摘要

随着世界模型在具身智能领域的快速发展，越来越多的工作开始探索将视频基础模型作为生成式世界模型，应用于机器人规划、模型预测控制、四维生成、交互式仿真及其他下游具身任务。然而，对于当前的探索仍存在两个关键问题未回答：一，现有具身世界模型的生成泛化能力是否足以在多样场景中维持人类观察者眼中的视觉保真度；二，这类模型是否足够真实且可执行，能够作为现实世界具身智能体的通用先验模型。目前尚缺乏一套系统、标准化的框架对模型在具身任务上的能力进行评测。

针对上述问题，本文提出具身世界模型的图灵测试评测基准 EWM-TT。该基准基于 927 条机器人操作数据，从感知、规划、预测、泛化与执行五大核心能力对视频生成模型、世界基础模型与具身世界模型进行系统评估，并设计包含 22 项指标用于量化模型生成能力的综合评测方案。为了进一步评估模型生成结果在主观真实性和真实可执行性层面的可信程度，本文在此评测基准的基础上进行扩展，设计了两类图灵测试：一类是面向人类观察者的人类图灵测试；另一类是基于逆向动力学模型的机器图灵测试。两者分别衡量生成视频是否能够骗过人类让人类真假难分和是否能够诱导逆向动力学模型生成让真实机器人可执行的动作。

本文对三类模型分别进行评测，其中包括闭源的商业视频生成模型、开源的世界基础模型，还有具身世界模型。从实验得到的结果来看，评测基准的综合得分与人类偏好评分之间有较高的一致性，它们之间的皮尔逊相关系数达到了 0.97，这给人类图灵测试的可靠性奠定了可量化的基础。通过 EWM-TT 的实验分析可以发现，现有模型在长程任务规划上的得分较弱，最高分也只有 27.87；在物理规律一致性方面，表现最好的模型得分也仅为 73.29，这说明当前的具身世界模型在时空一致性和物理推理上还存在明显的局限性。更进一步，在机器图灵测试里，本文用逆向动力学模型把生成视频翻译出动作序列，在真实环境中去执行，计算执行成功率，结果发现大部分模型在真实执行任务里几乎完全失败，只有少数几个模型拿到了非零的成功率。其中，以 Wan2.1 作为骨干模型训练的具身世界模型 WoW-wan，它的真实执行成功率是 40.74%。这个结果也揭示了，当前具身世界模型在连接“可视世界”和“可执行世界”方面存在的鸿沟。

本研究给具身世界模型提供了一套统一的评测基准和双视角图灵测试框架，也给未来设计物理上更合理、规划推理更强、真正能执行的通用具身世界模型，提供了实验依据与优化方向。

关键词：具身智能，世界模型，评测基准，图灵测试