基于大模型的世界模型研究综述(特邀)

赵翔[1],黑梦哲[2],李家旭[1],庞宁[3],陈子阳[1]

(1. 国防科技大学大数据与决策国家级重点实验室,湖南 长沙 410005;
2. 国防科技大学信息系统工程国家重点实验室,湖南 长沙 410005;3. 空军航空大学,吉林 长春 130000)

摘 要: 一般认为,世界模型理解并表示外部世界,同时根据当前的世界状态和动作预测世界的未来状态。大模型依靠海量的训练数据和庞大的参数规模,拥有出众的文本知识学习、理解表示和生成能力,例如语言大模型 GPT-4、LLaMA 等。近年来,世界模型研究备受工业界和学术界的关注,涌现出了一大批包括自动驾驶、社会模拟、具身智能和视频生成的研究和商业成果,并且研究者将各类大模型的出色成果应用在世界模型上,使世界模型的效果得到了进一步提升。本文对利用大模型构建的各领域世界模型进行了全面综述,包括基于语言大模型和基于视觉大模型(VLM),并且选取了数个重要的应用领域对相关模型进行介绍,包括具身智能、智慧城市、社会模拟和物理环境模拟。本文首先基于大模型的模态对世界模型进行分类,指出了基于不同模态的世界模型在功能上的不同;随后给出了世界模型重要的开源资源和基准,帮助相关领域的研究人员快速了解和使用世界模型;最后对文章进行总结,并对未来研究方向进行展望。

关键词: 世界模型;大模型;生成式大模型;模拟;具身智能

中图分类号: TP18 **文献标志码:** A **DOI:** 10.19678/j.issn.1000-3428.0260356

Survey of World Models Based on Large Models (Invited)

ZHAO Xiang[1], HEI Mengzhe[2], LI Jiaxu[1], PANG Ning[3], CHEN Ziyang[1]

(1. National Key Laboratory of Big Data and Decision, National University of Defense, Changsha 410005, Hunan, China;
2. National Key Laboratory of Information Systems Engineering, National University of Defense, Changsha 410005, Hunan, China;
3. Aviation University of Air Force, Changchun 130000, Jilin, China)

【Abstract】 World models are generally believed to understand and represent the external world and predict future states based on current world states and actions. Large models leverage massive training data and vast parameter scales to exhibit outstanding capabilities in learning, understanding, representing, and generating textual knowledge, as exemplified by language large models such as GPT-4 and LLaMA. In recent years, research on world models has attracted significant attention from both industry and academia, leading to significant research and commercial achievements in domains such as autonomous driving, social simulation, embodied intelligence, and video generation. Moreover, researchers have applied the remarkable results of various large models to world models, further enhancing their performance. This paper comprehensively reviews world models built using large models across different domains, covering both language large model- and Vision Large Model (VLM)-based approaches. Several important application areas, including embodied intelligence, smart cities, social simulation, and physical environment simulation, are selected to introduce relevant models. This paper classifies world models based on the modality of the large models used, highlighting the functional differences between world models based on different modalities. Subsequently, important open-source resources and benchmarks for world models are presented to help researchers in related fields understand and utilize world models quickly. Finally, this paper is summarized and future research directions are presented.

【Key words】 world model; large model; generative large model; simulation; embodied intelligence

0 引言

人工智能领域长期致力于开发能够再现世界规律、动态和现象的通用模型。近年来,一种能够表示外部环境动态变化并预测未来世界的模型吸引了研究者的注意,这种模型被称为世界模型。借助世界模型,可对世界进行预测、规划和推理。目前,世界模型尚无明确、通用的定义,但是一般认为其功能是理解世界的当前状态,并依据状态和动作预测世界的未来状态[1-2]。在早期的世界模型工作中,研究者

基金项目: 国家自然科学基金(U25B2047, 62272469, 72501299)。

作者简介: 赵翔,男,教授、博士,主研方向为大数据知识工程;黑梦哲(共同一作),男,博士研究生,主研方向为世界模型、热点事件态势预测;李家旭,博士研究生;庞宁,讲师、博士;陈子阳,博士研究生。

收稿日期: 2026-03-20 **修回日期:** 2026-04-22 **E-mail:** xiangzhao@nudt.edu.cn

更聚焦于生成当前世界的表示[3]，在近期的工作中更聚焦于在感知理解的基础上对世界的未来状态进行预测[4-6]。世界模型的应用范围十分广泛，对感知理解和预测未来能力有需求的领域都会依据本领域的数据和任务构建类似功能的模型，例如在机器人领域，世界模型需要加强机器人对外部世界的感知以理解环境并做出决策[7-8]，在模拟社会环境中需要模拟社会环境和社会中人类的互动模式[9-10]等。

大模型领域的突破对世界模型的发展起到了重要推动作用，大模型凭借其庞大的训练数据和参数规模，可以学习到更丰富、更全面的知识以及规律和规则，因此许多世界模型领域的工作都基于大模型开展。为此，本文从世界模型和大模型的关系角度出发，对世界模型的最新进展、应用以及未来的方向进行系统性的回顾。

现有世界模型综述一般根据世界模型的应用场景来划分。DING 等[11]将世界模型按照生成外部世界的表示和预测未来状态的功能分为两类，并基于这种分类进行介绍。XU 等[1]将世界模型分为通用性世界模型和专用性世界模型，并基于各个领域功能和作用进行介绍。也有工作对特定领域的世界模型进行介绍，如自动驾驶领域[12-13]、三维(3D)和四维(4D)场景建模[14]等。与这些工作不同的是，本文抓住近年来世界模型快速发展的原因之一，即各类大模型的成果被应用在了世界模型中，并且依据世界模型使用的大模型的模态对其进行分类介绍，这种做法避免了对年代距离较远、功能性能较弱的世界模型过多介绍。除此之外，考虑到不同领域的世界模型存在诸多不同，本文还依据世界模型的重要应用场景进行介绍。本文主要贡献可概括为：1)基于世界模型使用的大模型模态进行分类，揭示世界模型与基于不同模态大模型之间的关系；2)基于该分类体系，梳理了世界模型的 4 个主要用途，即具身智能、社会模拟、智慧城市和物理环境模拟，展示了世界模型在不同领域的侧重；3)指明了世界模型未来的发展趋势。

本文后续内容组织如下：第 1 章介绍基于不同模态大模型的世界模型；第 2 章介绍世界模型的关键应用领域；第 3 章介绍目前世界模型领域中重要的开源工具和基准；第 4 章做出总结并且指出了未来世界模型的发展方向。本文的文献检索以 Google Scholar 和 IEEE Xplore 为主要数据库，并辅以 Web of Science、arXiv 及 Scopus 进行补充。检索关键词围绕世界模型、具身智能、社会模拟和自动驾驶等。初步检索共获得文献约 500 篇，随后根据相关性和实际研究内容进行筛选，最终纳入文献约 125 篇用于系统分析。

1 基于不同模态大模型的世界模型

大模型凭借其强大的学习能力，可高效、高质量学习不同领域的知识、规律和规则，因此许多世界模型领域的工作都基于大模型开展。图 1 展示了基于大模型的世界模型分类情况(彩色效果见《计算机工程》官网 HTML 版，下同)。本章以大模型的模态作为划分依据，介绍基于大模型的世界模型。

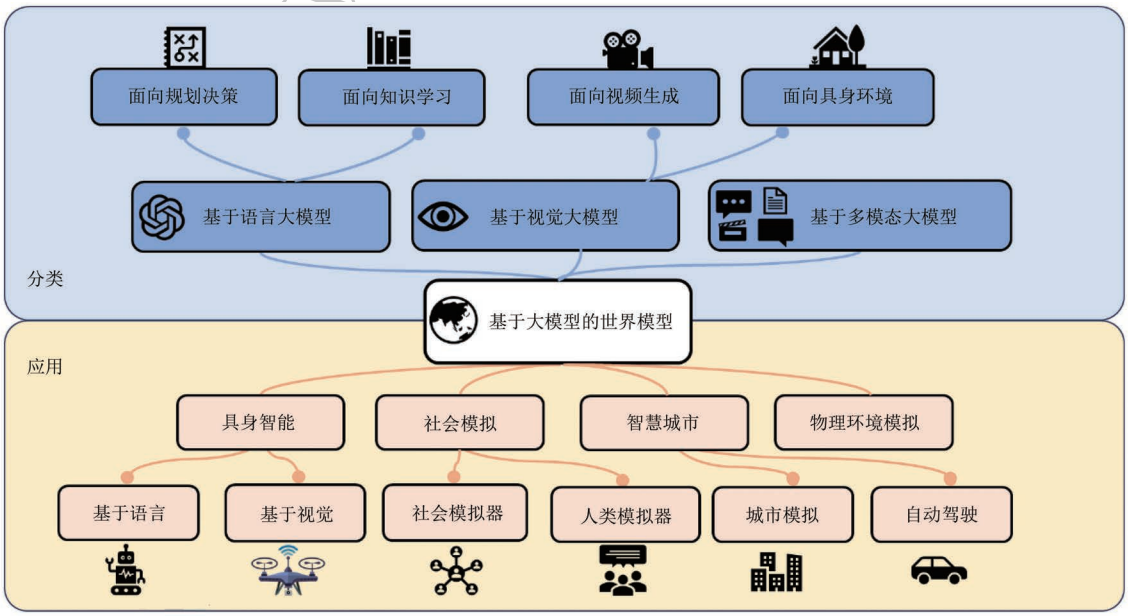

图 1 基于大模型的世界模型的分类与应用

Fig. 1 Classification and application of large model-based world models

1.1 基于语言大模型的世界模型

依靠海量文本数据训练的语言大模型可以更有效地学习世界的基本原理和规律，如 LLaMA、GPT 等。GURNEE 等[15]已经证明，语言大模型可以通过学习语料库中的数据习得跨越多个尺度的时空线性表示，进而掌握构成世界模型的基本要素，而不仅仅是浅层的文字表示。这种对世界知识和规律的理解能力使得语言大模型有潜力构建高效全面的符号化世界模型。本节将这些世界模型分为面向规划和决策与面向领域知识学习两种。

1.1.1 面向规划和决策的世界模型

语言作为通用信息传递和信息表示的媒介，基于语言大模型的世界模型在诸多决策任务中展现出了巨大潜力。同时，规划和决策也是语言大模型的重要应用之一，基于语言大模型的世界模型可以通过理解复杂的任务背景，生成完成任务的规划或者决策。本领域的代表性模型信息如表 1 所示。其中 RAP 模型[16]以 GPT-4 模型和树形推理结构为基础，在处理任务时随着推理步数的增加耗时在数秒到数分钟不等。

HU 等[2]指出相比于语言大模型，世界模型和智能体模型更符合人为决策的方式，因为它们包含诸如目标、背景知识等关键组分，语言大模型更适合作为后端支持整个决策过程，如图 2 所示。该团队具体分析了语言大模型如何支持此框架中的各个组分。其中最具代表性的就是直接将大语言模型（LLM）作为世界模型，从而进行推理和决策。HAO 等[16]提出的基于 GPT-4 的 RAP（Reasoning via Planning）推理框架与单纯提示大语言模型行动步骤的思维链（CoT）不同，该工作引导大模型预测采取某一步动作之后的状态，进而计算这步动作的奖励，从而制定策略完成推理。具体而言，该工作采用蒙特卡罗树搜索算法探索动作与状态空间。LONG 等[17]采取类似的做法基于 GPT 模型设计了一套类似专家会议的语言导航方案 DiscussNav，该模型能在每次移动前主动与语言大模型扮演的专家讨论以收集关键信息，完成指令理解、环境感知和完成度预估等核心导航子任务。ZHAO 等[18]进一步将语言大模型与开放词汇检测技术相融合，通过蒸馏导航数据中的关键信息建立多模态信号间的映射关系，构建导航中多模态信号与关键信息的关系，并且该工作提出全向图（Omni-graph）表示方法作为导航任务的世界模型，使得模型可以执行更精确的导航动作。YANG 等[19]利用语言大模型作为规划器主动探索环境，并且该模型能自适应评估并剔除不准确的感知描述，该工作还使用语言大模型绘制房间布局图，以此作为世界模型来增强全局环境理解能力。

表 1 代表性规划与决策模型概览

Table 1 Overview of representative planning and decision-making models

模型名称	年份与出处	任务场景	语言大模型用法	核心框架
LAW[2]	2023 arXiv	任务规划决策	大模型直接决策	LLM
RAP[16]	2023 EMNLP	任务规划决策	大模型直接决策	LLM、蒙特卡罗树
DiscussNav[17]	2024 ICRA	视觉语言导航	大模型直接决策	LLM
OVER-NAV[18]	2024 CVPR	导航	大模型直接决策	LLM、全向图
RILA[19]	2024 CVPR	视觉语言导航	大模型直接决策	LLM
WKM[20]	2024 NeurIPS	任务规划决策	作为世界知识模型	LLM
Webagent[21]	2025 arXiv	网页导航	生成训练数据	LLM

前文介绍的工作大多直接将大语言模型作为世界模型进行规划或决策，优点是简单易用，但是缺点在于对世界知识的感知理解不足，导致这些模型在全局规划中仍然存在盲目试错、在局部规划中产生幻觉行为等问题。为了解决这些问题，CHEN 等[20]受心智世界知识模型的启发，在模型执行任务时维护一个参数化的世界知识模型为智能体提供全局先验知识和局部动态知识，从而避免直接使用语言大模型的生成结果，有效缓解盲目试错和幻觉行为问题，并且提高了模型在未见任务上的泛化性。类似地，CHAE 等[21]注意到大语言模型不能准确预测在网页导航决策过程中每一步动作的结果，故开发了一种世界模型增强的智能体，并且提出一种以状态转移为中心的观测抽象方法来预测行动结果，解决了诸如购买不可退款机票等不可逆行动所带来的复杂问题。

有些研究则使用语言大模型生成决策环境和任务。HU 等[22]为了解决文本规划问题面临的人为设计的决策环境固定且耗费劳力问题，使用语言大模型为智能体自动生成灵活多样的决策环境，并且采用一种从易到难的任务设计生成方法加强基于语言大模型智能体的学习效率。WANG 等[23]研究如

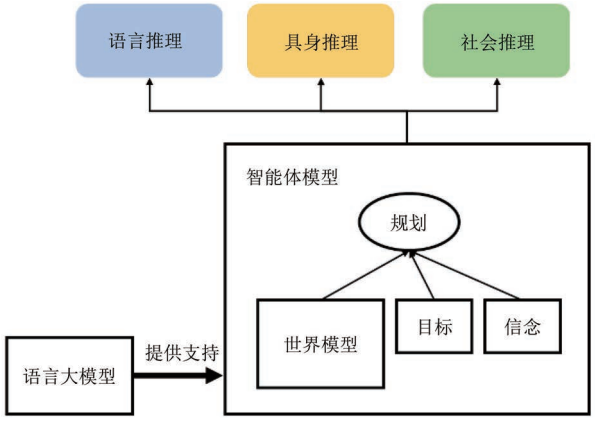

图 2 世界模型、语言大模型和智能体模型的关系[2]

Fig. 2 Relationship among world models, language large models, and agent models[2]

何让语言大模型生成可交互和可解释的世界模型，并将这个任务转化为语言大模型生成基于 Python 语言的文字游戏，同时提出了一个预定义的专注于推理任务的文本游戏语料库。这部分工作的评价指标主要是推理任务的正确率等。

1.1.2 面向领域知识学习的世界模型

基于语言大模型的世界模型也可用于学习大量的世界知识和常识，这些知识对模型高效解决多种场景下的真实世界任务来说至关重要。GURNEE 等[15]的工作首次证明了 LLaMA2 这样的语言大模型通过对海量文本数据的学习能习得对世界的时空认识。具体而言，该团队发现在 LLaMA2 模型[24]中存在独特的“空间神经元”和“时间神经元”分别学习跨多尺度空间与时间的线性表示。与注重学习通用知识和常识的大语言模型不同，世界模型更注重学习领域深度知识。部署此类模型往往需要对大模型进行微调，以 CityGPT 为例，模型参数为 7×10^9，实验在 4 块 A800 GPU 上进行，耗时为数小时到数天。

部分研究者聚焦使用语言大模型学习物理世界知识。FENG 等[25]为了解决大语言模型在处理城市环境中的真实地理空间任务时的不足，提出 CityGPT 这一城市级世界模型，通过在训练过程中加入物理世界知识及相关数据增强其对城市空间的理解，并提升其解决相关城市任务的能力。ROBERTS 等[26]基于 GPT-4 探索了语言大模型获取地理事实知识的程度及其运用该知识进行解释性推理的能力，并尝试使用语言大模型解决地理空间分析、供应链管理及灾害响应等任务。基于此类模型中的大范围地理空间知识，研究者可开展全球移动预测和地理定位等研究[27-28]。

除了大范围的地理空间知识以外，更细致的人类日常生活环境和常见的现实任务所组成的局部物理世界也包含丰富的世界知识，因此也需要世界模型对局部物理世界的规律和知识进行学习。JIN 等[29]发现代码语言模型可以学习到程序语义的涌现现象，即分类器能在训练过程中从语言模型隐藏状态中提取出愈发精确的世界状态表示，这表明该模型获得了形式意义上解释程序的能力。GAO 等[30]借鉴比较心理学与认知科学理论，提出两阶段评估框架，通过感知能力与预测能力实现对大模型作为世界模型的建模评估，实验表明大语言模型的方法在构建精确模型方面存在较大缺陷，如无法区分运动轨迹和缺乏解耦式的理解能力。以上介绍的模型更聚焦于静态的知识学习，因而无法处理更复杂、可交互的世界知识。为此，LIN 等[4]想让智能体模型理解包含世界状态、通用常识和交互反馈的复杂语言，例如“遥控器上的按钮可以改变电视的状态”，并且智能体可以用这些复杂语言作为预测未来的信号，即理解指令并且判断行动会如何改变世界的状态，以及如何获得奖励。基于此，该团队提出了 Dynalang 框架，智能体通过学习世界模型来预测未来表示，并且从中推演行动策略。

除了真实的物理环境以外，理解人类社会环境的规则和知识也是世界模型的一个重要部分，近期有许多研究讨论了该如何构建这种模仿人类心智或者人类社会的世界模型。CHEN 等[31]使用故事类基准测试，深入剖析了如何提升和评估语言大模型的心智能力。SAP 等[32]通过两项任务检测 GPT-3 理解并驾驭日常社交互动的能力和能否推断情境参与者的心理状态与现实认知。STRACHAN 等[33]注重考察大语言模型追踪他人心理状态的能力，通过模型在理解错误信念、解读隐含请求、识别反语与失礼行为等维度上的表现，测试了 GPT 系列和 LLaMA 系列的语言模型的能力，结果表明 GPT-4 和 LLaMA2 都在不同的方面超过了人类的表现水平。除了试图复现人类个体的心智能力以外，一些世界模型还尝试构建人类社会整体的规则。PARK 等[34]仿照《模拟人生》的理念创建了一种交互式沙盒环境，基于大语言模型构建了 25 个智能体，这些智能体可以在虚拟的小镇内互动，并且可以根据用户指令涌现出社会行为。BAKLASHKIN 等[35]考察语言大模型在复杂的环境中的伦理判断和决策，研究表明情绪可以显著影响大模型的伦理判断，强调建立机制保证伦理标准。VAYANI 等[36]为了构建全球化的社交模型，对涵盖百种语言的大模型进

行了评估,更彰显了文化与语言包容性的重要性。为提升语言大模型在复杂社交情景下的心智能力,也有研究者提出新的方法。WANG 等[23]提出了首个面向机器心智理论的认知知识图谱 COKE,该模型构建了 4.5 万多条人工验证的认知链集合,刻画了人类在特定社会情境中面临的心理活动及其后续行为反应,凭借这些认知链该团队构建出专为认知推理设计的生成模型 COLM,显著提高了模型的心智能力。

1.2 基于视觉大模型(VLM)的世界模型

视觉大模型与传统的单模态模型相比更注重视觉信息的处理,能够凭借多模态信息处理更复杂的任务。其中与世界模型关联最紧密的视觉大模型技术是视频生成和具身环境。

1.2.1 面向视频生成的世界模型

视频生成技术可直接捕捉真实世界中连续的时空动态,因此基于视频生成技术的世界模型可以处理更复杂的动态世界环境。最近的视频生成模型主要有 Sora[37]、Keling[38]和 Gen 系列模型[39],这些模型都可以根据输入的文本指令和图像数据生成高质量的视频。Sora[37]是一款大型视频生成模型,可以生成长达 1 min 的高质量、时序连贯且主体一致的视频。这些画面非常符合物理规律,比如正在流淌的水滴、反射的镜面和融化的物体。这表明 Sora 具备作为世界模型的潜力。简单地说,Sora 通过扩散变换器模型(Diffusion Transformer Model)[40]来处理输入数据和生成视频。在训练过程中,扩散 Transformer 通过学习输入的视觉数据的分布,将这些分布映射到低维空间,从而实现对视频的压缩和重构。随着 Sora 模型在视频生成领域的成功,近些年也出现了很多视频生成模型。ZHENG 等[41]开源了 Open-Sora,该模型可生成时长为 15 s、最高分辨率为 720 像素的视频内容,模型特别引入了时空扩散变换器(Spatial-Temporal Diffusion Transformer),这同样也是一种视频扩散框架,并且还采用了 3D 自编码器和定制化训练策略加快训练速度。YANG 等[42]开发了 GogVideoX,该模型能够生成时长为 10 s、分辨率达 768×1 360 像素的连续视频。该模型采用了专家自适应层归一化(LayerNorm)来确保生成的视频和文本内容对齐,并且使用了渐进式训练与多分辨率帧包技术,保证了视频的时长和结构连贯性。英伟达公司研发了 Cosmos[43]世界模型,与更加侧重视频生成质量的 Sora 相比,Cosmos 更加注重生成内容对物理规律的理解和遵守,该模型同样基于扩散变换器框架,可为机器人控制和智能驾驶生成宝贵的领域合成高保真视频数据。谷歌公司的 DeepMind 团队研发的 Genie2[44]和 Genie3[45]则更加注重游戏环境的生成和实时交互。本节在表 2 中列举了常见的视频生成模型的关键特征。值得注意的是,Sora 模型的参数规模约为 30×10^9,属于中等规模的扩散型 Transformer 视频生成模型。商用视频生成模型在预训练时耗时长、所需算力较高,在用户使用时生成一段 10 s 的视频就需要 3～5 min。

表 2 代表性视频生成模型概览

Table 2 Overview of representative video generation models

模型名称	年份	类型	时长	分辨率/像素	核心框架
Sora[37]	2024	视频生成	60 s	1 080	DiT
Keling[38]	2024	视频生成	120 s	1 080	3D VAE + DiT
Gen-3[39]	2024	视频生成	30 s	3 840×2 160	Multi-stage Diffusion Pipeline
Open-Sora[41]	2024	视频生成	15 s	720	Spatial-Temporal DiT
GogVideoX[42]	2025	视频生成	10 s	768×1 360	3D VAE
Cosmos[43]	2025	物理环境生成	25 s	1 280×704	VAE + DiT
Genie2[44]	2024	交互环境生成	数分钟	720	自回归的潜在扩散模型
Genie3[45]	2025	交互环境生成	数分钟	720	自回归的潜在扩散模型

尽管以 Sora 为代表的面向视频生成的世界模型取得了突破性的进展,这些模型能够生成高质量的模拟动态视频来体现对世界知识和规则的理解。然而,这些模型在很多方面仍存在局限性,所以很多研究工作聚焦于如何让这类模型的生成内容更加符合世界规律。DING 等[11]和 YIN 等[46]的工作聚焦于如何让模型生成的视频更长,前者使用分块环形注意力机制(Blockwise Ring Attention)来解决长序列数据带来的计算资源匮乏和上下文数据缺失的问题,后者提出了一种新的扩散式架构保证视频长度,即先由全局扩散模型生成全时域关键帧,再由局部扩散模型填充相邻帧间内容。BRUCE 等[47]和 YANG 等[48]的工作则强调提高世界模型的可交互性,前者通过无监督方式从无标签视频中训练生成

式交互环境，后者通过引入强化学习和上下文学习等技术提高模型的可交互性。还有一些研究旨在提高模型的泛化性，希望模型能够解决多个场景的任务[46,49]。YU 等[50]提出一种混合式空间记忆方法，以实现可靠的定位与目标检索，进而使得结果更符合相机位姿。

1.2.2 面向具身环境的世界模型

面向具身环境的世界模型旨在为智能体提供交互和训练的外部环境。与视频生成技术不同，具身环境不仅要模拟外部世界的视觉信息，还需要整合其他动态特性。本节在表 3 中列举了该领域的代表性模型。部署该领域模型往往涉及大模型训练，以 Aether 为例，该模型参数规模为 5×10^9，在 80 块 A100-80 GB GPU 上需要约两周。

表 3 代表性具身环境模型概览

Table 3 Overview of representative embodied environment models

模型名称	年份	场景	类型
Aether[5]	2025	室内、室外	动态
Matterport3D[51]	2017	室内	静态
UrbanWorld[52]	2024	室外	静态
MineDojo[53]	2022	室外	静态
UniSim[54]	2023	室内、室外	动态
EVA[55]	2024	室内	动态
Pandora[56]	2024	室内、室外	动态
TessrAct[57]	2025	室内	动态

该领域的工作从一开始提供静态、预设好的环境转向提供动态、开放的环境。早期工作 Matterport3D[51]主要通过 RGB 数据集生成室内场景，支持智能体进行多种训练方法。但是当需要模拟室外场景时，模型需要面对更大的数据规模和更多的环境动态性。SHANG 等[52]提出的 UrbanWorld 则提出了一种基于渐进式扩散的渲染方法，将其用于生成高质量的 3D 城市环境，并且还设计了一种基于真实街景加文本的城市环境大模型，用于监督和引导扩散模型的生成过程。ANANDKUMAR 等[53]提出的 MineDojo 则更注重虚拟开放环境的生成，利用视觉大模型为虚拟环境的智能体设计奖励函数，以支持智能体的各种任务。

以上方法只能提供静态的具身环境，无法支持智能体在更复杂、更灵活的任务场景下进行训练。随着生成式模型的进展，更多研究者使用生成式大模型创建更灵活、更动态的环境。YANG 等[54]提出的 UniSim 探索了如何通过生成式建模来学习通用的现实交互模拟器，该模型利用海量的人类活动视频、机器人数据、全景扫描图以及互联网图文数据，能够为智能体生成各种场景下的逼真交互体验。DENG 等[58]基于视频扩散模型构建长序列回归框架，并且利用新型时间差值方法防止模型失真，模型在谷歌街景数据集上训练，最终获得可以任意调整视角的街景世界模型。但是，单纯通过基于生成式模型构建的世界模型往往面临泛化能力不佳的问题，往往只能生成预设条件下的具身环境，为此有些研究者开始注重提升模型的泛化能力。XIANG 等[56]提出的 Pandora 模型使用混合的自回归加扩散模型框架，通过生成视频模拟世界状态，并支持文本指令实时操控。通过大规模预训练与指令微调，Pandora 模型实现了领域泛化性、视频一致性与可控性。CHI 等[55]为了加强环境视频生成模型的分布外泛化能力，采用视频生成模型加视觉语言大模型的互补做法构建具身视频预测世界模型，实验表明该模型在机器人控制等下游任务上效果较好。ZHU 等[5]认为过去的模型缺乏从合成数据到真实数据的零样本泛化能力，并且难以结合几何重建和生成建模，为此该团队提出名为 Aether 的世界模型，该模型可利用 4D 数据和相机轨迹数据建模几何知识，再通过视觉大模型实现基于动作的预测和视觉规划，可在零样本的条件下完成 4D 场景生成、条件化动作预测和视觉规划等任务。

还有一些研究者通过提出不同的约束条件来使生成结果更符合现实世界。RIGTER 等[59]基于 UniSim 模型的框架，在标注了动作的小领域专用数据集上实现交互模拟器，并且该模型使用一种可学的掩码修改预训练模型的中间输出，从而得到动作条件确定的视频。ZHEN 等[57]提出的 TessrAct 模型除了使用扩散模型生成世界场景以外，还利用地图来约束模型的生成结果，使得模型结果与现实世界景观更接近。

1.3 基于其他多模态大模型的世界模型

除了上述的两种大模型以外，还有一些多模态大模型使用文本、视频和音频等模态信息作为训练输入，使模型能更好地应对复杂场景下的任务，处理并整合视觉、文本及其他感官模态的信息，构建更丰富全面的世界表征，拓展世界模型的建模潜力[60-64]。厘清这些模型如何处理、表征和应用多模态知识和世界知识，能帮助构建高效的世界模型。其中具有代表性的模型如表 4 所示。

表 4 基于其他模态大模型的世界模型概览

Table 4 Overview of world models based on large models of other modalities

模型名称	类型	包含的模态
WorldGPT[61]	预测世界未来状态	文本、图片、音频、视频
Finsta[64]	文本、视频时空对齐	文本、视频
AVLEN[65]	多模态路径导航	文本、语音、位置信息
iGibson[66]	交互场景生成	RGB、空间信息、激光雷达信号
MUVO[67]	自动驾驶	视频、激光雷达信号

有些研究聚焦于能够处理多模态信息的世界模型。GE 等[61]为了解决世界模型支持的模态数量少的问题，提出 WorldGPT 这一世界模型，该模型采用渐进式状态转换训练(Progressive State Transition Training)方法，将不同模态的信息统一起来，从而能够处理多模态数据。该模型可以处理来自文本、视频、图片以及音频模态的数据，提高了领域外预测的能力，并且更加符合真实世界包含多种模态信息的特点。FEI 等[64]为了解决多模态世界模型的模态对齐、视角割裂等问题，提出了基于精细化的结构时空对齐学习方法来强化大模型。除此之外，有些与环境模拟的世界模型需要接收激光雷达信号。CHERIAN 等[65]提出了名为 AVLEN 的多模态导航世界模型，该模型下的智能体能够接收人类发出的语音、文本和定位等信息。SHEN 等[66]提出的 iGibson 模型通过整合 RGB、深度和激光雷达信号等模态的信息可为真实场景交互智能体提供仿真环境。BOGDOLL 等[67]提出的 MUVO 多模态世界模型结合摄像头提供的视觉数据和激光雷达数据与 3D 占用预测，为自动驾驶任务提供了更优的模型支持。

2 世界模型的应用

世界模型有较强的领域性，即各个领域为了满足本领域中感知环境状态和预测未来的需求都会设计各自的世界模型，本章从多个比较突出的应用领域出发介绍世界模型的应用情况。

2.1 具身智能

具身智能是人工智能与机器人学交叉的前沿领域，强调智能体对外部环境的感知、理解和交互。显而易见，如何让智能体或者机器人感知理解外部环境，并且稳定做出规划决策完成任务是该领域的核心挑战。世界模型感知环境与预测未来的特性与具身智能的要求不谋而合，可赋予智能体急需的感知力和预测能力。具体而言，具身智能模型在着重解决任务规划时，更关注世界模型的感知能力，而在构建智能体环境和预测外部世界动态变化时，更关注世界模型的预测能力。本节依旧从模态的角度出发，对该领域的工作进行介绍。

2.1.1 基于语言大模型的具身智能模型

在具身智能领域的工作中，语言大模型被运用在策略学习、任务规划等任务中。WANG 等[68]提出的 GenSim 为了解决策略学习模拟数据少的问题，给 GPT-4 模型提供设定的智能体任务，并且引导大模型渐进式地从过去的任务中生成新任务，以达到为智能体生成新颖灵活的环境的目的。LIN 等[69]提出的 Text2Motion 尝试让智能体理解耗时较长的推理任务，该模型以自然语言为输入，构建了任务级加动作级的计划框架，引导大型语言模型进行任务规划，并且该模型在规划过程中通过执行几何可行性规划，主动解决跨越技能序列的几何依赖关系。WANG 等[70]为了解决智能体在开放环境下执行复杂任务时遇到的多子任务顺序排列问题，提出了一种基于语言大模型的可交互的规划方法，该模型可通过分析计划执行过程，在扩展计划阶段对遭遇的失败做出反馈，从而使大模型生成计划错误的修正。MAO 等[71]为了加强机器人等智能体在复杂 3D 场景中的空间推理能力，提出了一种名为 SpatialLM 的 3D 大型语言模型，该模型旨在处理 3D 点云数据并生成结构化的场景理解输出，包括建筑元素和定向物体边界，可有效提高对 3D 数据的理解能力。

以上都是直接利用语言大模型加强智能体表现的工作，它们往往会面临泛化性差，容易被大模型幻觉误导的问题。另一种做法是尝试将大语言模型融入决策框架中。LIU 等[72]提出的 LLM+P 框架将经典规划器融入语言大模型中，LLM+P 接收规划问题的自然语言描述，随后以自然语言形式返回解决该问题的正确方案，以解决机器人的长时域规划问题。GUAN 等[73]为解决语言大模型作为规划器时面临的正确率有限、无法有效利用反馈等问题，提出了一种新颖的范式，即在规划领域定义语言(PDDL)中构建显式的世界模型，然后使用该模型通过健全的领域无关规划器进行规划。

2.1.2 基于视觉大模型的具身智能模型

基于视觉大模型的具身智能世界模型常被用在预测外部环境的未来状态上，也有研究将重点放在真实世界模拟上。

由于视觉大模型的突破性研究，具身智能领域有很多工作基于视频预测方法进行外部环境预测。ABBEEL 等[74]为了给智能体设计复杂行为中的奖

励信号，提出视频预测奖励（VIPER）算法，该算法利用预训练视频预测模型的似然值作为强化学习智能体的奖励信号。该模型凭借大模型的泛化能力可为分布外的环境生成奖励，提高了跨实体的泛化能力。HU 等[7]认为视频扩散模型能够生成兼具当前静态信息与预测未来动态的视觉表征，并提出视频预测策略（VPP），该策略在视频扩散模型中学习基于预测未来表示的隐式逆动力学模型，以提高视频预测的效果。

以上模型由于模型建模能力不足，往往在未见任务下表现不佳，而为提升模型的泛化能力，许多研究者开始加大模型规模或改变建模方式。CHEANG 等[75]提出的 GR-2 模型在预训练阶段使用了 3 800 万个视频片段和超过 500 亿个词元作为输入，在后续的策略学习过程中能处理多种机器人任务和环境，并且可以生成机器人轨迹视频并支持动作预测微调。ZHOU 等[76]为了提升机器人控制场景下世界模型的泛化能力，提出 RoboDreamer 模型，通过分解视频生成过程来学习组合式世界模型，模型将指令解析为一组低级原始操作，并通过多模型协同生成视频。ASSRAN 等[8]提出一种基于自监督方法的 V-JEPA2，利用联合嵌入预测架构在包含逾百万小时互联网视频的图像视频数据集上进行预训练，模型在动作理解、人类动作预测和视频问答等任务上实现了较好的效果，显著提升了世界模型对物理世界的感知能力。ABBEEL 等[77]将序列决策问题转化为文本条件下的视频生成问题，在生成文本条件下的视频之后，从生成的视频中提取控制动作。该工作进一步将不同状态和动作空间的环境统一表示为图像空间，可实现跨多种机器人操作任务的学习与泛化。

还有一些研究使用世界模型作为环境模拟器，为智能体模型提供学习策略的环境。ZHU 等[78]提出的 IRASim 可为实体机器人提供交互式模拟器，该模型利用视频生成式模型，设计了视频级别和帧级别的适配器，为特定动作轨迹的机械臂生成高保真视频。SHANG 等[79]聚焦于提升世界模型的物理感知能力，为此提出一种物理感知世界模型 RoboScape，该模型可以通过框架设计同时学习 RGB 物理生成和物理知识，并且设计了时序深度预测和动力学学习两个联合训练任务，提高了世界模型对物理属性的感知能力。

除此之外，MAJUMDAR 等[80]为了测试智能体的具身智能水平，提出了名为 OpenEQA 场景开放问答数据集，该数据集包含来自 180 多个真实环境的人工生成的 1 600 个问题。

2.2 社会模拟

社会模拟是指利用世界模型模拟社会群体所处的社会环境，并且可以通过模拟社会个体之间的活动实现预测未来可能的社会状态。社会模拟在早期主要依靠计算机仿真完成，而传统方法中的个体和环境都不可避免地较为单调和相似，无法完全模拟多样、异构并且广泛联系的真实社会和个体。随着语言大模型和智能体模型的出现，构建更贴近真实的大型社会模拟器成为可能。当社会模拟模型需要学习人类社会心智知识或社会个体交互规律时，模型更关注世界模型的感知、表示能力；当预测社会动态变化或个体意见时，模型更关注的是世界模型的预测能力。

2.2.1 面向社会模拟器的世界模型

随着语言大模型和智能体模型的飞速发展，构建大型逼真的社会模拟器成为可能。PARK 等[34]在一个沙盒环境中构建了 25 个代表居民的智能体，智能体基于语言大模型可以相互交互，研究者发现这个仅有几十个人的小镇可以依据用户的行为涌现出有趣的社会行为，如用户在小镇中发出举行派对的邀请，则智能体会在数日之后如期举行派对。这样的一个早期的社会模拟器证明了大模型在模拟社会环境上的潜力。有些研究者从社会网络着手研究社会模拟器。GAO 等[81]构建了社交网络模拟系统 S^3，该模型使用基于智能体模型和语言大模型的微调和提示技术，保证智能体对用户的情绪、态度和交互行为的高保真模拟，基于此模拟系统可观察出群体层面的现象涌现效应，包括信息传播、态度传播和情绪传播。PAPACHRISTOU 等[82]通过研究社会网络中智能体的自发组成行为来模拟人类社会中的自组织行为，该模型基于语言大模型添加了贴合真实世界的设置，如智能体之间的关系、雇佣行为和联系等。借助智能体模型和语言大模型，有些研究从对社会中微观现象的研究转向宏观现象的研究，如关注宏观经济趋势的 EconAgent[83]、关注政策影响的 SRAP-Agent[84]和关注群体对税收政策反应的 Project Sid[85]。

最近的工作则向智能体更多、更多样、更大规模的社会模拟器发展。PIAO 等[86]提出了名为 AgentSociety 的大型社会模拟器，该模拟器包含了基于语言大模型的智能体、逼真的社会环境及强大的大规模模拟引擎，该模拟器为万名智能体生成行动轨迹，并且模拟了数百万次交互，基于此系统可进行大量社会预测。ZHANG 等[9]提出的名为

SocioVerse 的社会模拟器基于海量真实的社交网络平台用户数据，并为每一个智能体设定了不同的智能体信息，还可以自定义社会背景，从而支持不同社会背景下的社会实验。基于该系统可以以非常低的成本进行大规模的模拟问卷调查等经典社会实验。WANG 等[10]提出了一款名为“玉兰-万象”的社会模拟器，基于语言大模型研究者设计了涵盖面极大的模拟社会环境，包括经济学、社会学、政治学、心理学等，为不同领域的社会研究者提供低成本的社会实验机会。该模型还可以接受外部反馈，以自动优化模型以提高模型精度。

总之，这些工作表明基于语言大模型的世界模型可以生成对外部社会环境的高质量模拟，基于这些模拟器可以进行低成本的社会科学研究。

2.2.2 面向人类模拟器的世界模型

在社会模拟领域，研制出可像人类一样进行决策、广泛参与社会交互的智能体一直是巨大的挑战。利用语言大模型所学习的社会知识与常识，研制出人类级别的智能体正变得越来越可行。

QIAN 等[87]为了模拟人类学习和积累经验的过程，提出了一种名为“探究-整合-利用”(ICE)的新型策略，该策略可通过跨任务自我进化来增强智能体的适应性和灵活性。ZHANG 等[88]探讨语言大模型能否模拟人类在合作协同方面的特性，为每一个智能体设定了个性化特征来模拟真实环境下人类的特质和思维模式，并按照此类特点展开智能体之间的合作。实验发现智能体展现出从众、达成共识等人性化社会行为。ZHANG 等[89]研发的 Agent-pro 聚焦于让智能体学习人类参加大型交互性活动的能力，该模型能使智能体通过生成动态的信念不断反思，实现策略的进化，实验表明 Agent-pro 在两类牌类游戏中表现出色。

总之，以上模型通过模拟社会人的某个方面，学习人类在社会环境中的互动和生成行动决策的过程。

2.3 智慧城市

城市是一个由经济、社会和民生等要素组成的复杂巨系统，因而城市规划的相关任务总是面临滞后性、效率低和片面化的问题。随着大模型和世界模型技术的进步，世界模型可为城市生成高保真的环境模拟，并预测城市未来状态，进一步可进行灾害救援规划、城市交通规划、绿色城市建设和智能驾驶等高效的城市规划任务[90-91]。一般而言，城市建模中的城市级知识学习和建模更关注世界模型的表示和感知能力，而自动驾驶模型更关注世界模型的预测能力。

2.3.1 面向城市模拟的世界模型

面向城市模拟的世界模型旨在生成城市的模拟仿真场，为下游城市规划任务提供学习环境。有些模型使用基于渲染和随机噪声的方法生成城市景象，如 CHEN 等[92]研发的 SceneDreamer 基于大量的 2D 图像知识渲染器和单纯形噪声的高效鸟瞰视图(BEV)表示，可生成逼真的模拟常见场景，并且从随机噪声中合成大规模 3D 景观。LIN 等[93]提出的 InfiniCity 框架可构建并渲染出无限制规模的 3D 城市环境。InfiniCity 利用二维(2D)与 3D 数据，先从鸟瞰视角生成任意尺度的 2D 地图，然后基于八叉树的体素补全模块将生成的 2D 地图提升为 3D 八叉树结构，最后赋予纹理并渲染 2D 图像。CityDreamer 模型为了解决城市中存在的复杂结构，如建筑和车辆，将城市中的动态物体和静态物体分开，然后分别使用神经场渲染进行渲染[94]。

受限于模型的建模能力，以上模型往往只能进行 2D 建模，而随着视频生成模型的发展，很多工作也使用生成式模型进行场景生成，如 CityGen[95]，该框架引入无限扩展模块将局部布局扩展至城市级布局，并通过多尺度模块实现布局的超采样与精细化处理，最后基于扩散模型生成端到端的 3D 城市建模。UrbanWorld 模型[52]使用渐进式扩散渲染方法，还有辅助生成的城市专用语言大模型，可以自动生成高质量可交互的 3D 模拟城市，同时支持灵活的控制条件。FENG 等[96]利用 GPS 信号控制视频的生成，训练了从 GPS 信号到图像的生成模型，并且同样使用了扩散模型，使模型可以兼顾 GPS 和文本信息的指导，该模型可用于生成街区、公园和地标的独特外观特征。除了依靠视觉模型的场景生成工作以外，还有依靠语言大模型学习城市地理知识的 CityGPT[25]和 GeoLLM[97]。

总之，这些模型通过各类模型模拟了城市环境的外部特征，为下游的城市规划任务提供了良好的实验环境。

2.3.2 面向自动驾驶的世界模型

自动驾驶技术十分需要一种能预测未来世界状态并评估其影响的能力，这对驾驶安全和行驶效率至关重要。面向自动驾驶的世界模型注重当前状态感知和未来状态预测，整合并解读来自自动驾驶系统传感器的各种信息，从而预测车辆的未来场景并做出适当的规划[12]。该领域的世界模型可依据输入的形式分为基于图像、基于占据栅格、基于鸟瞰图和基于点云，有些工作也会在训练时加入驾驶员的操作数据和激光雷达信号；按照模型框架可以分为基于 Transformer、基于扩散模型、基于语言大模型或多模

态大模型等;按照功能可分为用于场景感知理解、用于行驶道路规划、用于车辆运动预测和用于驾驶控制[98]。表 5 给出了代表性的自动驾驶模型的核心信息。

表 5 代表性自动驾驶模型概览

Table 5 Overview of representative autonomous driving models

模型名称	年份	核心框架	输入形式	输出形式
DriveDreamer[6]	2024	扩散模型	文本、图像、HD 地图、边界框和动作	图像和动作
WorldDreamer[49]	2024	语言大模型	文本、视频、图像和动作	图像
MUVO[67]	2025	Transformer	图像、激光雷达信号和动作	图像和占据栅格
DriveDreamer-2[99]	2025	语言大模型+扩散模型	文本、HD 地图和边界框	图像
DriveDreamer4D[100]	2025	扩散模型	文本、图像和轨迹	图像
Drive-WM[101]	2024	扩散模型	文本、图像、地图、边界框和动作	图像
Vista[102]	2024	扩散模型	图像和动作	图像
GAIA-1[103]	2023	Transformer	文本、图像和动作	图像
GAIA-2[104]	2025	扩散模型	文本、图像和动作	图像
OccWorld[105]	2024	Transformer	占据栅格和轨迹	占据栅格和轨迹
OccSora[106]	2024	扩散模型	占据栅格和轨迹	占据栅格

DriveDreamer[6]将 Dreamer 系列模型扩展到自动驾驶领域,该模型为了生成贴近真实的驾驶场景,基于扩散模型设计了两阶段训练任务,首先让模型深度理解结构化交通规则约束,然后训练其预测未来状态的能力,最终该模型可在交通密集的环境下生成可控视频。但是,该模型输出的 2D 视频时空一致性较差,后续的 DriveDreamer-2[99]模型引入了语言大模型的提示机制,在原本模型的基础上提升了视频的一致性,并且加强了交互性。后续随着建模能力的提升,DriveDreamer4D[100]可生成时空一致的 4D 驾驶视频。WorldDreamer[49]则更关注对通用世界物理特性与运动规律的全面理解,该模型将世界建模转化为无监督视觉序列建模问题,采用掩码标记预测进行多模态生成,拓展了世界模型的适用性。Dreamer 系列模型的发展呈现出从 2D 向多维视频、从特定任务到通用任务和从人为设置条件到开放条件的转变规律。

Drive-WM[101]和 Vista[102]是两类聚焦高保真时空建模的世界模型,它们使用扩散模型,并使用更长的时间跨度和更高的空间分辨率,显著提升了输出视频的连贯性和一致性。Drive-WM 同样基于扩散模型,可生成多个视角的高保真视频,并且根据图像奖励机制确定最优驾驶轨迹,提高驾驶的安全性[101]。Vista 为了提升未来驾驶视频的准确性和可控性,设计了两种新型损失函数以促进移动车辆和结构信息的学习,并且该模型将历史帧作为先验信息输入,延长了视频长度,还添加了汽车操控集来学习驾车的不同操控策略[102]。GAIA-1[103]和 GAIA-2[104]属于大规模的预训练框架,着眼于构建功能级别的生成式世界模型,GAIA-1 模型着眼于对车辆附近景观的理解和生成精细的未来景象,在由景观图像数据、汽车操控数据和文本数据组成的数据集上,使用自回归预测和视频生成模型生成未来驾驶场景的视频[103]。与 GAIA-1 相比,GAIA-2 使用了潜在的扩散模型,大大提升了生成视频场景的丰富性和真实性[104]。除此之外,VectorWorld[107]为了避免生成式世界模型在闭环中效果下降,在推演过程中增量式地生成以自车为中心的"车道-智能体"矢量图块,并通过一个门控 VAE 对齐历史策略信息,进而提高模型的保真度和稳定性。

除了基于视频、图像输入生成视频的模型以外,还有一些研究通过激光雷达信号[67]、3D 占用信号[106]等其他模态的信息进行世界模型构建。ZHENG 等[105]提出的 OccWorld 利用 3D 占用数据和激光雷达数据学习一种基于重建的场景分词器,以获得离散的场景词元,然后使用类 GPT 的时空生成式变换器解码未来空间占用与车辆轨迹。OccSora[106]获取行驶场景中的 4D 场景的时空表示,用于训练扩散变换器,并且结合轨迹提示生成 4D 占用数据,实现了符合真实驾驶的 3D 布局的视频。世界模型除了可以模拟驾驶场景以外,还可以发挥领域学习能力,以及任务规划和控制的作用。CIPOLLA 等[108]提出基于模型的模仿学习方法,该方法从离线数据集中学习高度紧凑的潜在空间,达到同时学习场景世界模型和自动驾驶策略的目的,因此可以构建车辆驾驶环境中的静态与动态环境,并可以生成行车方案。

总之,在自动驾驶领域,世界模型发挥了其理解环境和预测状态的功能,提高了自动驾驶模型的模拟精度和制定决策的准确性。

2.4 物理环境模拟

面向物理的世界模型强调对物理世界规律或定律的高度遵循,注重可解释性、可靠性和高保真性。此类模型往往直接通过物理定理建立世界模型,适用于物理工程领域的高精度场景建模。此时,模型需要世界模型的表示学习能力和对未来物理过程的预测能力。

对特定物理领域,模拟器往往应用于面向工程的仿真,这是因为进行物理实验成本高昂,并且高度危险。对于可微分的物理过程,模拟工具采用偏微分方程对物理过程建模,并通过数学方法对过程进行求解。如有限元分析工具 Ansys、Abaqus、COMSOL 和 FEniCS[109]等,可对结构变形、断裂力学、传热、流体流动等物理场进行精确模拟。近些年,英伟达公司开发了 Omniverse,使用 GPU 加速数值模拟,证明了经典数学物理工具也可与 GPU 结合,共同解决物理环境模拟问题。然而,此类方法无法解决动态场景或者信息不足的情形,而且此类方法往往需要很多计算资源才能满足复杂状态下的需求。为了解决这些问题,研究者提出了通用性模拟方法,此类方法不追求高保真度,而是更强调效率和通用性,例如,为机器人领域设计的 PyBullet[110]、基于模型控制的 MuJoCo[111]等。

近年来,可微分仿真技术为物理环境模拟提供了更高效的途径,该方法可使梯度贯穿仿真过程以实现高效模拟。HU 等[112]提出高性能可微物理仿真编程语言 DiffTaichi,通过保留算术密集度和并行性的源代码变换生成仿真步骤的梯度,实现端到端反向传播,实验表明该语言在 10 种物理过程模拟器上的性能和生产力比现有的研究更加优异。该团队针对可变形物体的仿真,提出基于移动最小二乘材料点法(MLS-MPM)的实时可微分混合拉格朗日-欧拉物理仿真器 ChainQueen,可实现带碰撞的柔性物体仿真,并能无缝集成至软体机器人系统。还有模型将物理先验知识融入学习过程,例如,BILLARD 等[113]融合机器人任务场景的物理知识,为机器人设计了兼顾针对性、泛化性与可移植性的机器人平台算法,并且设计了涵盖面广、可持续更新的面向机器人任务的大型数据集。还有研究通过压缩精度以实现规模更大的模拟,例如,LIU 等[114]基于用户给出的误差上限和内存压缩率生成压缩方案,该方案利用自动微分技术估算每次压缩操作对总误差的贡献,进而将该任务转化为约束优化问题,通过线性化目标函数推导的解析公式进行高效求解,显著提高了内存压缩率。

除了以上提到的模拟方法以外,物理环境模拟方法还包括基于数据驱动和基于物理数据的混合方法等。总之,这类强调对物理世界规则遵循的世界模型为各类工程研究提供了理想的实验平台。

3 世界模型领域的测评集和资源

由于世界模型的应用场景太过广泛,且不同领域构建世界模型的方法和结构各不相同,因此没有通用的世界模型效果验证方法,目前也不存在各个领域都表现良好的通用世界模型。本章将介绍世界模型领域中重要的测评集和学习资源。

3.1 世界模型基准

本节从本文介绍过的部分世界模型功能出发介绍世界模型的测评集。

在领域知识学习方面,有测评集聚焦测试模型的学习能力,EWoK[62]主要用于测试模型对世界建模所涉及的概念性知识的理解程度,如社会互动和空间关系,包括 4 374 个题目且覆盖 11 个世界知识领域。KoLA[63]仿照人类认知模型构建了一个涵盖 19 个任务的四级知识相关能力的分类系统用以测评大模型的建模能力。还有测评集重点测试模型在多元文化背景下的跨语言能力[36,115]、地理知识学习能力[26,36]、社会规则和知识学习能力[33,35,116]。

在基于视觉大模型的世界模型的应用中,测评集主要测评模型的视觉感知和预测能力[30,117]、视频生成能力[118]、对世界的模拟能力[119-122]。根据世界模型的应用领域,各个领域有各自的测评集,包括具身智能领域[11,79,123]、城市模拟领域[124-125]、自动驾驶领域[126-127],如表 6 所示。

表 6 世界模型测评集

Table 6 Evaluation sets of world models

应用领域	评价内容	相关文献
领域知识学习	知识学习能力	文献[62-63]
	跨语言能力	文献[36,115]
	地理知识学习能力	文献[26]
	社会规则学习能力	文献[33,35,116]
视频生成	感知和预测能力	文献[30,117]
	视频生成能力	文献[118]
	世界模拟能力	文献[119-122]
具身智能	机器人策略规划能力	文献[11,79,123]
城市模拟	不同粒度的城市模拟能力	文献[124-125]
自动驾驶	路线规划能力 车辆操作规划能力	文献[126-127]

总之，世界模型的测评集是高度领域化的，并且缺乏通用性。各个测评集只能测试相似的功能，而不能通用地测试某类世界模型的好坏。例如，自动驾驶测评集中存在基于占据栅格和基于图像的测评集，使用占据栅格的模型不能在基于图像的测评集上测试，也不能在社会模拟的测评集上测评。

3.2 学习资源

为帮助读者快速熟悉世界模型，本节介绍一些学习资源和可以直接使用的世界模型平台，如表 7 所示。

表 7 世界模型学习资源

Table 7 World model learning resources

名称	类型	来源
I-JEPA	图像世界模型	https://github.com/facebookresearch/ijepa
V-JEPA	视频世界模型	https://github.com/facebookresearch/jepa
CWM	代码世界模型	https://github.com/facebookresearch/cwm
Spark	3D 场景生成	https://github.com/sparkjsdev/spark
WorldGen	3D 场景生成	https://github.com/ZiYang-xie/WorldGen
SocioVerse	社会模拟器	https://github.com/FudanDISC/SocioVerse
玉兰-万象	社会模拟器	https://github.com/RUC-GSAI/YuLan-OneSim
混元世界	3D 场景生成	https://github.com/Tencent-Hunyuan/HunyuanWorld-1.0
Matrix-Game 2.0	视频生成	https://github.com/SkyworkAI/Matrix-Game

4 结束语

综上所述，开发能够理解当前环境并对未来做出准确预测的世界模型将一直是相关领域研究者的目标。本文从世界模型使用的大模型模态出发，对借助了大模型的世界模型进行了详尽介绍，还分析了多个重要应用领域的情况，最后详述了世界模型的测评集和学习资源，可为刚接触世界模型的研究者和学者提供帮助。

未来的世界模型可能会与神经符号大模型相结合，增强对世界规则与知识的表示及学习能力，从而对未来做出更准确的预测。同时，未来的世界模型会向更加通用的方向发展，这是由于大模型的表示能力更强，可以将区别较小的领域结合，先实现在数个领域通用的世界模型，再实现较为通用的底座世界模型。此外，未来的世界模型也可能会更关注技术应用的落地，使世界模型更加易用高效，也更关注用户的隐私和数据安全。

参考文献

[1] XU K, ZHAO H, HU R, et al. From specialist to generalist: a comprehensive survey on world models[EB/OL]. [2026-02-27]. https://zenodo.org/records/18050668.

[2] HU Z T, SHU T M. Language models, agent models, and world models: the LAW for machine reasoning and planning[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2312.05230.

[3] HA D, SCHMIDHUBER J. World models[EB/OL]. [2026-02-27]. https://arxiv.org/abs/1803.10122.

[4] LIN J, DU Y Q, WATKINS O, et al. Learning to model the world with language[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2308.01399.

[5] ZHU H, WANG Y, ZHOU J, et al. Aether: geometric-aware unified world modeling[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Washington D.C., USA: IEEE Press, 2025: 8535-8546.

[6] WANG X F, ZHU Z, HUANG G, et al. DriveDreamer: towards real-world-drive world models for autonomous driving[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2024: 55-72.

[7] HU Y C, GUO Y J, WANG P C, et al. Video prediction policy: a generalist robot policy with predictive visual representations[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2412.14803.

[8] ASSRAN M, BARDES A, FAN D, et al. V-JEPA2: self-supervised video models enable understanding, prediction and planning[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2506.09985.

[9] ZHANG X N, LIN J Y, MOU X Y, et al. SocioVerse: a world model for social simulation powered by LLM agents and a pool of 10 million real-world users[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2504.10157.

[10] WANG L, GAO H Y, BO X H, et al. YuLan-OneSim: towards the next generation of social simulator with large language models[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2505.07581.

[11] DING J T, ZHANG Y K, SHANG Y, et al. Understanding world or predicting future? A comprehensive survey of world models[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2411.14499.

[12] GUAN Y C, LIAO H C, LI Z N, et al. World models for autonomous driving: an initial survey[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2403.02622.

[13] TU S F, ZHOU X, LIANG D K, et al. The role of world models in shaping autonomous driving: a comprehensive survey[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2502.10498.

[14] KONG L D, YANG W, MEI J B, et al. 3D and 4D world modeling: a survey[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2509.07996.

[15] GURNEE W, TEGMARK M. Language models represent space and time[EB/OL]. [2026-02-27]. https://www.

semanticscholar.org/paper/Language-Models-Represent-Space-and-Time-Gurnee-Tegmark/740c783ac07039cf30b6d8a8f95e775b3297c79e.

[16] HAO S, GU Y, MA H, et al. Reasoning with language model is planning with world model[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023: 8154-8173.

[17] LONG Y X, LI X Q, CAI W Z, et al. Discuss before moving: visual language navigation via multi-expert discussions[C]//Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Washington D.C., USA: IEEE Press, 2024: 17380-17387.

[18] ZHAO G L, LI G B, CHEN W K, et al. OVER-NAV: elevating iterative vision-and-language navigation with open-vocabulary detection and structured representation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2024: 16296-16306.

[19] YANG Z Y, LIN J G, CHEN P H, et al. RILA: reflective and imaginative language agent for zero-shot semantic audio-visual navigation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2024: 16251-16261.

[20] CHEN H J, CHEN X, DENG S M, et al. Agent planning with world knowledge model[C]//Proceedings of the Advances in Neural Information Processing Systems 37. Vancouver, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 114843-114871.

[21] CHAE H, KIM N, ONG K T, et al. Web agents with world models: learning and leveraging environment dynamics in web navigation[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2410.13232.

[22] HU M K, ZHAO P, XU C, et al. AgentGen: enhancing planning abilities for large language model based agent via environment and task generation[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2408.00764.

[23] WANG R Y, TODD G, YUAN X D, et al. ByteSized32: a corpus and challenge task for generating task-specific world models expressed as text games[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023: 13455-13471.

[24] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2302.13971.

[25] FENG J, LIU T H, DU Y W, et al. CityGPT: empowering urban spatial cognition of large language models[C]//Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2025: 591-602.

[26] ROBERTS J, LÜDDECKE T, DAS S, et al. GPT4GEO: how a language model sees the world's geography[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2306.00020.

[27] FENG J, DU Y W, ZHAO J, et al. AgentMove: a large language model based agentic framework for zero-shot next location prediction[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Philadelphia, USA: Association for Computational Linguistics, 2025: 1322-1338.

[28] LI L, ZHOU Y, LIANG Y X, et al. Recognition through reasoning: reinforcing image geo-localization with large vision-language models[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2506.14674.

[29] JIN C, RINARD M. Emergent representations of program semantics in language models trained on programs[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2305.11169.

[30] GAO Q Y, PI X Y, LIU K, et al. Do vision-language models have internal world models? Towards an atomic evaluation[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2506.21876.

[31] CHEN R R, JIANG W F, QIN C W, et al. Theory of mind in large language models: assessment and enhancement[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2505.00026.

[32] SAP M, LE BRAS R, FRIED D, et al. Neural theory-of-mind? On the limits of social intelligence in large LMs[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics, 2022: 3762-3780.

[33] STRACHAN J W A, ALBERGO D, BORGHINI G, et al. Testing theory of mind in large language models and humans[J]. Nature Human Behaviour, 2024, 8(7): 1285-1295.

[34] PARK J S, O'BRIEN J, CAI C J, et al. Generative agents: interactive simulacra of human behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. New York, USA: ACM Press, 2023: 1-22.

[35] BAKLASHKIN M, BODISHTIANU V, GLUSHANINA M, et al. EAI: emotional decision-making of LLMs in strategic games and ethical dilemmas[C]//Proceedings of the Advances in Neural Information Processing Systems 37. Vancouver, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 53969-54002.

[36] VAYANI A, DISSANAYAKE D, WATAWANA H, et al. All languages matter: evaluating LMMs on culturally diverse 100 languages[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2025: 19565-19575.

[37] Sora | OpenAI[EB/OL]. [2026-02-27]. https://openai.com/zh-Hans-CN/sora/.

[38] 可灵 AI——新一代 AI 创意生产力平台[EB/OL]. [2026-02-27]. https://klingai.com/cn/.
Keling AI—next-generation AI creative productivity platform[EB/OL]. [2026-02-27]. https://klingai.com/cn/. (in Chinese)

[39] Runway | AI image and video generator[EB/OL]. [2026-02-27]. https://runwayml.com/product.

[40] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 6000-6010.

[41] ZHENG Z W, PENG X Y, YANG T J, et al. Open-Sora: democratizing efficient video production for all[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2412.20404.

[42] YANG Z Y, TENG J Y, ZHENG W D, et al. CogVideoX: text-to-video diffusion models with an expert Transformer[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2408.06072.

[43] AGARWAL N, ALI A, BALA M, et al. Cosmos world foundation model platform for physical AI[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2501.03575.

[44] PARKER-HOLDER J, BALL P, BRUCE J, et al. Genie2: a large-scale foundation world modelEB/OL]. [2026-02-27]. https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model.

[45] Genie3: 通往 AGI 的世界模型新前沿 | Google DeepMind[EB/OL]. [2026-02-27]. https://www.genie3.cloud/zh#feature.
Genie3: a new frontier of world models toward AGI |

Google DeepMind[EB/OL]. [2026-02-27]. https://www.genie3.cloud/zh#feature. (in Chinese)
[46] YIN S M, WU C F, YANG H, et al. NUWA-XL: diffusion over diffusion for extremely long video generation [C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, USA: Association for Computational Linguistics, 2023: 1309-1320.
[47] BRUCE J, DENNIS M, EDWARDS A, et al. Genie: generative interactive environments [EB/OL]. [2026-02-27]. https://arxiv.org/abs/2402.15391.
[48] YANG S, WALKER J, PARKER-HOLDER J, et al. Video as the new language for real-world decision making [EB/OL]. [2026-02-27]. https://arxiv.org/abs/2402.17139.
[49] WANG X F, ZHU Z, HUANG G, et al. WorldDreamer: towards general world models for video generation via predicting masked tokens[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2401.09985.
[50] YU W, QIAN R J, LI Y M, et al. MosaicMem: hybrid spatial memory for controllable video world models [EB/OL]. [2026-02-27]. https://arxiv.org/abs/2603.17117.
[51] CHANG A, DAI A, FUNKHOUSER T, et al. Matterport3D: learning from RGB-D data in indoor environments[EB/OL]. [2026-02-27]. https://arxiv.org/abs/1709.06158.
[52] SHANG Y, LIN Y M, ZHENG Y, et al. UrbanWorld: an urban world model for 3D city generation[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2407.11965.
[53] ANANDKUMAR A, FAN L X, HUANG D A, et al. MineDojo: building open-ended embodied agents with Internet-scale knowledge[C]//Proceedings of the Advances in Neural Information Processing Systems 35. New Orleans, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022: 18343-18362.
[54] YANG S, DU Y L, GHASEMIPOUR K, et al. Learning interactive real-world simulators[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2310.06114.
[55] CHI X W, FAN C K, ZHANG H Y, et al. EVA: an embodied world model for future video anticipation [EB/OL]. [2026-02-27]. https://arxiv.org/abs/2410.15461.
[56] XIANG J N, LIU G Y, GU Y, et al. Pandora: towards general world model with natural language actions and video states[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2406.09455.
[57] ZHEN H Y, SUN Q, ZHANG H X, et al. TesserAct: learning 4D embodied world models [EB/OL]. [2026-02-27]. https://arxiv.org/abs/2504.20995.
[58] DENG B Y, TUCKER R, LI Z Q, et al. Streetscapes: large-scale consistent street view generation using autoregressive video diffusion [C]// Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference. New York, USA: ACM Press, 2024: 1-11.
[59] RIGTER M, GUPTA T, HILMKIL A, et al. AVID: adapting video diffusion models to world models[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2410.12822.
[60] HU M, CHEN T, ZOU Y, et al. Text2World: benchmarking large language models for symbolic world model generation [C]// Proceedings of Findings of the Association for Computational Linguistics: ACL 2025. Philadelphia, USA: Association for Computational Linguistics, 2025: 26043-26066.
[61] GE Z Q, HUANG H Z, ZHOU M Z, et al. WorldGPT: empowering LLM as multimodal world model [EB/OL]. [2026-02-27]. https://arxiv.org/abs/2404.18202.
[62] IVANOVA A A, SATHE A, LIPKIN B, et al. Elements of World Knowledge (EWoK): a cognition-inspired framework for evaluating basic world knowledge in language models[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2405.09605.
[63] YU J F, WANG X Z, TU S Q, et al. KoLA: carefully benchmarking world knowledge of large language models [EB/OL]. [2026-02-27]. https://arxiv.org/abs/2306.09296.
[64] FEI H, WU S Q, ZHANG M S, et al. Enhancing video-language representations with structural spatio-temporal alignment[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 7701-7719.
[65] CHERIAN A, PAUL S, ROY-CHOWDHURY A. AVLEN: audio-visual-language embodied navigation in 3D environments[C]//Proceedings of the Advances in Neural Information Processing Systems 35. New Orleans, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022: 6236-6249.
[66] SHEN B K, XIA F, LI C S, et al. iGibson 1.0: a simulation environment for interactive tasks in large realistic scenes [C]// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Washington D.C., USA: IEEE Press, 2021: 7520-7527.
[67] BOGDOLL D, YANG Y T, JOSEPH T, et al. MUVO: a multimodal generative world model for autonomous driving with geometric representations[C]//Proceedings of the IEEE Intelligent Vehicles Symposium (IV). Washington D.C., USA: IEEE Press, 2025: 2243-2250.
[68] WANG L R, LING Y Y, YUAN Z C, et al. GenSim: generating robotic simulation tasks via large language models[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2310.01361.
[69] LIN K, AGIA C, MIGIMATSU T, et al. Text2Motion: from natural language instructions to feasible plans [J]. Autonomous Robots, 2023, 47(8): 1345-1365.
[70] WANG Z H, CAI S F, CHEN G Z, et al. Describe, explain, plan and select: interactive planning with large language models enables open-world multi-task agents[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2302.01560.
[71] MAO Y S, ZHONG J H, FANG C, et al. SpatialLM: training large language models for structured indoor modeling[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2506.07491.
[72] LIU B, JIANG Y Q, ZHANG X H, et al. LLM+P: empowering large language models with optimal planning proficiency[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2304.11477.
[73] GUAN L, KAMBHAMPATI S, SREEDHARAN S, et al. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning[C]// Proceedings of the Advances in Neural Information Processing Systems 36. New Orleans, USA. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 79081-79094.
[74] ABBEEL P, ADENIJI A, ESCONTRELA A, et al. Video prediction models as rewards for reinforcement learning[C]// Proceedings of the Advances in Neural Information Processing Systems 36. New Orleans, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 68760-68783.
[75] CHEANG C L, CHEN G Z, JING Y, et al. GR-2: a generative video-language-action model with Web-scale knowledge for robot manipulation[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2410.06158.
[76] ZHOU S Y, DU Y L, CHEN J B, et al. RoboDreamer: learning compositional world models for robot imagination [EB/OL]. [2026-02-27]. https://arxiv.org/abs/2404.

12377.
[77] ABBEEL P, DAI B, DAI H J, et al. Learning universal policies via text-guided video generation[C]//Proceedings of the Advances in Neural Information Processing Systems 36. New Orleans, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 9156-9172.
[78] ZHU F, WU H, GUO S, et al. IRASim: learning interactive real-robot action simulators[EB/OL]. [2026-02-27]. https://arxiv.org/html/2406.14540v1.
[79] SHANG Y, ZHANG X, TANG Y Z, et al. RoboScape: physics-informed embodied world model[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2506.23135.
[80] MAJUMDAR A, AJAY A, ZHANG X H, et al. OpenEQA: embodied question answering in the era of foundation models [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2024: 16488-16498.
[81] GAO C, LAN X C, LU Z H, et al. S^3: social-network simulation system with large language model-empowered agents[EB/OL]. [2026-02-27]. https: // arxiv. org/abs/2307.14984.
[82] PAPACHRISTOU M, YUAN Y. Network formation and dynamics among multi-LLMs [EB/OL]. [2026-02-27]. https://arxiv.org/abs/2402.10659.
[83] LI N, GAO C, LI M Y, et al. EconAgent: large language model-empowered agents for simulating macroeconomic activities[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2310.10436.
[84] JI J, LI Y, LIU H, et al. SRAP-Agent: simulating and optimizing scarce resource allocation policy with llm-based agent[C]//Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2024. Miami, USA: Association for Computational Linguistics, 2024: 267-293.
[85] AL A, AHN A, BECKER N, et al. Project Sid: many-agent simulations toward AI civilization[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2411.00114.
[86] PIAO J H, YAN Y W, ZHANG J, et al. AgentSociety: large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2502.08691.
[87] QIAN C, LIANG S H, QIN Y J, et al. Investigate-consolidate-exploit: a general strategy for inter-task agent self-evolution[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2401.13996.
[88] ZHANG J T, XU X, ZHANG N Y, et al. Exploring collaboration mechanisms for LLM agents: a social psychology view[EB/OL]. [2026-02-27]. https: // arxiv.org/abs/2310.02124.
[89] ZHANG W Q, TANG K, WU H, et al. Agent-pro: learning to evolve via policy-level reflection and optimization [EB/OL]. [2026-02-27]. https: // arxiv. org/abs/2402.17574.
[90] 世界模型重塑未来城市，生成式 AI 如何构建智慧规划新范式[EB/OL]. [2026-02-07]. https://www.aigc.cn/91864.html.
World models reshape the future of cities: how generative AI is building a new paradigm for smart planning[EB/OL]. [2026-02-07]. https: // www. aigc. cn/91864. html. (in Chinese)
[91] 魏天呈，郭真，杨云龙. 大模型赋能智慧城市建设的路径与策略研究[J]. 信息通信技术与政策，2025，51(8)：91-96.
WEI T C, GUO Z, YANG Y L. Research on the paths and strategies of empowering smart city construction with large language models [J]. Information and Communications Technology and Policy, 2025, 51(8): 91-96. (in Chinese)
[92] CHEN Z X, WANG G C, LIU Z W. SceneDreamer: unbounded 3D scene generation from 2D image collections [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15562-15576.
[93] LIN C H, LEE H Y, MENAPACE W, et al. InfiniCity: infinite-scale city synthesis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Washington D. C., USA: IEEE Press, 2024: 22751-22761.
[94] XIE H Z, CHEN Z X, HONG F Z, et al. CityDreamer: compositional generative model of unbounded 3D cities[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2024: 9666-9675.
[95] DENG J, CHAI W H, DENG J, et al. CityGen: infinite and controllable city layout generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Washington D. C., USA: IEEE Press, 2025: 1986-1996.
[96] FENG C, CHEN Z Y, HOŁYŃSKI A, et al. GPS as a control signal for image generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2025: 2766-2778.
[97] MANVI R, KHANNA S, MAI G C, et al. GeoLLM: extracting geospatial knowledge from large language models [EB/OL]. [2026-02-27]. https: // arxiv. org/abs/2310.06213.
[98] FENG T, WANG W G, YANG Y. A survey of world models for autonomous driving[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2501.11260.
[99] ZHAO G S, WANG X F, ZHU Z, et al. DriveDreamer-2: LLM-enhanced world models for diverse driving video generation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2025: 10412-10420.
[100] ZHAO G S, NI C J, WANG X F, et al. DriveDreamer4D: world models are effective data machines for 4D driving scene representation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2025: 12015-12026.
[101] WANG Y Q, HE J W, FAN L, et al. Driving into the future: multiview visual forecasting and planning with world model for autonomous driving [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2024: 14749-14759.
[102] CHEN L, CHITTA K, GAO S Y, et al. Vista: a generalizable driving world model with high fidelity and versatile controllability[C]//Proceedings of the Advances in Neural Information Processing Systems 37. Vancouver, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 91560-91596.
[103] HU A, RUSSELL L, YEO H, et al. GAIA-1: a generative world model for autonomous driving[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2309.17080.
[104] RUSSELL L, HU A, BERTONI L, et al. GAIA-2: a controllable multi-view generative world model for autonomous driving [EB/OL]. [2026-02-27]. https: // arxiv.org/abs/2503.20523.
[105] ZHENG W Z, CHEN W L, HUANG Y H, et al. OccWorld: learning a 3D occupancy world model for autonomous driving [C] // Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2024: 55-72.
[106] WANG L N, ZHENG W Z, REN Y L, et al. OccSora: 4D occupancy generation models as world simulators for autonomous driving [EB/OL]. [2026-02-27]. https: //

arxiv.org/abs/2405.20337.
[107] JIANG C K, ZHOU D S, LIU J M, et al. VectorWorld: efficient streaming world model via diffusion flow on vector graphs[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2603.17652.
[108] CIPOLLA R, CORRADO G, GRIFFITHS N, et al. Model-based imitation learning for urban driving[C]// Proceedings of the Advances in Neural Information Processing Systems 35. New Orleans, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022: 20703-20716.
[109] LOGG A, MARDAL K A, WELLS G. Automated solution of differential equations by the finite element method: the FEniCS book[M]. Berlin, Germany: Springer, 2012.
[110] COUMANS E, BAI Y P. PyBullet, a Python module for physics simulation for games, robotics and machine learning [EB/OL]. [2026-02-27]. http://pybullet.org.
[111] TODOROV E, EREZ T, TASSA Y. MuJoCo: a physics engine for model-based control[C]// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Washington D.C., USA: IEEE Press, 2012: 5026-5033.
[112] HU Y M, LIU J C, SPIELBERG A, et al. ChainQueen: a real-time differentiable physical simulator for soft robotics[C]// Proceedings of the International Conference on Robotics and Automation (ICRA). Washington D.C., USA: IEEE Press, 2019: 6265-6271.
[113] BILLARD A, ALBU-SCHAEFFER A, BEETZ M, et al. A roadmap for AI in robotics[J]. Nature Machine Intelligence, 2025, 7(6): 818-824.
[114] LIU J F, SHI H Y, ZHANG S Y, et al. Automatic quantization for physics-based simulation[J]. ACM Transactions on Graphics, 2022, 41(4): 1-16.
[115] ANTYPAS D, AYELE A, BORKAKOTY H, et al. BLEnD: a benchmark for LLMs on everyday knowledge in diverse cultures and languages[C]// Proceedings of the Advances in Neural Information Processing Systems 37. Vancouver, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 78104-78146.
[116] STREET W, SIY J O, KEELING G, et al. LLMs achieve adult human performance on higher-order theory of mind tasks[J]. Frontiers in Human Neuroscience, 2025, 19: 1633272.
[117] YANG J H, YANG S S, GUPTA A W, et al. Thinking in space: how multimodal large language models see, remember, and recall spaces[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2025: 10632-10643.
[118] GUO X Y, HUO J Y, SHI Z M, et al. T2VPhysBench: a first-principles benchmark for physical consistency in text-to-video generation[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2505.00337.
[119] QIN Y R, SHI Z L, YU J W, et al. WorldSimBench: towards video generation models as world simulators[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2410.18072.
[120] DUAN H Y, YU H X, CHEN S R, et al. WorldScore: a unified evaluation benchmark for world generation[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2504.00983.
[121] HUANG Z Q, HE Y N, YU J S, et al. VBench: comprehensive benchmark suite for video generative models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2024: 21807-21818.
[122] ZHENG D, HUANG Z Q, LIU H B, et al. VBench-2.0: advancing video generation benchmark suite for intrinsic faithfulness[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2503.21755.
[123] Evaluating robot policies in a world model[EB/OL]. [2026-02-27]. https://arxiv.org/html/2506.00613v1.
[124] FENG J, ZHANG J, LIU T H, et al. CityBench: evaluating the capabilities of large language models for urban tasks[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2406.13945.
[125] ZHAO B N, FANG J J, DAI Z C, et al. UrbanVideo-bench: benchmarking vision-language models on embodied intelligence with video data in urban spaces[EB/OL]. [2026-02-27]. https://arxiv.org/abs/2503.06157.
[126] MA Y S, CUI C, CAO X, et al. LaMPilot: an open benchmark dataset for autonomous driving with language model programs[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2024: 15141-15151.
[127] DARGAHI NOBARI K, BERTRAM T. A multimodal driver monitoring benchmark dataset for driver modeling in assisted driving automation[J]. Scientific Data, 2024, 11: 327.

文字编辑 陆燕菲
栏目编辑 宋 圆