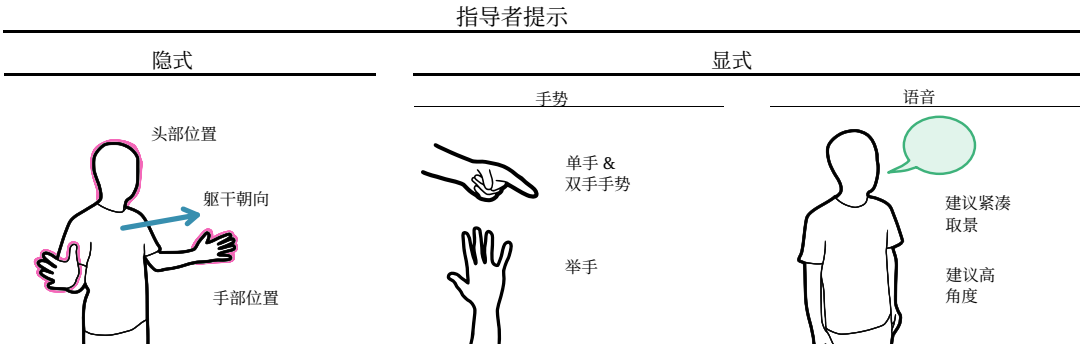

图 3：影响 Stargazer 行为的讲师提示。

以高角度取景，将指导者的动作和物品一并纳入画面，让观众看得更清楚。

我们设计交互词汇的目标，是找出一些流畅、含蓄的信号，而且这些信号可以自然地融入讲师与视频观众之间的对话或互动中。我们的设计避免了讲师需要脱离观众、单独向摄像机器人传达信息；相反，这些信号巧妙地融入了教程演示中。

4.1 能力与需求

我们对教程视频的分析表明，讲师需要多种视频镜头和能力。这些构成了 Stargazer 功能的基础。我们根据第 3 节提出的交互词汇和编码方案对这些内容进行归类。

- 拍摄对象：摄像机需要能够拍摄教程视频中常见的拍摄对象。

  指导者镜头主要以讲师为画面主体，让其直接面向观众讲话——有时也会在展示物体的同时进行讲解。（第 3 节中的讲师和讲师 + 物体。）

  动作镜头拍摄讲师的手部在演示物理过程时操作各种工具和物体。（第 3 节中的手部 + 物体。）**物体镜头**聚焦于讲师指向的一个或多个物体（第 3 节中的物体。）

- 跟随：由于拍摄对象可能会移动，机器人应自动定位并跟随拍摄对象，在当前镜头类型和距离的限制下保持这一画面。
- 摄像机取景/角度：指导者应能够在普通取景和紧凑（放大）取景之间进行选择和切换，也应能够在平视角度和高角度之间进行选择和切换。
- 摄像机运动：指导者还应能够发起常见的电影摄影机运动（例如环绕运动、横向移动）。
- 反馈：与其他自主运行的智能体和人工制品一样，[28, 52], 机器人需要向讲师反馈其当前状态。

4.2 交互词汇

Stargazer 会响应经过挑选、力求不显眼的指导者信号（图 3）。我们的目标是定义一套信号词汇，使其理想情况下能够融入内容呈现的自然流程，既不突兀，也不打扰观众和讲师。绝大多数时候，人们认为 Stargazer 正以合理的方式自主运行（即以给定的取景跟踪并跟随拍摄对象），并在不同镜头间平滑切换，以提升画面的清晰度。默认情况下，摄像机行为会跟随讲师的隐式身体动作以及物体被操作时的移动。摄像机位置和朝向的变化由指导者发出的信号驱动——机器人持续监测讲师的手部、头部和身体姿势，以寻找这些信号。讲师面向观众做出的手势和语音信号，会向 Stargazer 发出改变高层摄像机行为的指令，包括镜头拍摄对象、摄像机取景和摄像机角度。我们的设计选择了能够在信号不显眼和误触发可能性之间取得平衡的提示。

4.2.1 身体动作提示：手部。 绝大多数教程视频镜头拍摄的都是讲师前方的球形区域，即讲师在此处演示或操作物体；这是 Stargazer 的默认状态。对于此类动作镜头，Stargazer 会定位讲师的双手，并尝试：(1) 将双手置于摄像机画面中央附近；(2) 确定合适的距离，以便在适应近期手部动作的情况下拍摄双手（图 4 A-D）。Stargazer 会跟踪双手的中点，并建立一个球形区域，用于表示过去 5 秒内的手部动作范围。这可以应对双手之间的距离快速变化的情况——例如，讲师需要选取并使用放置在所操作物体旁边的工具。摄像机的摆放方式应使其对准双手的中点，并保持一定距离，使该球形区域覆盖摄像机画面宽度的三分之一。

4.2.2 身体动作提示：头部。 在讲师镜头中，讲师的面部在画面中水平居中，摄像机与讲师的躯干垂直（图 4 E 和 F）。根据电影摄影原则中的三分法构图 [7],Stargazer 定位摄像机，使讲师的眼睛位于摄像机画面自上而下三分之一处，并努力保持这一取景和朝向。

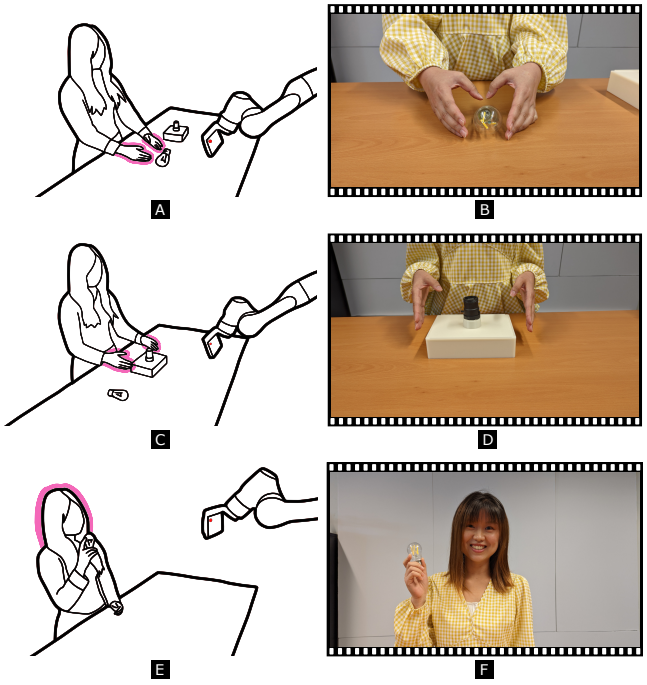

图 4：Stargazer 会对哪些身体运动线索做出反应。（A 和 C）在动作镜头中，Stargazer 会跟随讲师的手部，并调整与手部的距离。（B 和 D）展示摄像机在 A 和 C 中拍摄到的画面。（E 和 F）在指导者镜头中，Stargazer 会跟随讲师的头部位置。

4.2.3 身体动作提示：躯干朝向。 在讲师镜头和动作镜头中，Stargazer 将摄像机垂直对准连接讲师两肩的线段，以提高主体可见性。当讲师的躯干朝向发生变化时，摄像机器人会重新调整自身朝向（图 5）。

4.2.4 手势线索：单手指向。 讲师可以用食指指向某个物体，从而引导观众的注意力——这是一种在日常生活 [19] 和教学场景中都很常见的交流行为 [35]。Stargazer 通过跟踪讲师指向方向上、靠近指尖的位置，将该动作解读为切换至物体镜头（图6）。在讲师指向时，Stargazer 会继续跟随讲师的手部（例如，这使得他们能够沿着一条轨迹进行指向，

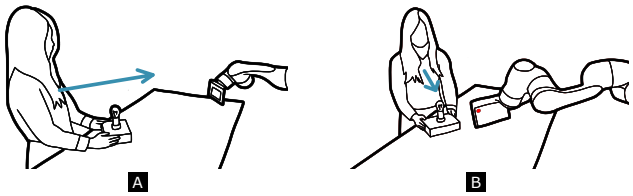

图 5：在指导者镜头和动作镜头中，Stargazer 会调整摄像机的朝向，使其垂直于连接讲师双肩的直线。

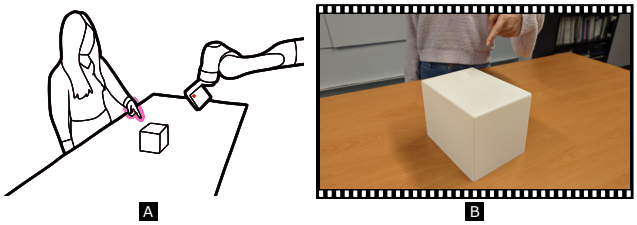

图 6：Stargazer 通过跟随讲师指向的手，切换至物体镜头。

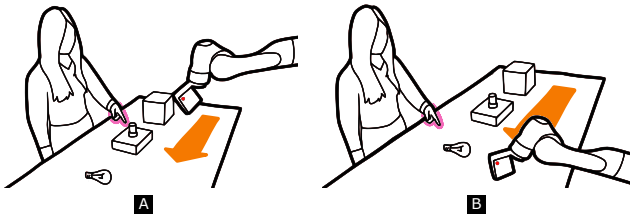

图 7：讲师可以沿水平线移动手指，触发摄像机的“横向移动”效果（即摄像机沿垂直于镜头方向的轴水平移动）。

）。一旦讲师停止指向，Stargazer 就会恢复捕捉动作镜头，即跟随手部。

讲师还可以在做出指向手势的同时，让食指沿水平线快速移动。Stargazer 会将此解读为执行线性跟踪效果（“横向移动”）的信号。其摄像机在与手指移动轨迹垂直的方向上跟随手指的位置（图7）。

4.2.5 手势线索：举手。 举手并挥手是常见的问候手势，也可用于引导注意力 [33],，例如开启对话。Stargazer 会将这一挥手信号解读为切换到讲师镜头的指令（图8）。具体而言，Stargazer 能识别举过肩膀并保持短暂时间（1秒）的手。

4.2.6 手势线索：双手指向。 讲师可以通过双手各伸出一根手指，做出仿佛在向观众展示某个物体的动作（图 9），从而启动摄像机环绕运动。Stargazer 将这一手势解读为围绕两根手指指尖

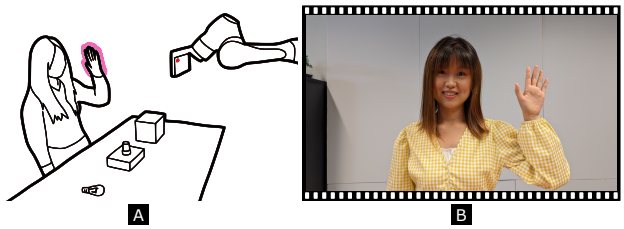

图8：讲师举手，示意 Stargazer 切换到讲师镜头。

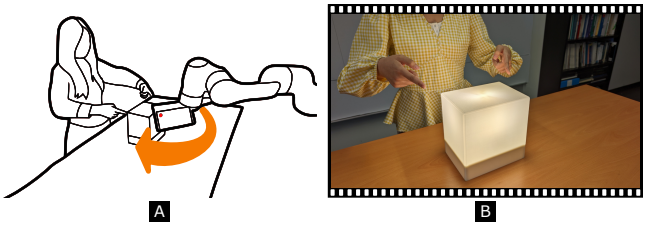

图 9：讲师做出双手指向手势，使摄像机围绕两根指尖的中点进行环绕运动。

中点环绕的信号，同时在摄像机沿弧线移动时提供扩展镜头。

4.2.7 语音提示：通过语言建议采用紧凑取景和俯拍角度。在面对面的教学环境中，讲师可能会建议学习者靠近一些（例如："如果你仔细看看……"），或从不同的视角进行观察（例如："你可以从上面看到……"）。这类言语行为对于运动技能学习十分必要 [13]，，因为它们会引导观众注意特定内容。当对话似乎是在向观众提出这类建议时，Stargazer 会将讲师预先定义的特定语音信号解读为关于摄像机取景和角度的建议。我们实现了两种由语音触发的具体行为。

Stargazer 利用先进的大语言模型推断讲师语音中的意图，而无需强制使用严格的模板。讲师可以通过说出建议仔细观察的句子来控制放大。例如："如果你仔细看看，就能看到这个插座是六边形的"，或"更加注意我是如何取下盖子的。"（图 10A 和 B）。同样，讲师可以通过建议从更高位置观察的句子来控制摄像机角度（高角度镜头与标准角度镜头）。例如："最好从上面看，这样才能看到我是如何取下头带的"，或"现在我希望你采用俯视视角，这样就能看到模型的全貌。"（图 10C 和 D）。若要将摄像机角度重置为平视角度，或将取景恢复为正常状态，讲师会将一只手藏到身体后面。

Stargazer 以这种方式使用语音，让讲师的双手得以解放；在教程视频中，讲师的双手通常都忙着操作。这与我们使用手势来触发涉及拍摄对象切换的镜头切换截然不同。这类切换通常发生在一个完整步骤开始前或结束后，此时讲师可以选择腾出双手。

4.3 向讲师提供反馈

先前研究表明，向与自主代理交互的人们提供有关其状态的反馈有诸多益处 [28, 52]。我们的早期试点测试也表明，讲师需要知道摄像机"看到"什么，并能够理解 Stargazer 的状态以及即将采取的行动。Stargazer通过安装在相机支架上、面向讲师的显示器提供这些反馈（图11）。显示器会显示实时相机预览，并额外显示当前镜头类型的图标，以及两个指示器，用于显示是否检测到左手和右手的指向手势。

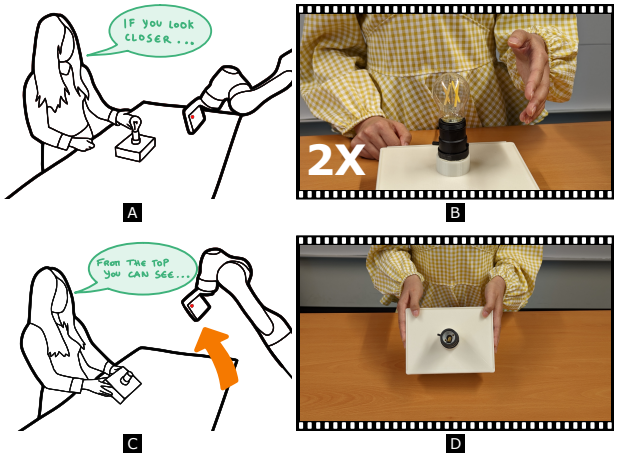

图 10：讲师可以（A 和 B）通过语音建议摄像机拉近查看，从而将画面放大 2 倍；也可以（C 和 D）通过语音建议从更高的位置拍摄，以获得高角度镜头。

5 实现

Stargazer原型由三个主要部分组成：机械臂、摄像机和传感器（图12）。机械臂是一台七自由度的 Franka Emika Panda，由运行在 Linux工作站（机器人工作站）上的一组机器人操作系统（ROS）节点控制。传感器包括一台 Kinect v2 深度摄像头、一台 RGB 网络摄像头和一支无线麦克风，用于感知讲师的骨骼、指向手势和语音。所有传感器都连接到一台 Windows工作站（传感器工作站），该工作站会处理传感器数据，并通过有线局域网，使用用户数据报协议将数据发送到机器人工作站。摄像机和摄像机监视器都通过定制支架安装在机器人上。我们原型的软件和硬件设计已在网上开源[3]。

5.1 摄像机硬件与软件

我们使用 Android手机（Google Pixel 6 Pro）的后置摄像头作为 Stargazer 的摄像机。该手机运行一个定制软件应用程序，用于录制视频，并通过无线UDP连接与机器人工作站通信。该应用程序显示手机相机视频流的预览，以及多个显示当前镜头类型和手势检测结果的小部件，如第4.3节所述。机器人工作站向该应用程序发送消息，以控制摄像机的变焦级别并更新状态小部件。摄像机监视器将手机屏幕镜像显示出来，为讲师提供反馈（更多详情见第5.1.1节）。手机插入一个3D打印支架（图11）中，该支架连接到机械臂上，作为定制的末端执行器。

5.1.1 摄像机监视器。 我们将一台iPod Touch安装在摄像机支架朝向讲师的一侧，作为摄像机监视器。它显示手机摄像机视频流的预览和机器人状态，并通过本地无线网络镜像手机屏幕。我们

[3]https://github.com/jchrisli/stargazer-chi23

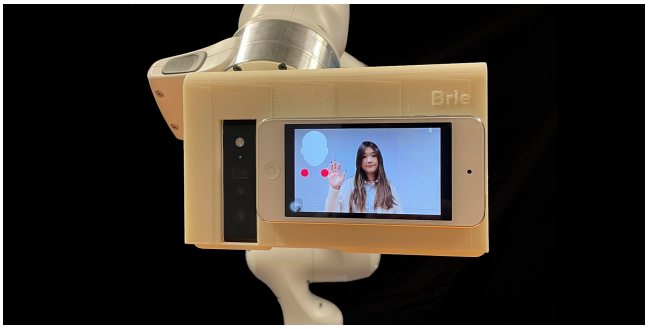

图 11：摄像机监视器向讲师提供反馈，显示摄像机预览、当前镜头类型，以及是否检测到指向手势。在这张照片中，监视器显示了一个头部图标，表示当前镜头类型为讲师镜头；两个红点则表示双手都没有做出指向动作。

使用传感器工作站作为屏幕镜像的中继点。传感器工作站首先通过 scrcpy[4] 获取手机的屏幕投放，然后通过 Moonlight[5]将屏幕投放传输到 iPod Touch。

5.2 感知指导者的动作

Stargazer原型结合多种传感器来检测讲师的身体姿态、手势和语音。未来的工作可以探索利用机器人自身摄像机的感知能力。

*5.2.1 身体姿态。*我们将 Kinect v2 深度摄像头对准讲师，并使用 Windows 版 Kinect SDK 推断讲师的身体姿态。Stargazer 使用头部关节作为面部的代理，使用指尖关节表示手部，并将腕关节到相应指尖关节的向量作为指向方向。在物体镜头中，我们追踪沿指向方向、距指尖 12 厘米的位置。为了估算包含双手的球形区域，我们对最近 5 秒内双手之间的距离计算滚动平均值。我们使用 Creepy Tracker工具包 [54] 对数据进行过滤和流式传输。

5.2.2 手部手势。 我们将一台 720p RGB 网络摄像头对准工作台中央，以检测讲师的指向手势。我们使用深度学习检测器 [34] 定位讲师双手的二维关键点位置，并通过自定义模型判断手部是否正在执行指向手势（食指向前伸展，而中指、无名指和小指蜷曲）。每只手的 21 个关键点位置在分类前都已进行归一化处理。手势检测模型是一个使用 Keras 实现的四层全连接神经网络，使用从研究团队收集的数据进行训练。为减少分类歧义，只有当模型的置信度至少达到 85% 时，我们才将当前手势分类为指向手势。我们还采用低通滤波来消除零星的误报。

[4]https://github.com/Genymobile/scrcpy [5] https://moonlight-stream.org/

5.2.3 语音理解。 我们使用夹在讲师身上的无线麦克风录制讲师的语音，并将其发送至语音识别 API（Microsoft Azure 语音转文本[6]）。转录文本连同自定义提示词7随后被发送至大语言模型（GPT-3 [8]），由其标注讲师的意图（即更紧的取景/高角度/正常）。

5.3 机器人控制

Stargazer原型使用 ROS 与机械臂通信并控制其运动。机器人工作站运行一个 ROS 节点，根据传感器输入计算下一个目标相机位姿。我们在第 5.4 节中介绍规划下一个相机位姿的算法。给定目标相机位姿后，我们使用 PID 控制器获取所需的笛卡尔线速度和角速度。实时伺服库 MoveIt Servo 将笛卡尔速度转换为关节速度。最后，我们通过 Franka ROS 接口设置关节速度[8]。下一个相机位姿以 5 Hz 的频率计算。PID 控制回路以 100 Hz 的频率运行。

我们监测机器人的状态，以确定其是否接近运动学奇异点或关节极限，因为这些情况会导致机器人无法正常移动。如果是这样，控制程序会暂停实时速度控制，并通过将机器人返回安全中性位置，使其从错误状态中恢复。

5.4 机器人运动规划

虽然相机位姿有六个自由度，但 Stargazer 的摄像机始终瞄准已知位置（例如讲师的头部），并保持像平面的水平轴与地面平行。这使我们的运动规划问题从 $\mathbb{R}^6$ 简化为 $\mathbb{R}^3$。我们的目标是在时间步 t，在运动平滑性、与拍摄对象的距离等若干约束条件下，为位于位置 $\mathbf{x_s} \in \mathbb{R}^3$ 的拍摄对象（例如面部、双手）规划下一个目标相机位置 $\mathbf{x_{t+1}} \in \mathbb{R}^3$。除规划环绕运动的轨道路径外，我们将下一个位置的选择表述为一个优化问题，沿用自主无人机电影摄影领域的既有研究 [41]。如第 5.4.2 节所述，我们将轨道路径直接计算为一系列航路点。

*5.4.1 运动规划代价函数。*代价函数 J 由以下代价值的加权和构成。

运动平滑性：为了实现平滑的摄影机运动，我们通过代价 $c_{ms} = (\mathbf{x_{t+1}} - \mathbf{x_t})^2$ 对相机位置的大幅位移进行惩罚。

期望距离：我们首先根据当前镜头类型，计算拍摄对象的期望拍摄距离 d。对于动作镜头，拍摄对象应占摄像机画面宽度的三分之一。因此，对于半径为 r、摄像机视场角为 α的拍摄对象，期望距离为 $d = 2r\tan(\alpha/2)/3$。对于讲师镜头和物体镜头，我们通过经验分别将 d设为 0.20m 和 0.12m。给定 d，期望距离代价为 $c_{dd} = (\|\mathbf{x_{t+1}} - \mathbf{x_s}\| - d)^2$。对于俯拍镜头，我们不施加距离代价，因为这往往会导致机器人进入运动学奇异点。

俯仰角：对于讲师镜头和采用高角度设置的镜头，我们分别建议将相机俯仰角设为接近 $\pi/2$ 和 0。

[6]https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/ [7] 我们附上了在补充材料中使用的提示词。 [8] https://frankaemika.github.io/docs/index.html

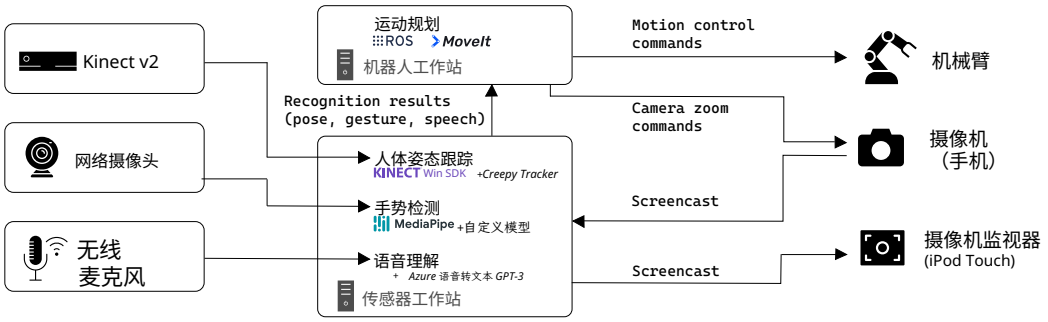

图12： Stargaze r 原型的组件概览。

我们将目标俯仰角 β 的俯仰角代价计算为 $c_p = ((x_s - x_{t+1})/\|x_s - x_{t+1}\| \cdot v_g - \arccos(\pi/2 - \beta))_2$，其中 v_g 表示重力方向（0，0，−1）（见图 13A）。

朝向： 我们在两种情况下加入相机朝向项 c_o。对于面部镜头和动作镜头，我们鼓励摄像机垂直对准连接讲师双肩的直线。在横移运动中，我们引导摄像机垂直对准手指运动方向。由于我们已经有一个用于控制相机俯仰角的代价项，因此只需惩罚下一步的相机朝向向量与地平面上的目标相机方向向量之间的差异（图 13B）。详细推导请参阅附录 A.1 中的 c_o。

总之，用于选择下一步相机位置的代价函数为

$$J = \omega_{sm} c_{sm} + \omega_{dd} q_{dd} c_{dd} + \omega_p q_p c_p + \omega_o q_o c_o$$

其中 ω_{sm}、ω_{dd}、ω_p、ω_o 是各成本项的权重。如果当前镜头不是高角度镜头，则 $q_{dd} = 1$，否则为 0。如果当前摄像机角度设置为俯拍角度，或当前镜头为讲师镜头，则设为 $q_p = 1$，否则为 0。如果当前镜头是手部镜头或讲师镜头，或者摄像机正在进行横移运动，则 $q_o = 1$，否则为 0。对于这些权重，我们通过经验将 $\omega_{sm} = 1.0$、$\omega_{dd} = 0.2$ $\omega_p = 1.0$ $\omega_o = 0.5$ 、设定为，以在响应性和平滑跟踪之间取得平衡。为简化变量的界限，我们将 x_{t+1} 重新参数化为 $(\theta_{t+1}, \psi_{t+1}, R_{t+1})$，采用极坐标系。我们使用这一非线性优化问题

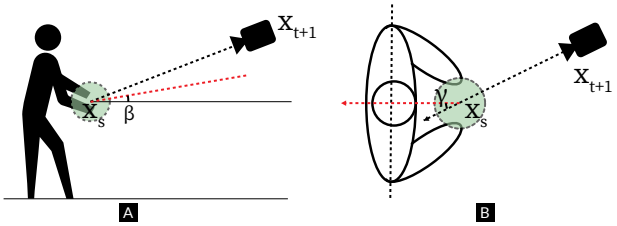

图 13：（A）下一步相机位置 x_{t+1} 的相机俯仰角与期望俯仰角（红色）之间差异的代价 β。（B）期望相机朝向（红色）与下一步相机位置 x_{t+1}的朝向之间差异的代价。γ 表示两者之间的差异。

通过 SciPy 中的 SLSQP 求解器进行求解。关于所使用的极坐标系的详细信息以及变量的界限，请参阅附录 A.2。

5.4.2 轨道路径。给定一个轨道中心，Stargazer 相机始终沿着半径为 $\pi/4$ 的弧线移动，并以 $\pi/6$ 的俯仰角，在距离轨道中心 60cm 处朝向轨道中心。

5.4.3 实现细节。机器人工作站运行 Ubuntu 20.04 和 ROS 1 Neotic，配备 Intel Xeon W2295 CPU 和 503 GB 内存。传感器工作站运行 Windows 11 7 8750 16，配备 Intel i CPU 和 GB 内存。我们用 C++ 实现了一个对性能要求较高的 ROS 节点，用于计算末端执行器速度；其他节点则用 Python 实现。语音检测和标注程序用 C# 实现。其他传感器对应的程序用 Python 实现。

6 评估

我们开展了一项初步用户评估，邀请六位讲师创作视频教程，演示各类身体技能。此次评估旨在初步了解 Stargazer 支持拍摄教程视频的能力，以及讲师与 Stargazer 合作时的感受。

6.1 研究设计与参与者

我们邀请了六位身体技能指导者（四女两男），请他们为自己擅长的技能制作视频教程。三位参与者（P1、P5 和 P6）是艺术家，他们以专业身份从事教程中讲解的手工艺。其余三位参与者是经验丰富的业余爱好者，在所教授的技能方面拥有一年以上的经验。两位参与者（P2 和 P5）曾通过在线视频或直播教授身体技能。P5 和 P6 是专业电影制作人。参与者的更多详情请见表 1。参与者获得了相应的时间报酬。

本研究方案已获我们所在机构的伦理审查委员会批准。

6.2 任务与流程

我们请参与者不受限制地自由展示自己的动手技能。唯一的限制是规模（任务在工作台上演示）和视频教学的时长（每次约5分钟）。参与者自备材料。图14展示了一位参与者正在用摄像机器人 Stargazer 拍摄视频



图14：参与者正在录制滑板维护教程视频。

。每次体验约持续90分钟，分为三个阶段：

培训。 完成一份人口统计调查后，参与者参加了培训环节，了解 Stargazer 的功能，包括其提供的镜头类型，以及用于改变摄像机行为的交互方式。参与者体验了所有功能，并重新制作了一段简短的乐高组装教程视频，过程中要求至少使用每项功能一次。

彩排与回顾。成功重新制作乐高教程后，参与者使用 Stargazer 拍摄了自己的实用技能教程的排练版本。拍摄期间，实验人员未提供任何指导或反馈。随后，参与者与实验人员一起回顾排练版本，并通过出声思考对视频进行评论，反思哪些部分可以改进。我们设置彩排环节，是为了让参与者练习讲解教程。

拍摄与回顾。随后，参与者制作了教程视频的第二次录制版本，纳入了他们在回顾排练版本后计划对内容和教学实施作出的修改。同样，参与者采用出声思考法来回顾结果，并评论

ID	背景	所教授的技能
P1（专业人士）	数字制造艺术家。	摄影测量与三维打印。
P2（资深爱好者）	经营滑板网红账号一个 Instagram 频道，教授滑板技巧。	滑板维护。
P3（资深爱好者）	虚拟现实研究员。	头戴式显示器设置。
P4（资深爱好者）	接受过艺术训练的手工爱好者。	黏土模型制作。
P5（专业人士）	电影制作人，教授定格动画电影制作方法，为大学生和公众提供教学。	定格动画木偶制作。
P6（专业人士）	数字艺术家和电影制作人。	互动雕塑制作。

表 1：评估参与者（Pro - 专业人士；EH - 资深爱好者）

视频质量以及拍摄时的体验。回顾阶段结束后，实验人员对参与者进行了半结构化访谈，重点探讨视频拍摄的替代方式、机器人如何融入他们当前的教程创作工作流程，以及将 Stargazer 与现有的视频采集解决方案进行比较，例如用于在多个摄像机信号源之间实时切换的 OpenBroadcast Studio（OBS）。最后，参与者填写了一份问卷，其中包括采用 7 点李克特量表的问题，内容涉及他们对机器人行为、交互以及所制作视频的感知。我们记录了机器人运动和用户输入日志，以进行初步性能分析。作为构建用于拍摄教程视频的交互式摄像机器人的第一步，我们的数据收集聚焦于指导者的视角，以获取他们对 Stargazer 的交互设计可能如何影响教学实施和视频质量的第一手描述。未来的工作则可以研究外部专家和学习者对所制作视频的视角。

6.3 设置与设备

我们在受控实验室环境中开展了这项研究。参与者展示技能所用的工作台尺寸为 1.6 米 × 0.8 米，距离机器人底座 1.2 米。我们选择了这一实体布置，并在运动规划中设置了限制，以确保机器人与讲师保持足够距离，避免发生碰撞。一位研究人员引导参与者完成实验并进行了访谈。机器人的设置如第 5 节所述。

6.4 结果

在本节中，我们将介绍参与者制作的视频结果、他们与 Stargazer 交互时的感受，以及初步的技术性能指标。

6.4.1 培训。所有参与者在培训环节拍摄乐高组装教程时，都能成功使用 Stargazer 的所有功能。参与者很快就掌握了基于手势的交互，但需要尝试 2-3 次才能学会将语音命令加入指导内容中。

6.4.2 视频内容。所有参与者都在第一次（彩排）或第二次拍摄后完成了教程视频。两位参与者（P4 和 P6）对第一次拍摄的结果感到满意，于是停止拍摄。其余参与者进行了第二次拍摄。以下对视频内容的分析基于最终的视频结果。

图15展示了参与者制作的每个视频的关键帧和时间线摘要。最终视频时长从245秒（4分5秒）到451秒（7分31秒）不等（M=320s，SD=69s）。各个视频所展示的技能在所需人工制品和流程方面差异显著，参与者也在内容呈现中融入了个人风格。总体而言，视频总时长的16.1%为指导者镜头，78.4%为动作镜头，5.5%为物体镜头。参与者通过灵活组合拍摄对象、摄像机取景和摄像机角度，呈现了丰富多样的镜头（见图15）。平均而言，每个视频出现了7.7次主体切换（例如从动作镜头切换到物体镜头，SD=2.6），以及5.7次取景变化（例如放大，

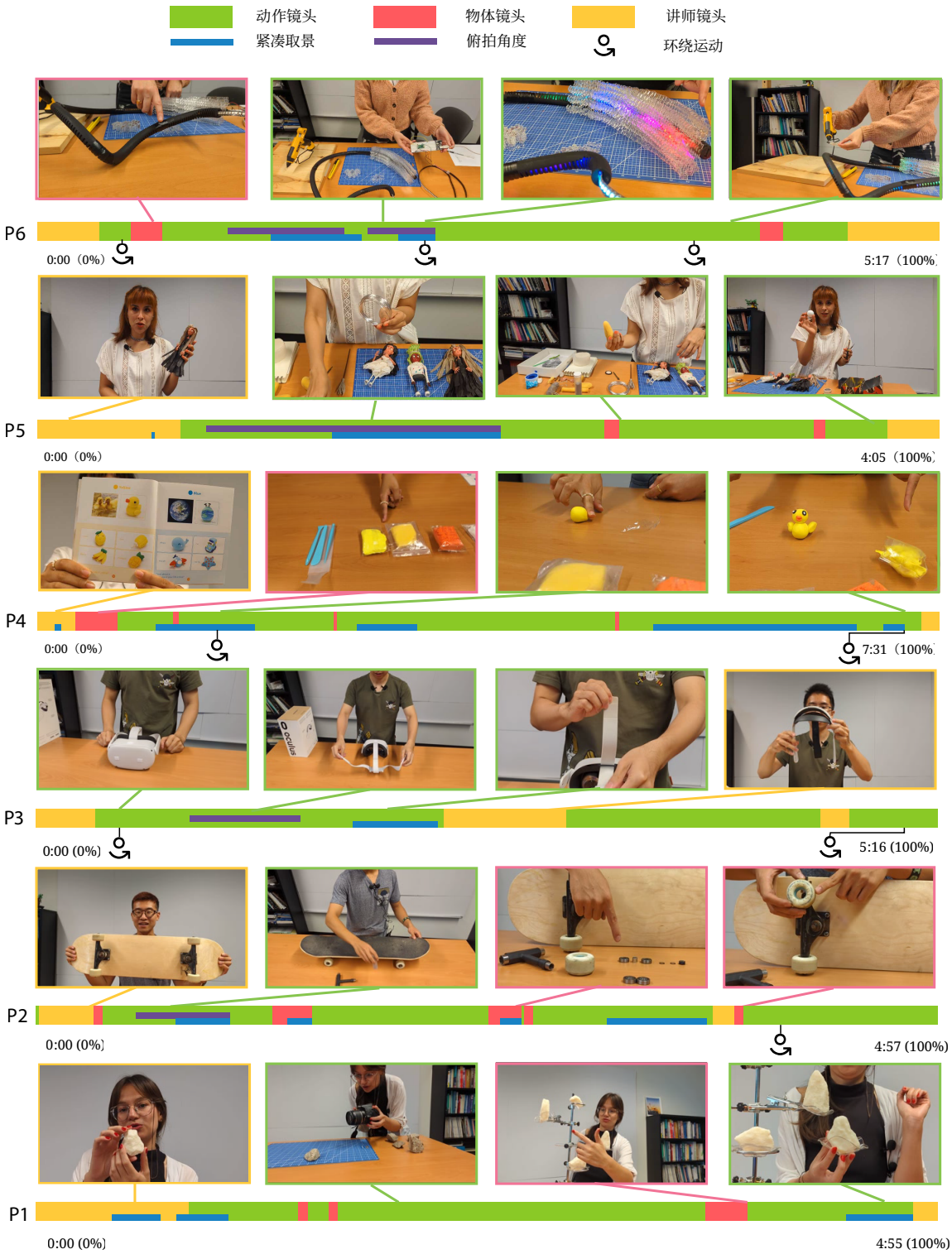

图15：参与者使用 Stargazer 拍摄的视频时间线。时间线按镜头类型、摄像机角度、摄像机取景和摄像机运动进行颜色编码。每段视频的代表性画面位于时间线上方。

SD=2.9），以及 1.7 次角度变化（例如从平视角度变为高角度，SD=1.5）。参与者也会在视频中加入摄像机运动，但使用程度各不相同。平均每个视频使用环绕运动 1.2 次（SD=1.2）。只有一位参与者使用过横移运动。我们在补充材料中附上了参与者制作的视频。

6.4.3 视频质量。参与者观看了自己制作的教程视频，并对成品质量进行了评价。所有参与者均表示，自己能够在制作的视频中成功呈现预先规划的教学内容（M=6.5，IQR=0.75）。此外，他们认为视频内容丰富（M=6.3，IQR=1.0），视觉质量良好（M=6.5，IQR=0.75），且具有良好的电影质感（M=6.0，IQR=1.5）。一位身为专业电影制作人的参与者（P6）强调了 Stargazer 拍摄的"一镜到底"视频的视觉特征："（一镜到底电影）背后有一种有趣的理论，关乎连贯性与缓慢感，使观众能够全身心投入其中。"事实上，专业人士长期以来都重视电影剪辑中的电影式连贯性，而认知科学研究则认为，连贯性有助于引导观众的视觉注意力 [53]。人机交互与教育心理学领域一个值得探讨的问题是，学习材料中的电影式连贯性会如何影响身体技能学习。

6.4.4 参与者的总体感知。参与者认为 Stargazer 对制作教程视频内容很有用（M=6.3，IQR=0.75）。专业艺术家参与者（P1、P5、P6）表示，他们经常需要通过视频记录自己的工作流程，"我们会请同事帮忙（拍摄），但他们并不总是有空"（P1）。在将 Stargazer 与他们当前的记录工作流程进行比较时，他们很欣赏 Stargazer 能够减轻架设摄像机和后期编辑的负担，"我做自己的事情，拍摄就会自动进行"（P5）。

6.4.5 机器人行为与指导者控制。我们综合了访谈数据、问卷回答，以及参与者观看所拍摄视频时的出声思维评论，对 Stargazer 的机器人行为及相关交互如何为参与者发挥作用进行了深入分析。

镜头类型。所有参与者都认为，动作镜头、讲师镜头和物体镜头这三种镜头类型足以满足他们制作教程的需求（M=6.0，IQR=1.5）。正如预期，参与者主要使用讲师镜头来介绍自己、任务和某些物体。他们使用动作镜头来展示步骤，并使用物体镜头来突出物体及其细节。然而，我们还观察到，参与者能够灵活调整机器人的行为，以满足其现场拍摄需求，从而模糊了预期镜头用途之间的界限。例如，在她制作的定格动画木偶教程中，P5 经常一边将某个部件拿近胸前，一边对其进行描述。这一手部位置会使摄像机在机器人仍处于拍摄动作镜头模式时，同时捕捉到她的面部和手部。她解释说，她引导机器人的摄像机进入这一位置，是因为她希望更频繁地与观众面对面交流。

摄像机取景和角度。大多数参与者（5/6）表示，他们能够以期望的取景和角度拍摄教程内容（取景：M=5.3，IQR=2.5；角度：M=5.1，IQR=1.0 对于角度）。所有参与者都认为，使用语音调整取景（M=6.6，IQR=0.75）和角度（M=6.5，IQR=1.0）的功能具有明显的实用价值。

参与者大多采用常规取景（讲师镜头使用中景，动作与物体镜头使用特写），占视频总时长的78.9%；但他们也会切换到更紧的取景，以突出细节。对取景的控制尤其有助于涉及更多细微尺度动作和物体的任务。P4在其完整视频的43%时长中采用了更紧的取景，以确保黏土模型的各个部分（包括面部特征等较小部件）都能被观众清楚看见（例如，图15中P4时间线上的最后一个关键帧）。参与者使用俯拍镜头来展示工作台的全貌（例如，图15中P2时间线上的第二个关键帧），或从不同且具有视觉趣味的视角拍摄同一物体（例如，P5和P6在图15各自时间线上的第二个关键帧）。在展示各个部件如何构成一件雕塑后，P6让机器人向下看，以展示雕塑的完整结构。

将摄像机控制融入教学。总体而言，参与者发现，他们可以将摄像机控制流畅地融入教学活动中。一位参与者（P5）借鉴了自己通过视频录制工具（Open Broadcaster Software，或 OBS）教授远程艺术课程的经验。该工具支持在视频源之间进行实时切换；她表示，Stargazer 的控制不会迫使她"每次都停下手头的工作，伸手去够那个盒子（用于在 *OBS* 中切换摄像机信号源的键盘）"。

四位参与者明确表示，使用指向手势来触发物体镜头很方便。他们发现，这种手势能够自然地融入自己的动作，不会打断讲解（M=6.7，IQR=0.75）。正如 P6 所说："不用特意去想"。P4 喜欢通过指向，快速切换摄像机拍摄工具或手头正在进行的工作。不过，P1 认为指向手势转瞬即逝，在一定程度上限制了她构图物体镜头的自由。为了让摄像机持续对准某个物体，同时腾出手去拿其他工具，她必须一只手持续指着该物体，另一只手拿取工具。她建议机器人应该"除非我让它停止，否则就一直对准某个物体"。

参与者表示，举手触发讲师镜头并不会打断他们的讲解节奏，但这种方式不如指向自然（M=5.2，IQR=2.25）。P3 解释道："手里有东西要展示时，举手会让人觉得很自然；但如果只是想快速和观众交谈，就没那么自然了"。P6 提出了注视的潜力："我希望看向摄像机时，它能转向我的面部；当我移开视线时，它再转回我的手部"。

六名参与者中的五名认为，使用语音控制摄像机取景和角度并不会干扰他们的教学实施（M=5.5，IQR=1.75）。P2认为这是他最喜欢的功能，因为"我可以随时呼叫它进行放大，即使我的双手都被占用……能够展示细节让我更想进行讲解，因为他们（观众）可以看到这些细节。"参与者喜欢 Stargazer 在选择语音命令方面提供的灵活性；然而，他们也表示，对于"机器人究竟能理解什么"（P1）的不确定性，可能会让他们反复思考该说些什么。

6.4.6 初步性能指标。我们根据数据

日志和研究环节录制视频的人工标注，报告了 Stargazer 原型若干方面的客观指标。请注意，这些指标仅用于初步了解我们所采用的技术方法。我们将彩排录像和最终拍摄片段（如有）纳入分析。分析的视频总时长为71分22秒。

指向手势。 我们比较了参与者手势的人工标注与检测结果日志。指向检测的精确率为91.2%，召回率为86.6%。

语音命令 我们计算了语音命令的指标，并以每个句子为单位进行标注（即，“你可以在这里进行近距离拍摄”视为一个句子）。人工共标注出41个建议采用“紧凑取景”的句子，Stargazer则标注出40个。该标签的精确率为90.0%，召回率为87.8%。人工共标注出18个建议采用“俯拍角度”的句子，Stargazer则标注出17个。该标签的精确率为94.1%，召回率为88.9%。我们还计算了参与者说完带有特定意图（例如紧凑取景）的句子，到机器人接收到相应状态变更命令之间的延迟。平均延迟为2.4秒（标准差=2.3秒）。较大的方差可能是由于语音识别结果的差异所致。

Stargazer原型的目标是演示并评估这样一种概念：通过跟随指导者提示来操控摄像机操作。我们预计，随着机器学习和感知技术的发展，我们所采用技术的性能将持续提升。

7 讨论

基于评估结果和观察，本节将讨论为摄像机器人解读指导者意图所面临的挑战、支持其他内容创作工作流的途径，以及最后的局限性与未来工作。

7.1 解读指导者意图

我们针对 Stargazer 的交互设计的目标之一，是识别并利用能够清晰表达指导者在摄影机配置方面意图的指导者提示。尽管我们的设计从评估中获得了令人鼓舞的反馈，但我们也注意到，未来能够解读指导者或其他参与者意图、以协助视频创作的摄像机器人还面临两个普遍挑战。

第一个挑战是理解未被清晰表达的指导者意图。尽管参与者通常并不认为加入摄影机控制提示会干扰教学，但他们仍然必须“主动记住什么时候该使用什么”（P3）。P3希望未来的机器人能够更加自主，“（内容创作者）只需做自己的事情，而不用考虑有一台摄像机”（P3）。P6建议，未来的机器人可以在没有指导者指令的情况下，根据当前场景引入更多合适的摄影机运动变化。能够在不需要人类刻意输入的情况下生成高质量镜头素材的摄像机器人，应当理解需要拍摄什么以及如何拍摄。然而，指导者可能不想明确指定这些内容，或者不知道如何清楚地表达它们。要从指导者提示中推断出此类意图可能很困难，因为手势和语音等提示脱离其语义（例如当前任务是什么）和时间（例如指导者迄今为止的动作）上下文后，往往具有歧义 [18]。应对这一挑战的一个令人振奋的机会可能是采用数据驱动方法，根据以脚本、分镜脚本或指导者迄今为止与之交互过的物体历史记录等形式呈现的教程语义内容来推断意图。

另一个挑战是构建能够适应指导者偏好和风格的个性化机器人行为。试图探索前沿的内容创作者可能会寻求“分布外”的镜头类型，而不是最常见的镜头类型。对于这些探索者而言，示教编程或其他更显式的机器人编程方法，可能有助于补充数据驱动或基于启发式的方法。

7.2 支持其他内容创作工作流程

Stargazer 目前通过单次连续拍摄来录制视频。参与者之一（P6）是一位电影制作人，他强调了一镜到底视频的美学特点，但这并非教学视频创作者采用的唯一形式。例如，经营滑板社交媒体频道的 P2 谈到了使用 Stargazer 拍摄滑板训练的可能性，并指出他的大多数视频都是经过大量剪辑、适合移动端观看的短片。他还必须多次完成并拍摄同一套训练动作，直到获得满意的效果。

许多内容创作者会剪辑原始素材，以使其更加简洁、清晰，并提升视觉质量。例如，有些内容创作者可能希望删去机器人从跟踪一个拍摄对象切换到另一个拍摄对象时的片段。未来的摄像机器人可以利用自身的数据日志辅助他们的工作流程。算法可以根据机器人的状态历史，在后期处理中自动删去视频中的这些片段。或者，机器人切换期间，系统可以切换到第二台固定摄像机的画面，以避免延迟。对于需要多次拍摄同一流程的讲师，未来的系统可以提供界面，让讲师根据首次拍摄的记录来规划后续拍摄中的机器人运动。

7.3 局限性与未来工作

机械臂的运动学约束，包括其固定底座和有限工作范围，限制了可用摄像机位置的范围。Stargazer 的可用摄像机位置范围足以应对桌面规模活动，但无法满足涉及房间内多个相距较远位置的任务。这些约束还限制了某些摄像机角度（例如主观视角、俯视）和摄像机运动（例如 180°环绕运动）的使用。未来，我们希望探索摄像无人机 [50] 和移动操作机器人 [3]在更大环境中从更多不同角度协助拍摄任务的潜力。

Stargazer 所响应的线索并未涵盖讲师意图中丰富的细微差别。例如，一些参与者试图通过指示手势触发物体镜头，比如将物体递给摄像机或展示给摄像机，但这些并不属于 Stargazer 能识别的线索。我们观察到，大多数参与者在训练阶段或首次拍摄时，都逐渐改变了自己对如何与 Stargazer 合作的心智模型。他们开始有意识地考虑机器人在讲解教程时能够理解什么，正如 P5 所说：“我开始觉得，这是我和机器人共同完成的一场表演。”未来研究可以

探索识别多样且细微意图的方法，或许可以结合注视、姿势和语音等多模态信号。

我们在 Stargazer 的视频分析和交互设计中，重点是理解和制作突出教学内容的视频。其他类型的体能技能类视频，例如创意直播，可能包含大量非教学内容，如社交互动和表演 [17]，而且时长可能长达数小时。我们计划进一步研究这些视频的内容及其创作者的摄像机控制需求，以确定更多线索和机器人行为，让 Stargazer 更好地支持非教学内容。

我们的视频分析所用数据集以观看次数作为视频质量达标的代理指标，但并未直接评估这些视频的质量。因此，我们的数据包含了常见的讲师做法，但未必是最佳实践。未来可以构建包含全面质量指标的数据集，为教程视频拍摄指南提供依据。

作为研究原型，目前的 Stargazer原型依赖昂贵的通用机械臂和一套外部传感器。随着机器人辅助拍摄日益普及，市场可能会需要价格更实惠、更易使用且更轻便的专用内容采集机器人。未来的工作可以探索此类机器人的技术和交互设计，以及它们对更广泛内容创作的影响。另一种方法是将 Stargazer 的部分摄像机控制交互应用于现成技术，例如由多台固定摄像机构成的摄影机配置。

我们采用恢复运动，使 Stargazer 机器人脱离接近运动学奇异点 [15] 或关节限位的构型。尽管恢复运动持续时间很短，但会破坏电影摄影的流畅连贯性。在实时机器人运动合成领域，计算完全无奇异点的运动仍是一个活跃的研究方向 [45]。解决这一问题的一个可行方案，是在运动规划中放宽对一个或多个自由度的约束。在机器人电影摄影的语境下，这种方法的一个例子是忽略摄像机与主体的距离。

尽管讲师本人对 Stargazer 拍摄的视频反响积极，但我们意识到，参与者可能会对自己的作品产生自我评价偏差。进一步的评估可以邀请所教授技能领域的外部专家以及电影摄影专家参与，以抵消这种潜在偏差。我们还注意到，Stargazer 的设计与评估尚未考虑教程视频最终消费者的视角。未来的工作将研究学习者对机器人拍摄视频的看法，并将其与使用固定摄像机拍摄的视频等基准方法进行比较。我们还对交互设计感兴趣，希望让学习者能够影响摄像机器人在实时教程中的行为。

8 结论

本文提出了 Stargazer，这是一种利用摄像机器人拍摄动态教程视频内容的新方法，能够响应讲师发出的细微提示。Stargazer 的摄像机可自动跟随讲师的手部或面部，并根据讲师的暗示在这些关注对象之间切换。讲师可以使用手势和语音，表面上是对观众说话，借此向 Stargazer 发出信号，改变它所跟随的关注对象、摄像机取景和摄像机角度。我们通过分析 50 部在线教程视频，为 Stargazer 的交互设计提供依据。一项有六名参与者（包括两名专业电影制作人）参与的用户研究表明，身体技能指导者能够使用 Stargazer 成功创作教程视频内容，并对成片质量感到满意。他们也认为这种交互不会干扰教学实施。Stargazer 探索了机器人辅助视频拍摄在教程视频领域的潜力，我们希望它能启发未来在人机协作内容创作方面的研究。

致谢

作者感谢参与者、匿名审稿人以及加拿大自然科学与工程研究委员会资助项目 RGPIN-2017-06415、DGDND-2017-00093、IRCPJ 545100-18、SMFSU 549149-2020 和 RGPIN-2017-04883 的支持。

参考文献

[1] 2022. Bolt：高速小型电影摄影机器人：Affordable Robotic Solutions。https: //www.mrmoco.com/motion-control/bolt/[2] Henny Admoni 和 Siddhartha Srinivasa。2016。通过眼神注视预测用户意图以实现共享自主性。载于 *2016*年*AAAI* 秋季研讨会系列。[3] Michael Ahn、Anthony Brohan、Noah Brown、Yevgen Chebotar、Omar Cortes、Byron David、Chelsea Finn、Keerthana Gopalakrishnan、Karol Hausman、Alex Herzog 等。2022。照我能做的做，而不是照我说的做：将语言建立在机器人可供性之上。*arXiv*预印本 *arXiv:2204.01691*（2022）。[4] Reuben M Aronson 和 Henny Admoni。2020。用于辅助操作的眼神注视。载于 *2020*年*ACM/IEEE*人机交互国际会议论文集。552–554。[5] Reuben M Aronson、Nadia Almutlak 和 Henny Admoni。2021。在遥操作期间通过注视推断目标。载于 *2021*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*）。IEEE，7307–7314。[6] Rogerio Bonatti、Wenshan Wang、Cherie Ho、Aayush Ahuja、Mirko Gschwindt、Efe Camci、Erdal Kayacan、Sanjiban Choudhury 和 Sebastian Scherer。2020。在非结构化环境中通过学习艺术性决策实现自主航拍电影摄影。《野外机器人学杂志》37，4（2020），606–641。[7] Blain Brown。2016。《电影摄影理论与实践：电影摄影师与导演的影像创作》。Routledge。[8] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020。语言模型是少样本学习器。《神经信息处理系统进展》33（2020），1877–1901。[9] William Cerbin。2018。改进学生从讲座中的学习效果。《心理学教学与学习研究》4，3（2018），151。[10] Minsuk Chang、Leonore V. Guillain、Hyeungshik Jung、Vivian M. Hare、Juho Kim 和 Maneesh Agrawala。2018。RecipeScape：用于大规模分析烹饪说明的交互式工具。载于 *2018*年*CHI*计算系统中的人为因素会议论文集（加拿大魁北克省蒙特利尔）*(CHI '18)*。美国计算机协会，纽约，纽约州，美国，1–12。https://doi.org/10.1145/ 3173574.3174025[11] Minsuk Chang、Mina Huh 和 Juho Kim。2021。RubySlippers：为教程视频提供基于内容的语音导航支持。载于 *2021*年*CHI*计算系统中的人为因素会议论文集（日本横滨）*(CHI '21*97, 14 10 1145 3411764*)*。美国计算机协会，纽约，纽约州，美国，文章页码。 3445151 https://doi.org/ . / .[12] Pei-Yu Chi、Joyce Liu、Jason Linder、Mira Dontcheva、Wilmot Li 和 Bjoern Hartmann。2013。DemoCut：为实体演示生成简洁的教学视频。载于 第*26*届*ACM*用户界面软件与技术年度研讨会论文集（英国苏格兰圣安德鲁斯）*(UIST '13)*。美国计算机协会，纽约，纽约州，美国，141–150。https: //doi.org/ 10 1145 2501988 2502052 . / .[13] Gibson Darden 和 Jane Shimon。2000。重新审视“老式”技术——用于运动技能学习和表现的录像反馈。《策略》13，4（2000），17–21。[14] Björn B. de Koning、Vincent Hoogerheide 和 Jean-Michel Boucheix。2018。通过教学视频学习的发展与趋势。《人类行为中的计算机》 89 2018 395 398 10 1016 2018 08 055 ( ), – 。[15] 2010 https://doi.org/ . /j.chb. . . Peter Donelan。。机器人操作臂的运动学奇异点。载于《机器人操作臂进展》。InTechopen。

[16] Maxwell Forbes、Rajesh PN Rao、Luke Zettlemoyer 和 Maya Cakmak。2015。通过结合情境的空间语言理解进行示范式机器人编程。载于2015年IEEE机器人与自动化国际会议（ICRA）。IEEE，2014–2020。[17] C. Ailie Fraser、Joy O. Kim、Alison Thornsberry、Scott Klemmer 和 Mira Dontcheva。2019。共享工作室：创意直播如何激发灵感、寓教于乐并促进参与。载于2019年创造力与认知会议论文集（美国加利福尼亚州圣迭戈）（C&C '19）。美国计算机协会，纽约，纽约州，美国，144 155 10 1145 3325480 3325485 – 。https://doi.org/ . / .
[18] Euan Freeman、Stephen Brewster 和 Vuokko Lantz。2016。“在那里做”：一种用于空中手势系统的交互技术。载于2016年CHI计算系统中的人为因素会议论文集（美国加利福尼亚州圣何塞）（CHI '16）。美国计算机协会，纽约，纽约州，美国，2319 2331 10 1145 2858036 2858308 – 。https://doi.org/ . / .
[19] Kathleen R Gibson、Kathleen Rita Gibson 和 Tim Ingold。1994。人类进化中的工具、语言与认知。剑桥大学出版社。[20] Brian Gleeson、Karon MacLean、Amir Haddadi、Elizabeth Croft 和 Javier Alcazar。2013。面向工业应用的手势：通过观察人类实现直观的人机通信。载于2013年第8届ACM/IEEE人机交互国际会议（HRI）。IEEE，349–356。[21] Pavel Gurevich、Joel Lanir、Benjamin Cohen 和 Ran Stone。2012。TeleAdvisor：一种多功能增强现实远程协助工具。载于SIGCHI计算系统中的人为因素会议论文集（美国得克萨斯州奥斯汀）（CHI '12）。美国计算机协会，纽约，纽约州，美国，619–622。https://doi.org/ 10 1145 2207676 2207763 . / .
[22] D. Brian Hall、Lyn Bartram 和 Matthew Brehmer。2022。用于向远程受众展示数据的增强手势学。载于第35届ACM用户界面软件与技术年度研讨会论文集（UIST '22）。美国计算机协会，纽约，纽约州，美国。[23] Vincent Hoogerheide、Margot van Wermeskerken、Hilke van Nassau 和 Tamara van Gog。2018。视频示范示例学习中的示范者与观察者相似性及任务适切性：示范者和学生的性别是否会影响测试表现、自我效能和感知能力？《人类行为中的计算机》89（2018），457–464。[24] Chong Huang、Fei Gao、Jie Pan、Zhenyu Yang、Weihao Qiu、Peng Chen、Xin Yang、Shaojie Shen 和 Kwang-Ting Cheng。2018。ACT：用于动作场景的自主无人机电影摄影系统。载于2018年IEEE机器人与自动化国际会议（ICRA）。7039–7046。https://doi.org/10.1109/ICRA. 2018.8460703[25] Chien-Ming Huang 和 Bilge Mutlu。2016。用于高效人机协作的预期式机器人控制。载于2016年第11届ACM/IEEE人机交互国际会议（HRI）。IEEE，83–90。[26] Kevin Huang、Jiannan Li、Mauricio Sousa 和 Tovi Grossman。2022。Immer-sivePOV：使用头戴式360°运动相机拍摄教程视频。载于2022年CHI计算系统中的人为因素会议论文集（美国路易斯安那州新奥尔良）（CHI '22）。美国计算机协会，纽约，纽约州，美国，文章91, 13 10 1145 3491102 3517468 页码。
[27] https://doi.org/ . / . Niels Joubert、Dan B Goldman、Floraine Berthouzoz、Mike Roberts、James A Lan-day、Pat Hanrahan 等。2016。迈向无人机摄影师：运用视觉构图原则引导四旋翼摄像机。arXiv预印本 arXiv:1610.01691（2016）。[28] Wendy Ju。2015。隐式交互设计。人本信息学综合讲座8，2（2015），1–93。[29] Juho Kim、Phu Tran Nguyen、Sarah Weir、Philip J. Guo、Robert C. Miller 和 Krzysztof Z. Gajos。2014。众包分步信息提取，以增强现有教程视频。载于SIGCHI计算系统中的人为因素会议论文集（加拿大安大略省多伦多）（CHI '14）。美国计算机协会，纽约，纽约州，美国，4017–4026。https: //doi.org/ 10 1145 2556288 2556986 . / . Ryan Knott。2022。你需要了解的视频统计数据、习惯与趋势。https: //www.techsmith.com/blog/video-statistics/[31] Jiannan Li、Maurício Sousa、Chu Li、Jessie Liu、Yan Chen、Ravin Balakrishnan 和 Tovi Grossman。2022。ASTEROIDS：探索用于实体技能示范的微型远程呈现机器人群。载于2022年CHI计算系统中的人为因素会议论文集（美国路易斯安那州新奥尔良）（CHI '22）。美国计算机协会，纽约，纽约州，美国，文章111，14页。https://doi.org/ 10 1145 3491102 3501927 . / . Jian Liao、Adnan Karim、Shivesh Jadon、Rubaiat Habib Kazi 和 Ryo Suzuki。2022。RealityTalk：用于AR实时讲故事的实时语音驱动增强演示。载于第35届ACM用户界面软件与技术年度研讨会论文集（UIST '22）。美国计算机协会，纽约，纽约州，美国。[33] Christian Licoppe。2017。Skype出场、多重问候与“coucou”：视频媒介对话开场的序列组织。《语用学》27，3（2017），351–386。[34] Camillo Lugaresi、Jiuqiang Tang、Hadon Nash、Chris McClanahan、Esha Uboweja、Michael Hays、Fan Zhang、Chuo-Ling Chang、Ming Guang Yong、Juhyun Lee 等。2019。Mediapipe：用于构建感知管线的框架。arXiv
[32] arXiv预印本 arXiv:1906.08172（2019）。[35] Lucy Macken 和 Paul Ginns。2014。指向和描摹手势可能有助于解剖学与生理学的学习。《医学教师》36，7（2014），596–601。[36] Cynthia Matuszek、Liefeng Bo、Luke Zettlemoyer 和 Dieter Fox。2014。通过未编排的指示手势和语言进行学习，以实现人机交互。载于《AAAI人工智能大会论文集》第28卷。[37] Richard E Mayer、Logan Fiorella 和 Andrew Stull。2020。提升教学视频效果的五种方法。《教育技术研究与发展》68，3（2020），837–852。[38] Antoine Miech、Dimitri Zhukov、Jean-Baptiste Alayrac、Makarand Tapaswi、Ivan Laptev 和 Josef Sivic。2019。Howto100m：通过观看一亿段带解说的视频片段，学习文本-视频嵌入。载于《IEEE/CVF国际计算机视觉大会论文集》。2630–2640。[39] Terrance Mok 和 Lora Oehlberg。2017。为远程受众评议实体原型。载于《2017年交互系统设计大会论文集》。1295–1307。[40] Benjamin A Newman、Abhijat Biswas、Sarthak Ahuja、Siddharth Girdhar、Kris K Kitani 和 Henny Admoni。2020。考察预测式机器人辅助对人类决策的影响。载于《国际社交机器人大会》。Springer，590–603。[41] Tobias Nägeli、Javier Alonso-Mora、Alexander Domahidi、Daniela Rus 和 Otmar Hilliges。2017。通过动态避障和视点优化实现航拍摄像的实时运动规划。《IEEE机器人与自动化快报》 2, 3 2017 1696 1703 10 1109 2017 2665693 ( ), –。
[42] https://doi.org/ 。/LRA. 。Nadya Peek、Jennifer Jacobs、Wendy Ju、Neil Gershenfeld 和 Tom Igoe。2021。远程创作：疫情期间的实践课程教学。载于《2021年CHI计算系统中的人为因素会议扩展摘要》。1–5。[43] Ken Perlin、Zhenyi He 和 Karl Rosenberg。2018。Chalktalk：一种可视化与交流语言——作为计算机科学教育领域的工具。arXiv预印本 arXiv:1809.07166（2018）。[44] Daniel Rakita、Bilge Mutlu 和 Michael Gleicher。2018。一种用于有效远程遥操作的自主动态摄像机方法。载于《2018年第13届ACM/IEEE人机交互国际会议（HRI）》。325–333。[45] Daniel Rakita、Bilge Mutlu 和 Michael Gleicher。2018。RelaxedIK：准确且可行的机械臂运动的实时合成。。载于《机器人学：科学与系统》。匹兹堡，宾夕法尼亚州，26–30。[46] Mose Sakashita、E. Andy Ricci、Jatin Arora 和 François Guimbretière。2022。RemoteCoDe：通过机器人具身化增强远程协作任务中的周边感知。《ACM人机交互汇刊》6，CSCW1，文章63（2022年4月），22页。https://doi.org/10.1145/3512910[47] Kenji Sakita、Koichi Ogawara、Shinji Murakami、Kentaro Kawamura 和 Katsushi Ikeuchi。2004。通过注视信息解读人类意图，实现人与机器人之间的灵活协作。载于《2004年IEEE/RSJ智能机器人与系统国际会议（IROS）（IEEE目录号04CH37566）》，第1卷。IEEE，846–851。[48] Nazmus Saquib、Rubaiat Habib Kazi、Li-Yi Wei 和 Wilmot Li。2019。用于增强视频表演的交互式身体驱动图形。载于《2019年CHI计算机系统人因大会论文集》（英国苏格兰格拉斯哥）（CHI '19）。美国计算机协会，纽约，纽约州，美国，1–12。https://doi.org/ 10 1145 3290605 3300852 。/。Allison Sauppé 和 Bilge Mutlu。。
[49] 2014 机器人指示表达：手势与情境如何塑造指称交流。载于《2014年第九届ACM/IEEE人机交互国际会议（HRI）》。IEEE，342–349。[50] Emmanuel Senft、Michael Hagenow、Pragathi Praveena、Robert Radwin、Michael Zinn、Michael Gleicher 和 Bilge Mutlu。2022。用于支持远程机器人操作的自动化无人机视点方法。arXiv预印本 arXiv:2208.04391（2022）。[51] Emmanuel Senft、Michael Hagenow、Robert Radwin、Michael Zinn、Michael Gleicher 和 Bilge Mutlu。2021。面向人机协作的情境化实时编程。载于《第34届ACM用户界面软件与技术年度研讨会》（线上活动，美国）（UIST '21）。美国计算机协会，纽约，纽约州，美国，613 625 10 1145 3472749 3474773 –。
[52] 1986 https://doi.org/ 。/。Thomas Sheridan。。机器人系统的人类监督控制。载于《论文集：1986年IEEE机器人与自动化国际会议》，第3卷。IEEE，808–812。[53] Tim J Smith。2012 6，1 2012 1 27。《电影连续性的注意理论》。《投影》（），–。[54] Maurício Sousa、Daniel Mendes、Rafael Kuffner Dos Anjos、Daniel Medeiros、Alfredo Ferreira、Alberto Raposo、João Madeiras Pereira 和 Joaquim Jorge。2017。用于情境感知界面的 Creepy Tracker 工具包。载于《2017年ACM交互表面与空间国际会议论文集》（英国布莱顿）（ISS '17）。ACM，纽约，纽约州，美国，191–200。https://doi.org/10.1145/3132272.3134113[55] Ryo Suzuki、Rubaiat Habib Kazi、Li-yi Wei、Stephen DiVerdi、Wilmot Li 和 Daniel Leithinger。2020。RealitySketch：通过动态绘制将响应式图形与可视化嵌入增强现实。载于《第33届ACM用户界面软件与技术年度研讨会论文集》（线上

活动，美国）*(UIST '20)*。美国计算机协会，纽约，纽约州，美国，
166 181 10 1145 3379337 3415892 –。https://doi.org/ 。/。
[56] Stefanie Tellex、Nakul Gopalan、Hadas Kress-Gazit 和 Cynthia Matuszek。2020。使用语言的机器人。《控制、机器人与自主系统年度评论》3，1（2020）。[57] Katherine M Tsui、Munjal Desai、Holly A Yanco 和 Chris Uhlik。2011。探索远程呈现机器人的应用场景。载于《*2011*年第六届*ACM/IEEE*人机交互国际会议（*HRI*）》。IEEE，11–18。[58] Sylvaine Tuncer、Barry Brown 和 Oskar Lindwall。2020。暂停：如何利用在线教学视频完成实践任务。载于《*2020*年*CHI*计算系统中的人为因素会议论文集》（美国夏威夷州檀香山）（*CHI '20*）。美国计算机协会，纽约，纽约州，美国，1–12。https://doi.org/10.1145/3313831.3376759[59] Ana M Villanueva、Ziyi Liu、Zhengzhe Zhu、Xin Du、Joey Huang、Kylie A Peppler 和 Karthik Ramani。2021。RobotAR：一种兼容增强现实的远程咨询机器人工具包，用于增强型创客空间体验。载于《*2021*年*CHI*计算系统中的人为因素会议论文集》（日本横滨）（*CHI '21*）。美国计算机协会，纽约，纽约州，美国，文章
477, 13 10 1145 3411764 3445726 页。https://doi.org/ 。/。
[60] Kenta Yamamoto、Yuki Koyama 和 Yoichi Ochiai。2022。摄影师参与的贝叶斯优化在摄影灯光设计中的应用。载于《第*35*届*ACM*用户界面软件与技术年度研讨会论文集》（美国俄勒冈州本德）（*UIST '22*）。美国计算机协会，纽约，纽约州，美国，文章
92, 11 10 1145 3526113 3545690 页。
https://doi.org/ 。/。[61] Yaxi Zhao、Razan Jaber、Donald McMillan 和 Cosmin Munteanu。2022。。“倒回到肉块抖动的部分”：理解日常任务中教学视频的语音控制。载于《*2022*年*CHI*计算系统中的人为因素会议论文集》（美国路易斯安那州新奥尔良）（*CHI '22*）。美国计算机协会，纽约，纽约州，美国，文章58，11页。https://doi.org/10.1145/ 3491102.3502036

A 机器人运动规划的补充细节

除了 5.4 节中的内容外，我们还进一步介绍运动规划的细节，包括方向代价的计算，以及我们在优化中使用的坐标系。

A.1 方向代价

如 5.4.1 节所述，对于方向代价项，我们根据下一相机位置的朝向$\mathbf{v}_{t+1}$与目标相机方向$\mathbf{v}_o$之间的差异进行计算。这两个向量都位于地平面上（图 16 中的 x-y 平面）。我们以摄像机应垂直于指导者的两个肩关节之间的连线（v_{sh} 表示其方向）为例，说明这两个向量的计算过程。对于横移运动，摄像机方向应垂直于手指轨迹，其计算方法类似。

为计算 $\mathbf{v}_o$，我们首先得到一个特定向量 $v_{ba} = v_{sh} \times v_g$，它与 v_{sh} 垂直。$v_g = (0, 0, -1)$ 是重力方向。然后，v_o 是 v_{ba} 在地平面上的投影，并已归一化。为计算 $\mathbf{v}_{t+1}$，我们只需将摄像机朝向 x_{t+1} 减去 x_s 投影到地平面上。随后，方向代价为这两个向量之差的范数的平方。即 $c_o = \|\mathbf{v}_{t+1} - \mathbf{v}_o\|^2$

A.2 用于约束优化的极坐标系

为了简化优化中的约束条件，我们将摄像机的下一个位置从笛卡尔坐标转换为极坐标。极坐标系的原点位于机器人的第二个关节处（图 16）。根据 Panda机器人的规格，该点在机器人坐标系中的笛卡尔坐标为 (0, 0, 0.333m)。极坐标系的 x、y、z 轴与机器人坐标系

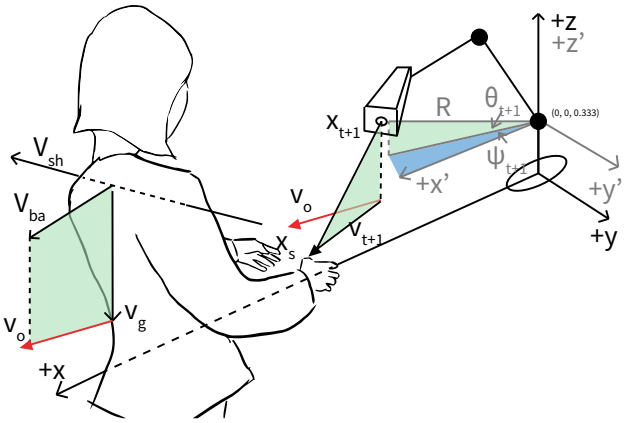

图16：用于优化的期望摄像机方向v_o（红色）以及极坐标系（坐标轴为灰色）的计算。垂直平面中的角度为绿色，水平面中的角度为蓝色。

平行。 R是摄像机与原点之间的距离。 θ是垂直平面上的角度。 ψ是水平面上的角度。我们选择 $R \in (0.36m,\ 0.66m)$、 $\theta \in (-\pi/10,\ 0.4\pi)$ 和 $\psi \in (-0.4\pi,\ 0.4\pi)$ 作为优化中的边界。