用于三维严肃游戏自动化任务表现评估的评估 HTN（A-HTN）

Kevin Desai*计算机科学系 德克萨斯大学圣安东尼奥分校

Omeed Ashtiani†计算机科学系 德克萨斯大学达拉斯分校

Balakrishnan Prabhakaran‡计算机科学系 德克萨斯大学达拉斯分校

arXiv:2302.05795v1 [cs.HC] 11 Feb 2023

摘要

近年来，人们开发了各种三维混合现实严肃游戏，用于体能训练、康复和教育等不同应用。严肃游戏中的任务表现是衡量用户完成游戏目标的效率和准确度的一项指标。以往研究采用基于图的任务表示方法，例如分层任务网络（HTN），但这类方法仅对游戏任务建模，并不进行评估。本文提出评估 HTN（A-HTN），既能高效地对任务建模，也能将评估逻辑纳入游戏目标中。根据任务表现的评估方式，A-HTN 可自动执行：(a) 通过比较物体操作进行任务级评估；(b) 通过比较运动轨迹进行动作级评估。该系统还可根据评估对象，将任务表现评估分为单个用户或多用户评估。我们通过两款三维虚拟现实严肃游戏展示 A-HTN 的有效性：一个比重计实验和一个多用户化学实验。A-HTN 实验显示，教师评分与系统生成的分数高度相关，这表明所提出的 A-HTN 在自动评估方面达到了与领域专家相当的水平。

索引词： 以人为中心的计算—混合/增强现实—；以人为中心的计算—协作交互—；应用计算—交互式学习环境—；社会与专业主题—学生评估—；以人为中心的计算—用户模型—

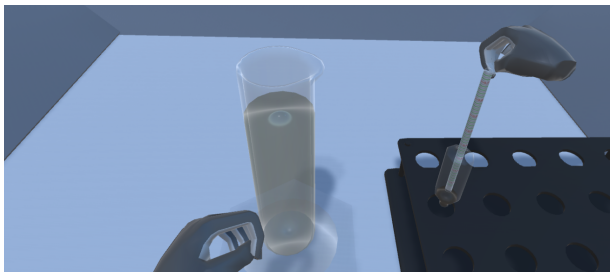
1 引言

三维游戏已开始应用于虚拟实验室，旨在传授科学、技术、工程和数学（STEM）领域的知识与实践技能。由于近期的新冠疫情，许多大学开始提供居家虚拟课程，而非线下课程。新罕布什尔大学等机构已采用虚拟现实教室，例如其 Parpak 项目。 [1]。许多此类课堂环境（如化学实验室课程）都需要三维空间，以便充分操控必要的工具。例如，考虑化学比重计实验：学生需要将易碎的比重计放入量筒，同时确保比重计不接触量筒壁。在这种情况下，深度感知至关重要，因此必须将该实验设计为三维环境。随着此类三维虚拟环境的普及，对其中学生能力进行自动评估的需求也随之增加。

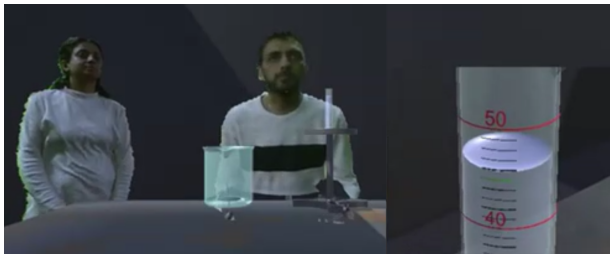
*电子邮箱：kevin.desai@utsa.edu

†电子邮箱：omeed.ashtiani@utdallas.edu

‡电子邮箱：bprabhakaran@utdallas.edu



(a)



(b)

图1：严肃游戏案例研究——(a) 比重计实验 (b) 协作化学实验

严肃游戏中的任务表现，是指用户完成任务以实现总体目标时的效率和准确性。要评估用户在游戏中的任务表现，必须将每项任务的细节融入游戏设计中。在现实世界中，通常会有领域专家（SME），例如治疗师或指导教师，负责评估用户的任务表现。为了实现自动评估，游戏作者应与领域专家合作，采用有原则的评估设计框架，并将每项任务的细节纳入其中。一种此类框架称为隐性评估 [23]，，它衡量能力这一概念构念（即技能与能力）、游戏内任务，以及在游戏过程中收集到的能力证据。

1.1 任务表现评估分类

严肃游戏（例如用于 STEM 培训的游戏）可以是单人游戏，也可以是多人游戏。根据评估的“对象”是谁，任务表现评估可分为：

- **单个用户**——评估用户在单人环境中的严肃游戏任务表现。
- **群体** ——评估严肃游戏中的多个参与者作为一个整体的表现；每次评估整个群体。
- **群体中的单个用户**——评估多人游戏环境中单个用户的严肃游戏任务表现。

除了要评估“谁”之外，我们还需要了解如何进行评估。在机器人学领域，示教学习（LfD）系统旨在让非专业人士教机器人执行新任务。如 [5],所述，LfD 方法通常可分为两大类：(a) 任务级学习： [4, 17, 19] 高层次的复杂任务，例如堆叠积木或整理对象，并假设机器人已掌握完成这些任务所需的基本动作；(b) 动作级学习： [3, 20, 26] 教机器人学习低层次的操作轨迹，例如抓取或拾取对象。我们将任务表现评估分为：

- **任务级评估：**通过比较不同的虚拟**对象操作**——位置、方向、碰撞、附着、文本输入等，来评估用户的任务表现。这对于评估用户的能力和对任务的了解至关重要。

- **动作级评估：**不只是比较起始位置或最终位置，而是利用对象的**运动轨迹**来评估任务表现。这有助于了解某些安全操作规范，例如如何正确握持试管、化学品与面部之间应保持的距离，或如何将工具递给其他用户。

1.2 提出的方法

要在严肃游戏中实现自动化任务表现评估，我们需要同时表示任务和评估逻辑。此前已有研究提出使用分层任务网络（HTN） [18] 和动态贝叶斯网络（DBN） [13]作为严肃游戏的建模策略。然而，自动化评估的适用范围非常有限，主要局限于由单个用户 [11] 体验的特定严肃游戏，或在线二维游戏 [6, 10, 27]。据我们所知，三维严肃游戏中的多用户自动化任务表现评估尚未得到探索。HTN 模型 [22] 仅用于任务的图形化表示，并未用于表现评估。本文提出评估 HTN（A-HTN），它不仅表示任务，还能执行自动化任务表现评估。通过这一扩展，我们还将严肃游戏任务和表现评估的适用范围扩展到任意数量的用户，而不再局限于单个用户。A-HTN 应适用于任何类型的游戏，例如一维、二维或三维游戏；不过，我们通过三维虚拟现实严肃游戏案例研究展示了 A-HTN 的有效性，如图 1 所示。

2 相关工作

针对严肃游戏中的任务建模，已经开展了大量研究。 [25] 介绍了智能辅导系统（ITSs）在课程教学和教育研究中的不同应用背景。通用智能辅导框架（GIFT）是一个开源、领域无关的智能辅导系统框架，用于创建课堂使用的自适应辅导内容。GIFT 为指导教师和学科专家（SMEs）提供了创作工具；使用这些工具无需具备计算机科学背景，即可创建完全自适应的计算机辅助课程。指导教师可以确定课程的组成部分（如调查问卷、测验、课程材料和视频）以及课程路径。 [7] 对 33 篇论文进行了系统性文献综述；这些论文从 2009 年至 2013 年发表在七个数字图书馆中的 4622 篇论文中筛选并最终确定，旨在了解非程序员作者如何使用创作工具进行智能辅导系统设计。 [6] 介绍了领域无关的基于 Web 的 Siette 评估环境，该环境可以与智能系统或大型学习管理系统（LMSs）（如 Moodle）进行语义集成。该环境集成了多种支架功能，如提示、反馈和错误观念，此外还涵盖了不同的教育需求和技术，例如间隔重复、协作测试或泛在学习。以及其他涵盖不同教育需求和技术的功能，例如间隔重复、协作测试或泛在学习。针对大规模开放在线课程（MOOCs）， [27] 提出了一种学习者冷启动算法，该算法基于行列式点过程对一组初始且多样化的问题进行采样，从而为学习者提供有用的反馈。 [21] 提到了一种自动化方法，用于为针对读写技能的教育干预研究制定可靠评分量规。该方法将成功的写作干预研究中使用的主旨评分量规，与一种用于评估机器生成摘要的高可靠性智慧群体内容评估方法进行了比较。一个集成了评估方法的教育系统被指导教师用于评估 400 名学生在学习搜索算法方面的表现 [10]。

据我们所知，目前只有少数建模与学习策略被用于形式化表示三维严肃游戏（如化学实验室 [8], ），并将评估纳入其中。尽管以下模型未用于三维虚拟现实游戏，但它们确实包含了评估功能。分层任务网络（HTN）是一种用于表示游戏中所执行任务的图形化方法，并已被正式纳入 ANSI/CEA-2018标准 [22]。 [18] 使用 HTN，通过单次演示进行交互式分层任务学习，并以更换轮胎作为应用示例。 [11] 展示了一种基于有色Petri网的模型，用于评估空中交通管制场景中人类操作员的认知表现。 [13] 和 [14]使用分层动态贝叶斯网络，提出一种计算框架，将复杂任务或动作序列的自动表现评估纳入其中。 [14] 使用分层动态贝叶斯网络，对学生在五种不同智能辅导系统中的知识进行建模，并进行隐性评估，以适应每位学生的需求。 [12] 更进一步，利用隐性评估框架比较了不同模型，例如单任务与多任务模型，以及静态与序列模型。 [15] 也使用隐性评估框架，并采用通过堆叠去噪自编码器预训练的深度神经网络，其表现优于支持向量机和朴素贝叶斯等经典机器学习技术。同样， [2] 利用长短期记忆网络改进隐性评估。隐性评估以证据中心设计（ECD）为基础，该方法通过任务、证据和能力模型，对学生的知识、技能和能力进行概率推理 [16]。在本研究中，我们将证据中心设计框架与分层任务网络（HTN）结合，构建用于自动化任务表现评估的评估 HTN（A-HTN）。

证据中心设计（ECD）是一种评估设计框架，包含多个要素，用于以简洁且可靠的方式对评估过程进行建模。这些要素包括：（1）能力模型：应评估哪些知识和技能；（2）证据模型：哪些行为或表现能够揭示这些能力；（3）任务模型：哪些任务能够引发这些行为 [23]。证据模型是能力模型与任务模型之间的纽带，它观察游戏内任务及其底层能力构念。 [16]描述了开发评估的一般方法论，我们将其与分层任务网络结合起来，以创建评估 HTN（A-HTN）。正如我们将通过 A-HTN 看到的，能力模型定义了应评估哪些知识和技能，即用户是否理解比重计实验，以及是否理解安全预防措施。为了展现这种能力，我们通过衡量用户在多大程度上遵循主题专家的物体操作和运动轨迹来收集证据。ECD 模型中的任务模型与我们的任务级评估和动作级评估直接相关。

3 评估 HTN（A-HTN）

分层任务网络（HTN） [22] 可以通过使用两种类型的节点——抽象节点和原始节点，以图形方式表示游戏中的任务。

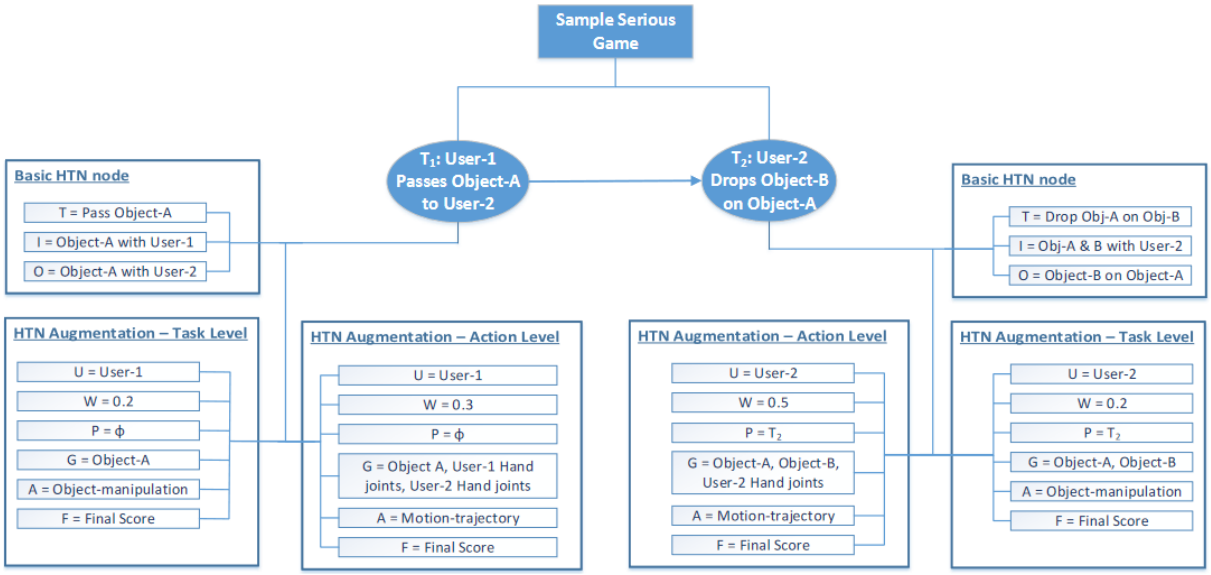


图2：样例严肃游戏的评估 HTN（A-HTN）

抽象节点是对所涉及不同任务的高层表示，由一个或多个原始节点以及零个或多个抽象节点组成。原始节点是叶节点，表示游戏中要执行的特定任务。任务的执行顺序由节点之间的箭头确定。每个原始 HTN 节点可以表示三项内容：(i) T = 要执行的任务，(ii) I = 任务的输入，以及 (iii) O = 任务的输出。因此，我们可以将 HTN 原始节点表示为 $N_i = \{T_i, I_i, O_i\}$。

设有一个由两名用户组成的游戏——U_1 和 U_2，，他们协作处理两个对象——A 和 B。需要执行两项任务：T_1，其中 U_1 将对象 A 传递给 U_2；以及 T_2，其中 U_2 将对象 B 放到对象 A 上。只有在 T_1 完成后，才能执行任务 T_2。需要根据每项任务中操作的虚拟对象，分别评估参与协作的用户 U_1 和 U_2，的任务表现。标准 HTN 表示法无法表示多个用户，也无法执行自动化任务表现评估。

3.1 对 HTN 的扩展

为了在多用户严肃游戏中纳入任务表现评估，我们提出了评估 HTN（A-HTN），其中原始节点 N_i 扩展了以下参数：

- U_i =在此任务中将接受任务表现评估的**用户**。这属于以下三种评估类别之一：单个用户、群体，或群体中的单个用户。
- W_i =针对所选用户进行总体表现评估时，特定任务的**权重**。
- P_i =**前置任务**，必须在此任务开始前完成。这些信息用于生成待执行任务的正确层次排序。
- G_i = 用于修改和执行此任务评估的**游戏对象**。
- A_i = **评估** 此任务所执行的评估类型。评估分为两种类型：使用对象操作进行任务级评估、使用运动轨迹进行动作级评估，或两者结合。用户的任务表现会与主题专家的参考表现进行比较，以评估任务并提供反馈。
  - **时间**可作为额外的表现特征加入比较。许多严肃游戏在设计中都固有时间限制。如果时间限制并非固有设定，则可将其纳入评估参数，以进行基于时间的评估。
- F_i = 向用户提供的**反馈**：实时 反馈或最终得分。实时表示在用户玩游戏时评估其表现并实时提供反馈。最终得分 表示在整个游戏过程中评估用户，并在任务完成后计算最终得分。

在对上述参数进行扩展后，原始 A-HTN 节点 N_i 现在可表示为 $N_i = \{T_i, I_i, O_i, U_i, W_i, P_i, G_i, A_i, F_i\}$。要为严肃游戏生成完整的 A-HTN，需要主题专家提供每项任务的详细信息。对于每项任务，主题专家首先提供任务名称，并将任务类型 选为抽象或原始。抽象任务 包含必须定义的子任务，因此不会分配其他 A-HTN 参数。另一方面，原始任务 需要分配所有 A-HTN 参数。主题专家还会选择前置任务 P_i，即在执行给定任务之前必须完成的任务。这些信息用于在 A-HTN 节点之间创建箭头。评估 A_i 可通过任务级评估、动作级评估或两者结合的方式进行。其余 A-HTN 参数则根据所选评估策略以不同方式分配。

图2展示了一个示例严肃游戏的 A-HTN，其中包含带有增强参数的原始节点。为清晰起见，增强参数中的任务级评估和动作级评估彼此分开，使指导教师或学科专家能够根据需要独立检查表现评估。

3.2 任务级评估

任务级评估取决于操作了哪些对象以及如何操作。为了评估用户的表现，主题专家以任务表现记录的形式提供参考数据。如果只提供一份参考记录，则将其视为完成任务的最佳方式；最差情况可能是不发生动作

或没有变化。遵循主题专家轨迹的任务可作为理解实验能力的证据。每项参考表现还会提供一个从 0.0 – 1.0 起的质量评分。对于用户在严肃游戏中执行的每项任务 T_i，根据其与主题专家参考记录的相对接近程度，从 0.0 – 1.0计算得出分数值 Ω_i。权重 W_i由主题专家提供，用于计算最终得分 Δ。

如果在主题专家（SME）的协助下正确定义，所提出的 A-HTN 可以纳入任何类型的物体操控检查。我们在案例研究中考虑的检查如下：

- **方向** ——通过比较用户执行任务时选定物体 G_j 的平均方向与主题专家提供的参考方向来计算分数。

- **位置** ——与方向检查类似，将用户操作物体 G_j 时的平均位置与参考位置进行比较，并计算基于距离的评分。

- **附着** ——确保物体 X 始终附着于另一个物体 Y。分数值根据X 附着于Y 的时长占完成任务总耗时的比例计算。

- **碰撞** ——确保物体 X 不会与另一个物体Y 或场景中的其他物体发生碰撞。根据碰撞的重要性，每次发生碰撞时，分数都可能被扣除预设的惩罚值，例如 0.01。

- **文本值输入** ——通过将用户使用键盘输入的值与参考值进行比较来计算分数。

3.3 动作级评估

对于某些类型的严肃游戏，例如比重计实验，运动轨迹非常重要，因此我们纳入了动作级评估策略。这一重要性源于正确的操作与安全要求：小瓶和工具应小心处理，并遵循适当的规程和安全预防措施。动作级评估关注玩家的运动轨迹，而非任务级物体操作。对于每项原始任务，主题专家不仅可以提供每个物品的参考信息，还可以提供以骨架和网格文件记录的运动轨迹信息。主题专家会选择要评估的任务所涉及的G 个身体部位，例如头部、手和手指，并使用用户的骨骼位置，根据参考数据计算用户的表现。我们通过使用面部与手部之间的身高比进行基于相关性的校正，将用户与主题专家身体构型之间的差异纳入考量。在用户实时操作过程中，系统会在参考关节位置生成气泡，提供视觉反馈。如果用户能够匹配主题专家的全部G_j 个骨骼参考点，气泡就会破裂，游戏随即进入下一帧。如果超过跳过 – 时间 仍未完成，例如 5 秒，我们便假定用户无法做出相应动作，并在适用时更新气泡位置，进入下一个动作。系统还会通过显示破裂／未命中气泡的数量以及已完成的重复次数，提供实时反馈。通过比较记录的参考骨骼数据与实时用户骨骼数据，还可以检测异常。在我们的案例研究中，我们检测了三类异常：

- 在特定游戏中，用户可能会在现实空间中绊倒并**摔倒**，从而可能受伤。这是一种严重的异常，应记录此类情况，以避免日后造成伤害，并确认物体的移动是否导致了摔倒。

- 我们会检查用户的**方向**，以确保用户面向工作站区域，且偏离任一方向不超过90度。

- 我们评估手**的位置**，以确保用户正确操作设备。

每个异常都会以实时反馈的形式显示在屏幕上，供用户查看，让用户在使用工具或移动对象时调整并纠正自己的姿势和动作。系统会记录这些异常，供主题专家日后参考和分析。如果任何异常持续时间超过异常等待时间（例如10秒），我们便会认为用户正在忙于处理外部事务、骨架记录不正确，或用户遇到了其他困难。为避免用户受到伤害并节省录制存储空间，游戏将退出。

4 案例研究

为了评估 A-HTN 的有效性，我们设计了两个严肃游戏案例研究：比重计实验和协作化学实验。这两项实验均使用 Unity3D 游戏引擎开发，采用 OpenGL 渲染，并使用 PhysX 实现物理交互。每个实验场地都配备了一台笔记本电脑，其处理器为 Intel i7 3.0 GHz，内存为 16 GB，显卡为 4 GB GTX 1070。比重计实验使用 Oculus Rift 作为显示设备，而协作化学实验使用 3D 电视。比重计实验中使用 Leap Motion 捕捉手部手指关节，以便与虚拟对象进行交互。由于协作化学实验需要捕捉用户的整个身体，我们使用了 Microsoft Kinect V2。用于生成三维人体模型的代码使用 C++ 编写，并在 Windows 上进行了 GPU优化。我们根据所执行评估的类型，将用户研究和评估分为两组——任务级操作评估和动作级评估。比重计实验和协作化学实验都包含这两种评估。严肃游戏中的任务表现评估视频见 https://youtu.be/Z0z20zUTELs。

4.1 比重计实验

比重计是一种用于测量液体比重（相对密度）的仪器，即液体密度与水的密度之比。我们创建了一个虚拟化学实验室来进行比重计实验，如图 1 所示。为成功完成实验，学生执行了以下任务：

- T_1 = 从上部杆身拿起比重计，以避免手上的油转移到比重计上。

- T_2 =将比重计移至量筒，并确保比重计不会从手中掉落或与量筒发生碰撞。

- T_3 =测量液面，待比重计在量筒中稳定后进行测量。为了准确测量，学生必须弯下身子，使头部与液面齐平。

- T_4 =输入测量值 到屏幕文本框中，并使用键盘输入。

图3展示了用于任务表现评估的 A-HTN，以及扩展后的节点参数。U = 单个用户，即一名学生，其**任务级评估** 基于实验中的A = 对象操作 进行，同时其**动作级评估** 则基于A = 运动轨迹进行。由于该实验是社区学院教学大纲的一部分，因此将F = 最终得分 作为反馈提供。每项任务都提供了前置任务 P_i。这些任务定义了任务顺序执行次序，并以箭头表示。学生操作游戏对象G = 比重计、量筒、手、头部/眼睛、文本值输入，并将其与主题专家的参考表现进行比较。

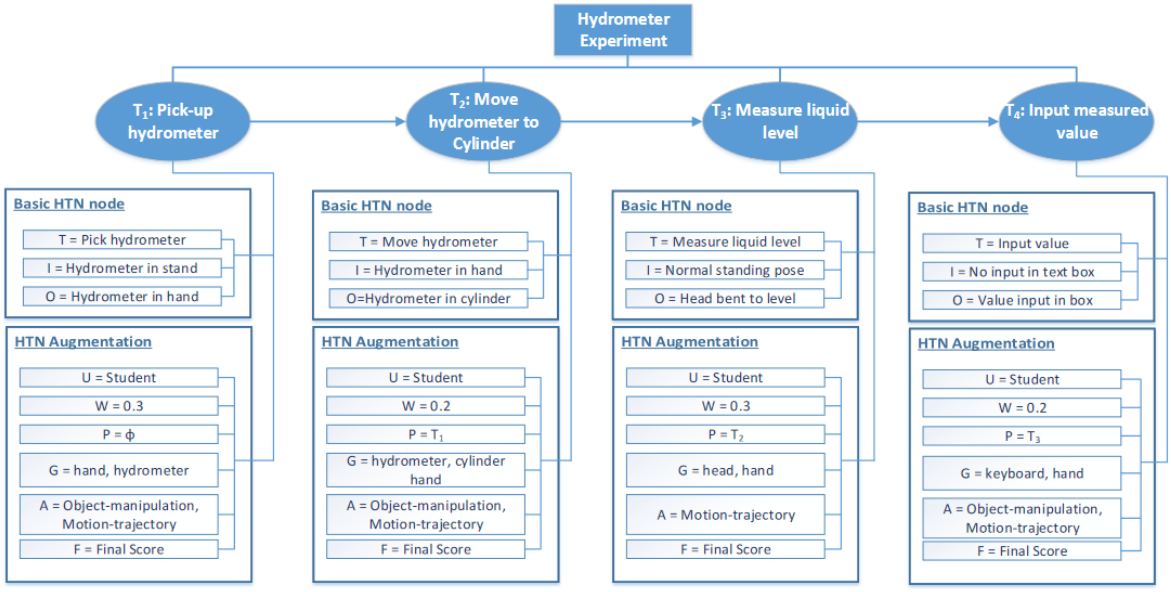


图3 ：用于比重计实验的评估 HTN（A-HTN） 验

• 对于任务*T*1，会检查用户手部的方向，以确保学生以正确的方式（即从上部）拿取比重计。

• 对于任务*T*2，执行了附着 检查，以确保比重计在移动到量筒的过程中与手相连。此外，还检查了比重计与量筒之间是否发生碰撞 。

• 对于任务*T*3，通过将学生的头部与量筒内液体的弯月面高度进行比较，对其进行了位置检查。

• 对于任务*T*4，通过比较学生提供的液体数值与主题专家提供的记录液体数值，执行了文本值输入 检查。

我们收集了这些任务级操作的数据，作为证据，用于评估用户理解各种液体密度的能力。虚拟现实游戏中的运动技能为用户的自动评估增加了额外的复杂性，因为解决方案并非简单的点击或按钮操作，因此我们也收集了动作级操作。通过遵循证据中心设计框架，我们旨在通过让用户执行（2）任务并分析（3）任务的（3）证据，来衡量用户的（1）能力。具体而言，在进行比重计实验时，我们评估用户在密度知识和设备操作安全方面的（1）能力。我们收集（2）任务级操作和动作级轨迹的数据，并将这些数据作为用户在知识和安全两方面的（3）能力证据。采用证据中心设计框架有助于我们更细致地了解用户在进行实验时可能欠缺哪些知识。

为每项任务提供了一个权重 W_i，并在比较后计算出分数 Ω_i。实验完成后，通过对所有单项任务分数进行加权和，计算得出总分 $\Delta = \sum_i W_i \cdot \Omega_i$。

4.2 协作化学实验

我们已展示如何将 A-HTN 用于单个用户的三维游戏。不过，许多环境和实验要求多人在同一地点共同执行任务。为此，我们采用了先前工作中展示的协作化学实验 [8]，即实验2。在该实验中，两名用户——一名指导教师和一名学生——虽身处三维环境中，但位于不同的实际地点，他们协作完成一项化学实验。要成功完成该实验，需要执行以下任务：

• T_1 = 指导教师将装有一定液位液体的量筒-*A* 递给学生。

• T_2 = 学生使用键盘将量筒-*A* 中的液面调整到指导教师建议的高度。学生必须以适当的方式倾斜并测量量筒，确认液面达到指定高度。

• T_3 = 量筒-*A* 中的液面达到正确高度后，指导教师将量筒-*B* 递给学生。

• T_4 = 学生以与*T*2 相同的方式调整量筒-B 中的液面。

• T_5 = 学生依次将量筒-*A* 和量筒-*B* 中的液体倒入烧杯并混合。化学反应会显示粒子效果。

协作化学实验的 A-HTN 如图4所示。如箭头所示，首先执行量筒-A 的任务，随后执行量筒-B 的任务，最后将两者中的液体混合。任务中修改的游戏对象包括 G =量筒-*A*、量筒-*B*、烧杯、头部/眼睛、键盘。我们利用此实验展示如何评估群体中的个体，并评估一名用户 U =学生的表现。任务的权重反映了通过赋予 T_2, T_4 和 T_5 $W_i > 0$的权重来评估学生的决策。在 T_2，中，学生需要先将量筒-A 中的液面设定到正确高度，然后才能继续。为完成此任务，系统会提供F = 实时反馈，以便同时评估学生并指导其如何正确完成任务

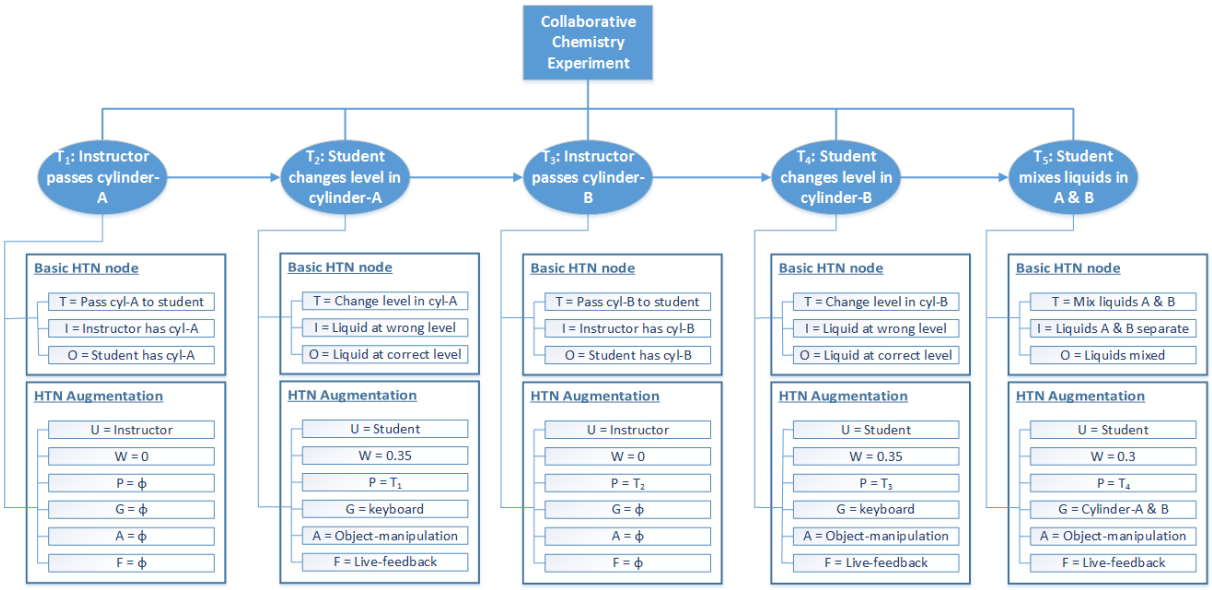


图4：协作实验的评估 HTN（A-HTN）

。通过将学生的A = 物体操作与主题专家的参考数据进行比较，对学生的表现开展任务级操作评估。

5 实验结果

这项比重计实验 是在当地一所社区学院的实验课中进行的，共有45名学生参加，由一位经验丰富的化学教授担任主题专家并负责指导。实验设备包括一台笔记本电脑、Oculus Rift 和 Leap Motion。主题专家提供了用于比较和表现评估的参考数据。每名学生都进行了实验，其任务和动作表现均由指导教师和系统进行评估。学生首先获准在示例虚拟现实房间中体验一段时间，以熟悉环境。对于每名学生，液体比重都是随机选取的，并相应地改变颜色。

我们对对象操作进行任务级评估，并对运动轨迹进行动作级评估，以评价学生对比重计实验及其安全要求的理解能力。由于反馈类型F 是最终得分，指导教师和系统分别计算了五项不同的分数：四项任务各一项，以及一项总最终得分。图5展示了五幅不同的散点图，以拟合直线的形式呈现系统生成的分数与教师评分之间的归一化相关性。我们不假设系统和指导教师给出的最终评估分数之间存在线性或单调关系，因此报告了三个相关系数——皮尔逊、斯皮尔曼和肯德尔，如表1所示。所有数值均使用 MATLAB 统计与机器学习工具箱提供的函数计算。

表1：比重计实验：系统生成的分数与教师评分之间的相关性

Task	皮尔逊	斯皮尔曼	肯德尔
T_1	87.54	88.12	75.28
T_2	95.45	89.11	85.41
T_3	93.78	91.77	81.95
T_4	82.56	85.45	82.00
Δ	91.83	90.69	77.64

从这些数值可以清楚看出，系统生成的分数与教师评分相关性良好。关于T_1, , 经主题专家审查后发现，分数差异的原因之一可能是用户从支架上取下比重计的方式不同。我们的系统可能按照关节结构的精确匹配进行评分，几乎没有留出变化空间，因此判定得更加严格。不过，指导教师对系统给出的整体评估表示满意，并同意未来可以大幅减少评分工作。

这项协作化学实验也是在与比重计实验相同的社区学院实验课中进行的。实验设置与比重计实验类似，不同之处在于需要两个场地，而不是一个。化学实验指导教师在其中一个场地，31名学生则分别在另一个场地参与实验。对每名学生都进行了物体操作评估，并在屏幕上提供实时反馈。指导教师和学生都能看到反馈。如有必要，指导教师可以通过语音和视觉通信提供额外帮助。不过，指导教师发现该系统非常直观，能够准确提供正确反馈，因此很少需要介入。

6 讨论

据我们所知，目前用于三维严肃游戏任务表现评估的模型和系统寥寥无几。大多数框架，例如 [6] 和 [27], , 只关注在线一维（文本）或二维（图像）内容的评估。一些详细且有效的框架专为特定案例研究而设计 [10]。比重计实验等实验要求学生在深度方向上操作，以便将易碎的比重计安全地放入量筒，同时不碰到筒壁。协作化学实验表明，即使学生属于一个群体，也有必要对其进行个别评估。对此类三维虚拟环境的需求不断增长，最终也推动了对自动化任务表现评估的需求。

三维严肃游戏中的评估仅在少数研究中得到探讨。 [13] 展示了一种三维严肃游戏评估系统，与我们的工作最为相关。该系统使用动态贝叶斯

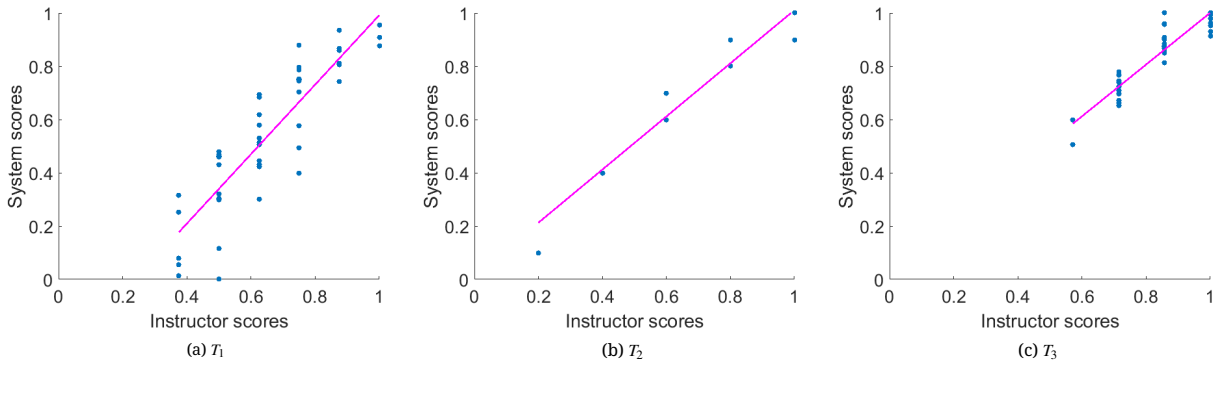


图 5：比重计实验：系统与教师评分分别针对（a - d）4 个原始任务 T_1 - T_4 以及（e）最终评估分数 Δ的散点图。

网络（DBN）评估一项火灾模拟研究中的任务表现。这项工作是迈向三维游戏自动化评估的重要一步，但其针对的是特定用例场景。尽管其他研究已针对各种数据集采用了不同的泛化策略， [9],我们将这些模型扩展到更广泛的用例，不仅涵盖三维游戏，也涵盖虚拟现实环境。借助 A-HTN模型，我们能够纳入多种（甚至可能是所有类型的）三维严肃游戏中的任务表现和动作表现，无论这些游戏是否基于虚拟现实。此类严肃游戏包括但不限于教育实验室和健身游戏。由于该方法同时涵盖任务级评估和动作级评估，目前没有其他系统能够与我们提出的 A-HTN直接或轻易地进行比较。

我们工作的另一项重要意义在于，我们评估的不仅是用户对实验相关知识的掌握能力，也评估其遵守安全预防措施的能力。例如，设想一个化学实验场景，研究人员使用了 2-巯基乙醇等化学品。重要的不仅是记录和评估他们如何使用该化学品，还要关注化学品与面部之间的关系。吸入 2-巯基乙醇可能会损害黏膜，并导致喉痉挛、肺炎和肺水肿。为了在虚拟场景中捕捉这一点，我们还通过动作级评估，根据比重计与面部之间的距离来评估用户的运动轨迹。

6.1 多用户或群体评估

有些三维严肃游戏允许多名用户身处不同地点，同时协作完成任务。我们的协作化学实验就是一个例子：两名学生身处不同地点，协作完成一项化学实验。根据任务的不同，主题专家可以选择评估谁应接受评估——学生$_1$, 学生$_2$ 或整个群体。我们提出的 A-HTN 可以轻松处理这三种情况中的任意一种。对于每项任务 T_i，变量 U_i 的值可以更改为正在接受任务表现评估的用户／群体。我们在实验中展示了评估群体中个体的能力。不过，通过调整指导教师的权重，我们表明所提出的 A-HTN 也能反映群组评分。

6.2 局限性与未来工作

在我们的工作中，可以通过任务级评估（对象操作）和动作级评估（运动轨迹）来评估表现。然而，也有一些游戏包含创造性环节 [24]。例如，设想一个物理环境，用户必须将一个物体运过障碍物。由于我们的系统在任务级评估（即物体从位置 A 移动到位置 B）和动作级评估（即角色如何转移物体）中都使用主题专家，因此用户的创造力或偏离主题专家所设路径的行为很可能会导致评估结果较低。未来，A-HTN 可以加入另一种评估类型，即创造力测量。然而，在我们的环境中，安全至上。为了确保安全得到教授和评估，我们牺牲了创造力。

目前，所提出的 A-HTN 和创作机制尚未尝试纳入可能涉及其他环境的创造性成分。我们将把这作为未来可能开展的工作。同样，我们尚未考虑任何非严肃（或仅供“娱乐”的）三维游戏；A-HTN 可能能够，也可能无法对这类游戏进行建模。

7 结论

本文提出了评估 HTN（A-HTN），用于评估三维严肃游戏中的任务表现。以往研究（包括 HTN 模型）主要侧重于对任务进行建模。然而，借助对 HTN 所作的扩展，我们还可以将任务表现评估逻辑纳入任何严肃游戏。我们使用 A-HTN 展示了两种不同的评估策略：（1）通过比较物体操作进行任务级评估；（2）通过比较运动轨迹进行动作级评估。A-HTN 可用于评估单个用户、用户群体或群体中的个别用户的任务表现。我们通过两个严肃游戏案例研究评估了所提出的 A-HTN，分别是比重计实验和协作化学实验。严肃游戏案例研究的视频可在 https://youtu.be/Z0z20zUTELs 查看。案例研究的实验结果展示了 A-HTN 在将评估逻辑纳入三维虚拟现实严肃游戏并执行自动化任务表现评估方面的实用性和有效性。

8 致谢

本材料基于美国国家科学基金会资助项目（编号：2153249）所支持的工作。本文所表达的任何观点、发现、结论或建议均属于作者个人意见，不一定反映美国国家科学基金会的观点。

参考文献

[1] UNH 虚拟现实 - Parpak Animated。https://vr.unh.edu/demos/parpak-animated. [2] B. Akram、W. Min、E. Wiebe、B. Mott、K. E. Boyer 和 J. Lester。利用基于 LSTM 的分析改进基于游戏的学习中的隐性评估。载于国际教育数据挖掘会议，2018年。[3] C. Bowen、G. Ye 和 R. Alterovitz。利用时变代价地图为已学习的任务进行渐近最优运动规划。*IEEE*《自动化科学与工程汇刊》，12 182，2014 1 171( ): – .[4] G. Chang 和 D. Kuli´c。使用佩特里网通过演示学习机器人任务。载于*2013*年 *IEEE RO-MAN*，第31–36页。IEEE，2013年。[5] S. Chernova 和 A. L. Thomaz。从人类教师处学习机器人。《人工智能与机器学习综合讲座》，8(3):1–121，2014年。[6] R. Conejo、E. Guzm´an 和 M. Trella。Siette 自动评估环境。《国际人工智能教育杂志》，

26 1 270 292, 2016 10 1007 40593 015 0078 4 ( ): – [7]

Mar . doi: . /s - - - D. Dermeval 、 R. Paiva 、 I. I. Bittencourt 、 J. Vassileva 和 D. Borges。用于设计智能辅导系统的创作工具：文献系统综述。《国际人工智能教育杂志》，28(3):336–384，2018年9月。doi: 10.1007/s40593-017 -0157-9[8] K. Desai、U. H. H. Belmonte、R. Jin、B. Prabhakaran、P. Diehl、V. A. Ramirez、V. Johnson 和 M. Gans。多模态协作虚拟实验室（ mmcvl ）的实践经验。载于*2017*年 *IEEE* 第三届多媒体大数据国际会议（*BigMM*），第376–383页。IEEE，2017年。[9] K. Georgiadis、G. van Lankveld、K. Bahreini 和 W. Westera。隐性评估的稳健性。*IEEE*《游戏汇刊》，13(2):180–192，2020年。[10] F. Grivokostopoulou、I. Perikos 和 I. Hatzilygeroudis。用于学习搜索算法并自动评估学生表现的教育系统。《国际人工智能教育杂志》，27(1):207–240，2017年3月。doi: 10.1007/s40593-016-0116 -x[11] A. Hasselberg 和 D. S¨offker。基于佩特里网的建模：以空中交通管制为例，对人类操作员的规划进行建模，以评估任务表现。*IEEE*《人机系统汇刊》，45(6):676–685，2015年12月。doi: 10.1109/THMS.2015.2456104

[12] N. Henderson、V. Kumaran、W. Min、B. Mott、Z. Wu、D. Boulden、T. Lord、F. Reichsman、C. Dorsey、E. Wiebe 等。采用混合式隐性评估框架，增强基于游戏的学习中的学生能力模型。国际教育数据挖掘学会，2020年。[13] M. R. Iseli、A. D. Koenig、J. J. Lee 和 R. Wainess。对游戏和模拟中的复杂任务表现进行自动化评估。载于军种/工业界训练、模拟与教育会议论文集，奥兰多，佛罗里达州，2010年。[14] T. K¨aser、S. Klingler、A. G. Schwing 和 M. Gross。用于学生建模的动态贝叶斯网络。*IEEE*《学习技术汇刊》，10(4):450–462，2017年。[15] W. Min、M. H. Frankosky、B. W. Mott、J. P. Rowe、E. Wiebe、K. E. Boyer 和 J. C. Lester。DeepStealth：利用深度学习模型，在基于游戏的学习环境中进行隐性评估。载于《教育中的人工智能：第*17*届国际会议，*AIED 2015*，马德里，西班牙，*2015*年*6*月*22*日至*26*日。论文集*17*》，第277–286页。施普林格，2015年。[16] R. J. Mislevy、L. S. Steinberg 和 R. G. Almond。专题文章：论教育评估的结构。《测量：跨学科研究与观点》，1(1):3–62，2003年。[17] A. Mohseni-Kabir、C. Rich、S. Chernova、C. L. Sidner 和 D. Miller。通过单次演示进行交互式分层任务学习。载于第十届 *ACM/IEEE* 人机交互国际会议论文集，第205–212页，2015年。[18] A. Mohseni-Kabir、C. Rich、S. Chernova、C. L. Sidner 和 D. Miller。通过单次演示进行交互式分层任务学习。载于第十届 *ACM/IEEE* 人机交互国际会议论文集，HRI '15，第205–212页。ACM，纽约，纽约州，美国，

2015 10 1145 2696454 2696474 。doi: . / . S. [19]

Niekum、S. Osentoski、G. Konidaris、S. Chitta、B. Marthi 和 A. G. Barto。从非结构化演示中学习基于实体的有限状态表示。《国际机器人研究杂志》，34(2):131–157，2015年。[20] T. Osa、N. Sugita 和 M. Mitsuishi。用于外科手术任务自动化的在线轨迹规划与力控制。*IEEE*《自动化科学与工程汇刊》，15(2):675–691，2017年。[21] R. J. Passonneau、A. Poddar、G. Gite、A. Krivokapic、Q. Yang 和 D. Perin。Wise crowd 内容评估与教育评分量规。《国际人工智能教育杂志》，28(1):29–55，2018年3月。doi: 10.1007/s40593-016-0128-6[22] C. Rich。使用 ANSI/CEA-2018 构建基于任务的用户界面。《计算机》，

42 8 20 27, 2009 10 1109 2009 247 ( ): – Aug . [23]

doi: . /MC. . V. J. Shute。通过基于计算机的游戏进行隐性评估以支持学习。《计算机游戏与教学》，55(2):503–524，2011年。[24] V. J. Shute 和 S. Rahimi。在物理视频游戏中对创造力进行隐性评估。《计算机在人类行为中的应用》，116:106647，2021年。[25] A. M. Sinatra、S. Ososky、R. Sottilare 和 J. Moss。关于在课堂和教育研究中使用自适应辅导系统的建议。载于 D. D. Schmorrow 和 C. M. Fidopiastis 编，《 style id='42'>增强认知：复杂人类环境中的认知与行为增强，第223–236页。施普林格国际出版社，查姆，2017年。[26] A. Vakanski、I. Mantegh、A. Irish 和 F. Janabi-Sharifi。使用隐马尔可夫模型和动态时间规整，通过示教学习机器人编程所需的轨迹。*IEEE*系统、人类与控制论汇刊*B*辑（控制论），42(4):1039–1052，2012年。[27] J.-J. Vie、F. Popineau、 ´E. Bruillard 和 Y. Bourda。通过自动化试题组卷解决大规模评估中的学习者冷启动问题。《国际人工智能教育杂志》，28(4):616–631，2018年12月。doi: 10.1007/s40593-017-0163-y