

AGENT AI: SURVEYING THE HORIZONS OF MULTIMODAL INTERACTION

AGENT AI: 多模态交互视野的调查

Zane Durante ^{1†*}, Qiuyuan Huang ^{2‡*}, Naoki Wake ^{2*},
 Zane Durante ^{1†*}, Qiuyuan Huang ^{2‡*}, Naoki Wake ^{2*},

Ran Gong ^{3†}, Jae Sung Park ^{4†}, Bidipta Sarkar ^{1†}, Rohan Taori ^{1†}, Yusuke Noda ⁵, Demetri Terzopoulos ³, Yejin Choi ⁴, Katsushi Ikeuchi ², Hoi Vo ⁵, Li Fei-Fei ¹, Jianfeng Gao ²

Ran Gong ^{3†}, Jae Sung Park ^{4†}, Bidipta Sarkar ^{1†}, Rohan Taori ^{1†}, Yusuke Noda ⁵, Demetri Terzopoulos ³, Yejin Choi ⁴, Katsushi Ikeuchi ², Hoi Vo ⁵, Li Fei-Fei ¹, Jianfeng Gao ²

¹ Stanford University; ² Microsoft Research, Redmond;

¹ 斯坦福大学; ² 微软研究院, 雷德蒙德;

³ University of California, Los Angeles; ⁴ University of Washington; ⁵ Microsoft Gaming

³ 加州大学洛杉矶分校; ⁴ 华盛顿大学; ⁵ 微软游戏

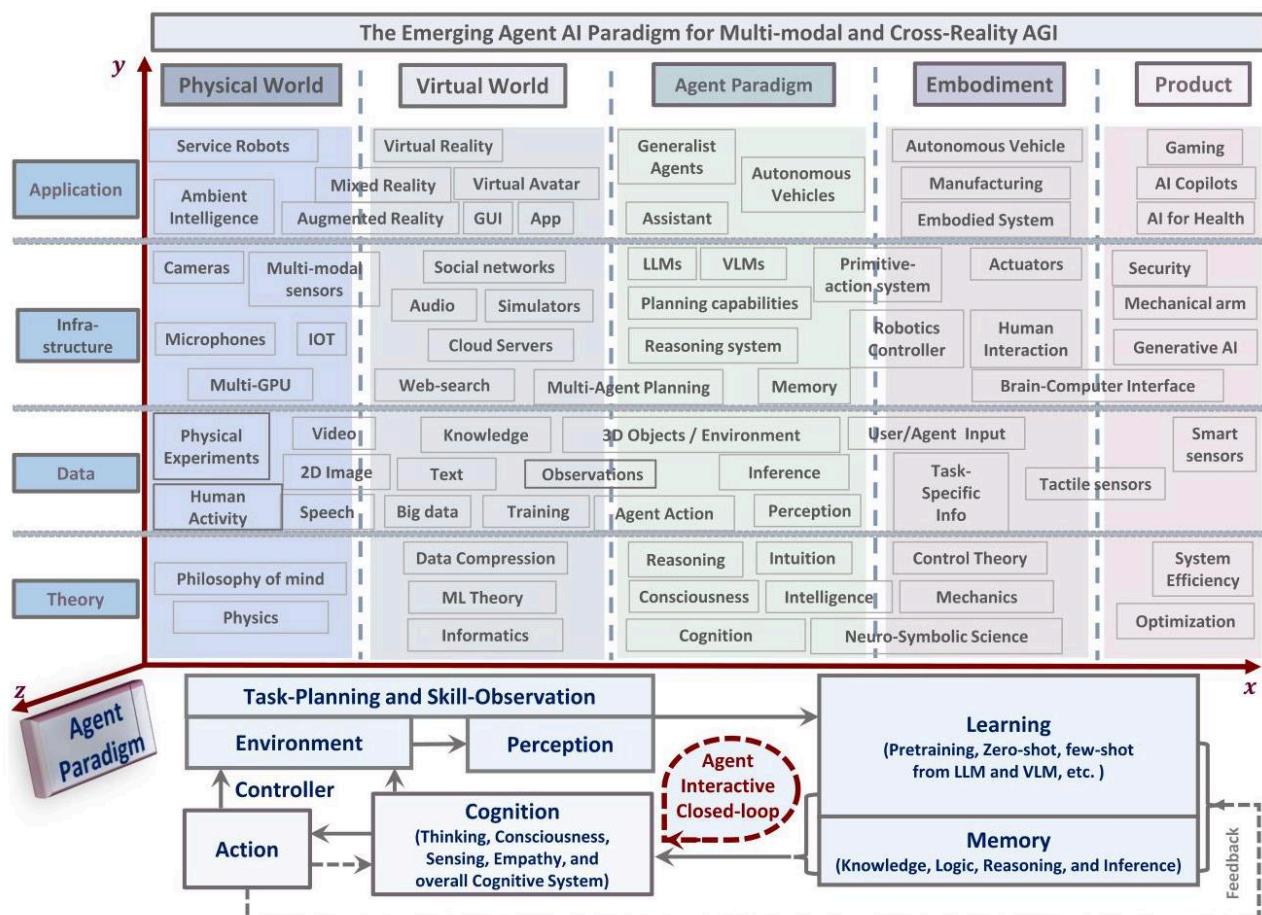


Figure 1: Overview of an Agent AI system that can perceive and act in different domains and applications. Agent AI is emerging as a promising avenue toward Artificial General Intelligence (AGI). Agent AI training has demonstrated the capacity for multi-modal understanding in the physical world. It provides a framework for reality-agnostic training by leveraging generative AI alongside multiple independent data sources. Large foundation models trained for agent and action-related tasks can be applied to physical and virtual worlds when trained on cross-reality data. We present the general overview of an Agent AI system that can perceive and act in many different domains and applications, possibly serving as a route towards AGI using an agent paradigm.

图1：Agent AI系统概览，该系统能够在不同领域和应用中感知和行动。Agent AI正成为迈向人工通用智能（Artificial General Intelligence, AGI）的有前景的途径。Agent AI训练已展示了在物理世界中实现多模态理解的能力。它通过结合生成式人工智能与多个独立数据源，提供了一个与现实无关的训练框架。针对代理和动作相关任务训练的大型基础模型，在跨现实数据训练后，可应用于物理和虚拟世界。我们展示了一个能够在多种不同领域和应用中感知和行动的Agent AI系统的总体概览，可能作为采用代理范式迈向AGI的路径。

*Equal Contribution. ‡ Project Lead. † Work done while interning at Microsoft Research, Redmond.

*同等贡献。‡ 项目负责人。† 工作完成于微软研究院雷德蒙德实习期间。

Agent AI:

Agent AI:

Surveying the Horizons of Multimodal Interaction

多模态交互视野的调查

1 Abstract

2 摘要

Multi-modal AI systems will likely become a ubiquitous presence in our everyday lives. A promising approach to making these systems more interactive is to embody them as agents within physical and virtual environments. At present, systems leverage existing foundation models as the basic building blocks for the creation of embodied agents. Embedding agents within such environments facilitates the ability of models to process and interpret visual and contextual data, which is critical for the creation of more sophisticated and context-aware AI systems. For example, a system that can perceive user actions, human behavior, environmental objects, audio expressions, and the collective sentiment of a scene can be used to inform and direct agent responses within the given environment. To accelerate research on agent-based multimodal intelligence, we define "Agent AI" as a class of interactive systems that can perceive visual stimuli, language inputs, and other environmentally-grounded data, and can produce meaningful embodied actions. In particular, we explore systems that aim to improve agents based on next-embody action prediction by incorporating external knowledge, multi-sensory inputs, and human feedback. We argue that by developing agentic AI systems in grounded environments, one can also mitigate the hallucinations of large foundation models and their tendency to generate environmentally incorrect outputs. The emerging field of Agent AI subsumes the broader embodied and agentic aspects of multimodal interactions. Beyond agents acting and interacting in the physical world, we envision a future where people can easily create any virtual reality or simulated scene and interact with agents embodied within the virtual environment.

多模态人工智能系统很可能成为我们日常生活中的普遍存在。使这些系统更具交互性的一个有前景的方法是将其具象化为物理和虚拟环境中的代理。目前，系统利用现有的基础模型作为构建具象代理的基本模块。将代理嵌入此类环境，有助于模型处理和解释视觉及上下文数据，这对于创建更复杂且具上下文感知的人工智能系统至关重要。例如，一个能够感知用户行为、人类行为、环境物体、音频表达及场景整体情绪的系统，可以用来指导代理在特定环境中的响应。为了加速基于代理的多模态智能研究，我们将“Agent AI”定义为一类能够感知视觉刺激、语言输入及其他环境基础数据，并能产生有意义具象动作的交互系统。特别地，我们探讨旨在通过整合外部知识、多感官输入和人类反馈，基于下一步具象动作预测来提升代理性能的系统。我们认为，通过在有根环境中开发代理式人工智能系统，还可以减轻大型基础模型的幻觉现象及其生成环境不符输出的倾向。新兴的Agent AI领域涵盖了多模态交互中更广泛的具象和代理特性。除了代理在物理世界中的行动和交互外，我们设想未来人们可以轻松创建任何虚拟现实或模拟场景，并与虚拟环境中具象的代理进行交互。

3 Contents

4 目录

1 Introduction 5

1 引言 5

1.1 Motivation 5

1.1 动机 5

1.2 Background 5

1.2 背景 5

1.3 Overview 6

1.3 概述 6

2 Agent AI Integration 7

2 代理人工智能集成 7

2.1 Infinite AI agent 7

2.1 无限代理人工智能 7

2.2 Agent AI with Large Foundation Models 8

2.2 基于大型基础模型的代理人工智能 8

2.2.1 Hallucinations 8

2.2.1 幻觉现象 8

2.2.2 Biases and Inclusivity 9

2.2.2 偏见与包容性 9

2.2.3 Data Privacy and Usage 10

2.2.3 数据隐私与使用 10

2.2.4 Interpretability and Explainability 11

2.2.4 可解释性与可说明性 11

2.2.5 Inference Augmentation 12

2.2.5 推理增强 12

2.2.6 Regulation 13

2.2.6 监管 13

2.3 Agent AI for Emergent Abilities 14

2.3 代理人工智能的突现能力 14

3 Agent AI Paradigm 15

3 代理人工智能范式 15

3.1 LLMs and VLMs 15

3.1 大型语言模型(LLMs)与视觉语言模型(VLMs) 15

3.2 Agent Transformer Definition 15

3.2 代理变换器定义 15

3.3 Agent Transformer Creation 16

3.3 代理变换器创建 16

4 Agent AI Learning 17

4 代理人工智能学习 17

4.1 Strategy and Mechanism 17

4.1 策略与机制 17

4.1.1 Reinforcement Learning (RL) 17

4.1.1 强化学习 (Reinforcement Learning, RL) 17

4.1.2 Imitation Learning (IL) 18

4.1.2 模仿学习 (Imitation Learning, IL) 18

4.1.3 Traditional RGB 18

4.1.3 传统RGB 18

4.1.4 In-context Learning 18

4.1.4 上下文学习 18

4.1.5 Optimization in the Agent System 18

4.1.5 代理系统中的优化 18

4.2 Agent Systems (zero-shot and few-shot level) 19

4.2 代理系统 (零样本和少样本级别) 19

4.2.1 Agent Modules 19

4.2.1 代理模块 19

4.2.2 Agent Infrastructure 19

4.2.2 代理基础设施 19

4.3 Agentic Foundation Models (pretraining and finetune level) 19

4.3 具代理性的基础模型 (预训练和微调级别) 19

5 Agent AI Categorization 20 5.1 Generalist Agent Areas 20

5 代理人工智能分类 20 5.1 通用代理领域 20

5.2 Embodied Agents 20

5.2 具身代理 20

5.2.1 Action Agents 20

5.2.1 行动代理 20

5.2.2 Interactive Agents	21
5.2.2 交互代理	21
5.3 Simulation and Environments Agents	21
5.3 仿真与环境代理	21
5.4 Generative Agents	21
5.4 生成式代理	21
5.4.1 AR/VR/mixed-reality Agents	22
5.4.1 增强现实/虚拟现实/混合现实代理	22
5.5 Knowledge and Logical Inference Agents	22
5.5 知识与逻辑推理代理	22
5.5.1 Knowledge Agent	23
5.5.1 知识代理	23
5.5.2 Logic Agents	23
5.5.2 逻辑代理	23
5.5.3 Agents for Emotional Reasoning	23
5.5.3 情感推理代理	23
5.5.4 Neuro-Symbolic Agents	24
5.5.4 神经符号代理	24
5.6 LLMs and VLMs Agent	24
5.6 大型语言模型(LLMs)与视觉语言模型(VLMs)代理	24
6 Agent AI Application Tasks	24
6 代理人工智能应用任务	24
6.1 Agents for Gaming	24
6.1 游戏代理	24
6.1.1 NPC Behavior	24
6.1.1 非玩家角色(NPC)行为	24
6.1.2 Human-NPC Interaction	25
6.1.2 人类与NPC交互	25
6.1.3 Agent-based Analysis of Gaming	25
6.1.3 基于代理的游戏分析	25
6.1.4 Scene Synthesis for Gaming	27
6.1.4 游戏场景合成	27
6.1.5 Experiments and Results	27
6.1.5 实验与结果	27
6.2 Robotics	28
6.2 机器人技术	28
6.2.1 LLM/VLM Agent for Robotics.	30
6.2.1 用于机器人技术的大型语言模型(LLM)/视觉语言模型(VLM)代理	30

6.2.2 Experiments and Results. 31

6.2.2 实验与结果 31

6.3 Healthcare 35

6.3 医疗保健 35

6.3.1 Current Healthcare Capabilities 36

6.3.1 当前医疗保健能力 36

6.4 Multimodal Agents 36

6.4 多模态智能体 36

6.4.1 Image-Language Understanding and Generation 36

6.4.1 图像-语言理解与生成 36

6.4.2 Video and Language Understanding and Generation 37

6.4.2 视频与语言理解与生成 37

6.4.3 Experiments and Results 39

6.4.3 实验与结果 39

6.5 Video-language Experiments 41

6.5 视频语言实验 41

6.6 Agent for NLP 45

6.6 自然语言处理智能体 45

6.6.1 LLM agent 45

6.6.1 大型语言模型（LLM）智能体 45

6.6.2 General LLM agent 45

6.6.2 通用大型语言模型（LLM）智能体 45

6.6.3 Instruction-following LLM agents 46

6.6.3 遵循指令的大型语言模型（LLM）智能体 46

6.6.4 Experiments and Results 46

6.6.4 实验与结果 46

7 Agent AI Across Modalities, Domains, and Realities 48

7 跨模态、跨领域与跨现实的智能体人工智能 48

7.1 Agents for Cross-modal Understanding 48

7.1 跨模态理解智能体 48

7.2 Agents for Cross-domain Understanding 48

7.2 跨领域理解智能体 48

7.3 Interactive agent for cross-modality and cross-reality 49

7.3 跨模态与跨现实的交互式智能体 49

7.4 Sim to Real Transfer 49

7.4 从仿真到现实的迁移 49

8 Continuous and Self-improvement for Agent AI 49

8 代理人工智能的持续与自我提升 49

8.1 Human-based Interaction Data	49
8.1 基于人类的交互数据	49
8.2 Foundation Model Generated Data	50
8.2 基础模型（Foundation Model）生成的数据	50
9 Agent Dataset and Leaderboard	50
9 代理数据集与排行榜	50
9.1 "CuisineWorld" Dataset for Multi-agent Gaming	50
9.1 多代理游戏的“CuisineWorld”数据集	50
9.1.1 Benchmark	51
9.1.1 基准测试	51
9.1.2 Task	51
9.1.2 任务	51
9.1.3 Metrics and Judging	51
9.1.3 指标与评判	51
9.1.4 Evaluation	51
9.1.4 评估	51
9.2 Audio-Video-Language Pre-training Dataset.	51
9.2 音频-视频-语言预训练数据集	51
10 Broader Impact Statement	52
10 更广泛的影响声明	52
11 Ethical Considerations	53
11 伦理考量	53
12 Diversity Statement	53
12 多样性声明	53
References	55
参考文献	55
Appendix	69
附录	69
A GPT-4V Agent Prompt Details	69
GPT-4V代理提示详情	69
B GPT-4V for Bleeding Edge	69
GPT-4V用于前沿技术	69
C GPT-4V for Microsoft Fight Simulator	69
GPT-4V用于微软格斗模拟器	69
D GPT-4V for Assassin's Creed Odyssey	69
GPT-4V用于《刺客信条：奥德赛》	69
E GPT-4V for GEARS of WAR	69
GPT-4V用于《战争机器4》	69

F GPT-4V for Starfield 75

GPT-4V用于《星空》 75

Author Biographies 77

作者简介 77

Acknowledgements 80

致谢 80

5 1 Introduction

6 1 引言

6.1 1.1 Motivation

6.2 1.1 动机

Historically, AI systems were defined at the 1956 Dartmouth Conference as artificial life forms that could collect information from the environment and interact with it in useful ways. Motivated by this definition, Minsky's MIT group built in 1970 a robotics system, called the "Copy Demo," that observed "blocks world" scenes and successfully reconstructed the observed polyhedral block structures. The system, which comprised observation, planning, and manipulation modules, revealed that each of these subproblems is highly challenging and further research was necessary. The AI field fragmented into specialized subfields that have largely independently made great progress in tackling these and other problems, but over-reductionism has blurred the overarching goals of AI research.

历史上，人工智能（AI）系统在1956年达特茅斯会议上被定义为能够从环境中收集信息并以有用方式与之交互的人工生命形式。基于这一定义，明斯基（Minsky）领导的麻省理工学院团队于1970年构建了一个名为“复制演示”

（Copy Demo）的机器人系统，该系统观察“积木世界”场景并成功重建所观察到的多面体积木结构。该系统包含观察、规划和操作模块，揭示了这些子问题各自的高度挑战性，表明需要进一步研究。人工智能领域因此分化为多个专业子领域，这些子领域在解决这些及其他问题上各自取得了重大进展，但过度简化导致人工智能研究的总体目标变得模糊。

To advance beyond the status quo, it is necessary to return to AI fundamentals motivated by Aristotelian Holism. Fortunately, the recent revolution in Large Language Models (LLMs) and Visual Language Models (VLMs) has made it possible to create novel AI agents consistent with the holistic ideal. Seizing upon this opportunity, this article explores models that integrate language proficiency, visual cognition, context memory, intuitive reasoning, and adaptability. It explores the potential completion of this holistic synthesis using LLMs and VLMs. In our exploration, we also revisit system design based on Aristotle's Final Cause, the teleological "why the system exists", which may have been overlooked in previous rounds of AI development.

为了超越现状，有必要回归以亚里士多德整体论（Aristotelian Holism）为动力的人工智能基础。幸运的是，近期大型语言模型（LLMs）和视觉语言模型（VLMs）的革命使得创建符合整体理想的新型人工智能代理成为可能。抓住这一机遇，本文探讨了整合语言能力、视觉认知、上下文记忆、直觉推理和适应性的模型。文章探讨了利用LLMs和VLMs完成这一整体综合的潜力。在探索过程中，我们还重新审视了基于亚里士多德终极因（Final Cause）——即系统存在的目的（teleological "why the system exists"）——的系统设计，这一点在以往的人工智能发展中可能被忽视。

With the advent of powerful pretrained LLMs and VLMs, a renaissance in natural language processing and computer vision has been catalyzed. LLMs now demonstrate an impressive ability to decipher the nuances of real-world linguistic data, often achieving abilities that parallel or even surpass human expertise (OpenAI, 2023). Recently, researchers have shown that LLMs may be extended to act as agents within various environments, performing intricate actions and tasks when paired with domain-specific knowledge and modules (Xi et al., 2023). These scenarios, characterized by complex reasoning, understanding of the agent's role and its environment, along with multi-step planning, test the agent's ability to make highly nuanced and intricate decisions within its

environmental constraints (Wu et al., 2023; Meta Fundamental AI Research (FAIR) Diplomacy Team et al., 2022). 随着强大预训练LLMs和VLMs的出现，自然语言处理和计算机视觉领域迎来了复兴。LLMs现已展现出解读现实世界语言数据细微差别的惊人能力，常常达到甚至超越人类专家水平 (OpenAI, 2023)。近期研究表明，LLMs可扩展为在各种环境中作为代理执行复杂动作和任务，前提是结合领域特定知识和模块 (Xi等, 2023)。这些场景以复杂推理、对代理角色及其环境的理解以及多步规划为特征，考验代理在环境约束下做出高度细致复杂决策的能力 (Wu 等, 2023; Meta基础人工智能研究 (FAIR) 外交团队等, 2022)。

Building upon these initial efforts, the AI community is on the cusp of a significant paradigm shift, transitioning from creating AI models for passive, structured tasks to models capable of assuming dynamic, agentic roles in diverse and complex environments. In this context, this article investigates the immense potential of using LLMs and VLMs as agents, emphasizing models that have a blend of linguistic proficiency, visual cognition, contextual memory, intuitive reasoning, and adaptability. Leveraging LLMs and VLMs as agents, especially within domains like gaming, robotics, and healthcare, promises not just a rigorous evaluation platform for state-of-the-art AI systems, but also foreshadows the transformative impacts that Agent-centric AI will have across society and industries. When fully harnessed, agentic models can redefine human experiences and elevate operational standards. The potential for sweeping automation ushered in by these models portends monumental shifts in industries and socio-economic dynamics. Such advancements will be intertwined with multifaceted leader-board, not only technical but also ethical, as we will elaborate upon in Section 11. We delve into the overlapping areas of these sub-fields of Agent AI and illustrate their interconnectedness in Fig.1.

基于这些初步努力，人工智能社区正处于重大范式转变的边缘，从为被动、结构化任务创建模型，转向能够在多样复杂环境中承担动态代理角色的模型。在此背景下，本文探讨了利用LLMs和VLMs作为代理的巨大潜力，强调具备语言能力、视觉认知、上下文记忆、直觉推理和适应性混合特征的模型。将LLMs和VLMs作为代理，尤其是在游戏、机器人和医疗等领域，不仅为最先进人工智能系统提供了严格的评估平台，也预示着以代理为中心的人工智能将在社会和产业中带来变革性影响。充分利用代理模型可重新定义人类体验并提升运营标准。这些模型带来的广泛自动化潜力预示着产业和社会经济格局的重大变革。此类进展将伴随多维度的排行榜，不仅涵盖技术层面，还涉及伦理层面，详见第11节。我们探讨了代理人工智能子领域的交叉区域，并在图1中展示了它们的相互关联性。

6.3 1.2 Background

6.4 1.2 背景

We will now introduce relevant research papers that support the concepts, theoretical background, and modern implementations of Agent AI.

接下来我们将介绍支持代理人工智能概念、理论背景及现代实现的相关研究论文。

Large Foundation Models: LLMs and VLMs have been driving the effort to develop general intelligent machines (Bubeck et al., 2023; Mirchandani et al., 2023). Although they are trained using large text corpora, their superior problem-solving capacity is not limited to canonical language processing domains. LLMs can potentially tackle complex tasks that were previously presumed to be exclusive to human experts or domain-specific algorithms, ranging from mathematical reasoning (Imani et al., 2023; Wei et al., 2022; Zhu et al., 2022) to answering questions of professional law (Blair-Stanek et al., 2023; Choi et al., 2023; Nay, 2022). Recent research has shown the possibility of using LLMs to generate complex plans for robots and game AI (Liang et al., 2022; Wang et al., 2023a,b; Yao et al., 2023a; Huang et al., 2023a), marking an important milestone for LLMs as general-purpose intelligent agents.

大型基础模型：大型语言模型（LLMs）和视觉语言模型（VLMs）一直推动着通用智能机器的发展 (Bubeck 等, 2023; Mirchandani 等, 2023)。尽管它们是通过大规模文本语料库训练的，但其卓越的问题解决能力并不限于传统的语言处理领域。LLMs有潜力处理此前被认为仅限于人类专家或特定领域算法的复杂任务，涵盖数学推理 (Imani 等, 2023; Wei 等, 2022; Zhu 等, 2022) 到专业法律问答 (Blair-Stanek 等, 2023; Choi 等, 2023; Nay, 2022)。最新研究表明，LLMs可用于生成机器人和游戏人工智能的复杂计划 (Liang 等, 2022; Wang 等, 2023a,b; Yao 等, 2023a; Huang 等, 2023a)，标志着LLMs作为通用智能代理的重要里程碑。

Embodied AI: A number of works leverage LLMs to perform task planning (Huang et al., 2022a; Wang et al., 2023b; Yao et al., 2023a; Li et al., 2023a), specifically the LLMs' WWW-scale domain knowledge and emergent zero-shot embodied abilities to perform complex task planning and reasoning. Recent robotics research also leverages LLMs to perform task planning (Ahn et al., 2022a; Huang et al., 2022b; Liang et al., 2022) by decomposing natural language instruction into a sequence of subtasks, either in the natural language form or in Python code, then using a low-level controller to execute these subtasks. Additionally, they incorporate environmental feedback to improve task performance (Huang et al., 2022b), (Liang et al., 2022), (Wang et al., 2023a), and (Ikeuchi et al., 2023).

具身人工智能：许多研究利用LLMs进行任务规划（Huang 等, 2022a; Wang 等, 2023b; Yao 等, 2023a; Li 等, 2023a），特别是LLMs在万维网规模的领域知识和新兴的零样本具身能力，用于执行复杂的任务规划和推理。近期机器学习研究也利用LLMs进行任务规划（Ahn 等, 2022a; Huang 等, 2022b; Liang 等, 2022），通过将自然语言指令分解为一系列子任务，形式可以是自然语言或Python代码，然后使用低级控制器执行这些子任务。此外，他们还结合环境反馈以提升任务表现（Huang 等, 2022b）、（Liang 等, 2022）、（Wang 等, 2023a）和（Ikeuchi 等, 2023）。

Interactive Learning: AI agents designed for interactive learning operate using a combination of machine learning techniques and user interactions. Initially, the AI agent is trained on a large dataset. This dataset includes various types of information, depending on the intended function of the agent. For instance, an AI designed for language tasks would be trained on a massive corpus of text data. The training involves using machine learning algorithms, which could include deep learning models like neural networks. These training models enable the AI to recognize patterns, make predictions, and generate responses based on the data on which it was trained. The AI agent can also learn from real-time interactions with users. This interactive learning can occur in various ways: 1) Feedback-based learning: The AI adapts its responses based on direct user feedback (Li et al., 2023b; Yu et al., 2023a; Parakh et al., 2023; Zha et al., 2023; Wake et al., 2023a,b,c). For example, if a user corrects the AI's response, the AI can use this information to improve future responses (Zha et al., 2023; Liu et al., 2023a). 2) Observational Learning: The AI observes user interactions and learns implicitly. For example, if users frequently ask similar questions or interact with the AI in a particular way, the AI might adjust its responses to better suit these patterns. It allows the AI agent to understand and process human language, multi-model setting, interpret the cross reality-context, and generate human-users' responses. Over time, with more user interactions and feedback, the AI agent's performance generally continuous improves. This process is often supervised by human operators or developers who ensure that the AI is learning appropriately and not developing biases or incorrect patterns.

交互式学习：为交互式学习设计的人工智能代理结合了机器学习技术和用户交互。最初，AI代理在大规模数据集上进行训练。该数据集包含多种信息类型，取决于代理的预期功能。例如，针对语言任务设计的AI会在庞大的文本语料库上训练。训练过程采用机器学习算法，包括神经网络等深度学习模型。这些训练模型使AI能够识别模式、做出预测并基于训练数据生成响应。AI代理还可以通过与用户的实时交互进行学习。交互式学习有多种方式：1) 基于反馈的学习：AI根据用户的直接反馈调整响应（Li 等, 2023b; Yu 等, 2023a; Parakh 等, 2023; Zha 等, 2023; Wake 等, 2023a,b,c）。例如，当用户纠正AI的回答时，AI可利用该信息改进后续响应（Zha 等, 2023; Liu 等, 2023a）。2) 观察学习：AI观察用户交互并隐式学习。例如，如果用户频繁提出类似问题或以特定方式与AI互动，AI可能调整响应以更好地适应这些模式。它使AI代理能够理解和处理人类语言、多模态环境，解释跨现实上下文，并生成符合人类用户的回应。随着用户交互和反馈的增加，AI代理的性能通常持续提升。该过程通常由人工操作员或开发者监督，确保AI适当学习，避免产生偏见或错误模式。

6.5 1.3 Overview

6.6 1.3 概述

Multimodal Agent AI (MAA) is a family of systems that generate effective actions in a given environment based on the understanding of multimodal sensory input. With the advent of Large Language Models (LLMs) and Vision-Language Models (VLMs), numerous MAA systems have been proposed in fields ranging from basic research to applications. While these research areas are growing rapidly by integrating with the traditional technologies of each domain (e.g., visual question answering and vision-language navigation), they share common interests such as data

collection, benchmarking, and ethical perspectives. In this paper, we focus on the some representative research areas of MAA, namely multimodality, gaming (VR/AR/MR), robotics, and healthcare, and we aim to provide comprehensive knowledge on the common concerns discussed in these fields. As a result we expect to learn the fundamentals of MAA and gain insights to further advance their research. Specific learning outcomes include:
多模态代理人工智能（MAA）是一类基于对多模态感知输入理解，在特定环境中生成有效行动的系统。随着大型语言模型（LLMs）和视觉语言模型（VLMs）的出现，众多MAA系统在基础研究到应用领域被提出。尽管这些研究领域通过与各自传统技术（如视觉问答和视觉语言导航）融合迅速发展，但它们在数据收集、基准测试和伦理视角等方面具有共同关注点。本文聚焦于MAA的若干代表性研究领域，即多模态、游戏（虚拟现实/增强现实/混合现实）、机器人和医疗健康，旨在提供这些领域共同关注问题的全面知识。由此，我们期望掌握MAA的基础并获得推动其研究进一步发展的洞见。具体学习成果包括：

- MAA Overview: A deep dive into its principles and roles in contemporary applications, providing researcher with a thorough grasp of its importance and uses.
• MAA概述：深入探讨其原理及在当代应用中的作用，为研究人员提供对其重要性和用途的全面理解。
- Methodologies: Detailed examples of how LLMs and VLMs enhance MAAs, illustrated through case studies in gaming, robotics, and healthcare.
• 方法论：通过游戏、机器人和医疗保健的案例研究，详细展示大型语言模型（LLMs）和视觉语言模型（VLMs）如何增强多模态智能代理（MAAs）。
- Performance Evaluation: Guidance on the assessment of MAAs with relevant datasets, focusing on their effectiveness and generalization.
• 性能评估：提供使用相关数据集评估多模态智能代理（MAAs）效果和泛化能力的指导。
- Ethical Considerations: A discussion on the societal impacts and ethical leader-board of deploying Agent AI, highlighting responsible development practices.
• 伦理考量：讨论部署智能代理AI的社会影响及伦理排行榜，强调负责任的发展实践。
- Emerging Trends and Future leader-board: Categorize the latest developments in each domain and discuss the future directions.
• 新兴趋势与未来排行榜：分类各领域最新进展并探讨未来发展方向。

Computer-based action and generalist agents (GAs) are useful for many tasks. A GA to become truly valuable to its users, it can natural to interact with, and generalize to a broad range of contexts and modalities. We aims to cultivate a vibrant research ecosystem and create a shared sense of identity and purpose among the Agent AI community. MAA has the potential to be widely applicable across various contexts and modalities, including input from humans. Therefore, we believe this Agent AI area can engage a diverse range of researchers, fostering a dynamic Agent AI community and shared goals. Led by esteemed experts from academia and industry, we expect that this paper will be an interactive and enriching experience, complete with agent instruction, case studies, tasks sessions, and experiments discussion ensuring a comprehensive and engaging learning experience for all researchers.
基于计算机的行动和通用智能代理（GAs）在许多任务中非常有用。为了使通用智能代理真正对用户有价值，它需要能够自然交互，并能泛化到广泛的情境和模态。我们旨在培育一个充满活力的研究生态系统，并在智能代理AI社区中创造共同的身份认同和目标感。多模态智能代理（MAA）有潜力广泛应用于各种情境和模态，包括来自人类的输入。因此，我们相信该智能代理AI领域能够吸引多样化的研究者，促进一个充满活力的智能代理AI社区和共同目标。在学术界和工业界资深专家的领导下，我们期望本文将成为一次互动且丰富的体验，包含代理指令、案例研究、任务环节和实验讨论，确保为所有研究者提供全面且引人入胜的学习体验。

This paper aims to provide general and comprehensive knowledge about the current research in the field of Agent AI. To this end, the rest of the paper is organized as follows. Section 2 outlines how Agent AI benefits from integrating with related emerging technologies, particularly large foundation models. Section 3 describes a new paradigm and framework that we propose for training Agent AI. Section 4 provides an overview of the

methodologies that are widely used in the training of Agent AI. Section 5 categorizes and discusses various types of agents. Section 6 introduces Agent AI applications in gaming, robotics, and healthcare. Section 7 explores the research community's efforts to develop a versatile Agent AI, capable of being applied across various modalities, domains, and bridging the sim-to-real gap. Section 8 discusses the potential of Agent AI that not only relies on pre-trained foundation models, but also continuously learns and self-improves by leveraging interactions with the environment and users. Section 9 introduces our new datasets that are designed for the training of multimodal Agent AI. Section 11 discusses the hot topic of the ethics consideration of AI agent, limitations, and societal impact of our paper.

本文旨在提供关于智能代理AI领域当前研究的通用且全面的知识。为此，本文其余部分组织如下。第2节概述智能代理AI如何受益于与相关新兴技术的整合，特别是大型基础模型。第3节描述我们提出的训练智能代理AI的新范式和框架。第4节概述智能代理AI训练中广泛使用的方法论。第5节对各种类型的智能代理进行分类和讨论。第6节介绍智能代理AI在游戏、机器人和医疗保健中的应用。第7节探讨研究社区致力于开发多模态、跨领域且能弥合仿真与现实差距的通用智能代理AI的努力。第8节讨论不仅依赖预训练基础模型，还通过与环境和用户的交互持续学习和自我提升的智能代理AI潜力。第9节介绍我们为多模态智能代理AI训练设计的新数据集。第11节讨论智能代理AI伦理考量、局限性及其社会影响这一热点话题。

7 2 Agent AI Integration

8 2 智能代理AI整合

Foundation models based on LLMs and VLMs, as proposed in previous research, still exhibit limited performance in the area of embodied AI, particularly in terms of understanding, generating, editing, and interacting within unseen environments or scenarios (Huang et al., 2023a; Zeng et al., 2023). Consequently, these limitations lead to sub-optimal outputs from AI agents. Current agent-centric AI modeling approaches focus on directly accessible and clearly defined data (e.g. text or string representations of the world state) and generally use domain and environment-independent patterns learned from their large-scale pretraining to predict action outputs for each environment (Xi et al., 2023; Wang et al., 2023c; Gong et al., 2023a; Wu et al., 2023). In (Huang et al., 2023a), we investigate the task of knowledge-guided collaborative and interactive scene generation by combining large foundation models, and show promising results that indicate knowledge-grounded LLM agents can improve the performance of 2D and 3D scene understanding, generation, and editing, alongside with other human-agent interactions (Huang et al., 2023a). By integrating an Agent AI framework, large foundation models are able to more deeply understand user input to form a complex and adaptive HCI system. Emergent ability of LLM and VLM works invisible in generative AI, embodied AI, knowledge augmentation for multi-model learning, mix-reality generation, text to vision editing, human interaction for 2D/3D simulation in gaming or robotics tasks. Agent AI recent progress in foundation models present an imminent catalyst for unlocking general intelligence in embodied agents. The large action models, or agent-vision-language models open new possibilities for general-purpose embodied systems such as planning, problem-solving and learning in complex environments. Agent AI test further step in metaverse, and route the early version of AGI.

基于大型语言模型（LLMs）和视觉语言模型（VLMs）的基础模型，如先前研究所提出，在具身智能（embodied AI）领域仍表现有限，尤其是在理解、生成、编辑及在未见环境或场景中交互方面（Huang et al., 2023a; Zeng et al., 2023）。因此，这些限制导致智能代理输出效果不佳。目前以代理为中心的AI建模方法侧重于直接可访问且定义明确的数据（如世界状态的文本或字符串表示），通常利用其大规模预训练中学到的领域和环境无关模式来预测各环境的动作输出（Xi et al., 2023; Wang et al., 2023c; Gong et al., 2023a; Wu et al., 2023）。在（Huang et al., 2023a）中，我们通过结合大型基础模型，研究了知识引导的协作与交互式场景生成任务，展示了知识驱动的LLM代理能够提升2D和3D场景理解、生成与编辑的性能，以及其他人机交互（Huang et al., 2023a）。通过整合智能代理AI框架，大型基础模型能够更深入理解用户输入，形成复杂且自适应的人机交互系统。LLM和VLM的涌现能力在生成式AI、具身AI、多模态学习的知识增强、混合现实生成、文本到视觉编辑、游戏或机器人任务中的二维/三维仿真人机交互中发挥着无形作用。智能代理AI在基础模型上的最新进展为解锁具身代理中的通用智能提供了迫切的催化剂。大型动作模型或代理视觉语言模型为通用具身系统如规划、问题解决和复杂环境中的学习开辟了新可能。智能代理AI测试迈出元宇宙的进一步步伐，并引领通用人工智能（AGI）的早期版本。

8.1 2.1 Infinite AI agent

8.2 2.1 无限智能代理

AI agents have the capacity to interpret, predict, and respond based on its training and input data. While these capabilities are advanced and continually improving, it's important to recognize their limitations and the influence of the underlying data they are trained on. AI agent systems generally possess the following abilities: 1) Predictive Modeling: AI agents can predict likely outcomes or suggest next steps based on historical data and trends. For instance, they might predict the continuation of a text, the answer to a question, the next action for a robot, or the resolution of a scenario. 2) Decision Making: In some applications, AI agents can make decisions based on their inferences. Generally, the agent will base their decision on what is most likely to achieve a specified goal. For AI applications like recommendation systems, an agent can decide what products or content to recommend based on its inferences about user preferences. 3) Handling Ambiguity: AI agents can often handle ambiguous input by inferring the most likely interpretation based on context and training. However, their ability to do so is limited by the scope of their training data and algorithms. 4) Continuous Improvement: While some AI agents have the ability to learn from new data and interactions, many large language models do not continuously update their knowledge-base or internal representation after training. Their inferences are usually based solely on the data that was available up to the point of their last training update.

人工智能代理具备基于其训练和输入数据进行解释、预测和响应的能力。尽管这些能力先进且不断提升，但重要的是要认识到其局限性以及所依赖的训练数据的影响。人工智能代理系统通常具备以下能力：1) 预测建模：人工智能代理可以基于历史数据和趋势预测可能的结果或建议下一步操作。例如，它们可能预测文本的续写、问题的答案、机器人的下一步动作或场景的解决方案。2) 决策制定：在某些应用中，人工智能代理可以基于推断做出决策。通常，代理会基于最有可能实现特定目标的方案做出决策。对于推荐系统等人工智能应用，代理可以根据对用户偏好的推断决定推荐哪些产品或内容。3) 处理歧义：人工智能代理通常能够通过基于上下文和训练推断最可能的解释来处理模糊输入。然而，其能力受限于训练数据和算法的范围。4) 持续改进：虽然部分人工智能代理具备从新数据和交互中学习的能力，但许多大型语言模型在训练后并不持续更新其知识库或内部表示。它们的推断通常仅基于最后一次训练更新时可用的数据。

We show augmented interactive agents for multi-modality and cross reality-agnostic integration with an emergence mechanism in Fig. 2. An AI agent requires collecting extensive training data for every new task, which can be costly or impossible for many domains. In this study, we develop an infinite agent that learns to transfer memory information from general foundation models (e.g., GPT-X, DALL-E) to novel domains or scenarios for scene understanding, generation, and interactive editing in physical or virtual worlds.

我们在图2中展示了用于多模态和跨现实无关集成的增强交互代理及其涌现机制。人工智能代理需要为每个新任务收集大量训练数据，这在许多领域可能成本高昂或不可行。在本研究中，我们开发了一种无限代理，能够学习将通用基础模型（如GPT-X、DALL-E）的记忆信息迁移到新颖领域或场景，用于物理或虚拟世界中的场景理解、生成和交互编辑。



Figure 2: The multi-model agent AI for 2D/3D embodied generation and editing interaction in cross-reality.

图2：用于跨现实中二维/三维具身生成与编辑交互的多模态代理人工智能。

An application of such an infinite agent in robotics is RoboGen (Wang et al., 2023d). In this study, the authors propose a pipeline that autonomously run the cycles of task proposition, environment generation, and skill learning. RoboGen is an effort to transfer the knowledge embedded in large models to robotics.

这种无限代理在机器人领域的一个应用是RoboGen (Wang等, 2023d)。该研究提出了一个自动运行任务提议、环境生成和技能学习循环的流程。RoboGen旨在将大型模型中蕴含的知识转移到机器人领域。

8.3 2.2 Agent AI with Large Foundation Models

8.4 2.2 基于大型基础模型的代理人工智能

Recent studies have indicated that large foundation models play a crucial role in creating data that act as benchmarks for determining the actions of agents within environment-imposed constraints. For example, using foundation models for robotic manipulation (Black et al., 2023; Ko et al., 2023) and navigation (Shah et al., 2023a; Zhou et al., 2023a). To illustrate, Black et al. employed an image-editing model as a high-level planner to generate images of future sub-goals, thereby guiding low-level policies (Black et al., 2023). For robot navigation, Shah et al. proposed a system that employs a LLM to identify landmarks from text and a VLM to associate these landmarks with visual inputs, enhancing navigation through natural language instructions (Shah et al., 2023a).

近期研究表明，大型基础模型在生成作为基准数据以确定代理在环境约束下行动方面发挥着关键作用。例如，利用基础模型进行机器人操作 (Black等, 2023; Ko等, 2023) 和导航 (Shah等, 2023a; Zhou等, 2023a)。举例来说，Black等人采用图像编辑模型作为高层规划器，生成未来子目标的图像，从而指导低层策略 (Black等, 2023)。在机器人导航方面，Shah等人提出了一个系统，利用大型语言模型 (LLM) 从文本中识别地标，并利用视觉语言模型 (VLM) 将这些地标与视觉输入关联，通过自然语言指令增强导航能力 (Shah等, 2023a)。

There is also growing interest in the generation of conditioned human motions in response to language and environmental factors. Several AI systems have been proposed to generate motions and actions that are tailored to specific linguistic instructions (Kim et al., 2023; Zhang et al., 2022; Tevet et al., 2022) and to adapt to various 3D scenes (Wang et al., 2022a). This body of research emphasizes the growing capabilities of generative models in enhancing the adaptability and responsiveness of AI agents across diverse scenarios.

针对语言和环境因素生成条件化人体动作的研究也日益受到关注。已有多个AI系统被提出，用于生成符合特定语言指令（Kim等，2023；Zhang等，2022；Tevet等，2022）并适应各种三维场景（Wang等，2022a）的动作和行为。这些研究强调了生成模型在提升人工智能代理在多样场景中的适应性和响应能力方面的不断增强。

8.4.1 2.2.1 Hallucinations

8.4.2 2.2.1 幻觉现象

Agents that generate text are often prone to hallucinations, which are instances where the generated text is nonsensical or unfaithful to the provided source content (Raunak et al., 2021; Maynez et al., 2020). Hallucinations can be split into two categories, intrinsic and extrinsic (Ji et al., 2023). Intrinsic hallucinations are hallucinations that are contradictory to the source material, whereas extrinsic hallucinations are when the generated text contains additional information that was not originally included in the source material.

生成文本的代理常常容易出现幻觉现象，即生成的文本无意义或与提供的源内容不符（Raunak等，2021；Maynez等，2020）。幻觉可分为两类：内在幻觉和外在幻觉（Ji等，2023）。内在幻觉指与源材料矛盾的内容，而外在幻觉则是生成文本包含了源材料中未包含的额外信息。

Some promising routes for reducing the rate of hallucination in language generation involve using retrieval-augmented generation (Lewis et al., 2020; Shuster et al., 2021) or other methods for grounding natural language outputs via external knowledge retrieval (Dziri et al., 2021; Peng et al., 2023). Generally, these methods seek to augment language generation by retrieving additional source material and by providing mechanisms to check for contradictions between the generated response and the source material.

减少语言生成中幻觉率的一些有前景的方法包括使用检索增强生成（Lewis等，2020；Shuster等，2021）或通过外部知识检索为自然语言输出提供依据的其他方法（Dziri等，2021；Peng等，2023）。通常，这些方法旨在通过检索额外的源材料并提供机制以检查生成响应与源材料之间的矛盾，从而增强语言生成的可靠性。

Within the context of multi-modal agent systems, VLMs have been shown to hallucinate as well (Zhou et al., 2023b). One common cause of hallucination for vision-based language-generation is due to the over-reliance on co-occurrence of objects and visual cues in the training data (Rohrbach et al., 2018). AI agents that exclusively rely upon pretrained LLMs or VLMs and use limited environment-specific finetuning can be particularly vulnerable to hallucinations since they rely upon the internal knowledge-base of the pretrained models for generating actions and may not accurately understand the dynamics of the world state in which they are deployed.

在多模态智能体系统的背景下，视觉语言模型（VLMs）也被发现会产生幻觉（Zhou等，2023b）。基于视觉的语言生成产生幻觉的一个常见原因是过度依赖训练数据中对象和视觉线索的共现（Rohrbach等，2018）。仅依赖预训练大型语言模型（LLMs）或视觉语言模型（VLMs）并进行有限环境特定微调的人工智能智能体尤其容易出现幻觉，因为它们依赖预训练模型的内部知识库来生成动作，可能无法准确理解其部署环境中的世界状态动态。

8.4.3 2.2.2 Biases and Inclusivity

8.4.4 2.2.2 偏见与包容性

AI agents based on LLMs or LMMs (large multimodal models) have biases due to several factors inherent in their design and training process. When designing these AI agents, we must be mindful of being inclusive and aware of the needs of all end users and stakeholders. In the context of AI agents, inclusivity refers to the measures and principles employed to ensure that the agent's responses and interactions are inclusive, respectful, and sensitive to a wide range of users from diverse backgrounds. We list key aspects of agent biases and inclusivity below.

基于大型语言模型（LLMs）或大型多模态模型（LMMs）的人工智能智能体因其设计和训练过程中的多种固有因素而存在偏见。在设计这些智能体时，我们必须关注包容性，意识到所有终端用户和利益相关者的需求。在人工智能智

能体的语境中，包容性指的是确保智能体的响应和交互包容、尊重并敏感对待来自多元背景的广泛用户所采取的措施和原则。以下列出智能体偏见与包容性的关键方面。

- Training Data: Foundation models are trained on vast amounts of text data collected from the internet, including books, articles, websites, and other text sources. This data often reflects the biases present in human society, and the model can inadvertently learn and reproduce these biases. This includes stereotypes, prejudices, and slanted viewpoints related to race, gender, ethnicity, religion, and other personal attributes. In particular, by training on internet data and often only English text, models implicitly learn the cultural norms of Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies (Henrich et al., 2010) who have a disproportionately large internet presence. However, it is essential to recognize that datasets created by humans cannot be entirely devoid of bias, since they frequently mirror the societal biases and the predispositions of the individuals who generated and/or compiled the data initially.
- 训练数据：基础模型在大量来自互联网的文本数据上训练，包括书籍、文章、网站及其他文本来源。这些数据往往反映了人类社会中的偏见，模型可能无意中学习并再现这些偏见，包括与种族、性别、族裔、宗教及其他个人属性相关的刻板印象、偏见和倾斜观点。特别是，由于训练数据多来自互联网且通常仅包含英文文本，模型隐式学习了西方、受教育、工业化、富裕和民主（WEIRD）社会（Henrich 等, 2010）的文化规范，这些社会在互联网中占据不成比例的主导地位。然而，必须认识到，由人类创建的数据集不可能完全无偏，因为它们常常反映了最初生成和/或编纂数据的个体的社会偏见和倾向。
- Historical and Cultural Biases: AI models are trained on large datasets sourced from diverse content. Thus, the training data often includes historical texts or materials from various cultures. In particular, training data from historical sources may contain offensive or derogatory language representing a particular society's cultural norms, attitudes, and prejudices. This can lead to the model perpetuating outdated stereotypes or not fully understanding contemporary cultural shifts and nuances.
- 历史与文化偏见：人工智能模型训练于来源多样的大型数据集，因此训练数据中常包含历史文本或来自不同文化的材料。尤其是历史来源的训练数据可能包含反映特定社会文化规范、态度和偏见的冒犯性或贬义语言，这可能导致模型延续过时的刻板印象，或无法充分理解当代文化的变迁和细微差别。
- Language and Context Limitations: Language models might struggle with understanding and accurately representing nuances in language, such as sarcasm, humor, or cultural references. This can lead to misinterpretations or biased responses in certain contexts. Furthermore, there are many aspects of spoken language that are not captured by pure text data, leading to a potential disconnect between human understanding of language and how models understand language.
- 语言与语境限制：语言模型可能难以理解和准确表达语言中的细微差别，如讽刺、幽默或文化引用，这可能导致在某些语境下产生误解或偏见的回应。此外，纯文本数据无法捕捉口语语言的许多方面，导致模型对语言的理解与人类存在潜在脱节。
- Policies and Guidelines: AI agents operate under strict policies and guidelines to ensure fairness and inclusivity. For instance, in generating images, there are rules to diversify depictions of people, avoiding stereotypes related to race, gender, and other attributes.
- 政策与指南：人工智能智能体在严格的政策和指南下运行，以确保公平和包容。例如，在生成图像时，有规则要求多样化人物描绘，避免与种族、性别及其他属性相关的刻板印象。
- Overgeneralization: These models tend to generate responses based on patterns seen in the training data. This can lead to overgeneralizations, where the model might produce responses that seem to stereotype or make broad assumptions about certain groups.
- 过度泛化：这些模型倾向于基于训练数据中观察到的模式生成响应，这可能导致过度泛化，模型可能产生看似刻板印象或对某些群体做出广泛假设的回答。
- Constant Monitoring and Updating: AI systems are continuously monitored and updated to address any

emerging biases or inclusivity issues. Feedback from users and ongoing research in AI ethics play a crucial role in this process.

- 持续监控与更新：人工智能系统持续接受监控和更新，以应对新出现的偏见或包容性问题。用户反馈和人工智能伦理领域的持续研究在此过程中发挥关键作用。
- Amplification of Dominant Views: Since the training data often includes more content from dominant cultures or groups, the model may be more biased towards these perspectives, potentially underrepresenting or misrepresenting minority viewpoints.
- 主流观点的放大：由于训练数据中往往包含更多来自主流文化或群体的内容，模型可能更偏向这些视角，可能导致少数群体观点的代表性不足或误读。
- Ethical and Inclusive Design: AI tools should be designed with ethical considerations and inclusivity as core principles. This includes respecting cultural differences, promoting diversity, and ensuring that the AI does not perpetuate harmful stereotypes.
- 伦理与包容性设计：人工智能工具应以伦理考量和包容性为核心原则进行设计，包括尊重文化差异、促进多样性，并确保人工智能不延续有害的刻板印象。
- User Guidelines: Users are also guided on how to interact with AI in a manner that promotes inclusivity and respect. This includes refraining from requests that could lead to biased or inappropriate outputs. Furthermore, it can help mitigate models learning harmful material from user interactions.
- 用户指南：用户也被指导以促进包容和尊重的方式与人工智能互动，包括避免提出可能导致偏见或不当输出的请求。此外，这有助于减少模型从用户交互中学习有害内容的风险。

Despite these measures, AI agents still exhibit biases. Ongoing efforts in agent AI research and development are focused on further reducing these biases and enhancing the inclusivity and fairness of agent AI systems. Efforts to Mitigate Biases:

尽管采取了这些措施，人工智能智能体仍然表现出偏见。智能体人工智能研究与开发的持续努力集中于进一步减少这些偏见，提升智能体系统的包容性和公平性。减轻偏见的努力包括：

- Diverse and Inclusive Training Data: Efforts are made to include a more diverse and inclusive range of sources in the training data.
- 多样且包容的训练数据：努力在训练数据中包含更多多样且包容的来源。
- Bias Detection and Correction: Ongoing research focuses on detecting and correcting biases in model responses.
- 偏见检测与纠正：持续的研究聚焦于检测和纠正模型响应中的偏见。
- Ethical Guidelines and Policies: Models are often governed by ethical guidelines and policies designed to mitigate biases and ensure respectful and inclusive interactions.
- 伦理指南与政策：模型通常受伦理指南和政策的约束，旨在减轻偏见，确保尊重且包容的交互。
- Diverse Representation: Ensuring that the content generated or the responses provided by the AI agent represent a wide range of human experiences, cultures, ethnicities, and identities. This is particularly relevant in scenarios like image generation or narrative construction.
- 多样化表现：确保AI代理生成的内容或提供的回应能够代表广泛的人类经验、文化、族群和身份。这在图像生成或叙事构建等场景中特别重要。
- Bias Mitigation: Actively working to reduce biases in the AI's responses. This includes biases related to race, gender, age, disability, sexual orientation, and other personal characteristics. The goal is to provide fair and balanced responses that do not perpetuate stereotypes or prejudices.
- 偏见缓解：积极减少AI回应中的偏见，包括与种族、性别、年龄、残疾、性取向及其他个人特征相关的偏见。目标是提供公平且平衡的回应，避免延续刻板印象或偏见。

- Cultural Sensitivity: The AI is designed to be culturally sensitive, acknowledging and respecting the diversity of cultural norms, practices, and values. This includes understanding and appropriately responding to cultural references and nuances.
- 文化敏感性：AI设计时注重文化敏感性，承认并尊重文化规范、习俗和价值观的多样性，包括理解并恰当回应文化引用和细微差别。
- Accessibility: Ensuring that the AI agent is accessible to users with different abilities, including those with disabilities. This can involve incorporating features that make interactions easier for people with visual, auditory, motor, or cognitive impairments.
- 无障碍性：确保AI代理对不同能力的用户均可访问，包括残障人士。这可能涉及加入便于视觉、听觉、运动或认知障碍者互动的功能。
- Language-based Inclusivity: Providing support for multiple languages and dialects to cater to a global user base, and being sensitive to the nuances and variations within a language (Liu et al., 2023b).
- 基于语言的包容性：支持多种语言和方言，以满足全球用户需求，并对语言内部的细微差别和变体保持敏感 (Liu et al., 2023b)。
- Ethical and Respectful Interactions: The Agent is programmed to interact ethically and respectfully with all users, avoiding responses that could be deemed offensive, harmful, or disrespectful.
- 伦理且尊重的互动：代理被编程为与所有用户进行伦理且尊重的互动，避免产生可能被视为冒犯、有害或不尊重的回应。
- User Feedback and Adaptation: Incorporating user feedback to continually improve the inclusivity and effectiveness of the AI agent. This includes learning from interactions to better understand and serve a diverse user base.
- 用户反馈与适应：结合用户反馈持续提升AI代理的包容性和有效性，包括通过交互学习更好地理解和服务多样化用户群体。
- Compliance with Inclusivity Guidelines: Adhering to established guidelines and standards for inclusivity in AI agent, which are often set by industry groups, ethical boards, or regulatory bodies.
- 遵守包容性指南：遵循由行业团体、伦理委员会或监管机构制定的AI代理包容性相关的既定指南和标准。

Despite these efforts, it's important to be aware of the potential for biases in responses and to interpret them with critical thinking. Continuous improvements in AI agent technology and ethical practices aim to reduce these biases over time. One of the overarching goals for inclusivity in agent AI is to create an agent that is respectful and accessible to all users, regardless of their background or identity.

尽管做出了这些努力，但仍需意识到回应中可能存在的偏见，并以批判性思维加以解读。AI代理技术和伦理实践的持续改进旨在逐步减少这些偏见。包容性AI代理的总体目标之一是创建一个尊重且对所有用户均可访问的代理，无论其背景或身份如何。

8.4.5 2.2.3 Data Privacy and Usage

8.4.6 2.2.3 数据隐私与使用

One key ethical consideration of AI agents involves comprehending how these systems handle, store, and potentially retrieve user data. We discuss key aspects below:

AI代理的一个关键伦理考量是理解这些系统如何处理、存储及可能检索用户数据。以下讨论关键方面：

Data Collection, Usage and Purpose. When using user data to improve model performance, model developers access the data the AI agent has collected while in production and interacting with users. Some systems allow users to view their data through user accounts or by making a request to the service provider. It is important to recognize what data the AI agent collects during these interactions. This could include text inputs, user usage patterns, personal preferences, and sometimes more sensitive personal information. Users should also understand how the

data collected from their interactions is used. If, for some reason, the AI holds incorrect information about a particular person or group, there should be a mechanism for users to help correct this once identified. This is important for both accuracy and to be respectful of all users and groups. Common uses for retrieving and analyzing user data include improving user interaction, personalizing responses, and system optimization. It is extremely important for developers to ensure the data is not used for purposes that users have not consented to, such as unsolicited marketing.

数据收集、使用及目的。在利用用户数据提升模型性能时，模型开发者会访问AI代理在生产环境中与用户交互时收集的数据。一些系统允许用户通过账户查看其数据或向服务提供商提出请求。重要的是明确AI代理在交互过程中收集了哪些数据，可能包括文本输入、用户使用模式、个人偏好，有时甚至更敏感的个人信息。用户还应了解其交互数据的使用方式。如果AI对某个人或群体持有错误信息，应有机制允许用户在发现后协助纠正。这对准确性及尊重所有用户和群体都至关重要。检索和分析用户数据的常见用途包括改善用户交互、个性化回应和系统优化。开发者必须确保数据不会被用于用户未同意的目的，如未经请求的营销。

Storage and Security. Developers should know where the user interaction data is stored and what security measures are in place to protect it from unauthorized access or breaches. This includes encryption, secure servers, and data protection protocols. It is extremely important to determine if agent data is shared with third parties and under what conditions. This should be transparent and typically requires user consent.

存储与安全。开发者应明确用户交互数据的存储位置及所采取的安全措施，以防止未经授权的访问或泄露，包括加密、安全服务器和数据保护协议。确定代理数据是否与第三方共享及其条件也极为重要，应保持透明，通常需用户同意。

Data Deletion and Retention. It is also important for users to understand how long user data is stored and how users can request its deletion. Many data protection laws give users the right to be forgotten, meaning they can request their data be erased. AI agents must adhere to data protection laws like GDPR in the EU or CCPA in California. These laws govern data handling practices and user rights regarding their personal data.

数据删除与保留。用户还应了解数据存储期限及如何请求删除。许多数据保护法规赋予用户“被遗忘权”，即可请求删除其数据。AI代理必须遵守如欧盟GDPR或加州CCPA等数据保护法律，这些法律规范数据处理实践及用户对个人数据的权利。

Data Portability and Privacy Policy. Furthermore, developers must create the AI agent's privacy policy to document and explain to users how their data is handled. This should detail data collection, usage, storage, and user rights. Developers should ensure that they obtain user consent for data collection, especially for sensitive information. Users typically have the option to opt-out or limit the data they provide. In some jurisdictions, users may even have the right to request a copy of their data in a format that can be transferred to another service provider.

数据可携带性与隐私政策。此外，开发者必须制定AI代理的隐私政策，向用户说明数据的处理方式，详细描述数据收集、使用、存储及用户权利。开发者应确保获得用户对数据收集的同意，尤其是敏感信息。用户通常可选择退出或限制提供的数据。在某些司法管辖区，用户甚至有权请求以可转移格式获取其数据副本，以便转移至其他服务提供商。

Anonymization. For data used in broader analysis or AI training, it should ideally be anonymized to protect individual identities. Developers must understand how their AI agent retrieves and uses historical user data during interactions. This could be for personalization or improving response relevance.

匿名化。用于更广泛分析或AI训练的数据应尽可能匿名化以保护个人身份。开发者必须了解其AI代理在交互中如何检索和使用历史用户数据，可能用于个性化或提升回应相关性。

In summary, understanding data privacy for AI agents involves being aware of how user data is collected, used, stored, and protected, and ensuring that users understand their rights regarding accessing, correcting, and deleting their data. Awareness of the mechanisms for data retrieval, both by users and the AI agent, is also crucial for a comprehensive understanding of data privacy.

总之，理解AI代理的数据隐私涉及了解用户数据如何被收集、使用、存储和保护，并确保用户了解其访问、更正和删除数据的权利。了解用户和AI代理的数据检索机制对于全面理解数据隐私也至关重要。

8.4.7 2.2.4 Interpretability and Explainability

8.4.8 2.2.4 可解释性与可说明性

Imitation Learning → Decoupling. Agents are typically trained using a continuous feedback loop in Reinforcement Learning (RL) or Imitation Learning (IL), starting with a randomly initialized policy. However, this approach faces leader-board in obtaining initial rewards in unfamiliar environments, particularly when rewards are sparse or only available at the end of a long-step interaction. Thus, a superior solution is to use an infinite-memory agent trained through IL, which can learn policies from expert data, improving exploration and utilization of unseen environmental space with emergent infrastructure as shown in Fig. 3. With expert characteristics to help the agent explore better and utilize the unseen environmental space. Agent AI, can learn policies and new paradigm flow directly from expert data.

模仿学习→解耦。代理通常通过强化学习（Reinforcement Learning, RL）或模仿学习（Imitation Learning, IL）中的连续反馈循环进行训练，起始于随机初始化的策略。然而，该方法在陌生环境中获得初始奖励时面临挑战，尤其是在奖励稀疏或仅在长步骤交互结束时才有奖励的情况下。因此，更优的解决方案是使用通过IL训练的无限记忆代理，该代理能够从专家数据中学习策略，提升对未知环境空间的探索和利用，形成如图3所示的涌现基础设施。借助专家特性帮助代理更好地探索和利用未知环境空间。Agent AI能够直接从专家数据中学习策略和新范式流程。

Traditional IL has an agent mimicking an expert demonstrator's behavior to learn a policy. However, learning the expert policy directly may not always be the best approach, as the agent may not generalize well to unseen situations. To tackle this, we propose learning an agent with in-context prompt or a implicit reward function that captures key aspects of the expert's behavior, as shown in Fig. 3. This equips the infinite memory agent with physical-world behavior data for task execution, learned from expert demonstrations. It helps overcome existing imitation learning drawbacks like the need for extensive expert data and potential errors in complex tasks. The key idea behind the Agent AI has two parts: 1) the infinite agent that collects physical-world expert demonstrations as state-action pairs and 2) the virtual environment that imitates the agent generator. The imitating agent produces actions that mimic the expert's behavior, while the agent learns a policy mapping from states to actions by reducing a loss function of the disparity between the expert's actions and the actions generated by the learned policy.

传统的IL中，代理通过模仿专家示范者的行为来学习策略。然而，直接学习专家策略并不总是最佳方法，因为代理可能无法很好地泛化到未见过的情境。为此，我们提出通过上下文提示或隐式奖励函数来学习代理，该函数捕捉专家行为的关键方面，如图3所示。这为无限记忆代理配备了从专家示范中学习的物理世界行为数据，用于任务执行。它有助于克服现有模仿学习的缺点，如需要大量专家数据和复杂任务中的潜在错误。Agent AI的核心思想包括两部分：1) 收集物理世界专家示范作为状态-动作对的无限代理；2) 模拟代理生成器的虚拟环境。模仿代理产生模仿专家行为的动作，代理通过减少专家动作与学习策略生成动作之间差异的损失函数，学习从状态到动作的策略映射。

Decoupling → Generalization. Rather than relying on a task-specific reward function, the agent learns from expert demonstrations, which provide a diverse set of state-action pairs covering various task aspects. The agent then learns a policy that maps states to actions by imitating the expert's behavior. Decoupling in imitation learning refers to separating the learning process from the task-specific reward function, allowing the policy to generalize across different tasks without explicit reliance on the task-specific reward function. By decoupling, the agent can learn from expert demonstrations and learn a policy that is adaptable to a variety of situations. Decoupling enables transfer learning, where a policy learned in one domain can adapt to others with minimal fine-tuning. By learning a general policy that is not tied to a specific reward function, the agent can leverage the knowledge it acquired in one task to perform well in other related tasks. Since the agent does not rely on a specific reward function, it can adapt to changes in the reward function or environment without the need for significant retraining. This makes the learned policy more robust and generalizable across different environments. Decoupling in this context refers to the separation of two tasks in the learning process: learning the reward function and learning the optimal policy.

解耦→泛化。代理不依赖于特定任务的奖励函数，而是从专家示范中学习，这些示范提供了涵盖任务各方面的多样状态-动作对。代理通过模仿专家行为学习从状态到动作的策略。模仿学习中的解耦指的是将学习过程与特定任务奖励函数分离，使策略能够在不同任务间泛化，而无需显式依赖任务奖励函数。通过解耦，代理可以从专家示范中学习，并获得适应多种情境的策略。解耦促进迁移学习，使得在一个领域学到的策略能通过最小微调适应其他领域。学习不依赖特定奖励函数的通用策略，使代理能利用在一项任务中获得的知识，在其他相关任务中表现良好。由于代理不依赖特定奖励函数，它能适应奖励函数或环境的变化，无需大量重新训练，从而使学习到的策略在不同环境中更具鲁棒性和泛化能力。此处的解耦指学习过程中的两个任务分离：学习奖励函数和学习最优策略。

Generalization → Emergent Behavior. Generalization explains how emergent properties or behaviors can arise from simpler components or rules. The key idea lies in identifying the basic elements or rules that govern the behavior of the system, such as individual neurons or basic algorithms. Consequently, by observing how these simple components or rules interact with one another. These interactions of these components often lead to the emergence of complex behaviors, which are not predictable by examining individual components alone. Generalization across different levels of complexity allows a system to learn general principles applicable across these levels, leading to emergent properties. This enables the system to adapt to new situations, demonstrating the emergence of more complex behaviors from simpler rules. Furthermore, the ability to generalize across different complexity levels facilitates knowledge transfer from one domain to another, which contributes to the emergence of complex behaviors in new contexts as the system adapts.

泛化→涌现行为。泛化解释了复杂属性或行为如何从更简单的组成部分或规则中涌现。关键在于识别支配系统行为的基本元素或规则，如单个神经元或基本算法。通过观察这些简单组成部分或规则的相互作用，这些交互常常导致复杂行为的涌现，而单独考察个别组成部分无法预测这些行为。跨不同复杂度层次的泛化使系统能够学习适用于这些层次的通用原理，进而产生涌现属性。这使系统能够适应新情境，展现出由简单规则产生的更复杂行为。此外，跨复杂度层次的泛化能力促进了知识从一个领域向另一个领域的转移，随着系统适应新环境，复杂行为在新情境中得以涌现。

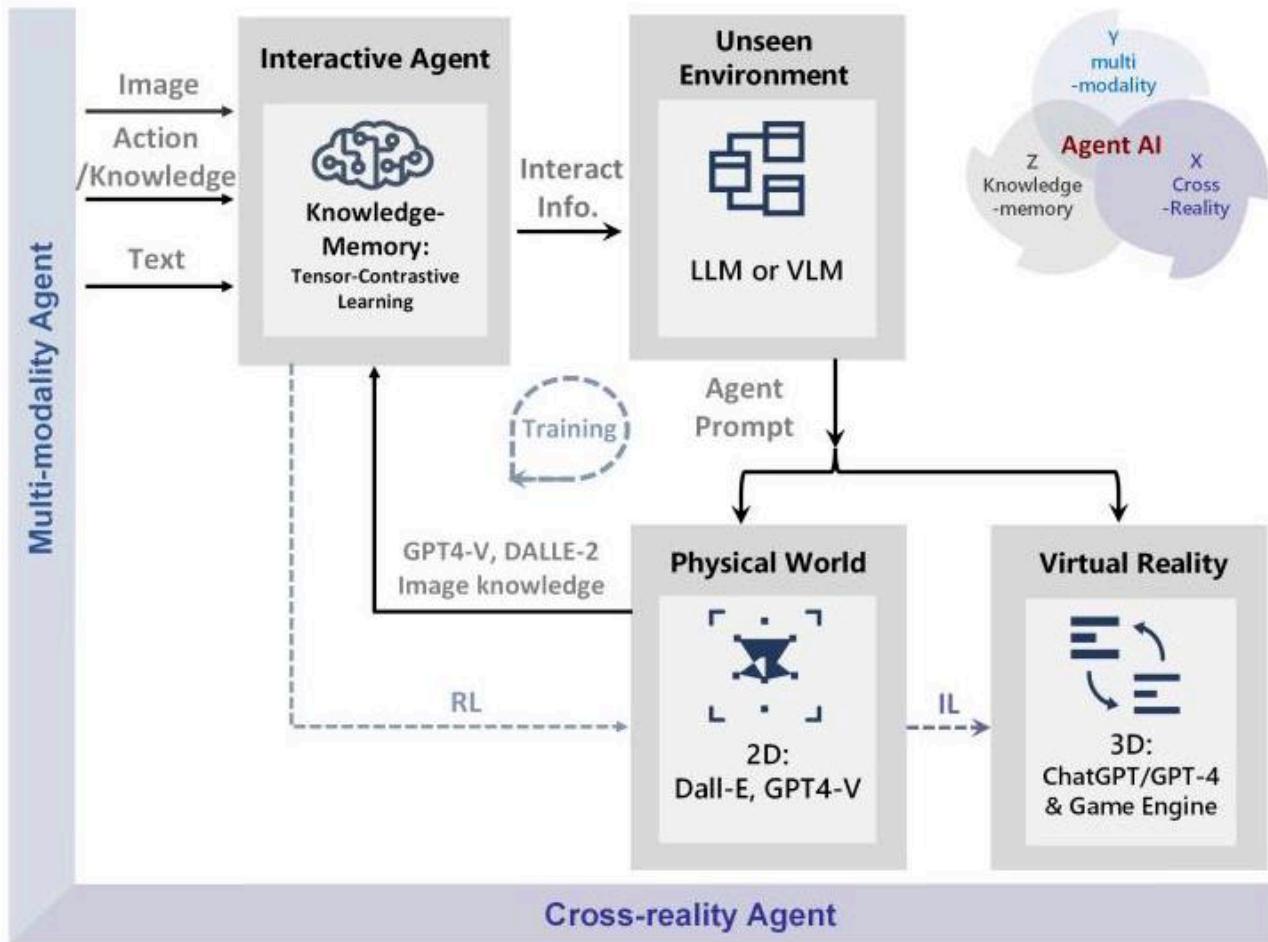


Figure 3: Example of the Emergent Interactive Mechanism using an agent to identify text relevant to the image from candidates. The task involves using a multi-modal AI agent from the web and human-annotated knowledge interaction samples to incorporate external world information.

图3：使用代理从候选文本中识别与图像相关文本的涌现交互机制示例。该任务涉及利用来自网络的多模态AI代理和人工注释的知识交互样本，融合外部世界信息。

8.4.9 2.2.5 Inference Augmentation

8.4.10 2.2.5 推理增强

The inference ability of an AI agent lies in its capacity to interpret, predict, and respond based on its training and input data. While these capabilities are advanced and continually improving, it's important to recognize their limitations and the influence of the underlying data they are trained on. Particularly, in the context of large language models, it refers to its capacity to draw conclusions, make predictions, and generate responses based on the data it has been trained on and the input it receives. Inference augmentation in AI agents refers to enhancing the AI's natural inference abilities with additional tools, techniques, or data to improve its performance, accuracy, and utility. This can be particularly important in complex decision-making scenarios or when dealing with nuanced or specialized content. We denote particularly important sources for inference augmentation below:

AI代理的推理能力在于其基于训练和输入数据进行解释、预测和响应的能力。尽管这些能力先进且不断提升，但必须认识到其局限性以及所依赖训练数据的影响。特别是在大型语言模型（Large Language Models, LLMs）的背景下，推理能力指的是基于训练数据和接收的输入得出结论、做出预测和生成响应的能力。AI代理中的推理增强指通过额外的工具、技术或数据来提升AI的自然推理能力，以提高其性能、准确性和实用性。这在复杂决策场景或处理细微或专业内容时尤为重要。以下列出推理增强中特别重要的来源：

Data Enrichment. Incorporating additional, often external, data sources to provide more context or background can help the AI agent make more informed inferences, especially in areas where its training data may be limited. For example, AI agents can infer meaning from the context of a conversation or text. They analyze the given information and use it to understand the intent and relevant details of user queries. These models are proficient at recognizing patterns in data. They use this ability to make inferences about language, user behavior, or other relevant phenomena based on the patterns they've learned during training.

数据丰富。引入额外的、通常是外部的数据源以提供更多上下文或背景，有助于AI代理做出更有依据的推理，尤其是在其训练数据有限的领域。例如，AI代理可以从对话或文本的上下文中推断含义。它们分析给定信息，用以理解用户查询的意图和相关细节。这些模型擅长识别数据中的模式，利用这一能力基于训练中学到的模式对语言、用户行为或其他相关现象进行推理。

Algorithm Enhancement. Improving the AI's underlying algorithms to make better inferences. This could involve using more advanced machine learning models, integrating different types of AI (like combining NLP with image recognition), or updating algorithms to better handle complex tasks. Inference in language models involves understanding and generating human language. This includes grasping nuances like tone, intent, and the subtleties of different linguistic constructions.

算法增强。改进AI的底层算法以实现更优推理。这可能包括使用更先进的机器学习模型，整合不同类型的AI（如结合自然语言处理（NLP）与图像识别），或更新算法以更好地处理复杂任务。语言模型中的推理涉及理解和生成自然语言，包括把握语气、意图及不同语言结构的细微差别。

Human-in-the-Loop (HITL). Involving human input to augment the AI's inferences can be particularly useful in areas where human judgment is crucial, such as ethical considerations, creative tasks, or ambiguous scenarios. Humans can provide guidance, correct errors, or offer insights that the agent would not be able to infer on its own. 人机协同（Human-in-the-Loop, HITL）。引入人工输入以增强AI推理，在需要人类判断的领域尤为有用，如伦理考量、创意任务或模糊场景。人类可以提供指导、纠正错误或提供AI自身无法推断的见解。

Real-Time Feedback Integration. Using real-time feedback from users or the environment to enhance inferences is another promising method for improving performance during inference. For example, an AI might adjust its recommendations based on live user responses or changing conditions in a dynamic system. Or, if the agent is taking actions in a simulated environment that break certain rules, the agent can be dynamically given feedback to help correct itself.

实时反馈整合。利用来自用户或环境的实时反馈来提升推理，是改进推理性能的另一有效方法。例如，AI可以根据用户的实时反应或动态系统中变化的条件调整其推荐。或者，当代理在模拟环境中执行违反规则的操作时，可以动态给予反馈以帮助其自我纠正。

Cross-Domain Knowledge Transfer. Leveraging knowledge or models from one domain to improve inferences in another can be particularly helpful when producing outputs within a specialized discipline. For instance, techniques developed for language translation might be applied to code generation, or insights from medical diagnostics could enhance predictive maintenance in machinery.

跨领域知识迁移。利用一个领域的知识或模型来提升另一个领域的推理，尤其在专业学科的输出中非常有帮助。例如，语言翻译中开发的技术可能应用于代码生成，或医学诊断中的洞见可增强机械设备的预测性维护。

Customization for Specific Use Cases. Tailoring the AI's inference capabilities for particular applications or industries can involve training the AI on specialized datasets or fine-tuning its models to better suit specific tasks, such as legal analysis, medical diagnosis, or financial forecasting. Since the particular language or information within one domain can greatly contrast with the language from other domains, it can be beneficial to finetune the agent on domain-specific information.

针对特定用例的定制。为特定应用或行业量身定制AI的推理能力，可能涉及在专业数据集上训练AI或微调模型以更好地适应特定任务，如法律分析、医学诊断或金融预测。由于某一领域的语言或信息与其他领域存在显著差异，针对领域特定信息微调代理是有益的。

Ethical and Bias Considerations. It is important to ensure that the augmentation process does not introduce new biases or ethical issues. This involves careful consideration of the sources of additional data or the impact of the new inference augmentation algorithms on fairness and transparency. When making inferences, especially about sensitive topics, AI agents must sometimes navigate ethical considerations. This involves avoiding harmful stereotypes, respecting privacy, and ensuring fairness.

伦理与偏见考量。确保增强过程不引入新的偏见或伦理问题至关重要。这需要对额外数据来源或新推理增强算法对公平性和透明度的影响进行审慎考量。在推理过程中，尤其涉及敏感话题时，AI代理必须处理伦理问题，包括避免有害刻板印象、尊重隐私和确保公平。

Continuous Learning and Adaptation. Regularly updating and refining the AI's capabilities to keep up with new developments, changing data landscapes, and evolving user needs.

持续学习与适应。定期更新和完善AI能力，以跟上新发展、变化的数据环境和不断演变的用户需求。

In summary, inference augmentation in AI agents involves methods in which their natural inference abilities can be enhanced through additional data, improved algorithms, human input, and other techniques. Depending on the use-case, this augmentation is often essential for dealing with complex tasks and ensuring accuracy in the agent's outputs.

总之，AI代理中的推理增强涉及通过额外数据、改进算法、人工输入及其他技术手段提升其自然推理能力。根据具体用例，这种增强通常是处理复杂任务和确保输出准确性的关键。

8.4.11 2.2.6 Regulation

8.4.12 2.2.6 规范

Recently, Agent AI has made significant advancements, and its integration into embodied systems has opened new possibilities for interacting with agents via more immersive, dynamic, and engaging experiences. To expedite the process and ease the cumbersome work in agent AI developing, we are proposing to develop the next-generation AI-empowered pipeline for agent interaction. Develop a human-machine collaboration system where humans and machines can communicate and interact meaningfully. The system can leverage the LLM's or VLM dialog capabilities and vast action to talk with human players and identify human needs. Then it will perform proper actions to help human players upon request.

近年来，Agent AI取得了显著进展，其在具身系统中的集成成为通过更沉浸式、动态且富有吸引力的体验与代理互动开辟了新可能。为加快进程并简化Agent AI开发中的繁琐工作，我们提议开发下一代赋能AI的代理交互流水线。构建一个人机协作系统，使人类与机器能够进行有意义的沟通和互动。该系统可利用大型语言模型（LLM）或视觉语言模型（VLM）的对话能力及丰富动作，与人类玩家交流并识别其需求，然后根据请求执行适当操作以协助玩家。

When employing LLM/VLMs for a human-machine collaboration system, it is essential to note that these operate as black boxes, generating unpredictable output. This uncertainty can become crucial in a physical setup, such as operating actual robotics. An approach to address this challenge is constraining the focus of the LLM/VLM through prompt engineering. For instance, in robotic task planning from instructions, providing environmental information within the prompt has been reported to yield more stable outputs than relying solely on text (Gramopadhye and Szafir, 2022). This report is supported by the Minsky's frame theory of AI (Minsky, 1975), suggesting that the problem space to be solved by LLM/VLMs is defined by the given prompts. Another approach is designing prompts to make LLM/VLMs include explanatory text to allow users understand what the model has focused on or recognized. Additionally, implementing a higher layer that allows for pre-execution verification and modification under human guidance can facilitate the operation of systems working under such guidance (Fig. 4). 在将大型语言模型（LLM）/视觉语言模型（VLM）应用于人机协作系统时，必须注意这些模型作为黑箱运行，输出结果具有不可预测性。这种不确定性在物理环境中尤为关键，例如操作实际机器人。解决该挑战的一种方法是通过提示工程限制LLM/VLM的关注范围。例如，在基于指令的机器人任务规划中，报告显示在提示中提供环境信息比仅依赖文本能产生更稳定的输出（Gramopadhye 和 Szafir, 2022）。该报告得到了明斯基（Minsky, 1975）提出的人工智能框架理论（frame theory）的支持，表明LLM/VLM解决的问题空间由给定的提示定义。另一种方法是设计提

示，使LLM/VLM包含解释性文本，帮助用户理解模型关注或识别的内容。此外，实施一个更高层次的机制，允许在执行前进行验证和在人工指导下修改，有助于促进在此类指导下运行的系统操作（图4）。

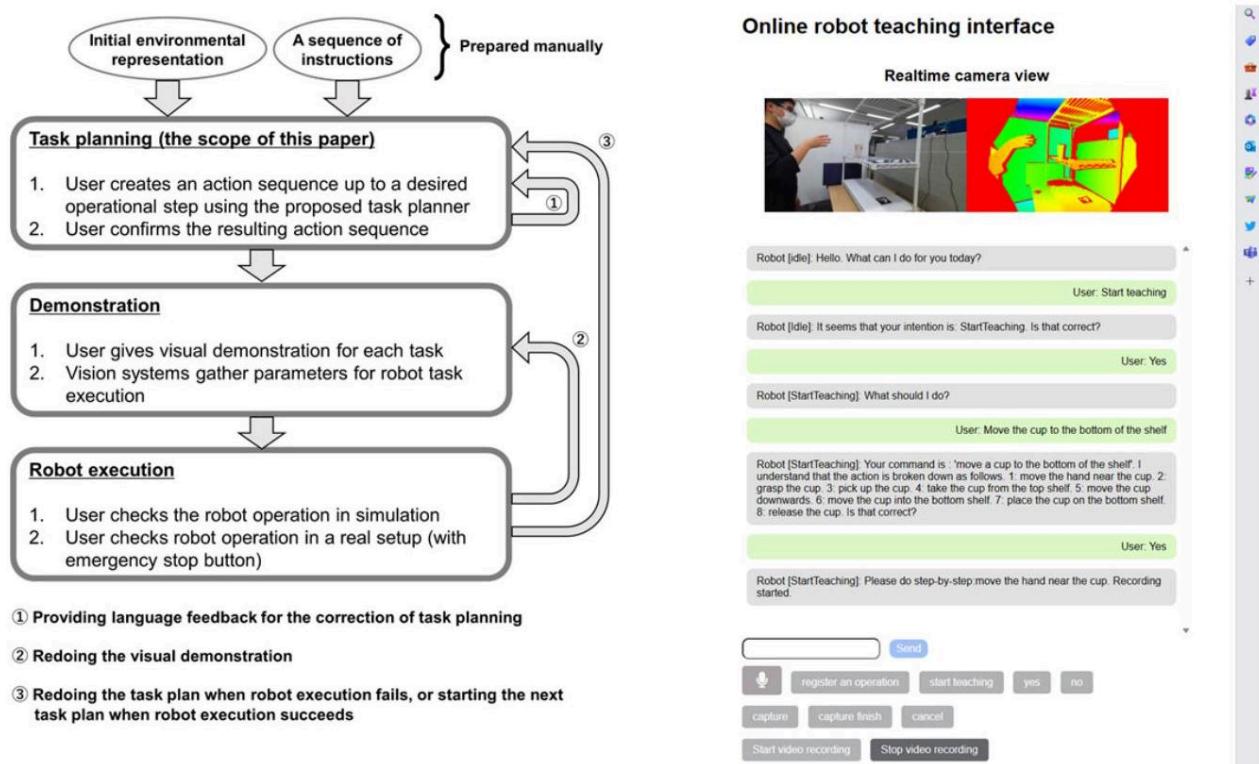


Figure 4: A robot teaching system developed in (Wake et al., 2023c). (Left) The system workflow. The process involves three steps: Task planning, where ChatGPT plans robotic tasks from instructions and environmental information; Demonstration, where the user visually demonstrates the action sequence. All the steps are reviewed by the user, and if any step fails or shows deficiencies, the previous steps can be revisited as necessary. (Right) A web application that enables uploading of demonstration data and the interaction between the user and ChatGPT. 图4：(Wake 等, 2023c) 开发的机器人教学系统。 (左) 系统工作流程。该过程包括三个步骤：任务规划，ChatGPT根据指令和环境信息规划机器人任务；示范，用户直观演示动作序列。所有步骤均由用户审核，如有任何步骤失败或存在不足，可根据需要回溯至前一步骤。 (右) 一个网页应用，支持上传示范数据及用户与ChatGPT之间的交互。

8.5 2.3 Agent AI for Emergent Abilities

8.6 2.3 具备新兴能力的智能体AI

Despite the growing adoption of interactive agent AI systems, the majority of proposed methods still face a challenge in terms of their generalization performance in unseen environments or scenarios. Current modeling practices require developers to prepare large datasets for each domain to finetune/pretrain models; however, this process is costly and even impossible if the domain is new. To address this issue, we build interactive agents that leverage the knowledge-memory of general-purpose foundation models (ChatGPT, Dall-E, GPT-4, etc.) for a novel scenario, specifically for generating a collaboration space between humans and agents. We discover an emergent mechanism- which we name Mixed Reality with Knowledge Inference Interaction-that facilitates collaboration with humans to solve challenging tasks in complex real-world environments and enables the exploration of unseen

environments for adaptation to virtual reality. For this mechanism, the agent learns i) micro-reactions in cross-modality: collecting relevant individual knowledge for each interaction task (e.g., understanding unseen scenes) from the explicit web source and by implicitly inferring from the output of pretrained models; ii) macro-behavior in reality-agnostic: improving interactive dimensions and patterns in language and multi-modality domains, and make changes based on characterized roles, certain target variable, influenced diversification of collaborative information in mixed-reality and LLMs. We investigate the task of knowledge-guided interactive synergistic effects to collaborated scene generation with combining various OpenAI models, and show promising results of how the interactive agent system can further boost the large foundation models in our setting. It integrates and improves the depth of generalization, conscious and interpretability of a complex adaptive AI systems.

尽管交互式代理人工智能系统的应用日益广泛，但大多数现有方法在未见环境或场景中的泛化性能仍面临挑战。当前的建模实践要求开发者为每个领域准备大量数据集以微调/预训练模型；然而，这一过程成本高昂，且在新领域中甚至不可能实现。为解决此问题，我们构建了利用通用基础模型（如ChatGPT、Dall-E、GPT-4等）知识记忆的交互式代理，专门针对新场景，特别是生成人人与代理协作空间。我们发现了一种新兴机制——我们称之为“基于知识推理交互的混合现实”，该机制促进人与代理的协作，以解决复杂现实环境中的挑战性任务，并支持对未见环境的探索，从而适应虚拟现实。针对该机制，代理学习i) 跨模态的微观反应：从显性网络资源收集每个交互任务（如理解未见场景）相关的个体知识，并通过预训练模型输出隐式推理；ii) 与现实无关的宏观行为：提升语言和多模态领域的交互维度与模式，并基于特定角色、目标变量及混合现实与大型语言模型（LLMs）中协作信息的多样化影响进行调整。我们研究了知识引导的交互协同效应任务，结合多种OpenAI模型进行协作场景生成，展示了交互代理系统如何在我们的设置中进一步提升大型基础模型的表现。该系统整合并提升了复杂自适应人工智能系统的泛化深度、意识性和可解释性。

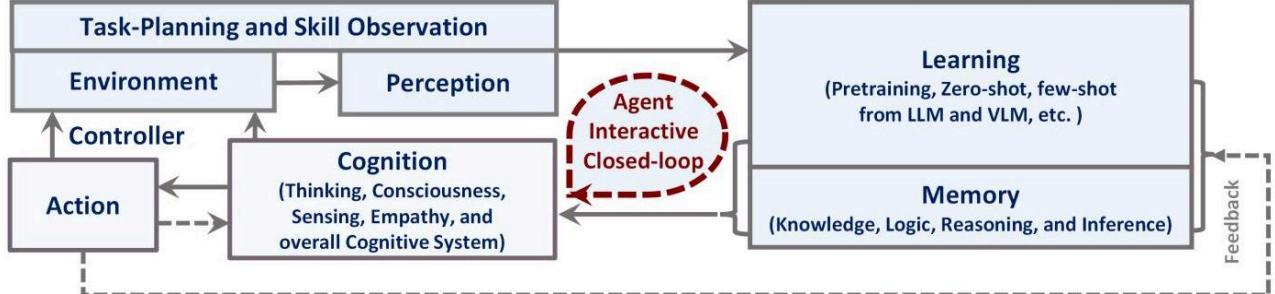


Figure 5: Our proposed new agent paradigm for a multi-modal generalist agent. There are 5 main modules as shown in the figures: 1) Environment and Perception with task-planning and skill observation; 2) Agent learning; 3) Memory; 4) Agent action; 5) Cognition.

图5：我们提出的多模态通用代理新范式。图中展示了5个主要模块：1) 环境与感知，包含任务规划和技能观察；2) 代理学习；3) 记忆；4) 代理动作；5) 认知。

9 3 Agent AI Paradigm

10 3 代理人工智能范式

In this section, we discuss a new paradigm and framework for training Agent AI. We seek to accomplish several goals with our proposed framework:

本节讨论训练代理人工智能的新范式和框架。我们希望通过所提框架实现以下几个目标：

- Make use of existing pre-trained models and pre-training strategies to effectively bootstrap our agents with effective understanding of important modalities, such as text or visual inputs.
- 利用现有预训练模型和预训练策略，有效引导代理对重要模态（如文本或视觉输入）进行有效理解。
- Support for sufficient long-term task-planning capabilities.
- 支持充分的长期任务规划能力。
- Incorporate a framework for memory that allows for learned knowledge to be encoded and retrieved later.
- 融入记忆框架，使学习到的知识能够被编码并在后续检索。
- Allow for environmental feedback to be used to effectively train the agent to learn which actions to take.
- 允许利用环境反馈，有效训练代理学习采取何种行动。

We show a high-level new agent diagram outlining the important submodules of such a system in Fig. 5.

我们在图5中展示了该系统重要子模块的高层次新代理示意图。

10.1 3.1 LLMs and VLMs

10.2 3.1 大型语言模型（LLMs）与视觉语言模型（VLMs）

We can use the LLM or VLM model to bootstrap the components of the Agent as showed in Fig. 5. In particular, LLMs have been shown to perform well for task-planning (Gong et al., 2023a), contain significant world knowledge (Yu et al., 2023b), and display impressive logical reasoning capabilities (Creswell et al., 2022). Additionally, VLMs such as CLIP (Radford et al., 2021) provide a general visual encoder that is language-aligned, as well as providing zero-shot visual recognition capabilities. For example, state-of-the-art open-source multi-modal models such as LLaVA (Liu et al., 2023c) and InstructBLIP (Dai et al., 2023) rely upon frozen CLIP models as visual encoders. 我们可以使用LLM或VLM模型来引导代理组件的构建，如图5所示。特别是，LLMs已被证明在任务规划（Gong等，2023a）、包含丰富世界知识（Yu等，2023b）以及展现出色逻辑推理能力（Creswell等，2022）方面表现优异。此外，诸如CLIP（Radford等，2021）等VLM提供了与语言对齐的通用视觉编码器，并具备零样本视觉识别能力。例如，最先进的开源多模态模型如LLaVA（Liu等，2023c）和InstructBLIP（Dai等，2023）均依赖冻结的CLIP模型作为视觉编码器。

10.3 3.2 Agent Transformer Definition

10.4 3.2 代理变换器定义

Instead of using frozen LLMs and VLMs for the AI agent, it is also possible to use a single-agent transformer model that takes visual tokens and language tokens as input, similar to Gato (Reed et al., 2022). In addition to vision and language, we add a third general type of input, which we denote as agent tokens. Conceptually, agent tokens are used to reserve a specific subspace of the input and output space of the model for agentic behaviors. For robotics or game playing, this may be represented as the input action space of the controller. When training agents to use specific tools, such as image-generation or image-editing models, or for other API calls, agent tokens can also be used. As showed in Fig. 7, we can combine the agent tokens with visual and language tokens to generate a unified interface for training multi-modal agent AI. Compared to using large, proprietary LLMs as agents, there are several advantages to using an agent transformer. Firstly, the model can be easily customized to very specific agentic tasks that may be difficult to represent in natural language (e.g. controller inputs or other specific actions). Thus, the agent can learn from environmental interactions and domain-specific data to improve performance. Secondly, it can be easier to understand why the model does or does not take specific actions by having access to the probabilities of the agent tokens. Thirdly,

除了使用冻结的LLMs和VLMs作为AI代理外，也可以使用单一代理变换器模型，该模型以视觉标记和语言标记作为输入，类似于Gato（Reed等，2022）。除了视觉和语言，我们还添加了第三类通用输入，称为代理标记。概念上，代理标记用于在模型的输入输出空间中保留特定子空间以实现代理行为。对于机器人或游戏玩法，这可能表现为控制

器的输入动作空间。在训练代理使用特定工具（如图像生成或图像编辑模型）或进行其他API调用时，也可使用代理标记。如图7所示，我们可以将代理标记与视觉和语言标记结合，生成用于训练多模态代理AI的统一接口。与使用大型专有LLMs作为代理相比，使用代理变换器有若干优势。首先，该模型可轻松定制为非常具体的代理任务，这些任务可能难以用自然语言表达（如控制器输入或其他特定动作）。因此，代理可以通过环境交互和领域特定数据学习以提升性能。其次，通过访问代理标记的概率，可以更容易理解模型为何采取或不采取特定动作。第三，

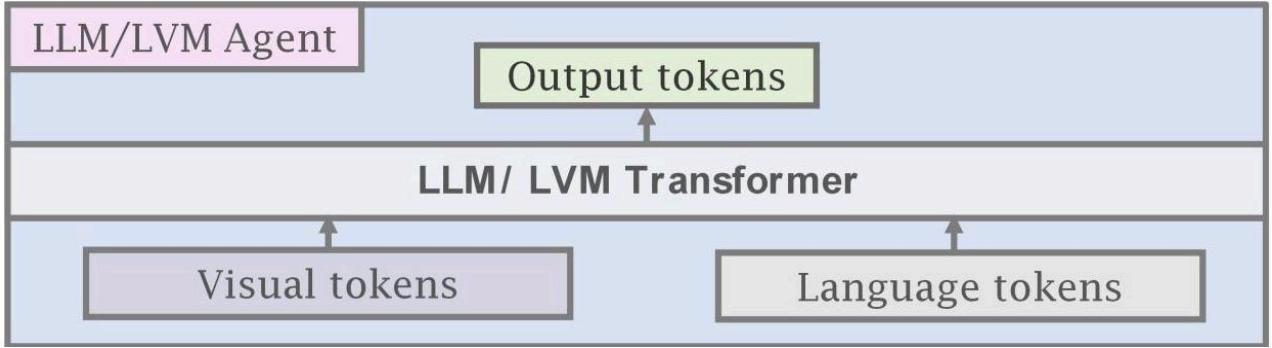


Figure 6: We show the current paradigm for creating multi-modal AI agents by incorporating a Large Language Model (LLM) with a Large Vision Model (LVM). Generally, these models take visual or language inputs and use pre-trained and frozen visual and language models, learning smaller sub-network that connect and bridge modalities. Examples include Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023c), InstructBLIP (Dai et al., 2023), and LLaVA (Liu et al., 2023c).

图6：我们展示了通过结合大型语言模型（LLM）与大型视觉模型（LVM）来创建多模态人工智能代理的当前范式。通常，这些模型接受视觉或语言输入，使用预训练且冻结的视觉和语言模型，学习连接和桥接模态的小型子网络。示例包括Flamingo（Alayrac等，2022）、BLIP-2（Li等，2023c）、InstructBLIP（Dai等，2023）和LLaVA（Liu等，2023c）。

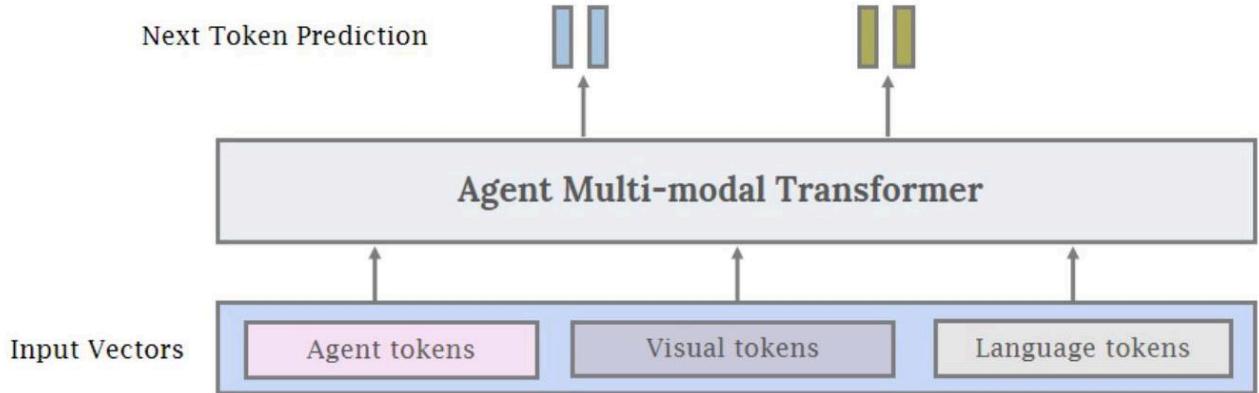


Figure 7: The unified agent multi-modal transformer model. Instead of connecting frozen submodules and using existing foundation models as building blocks, we propose a unified and end-to-end training paradigm for agent systems. We can still initialize the submodules with LLMs and LVMs as in Figure 6 but also make use of agent tokens, specialized tokens for training the model to perform agentic behaviors in a specific domain (e.g., robotics). For more details about agent tokens, see Section 3.2

图7：统一代理多模态变换器模型。我们提出了一种统一的端到端训练范式用于代理系统，而非连接冻结的子模块并

使用现有基础模型作为构建块。我们仍可如图6所示用LLM和LVM初始化子模块，同时利用代理标记（agent tokens），这些是用于训练模型在特定领域（如机器人学）执行代理行为的专用标记。关于代理标记的更多细节见第3.2节。

there are certain domains such as healthcare and law that have strict data privacy requirements. Finally, a relatively smaller agent transformer can potentially be significantly cheaper than a larger proprietary language model.

某些领域如医疗和法律具有严格的数据隐私要求。最后，相对较小的代理变换器在成本上可能远低于大型专有语言模型。

10.5 3.3 Agent Transformer Creation

10.6 3.3 代理变换器的创建

As shown above in Fig. 5, we can use the new agent paradigm with LLM and VLM-bootstrapped agents, as well as leveraging data generated from large foundation models to train the agent transformer model for learning to execute specific goals. Within this process, the agent model is trained to be specialized and tailored for specific tasks and domains. This approach allows you to leverage a pre-existing, foundation model's learned features and knowledge. We show a simplified overview of the process in two steps below:

如图5所示，我们可以使用新的代理范式，结合LLM和VLM引导的代理，以及利用大型基础模型生成的数据，训练代理变换器模型以学习执行特定目标。在此过程中，代理模型被训练为专门针对特定任务和领域定制。该方法允许利用预先存在的基础模型所学的特征和知识。以下以两个步骤简要概述该过程：

Define Objectives within the Domain. In order to train the agent transformer, the objectives and the action-space of the agent within the context of each specific environment needs to be clearly defined. This includes determining which specific tasks or actions the agent needs to perform and assigning unique agent tokens for each. Furthermore, any automatic rules or procedures that can be used to identify successful completion of tasks can significantly improve the amount of data available for training. Otherwise, foundation-model generated or human-annotated data will be required for training the model. After the data is collected and it is possible to evaluate the performance of the agent, the process of continuous improvement can begin.

在领域内定义目标。为了训练代理变换器，需要明确界定代理在每个具体环境中的目标和动作空间。这包括确定代理需要执行的具体任务或动作，并为每个任务分配唯一的代理标记。此外，任何可用于识别任务成功完成的自动规则或程序都能显著提升训练数据量。否则，需依赖基础模型生成的数据或人工标注数据进行训练。数据收集完成且能够评估代理性能后，持续改进过程即可开始。

Continuous Improvement. Continuous monitoring of the model's performance and collection of feedback are essential steps in the process. Feedback should be used for further fine-tuning and updates. It is also crucial to ensure that the model does not perpetuate biases or unethical outcomes. This necessitates a careful examination of the training data, regular checks for biases in outputs, and, if needed, training the model to recognize and avoid biases. Once the model achieves satisfactory performance, it can be deployed for the intended application.

Continuous monitoring remains vital to ensure that the model performs as expected and to facilitate necessary adjustments. More details on this process, sources of training data, and details surrounding continuous learning for agent AI can be found in Section 8.

持续改进。持续监控模型性能和收集反馈是关键步骤。反馈应被用于进一步微调和更新。确保模型不延续偏见或产生不道德结果也至关重要，这需要对训练数据进行仔细审查，定期检查输出中的偏见，并在必要时训练模型识别并避免偏见。一旦模型达到满意的性能，即可部署于预期应用。持续监控依然重要，以确保模型表现符合预期并便于进行必要调整。关于该过程、训练数据来源及代理AI持续学习的更多细节见第8节。

11 4 Agent AI Learning

12 4 代理人工智能学习

12.1 4.1 Strategy and Mechanism

12.2 4.1 策略与机制

The strategy of interactive AI on different domains which extends the paradigm of calling large foundation models with a trained agent that actively seeks to collect user feedback, action information, useful knowledge for generation and interaction. Some times, the LLM/VLM models are not need to trained again, and we improve their performance by providing improved contextual prompts at test time for an agent. On the other hand, it always involves a knowledge/reasoning/commonsense/inference interactive modeling through a combination of triple systems - one performing knowledge retrieval from multi-model query, second performing interactive generation from the relevant agent, and last one the trained a new, informative self-supervised training or pre-training with reinforcement learning or imitation learning with improved way.

交互式人工智能在不同领域的策略，扩展了调用大型基础模型的范式，配合训练有素的代理主动收集用户反馈、动作信息及生成和交互所需的有用知识。有时，LLM/VLM模型无需重新训练，我们通过在测试时为代理提供改进的上下文提示来提升其性能。另一方面，这总是涉及通过三重系统的组合进行知识/推理/常识/推断的交互建模——一是从多模态查询中执行知识检索，二是从相关代理执行交互生成，三是通过改进方式进行新的信息自监督训练或预训练，结合强化学习或模仿学习。

12.2.1 4.1.1 Reinforcement Learning (RL)

12.2.2 4.1.1 强化学习 (RL)

There is a rich history of leveraging reinforcement learning (RL) to train interactive agents that exhibits intelligent behaviors. RL is a methodology to learn the optimal relationship between states and actions based on rewards (or penalties) received as a result of its actions. RL is a highly scalable framework that has been applied to numerous applications including robotics, however, it generally faces several leader-board and LLM/VLMs have shown their potential to mitigate or overcome some of those difficulties:

利用强化学习 (RL) 训练展现智能行为的交互代理有着丰富的历史。RL是一种基于奖励（或惩罚）学习状态与动作之间最优关系的方法。RL是高度可扩展的框架，已应用于包括机器人学在内的众多领域，但通常面临若干挑战，LLM/VLM显示出缓解或克服这些困难的潜力：

- Reward designing The efficiency of policy learning greatly depends on the design of the reward function. Designing the reward function requires not only knowledge of RL algorithms but also a deep understanding of the nature of the task, and thus often necessitates crafting the function based on expert experience. Several studies explored the use of LLM/VLMs for designing reward functions (Yu et al., 2023a; Katara et al., 2023; Ma et al., 2023).
- 奖励设计 策略学习的效率极大依赖于奖励函数的设计。设计奖励函数不仅需要RL算法知识，还需深入理解任务本质，因此通常基于专家经验进行设计。多项研究探索了使用LLM/VLM设计奖励函数（Yu等，2023a；Katara等，2023；Ma等，2023）。
- Data collection and efficiency Given its exploratory nature, RL-based policy learning requires a significant amount of data (Padalkar et al., 2023). The necessity for extensive data becomes particularly evident when the policy involves managing long sequences or integrating complex actions. This is because these scenarios demand more nuanced decision-making and learning from a wider range of situations. In recent studies, efforts have been directed towards enhancing data generation to support policy learning (Kumar et al., 2023; Du et al., 2023). Additionally, in some studies, these models have been integrated into the reward function to improve policy learning (Sontakke et al., 2023). Parallel to these developments, another strand of research has focused

on achieving parameter efficiency in learning processes using VLMs (Tang et al., 2023; Li et al., 2023d) and LLMs (Shi et al., 2023)

- 数据收集与效率 鉴于其探索性特征，基于强化学习（RL）的策略学习需要大量数据（Padalkar 等, 2023）。当策略涉及管理长序列或整合复杂动作时，对大量数据的需求尤为明显。这是因为这些场景要求更细致的决策制定，并从更广泛的情境中学习。近期研究致力于增强数据生成以支持策略学习（Kumar 等, 2023; Du 等, 2023）。此外，在一些研究中，这些模型被集成到奖励函数中以提升策略学习效果（Sontakke 等, 2023）。与这些进展并行，另一研究方向聚焦于利用视觉语言模型（VLMs）（Tang 等, 2023; Li 等, 2023d）和大型语言模型（LLMs）（Shi 等, 2023）实现学习过程中的参数效率。
- Long-horizon steps In relation to the issue of data efficiency, RL becomes more challenging as the length of action sequences increases. This is due to the ambiguity in the relationship between actions and rewards, known as the credit assignment problem, and the increase in the number of states to be explored, necessitating a significant amount of time and data. One typical approach for long and complex tasks is to break them down into a sequence of subgoals and apply pretrained policies to solve each subgoal (e.g., (Takamatsu et al., 2022)). This idea falls within the framework called the task and motion planning (TAMP)(Garrett et al., 2021). TAMP is composed of two primary components: task planning, which entails identifying sequences of high-level actions, and motion planning, which involves finding physically consistent, collision-free trajectories to achieve the objectives of the task plan.
- 长期步骤 关于数据效率问题，随着动作序列长度的增加，强化学习变得更加具有挑战性。这是由于动作与奖励之间关系的模糊性，即所谓的信用分配问题，以及需要探索的状态数量增加，导致需要大量时间和数据。处理长且复杂任务的典型方法是将其分解为一系列子目标，并应用预训练策略解决每个子目标（例如，Takamatsu 等, 2022）。这一思路属于任务与运动规划（TAMP）（Garrett 等, 2021）框架。TAMP由两个主要部分组成：任务规划，涉及识别高层次动作序列；运动规划，涉及寻找物理一致且无碰撞的轨迹以实现任务规划的目标。

LLMs are well-suited to TAMP, and recent research has often adopted an approach where LLMs are used to execute high-level task planning, while low-level controls are addressed with RL-based policies (Xu et al., 2023; Sun et al., 2023a; Li et al., 2023b; Parakh et al., 2023). The advanced capabilities of LLMs enable them to effectively decompose even abstract instructions into subgoals (Wake et al., 2023c), contributing to the enhancement of language understanding abilities in robotic systems.

大型语言模型（LLMs）非常适合任务与运动规划（TAMP），近期研究常采用由LLMs执行高层任务规划，而低层控制则由基于强化学习的策略处理的方法（Xu 等, 2023; Sun 等, 2023a; Li 等, 2023b; Parakh 等, 2023）。LLMs的先进能力使其能够有效地将抽象指令分解为子目标（Wake 等, 2023c），促进机器人系统语言理解能力的提升。

12.2.3 4.1.2 Imitation Learning (IL)

12.2.4 4.1.2 模仿学习 (IL)

While RL aims to train a policy based on exploratory behavior and maximizing rewards through interactions with the environment, imitation learning (IL) seeks to leverage expert data to mimic the actions of experienced agents or experts. For example, in robotics, one of the major frameworks based on IL is Behavioral Cloning (BC). BC is an approach where a robot is trained to mimic the actions of an expert by directly copying them. In this approach, the expert's actions in performing specific tasks are recorded, and the robot is trained to replicate these actions in similar situations. Recent BC-based methods often incorporate technologies from LLM/VLMs, enabling more advanced end-to-end models. For example, Brohan et al. proposed RT-1 (Brohan et al., 2022) and RT-2 (Brohan et al., 2023), transformer-based models that output an action sequence for the base and arm, taking a series of images and language as input. These models are reported to show high generalization performance as the result of training on a large amount of training data.

强化学习旨在通过与环境的交互基于探索行为和最大化奖励来训练策略，而模仿学习（IL）则试图利用专家数据模仿经验丰富的代理或专家的行为。例如，在机器人领域，基于IL的主要框架之一是行为克隆（Behavioral Cloning），

BC)。BC是一种通过直接复制专家动作来训练机器人模仿专家行为的方法。在该方法中，记录专家执行特定任务时的动作，机器人则被训练在类似情境下复制这些动作。近期基于BC的方法常结合LLM/VLM技术，实现更先进的端到端模型。例如，Brohan 等提出了RT-1 (Brohan 等, 2022) 和RT-2 (Brohan 等, 2023)，这两种基于Transformer的模型以一系列图像和语言作为输入，输出底盘和机械臂的动作序列。据报道，这些模型因训练于大量数据而表现出较高的泛化能力。

12.2.5 4.1.3 Traditional RGB

12.2.6 4.1.3 传统RGB

Learning intelligent agent behavior leveraging image inputs has been of interest for many years (Mnih et al., 2015). The inherent challenge of using RGB input is the curse of dimensionality. To solve this problem, researchers either use more data (Jang et al., 2022; Ha et al., 2023) or introduce inductive biases into the model design to improve sample efficiency. In particular, authors incorporate 3D structures into the model architecture for manipulations (Zeng et al., 2021; Shridhar et al., 2023; Goyal et al., 2023; James and Davison, 2022). For robot navigation, authors (Chaplot et al., 2020a,b) leverage maps as a representation. Maps can either be learned from a neural network aggregating all previous RGB inputs or through 3D reconstruction methods such as Neural Radiance Fields (Rosinol et al., 2022).

利用图像输入学习智能代理行为多年来一直备受关注 (Mnih 等, 2015)。使用RGB输入的固有挑战是维度灾难。为解决此问题，研究者要么使用更多数据 (Jang 等, 2022; Ha 等, 2023)，要么在模型设计中引入归纳偏置以提升样本效率。特别地，作者将三维结构融入模型架构以实现操作 (Zeng 等, 2021; Shridhar 等, 2023; Goyal 等, 2023; James 和 Davison, 2022)。在机器人导航方面，作者 (Chaplot 等, 2020a,b) 利用地图作为表示。地图可以通过神经网络从所有先前的RGB输入中学习获得，或通过神经辐射场 (Neural Radiance Fields, NeRF) (Rosinol 等, 2022) 等三维重建方法生成。

To obtain more data, researchers synthesize synthetic data using graphics simulators (Mu et al., 2021; Gong et al., 2023b), and try to close the sim2real gap (Tobin et al., 2017; Sadeghi and Levine, 2016; Peng et al., 2018).

Recently, there has been some collective effort to curate large-scale dataset that aims to resolve the data scarcity problem (Padalkar et al., 2023; Brohan et al., 2023). On the other hand, to improve sample complexity, data augmentation techniques have been extensively studied as well (Zeng et al., 2021; Rao et al., 2020; Haarnoja et al., 2023; Lifshitz et al., 2023).

为了获取更多数据，研究人员使用图形模拟器合成合成数据 (Mu 等, 2021; Gong 等, 2023b)，并尝试缩小模拟到现实的差距 (sim2real gap) (Tobin 等, 2017; Sadeghi 和 Levine, 2016; Peng 等, 2018)。最近，已有一些集体努力致力于策划大规模数据集，旨在解决数据稀缺问题 (Padalkar 等, 2023; Brohan 等, 2023)。另一方面，为了提高样本复杂度，数据增强技术也得到了广泛研究 (Zeng 等, 2021; Rao 等, 2020; Haarnoja 等, 2023; Lifshitz 等, 2023)。

12.2.7 4.1.4 In-context Learning

12.2.8 4.1.4 上下文学习

In-context learning was shown to be an effective method for solving tasks in NLP with the advent of large language models like GPT-3 (Brown et al., 2020; Min et al., 2022). Few-shot prompts were seen to be an effective way to contextualize model output's across a variety of tasks in NLP by providing examples of the task within the context of the LLM prompt. Factors like the diversity of examples and quality of examples shown for the in-context demonstrations may improve the quality of model outputs (An et al., 2023; Dong et al., 2022). Within the context of multi-modal foundation models, models like Flamingo and BLIP-2 (Alayrac et al., 2022; Li et al., 2023c) have been shown to be effective at a variety of visual understanding tasks when given only a small number of examples. In context learning can be further improved for agents within environments by incorporating environment-specific feedback when certain actions are taken (Gong et al., 2023a).

随着大型语言模型 (LLM) 如 GPT-3 (Brown 等, 2020; Min 等, 2022) 的出现，上下文学习被证明是解决自然语言处理 (NLP) 任务的有效方法。少样本提示被视为一种有效的方式，通过在 LLM 提示中提供任务示例，使模型输

出在各种 NLP 任务中具有上下文相关性。示例的多样性和质量等因素可能提升上下文演示中模型输出的质量 (An 等, 2023; Dong 等, 2022)。在多模态基础模型的背景下, Flamingo 和 BLIP-2 (Alayrac 等, 2022; Li 等, 2023c) 等模型在仅给出少量示例时, 已被证明在多种视觉理解任务中表现出色。通过在环境中结合特定环境反馈, 当采取某些动作时, 可以进一步提升智能体的上下文学习能力 (Gong 等, 2023a)。

12.2.9 4.1.5 Optimization in the Agent System

12.2.10 4.1.5 智能体系统中的优化

The optimization of agent systems can be divided into spatial and temporal aspects. Spatial optimization considers how agents operate within a physical space to execute tasks. This includes inter-robot coordination, resource allocation, and keeping an organized space.

智能体系统的优化可以分为空间和时间两个方面。空间优化考虑智能体如何在物理空间中执行任务, 包括机器人间的协调、资源分配以及保持空间的有序性。

In order to effectively optimize agent AI systems, especially systems with large numbers of agents acting in parallel, previous works have focused on using large batch reinforcement learning (Shacklett et al., 2023). Since datasets of multi-agent interactions for specific tasks are rare, self-play reinforcement learning enables a team of agents to improve over time. However, this may also lead to very brittle agents that can only work under self-play and not with humans or other independent agents since they over-fit to the self-play training paradigm. To address this issue, we can instead discover a diverse set of conventions (Cui et al., 2023; Sarkar et al., 2023), and train an agent that is aware of a wide range of conventions. Foundation models can further help to establish conventions with humans or other independent agents, enabling smooth coordination with new agents.

为了有效优化智能体 AI 系统, 尤其是大量智能体并行行动的系统, 先前的工作集中于使用大批量强化学习 (Shacklett 等, 2023)。由于特定任务的多智能体交互数据集稀缺, 自我博弈强化学习使智能体团队能够随着时间提升性能。然而, 这也可能导致智能体过于脆弱, 仅能在自我博弈环境下工作, 无法与人类或其他独立智能体协作, 因为它们过拟合于自我博弈训练范式。为解决此问题, 我们可以发现多样化的约定集合 (Cui 等, 2023; Sarkar 等, 2023), 并训练能够识别广泛约定的智能体。基础模型还能进一步帮助与人类或其他独立智能体建立约定, 实现与新智能体的顺畅协调。

Temporal optimization, on the other hand, focuses on how agents execute tasks over time. This encompasses task scheduling, sequencing, and timeline efficiency. For instance, optimizing the trajectory of a robot's arm is an example of efficiently optimizing movement between consecutive tasks (Zhou et al., 2023c). At the level of task scheduling, methods like LLM-DP (Dagan et al., 2023) and ReAct (Yao et al., 2023a) have been proposed to solve efficient task planning by incorporating environmental factors interactively.

时间优化则关注智能体如何随时间执行任务, 涵盖任务调度、排序和时间线效率。例如, 优化机器人手臂轨迹是高效优化连续任务间运动的一个实例 (Zhou 等, 2023c)。在任务调度层面, 诸如 LLM-DP (Dagan 等, 2023) 和 ReAct (Yao 等, 2023a) 等方法被提出, 通过交互式地结合环境因素, 实现高效任务规划。

12.3 4.2 Agent Systems (zero-shot and few-shot level)

12.4 4.2 智能体系统 (零样本和少样本层面)

12.4.1 4.2.1 Agent Modules

12.4.2 4.2.1 智能体模块

Our foray into the agent paradigm involves the development of Agent AI "Modules" for interactive multi-modal agents using LLMs or VLMs. Our initial Agent Modules facilitate training or in-context learning and adopt a minimalist design for the purposes of demonstrating the agent's ability to schedule and coordinate effectively. We also explored initial prompt-based memory techniques that facilitate better planning and inform future actions approaches within the domain. To illustrate, our "MindAgent" infrastructure comprises 5 main modules: 1) environment perception with task planning, 2) agent learning, 3) memory, 4) general agent action prediction and 5)

cognition, as shown in Figure 5.

我们在智能体范式上的探索涉及使用大型语言模型（LLM）或视觉语言模型（VLM）开发用于交互式多模态智能体的智能体 AI “模块”。我们的初始智能体模块支持训练或上下文学习，并采用极简设计，以展示智能体有效调度和协调的能力。我们还探索了基于提示的初步记忆技术，以促进更好的规划并指导领域内未来的行动方法。举例来说，我们的“MindAgent”架构包含五个主要模块：1) 环境感知与任务规划，2) 智能体学习，3) 记忆，4) 通用智能体动作预测，5) 认知，如图5所示。

12.4.3 4.2.2 Agent Infrastructure

12.4.4 4.2.2 智能体基础设施

Agent-based AI is a large and fast-growing community within the domains of entertainment, research, and industry. The development of large foundation models has significantly improved the performance of agent AI systems.

However, creating agents in this vein is limited by the increasing effort necessary to create high-quality datasets and overall cost. At Microsoft, building high-quality agent infrastructure has significantly impacted multi-modal agent copilots by using advanced hardware, diverse data sources, and powerful software libraries. As Microsoft continues to push the boundaries of agent technology, AI agent platforms are poised to remain a dominant force in the world of multimodal intelligence for years to come. Nevertheless, agent AI interaction is currently still a complex process that requires a combination of multiple skills. The recent advancements in the space of large generative AI models have the potential to greatly reduce the current high cost and time required for interactive content, both for large studios, as well as empowering smaller independent content creators to design high quality experiences beyond what they are currently capable of. The current human-machine interaction systems inside multi-modal agents are primarily rule-based. They do have intelligent behaviors in response to human/user actions and possess web knowledge to some extent. However, these interactions are often limited by software development costs to enable specific behaviors in the system. In addition, current models are not designed to help human to achieve a goal in the case of users' inability to achieve specific tasks. Therefore, there is a need for an agent AI system infrastructure to analyze users behaviors and provide proper support when needed.

基于代理的人工智能（Agent-based AI）是娱乐、科研和工业领域中一个庞大且快速发展的社区。大型基础模型的发展显著提升了代理AI系统的性能。然而，创建此类代理受限于制作高质量数据集所需的日益增加的努力和整体成本。在微软，构建高质量的代理基础设施通过使用先进硬件、多样化数据源和强大软件库，极大地推动了多模态代理助手的发展。随着微软不断推动代理技术的边界，AI代理平台有望在未来多年内继续成为多模态智能领域的主导力量。然而，目前代理AI的交互仍然是一个复杂的过程，需要多种技能的结合。大型生成式AI模型领域的最新进展有望大幅降低当前交互内容所需的高昂成本和时间，不仅惠及大型工作室，也赋能小型独立内容创作者设计出超越现有能力的高质量体验。目前多模态代理中的人机交互系统主要基于规则。它们对人类/用户的行为有一定的智能响应，并在一定程度上具备网络知识。然而，这些交互常因软件开发成本限制而难以实现系统中特定行为。此外，现有模型并未设计用于在用户无法完成特定任务时帮助其达成目标。因此，亟需一种代理AI系统基础设施，能够分析用户行为并在必要时提供适当支持。

12.5 4.3 Agentic Foundation Models (pretraining and finetune level)

12.6 4.3 代理基础模型（预训练与微调层面）

The use of pre-trained foundation models offers a significant advantage in their wide applicability across diverse use cases. The integration of these models enables the development of customized solutions for various applications, circumventing the need for extensive labeled datasets for each specific task.

预训练基础模型的使用在其广泛适用性方面具有显著优势。这些模型的整合使得针对各种应用开发定制化解决方案成为可能，避免了为每个具体任务准备大量标注数据集的需求。

A notable example in the field of navigation is the LM-Nav system (Shah et al., 2023a), which incorporates GPT-3 and CLIP in a novel approach. It effectively uses textual landmarks generated by the language model, anchoring them in images acquired by robots for navigation. This method demonstrates a seamless fusion of textual and visual data, significantly enhancing the capabilities of robotic navigation, while maintaining wide applicability.

导航领域的一个显著例子是LM-Nav系统 (Shah等, 2023a) , 该系统创新性地结合了GPT-3和CLIP。它有效利用语言模型生成的文本地标，并将其锚定于机器人获取的图像中以实现导航。该方法展示了文本与视觉数据的无缝融合，显著提升了机器人导航能力，同时保持了广泛的适用性。

In robot manipulation, several studies have proposed the use of off-the-shelf LLMs (e.g., ChatGPT) while using open vocabulary object detectors. The combination of LLM and advanced object detectors (e.g., Detic (Zhou et al., 2022)) facilitates the understanding of human instruction while grounding the textual information in scenery information (Parakh et al., 2023). Furthermore, the latest advancements showcase the potential of using prompt engineering with advanced multi-modal models such as GPT-4V(ision) (Wake et al., 2023b). This technique opens avenues for multi-modal task planning, underscoring the versatility and adaptability of pre-trained models in a variety of contexts.

在机器人操作领域，多项研究提出了使用现成的大型语言模型（如ChatGPT）结合开放词汇对象检测器的方法。大型语言模型与先进对象检测器（如Detic（Zhou等, 2022））的结合，有助于理解人类指令，同时将文本信息与场景信息相结合（Parakh等, 2023）。此外，最新进展展示了利用提示工程与先进多模态模型如GPT-4V(ision) (Wake等, 2023b) 相结合的潜力。这一技术为多模态任务规划开辟了新途径，凸显了预训练模型在多种情境中的多样性和适应性。

13 5 Agent AI Categorization

14 5 代理AI分类

14.1 5.1 Generalist Agent Areas

14.2 5.1 通用代理领域

Computer-based action and generalist agents (GAs) are useful for many tasks. Recent progress in the field of large foundation models and interactive AI has enabled new functionalities for GAs. However, for a GA to become truly valuable to its users, it must be natural to interact with, and generalize to a broad range of contexts and modalities. We high-quality extended main Chapters on Agent foundation AI in Sec.6, especially in areas relevant to the themes in general of these topics:

基于计算机的动作代理和通用代理 (GAs) 在许多任务中非常有用。大型基础模型和交互式AI领域的最新进展为通用代理带来了新功能。然而，要使通用代理真正对用户有价值，必须具备自然的交互方式，并能广泛适应多种情境和模态。我们在第6节中对代理基础AI的主要章节进行了高质量扩展，特别是在与这些主题相关的领域：

Multimodal Agent AI (MMA) is an upcoming forum² for our research and industry communities to engage with each other and with the broader research and technology communities in Agent AI. Recent progress in the field of large foundation models and interactive AI has enabled new functionalities for generalist agents (GAs), such as predicting user actions and task planning in constrained settings (e.g., MindAgent (Gong et al., 2023a), fine-grained multimodal video understanding (Luo et al., 2022), Robotics (Ahn et al., 2022b; Brohan et al., 2023)), or providing a chat companion for users that incorporates knowledge feedback (e.g., website customer support for healthcare systems (Peng et al., 2023)). More details about the representative works and most recent representative works are shown below. We hope to discuss our vision for the future of MAA and inspire future researchers to work in this space. This article and our forum covers the following main topics, but is not limited exclusively to these:

多模态代理AI (MMA) 是一个即将到来的论坛²，旨在促进我们的研究和工业社区之间以及与更广泛的代理AI研究和技术社区的交流。大型基础模型和交互式AI领域的最新进展为通用代理 (GAs) 带来了新功能，如在受限环境中预测用户行为和任务规划（例如，MindAgent (Gong等, 2023a) 、细粒度多模态视频理解 (Luo等, 2022) 、机器人技术 (Ahn等, 2022b; Brohan等, 2023) ），或为用户提供包含知识反馈的聊天伴侣（例如，医疗系统网站客户支持 (Peng等, 2023) ）。以下展示了代表性工作及最新代表性成果的更多细节。我们希望讨论对MMA未来的愿景，并激励未来研究者投身该领域。本文及我们的论坛涵盖以下主要主题，但不限于此：

- Primary Subject Topics: Multimodal Agent AI, General Agent AI

- 主要主题：多模态代理AI，通用代理AI
- Secondary Subject Topics: Embodied Agents, Action Agents, Language-based Agents, Vision & Language Agents, Knowledge and Inference Agents, Agents for Gaming, Robotics, Healthcare, etc.
- 次要主题：具身代理、动作代理、基于语言的代理、视觉与语言代理、知识与推理代理、游戏代理、机器人技术、医疗等领域的代理
- Extend Subject Topics: Visual Navigation, Simulation Environments, Rearrangement, Agent Foundation Models, VR/AR/MR, Embodied Vision & Language.
- 拓展主题：视觉导航、仿真环境、重排、代理基础模型、虚拟现实/增强现实/混合现实（VR/AR/MR）、具身视觉与语言

Next, we present a specific lists of representative agent categories as follows:

接下来，我们展示一份具体的代表性代理类别列表如下：

14.3 5.2 Embodied Agents

14.4 5.2 具身代理

Our biological minds live in bodies, and our bodies move through a changing world. The goal of embodied artificial intelligence is to create agents, such as robots, which learn to creatively solve challenging tasks requiring interaction with the environment. While this is a significant challenge, important advances in deep learning and the increasing availability of large datasets like ImageNet have enabled superhuman performance on a variety of AI tasks previously thought intractable. Computer vision, speech recognition and natural language processing have experienced transformative revolutions at passive input-output tasks like language translation and image classification, and reinforcement learning has similarly achieved world-class performance at interactive tasks like game playing. These advances have supercharged embodied AI, enabling a growing collection of users to make rapid progress towards intelligent agents can interactive with machine.

我们的生物大脑生活在身体中，而我们的身体在不断变化的世界中移动。具身人工智能的目标是创建能够学习创造性地解决需要与环境交互的复杂任务的代理，例如机器人。虽然这是一个重大挑战，但深度学习的重要进展以及像ImageNet这样的大型数据集的日益普及，使得在许多此前被认为难以解决的人工智能任务上实现了超越人类的表现。计算机视觉、语音识别和自然语言处理在语言翻译和图像分类等被动输入输出任务上经历了变革性革命，强化学习在游戏等交互任务上同样达到了世界级水平。这些进展极大推动了具身人工智能的发展，使越来越多的用户能够快速推进智能代理与机器的交互能力。

14.4.1 5.2.1 Action Agents

14.4.2 5.2.1 行动代理

Action agents refer to the agents that need to execute physical actions in the simulated physical environment or real world. In particular, they need to be actively engaging in activities with the environment. We broadly classify action agents into two different categories based on their application domains: gaming AI and robotics.

行动代理指的是需要在模拟物理环境或现实世界中执行物理动作的代理。特别是，它们需要积极地与环境进行互动。我们根据应用领域将行动代理大致分为两类：游戏人工智能和机器人。

In gaming AI, the agents will interact with the game environment and other independent entities. In these settings, natural language can enable smooth communication between agents and humans. Depending on the game, there may be a specific task to accomplish, providing a true reward signal. For instance, in the competitive Diplomacy game, training a language model using human conversation data along with an action policy with RL enables human-level play (Meta Fundamental AI Research (FAIR) Diplomacy Team et al., 2022).

在游戏人工智能中，代理将与游戏环境及其他独立实体互动。在这些场景中，自然语言能够实现代理与人类之间的顺畅交流。根据游戏不同，可能存在特定的任务需要完成，从而提供真实的奖励信号。例如，在竞技性外交游戏中，利

用人类对话数据训练语言模型，并结合强化学习的行动策略，实现了人类水平的游戏表现（Meta Fundamental AI Research (FAIR) Diplomacy Team 等, 2022）。

² Current URL: <https://multimodalagentai.github.io/>

² 当前网址: <https://multimodalagentai.github.io/>

There are also settings where agents act as normal residents in a town (Park et al., 2023a), without trying to optimize a specific goal. Foundation models are useful in these settings because they can model interactions that appear more natural by mimicking human behavior. When augmented with external memory, they produce convincing agents that can have conversations, daily schedules, form relationships, and have a virtual life.

还有一些场景中，代理作为城镇中的普通居民行动（Park 等, 2023a），不试图优化特定目标。基础模型在这些场景中非常有用，因为它们通过模仿人类行为，能够模拟更自然的互动。结合外部记忆，这些模型能够生成可信的代理，进行对话、安排日常日程、建立关系并拥有虚拟生活。

14.4.3 5.2.2 Interactive Agents

14.4.4 5.2.2 交互代理

Interactive agents simply refer to agents that can interact with the world, a broader class of agents than action agents. Their forms of interaction do not necessarily require physical actions, but may involve communicating information to users or modifying the environment. For instance, an embodied interactive agent may answer a user's questions about a topic through dialogue or help users parse through existing information similar to a chatbot. By extending an agent's capabilities to include information sharing, the core designs and algorithms of Agent AI can be effectively adapted for a range of applications, such as diagnostic (Lee et al., 2023) and knowledge-retrieval (Peng et al., 2023) agents.

交互代理泛指能够与世界互动的代理，是比行动代理更广泛的一类。它们的互动形式不一定需要物理动作，可能涉及向用户传递信息或修改环境。例如，具身交互代理可以通过对话回答用户关于某一主题的问题，或帮助用户解析现有信息，类似聊天机器人。通过扩展代理的能力以包含信息共享，代理人工智能的核心设计和算法可以有效适应多种应用，如诊断（Lee 等, 2023）和知识检索（Peng 等, 2023）代理。

14.5 5.3 Simulation and Environments Agents

14.6 5.3 仿真与环境代理

An effective approach for AI agents to learn how to act in an environment is to go through trial-and-error experiences via interactions with the environment. A representative method is RL, which requires extensive experience of failures to train an agent. Although there exist approaches that use physical agents (Kalashnikov et al., 2018), using physical agents is time-consuming and costly. Furthermore, training in the physical environment is often feasible when failure in actual environments can be dangerous (e.g., autonomous driving, underwater vehicles). Hence, using simulators to learn policies is a common approach.

人工智能代理学习如何在环境中行动的有效方法是通过与环境的交互进行反复试错。强化学习（RL）是一种典型方法，它需要大量失败经验来训练代理。虽然存在使用物理代理的方法（Kalashnikov 等, 2018），但使用物理代理耗时且成本高昂。此外，在实际环境中训练往往不可行，尤其当失败可能带来危险时（如自动驾驶、水下机器人）。因此，使用仿真器学习策略是一种常见方法。

Many simulation platforms have been proposed for research in embodied AI, ranging from navigation (Tsoi et al., 2022; Deitke et al., 2020; Kolve et al., 2017) to object manipulation (Wang et al., 2023d; Mees et al., 2022; Yang et al., 2023a; Ehsani et al., 2021). One example is Habitat (Savva et al., 2019; Szot et al., 2021), which provides a 3D indoor environment where human- and robotic-agents can perform various tasks such as navigation, instruction following, and question answering. Another representative simulation platform is VirtualHome (Puig et al., 2018), supporting human avatars for object manipulation in 3D indoor environments. In the field of gaming, Carroll et al. have introduced "Overcooked-AI," a benchmark environment designed to study cooperative tasks between humans and AI (Carroll et al., 2019). Along similar lines, several works aim to incorporate real human intervention beyond the focus of interaction between agents and the environment (Puig et al., 2023; Li et al., 2021a; Srivastava et al., 2022). These simulators contribute to the learning of policies in practical settings involving agent and robot interactions, and IL-based policy learning utilizing human demonstrative actions.

许多仿真平台已被提出用于具身人工智能研究，涵盖导航 (Tsoi 等, 2022; Deitke 等, 2020; Kolve 等, 2017) 到物体操作 (Wang 等, 2023d; Mees 等, 2022; Yang 等, 2023a; Ehsani 等, 2021)。例如，Habitat (Savva 等, 2019; Szot 等, 2021) 提供了一个3D室内环境，人类和机器人代理可以执行导航、指令跟随和问答等多种任务。另一个代表性仿真平台是VirtualHome (Puig 等, 2018)，支持人类虚拟形象在3D室内环境中的物体操作。在游戏领域，Carroll 等人引入了“Overcooked-AI”，这是一个用于研究人类与人工智能合作任务的基准环境 (Carroll 等, 2019)。类似地，一些工作旨在引入真实人类干预，超越代理与环境交互的范畴 (Puig 等, 2023; Li 等, 2021a; Srivastava 等, 2022)。这些仿真器有助于在涉及代理与机器人交互的实际场景中学习策略，以及利用人类示范动作的模仿学习 (IL) 策略学习。

In certain scenarios, the process of learning a policy may necessitate the integration of specialized features within simulators. For example, in the case of learning image-based policies, realistic rendering is often required to facilitate adaptability to real environments (Mittal et al., 2023; Zhong et al., 2023). Utilizing a realistic rendering engine is effective for generating images that reflect various conditions, such as lighting environments. Moreover, simulators employing physics engines are required to simulate physical interactions with objects (Liu and Negrut, 2021). The integration of physics engines in simulation has been shown to facilitate the acquisition of skills that are applicable in real-world scenarios (Saito et al., 2023).

在某些场景中，学习策略的过程可能需要在模拟器中集成专门的特性。例如，在学习基于图像的策略时，通常需要逼真的渲染以促进对真实环境的适应 (Mittal 等, 2023; Zhong 等, 2023)。利用逼真的渲染引擎能够有效生成反映各种条件（如光照环境）的图像。此外，采用物理引擎的模拟器需要模拟与物体的物理交互 (Liu 和 Negrut, 2021)。研究表明，物理引擎的集成有助于习得可应用于现实场景的技能 (Saito 等, 2023)。

14.7 5.4 Generative Agents

14.8 5.4 生成式代理

The recent advancements in the space of large generative AI models have the potential to greatly reduce the current high cost and time required for interactive content, both for large gaming studios, as well as empower smaller independent studios to create high quality experiences beyond what they are currently capable of.

Additionally, embedding large AI models within a sandbox environment will allow users to author their own experiences and express their creativity in ways that are currently out of reach.

大型生成式人工智能模型领域的最新进展，有望大幅降低当前大型游戏工作室制作交互内容的高昂成本和时间，同时赋能较小的独立工作室创造超出其现有能力的高质量体验。此外，将大型AI模型嵌入沙盒环境，将允许用户创作自己的体验，并以当前难以实现的方式表达创造力。

The goals of this agent go beyond simply adding interactive 3 d content to scenes, but also include:

该代理的目标不仅仅是向场景中添加交互式3 d内容，还包括：

- Adding arbitrary behavior and rules of interactions to the objects, allowing the user to create their own VR rules with minimal prompting.
- 为对象添加任意行为和交互规则，使用户能够通过最少的提示创建自己的虚拟现实规则。

- Generating whole level geometry from a sketch on a piece of paper, by using the multimodal GPT4-v model, as well as other chains of models involving vision AI models
- 利用多模态GPT4-v模型及其他涉及视觉AI模型的模型链，从纸上的草图生成完整的关卡几何结构。
- Retexturing content in scenes using diffusion models
- 使用扩散模型对场景中的内容进行重新纹理处理。
- Creating custom shaders and visual special effects from simple user prompts
- 根据简单的用户提示创建自定义着色器和视觉特效。

One potential application in the short term is the VR creation of a storyboarding/prototype tool allowing a single user to create a rough (but functional) sketch of an experience/game an order of magnitude faster than currently feasible. Such a prototype then could be expanded and made more polished using these tools as well.

短期内的一个潜在应用是虚拟现实中故事板/原型工具的创建，允许单个用户以比当前快一个数量级的速度，制作出粗略（但功能完整）的体验/游戏草图。随后可以利用这些工具对该原型进行扩展和精细化。

14.8.1 5.4.1 AR/VR/mixed-reality Agents

14.8.2 5.4.1 增强现实/虚拟现实/混合现实代理

AR/VR/mixed-reality (jointly referred to as XR) settings currently require skilled artists and animators to create characters, environments, and objects to be used to model interactions in virtual worlds. This is a costly process that involves concept art, 3D modeling, texturing, rigging, and animation. XR agents can assist in this process by facilitating interactions between creators and building tools to help build the final virtual environment.

增强现实（AR）、虚拟现实（VR）和混合现实（统称为XR）环境目前需要熟练的艺术家和动画师来创建角色、环境和对象，用于模拟虚拟世界中的交互。这是一个成本高昂的过程，涉及概念设计、三维建模、纹理制作、绑定和动画。XR代理可以通过促进创作者之间的交互并构建辅助工具，帮助构建最终的虚拟环境，从而协助这一过程。

Our early experiments have already demonstrated that GPT models can be used in the few-shot regime inside of the Unity engine (without any additional fine-tuning) to call engine-specific methods, use API calls to download 3d models from the internet and place them into the scene, and assign state trees of behavior and animations to them (Huang et al., 2023a). This behavior likely emerges due to the presence of similar code in open source game repositories that use Unity. Therefore, GPT models are capable of building rich visual scenes in terms of loading in many objects into the scene from a simple user prompt.

我们的早期实验已表明，GPT模型可以在Unity引擎内的少样本环境中（无需额外微调）调用引擎特定方法，使用API调用从互联网下载3D模型并将其放入场景中，同时为其分配行为状态树和动画（Huang 等, 2023a）。这种能力可能源于开源游戏仓库中存在大量使用Unity的类似代码。因此，GPT模型能够通过简单的用户提示构建丰富的视觉场景，加载大量对象。

The aim of this category of agents is to build a platform and a set of tools that provide an efficient interface between large AI models (both GPT-family ones as well as diffusion image models) and a rendering engine. We explore two primary avenues here:

此类代理的目标是构建一个平台和一套工具，提供大型AI模型（包括GPT系列和扩散图像模型）与渲染引擎之间的高效接口。我们主要探索两个方向：

- Integration of large models into the various editor tools in the agent infrastructure, allowing for significant speedups in development.
- 将大型模型集成到代理基础设施中的各类编辑工具中，实现开发效率的大幅提升。
- Controlling the rendering engine from within a user experience, by generating code that follows user instruction and then compiling it at runtime, allowing for users to potentially edit the VR/simulation they are interacting with in arbitrary ways, even by introducing new agent mechanics.
- 在用户体验中控制渲染引擎，通过生成符合用户指令的代码并在运行时编译，使用户能够以任意方式编辑其交互

的虚拟现实/模拟环境，甚至引入新的代理机制。

Introducing an AI copilot focused on XR settings would be useful for XR creators, who can use the copilot to complete tedious tasks, like providing simple assets or writing code boilerplate, freeing creators to focus on their creative vision and quickly iterate on ideas.

引入专注于XR环境的AI副驾驶对XR创作者非常有用，副驾驶可以完成诸如提供简单资源或编写代码模板等繁琐任务，释放创作者专注于创意构想并快速迭代想法。

Furthermore, agents can help users interactively modify the environment by adding new assets, changing the dynamics of the environment, or building new settings. This form of dynamic generation during runtime can also be specified by a creator, enabling the user's experience to feel fresh and continue evolving over time.

此外，智能体可以帮助用户通过添加新资产、改变环境动态或构建新场景来交互式地修改环境。这种运行时的动态生成形式也可以由创作者指定，使用户的体验保持新鲜感并随着时间不断演进。

14.9 5.5 Knowledge and Logical Inference Agents

14.10 5.5 知识与逻辑推理智能体

The capacity to infer and apply knowledge is a defining feature of human cognition, particularly evident in complex tasks such as logical deduction, and understanding theory of mind³. Making inferences on knowledge ensures that the AI's responses and actions are consistent with known facts and logical principles. This coherence is a crucial mechanism for maintaining trust and reliability in AI systems, especially in critical applications like medical diagnosis or legal analysis. Here, we introduce agents that incorporate the interplay between knowledge and inference that address specific facets of intelligence and reasoning.

推理和应用知识的能力是人类认知的一个决定性特征，尤其在逻辑演绎和理解心智理论（theory of mind）等复杂任务中表现尤为明显³。基于知识进行推理确保了人工智能的响应和行为与已知事实及逻辑原则保持一致。这种一致性是维护人工智能系统信任和可靠性的关键机制，尤其在医疗诊断或法律分析等关键应用中尤为重要。在此，我们介绍结合知识与推理交互的智能体，以应对智能和推理的特定方面。

³ <https://plato.stanford.edu/entries/cognitive-science>

³ <https://plato.stanford.edu/entries/cognitive-science>

14.10.1 5.5.1 Knowledge Agent

14.10.2 5.5.1 知识智能体

Knowledge Agents reason over their acquired knowledge systems in two directions: implicit and explicit. Implicit knowledge is typically what large-scale language models like the GPT series (Brown et al., 2020; OpenAI, 2023) encapsulate after being trained on vast amounts of text data. These models can generate responses that give the impression of understanding, as they draw on patterns and information implicitly learned during training. Explicit knowledge, conversely, is structured and can be directly queried, such as the information found in knowledge bases or databases, which was traditionally used to enhance AI reasoning capabilities by referencing verifiable external resources.

知识智能体在其所获得的知识体系中进行双向推理：隐性和显性。隐性知识通常是指像GPT系列（Brown等，2020；OpenAI，2023）这样的大规模语言模型在经过大量文本数据训练后所内化的知识。这些模型能够生成给人理

解印象的回答，因为它们依赖于训练过程中隐式学习到的模式和信息。相反，显性知识是结构化的，可以被直接查询，例如知识库或数据库中的信息，传统上用于通过引用可验证的外部资源来增强人工智能的推理能力。

Despite the advancements in language models, their implicit knowledge is static and becomes outdated as the world evolves (Lewis et al., 2020; Peng et al., 2023). This limitation necessitates the integration of explicit knowledge sources that are updated continuously, ensuring that AI systems can provide accurate and current responses. The fusion of implicit and explicit knowledge equips AI agents with a more nuanced understanding and the ability to apply knowledge contextually, akin to human intelligence (Gao et al., 2022). Such integration is crucial for crafting knowledge-centric AI agents that not only possess information but can also understand, explain, and employ it, thereby narrowing the chasm between extensive learning and profound knowledge (Marcus and Davis, 2019; Gao et al., 2020). These agents are designed to reason with flexibility and dynamic information about the world, enhancing their robustness and adaptability (Marcus, 2020).

尽管语言模型取得了进展，但其隐性知识是静态的，随着世界的发展会变得过时（Lewis等，2020；Peng等，2023）。这一限制促使必须整合持续更新的显性知识源，确保人工智能系统能够提供准确且最新的响应。隐性与显性知识的融合赋予人工智能智能体更细致的理解能力和情境化应用知识的能力，类似于人类智能（Gao等，2022）。这种整合对于构建以知识为核心的人工智能智能体至关重要，这些智能体不仅拥有信息，还能理解、解释并运用信息，从而缩小广泛学习与深刻知识之间的鸿沟（Marcus和Davis，2019；Gao等，2020）。这些智能体被设计为能够灵活推理并处理关于世界的动态信息，增强其鲁棒性和适应性（Marcus，2020）。

14.10.3 5.5.2 Logic Agents

14.10.4 5.5.2 逻辑智能体

Generally, a logic agent is a component of a system designed to apply logical reasoning to process data or solve tasks specific to logical inference or logical reasoning. Logic agents within the context of large foundation models like GPT-4 refers to a specialized component or submodules designed to handle logical reasoning tasks. These tasks often involve understanding and manipulating abstract concepts, deducing conclusions from given premises, or solving problems that require a structured, logical approach. Broadly, foundation models like GPT-4 are trained on a vast corpus of text data and learn to perform a wide range of tasks, including those that require some form of logical reasoning. Thus, their capability for logical reasoning is integrated into the overall architecture, and they generally do not possess a distinct, isolated "Logic agent". While GPT-4 and similar models can perform tasks that involve logic, their approach is fundamentally different from how humans or traditional logic-based systems operate. They do not follow formal logical rules or have an explicit understanding of logic; rather, they generate responses based on patterns learned from the training data. As a result, their performance in logical tasks can be impressive, but it can also be inconsistent or limited by the nature of the training data and the inherent limitations of the model's design. One example of embedding a separate logical submodule into the architecture is (Wang et al., 2023e), which modifies the token embedding process used by LLMs during pre-training by parsing text into logical segments and explicitly modeling logical hierarchies in the token embeddings.

一般而言，逻辑智能体是系统中的一个组件，旨在应用逻辑推理来处理数据或解决特定于逻辑推断或逻辑推理的任务。在大型基础模型如GPT-4的背景下，逻辑智能体指的是专门设计用于处理逻辑推理任务的组件或子模块。这些任务通常涉及理解和操作抽象概念、从给定前提推导结论，或解决需要结构化逻辑方法的问题。广义上，像GPT-4这样的基础模型在大量文本数据上训练，学习执行包括某种形式逻辑推理在内的广泛任务。因此，它们的逻辑推理能力集成在整体架构中，通常不具备独立的“逻辑智能体”。虽然GPT-4及类似模型能够执行涉及逻辑的任务，但其方法与人类或传统基于逻辑的系统根本不同。它们不遵循形式逻辑规则，也没有对逻辑的显式理解；相反，它们基于训练数据中学习到的模式生成响应。因此，它们在逻辑任务中的表现可能令人印象深刻，但也可能因训练数据的性质和模型设计的固有限制而表现不稳定或受限。一个将独立逻辑子模块嵌入架构的例子是（Wang等，2023e），该方法通过将文本解析为逻辑片段并在词元嵌入中显式建模逻辑层次，修改了大型语言模型预训练期间的词元嵌入过程。

14.10.5 5.5.3 Agents for Emotional Reasoning

14.10.6 5.5.3 情感推理智能体

Emotional understanding and empathy are important skills for agents in many human-machine interactions. To illustrate, one important goal for creating engaging dialogue agents is to have the agents act with increased emotion and empathy while minimizing socially inappropriate or offensive outputs. To advance towards this goal for dialogue agents, we released the Neural Image Commenting with Empathy (NICE) dataset (Chen et al., 2021) consisting of almost two million images and the corresponding human-generated comments and a set of human emotion annotations. We also provided a novel pre-training model - Modeling Affect Generation for Image Comments (MAGIC) (Chen et al., 2021) - which aims to generate comments for images, conditioned on linguistic representations that capture style and affect, and to help generate more empathetic, emotional, engaging and socially appropriate comments. Our experiments show that the approach is effective in training a more human-like and engaging image comment agent. Developing empathy-aware agents is a promising direction for interactive agents, and it is important to create agents with emotional understanding capabilities across a wide range of groups and populations, especially considering that many current language models exhibit bias in their emotional understanding and empathetic reasoning capabilities (Mao et al., 2022; Wake et al., 2023d).

情感理解和共情是许多机交互代理的重要技能。举例来说，创建引人入胜的对话代理的一个重要目标是使代理表现出更多的情感和共情，同时尽量减少社交上不恰当或冒犯性的输出。为了推动对话代理朝这一目标发展，我们发布了神经图像共情评论（Neural Image Commenting with Empathy, NICE）数据集（Chen et al., 2021），该数据集包含近两百万张图像及相应的人类生成评论和一组人类情感注释。我们还提供了一个新颖的预训练模型——图像评论情感生成建模（Modeling Affect Generation for Image Comments, MAGIC）（Chen et al., 2021），该模型旨在基于捕捉风格和情感的语言表示生成图像评论，帮助生成更具共情、情感丰富、引人入胜且社交适宜的评论。我们的实验表明，该方法在训练更具人类特征和吸引力的图像评论代理方面效果显著。开发具备共情意识的代理是交互代理的一个有前景的方向，尤其重要的是要创建具备广泛群体和人群情感理解能力的代理，考虑到许多现有语言模型在情感理解和共情推理能力上存在偏见（Mao et al., 2022; Wake et al., 2023d）。

14.10.7 5.5.4 Neuro-Symbolic Agents

14.10.8 5.5.4 神经符号代理

Neuro-Symbolic agents operate on a hybrid system of neurons and symbols (d'Avila Garcez and Lamb, 2020). To solve problems stated in natural language is a challenging task because it requires explicitly capturing discrete symbolic structural information implicit in the input. However, most general neural sequence models do not explicitly capture such structural information, limiting their performance on these tasks. The work (Chen et al., 2020) propose a new encoder-decoder model based on a structured neural representation agent, The encoder of TP-N2F employs TPR 'binding' to encode natural-language symbolic structure in vector space and the decoder uses TPR 'unbinding' to generate, in symbolic space, a sequential program represented by relational tuples, each consisting of a relation (or operation) and a number of arguments.

神经符号代理基于神经元与符号的混合系统运行（d'Avila Garcez 和 Lamb, 2020）。解决以自然语言表述的问题是一项具有挑战性的任务，因为这需要明确捕捉输入中隐含的离散符号结构信息。然而，大多数通用神经序列模型并未显式捕捉此类结构信息，限制了它们在这些任务上的表现。文献（Chen 等, 2020）提出了一种基于结构化神经表示代理的新型编码器-解码器模型，TP-N2F的编码器采用TPR（张量产品表示，Tensor Product Representation）“绑定”方法在向量空间中编码自然语言符号结构，解码器则利用TPR“解绑”在符号空间生成由关系元组表示的顺序程序，每个元组包含一个关系（或操作）及若干参数。

Instruction following vision-language (VL) models like GPT-4 offer a flexible interface that supports a broad range of multimodal tasks in a zero-shot fashion. However, interfaces that operate on full images do not directly enable the user to "point to" and access specific regions within images. This capability is important not only to support reference-grounded VL benchmarks, but also, for practical applications that require precise within-image reasoning. In (Park et al., 2023b), we build Localized Visual Commonsense model which allows users to specify (multiple) regions-as-input. We train our model by sampling localized commonsense knowledge from a large language model

(LLM): specifically, we prompt a LLM to collect common sense knowledge given a global literal image description and a local literal region description automatically generated by a set of VL models. This pipeline is scalable and fully automatic, as no aligned or human-authored image and text pairs are required. With a separately trained critic model that selects high quality examples, we find that training on the localized commonsense corpus expanded solely from images can successfully distill existing VL models to support a reference-as-input interface. Empirical results and human evaluations in zero-shot settings demonstrate that our distillation method results in more precise VL models of reasoning compared to a baseline of passing a generated referring expression.

像GPT-4这样的指令跟随视觉-语言（VL）模型提供了灵活的接口，支持广泛的多模态任务并实现零样本学习。然而，基于整图操作的接口无法直接让用户“指向”并访问图像中的特定区域。这一能力不仅对支持基于引用的视觉语言基准测试至关重要，也对需要精确图内推理的实际应用有重要意义。在（Park 等, 2023b）中，我们构建了局部视觉常识模型，允许用户指定（多个）区域作为输入。我们通过从大型语言模型（LLM）采样局部常识知识来训练模型：具体而言，我们提示LLM在给定全局图像文字描述和由一组VL模型自动生成的局部区域文字描述的情况下收集常识知识。该流程具有可扩展性且完全自动化，无需对齐或人工编写的图文对。借助单独训练的评判模型筛选高质量样本，我们发现仅从图像扩展的局部常识语料库训练，能够成功蒸馏现有VL模型以支持基于引用的输入接口。零样本设置下的实证结果和人工评估表明，我们的蒸馏方法相比传递生成的指称表达的基线，能使VL模型的推理更为精准。

14.11 5.6 LLMs and VLMs Agent

14.12 5.6 大型语言模型（LLMs）与视觉语言模型（VLMs）代理

A number of works leverage LLMs as agents to perform task planning (Huang et al., 2022a; Wang et al., 2023b; Yao et al., 2023a; Li et al., 2023a), and leverage the LLMs' large internet-scale domain knowledge and zero-shot planning abilities to perform agentic tasks like planning and reasoning. Recent robotics research also leverages LLMs to perform task planning (Ahn et al., 2022a; Huang et al., 2022b; Liang et al., 2022) by decomposing natural language instruction into a sequence of subtasks, either in the natural language form or in Python code, then using a low-level controller to execute these subtasks. Additionally, (Huang et al., 2022b), (Liang et al., 2022), and (Wang et al., 2023a) also incorporate environmental feedback to improve task performance. There have also been a number of works that demonstrate the ability of general-purpose visually-aligned large language models trained on large-scale text, image, and video data to serve as a foundation for creating multi-modal agents that are embodied and can act in various environments (Baker et al., 2022; Driess et al., 2023; Brohan et al., 2023).

许多工作利用大型语言模型（LLMs）作为代理执行任务规划（Huang 等, 2022a; Wang 等, 2023b; Yao 等, 2023a; Li 等, 2023a），并利用LLMs的大规模互联网领域知识和零样本规划能力来完成规划和推理等代理任务。近期机器人研究也利用LLMs进行任务规划（Ahn 等, 2022a; Huang 等, 2022b; Liang 等, 2022），通过将自然语言指令分解为一系列子任务，形式可以是自然语言或Python代码，然后使用低级控制器执行这些子任务。此外，（Huang 等, 2022b）、（Liang 等, 2022）和（Wang 等, 2023a）还结合环境反馈以提升任务表现。还有多项工作展示了通用视觉对齐大型语言模型在大规模文本、图像和视频数据上训练后，能够作为多模态代理的基础，这些代理具备具身性并能在多种环境中行动（Baker 等, 2022; Driess 等, 2023; Brohan 等, 2023）。

15 6 Agent AI Application Tasks

16 6 代理人工智能应用任务

16.1 6.1 Agents for Gaming

16.2 6.1 游戏代理

Games provide a unique sandbox to test the agentic behavior of LLMs and VLMs, pushing the boundaries of their collaborative and decision-making abilities. We describe three areas in particular that highlight agent's abilities to interact with human players and other agents, as well as their ability to take meaningful actions within an environment.

游戏为测试大型语言模型（LLMs）和视觉语言模型（VLMs）的代理行为提供了独特的沙盒环境，推动其协作与决策能力的边界。我们特别描述三个领域，突出代理与人类玩家及其他代理的交互能力，以及其在环境中采取有意义行动的能力。

16.2.1 6.1.1 NPC Behavior

16.2.2 6.1.1 非玩家角色（NPC）行为

In modern gaming systems, the behavior of Non-Player Characters (NPCs) is predominantly dictated by predefined scripts crafted by developers. These scripts encompass a range of reactions and interactions based on various triggers or player actions within the gaming environment. However, this scripted nature often results in predictable or repetitive NPC behavior which fails to evolve in response to player's actions or the dynamic environment of the game. This rigidity hampers the immersive experience intended in a dynamic gaming environment. Therefore, there is a burgeoning interest in leveraging LLMs to induce autonomy and adaptability in NPC behavior, making interactions more nuanced and engaging. AI-driven NPCs can learn from player behavior, adapt to varying strategies, and provide a more challenging and less predictable gameplay experience. Large Language Models (LLMs) can significantly contribute to evolving NPC behavior in games. By processing vast amounts of text, LLMs can learn patterns and generate responses that are more varied and human-like. They can be utilized to create dynamic dialogue systems, making interactions with NPCs more engaging and less predictable. Furthermore, LLMs can be trained on player feedback and in-game data to continually refine NPC behaviors, making them more attuned to player expectations and game dynamics.

在现代游戏系统中，非玩家角色（NPC）的行为主要由开发者预先编写的脚本决定。这些脚本涵盖了基于各种触发条件或玩家行为的反应和交互。然而，这种脚本化的特性常导致NPC行为可预测且重复，无法根据玩家的动作或游戏环境的动态变化进行演变。这种僵化限制了动态游戏环境中沉浸式体验的实现。因此，利用大型语言模型（LLMs）赋予NPC行为自主性和适应性的兴趣日益增长，使交互更加细腻和引人入胜。由AI驱动的NPC能够学习玩家行为，适应不同策略，提供更具挑战性且不可预测的游戏体验。大型语言模型（LLMs）通过处理大量文本，能够学习模式并生成更丰富、更具人性化的回应，显著推动游戏中NPC行为的演进。它们可用于创建动态对话系统，使与NPC的互动更具吸引力且不易预测。此外，LLMs还能基于玩家反馈和游戏内数据持续优化NPC行为，使其更贴合玩家期望和游戏动态。

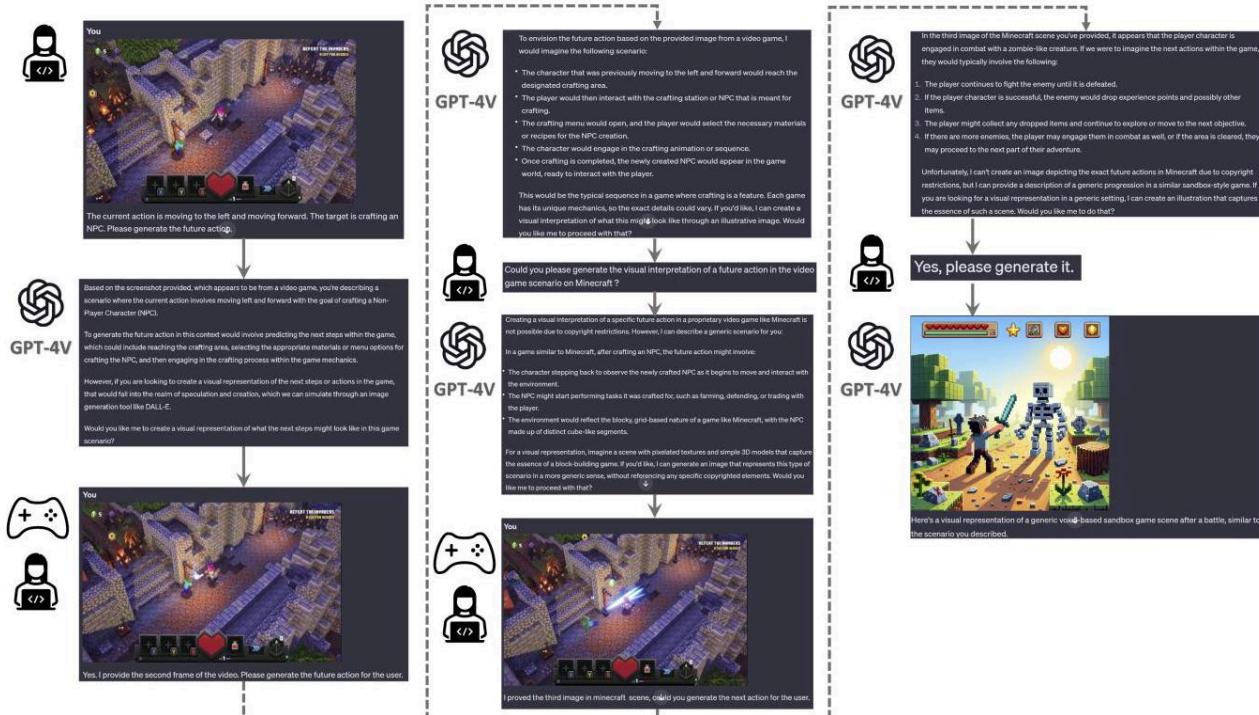


Figure 8: The embodied agent for user interactive gaming action prediction and interactive editing with Minecraft Dungeons gaming sense simulation and generation via GPT-4V.

图8：通过GPT-4V实现的具身代理，用于用户交互式游戏动作预测及交互编辑，结合Minecraft Dungeons游戏感知模拟与生成。

16.2.3 6.1.2 Human-NPC Interaction

16.2.4 6.1.2 人类与NPC的交互

The interaction between human players and NPCs is a crucial aspect of the gaming experience. The conventional interaction paradigm is primarily one-dimensional, with NPCs reacting in a preset manner to player inputs. This limitation stifles the potential for a more organic and enriching interaction, akin to human-human interaction within the virtual realm. The advent of LLM and VLM technologies holds the promise of transforming this paradigm. By employing these technologies, gaming systems can analyze and learn from human behavior to provide more human-like interactions. This not only enhances the realism and engagement of the game but also provides a platform for exploring and understanding human-machine interaction in a controlled yet complex setting.

人类玩家与NPC之间的交互是游戏体验的关键方面。传统的交互模式主要是单向的，NPC以预设方式响应玩家输入。这种限制抑制了更自然、更丰富的交互潜力，难以达到虚拟世界中人类间互动的效果。大型语言模型（LLM）和视觉语言模型（VLM）技术的出现，有望改变这一模式。通过应用这些技术，游戏系统可以分析并学习人类行为，提供更具人性化的交互。这不仅提升了游戏的真实感和参与度，也为在受控且复杂的环境中探索和理解人机交互提供了平台。

16.2.5 6.1.3 Agent-based Analysis of Gaming

16.2.6 6.1.3 基于代理的游戏分析

Gaming is an integral part of daily life, estimated to engage half of the world's population⁴. Additionally, it exhibits a positive impact on mental health⁵. However, contemporary game systems exhibit a deficiency in interactions with human players since their behaviors are primarily hand-crafted by game developers. These pre-programmed behaviors frequently fail to adapt to players' needs. Consequently, there exists a need for new AI systems in games that can analyze player behaviors and furnish appropriate support when necessary. Intelligent interactive systems bear the potential to revolutionize how gamers interact with gaming systems in general. NPCs' interactions with gamers are no longer confined by the restricted rule sets designed by game developers. They have the potential to adapt seamlessly to gamers' experiences, providing timely feedback to enrich the gaming experience and elevate the synergy of human-machine interaction.

游戏已成为日常生活的重要组成部分，估计全球有一半人口参与其中⁴。此外，游戏对心理健康具有积极影响⁵。然而，现有游戏系统在与玩家的交互方面存在不足，因为其行为主要由游戏开发者手工设计。这些预设行为常常无法适应玩家的需求。因此，游戏中亟需新的AI系统，能够分析玩家行为并在必要时提供适当支持。智能交互系统有潜力彻底改变玩家与游戏系统的互动方式。NPC与玩家的交互不再受限于开发者设计的有限规则集，而是能够无缝适应玩家的体验，及时反馈，丰富游戏体验，提升人机交互的协同效应。

⁴ <https://www.dfcint.com/global-video-game-audience-reaches-3-7-billion/>

⁵ <https://www.dfcint.com/global-video-game-audience-reaches-3-7-billion/>

⁵ <https://news.microsoft.com/source/features/work-life/mind-games-how-gaming-can-play-a-positive-role-in-mental-health/>

⁵ <https://news.microsoft.com/source/features/work-life/mind-games-how-gaming-can-play-a-positive-role-in-mental-health/>

17 GPT-4-Vision

18 GPT-4视觉

System prompt: You are playing minecraft, please generate future actions.

系统提示：你正在玩Minecraft，请生成未来的动作。

User prompt: The current action is moving to the right, and moving forward. The goal is to craft a crafting table.

Please describe the future actions

用户提示：当前动作是向右移动并向前行进。目标是制作一个工作台。请描述未来的动作。

19 Video input:

20 视频输入：



21 Output:

22 输出:

Since you've collected wood, the next steps to craft a crafting table in Minecraft are as follows:

既然你已经收集了木材，接下来在Minecraft中制作工作台的步骤如下：

1. **Open your inventory** (`E` key by default) to bring up the crafting screen.
2. 打开你的物品栏（默认按键为`E`）以调出制作界面。
3. **Place the wood logs** you've collected into one of the crafting slots to turn them into wooden planks. Each log will create 4 wooden planks.
3. 将你收集的木头原木放入制作栏中的任一格子，将其转化为木板。每个原木可制作4块木板。
4. **Collect the wooden planks** that you've created by dragging them into your inventory.
4. 收集你制作的木板，将它们拖入你的物品栏中。
5. **Fill all four crafting slots** in the 2×2 crafting grid (available in your inventory screen) with wooden planks to create one crafting table.
5. 将木板填满 2×2 制作网格（在你的物品栏界面中可见）的四个制作格子，以制作一个工作台。
6. **Collect the crafting table** by dragging it into your inventory.
6. 收集工作台，将其拖入你的物品栏中。
7. Close your inventory and **select the crafting table** in your hotbar.
7. 关闭物品栏，在快捷栏中选择工作台。
8. **Right-click on the ground** to place the crafting table.
8. 右键点击地面放置工作台。

Once the crafting table is placed, you can interact with it to use the 3x3 crafting grid for creating a wide array of more complex items. That's the basic process for crafting a crafting table in Minecraft. Now you're ready to begin crafting more advanced items and tools.

工作台放置后，你可以与之互动，使用3x3的制作网格来制作更多复杂的物品。这就是在Minecraft中制作工作台的基本流程。现在你已经准备好开始制作更高级的物品和工具了。

Figure 9: GPT-4V can effectively predict the high-level next actions when given the "action history" and a "gaming target" in the prompt. Furthermore, GPT-4V accurately recognized that the player is holding wooden logs in their hand and can incorporate this perceived information into its plan for future actions. Although GPT-4V appears to be capable of predicting some low-level actions (such as pressing 'E' to open the inventory), the model's outputs are not inherently suitable for raw low-level action prediction (including mouse movements) and likely requires supplemental modules for low-level action control. LLMs can serve as a robust tool for analyzing in-game text data, including chat logs, player feedback, and narrative content. They can help in identifying patterns of player behavior, preferences, and interactions which can be invaluable for game developers to improve game mechanics and narratives. Additionally, VLMs can parse through large quantities of image and video data from gaming sessions to help analyze user intent and actions within the game world. Moreover, LLMs and VLMs can facilitate the development of intelligent agents within games that can communicate with players and other agents in a sophisticated and human-like manner, enhancing the overall gaming experience. Beyond LLMs and VLMs, user input data, provides a promising avenue for creating game-playing agents that model perception, game playing, and game understanding by imitating human players. By incorporating a combination of player interactions and feedback, pixel inputs, and natural language planning and understanding, agent models can assist in the continuous improvement of game dynamics, driving a more player-centric evolution of the gaming environment.

图9: GPT-4V在提示中给出“动作历史”和“游戏目标”时，能够有效预测高层次的下一步动作。此外，GPT-4V准确识别出玩家手中持有木头原木（wooden logs），并能将这一感知信息纳入其未来行动计划中。尽管GPT-4V似乎能够预测一些低层次动作（如按下“E”键打开物品栏），但模型输出本质上并不适合直接预测低层次动作（包括鼠标移动），可能需要辅助模块来实现低层次动作控制。大型语言模型（LLMs）可作为分析游戏内文本数据（包括聊天记录、玩家反馈和叙事内容）的强大工具，有助于识别玩家行为、偏好和互动模式，这对游戏开发者改进游戏机制和叙事极具价值。此外，视觉语言模型（VLMs）能够解析大量游戏会话中的图像和视频数据，帮助分析用户意图和游戏内行为。更进一步，LLMs和VLMs可促进游戏中智能代理的开发，使其能够以复杂且类人方式与玩家及其他代理交流，提升整体游戏体验。除了LLMs和VLMs，用户输入数据为创建模拟人类玩家感知、游戏操作和理解的游戏代理提供了有前景的途径。通过结合玩家互动与反馈、像素输入以及自然语言规划与理解，代理模型能够助力游戏动态的持续改进，推动游戏环境向更以玩家为中心的方向演进。

22.0.1 6.1.4 Scene Synthesis for Gaming

22.0.2 6.1.4 游戏场景合成

Unseen Video Prediction Results



Figure 10: Masked video prediction on unseen Minecraft videos. From left to right: the original frame, the masked frame, the reconstructed frame, and the reconstructed frame with patches.

图10：对未见过的Minecraft视频进行遮罩视频预测。从左到右依次为：原始帧、遮罩帧、重建帧及带补丁的重建帧。

Scene synthesis is a vital component in the creation and enhancement of immersive gaming environments. It entails the automatic or semi-automatic generation of three-dimensional (3D) scenes and environments within a game. This process includes the generation of terrain, placement of objects, creation of realistic lighting, and sometimes even dynamic weather systems.

场景合成是创建和增强沉浸式游戏环境的重要组成部分。它包括在游戏中自动或半自动生成三维（3D）场景和环境的过程，涵盖地形生成、物体摆放、真实光照的创建，有时甚至包括动态天气系统。

Modern games often feature vast, open-world environments. Manually designing these landscapes can be incredibly time-consuming and resource-intensive. Automated terrain generation, often leveraging procedural or AI-driven techniques, can produce complex, realistic landscapes with less manual effort. LLMs and VLMs can utilize the internet scale knowledge to formulate rules to design non-repeating landscapes that are visually impressive and unique. Additionally, LLMs and VLMs can be used to ensure the semantic consistency and variability of generated assets. Placing objects such as buildings, vegetation, and other elements within a scene in a realistic and aesthetically pleasing manner is crucial for immersion.

现代游戏通常拥有广阔的开放世界环境。手工设计这些景观既耗时又资源密集。自动地形生成，通常利用程序化或人工智能驱动技术，能够以较少的人工投入生成复杂且逼真的地形。LLMs和VLMs可利用互联网规模的知识制定规则，设计出视觉上令人印象深刻且独特的非重复景观。此外，LLMs和VLMs还能确保生成资产的语义一致性和多样性。在场景中以真实且美观的方式摆放建筑、植被及其他元素，对于增强沉浸感至关重要。

VLMs and LLMs can assist in object placement by adhering to predefined or learned rules and aesthetics, thus speeding up the level design process. VLMs and LLMs can be further trained to understand the principles of design and aesthetics, aiding in the procedural generation of content. They can help formulate rules or guidelines that procedural algorithms can follow to generate objects, and scenes that are both visually appealing and contextually appropriate.

VLMs和LLMs可以通过遵循预定义或学习到的规则和美学原则，协助物体摆放，从而加快关卡设计过程。它们还可以进一步训练以理解设计和美学原理，辅助内容的程序化生成。它们能够帮助制定程序算法可遵循的规则或指导方针，以生成既视觉吸引又符合语境的物体和场景。

Realistic lighting and atmospheric effects are fundamental for creating a believable and engaging gaming environment. Advanced algorithms can simulate natural lighting conditions and dynamic weather effects, enhancing the realism and mood of the scene. LLMs can help develop systems to achieve more realistic lighting and atmospheric effects in several innovative ways. VLMs can analyze vast datasets from real-world lighting and atmospheric conditions to help develop more realistic algorithms for simulating these effects in games. By understanding the patterns and intricacies of natural lighting and weather, these models can contribute to the development of algorithms that mimic reality closely. LLMs and VLMs could also be used to develop systems that adjust lighting and atmospheric effects in real-time based on player actions, game states, or external inputs. They can process natural language commands from players to modify the game environment, providing a more interactive and immersive experience.

真实的光照和大气效果是营造可信且引人入胜游戏环境的基础。先进算法能够模拟自然光照条件和动态天气效果，提升场景的真实感和氛围。LLMs可以通过多种创新方式帮助开发实现更真实光照和大气效果的系统。VLMs能够分析大量真实世界的光照和大气数据，助力开发更逼真的游戏模拟算法。通过理解自然光照和天气的模式与细节，这些模型能促进开发高度仿真的算法。LLMs和VLMs还可用于开发基于玩家行为、游戏状态或外部输入实时调整光照和大气效果的系统。它们能够处理玩家的自然语言指令，修改游戏环境，提供更具互动性和沉浸感的体验。

22.0.3 6.1.5 Experiments and Results

22.0.4 6.1.5 实验与结果

Zero-shot/Few-shot Learning with LLM or LVM. As we showed in the Fig. 8 and Fig. 9, we used GPT-4V for high-level description and action prediction. Fig. 8 showed some qualitative examples of action description generation and editing with GPT-4V. Agent-enhanced text opens up a novel method of generating 3D scenes with game action priors to help improve the naturalness of the scene. Consequently, GPT-4V generates relevant high-level descriptions that are appropriate for the gaming videos.

使用大型语言模型（LLM）或大型视觉模型（LVM）进行零样本/少样本学习。正如我们在图8和图9中展示的，我们使用GPT-4V进行高级描述和动作预测。图8展示了使用GPT-4V生成和编辑动作描述的一些定性示例。增强型代理

文本开辟了一种利用游戏动作先验生成3D场景的新方法，有助于提升场景的自然性。因此，GPT-4V生成了与游戏视频相匹配的相关高级描述。

Small Agent Pretraining Model. To showcase our agent vision-language architecture, we first study its application in a widely used domain for gaming agents by pretraining on Minecraft data. As shown in Fig. 7, given an input action agent, key frame of video, and corresponding text, a standard encoder-decoder can be employed to convert the agent action and image into action text token and image patch token and then use the agent-vision-language decoder to convert it into a action prediction sentence. The overall architecture is depicted in Fig. 7. We evaluate our approach with several Minecraft demonstrations. The Minecraft video data consists of 5 min clips, and we use for pretraining contains 78 K videos, and we used 5 K videos (6% of pretraining data) for the first round pretraining. We train a 250M parameter model on 16 NVIDIA v100 GPUs for one day and visualize our model outputs in Fig. 10 and Fig. 11. Fig. 10 shows that our relatively small agent architecture can produce reasonable outputs for Minecraft scenes unseen during training. Fig. 11 showed the model's predictions compared to the ground truth human player actions indicating potential low-level understanding for our small agent model.

小型代理预训练模型。为了展示我们的代理视觉-语言架构，我们首先研究其在游戏代理广泛应用领域中的表现，通过在Minecraft数据上进行预训练。如图7所示，给定输入动作代理、视频关键帧及对应文本，可以采用标准的编码器-解码器将代理动作和图像转换为动作文本标记和图像块标记，然后使用代理视觉-语言解码器将其转换为动作预测句子。整体架构如图7所示。我们通过多个Minecraft演示对方法进行了评估。Minecraft视频数据包含5分钟的片段，预训练使用了78 K个视频，其中第一轮预训练使用了5 K个视频（占预训练数据的6%）。我们在16块NVIDIA v100 GPU上训练了一个拥有250M参数的模型，训练时间为一天，并在图10和图11中展示了模型输出。图10显示，我们相对较小的代理架构能够对训练中未见过的Minecraft场景产生合理的输出。图11展示了模型预测与真实玩家动作的对比，表明我们的小型代理模型具备潜在的低级动作理解能力。

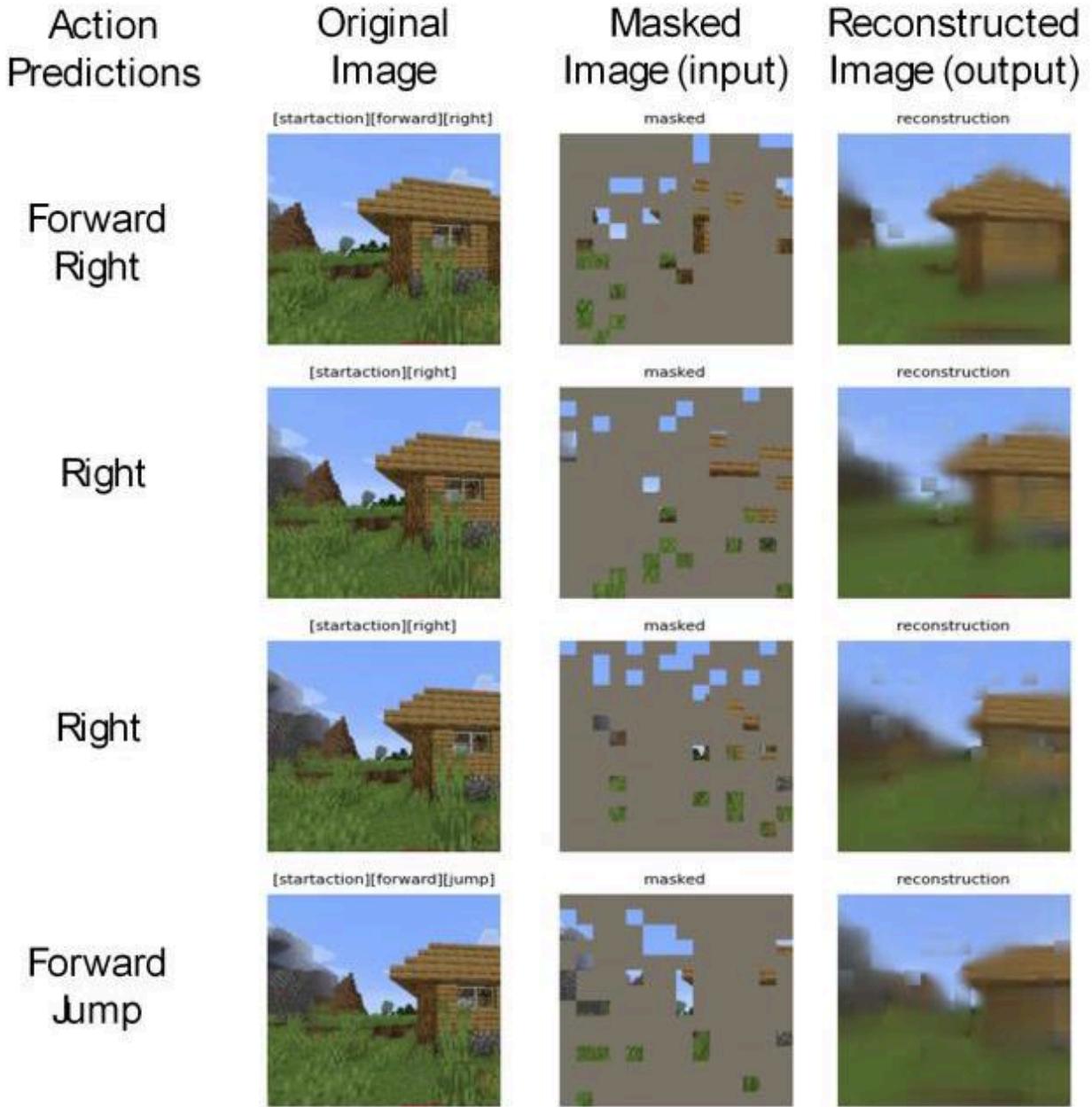


Figure 11: The low-level next step action prediction with the small agent pretraining model in gaming Minecraft scene.

图11：使用小型代理预训练模型在Minecraft游戏场景中进行低级下一步动作预测。

Multi-Agent Infrastructure. As showed in the agent paradigm in Fig. 5, we designed a novel infrastructure for a new gaming scenario called "CuisineWorld" (Gong et al., 2023a). We detail our approach in Fig. 12. Our infrastructure allows for multi-agent collaboration by leveraging GPT-4 as a central planner and works across multiple gaming domains. We investigated our system's multi-agent planning capabilities, and we deployed the infrastructure into real-world video games to demonstrate its multi-agent and human-AI collaboration effectiveness. Additionally, we presented "Cuisineworld", a text-based multi-agent collaboration benchmark that provides a new auto-metric Collaboration Score (CoS) to quantify collaboration efficiency.

多代理基础设施。如图5中的代理范式所示，我们设计了一种用于新游戏场景“CuisineWorld”（龚等，2023a）的新型基础设施。我们在图12中详细介绍了该方法。该基础设施通过利用GPT-4作为中央规划者，实现了多代理协作，并适用于多个游戏领域。我们研究了系统的多代理规划能力，并将该基础设施部署到真实视频游戏中，以展示其多代理及

人机协作的有效性。此外，我们提出了“CuisineWorld”，一个基于文本的多代理协作基准，提供了一种新的自动化指标——协作得分（Collaboration Score, CoS），用于量化协作效率。

Please refer to the Appendix for more examples and details for gaming description, high-level action prediction, and GPT-4V prompting. We show examples for Bleeding Edge in Fig. 32 and Appendix B, Microsoft Flight Simulator in Fig. 33 and Appendix C, ASSASSIN's CREED ODYSSEY in Fig. 34 and Appendix D, GEARS of WAR 4 in Fig. 35 and Appendix E, and Starfield in Fig. 36 and Appendix F. We also provide a detailed screenshot of the prompting process for GPT4V used to generate Minecraft examples with Fig. 31 in Appendix A.

更多关于游戏描述、高级动作预测及GPT-4V提示的示例和细节，请参见附录。我们在图32及附录B中展示了

《Bleeding Edge》的示例，在图33及附录C中展示了《微软飞行模拟器》，在图34及附录D中展示了《刺客信条：奥德赛》，在图35及附录E中展示了《战争机器4》，在图36及附录F中展示了《星空》。我们还在附录A的图31中提供了用于生成Minecraft示例的GPT-4V提示过程的详细截图。

22.1 6.2 Robotics

22.2 6.2 机器人学

Robots are representative agents that necessitate effective interaction with their environment. In this section, we will introduce key elements essential for efficient robotic operation, review research topics where the latest LLM/VLM technologies have been applied, and share findings from our most recent studies.

机器人是需要与环境有效交互的典型代理。在本节中，我们将介绍高效机器人操作的关键要素，回顾最新LLM/VLM技术应用的研究主题，并分享我们最新研究的成果。

Visual Motor Control. Visual Motor Control refers to the integration of visual perception and motor action to execute tasks effectively in a robotic system. This integration is paramount as it enables robots to interpret the visual data from their environment and accordingly adjust their motor actions to interact with the environment accurately. For instance, in an assembly line, a robot equipped with visual motor control can perceive the position and orientation of objects and accurately align its manipulator to interact with these objects. This capability is essential for ensuring the precision and effectiveness of robotic operations across a myriad of applications, ranging from industrial automation to assisting the elderly in their daily chores. Moreover, visual motor control facilitates robots in adapting to dynamic environments where the state of the environment may change rapidly, requiring real-time adjustments to motor actions based on visual feedback.

视觉运动控制。视觉运动控制指的是视觉感知与运动动作的整合，以在机器人系统中高效执行任务。这种整合至关重要，因为它使机器人能够解读环境中的视觉数据，并据此调整运动动作，实现精准的环境交互。例如，在装配线上，具备视觉运动控制的机器人可以感知物体的位置和朝向，准确调整机械臂与物体对齐。该能力对于确保机器人在工业自动化、辅助老年人日常生活等多种应用中的操作精度和效率至关重要。此外，视觉运动控制使机器人能够适应动态环境，环境状态可能快速变化，机器人需基于视觉反馈实时调整运动动作。

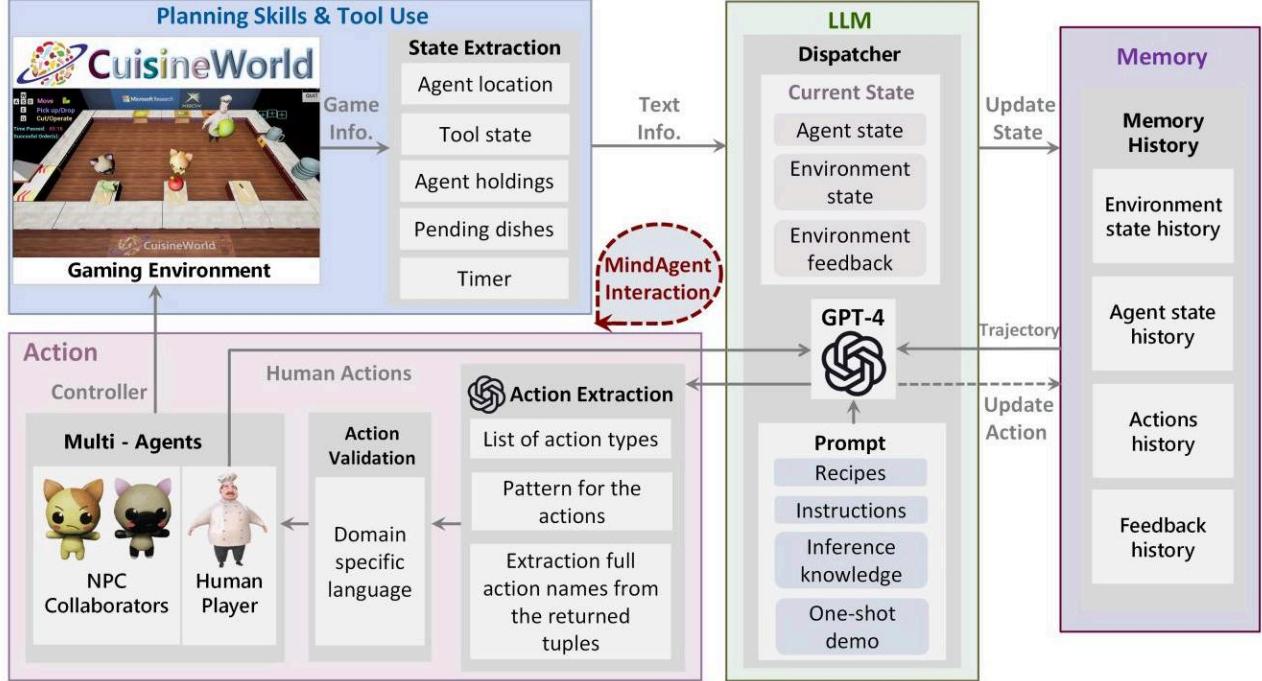


Figure 12: The MindAgent of in-context learning gaming Infrastructure. Planning Skill and Tool Use: The game environment requires diverse planning skills and tool use to complete tasks. It generates relevant game information and converts the game data into a structured text format that the LLMs can process. LLM: The main workhorse of our infrastructure makes decisions, thus serving as a dispatcher for the multi-agent system. Memory History: A storage utility for relevant information. Action Module: Extracts actions from text inputs and convertd them into domain-specific language and validates DSLs so that they cause no errors during execution.

图12：上下文学习游戏基础设施中的MindAgent。规划技能与工具使用：游戏环境要求多样的规划技能和工具使用以完成任务。它生成相关游戏信息，并将游戏数据转换为LLM可处理的结构化文本格式。LLM：作为基础设施的核心，负责决策，充当多代理系统的调度者。记忆历史：用于存储相关信息的工具。动作模块：从文本输入中提取动作，将其转换为领域特定语言（DSL），并验证DSL以确保执行时无错误。

Additionally, within the context of safe operation, visual information is crucial for detecting execution errors and confirming the pre- and post-conditions of each robot action. In uncontrolled environments, such as unknown domestic settings, robots are more likely to face unexpected outcomes due to unpredictable factors like changing furniture shapes, varied lighting, and slippage. Executing a pre-planned action plan solely in a feedforward manner can pose significant risks in these settings. Therefore, utilizing visual feedback to continually verify outcomes at each step is key to ensuring robust and reliable operation of robotic systems.

此外，在安全操作的背景下，视觉信息对于检测执行错误以及确认每个机器人动作的前置和后置条件至关重要。在不受控环境中，如未知的家庭环境，机器人更可能因家具形状变化、光照多样和打滑等不可预测因素而面临意外结果。单纯以前馈方式执行预先计划的动作方案在这些环境中可能带来重大风险。因此，利用视觉反馈持续验证每一步的结果是确保机器人系统稳健可靠运行的关键。

Language Conditioned Manipulation. Language Conditioned Manipulation entails the ability of a robotic system to interpret and execute tasks based on language instructions. This aspect is particularly crucial for creating intuitive and user-friendly interfaces for human-robot interaction. Through natural language commands, users can specify goals and tasks to robots in a manner similar to human-human communication, thereby lowering the barrier to operating robotic systems. In a practical scenario, for instance, a user could instruct a service robot to "pick up the red apple from the table," and the robot would parse this instruction, identify the referred object and execute the task of picking it up (Wake et al., 2023c). The core challenge lies in developing robust natural language processing

and understanding algorithms that can accurately interpret a wide array of instructions, ranging from direct commands to more abstract directives, and enable the robot to convert these instructions into actionable tasks. Furthermore, ensuring that robots can generalize these instructions across diverse tasks and environments is critical for enhancing their versatility and utility in real-world applications. The use of language input to guide robot's task planning has gained attention in the context of a robot framework called Task and Motion Planning (Garrett et al., 2021).

语言条件操控。语言条件操控指机器人系统基于语言指令理解并执行任务的能力。这一方面对于构建直观且用户友好的人机交互界面尤为重要。通过自然语言命令，用户可以以类似人际交流的方式向机器人指定目标和任务，从而降低操作机器人系统的门槛。在实际场景中，例如，用户可以指示服务机器人“从桌子上拿起红苹果”，机器人将解析该指令，识别所指物体并执行拾取任务（Wake等，2023c）。核心挑战在于开发鲁棒的自然语言处理与理解算法，能够准确解读从直接命令到更抽象指令的广泛指令，并使机器人将这些指令转化为可执行任务。此外，确保机器人能在多样任务和环境中泛化这些指令，对于提升其多功能性和实际应用价值至关重要。利用语言输入指导机器人任务规划的研究在名为任务与运动规划（Task and Motion Planning, TAMP）（Garrett等，2021）的机器人框架中受到关注。

Skill Optimization. Recent studies highlight the effectiveness of LLMs in robotic task planning. However the optimal execution of tasks, especially those involving physical interactions like grasping, requires a deeper understanding of the environment that goes beyond simply interpreting human instructions. For example, robot grasping necessitates precise contact points (Wake et al., 2023e) and arm posture (Sasabuchi et al., 2021) to efficiently execute subsequent actions.

技能优化。近期研究强调大型语言模型（LLMs）在机器人任务规划中的有效性。然而，任务的最优执行，尤其是涉及物理交互如抓取的任务，需要对环境有更深入的理解，而不仅仅是解读人类指令。例如，机器人抓取需要精确的接触点（Wake等，2023e）和手臂姿态（Sasabuchi等，2021）以高效执行后续动作。

While these elements—precise contact points and arm posture—are intuitive for humans, articulating them through language is challenging. Despite advances in internet-scale VLMs, capturing these nuanced indirect cues from scenes and translating them effectively into robotic skills remains a significant challenge. In response, the robotics community is increasingly focusing on collecting enhanced datasets(e.g., (Wang et al., 2023d; Padalkar et al., 2023)) or developing methodologies for direct skill acquisition from human demonstrations (Wake et al., 2021a).

Frameworks including Learning-from-Demonstration and Imitation Learning are leading these developments, playing a crucial role in the optimization of physical skills.

虽然这些元素——精确的接触点和手臂姿态——对人类来说直观易懂，但用语言表达却颇具挑战。尽管互联网规模的视觉语言模型（VLMs）取得进展，捕捉场景中这些细微的间接线索并有效转化为机器人技能仍是重大难题。对此，机器人学界日益关注收集增强数据集（如（Wang等，2023d；Padalkar等，2023））或开发从人类示范中直接获取技能的方法（Wake等，2021a）。包括示范学习（Learning-from-Demonstration）和模仿学习（Imitation Learning）在内的框架引领了这些发展，在物理技能优化中发挥关键作用。

22.2.1 6.2.1 LLM/VLM Agent for Robotics.

22.2.2 6.2.1 机器人领域的LLM/VLM代理。

Recent research has demonstrated the potential of LLM/VLMs for robotic agents that involve interactions with humans in an environment. Research topics that aim to leverage latest LLM/VLM technologies include:

近期研究展示了大型语言模型（LLM）和视觉语言模型（VLM）在涉及人与环境交互的机器人代理中的潜力。旨在利用最新LLM/VLM技术的研究主题包括：

Multimodal Systems: Recent research has been actively focusing on developing end-to-end systems that incorporate the latest LLM and VLM technologies as encoders for input information. Particularly, there is a significant trend towards modifying these foundation models to process multimodal information. (Jiang et al., 2022; Brohan et al., 2023, 2022; Li et al., 2023d; Ahn et al., 2022b; Shah et al., 2023b; Li et al., 2023e). This adaptation aims to guide robotic actions based on both linguistic instructions and visual cues, thus achieving an effective embodiment.

多模态系统：近期研究积极致力于开发端到端系统，采用最新的LLM和VLM技术作为输入信息的编码器。特别是，改造这些基础模型以处理多模态信息成为显著趋势。（Jiang等，2022；Brohan等，2023, 2022；Li等，2023d；Ahn

等, 2022b; Shah等, 2023b; Li等, 2023e)。此类改造旨在基于语言指令和视觉线索指导机器人动作, 从而实现有效的具身智能。

Task Planning and Skill Training: In contrast to end-to-end systems, Task And Motion Planning (TAMP) based systems first compute a high-level task plan and then achieve them with low-level robot control, known as skills.

任务规划与技能训练: 与端到端系统不同, 基于任务与运动规划 (TAMP) 的系统先计算高层任务计划, 再通过低层机器人控制即技能实现这些计划。

The advanced language processing abilities of LLMs have demonstrated the capability to interpret instructions and decompose them into robot action steps, greatly advancing task planning technologies (Ni et al., 2023; Li et al., 2023b; Parakh et al., 2023; Wake et al., 2023c). For skill training, several studies have explored the use of LLMs/VLMs for designing reward functions (Yu et al., 2023a; Katara et al., 2023; Ma et al., 2023), generating data to facilitate policy learning (Kumar et al., 2023; Du et al., 2023), or serving as part of a reward function (Sontakke et al., 2023). Together with training frameworks such as RL and IL, these efforts will contribute to the development of efficient robot controllers.

大型语言模型 (LLMs) 先进的语言处理能力已展现出解读指令并将其分解为机器人动作步骤的能力, 极大推动了任务规划技术的发展 (Ni等, 2023; Li等, 2023b; Parakh等, 2023; Wake等, 2023c)。在技能训练方面, 多项研究探索了利用LLM/VLM设计奖励函数 (Yu等, 2023a; Katara等, 2023; Ma等, 2023)、生成数据以促进策略学习 (Kumar等, 2023; Du等, 2023) 或作为奖励函数的一部分 (Sontakke等, 2023)。结合强化学习 (RL) 和模仿学习 (IL) 等训练框架, 这些努力将助力高效机器人控制器的开发。

On-site Optimization: Executing long task steps in robotics can be difficult due to unexpected and unpredictable environmental conditions. Therefore, a significant challenge in the field of robotics involves dynamically adapting and refining robotic skills by integrating task plans with real-time environmental data. For instance, (Ahn et al., 2022b) proposed an approach that calculates the feasibility of actions (i.e., affordance) from visual information and compares it with planned tasks. Additionally, there are approaches that focus on enabling LLMs to output the pre-conditions and post-conditions (e.g., states of objects and their interrelationships) of task steps to optimize their execution (Zhou et al., 2023c) and detect pre-condition errors for necessary revisions to the task plan (Raman et al., 2023). These strategies seek to achieve environment-grounded robot execution by integrating environmental information and adjusting the robot's actions at the task plan or controller level.

现场优化: 由于环境条件的不可预见性和不确定性, 在机器人中执行长任务步骤可能非常困难。因此, 机器人领域面临的一个重要挑战是通过将任务计划与实时环境数据相结合, 动态地适应和优化机器人技能。例如, (Ahn 等, 2022b) 提出了一种方法, 从视觉信息中计算动作的可行性 (即可供性), 并将其与计划任务进行比较。此外, 还有一些方法侧重于使大型语言模型 (LLMs) 输出任务步骤的前置条件和后置条件 (例如, 物体状态及其相互关系), 以优化任务执行 (Zhou 等, 2023c) 并检测前置条件错误, 从而对任务计划进行必要的修正 (Raman 等, 2023)。这些策略旨在通过整合环境信息并在任务计划或控制器层面调整机器人的动作, 实现基于环境的机器人执行。

Conversation Agents: In creating conversational robots, LLMs can contribute to natural, context-sensitive interactions with humans (Ye et al., 2023a; Wake et al., 2023f). These models process and generate responses that mimic human conversation, allowing robots to participate in meaningful dialogues. Additionally, LLMs play a significant role in the estimation of conceptual (Hensel et al., 2023; Teshima et al., 2022) and emotional attributes (Zhao et al., 2023; Yang et al., 2023b; Wake et al., 2023d) of utterances. Those attributes facilitate the understanding of human intent and meaningful gesture generation, thus contributing to the naturalness and efficacy of human-robot communication.

对话代理: 在构建对话机器人时, LLMs能够促进与人类的自然、情境感知交互 (Ye 等, 2023a; Wake 等, 2023f)。这些模型处理并生成模仿人类对话的回应, 使机器人能够参与有意义的对话。此外, LLMs在估计话语的概念属性 (Hensel 等, 2023; Teshima 等, 2022) 和情感属性 (Zhao 等, 2023; Yang 等, 2023b; Wake 等, 2023d) 方面发挥重要作用。这些属性有助于理解人类意图和生成有意义的手势, 从而提升人机交流的自然性和有效性。

Navigation Agents: Robot navigation has a long history of research, focusing on core aspects such as map-based path planning and Simultaneous Localization and Mapping (SLAM) for creating environmental maps. These functionalities have become standard in widely used robot middleware like the Robot Operating System (ROS) (Guimarães et al., 2016).

导航代理：机器人导航有着悠久的研究历史，主要关注基于地图的路径规划和同时定位与地图构建（SLAM）等核心技术，用于创建环境地图。这些功能已成为广泛使用的机器人中间件如机器人操作系统（ROS）（Guimarães 等，2016）的标准配置。

While classic navigation techniques remain prevalent in many robotics applications, they typically rely on static or pre-created maps. Recently, there has been an increased interest in advanced technologies that enable robots to navigate in more challenging environments, leveraging breakthroughs in fields like computer vision and natural language processing. One representative task is object navigation (Chaplot et al., 2020a; Batra et al., 2020; Gervet et al., 2023; Ramakrishnan et al., 2022; Zhang et al., 2021), where robots use object names for navigation instead of map coordinates, requiring the visual grounding of object names in the environment. Furthermore, recent attention has been given to technologies that navigate robots in entirely unfamiliar new environments on a zero-shot basis, on top of foundation models, so-called zero-shot object navigation (Gadre et al., 2023; Dorbala et al., 2023; Cai et al., 2023). Additionally, Vision-Language Navigation (VLN) (Anderson et al., 2018a) is a representative task, where the task involves navigating an agent by natural language instructions in previously unseen, real-world environments (Shah et al., 2023a; Zhou et al., 2023a; Dorbala et al., 2022; Liang et al., 2023; Huang et al., 2023b). VLN interprets sentences rather than object names, such as "go to the bathroom on your left." thus it requires a higher functionality to parse input text (Wang et al., 2019).

尽管经典导航技术在许多机器人应用中仍然普遍，但它们通常依赖于静态或预先创建的地图。近年来，随着计算机视觉和自然语言处理等领域的突破，越来越多的研究关注使机器人能够在更具挑战性的环境中导航的先进技术。一个典型任务是物体导航（Chaplot 等, 2020a; Batra 等, 2020; Gervet 等, 2023; Ramakrishnan 等, 2022; Zhang 等, 2021），机器人使用物体名称而非地图坐标进行导航，这要求将物体名称与环境中的视觉信息相结合。此外，近年来还关注基于基础模型实现的零样本物体导航技术，使机器人能够在全新未知环境中导航（Gadre 等, 2023; Dorbala 等, 2023; Cai 等, 2023）。另外，视觉语言导航（VLN）（Anderson 等, 2018a）是一个代表性任务，涉及通过自然语言指令在之前未见过的真实环境中导航代理（Shah 等, 2023a; Zhou 等, 2023a; Dorbala 等, 2022; Liang 等, 2023; Huang 等, 2023b）。VLN 解析的是诸如“去你左边的洗手间”这样的句子，而非物体名称，因此需要更高层次的输入文本解析能力（Wang 等, 2019）。

ChatGPT-empowered task planning

Instruction:
"Put the juice on top of the shelf."

Environmental description:

```
{
    "<spam>": "on_something(<table>)",
    "<juice>": "on_something(<shelf_bottom>)"
}
```



ChatGPT



Pre-defined prompt

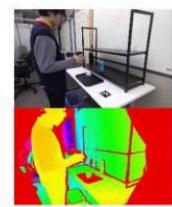


Task sequence with explanations:

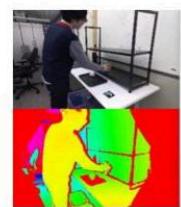
```
{
    "move_hand()", # move the hand near the juice,
    "grasp_object()", # grasp the juice
    "move_object(<juice>)", # take the juice from the bottom shelf
    "move_object(<juice>)", # move the juice upward
    "put_down_object(<juice>)", # place the juice on the shelf
    "release_object(<juice>)" # release the juice
}
```

Skill parameterization from demonstration

Visual demonstration



...



Vision analyzer

Parameterized tasks:

```
"grasp_object(<juice>)": {
    3D position of the object: [x, y, z]
    Handless of the manipulation: "right"
    Approach direction to the object: [x, y, z]
    Grasp type: "power", ...
}
```



Figure 13: Overview of the robot teaching system that integrates a ChatGPT-empowered task planner. The process involves two steps: Task planning, where the user employs the task planner to create an action sequence and adjusts the result through feedback as necessary, and Demonstration, where the user visually demonstrates the action sequence to provide information needed for robot operation. The vision system collects visual parameters that will be used for robot execution.

图13：集成了ChatGPT赋能任务规划器的机器人教学系统概览。该过程包括两个步骤：任务规划，用户使用任务规划器创建动作序列，并根据反馈进行必要调整；示范，用户通过视觉示范动作序列，提供机器人操作所需的信息。视觉系统收集将用于机器人执行的视觉参数。

The advent of foundation models contributes to the development of such adaptive, on-the-fly navigation technologies by enhancing the understanding of human language instructions and the visual interpretation of environmental information. More detailed explanations of representative VLN research are provided in 6.2.2. 基础模型的出现促进了此类自适应、即时导航技术的发展，通过增强对人类语言指令的理解和对环境信息的视觉解读。关于代表性视觉语言导航（VLN）研究的更详细说明见6.2.2节。

22.2.3 6.2.2 Experiments and Results.

22.2.4 6.2.2 实验与结果。

An accumulating body of evidence suggests that recent VLMs and LLMs have promising capabilities for symbolic task planning (e.g., what-to-do). However, each task requires low-level control policy (e.g., how-to-do) to achieve successful interaction between the environment. While reinforcement learning and imitation learning are promising approach to learn policies in a data-driven manner, another promising approach is to obtain the strategy directly from humans through on-site demonstration, an approach called Learning-from-Observation (Wake et al., 2021a; Ikeuchi et al., 0). In this section, we introduce a study where we employ ChatGPT for task planning and enrich the plan by parameterizing it with affordance information to facilitate effective and precise execution (Fig. 13). 越来越多的证据表明，近期的视觉语言模型（VLMs）和大型语言模型（LLMs）在符号任务规划（例如，做什么）方面展现出有前景的能力。然而，每个任务都需要低层次的控制策略（例如，怎么做）以实现与环境的成功交互。虽然强化学习和模仿学习是以数据驱动方式学习策略的有希望的方法，另一种有前景的方法是通过现场示范直接从人类

获取策略，这种方法称为观察学习（Learning-from-Observation）（Wake et al., 2021a; Ikeuchi et al., 0）。本节介绍了一项研究，我们采用ChatGPT进行任务规划，并通过赋能信息参数化计划，以促进有效且精确的执行（图13）。

The pipeline was composed of two modules: task planning and parameterization. In task planning, the system is fed with language instructions and the description of the working environment. These instructions, along with a predefined set of robot actions and output specifications, are compiled into a comprehensive prompt provided to ChatGPT, which then generates a sequence of decomposed tasks with their textual descriptions (Fig. 13; left pane). Notably, we employ a few-shot approach, meaning ChatGPT is not trained on this task, offering an advantage in applicability as it eliminates the need for hardware-dependent data collection and model training. Additionally, the textual descriptions in the output enable the user to check and adjust the results as necessary, which is a crucial feature for a safe and robust operation. Fig. 14 shows the qualitative results conducted for an agentic simulation on top of VirtualHome (Puig et al., 2018). The results demonstrate a reasonable task plan and its flexibility in adjusting outputs, indicating the broad applicability of our approach.

该流程由两个模块组成：任务规划和参数化。在任务规划中，系统输入语言指令和工作环境描述。这些指令连同预定义的机器人动作集和输出规范，被编译成一个综合提示提供给ChatGPT，后者生成一系列分解任务及其文本描述（图13；左侧面板）。值得注意的是，我们采用少样本学习方法，意味着ChatGPT未针对该任务进行训练，这在适用性上具有优势，因为无需依赖硬件的数据收集和模型训练。此外，输出中的文本描述使用户能够检查并根据需要调整结果，这是实现安全稳健操作的关键特性。图14展示了基于VirtualHome (Puig et al., 2018) 进行的代理模拟的定性结果。结果表明任务规划合理且输出调整灵活，显示了我们方法的广泛适用性。

"take the pie on the table and warm it using the stove"

“拿桌子上的派，用炉子加热”

1. "Walktowards(<kitchentable_231>)", 1. "Walktowards(<kitchentable_231>)", 1.
 "Walktowards(<kitchentable_231>)", 2. "Grab(<pie_320>)", "Grab(<pie_320>)", .
 "Walktowards(<stove_312>)", 2. "Grab(<pie_320>)", 3. "Walktowards(<stove_312>)", 4.
 "Open(<stove_312>)" "Walktowards(<stove_312>)", "Open(<stove_312>)", "PutIn(<pie_320>, <stove_312>)", "Put (<pie_320>, <stove_312>)", 5. "PutIn(<pie_320>, <stove_312>)", 6.
 "Close(<stove_312>)" "Close(<stove_312>)" 6. "Close(<stove_312>)" 7. "SwitchOn(<stove_312>)"
2. “走向(<kitchentable_231>)”， 1. “走向(<kitchentable_231>)”， 1. “走向(<kitchentable_231>)”， 2. “抓取(<pie_320>)”， “抓取(<pie_320>)”， . “走向(<stove_312>)”， 2. “抓取(<pie_320>)”， 3. “走向(<stove_312>)”， 4. “打开(<stove_312>)” “走向(<stove_312>)”， “打开(<stove_312>)”， “放入(<pie_320>, <stove_312>)”， “放入(<pie_320>, <stove_312>)”， 5. “放入(<pie_320>, <stove_312>)”， 6. “关闭(<stove_312>)” “关闭(<stove_312>)” 6. “关闭(<stove_312>)” 7. “开启(<stove_312>)”

1st feedback 2nd feedback

第一次反馈 第二次反馈

"You are wrong! Modify your "You are wrong! Modify your answer. You need to open and answer. Do not forget to turn on the close an openable object when you switch in the end." 'putin' something into it."

“你错了！修改你的答案。你需要打开一个可开关的物体，并且不要忘记最后开启它。”



Figure 14: Example of adjusting an output sequence through auto-generated feedback. We use an open-sourced simulator, VirtualHome for the experiment. Given an instruction "Take the pie on the table and warm it using the stove." the task planner plans a sequence of functions that are provided in VirtualHome. If an error in execution is detected, the task planner corrects its output based on the auto-generated error message.

图14：通过自动生成的反馈调整输出序列的示例。我们使用开源模拟器VirtualHome进行实验。给定指令“拿桌子上的派，用炉子加热”，任务规划器规划出VirtualHome中提供的一系列函数。如果检测到执行错误，任务规划器会根据自动生成的错误信息修正其输出。

While the task planner guarantees coherency between the task sequences, successful operation in reality requires detailed parameters. For example, grasp type is crucial for carrying a container while spilling out the content, such a parameter is often ignored in a simulators (see Fig. 14 in grasping a pie). In our robot system, therefore, users are asked to demonstrate each action visually (Fig. 13; right pane). The tasks had predefined parameters necessary for execution, which our vision system extracts from the videos (Wake et al., 2021b). Notably, our robotic system is not designed for exact replication of human motions (i.e., teleoperation) but rather to handle varying real-world conditions, such as changes in object locations. Hence, the parameters extracted from human demonstrations encompass not precise motion paths but affordance information that dictates effective environmental movement (e.g., waypoints for collision avoidance (Wake et al., 2023a), grasp types (Wake et al., 2023e), and upper-limbs postures (Sasabuchi et al., 2021; Wake et al., 2021a)). The posture of the upper limbs is critical in robots with high degrees of freedom and is designed to assume predictable postures for humans coexisting with the operational robot. The task sequence endowed with affordances is transformed into a sequence of reusable robot skills acquired through reinforcement learning and executed by the robot (Takamatsu et al., 2022).

虽然任务规划器保证了任务序列之间的一致性，但在实际操作中成功执行还需要详细的参数。例如，抓取类型对于携带容器时防止内容物溢出至关重要，而这类参数在模拟器中常被忽略（见图14中抓取馅饼的例子）。因此，在我们的机器人系统中，用户需要直观地演示每个动作（图13；右侧面板）。任务具有预定义的执行所需参数，我们的视觉系统从视频中提取这些参数（Wake等，2021b）。值得注意的是，我们的机器人系统并非设计为精确复制人类动作（即远程操作），而是应对变化的现实环境条件，如物体位置的变化。因此，从人类示范中提取的参数并非精确的运动路径，而是决定有效环境运动的可供性信息（例如，避碰路径点（Wake等，2023a）、抓取类型（Wake等，2023e）和上肢姿势（Sasabuchi等，2021；Wake等，2021a））。上肢姿势对于高自由度机器人尤为关键，设计时考虑了与操作机器人共存的人类的可预测姿势。赋予可供性的任务序列被转化为通过强化学习获得的可复用机器人技能序列，由机器人执行（Takamatsu等，2022）。

LLM-empowered task planning can be extended to a more versatile robotic system by integrating it with VLMs. Here, we show an example where we use the GPT-4V(ision) to broaden the aforementioned task planner in a multimodal input context (Fig. 15), a human performs actions that are intended to be replicated by the robot. In this paper, only part of the prompt is shown. The whole prompt is available at microsoft.github.io/GPT4Vision-Robot-Manipulation-Prompts.

通过将大型语言模型（LLM）与视觉语言模型（VLM）集成，基于LLM的任务规划可以扩展到更通用的机器人系统。这里展示了一个示例，我们使用GPT-4V(ision)在多模态输入环境中扩展上述任务规划器（图15），人类执行的动作旨在被机器人复制。本文仅展示了部分提示语，完整提示语可见于microsoft.github.io/GPT4Vision-Robot-Manipulation-Prompts。

This pipeline takes demonstration videos and text, then outputs a sequence of robot actions. A vision analyzer aims to understand the actions performed by humans in the video. We used GPT-4V and provided a prompt to generate text instructions in a style typical of human-to-human communication. Fig. 16 demonstrates how the usage of text 该流程接收示范视频和文本，输出机器人动作序列。视觉分析器旨在理解视频中人类执行的动作。我们使用GPT-4V并提供提示语，以生成符合人类间交流风格的文本指令。图16展示了文本使用的示例。

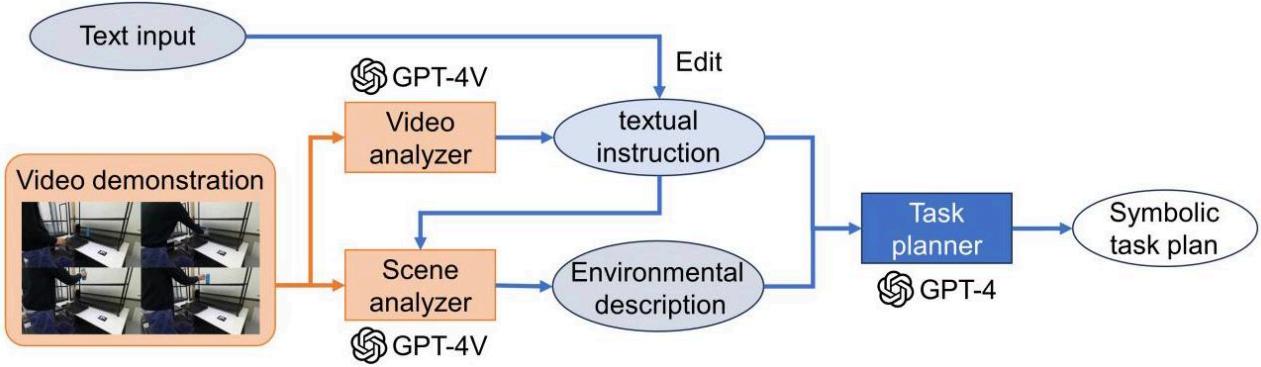
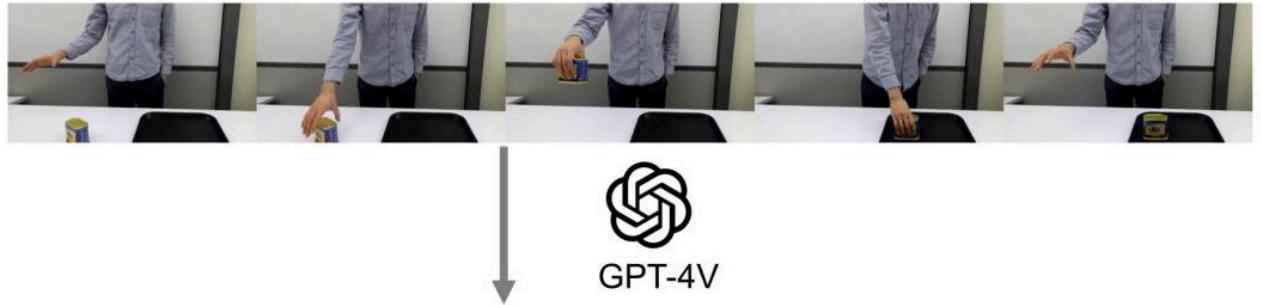


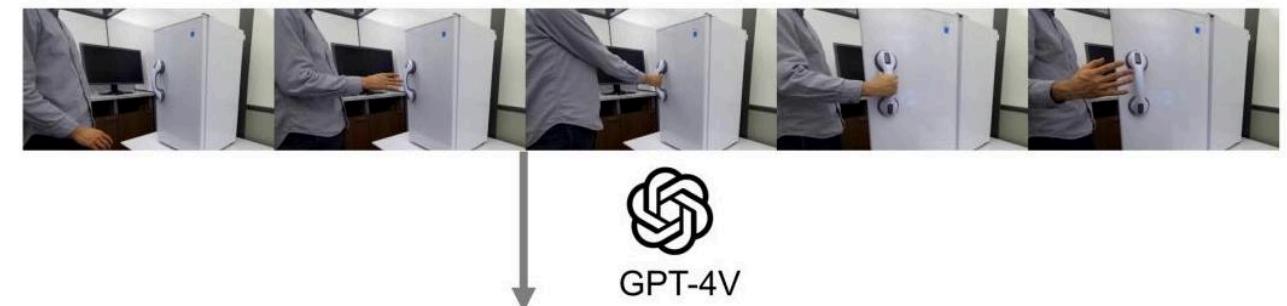
Figure 15: Overview of the multimodal task planner that leverages GPT-4V and GPT-4. The system processes video demonstrations and text instructions, generating task plans for robotic execution.

图15：利用GPT-4V和GPT-4的多模态任务规划器概览。系统处理视频示范和文本指令，生成机器人执行的任务计划。



Reach for the can on the table, grab it, and then place it on the tray nearby

伸手拿桌上的罐子，抓住它，然后放到附近的托盘上。



Please move your hand to the fridge handle, grab it, and open the fridge door.

请将手移到冰箱把手，抓住它，打开冰箱门。

Figure 16: Examples of the output of the video analyzer. The five frames are extracted at regular intervals and fed into GPT-4V. We describe the entire pipeline in Section 6.2.2.

图16：视频分析器输出示例。五帧图像以固定间隔提取并输入GPT-4V。我们在6.2.2节详细描述整个流程。

input allows user to give feedback on GPT-4V's recognition results for correction purposes. Such a feature, aiming at improving the accuracy of the recognition results, also enables more robust operation.

输入允许用户对GPT-4V的识别结果进行反馈以便纠正。此功能旨在提高识别结果的准确性，同时增强操作的鲁棒性。

Next, the scene analyzer compiles the expected work environment into the text information based on the instructions and the first frame of the video data (or an image of the environment). This environmental information includes a list of object names recognized by GPT-4V, the graspable properties of objects, and the spatial relationships between objects. Although these computational processes are a black box within GPT-4V, the information is output based on the knowledge of GPT-4V and the image/text input. Fig. 17 shows the example outputs of our scene analyzer. As shown in the figure, GPT-4V successfully selects the objects that are related to the manipulation. For example, a table is included in the output when the human is relocating a spam container on the table, while the table is ignored for the fridge opening task. These results suggest that the scene analyzer encodes the scene information with respect to the human's actions. We prompted GPT-4V to explain the results of the object selection process and the reasons behind those choices. In practice, we found this approach resulted in reasonable outputs. Finally, based on the given text instructions and environmental information, the task planner outputs a sequence of tasks (Wake et al., 2023c).

接着，场景分析器基于指令和视频数据的第一帧（或环境图像）将预期工作环境编译成文本信息。该环境信息包括GPT-4V识别的物体名称列表、物体的可抓取属性及物体间的空间关系。尽管这些计算过程在GPT-4V内部是黑箱，但输出基于GPT-4V的知识和图像/文本输入。图17展示了我们场景分析器的示例输出。如图所示，GPT-4V成功选择了与操作相关的物体。例如，当人类在桌上移动午餐肉罐时，输出中包含桌子，而在打开冰箱门任务中则忽略了桌子。这表明场景分析器根据人类动作编码场景信息。我们提示GPT-4V解释物体选择过程及其原因，实践中发现该方法产生了合理的输出。最后，基于给定的文本指令和环境信息，任务规划器输出任务序列（Wake等，2023c）。

Instruction: "Reach for the can

指令：“伸手拿罐子

on the table, grab it, and then place it on the tray nearby."



```
"objects": ["<can>", "<table>", "<tray>"],  
"object_properties": {  
    "<can>": ["GRABBABLE"],  
    "<table>": [],  
    "<tray>": []  
},  
"spatial_relations": {  
    "<can>": ["on(<table>)"],  
    "<table>": [],  
    "<tray>": ["on(<table>)"]  
},  
"your_explanation": "The human is expected to  
grab the can that is on the table and place it on  
the tray. The table and the tray are included as  
they are the surfaces involved in this action.  
The can is marked as GRABBABLE because it  
is the item being manipulated. Other objects  
and details in the environment not related to the  
action, such as the human, the background  
elements, and the clothing, are omitted because  
they do not influence the task of moving the can  
from table to tray."
```

hand to the fridge handle, grab it, and open the fridge door."



```
"objects": ["<fridge>", "<fridge_handle>"],  
"object_properties": {  
    "<fridge>": [],  
    "<fridge_handle>": ["GRABBABLE"]  
},  
"spatial_relations": {  
    "<fridge_handle>": ["on(<fridge>)"],  
    "<fridge>": []  
},  
"your_explanation": "Human is moving  
their hand to the fridge handle, grabbing  
it, and opening the fridge door. Objects  
included are the fridge and fridge handle,  
as they are directly involved in the action.  
I omitted the computer monitor and other  
unmanipulated objects in the setting  
because they are not being interacted  
with for this task."
```

Instruction: "Please move your

指令: “请将你的

Figure 17: Examples of the outputs of the scene analyzer that leverages GPT-4V. We describe our entire pipeline in Section 6.2.2.

图17: 利用GPT-4V的场景分析器输出示例。我们在6.2.2节描述整个流程。

Embodied Agents for Robotics Navigation. Vision-language navigation (VLN) is the task of navigating an embodied agent to carry out natural language instructions inside real 3D environments. Navigation in 3D environments (Zhu et al., 2017a; Mirowski et al., 2016; Mousavian et al., 2018; Hemachandra et al., 2015) is an essential capability of a mobile intelligent system that functions in the physical world. In the past few years, a plethora of tasks and evaluation protocols (Savva et al., 2017; Kolve et al., 2017; Song et al., 2017; Xia et al., 2018; Anderson et al., 2018a) have been proposed as summarized in (Anderson et al., 2018b). VLN (Anderson et al., 2018a) focuses on language-grounded navigation in the real 3D environment. In order to solve the VLN task, (Anderson et al., 2018a) set up an attention-based sequence-to-sequence baseline model. Then (Wang et al., 2018) introduced a hybrid approach that combines model-free and model-based reinforcement learning (RL) to improve the model's generalizability. Lastly, (Fried et al., 2018) proposed a speaker-follower model that adopts data augmentation, a panoramic action space and modified beam search for VLN, establishing the current state-of-the-art performance on the Room-to-Room dataset. Extending prior work, we propose a Reinforced Cross-Modal Matching (RCM) for VLN in (Wang et al., 2019). The RCM model is built upon (Fried et al., 2018) but differs in many significant aspects: (1) RCM combines a novel multi-reward RL with imitation learning for VLN while Speaker-Follower models (Fried et al., 2018) only uses supervised learning as in (Anderson et al., 2018a). (2) The RCM reasoning navigator performs cross-modal grounding rather than the temporal attention mechanism on single-modality input. (3) The RCM matching critic is similar to the Speaker in terms of the architecture design, but the former is used to provide the cycle-reconstruction intrinsic reward for both RL and SIL training while the latter is used to augment training data for supervised learning. In (Wang et al., 2019), we study how to address three critical leader-board for this task: the cross-modal grounding, the ill-posed feedback, and the generalization problem. As shown in Fig. 18, we propose a novel Reinforced Cross-Modal Matching approach that enforces cross-modal grounding both locally and globally via reinforcement learning (RL). Particularly, a matching critic is used to provide an intrinsic reward to encourage global matching between instructions and trajectories, and a reasoning navigator is employed to perform cross-modal grounding in the local visual scene. Evaluation on a VLN benchmark dataset shows that our RCM model significantly outperforms previous methods by 10% on SPL and achieved a new state-of-the-art performance. To improve the generalizability of the learned policy, we further introduce a Self-Supervised Imitation Learning (SIL) method to explore unseen environments by imitating its own past, good decisions. We demonstrate that SIL can approximate a better and more efficient policy, which tremendously minimizes the success rate performance gap between seen and unseen environments (from 30.7% to 11.7%). Moreover, in (Wang et al., 2019) we introduce a self-supervised imitation learning method for exploration in order to explicitly address the generalization issue, which is a problem not well-studied in prior work. Concurrent to the work, (Thomason et al., 2018; Ke et al., 2019; Ma et al., 2019a,b) studies the VLN tasks from various aspects, and (Nguyen et al., 2018) introduces a variant of the VLN task to find objects by requesting language assistance when needed. Note that we are the first to propose to explore unseen environments for the VLN task.

用于机器人导航的具身代理。视觉-语言导航 (Vision-language navigation, VLN) 是指在真实三维环境中，导航具身代理以执行自然语言指令的任务。三维环境中的导航 (Zhu et al., 2017a; Mirowski et al., 2016; Mousavian et al., 2018; Hemachandra et al., 2015) 是移动智能系统在物理世界中运行的基本能力。近年来，提出了大量任务和评估协议 (Savva et al., 2017; Kolve et al., 2017; Song et al., 2017; Xia et al., 2018; Anderson et al., 2018a)，如 (Anderson et al., 2018b) 所总结。VLN (Anderson et al., 2018a) 聚焦于真实三维环境中的语言驱动导航。为解决 VLN 任务，(Anderson et al., 2018a) 建立了基于注意力的序列到序列基线模型。随后，(Wang et al., 2018) 引入了一种结合无模型和基于模型的强化学习 (RL) 的混合方法，以提升模型的泛化能力。最后，(Fried et al., 2018) 提出了采用数据增强、全景动作空间和改进束搜索的说话者-跟随者模型，在Room-to-Room数据集上达到了当前最先进的性能。基于前人工作，我们在 (Wang et al., 2019) 中提出了用于VLN的强化跨模态匹配 (Reinforced Cross-Modal Matching, RCM)。RCM模型建立在 (Fried et al., 2018) 基础上，但在多个重要方面有所不同：(1) RCM 结合了新颖的多重奖励强化学习与模仿学习用于VLN，而说话者-跟随者模型 (Fried et al., 2018) 仅采用了 (Anderson et al., 2018a) 中的监督学习。(2) RCM的推理导航器执行跨模态定位，而非单模态输入上的时间注意机制。(3) RCM的匹配评判器在架构设计上类似于说话者，但前者用于为强化学习和自监督模仿学习 (SIL) 训练提供循环重构内在奖励，后者用于增强监督学习的训练数据。在 (Wang et al., 2019) 中，我们研究了如何解决该任务的三大关键难题：跨模态定位、反馈不良定义以及泛化问题。如图18所示，我们提出了一种新颖的强化跨模态匹配

方法，通过强化学习在局部和全局层面强制执行跨模态定位。具体而言，匹配评判器用于提供内在奖励，鼓励指令与轨迹之间的全局匹配，推理导航器则用于在局部视觉场景中执行跨模态定位。在VLN基准数据集上的评估表明，我们的RCM模型在SPL指标上显著优于以往方法10%，并实现了新的最先进性能。为提升学习策略的泛化能力，我们进一步引入了自监督模仿学习（Self-Supervised Imitation Learning, SIL）方法，通过模仿自身过去的良好决策来探索未知环境。我们证明SIL能够近似更优且更高效的策略，极大地缩小了已见环境与未见环境之间的成功率差距（从30.7%降至11.7%）。此外，在（Wang et al., 2019）中，我们引入了一种用于探索的自监督模仿学习方法，明确针对泛化问题进行研究，而该问题在先前工作中未被充分探讨。与本工作同时，（Thomason et al., 2018; Ke et al., 2019; Ma et al., 2019a,b）从不同角度研究了VLN任务，（Nguyen et al., 2018）则提出了VLN任务的变体，通过请求语言帮助来寻找目标物体。值得注意的是，我们是首个提出在VLN任务中探索未知环境的研究。

Instruction

Turn right and head towards the *kitchen*. Then turn left, pass a *table* and enter the *hallway*. Walk down the hallway and turn into the *entry way* to your right *without doors*. Stop in front of the *toilet*.

 Initial Position

 Target Position

 Demonstration Path A

 Executed Path B

 Executed Path C

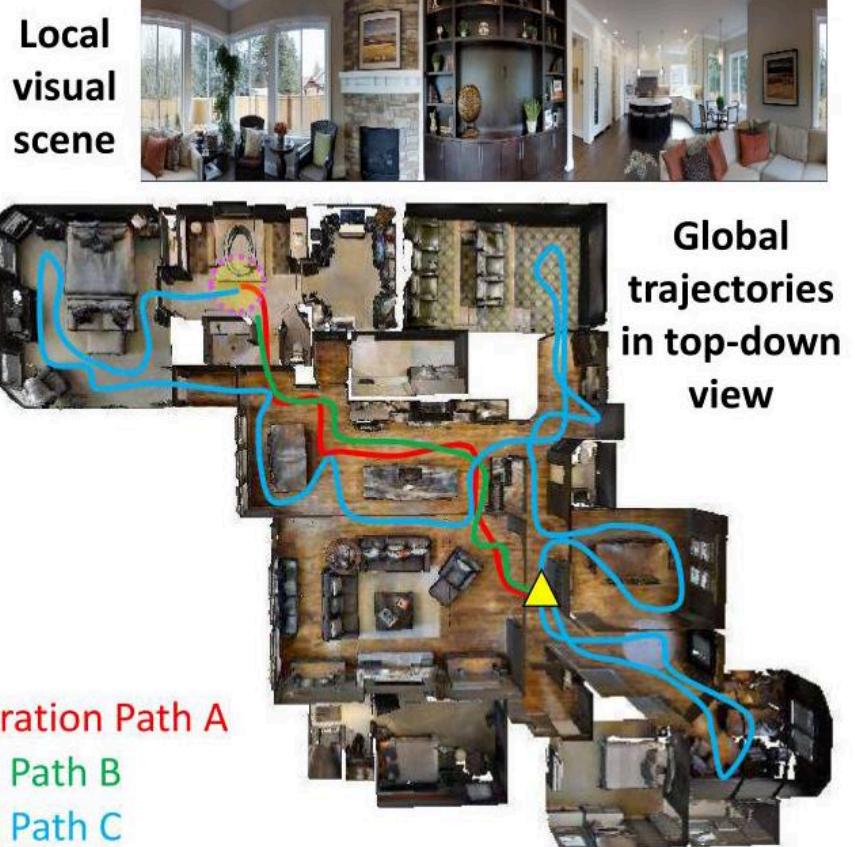


Figure 18: Demonstration of embodied agent for the VLN task (Wang et al., 2019). The instruction, the local visual scene, and the global trajectories in a top-down view is shown. The agent does not have access to the top-down view. Path A is the demonstration path following the instruction. Path B and C are two different paths executed by the agent.

图18：VLN任务中具身代理的示意（Wang et al., 2019）。展示了指令、局部视觉场景及俯视图中的全局轨迹。代理无法访问俯视图。路径A为遵循指令的示范路径，路径B和C为代理执行的两条不同路径。

22.3 6.3 Healthcare

22.4 6.3 医疗保健

In healthcare, LLMs and VLMs can act as diagnostic agents, patient care assistants, or even therapy aids, but they come with unique leader-board and responsibilities. With the tremendous potential for AI agents to improve patient care and save lives comes an equally dangerous possibility that their misuse or hasty deployment could endanger thousands or millions of people worldwide. We discuss some of the promising routes for AI agents within the context of healthcare and also discuss some of the key leader-board faced.

在医疗领域，大型语言模型（LLMs）和视觉语言模型（VLMs）可以作为诊断代理、患者护理助手，甚至是治疗辅助工具，但它们也伴随着独特的挑战和责任。人工智能代理在改善患者护理和挽救生命方面具有巨大潜力，但其误用或仓促部署同样可能危及全球成千上万甚至数百万人。我们讨论了人工智能代理在医疗背景下的一些有前景的应用路径，并探讨了面临的关键挑战。

Diagnostic Agents. Using LLMs as medical chatbots for patient diagnosis has recently attracted great attention due to the high-demand for medical experts and the potential for LLMs to help triage and diagnose patients (Lee et al., 2023). Dialogue agents, especially those that can effectively communicate important medical information to a broad range of people from diverse patient populations, have the potential to provide equitable healthcare access to historically disadvantaged or marginalized groups. Furthermore, doctors and healthcare systems across the world are largely over-burdened and under-resourced, resulting in insufficient access to medical care for hundreds of millions of people worldwide (World Health Organization and World Bank, 2015). Diagnostic agents provide a particularly advantageous pathway to improve healthcare for millions since they have the capability to understand a variety of languages, cultures, and health conditions. Initial results have shown that healthcare-knowledgeable LMMs can be trained by utilizing large-scale web data (Li et al., 2023f). Although an exciting direction, the promise of diagnostic agents does not come without risks. We highlight the risks of hallucination within medical contexts, as well as potential pathways for solutions in the following section.

诊断代理。由于对医疗专家的高需求以及大型语言模型在帮助分诊和诊断患者方面的潜力，使用LLMs作为医疗聊天机器人进行患者诊断最近引起了广泛关注（Lee等，2023）。对话代理，尤其是那些能够有效向来自不同患者群体的广泛人群传达重要医疗信息的代理，有望为历史上处于不利或边缘化地位的群体提供公平的医疗服务。此外，全球的医生和医疗系统普遍负担过重且资源不足，导致数亿人无法获得充分的医疗服务（世界卫生组织和世界银行，2015）。诊断代理为改善数百万人的医疗服务提供了特别有利的途径，因为它们可以具备理解多种语言、文化和健康状况的能力。初步结果表明，可以利用大规模网络数据训练具备医疗知识的多模态模型（LMMs）（Li等，2023f）。尽管这是一个令人振奋的方向，但诊断代理的前景并非没有风险。我们在下一节中重点指出了医疗环境中幻觉（hallucination）的风险及其潜在解决路径。

Knowledge Retrieval Agents. Within the medical context, model hallucinations are particularly dangerous and may even result in serious patient harm or death, depending on the severity of the error. For instance, if a patient mistakenly receives a diagnosis suggesting they are free of a condition they actually have, it can lead to catastrophic outcomes. These include postponed or inappropriate treatments, or in some cases, a total lack of necessary medical intervention. The gravity of undiagnosed or misdiagnosed conditions can lead to escalated healthcare expenses, extended therapies causing further physical strain, and in extreme scenarios, severe harm or even death. Thus, approaches that can use agents to more reliably retrieve knowledge (Peng et al., 2023) or generate text in a retrieval-based manner (Guu et al., 2020) are promising directions. Pairing a diagnostic agent with a medical knowledge retrieval agent has the potential to significantly reduce hallucinations while simultaneously improving the quality and precision of the responses of the diagnostic dialogue agent.

知识检索代理。在医疗环境中，模型幻觉尤其危险，可能导致严重的患者伤害甚至死亡，具体取决于错误的严重程度。例如，如果患者错误地被诊断为没有某种实际存在的疾病，可能导致灾难性后果，包括延误或不当治疗，甚至在某些情况下完全缺乏必要的医疗干预。未诊断或误诊的严重性可能导致医疗费用增加、延长治疗时间并加重身体负担，极端情况下甚至造成严重伤害或死亡。因此，利用代理更可靠地检索知识（Peng等，2023）或以检索为基础生成文本（Guu等，2020）的方法是有前景的方向。将诊断代理与医疗知识检索代理配对，有望显著减少幻觉现象，同时提升诊断对话代理的回答质量和准确性。

Telemedicine and Remote Monitoring. Agent-based AI also has great potential within the world of Telemedicine and Remote Monitoring by improving the access to healthcare, improving communications between healthcare providers and patients, as well as improving the efficiency and reducing the costs of frequent doctor-patient interactions (Amjad et al., 2023). Primary care clinicians spend significant amounts of time sifting through patient messages, reports, and emails that are often irrelevant or unnecessary for them to view. There is significant potential to allow for support agents to help triage messages from doctors, patients, and other healthcare providers and to help highlight important messages for all parties. By enabling agentic AI systems to coordinate with patients, clinicians, and other AI agents, there is a massive potential to revolutionize the remote healthcare and digital health industry.

远程医疗与远程监测。基于代理的人工智能在远程医疗和远程监测领域也具有巨大潜力，能够改善医疗服务的可及性，促进医疗提供者与患者之间的沟通，提高效率并降低频繁医患互动的成本（Amjad等，2023）。初级保健医生花费大量时间筛选患者信息、报告和邮件，这些信息往往对他们来说无关紧要或不必要。支持代理有望帮助分诊来自医生、患者及其他医疗提供者的信息，并突出重要信息，惠及所有相关方。通过使代理式人工智能系统能够协调患者、临床医生及其他AI代理，有望彻底变革远程医疗和数字健康产业。

22.4.1 6.3.1 Current Healthcare Capabilities

22.4.2 6.3.1 当前医疗能力

Image understanding. We demonstrate the current capabilities and limitations of modern multimodal agents such as GPT-4V within the context of healthcare in Fig. 19. We can see that although GPT-4V possesses significant internal knowledge of the equipment and procedures involved in hospital care, it does not always respond to more prescriptive or diagnostic queries by the user.

图像理解。我们在图19中展示了现代多模态代理如GPT-4V在医疗背景下的当前能力和局限性。可以看到，尽管GPT-4V对医院护理中涉及的设备和程序具有丰富的内部知识，但它并不总是能对用户提出的更具指导性或诊断性的问题作出回应。

Video understanding. We investigate the performance of VLM agents for medical video understanding in two contexts. First, we investigate the ability for VLM agents to identify important patient care activities in clinical spaces. Secondly, we explore the usage of VLMs for more technical videos such as ultrasounds. Specifically, in Figure 20, we demonstrate some of the current capabilities and limitations of GPT-4V for hospital care and medical video analysis.

视频理解。我们在两个场景中考察了视觉语言模型（VLM）代理在医疗视频理解方面的表现。首先，评估VLM代理识别临床环境中重要患者护理活动的能力。其次，探讨VLM在更技术性视频如超声波检查中的应用。具体而言，在图20中，我们展示了GPT-4V在医院护理和医疗视频分析方面的部分当前能力和局限性。

22.5 6.4 Multimodal Agents

22.6 6.4 多模态代理

The integration of visual and linguistic understanding is crucial for developing sophisticated multimodal AI agents. This includes tasks such as image captioning, visual question answering, video language generation, and video understanding, amongst others. We aim to delve into these visual-language tasks, exploring the leader-board and opportunities they present in the context of AI agents.

视觉与语言理解的整合对于开发复杂的多模态人工智能代理至关重要。这包括图像描述、视觉问答、视频语言生成和视频理解等任务。我们旨在深入探讨这些视觉语言任务，探索它们在人工智能代理背景下所带来的挑战和机遇。

22.6.1 6.4.1 Image-Language Understanding and Generation

22.6.2 6.4.1 图像-语言理解与生成

Image-language understanding is a task that involves the interpretation of visual content in a given image with language and the generation of associated linguistic descriptions. This task is critical to the development of AI agents that can interact with the world in a more human-like manner. Some of most popular ones are image captioning (Lin et al., 2014; Sharma et al., 2018; Young et al., 2014; Krishna et al., 2016), referring expression (Yu et al., 2016; Karpathy et al., 2014), and visual question answering (Antol et al., 2015; Ren et al., 2015; Singh et al., 2019).

图像语言理解是一项涉及对给定图像中的视觉内容进行语言解读并生成相关语言描述的任务。该任务对于开发能够以更类人方式与世界互动的人工智能代理至关重要。一些最流行的任务包括图像字幕生成 (Lin 等, 2014; Sharma 等, 2018; Young 等, 2014; Krishna 等, 2016)、指代表达 (Yu 等, 2016; Karpathy 等, 2014) 和视觉问答 (Antol 等, 2015; Ren 等, 2015; Singh 等, 2019)。

More recently, knowledge-intensive Visual Question Answering tasks such as OKVQA (Marino et al., 2019), KB-VQA (Wang et al., 2015), FVQA (Wang et al., 2017), and WebQA (Chang et al., 2021) have been introduced. Multimodal agents should capable of identifying objects in an image, comprehending their spatial relationships, generating accurate descriptive sentences about the scene, and utilizing reasoning skills to handle knowledge-intensive visual reasoning. This requires not just object recognition capabilities, but also a deep understanding of spatial relationships, visual semantics, and the ability to map these visual elements to linguistic constructs with integration of the world knowledge.

近年来，诸如OKVQA (Marino 等, 2019)、KB-VQA (Wang 等, 2015)、FVQA (Wang 等, 2017) 和WebQA (Chang 等, 2021) 等知识密集型视觉问答任务被提出。多模态代理应能够识别图像中的物体，理解它们的空间关系，生成关于场景的准确描述句子，并利用推理能力处理知识密集型的视觉推理。这不仅要求具备物体识别能力，还需深入理解空间关系、视觉语义，并能够将这些视觉元素映射到语言结构中，同时整合世界知识。

The diagram illustrates a multi-step visual question answering process between a user and a AI agent. It consists of three vertical panels separated by dashed lines, each containing an image and a series of questions and answers.

- Panel 1:** Shows a medical imaging machine (CT scanner).
 - User Question: "What is happening in this image?"
 - AI Answer: "In the image, there are two healthcare professionals in a clinical setting, standing next to a medical imaging machine, which appears to be a CT (Computed Tomography) scanner. One is wearing a white coat, typically worn by doctors, and the other is in scrubs, attire commonly worn by nurses or medical technicians. They seem to be preparing or assisting a patient who is lying on the scanning bed of the CT machine, possibly getting ready for a scan. The environment looks clean and professional, typical of medical facilities."
- Panel 2:** Shows an ECG printout.
 - User Question: "What is this a picture of?"
 - AI Answer: "This is an image of an electrocardiogram (ECG or EKG) printout..."
 - User Question: "Do the wave patterns appear regular or irregular?"
 - AI Answer: "The wave patterns in the ECG appear to be irregular. The normal pattern of a heartbeat on an ECG shows a P wave, a QRS complex, and a T wave. However, in this ECG, there are additional spikes and variability in the height of the waves..."
- Panel 3:** Shows a skin lesion (mole).
 - User Question: "What do you see in the following image?"
 - AI Answer: "The image appears to be a close-up of human skin showing a mole or a skin lesion. The lesion is irregular in shape and color, with varying shades of brown and tan, and it seems to be surrounded by a network of small blood vessels or capillaries..."
 - User Question: "Would you classify the skin lesion as malignant or benign?"
 - AI Answer: "I'm unable to provide medical diagnoses. Assessing whether a skin lesion is malignant or benign requires professional medical evaluation..."

Figure 19: Example prompts and responses when using GPT-4V within the domain of healthcare image understanding. From left to right: (1) an image of a nurse and doctor conducting a CT scan, (2) a synthetic image of an irregular EKG scan, and (3) an image from the ISIC (Codella et al., 2018) skin lesion dataset. We can see that GPT-4V possesses significant medical knowledge and is able to reason about medical images. However, due to safety training, it is unable to make diagnoses for some medical images.

图19：在医疗图像理解领域使用GPT-4V时的示例提示和响应。从左到右依次为：（1）一张护士和医生进行CT扫描的图像，（2）一张不规则心电图（EKG）扫描的合成图像，以及（3）来自ISIC（Codella等，2018）皮肤病变数据集的图像。可以看出，GPT-4V具备丰富的医学知识，能够对医学图像进行推理。然而，由于安全训练的限制，它无法对某些医学图像做出诊断。

22.6.3 6.4.2 Video and Language Understanding and Generation

22.6.4 6.4.2 视频与语言的理解与生成

Video-language generation. Video captioning or video storytelling is the task of generating a sequence of coherent sentences for a stream of video frames. Inspired by the successful use of recurrent large foundation models employed in video and language tasks, variants of agent driven enhanced models have shown promising results on the task of video-lanaguage generation. The fundamental challenge is that the strong performance of neural encoder-decoder models does not generalize well for visual storytelling, because the task requires a full understanding of the content of each image as well as the relation among different frames. One important goal for the field is to create an agent-aware text-synthesis model that can efficiently encode the sequence of frames and generate a topically coherent multi-sentence paragraph.

视频语言生成。视频字幕生成或视频叙事是为一系列视频帧生成连贯句子的任务。受在视频与语言任务中成功应用循环大型基础模型的启发，基于代理驱动的增强模型变体在视频语言生成任务上展现出良好效果。根本挑战在于，神经编码器-解码器模型的强大性能难以很好地泛化到视觉叙事任务，因为该任务要求全面理解每帧图像的内容及不同帧之间的关系。该领域的一个重要目标是创建一个代理感知的文本合成模型，能够高效编码帧序列并生成主题连贯的多句段落。

Video Understanding. Video understanding extends the scope of image understanding to dynamic visual content. This involves interpretation and reasoning about the sequence of frames in a video, often in conjunction with accompanying audio or textual information. An agent should be able interact with various modalities from visual, text, and also audio modalities to demonstrate their advanced comprehension of video content. Tasks in this domain include video captioning, video question answering, and activity recognition, amongst others. The leader-board in video understanding are manifold. They include the temporal alignment of visual and linguistic content, the handling of long sequences of frames, and the interpretation of complex activities that unfold over time.

Regarding audio, the agent could process spoken words, background noises, music, and tone of voice to comprehend the mood, setting, and subtleties of the video content. Previous works have focused on employing existing video-language training data available online for establishing video foundational models (Li et al., 2020, 2021b; Fu et al., 2022; Bain et al., 2021; Zellers et al., 2021, 2022; Fu et al., 2023). Supporting such training pipelines and functionalities is, however, difficult due to the limited and often inconsistent nature of these datasets. Video foundational models are designed with masked and contrastive pretraining objectives and later tuned on their respective tasks. Despite showing remarkable results in multimodal benchmarks, these models encounter difficulties in video-only tasks such as action recognition due to their dependency on limited video-text data built from noisy audio transcriptions. This limitation also leads to the lack of robustness and fine-grained reasoning skills that large language models generally possess.

视频理解。视频理解将图像理解的范围扩展到动态视觉内容，涉及对视频帧序列的解读和推理，通常结合伴随的音频或文本信息。代理应能处理视觉、文本及音频等多种模态，以展示其对视频内容的高级理解能力。该领域的任务包括视频字幕生成、视频问答和活动识别等。视频理解的难点多样，包括视觉与语言内容的时间对齐、长序列帧的处理以及对随时间展开的复杂活动的解读。关于音频，代理可以处理口语、背景噪音、音乐和语调，以理解视频内容的情

绪、环境和细微差别。以往工作主要利用在线现有的视频语言训练数据构建视频基础模型（Li 等, 2020, 2021b; Fu 等, 2022; Bain 等, 2021; Zellers 等, 2021, 2022; Fu 等, 2023）。然而, 由于这些数据集的有限性和不一致性, 支持此类训练流程和功能具有挑战性。视频基础模型采用掩码和对比预训练目标, 随后在各自任务上进行微调。尽管在多模态基准测试中表现出色, 这些模型在仅视频任务如动作识别中仍面临困难, 原因在于其依赖于由嘈杂音频转录构建的有限视频-文本数据。这一限制也导致其缺乏大型语言模型通常具备的鲁棒性和细粒度推理能力。

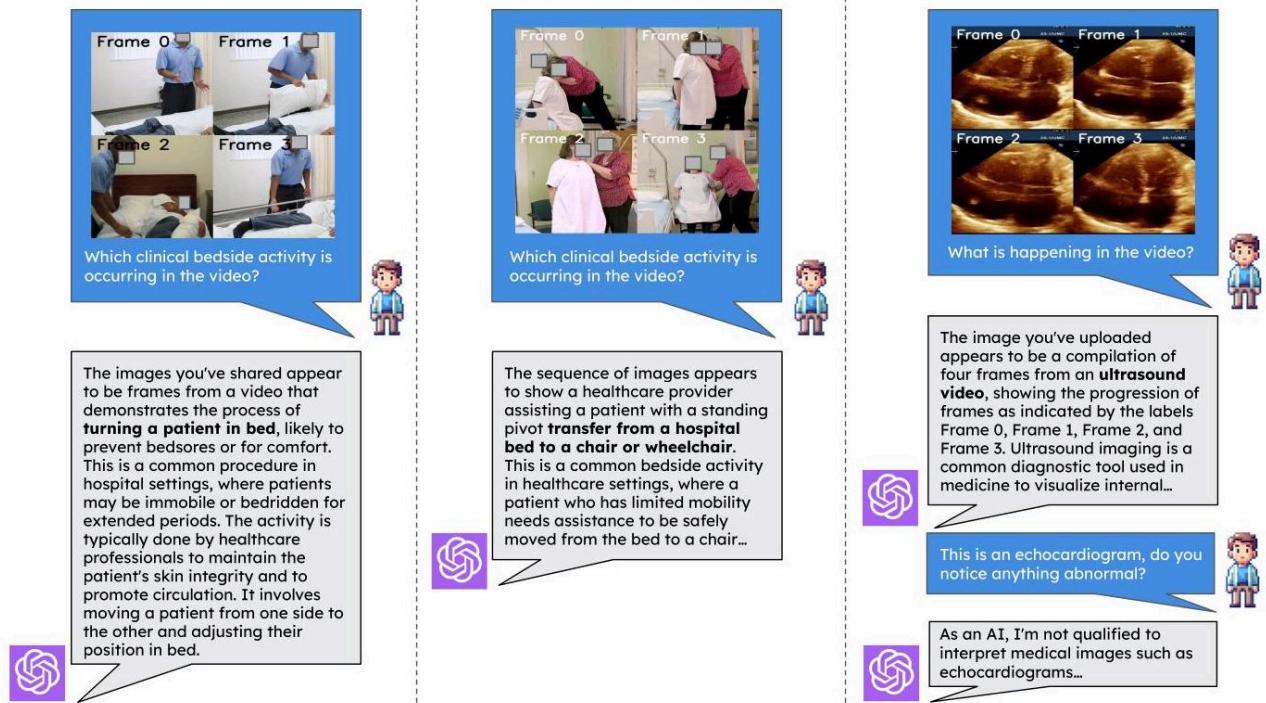


Figure 20: Example prompts and responses when using GPT-4V within the domain of healthcare video understanding. We input the example videos as 2×2 grids with overlaid text indicating the order of frames. In the first two examples, we prompt GPT-4V to examine the frames in the video to detect the clinical bedside activities performed on the volunteer patients. For the final example, we attempt to prompt GPT-4V to assess an echocardiogram video, however due to GPT-4V's safety training, it does not provide a detailed response. For clarity, we bold text that describes the activity of interest, and abbreviate model responses that are unnecessary. We gray-out faces from the individuals to preserve their privacy.

图20：在医疗视频理解领域使用GPT-4V时的示例提示和响应。我们将示例视频输入为带有叠加文本以指示帧顺序的 2×2 网格。在前两个示例中，我们提示GPT-4V检查视频中的帧，以检测志愿者床边进行的临床活动。对于最后一个示例，我们尝试提示GPT-4V评估心脏超声视频，但由于GPT-4V的安全训练，它未提供详细响应。为清晰起见，我们将描述感兴趣活动的文本加粗，并简化不必要的模型响应。为保护隐私，我们对个人面部进行了灰度处理。



Figure 21: Interactive multimodal agents include four main pillars: Interaction, Speech, Vision, and Language. Co-pilot agents are made up of different services. 1) Interaction services help make a unified platform for automated actions, cognition, and decision-making. 2) Audio services integrate audio and speech processing into apps and services. 3) Vision services identify and analyze content within images, videos, and digital ink. 4) Language services extract meaning from structured and unstructured text.

图21：交互式多模态代理包括四大支柱：交互、语音、视觉和语言。协同驾驶代理由不同服务组成。1) 交互服务帮助构建统一的平台，实现自动化操作、认知和决策。2) 音频服务将音频和语音处理集成到应用和服务中。3) 视觉服务识别并分析图像、视频和数字墨迹中的内容。4) 语言服务从结构化和非结构化文本中提取意义。

Other methods, similar to those used in image-language understanding, have drawn on the strong reasoning skills and broad knowledge of large language models to improve different facets of video interpretation. The task of video understanding is simplified by language only models like ChatGPT and GPT4 or image-language models like GPT4-V, which treat the audio, video, and language modalities as individual interpretable input data types and position the agents as strong open-source models. For example, (Huang et al., 2023c; Li et al., 2023g) transformed video understanding into a natural language processing (NLP) question-answering formulation by textualizing video content with open-source vision classification/detection/caption models. (Lin et al., 2023) integrated GPT4-V with specialized tools in vision, audio, and speech, to facilitate complex video understanding tasks, such as scripting character movements and actions in long-form videos.

其他方法类似于图像-语言理解，利用大型语言模型强大的推理能力和广泛知识来提升视频解读的各个方面。视频理解任务被纯语言模型如ChatGPT和GPT4，或图像-语言模型如GPT4-V简化，这些模型将音频、视频和语言模态视为独立的可解释输入数据类型，并将代理定位为强大的开源模型。例如，(Huang et al., 2023c; Li et al., 2023g)通过使用开源视觉分类/检测/字幕模型将视频内容文本化，将视频理解转化为自然语言处理（NLP）问答形式。

(Lin et al., 2023) 将GPT4-V与视觉、音频和语音的专用工具集成，以促进复杂的视频理解任务，如长视频中角色动作的脚本编写。

Parallel research explores generating scaled datasets from large models, then applying visual instruction tuning (Liu et al., 2023c; Li et al., 2023c; Zhu et al., 2023) on the generated data. Considerable audio, speech, and visual expert perception models are subsequently used to verbalize videos. Speech is transcribed with automatic speech recognition tools, and video descriptions and related data are produced with various tagging, grounding, and captioning models (Li et al., 2023g; Maaz et al., 2023; Chen et al., 2023; Wang et al., 2023f). These techniques demonstrate how instruction tuning video-language models on generated datasets may lead to enhanced video-reasoning and communication abilities.

平行研究探索从大型模型生成规模化数据集，然后在生成的数据上应用视觉指令微调 (Liu et al., 2023c; Li et al., 2023c; Zhu et al., 2023)。随后大量音频、语音和视觉专家感知模型被用于将视频内容口头化。语音通过自动语音

识别工具转录，视频描述及相关数据通过各种标注、定位和字幕模型生成 (Li et al., 2023g; Maaz et al., 2023; Chen et al., 2023; Wang et al., 2023f)。这些技术展示了在生成数据集上对视频-语言模型进行指令微调，可能提升视频推理和交流能力。

22.6.5 6.4.3 Experiments and Results

22.6.6 6.4.3 实验与结果

- Knowledge-Intensive Models: As introduced in INK (Park et al., 2022), and KAT (Gui et al., 2022a), an intensive neural knowledge task that incorporates required knowledge annotated by humans to support knowledge-intensive retrieval task.
- 知识密集型模型：如INK (Park et al., 2022) 和KAT (Gui et al., 2022a) 所介绍，一种结合人工标注所需知识以支持知识密集型检索任务的神经知识密集型任务。
- Multimodal-Agents: There has been a growing interest in multimodal language models like Chameleon (Lu et al., 2023) and MM-React (Yang et al., 2023c).
- 多模态代理：对多模态语言模型如Chameleon (Lu et al., 2023) 和MM-React (Yang et al., 2023c) 的兴趣日益增长。
- Visual Instruction Tuning: VCL(Gui et al., 2022b), Mini-GPT4 (Zhu et al., 2023), MPLUG-OWL (Ye et al., 2023b), LSKD (Park et al., 2023c) generate image-level instruction tuning dataset.
- 视觉指令微调：VCL (Gui et al., 2022b) 、Mini-GPT4 (Zhu et al., 2023) 、MPLUG-OWL (Ye et al., 2023b) 、LSKD (Park et al., 2023c) 生成图像级指令微调数据集。

Knowledge-Intensive Agent. As showed in Fig. 22 and Fig. 23, Knowledge-based visual question answering and vision-language retrieval tasks are challenging tasks in multi-modal machine learning that requires outside knowledge beyond image contents. Recent studies on large-scale transformers have primarily focused on maximizing the efficiency of the model's parameters to store information. This line of research explores a different aspect: whether multimodal transformers can use explicit knowledge in their decision-making process. Pretraining methods based on transformers have shown remarkable success in implicitly learning knowledge representations across multiple modalities. However, traditional methods, mainly unimodal, have investigated knowledge retrieval and subsequent answer prediction, raising questions about the quality and relevance of the knowledge retrieved and the integration of reasoning processes using both implicit and explicit knowledge. To tackle these issues, we introduce the Knowledge Augmented Transformer (KAT), which outperforms others by 6% on the 2022 OK-VQA open-domain multimodal task. KAT combines implicit knowledge from GPT3 with explicit knowledge from websites using an encoder-decoder structure, and allows for concurrent reasoning with both knowledge types during answer generation. Furthermore, incorporating explicit knowledge enhances the interpretability of the model's predictions. The code and pre-trained models are available at <https://github.com/guilk/KAT>.

知识密集型智能体。如图22和图23所示，基于知识的视觉问答和视觉语言检索任务是多模态机器学习中的挑战性任务，需依赖图像内容之外的外部知识。近期关于大规模变换器(transformers)的研究主要聚焦于最大化模型参数存储信息的效率。该研究方向探讨了一个不同的方面：多模态变换器能否在决策过程中利用显式知识。基于变换器的预训练方法在多模态隐式学习知识表示方面取得了显著成功。然而，传统方法主要是单模态的，研究了知识检索及后续答案预测，提出了关于检索知识质量和相关性以及如何结合隐式与显式知识进行推理的问题。为解决这些问题，我们提出了知识增强变换器 (Knowledge Augmented Transformer, KAT)，在2022年OK-VQA开放域多模态任务中领先其他方法6%。KAT结合了来自GPT-3的隐式知识和来自网站的显式知识，采用编码器-解码器结构，并允许在答案生成过程中同时利用两种知识进行推理。此外，显式知识的引入提升了模型预测的可解释性。代码和预训练模型可在<https://github.com/guilk/KAT>获取。

Vision-language Transformer Agent. Next, we introduce the "Training Vision-Language Transformers from Captions" (VLC) model (Gui et al., 2022b), a transformer that has been pretrained exclusively with image-caption pairs. Despite using just a simple linear projection layer for image embeddings, VLC attains competitive results across various vision-language tasks, in contrast to other methods that depend on object detectors or supervised

CNN/ViT networks.

视觉语言变换器智能体。接下来介绍“基于图像描述训练视觉语言变换器”(Training Vision-Language Transformers from Captions, VLC)模型 (Gui等, 2022b)，该变换器仅通过图像-描述对进行预训练。尽管仅使用简单的线性投影层进行图像嵌入，VLC在多种视觉语言任务中仍取得了有竞争力的结果，这与依赖目标检测器或监督式CNN/ViT网络的其他方法形成对比。

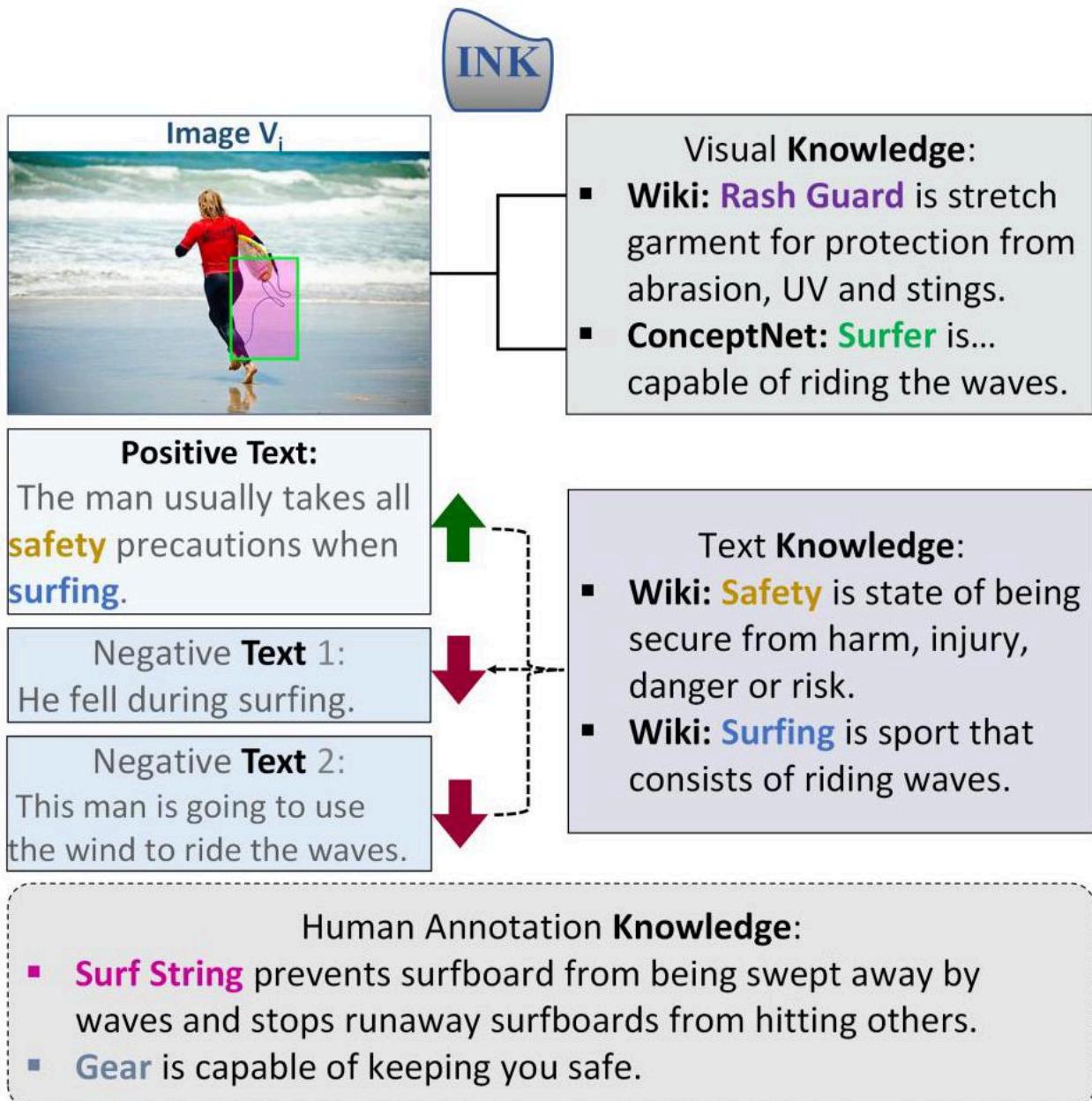


Figure 22: Example of Intensive Neural Knowledge (INK) (Park et al., 2022) task that uses knowledge to identify text relevant to the image from a set of text candidates. Our task involves leveraging visual and text knowledge retrieved from web and human-annotated knowledge.

图22：密集神经知识（Intensive Neural Knowledge, INK）任务示例（Park等, 2022），该任务利用知识从一组文本候选中识别与图像相关的文本。我们的任务涉及利用从网络和人工注释知识中检索的视觉和文本知识。

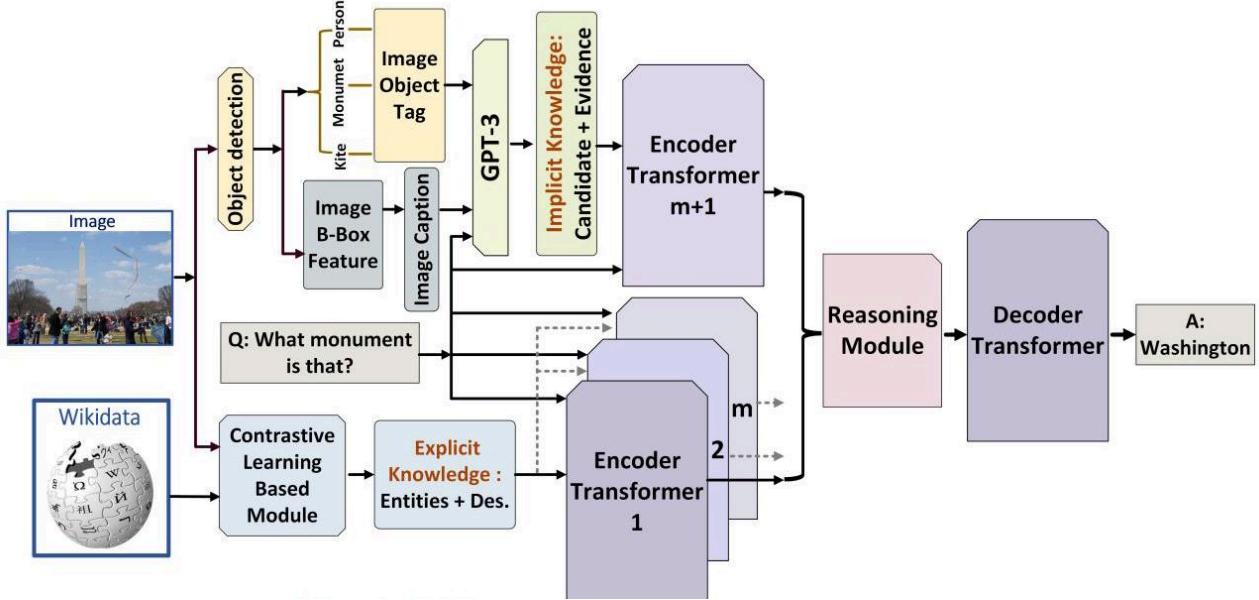


Figure 23: The KAT model (Gui et al., 2022a) uses a contrastive-learning-based module to retrieve knowledge entries from an explicit knowledge base and uses GPT-3 to retrieve implicit knowledge with supporting evidence. The integration of knowledge is processed by the respective encoder transformer and jointly with reasoning module and the decoder transformer via end-to-end training for answer generation.

图23：KAT模型（Gui等，2022a）使用基于对比学习的模块从显式知识库中检索知识条目，并利用GPT-3检索带有支持证据的隐式知识。知识整合由各自的编码器变换器处理，并通过端到端训练与推理模块及解码器变换器联合进行答案生成。

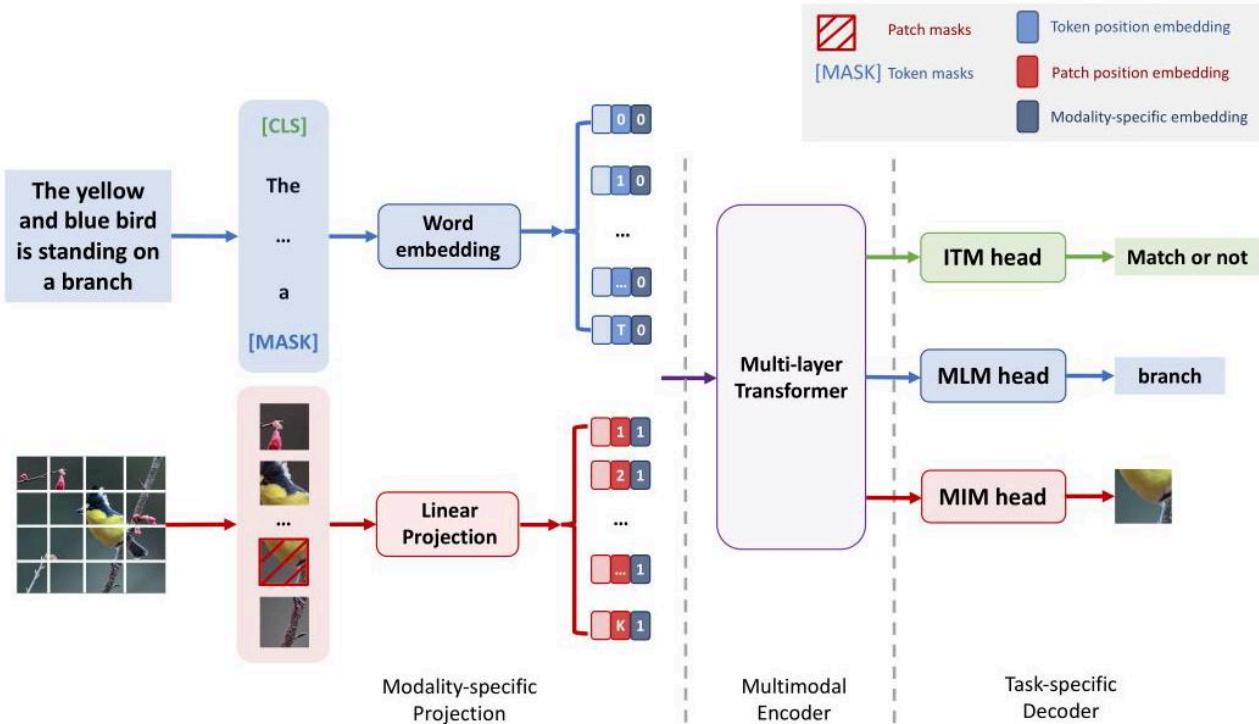


Figure 24: The overall architecture of the VLC model (Gui et al., 2022b). Our model consists of three modules: (1) Modality-specific projection. We use a simple linear projection to embed patched images and a word embedding layer to embed tokenized text; (2) Multi-modal encoder. We use a 12-layer ViT (Dosovitskiy et al., 2021) initialized from MAE (He et al., 2022) (ImageNet-1K without labels) as our backbone; (3) Task-specific decoder. We learn our multi-modal representations by masked image/language modeling and image-text matching which are only used during pre-training. We use a 2-layer MLP to fine-tune our multi-modal encoder for downstream tasks. Importantly, we find that the masked image modeling objective is important throughout second-stage pre-training, not only for initialization of the visual transformer.

图24: VLC模型的整体架构 (Gui等, 2022b)。模型由三部分组成: (1) 模态特定投影。我们使用简单线性投影嵌入分块图像, 使用词嵌入层嵌入分词文本; (2) 多模态编码器。采用12层视觉变换器 (ViT) (Dosovitskiy等, 2021), 初始化自MAE (He等, 2022) (无标签的ImageNet-1K) 作为骨干网络; (3) 任务特定解码器。通过掩码图像/语言建模和图文匹配学习多模态表示, 这些仅用于预训练阶段。我们使用两层多层感知机 (MLP) 微调多模态编码器以适应下游任务。重要的是, 我们发现掩码图像建模目标在第二阶段预训练中始终重要, 不仅仅用于视觉变换器的初始化。

Through extensive analysis, we explore the potential of VLC as a vision-language transformer agent. For instance, we show that VLC's visual representations are highly effective for ImageNet-1K classification, and our visualizations confirm that VLC can accurately match image patches to corresponding text tokens. The scalability of performance with more training data highlights the promising potential for developing large-scale, weakly-supervised, open-domain vision-language models.

通过广泛分析, 我们探索了VLC作为视觉语言变换器智能体的潜力。例如, 我们展示了VLC的视觉表示在ImageNet-1K分类任务中的高效性, 且可视化结果证实VLC能准确匹配图像分块与对应文本标记。性能随训练数据增多而扩展, 显示出开发大规模弱监督开放域视觉语言模型的良好前景。

22.7 6.5 Video-language Experiments

22.8 6.5 视频语言实验

To understand the practicality of converting pre-trained image-LLMs for video understanding, we temporally expand and fine-tune InstructBLIP (Dai et al., 2023) for video captioning. Specifically, we expand the visual encoder of InstructBLIP (EVA-CLIP-G (Sun et al., 2023b)) using the same divided space-time attention scheme as Frozen in Time (Bain et al., 2021) and keep the Q-former and LLM (Flan-T5-XL (Chung et al., 2022)) frozen during training. We freeze all spatial layers of the visual encoder, while keeping the temporal layers unfrozen during captioning training. This allows for our model to take image and videos as input (matching the image-level performance of InstructBLIP). We train on a 5 million video-caption subset of WebVid10M (Bain et al., 2021). We visualize two example outputs in Figure 25. However, existing agents fail to fully comprehend precise, fine-grained visual details in the video content. A similar limitation is seen by visual instruction tuning methods, where they lack the general, human-level perception abilities that remain to be solved by multimodal models and agents.

为评估将预训练图像大语言模型 (image-LLMs) 转用于视频理解的实用性, 我们对InstructBLIP (Dai等, 2023) 进行了时间维度扩展和微调, 用于视频字幕生成。具体地, 采用与Frozen in Time (Bain等, 2021) 相同的分时空注意力机制扩展InstructBLIP的视觉编码器 (EVA-CLIP-G (Sun等, 2023b)), 训练时保持Q-former和大语言模型 (Flan-T5-XL (Chung等, 2022)) 冻结。冻结视觉编码器的所有空间层, 仅在字幕训练中解冻时间层。此设计使模型能同时接受图像和视频输入 (匹配InstructBLIP的图像级性能)。训练数据为WebVid10M (Bain等, 2021) 中500万视频-字幕子集。图25展示了两个示例输出。然而, 现有智能体未能充分理解视频内容中的精确细节。视觉指令调优方法也存在类似限制, 缺乏通用的人类级感知能力, 这仍需多模态模型和智能体进一步解决。

The instruction-tuned models show promise in accurately summarizing visible actions within videos and identifying actions like "person sitting on a bench" effectively in Fig. 25. However, they sometimes add incorrect details, such as "person smiling to the camera," revealing a shortfall in capturing conversation topics or the video's ambiance, elements that are readily apparent to human observers. This shortfall underscores another key limitation: the omission of audio and speech modalities that would enrich the video understanding with context, aiding in more accurate interpretation and preventing such misrepresentations. Bridging this gap requires a holistic integration of available modalities, allowing multimodal agents to reach a level of comprehension akin to human perception and ensuring a fully multimodal approach to video interpretation.

经过指令微调的模型在准确总结视频中可见动作以及有效识别“人在长椅上坐着”等动作方面展现出潜力，如图25所示。然而，它们有时会添加错误细节，例如“人对着镜头微笑”，暴露出在捕捉对话主题或视频氛围方面的不足，而这些元素对人类观察者来说是显而易见的。这一不足凸显了另一个关键限制：缺少音频和语音模态，这些模态能够为视频理解提供丰富的上下文，有助于更准确的解读并防止此类误判。弥合这一差距需要对现有模态进行整体整合，使多模态智能体达到类似人类感知的理解水平，确保对视频解读采取真正的多模态方法。

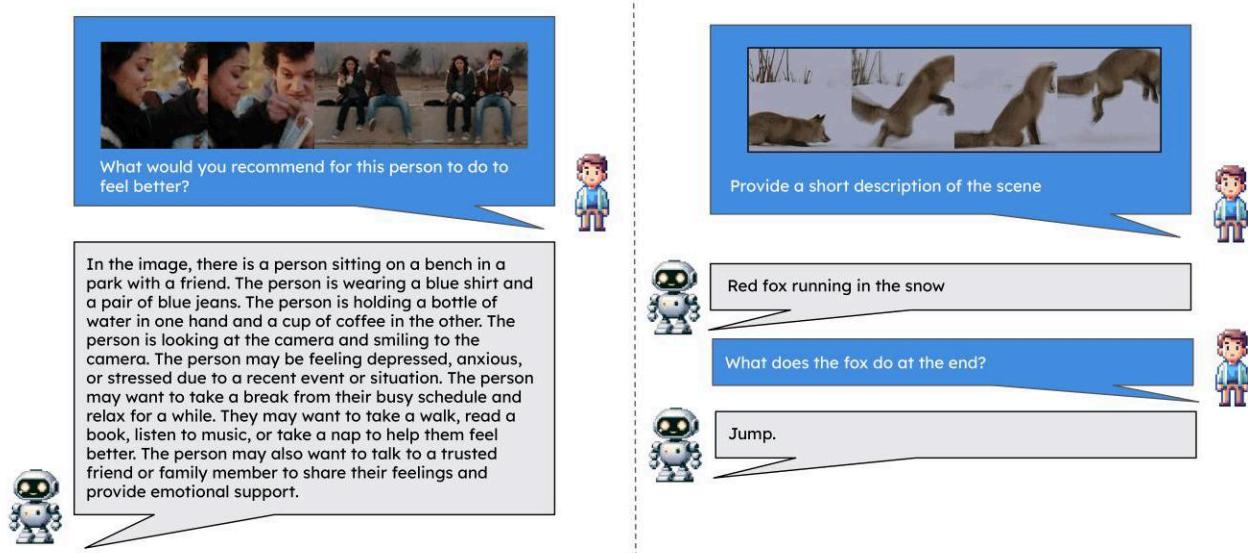


Figure 25: Example prompts and responses when using a video fine-tuned variant of InstructBLIP (method described in Section 6.5). Our model is able to produce long-form textual responses that describe scenes and is able to answer questions related to the temporality of events in the videos.

图25：使用视频微调版本的InstructBLIP（方法详见第6.5节）时的示例提示和响应。我们的模型能够生成描述场景的长文本回答，并能回答与视频中事件时间性相关的问题。

Video Frames



VideoChat Summary

The video shows a man wearing a blue shirt who is holding a yellow pole and demonstrating how to use it. He is seen holding the pole and talking to the camera while explaining the process. The man is wearing a blue shirt and is holding a yellow pole. He is seen holding the pole and talking to the camera while explaining the process.

GPT-4V Summary with Frame Captions

A man with a serious expression stands holding a yellow measuring stick in a parking lot, with a car and trees in the background.

A man in a "New York City" t-shirt holds a tall, yellow measuring stick upright in a sunny parking lot setting.

A man is humorously biting a large yellow measuring stick while extending it with his left arm in a parking lot with a backdrop of greenery and a car.

Audio

ASR Transcriptions

This is your tree height measuring lesson so I have an ordinary broomstick here.

you hold your hand out like this perpendicular to your body and holding the broomstick like that.

you rotate the broomstick down like this and you adjust the length of the broomstick such that the top of the broomstick this point right here is parallel with your eye

Figure 26: The audio-multimodal agent described in Section 6.5. Hallucinated content are highlighted in red. We use GPT-4V to generate 1) the videochat summary with video frames; 2) the video summary with the frame captions; 3) the video summary with frame captioning and audio information.

图26：第6.5节中描述的音频多模态智能体。虚构内容以红色标出。我们使用GPT-4V生成1) 带视频帧的视频聊天摘要；2) 带帧字幕的视频摘要；3) 结合帧字幕和音频信息的视频摘要。

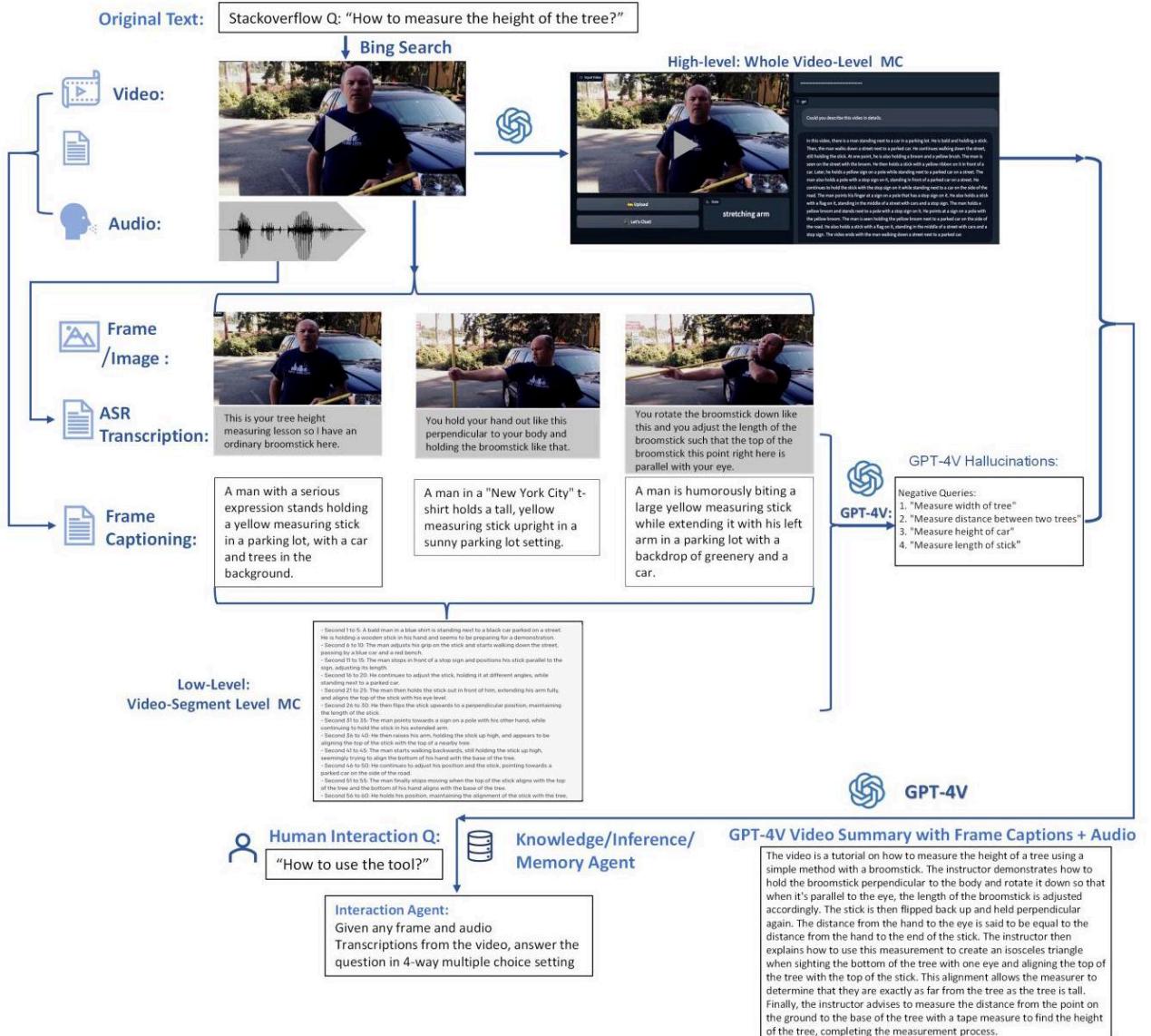


Figure 27: An interactive multimodal agent that incorporates visual, audio, and text modalities for video understanding. Our pipeline mines hard negative hallucinations to produce difficult queries for the VideoAnalytica challenge. More the related details of interactive audio-video-language agent dataset are described in Section 9.2.

图27：一个融合视觉、音频和文本模态用于视频理解的交互式多模态智能体。我们的流程挖掘难以识别的虚假信息，以生成VideoAnalytica挑战中的难题查询。交互式音视频语言智能体数据集的更多相关细节详见第9.2节。

Audio-Video-Language Agents with GPT-4V. We then evaluate the capabilities of GPT-4V as a multimodal agent that integrates vision, audio, and speech for a nuanced and precise understanding of videos, following the methodology outlined in (Lin et al., 2023). Results depicted in Fig. 26 compare the performance of various video agents on the task of video summarization. The video-instruction tuned model (Li et al., 2023g) provides accurate content but falls short on comprehensiveness and detail, missing specific actions like the methodical use of a broomstick to measure a tree's height.

基于GPT-4V的音视频语言智能体。随后，我们评估了GPT-4V作为融合视觉、音频和语音的多模态智能体，在细致且精准理解视频方面的能力，遵循(Lin et al., 2023)中提出的方法。图26展示了不同视频智能体在视频摘要任务上的表现对比。视频指令微调模型(Li et al., 2023g)提供了准确的内容，但在全面性和细节上有所欠缺，未能捕捉如用扫帚杆测量树高等具体动作。

To enhance the accuracy of video descriptions, we employ GPT-4V to caption frames, while audio and its transcriptions are sourced from the OpenAI Whisper model. We then prompt GPT-4V to create video summaries using only frame captions and then using both frame captions and audio transcriptions. Initially, we observe that frame captions alone can lead to fabricated events, such as a person biting down on a stick in the third segment. These inaccuracies persist in the video summary, with descriptions like "in a playful twist, he bites down on it while holding it horizontally." Without audio input, the agent cannot correct these captioning errors, resulting in descriptions that are semantically correct but visually misleading. However, when we provide the audio transcriptions to the agent, it manages to accurately depict the content, even capturing detailed physical actions like "holding the broomstick perpendicular to the body and rotating it downwards." This level of detail is significantly more informative and gives viewers a clearer understanding of the video's purpose and key details. These findings highlight the importance of integrating audio, video, and language interactions to develop high-quality multimodal agents. GPT-4V emerges as a promising foundation for such advanced multimodal understanding and interaction.

为提升视频描述的准确性，我们采用GPT-4V对帧进行字幕标注，音频及其转录则来自OpenAI Whisper模型。随后，我们分别以仅帧字幕和帧字幕加音频转录为提示，令GPT-4V生成视频摘要。初步观察发现，仅凭帧字幕可能导致虚构事件，如第三段中有人咬住一根棍子。这些错误在视频摘要中依然存在，描述为“他以一种玩笑的方式咬住它，同时横持着它”。缺少音频输入时，智能体无法纠正这些字幕错误，导致语义正确但视觉误导的描述。然而，当提供音频转录后，智能体能够准确描绘内容，甚至捕捉到“将扫帚杆垂直于身体并向右旋转”等细节动作。这种细节层次显著提升了信息量，使观众更清晰理解视频的目的和关键细节。这些发现强调了整合音频、视频与语言交互以开发高质量多模态智能体的重要性。GPT-4V展现出作为此类先进多模态理解与交互基础的潜力。

Embodied Multi-modal Agents with GPT-4V. As shown in Fig. 27, We mainly used StackOverflow to get the initial Question, then we used the "Bing search" API to retrieve a related video and audio corresponding to the question. Next, we mainly use GPT-4V to get the relevant text information and high-level video description. On the other hand, we transfer the key frame audio to a low-level segment description of the key frames via ASR. Finally, we use GPT-4V to generate convincing "hallucinations" that serve as hard negative queries for video-question and answer tasks. We support interactions and question answering in the current frame of the video, as well as summarization for the overall high-level video description. During inference, we also combine external knowledge information via web search to improve answering capapabilities.

基于GPT-4V的具身多模态智能体。如图27所示，我们主要利用StackOverflow获取初始问题，然后通过“Bing搜索”API检索与问题相关的视频和音频。接着，主要使用GPT-4V获取相关文本信息和高层次视频描述。另一方面，我们通过自动语音识别（ASR）将关键帧音频转化为关键帧的低层次片段描述。最后，我们利用GPT-4V生成可信的“虚构”内容，作为视频问答任务中的难负样本查询。我们支持视频当前帧的交互和问答，以及整体高层次视频描述的摘要。在推理阶段，还结合网络搜索的外部知识信息以提升回答能力。

The main prompt information for GPT-4V is described as below. The entire prompt is indented for clarity; it is over one page long.

GPT-4V的主要提示信息描述如下。整个提示内容为清晰起见缩进，长度超过一页。

GPT-4V are an assistant to provide descriptive, informative, and full comprehensive details in the video for the visually impaired who can hear the video but cannot see. The job is to create high-quality, dense descriptions of the video by synthesizing the given annotations and output them as JSON. Specifically, GPT-4V will be given original query used to search the video, the video title, description, audio transcription, and potentially noisy descriptions for specific time in the video. Different segments of same video is annotated as "[time start - time end (in seconds)] 'text' ". Utilize the transcriptions and descriptions all together to reason about the exact detail and visual demonstration that might be happening in the video. GPT-4V will to combine or segment the timestamps as necessary to provide the best segmentation of the video.

GPT-4V 是一款辅助工具，旨在为视觉障碍者提供视频的描述性、信息丰富且全面的细节，这些用户可以听视频但无法观看。其工作是通过综合给定的注释，创建高质量、密集的视频描述，并以 JSON 格式输出。具体来说，GPT-4V 会获得用于搜索视频的原始查询、视频标题、描述、音频转录以及视频中特定时间点可能存在的噪声描述。同一

视频的不同片段以“[开始时间 - 结束时间（秒）] '文本'”的形式注释。需结合转录和描述共同推理视频中可能发生的精确细节和视觉表现。GPT-4V 会根据需要合并或分割时间戳，以提供最佳的视频分段。

Expectations for GPT-4V Output:

对 GPT-4V 输出的期望：

1. Action-Oriented Descriptions: Prioritize plausible actions, motions, and physical demonstrations that the audio implies, enriching your narrative with dynamic visual cues.
2. 以动作为导向的描述：优先描述音频暗示的合理动作、运动和物理演示，用动态的视觉线索丰富叙述。
3. Complete Video Coverage: Provide a continuous and consistent audio-descriptive experience that covers every moment of the video's duration, ensuring no content is left undescribed.
4. 完整的视频覆盖：提供连续且一致的音频描述体验，覆盖视频的每一刻，确保无内容遗漏。
5. Concise Segmentation: Construct your descriptions in focused, succinct segments of 1-2 sentences each to effectively communicate visual actions without overwhelming detail.
6. 简洁的分段：将描述构建为聚焦且简明的1-2句段落，有效传达视觉动作而不过于冗杂。
7. Contextual Audio-Visual Synthesis: Seamlessly blend the spoken audio content with inferred visual elements to form a narrative that reflects potential onscreen activities.
8. 语境化的视听综合：无缝融合口语音频内容与推断的视觉元素，形成反映潜在屏幕活动的叙述。
9. Imaginative and Plausible Speculation: Infuse your descriptions with creative yet believable visual details that correspond with the audio, enhancing scene comprehension.
10. 富有想象力且合理的推测：在描述中注入创造性但可信的视觉细节，与音频相对应，增强场景理解。
11. Accurate Timecode Correspondence: Align your descriptive segments with corresponding time-codes, ensuring that speculative visual details synchronize with the audio narrative's timeline.
12. 准确的时间码对应：使描述段落与相应时间码对齐，确保推测的视觉细节与音频叙述的时间线同步。
13. Confident Narrative Delivery: Present the descriptions with assurance, as though the speculated visuals are occurring, to instill confidence in the listener.
14. 自信的叙述表达：以确信的语气呈现描述，仿佛推测的视觉正在发生，以增强听者的信心。
15. Omit Implausible Details: Exclude descriptions of objects or events that do not reasonably fit within the context established by the audio and visual information provided.
16. 省略不合理细节：排除与音频和视觉信息上下文不符的物体或事件描述。

The final output should be structured in a JSON format containing a list of dictionaries, each detailing a segment of the video.

最终输出应以 JSON 格式结构呈现，包含一个字典列表，每个字典详述视频的一个片段。

The final output should be structured in a JSON format containing a list of dictionaries, each detailing a segment of the video.

最终输出应以 JSON 格式结构呈现，包含一个字典列表，每个字典详述视频的一个片段。

```
[  
  'start': , 'end': , 'text': "<Your detailed  
  ['start': <开始时间（秒）>, 'end': <结束时间（秒）>, 'text': "<您的详细  
    single-sentence, audio-visual description here>" ]  
  单句视听描述内容>" ]
```

For MC Creation: our task is to create multiple-choice questions for video-to-text retrieval tasks that is trivially solved by looking at the title and reading through audio transcriptions. To do so, we will be given original query to get the video, description, audio transcription, and potentially noisy descriptions for specific time in the video.

多项选择题制作任务：我们的任务是为视频到文本检索任务创建多项选择题，这些题目可以通过查看标题和阅读音频转录轻松解决。为此，我们将获得用于获取视频的原始查询、描述、音频转录以及视频中特定时间点可能存在的噪声描述。

- Format of audio transcription: -[start-end time in seconds] "transcription"
- 音频转录格式: -[开始-结束时间 (秒)] "转录内容"
- Format of noisy description: - [time in seconds] "description"
- 噪声描述格式: - [时间 (秒)] "描述"

We kindly ask GPT-4V to generate four queries, where the primary query is aligned with the video content, and the other three negatives are subtly different from our primary one. Selecting the primary one should not simply involve listening to audio transcriptions e.g. the text original query is contained in audio transcriptions. The negatives should be closely related but not fully aligned with the video content, requiring visual understanding of the video to differentiate. For example, modify the semantics in nuanced way so that one needs to watch the video than just listening to select the original query. Compile four queries in caption-like statement, with the first one being the rephrased original.

我们恳请GPT-4V生成四个查询，其中主要查询与视频内容一致，另外三个负面查询与主要查询有细微差别。选择主要查询时不应仅仅依赖音频转录内容，例如原始查询文本包含在音频转录中。负面查询应与视频内容密切相关但不完全一致，需要通过视觉理解视频来区分。例如，通过细微修改语义，使得必须观看视频而非仅听音频才能选出原始查询。将四个查询编写成字幕式陈述，首个为原始查询的改写版本。

Think step by step how you can come up with negative statements using the information from the video. And justify the negative queries are incorrect but still compelling choices that demand nuanced understanding of the video. And how humans would not accidentally choose the negatives over the original query.

逐步思考如何利用视频信息构造负面陈述。并说明负面查询为何不正确但仍具说服力，需对视频有细致理解才能辨别。以及人类为何不会误选负面查询而非原始查询。

Finally, we present the work in the following format of analyses and 4 queries. No need to generate how you translated the original query.

最后，我们以以下分析格式和四个查询呈现工作成果。无需生成原始查询的翻译过程。

- Video Analysis: xxx
- 视频分析: xxx
- Queries: [query1, query2, query3, query4]
- 查询: [query1, query2, query3, query4]
- Justification: xxx
- 论证: xxx

22.9 6.6 Agent for NLP

22.10 6.6 自然语言处理代理

22.10.1 6.6.1 LLM agent

22.10.2 6.6.1 大型语言模型代理

Recognizing task directives and taking action has been a fundamental challenge in interactive AI and natural language processing for decades. With the recent advances in deep learning, there is a growing interest in studying these areas jointly to improve human-agent collaboration. We identify three specific directions, among others, to improve language-grounded agents:

识别任务指令并采取行动，长期以来一直是交互式人工智能和自然语言处理领域的核心挑战。随着深度学习的最新进展，研究这两个领域的结合以提升人机协作的兴趣日益增长。我们确定了三条具体方向，以改进基于语言的代理：

- Tool use and querying from knowledge bases. This direction emphasizes the importance of integrating external knowledge bases, web search, or other helpful tools into the reasoning processes of AI agents. By leveraging structured and unstructured data from various sources, agents can enhance their understanding and provide more accurate and context-aware responses. Furthermore, it fosters the agent's ability to proactively seek out information when faced with unfamiliar scenarios or queries, ensuring more comprehensive and informed responses. Examples include Toolformer (Schick et al., 2023) and Retrieve What You Need (Wang et al., 2023g).
- 工具使用与知识库查询。该方向强调将外部知识库、网络搜索或其他辅助工具整合进AI代理推理过程的重要性。通过利用来自多源的结构化和非结构化数据，代理能够增强理解力，提供更准确且具上下文感知的响应。此外，这促进了代理在面对未知场景或查询时主动寻求信息的能力，确保响应更全面且信息充足。示例包括Toolformer (Schick等, 2023) 和Retrieve What You Need (Wang等, 2023g)。
- Improved agent reasoning and planning. Enhancing the agent's ability to reason and plan is pivotal for effective human-agent collaboration. This involves the development of models that can understand complex instructions, infer user intentions, and predict potential future scenarios. This can be accomplished by asking the agent to reflect on past actions and failures as in ReAct (Yao et al., 2023a), or by structuring the agent thought process as a form of search (Yao et al., 2023b). By simulating different outcomes and assessing the ramifications of various actions, agents can make more informed context-aware decisions.
- 改进代理推理与规划。提升代理的推理和规划能力对于有效的人机协作至关重要。这涉及开发能够理解复杂指令、推断用户意图并预测潜在未来情景的模型。可通过让代理反思过去的行为和失败，如ReAct (Yao等, 2023a)，或将代理思维过程结构化为一种搜索形式 (Yao等, 2023b) 来实现。通过模拟不同结果并评估各种行动的后果，代理能够做出更具信息量且符合上下文的决策。
- Incorporating system and human feedback. AI agents can frequently operate in two primary contexts: environments that provide explicit signals about the effectiveness of their actions (system feedback), and settings where they collaborate with humans who can offer verbal critiques (human feedback). This direction underscores the need for adaptive learning mechanisms that allow agents to refine their strategies and rectify mistakes, such as in AutoGen (Wu et al., 2023). The ability to continuously learn and adapt from diverse feedback sources ensures that agents remain helpful and aligned for user needs.
- 融合系统与人类反馈。AI代理通常在两种主要环境中运行：一是提供其行为效果明确信号的系统反馈环境，二是与能提供口头批评的人类协作的环境。该方向强调需要自适应学习机制，使代理能够优化策略并纠正错误，如AutoGen (Wu等, 2023) 所示。持续从多样反馈源学习和适应的能力，确保代理始终对用户需求保持有用性和一致性。

22.10.3 6.6.2 General LLM agent

22.10.4 6.6.2 通用大型语言模型代理

Recognizing and understanding agent content and natural language has been a fundamental challenge in interactive AI and natural language processing for decades. With the recent advances in deep learning, there is a growing interest in studying these two areas jointly for deep understanding of both agent planning or human feedback for knowledge-inference and natural language generation. These are the key components of many human-machine-

interaction agents, such as "AutoGen"(Wu et al., 2023) and "Retrieve What You Need"(Wang et al., 2023g).

识别和理解代理内容及自然语言，长期以来一直是交互式人工智能和自然语言处理的核心挑战。随着深度学习的最新进展，研究这两个领域的结合以深入理解代理规划或人类反馈对于知识推理和自然语言生成的作用，兴趣日益增长。这些是许多机交互代理的关键组成部分，如“AutoGen”（Wu等，2023）和“Retrieve What You Need”（Wang等，2023g）。

General Alpaca Training Framework

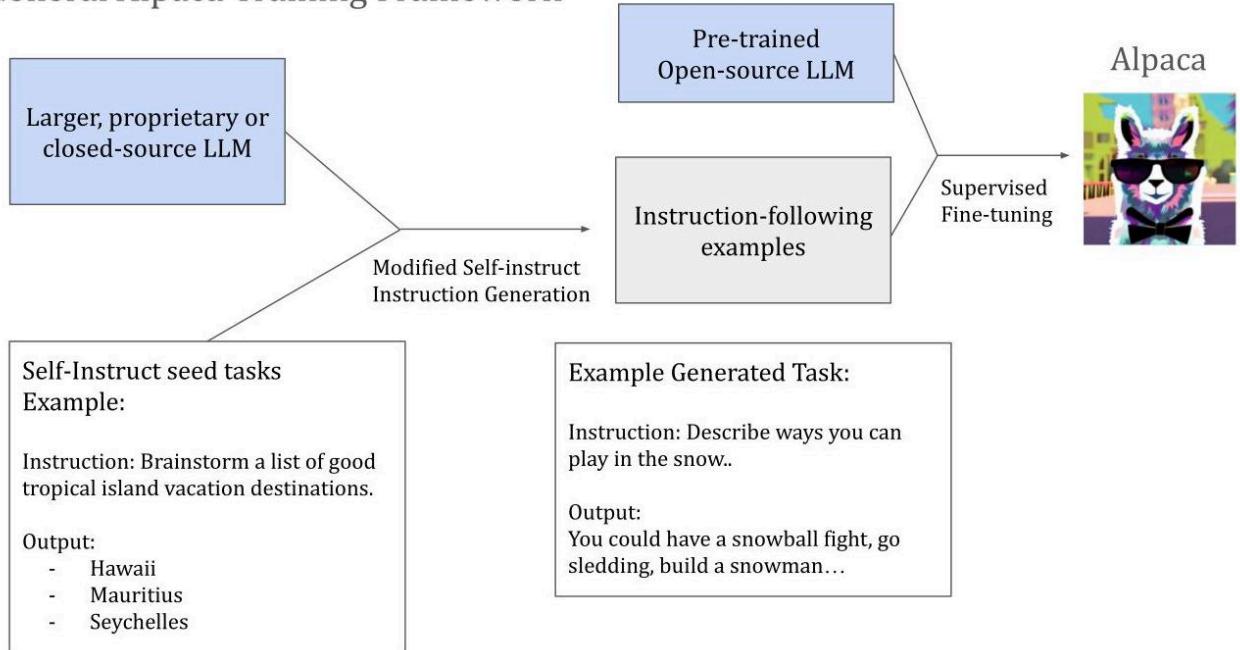


Figure 28: The training recipe used to train the Alpaca model (Taori et al., 2023). At a high level, existing LLMs are used to generate a large pool of instruction-following examples from a smaller set of seed tasks. The generated instruction-following examples are then used to instruction-tune an LLM where the underlying model weights are available.

图28：用于训练Alpaca模型（Taori等，2023）的训练方案。总体而言，现有大型语言模型用于从较小的种子任务集中生成大量遵循指令的示例。生成的遵循指令示例随后用于对可访问底层模型权重的大型语言模型进行指令微调。

22.10.5 6.6.3 Instruction-following LLM agents

22.10.6 6.6.3 遵循指令的大型语言模型代理

Furthermore, the creation of LLM Agents that can be trained to effectively follow human instructions has become an important area of research. Initial models used human feedback to train a proxy reward model to simulate human preferences, through a process known as Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022). This process produced models such as InstructGPT and ChatGPT. In order to more efficiently train instruction-following LLM agents without needing human labels, researchers developed a more efficient method for instruction-tuning that trains the LLM agent directly on instruction/response pairs, either generated by humans like Dolly 2.0⁶ or automatically from LLMs like Alpaca (Taori et al., 2023). We show the overall Alpaca training pipeline in Figure 28.

此外，能够训练出有效遵循人类指令的大型语言模型（LLM）代理的研究已成为一个重要领域。最初的模型通过人类反馈训练代理奖励模型以模拟人类偏好，这一过程称为带有人类反馈的强化学习（Reinforcement Learning with

Human Feedback, RLHF) (Ouyang 等, 2022)。该过程产生了如 InstructGPT 和 ChatGPT 等模型。为了更高效地训练遵循指令的 LLM 代理而无需人类标注, 研究人员开发了一种更高效的指令调优方法, 直接在由人类(如 Dolly 2.0⁶)或自动生成的 LLM(如 Alpaca (Taori 等, 2023))提供的指令/响应对上训练 LLM 代理。我们在图 28 中展示了整体的 Alpaca 训练流程。

22.10.7 6.6.4 Experiments and Results

22.10.8 6.6.4 实验与结果

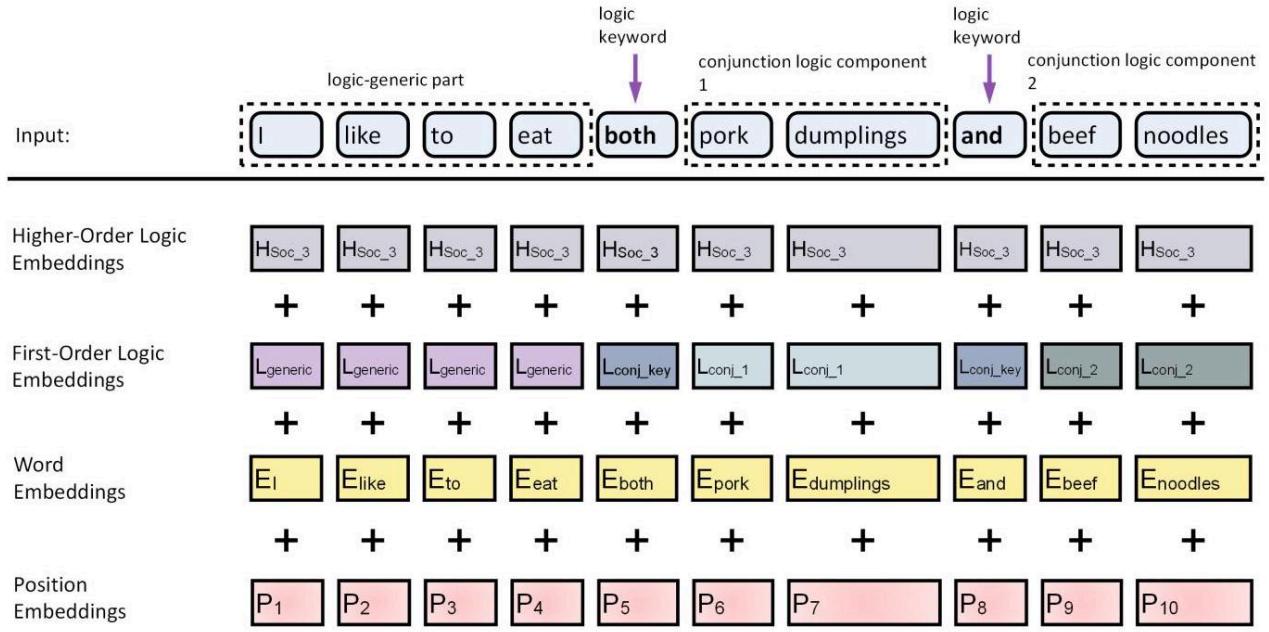
Despite the growing adoption of conversational and self-feedback systems, these forms of AI still do not perform well with regard to generating factually correct responses from their own implicit knowledge and therefore often use external tools like web search and knowledge retrieval mechanisms at inference-time to augment their response as a consequence. Addressing this would help create more engaging experiences for users in many real-life applications.

In social conversations (such as those on social media platforms like Instagram and Facebook), or with Q+A websites (such as Ask or Quora), people usually engage with others through a series of comments and by web-searching for information and knowledge relevant to the discussion. Thus, the task of generating conversational turns in this context is not to simply bootstrap upon traditional NLP models and tasks, but to use agents to generate dialogue through intelligent behaviors that reflect knowledge search and acquisition (Peng et al., 2023). In this way, intelligent agents for

尽管对话和自我反馈系统的应用日益广泛, 这些形式的人工智能在基于自身隐含知识生成事实正确的回答方面仍表现不佳, 因此常在推理时使用诸如网络搜索和知识检索等外部工具来增强回答。解决这一问题将有助于在许多现实应用中为用户创造更具吸引力的体验。在社交对话(如 Instagram 和 Facebook 等社交媒体平台)或问答网站(如 Ask 或 Quora)中, 人们通常通过一系列评论互动, 并通过网络搜索获取与讨论相关的信息和知识。因此, 在此背景下生成对话轮次的任务不仅仅是基于传统自然语言处理模型和任务的启动, 而是利用代理通过反映知识搜索与获取的智能行为来生成对话 (Peng 等, 2023)。通过这种方式, 智能代理用于

⁶ Dolly 2.0 blogpost link

⁶ Dolly 2.0 博客链接



These logic embeddings vectors are fully parameterized and are fully trainable during the fine-tuning training process on dialogue summarization datasets. The only restrictions that we put on their values are:

这些逻辑嵌入向量是完全参数化的，并且在对话摘要数据集的微调训练过程中完全可训练。我们对其数值施加的唯一限制是：

- Logic embeddings with same color (i.e. same logic role and type) must have the same vector values;
- 具有相同颜色（即相同逻辑角色和类型）的逻辑嵌入必须具有相同的向量值；
- Logic embeddings with different colors (i.e. different logic roles and types) can have different vector values.
- 具有不同颜色（即不同逻辑角色和类型）的逻辑嵌入可以具有不同的向量值。

Figure 29: The logic transformer agent model (Wang et al., 2023e). We integrate a logical reasoning module into the transformer-based abstractive summarization model in order to endow the logic agent the ability to reason over text and dialogue logic, so that it can generate better-quality abstractive summarizations and reduce factuality errors.

图 29：逻辑变换器代理模型（Wang 等，2023e）。我们将逻辑推理模块集成到基于变换器的抽象摘要模型中，以赋予逻辑代理对文本和对话逻辑进行推理的能力，从而生成更高质量的抽象摘要并减少事实性错误。

NLP tasks extends the task description and improves upon the interpretability of the response by adding an explicit knowledge search and retrieval step during dialogue. Incorporating these web search and retrieval agents as feedback during dialogue will help to engage further and deeper the social interactions between humans and agents (Wang et al., 2023e). As the Fig 29 showed, we introduced a new modeling paradigm for transformer language models that detects and extracts important logical structures and information from input texts and then integrates them into the input embeddings through carefully designed multi-layer hierarchical logical projections to infuse logical structures into pre-trained language models as one kind of NLP agent. (Wang et al., 2023e) propose a novel approach to construct logic-aware input embeddings for transformer language models through a combination of logic detection, logic mapping and hierarchical logical projections, and then develop a corresponding new modeling paradigm that can upgrade all existing transformer language models into logical transformers to consistently boost their performance. The proposed logical transformer agent consistently achieve superior performance over their baseline transformer models through a deeper understanding of the logical structures of texts. To human users, it is often these aspects that are more important for delivering a meaningful and interesting conversation via a agent-based coordination between dialogue and information retrieval. Delving deep into natural language processing, this

topic will discuss the advancements and leader-board in making LLMs more agentic and better suited for various language-centered tasks.

自然语言处理任务通过在对话中增加显式的知识搜索和检索步骤，扩展了任务描述并提升了回答的可解释性。在对话过程中将这些网络搜索和检索代理作为反馈纳入，将有助于加深人与代理之间的社交互动（Wang 等, 2023e）。如图 29 所示，我们引入了一种针对变换器语言模型的新建模范式，该范式检测并提取输入文本中的重要逻辑结构和信息，然后通过精心设计的多层次层级逻辑投影将其整合到输入嵌入中，以将逻辑结构注入预训练语言模型，作为一种自然语言处理代理。（Wang 等, 2023e）提出了一种通过逻辑检测、逻辑映射和层级逻辑投影相结合的方法，构建逻辑感知的输入嵌入，并开发了相应的新建模范式，能够将所有现有变换器语言模型升级为逻辑变换器，从而持续提升其性能。所提出的逻辑变换器代理通过对文本逻辑结构的更深理解，持续优于其基线变换器模型。对人类用户而言，正是这些方面对于通过基于代理的对话与信息检索协调，提供有意义且有趣的对话更为重要。深入自然语言处理领域，本主题将讨论使大型语言模型更具代理性并更适合各种语言中心任务的最新进展和排行榜。

An open-domain question answering (QA) system usually follows a retrieve-then-read paradigm, in which a retriever is used to retrieve relevant passages from a large corpus, and then a reader generates answers based on the retrieved passages and the original question. In (Wang et al., 2023g), we propose a simple and novel mutual learning framework to improve the performance of retrieve-then-read-style models via an intermediate module named the knowledge selector agent, which we train with reinforcement learning. The fine-grained knowledge selector into the retrieve-then-reader paradigm, whose goal is to construct a small subset of passages which retain question-relevant information. As showed in Figure 30, The knowledge selector agent is trained as a component of our novel mutual learning framework, which iteratively trains the knowledge selector and the reader. We adopt a simple and novel approach employing policy gradients to optimize the knowledge selector agent, using feedback from the reader to train it to select a small and informative set of passages. This approach avoids brute-force search or manually-designed heuristics, without requiring any annotated query-document pairs for supervision. We show that iteratively training the reader and the knowledge selector agent leads to better predictive performance on some public open-domain question answering benchmarks.

开放域问答 (QA) 系统通常遵循检索-阅读的范式，其中检索器用于从大规模语料库中检索相关段落，然后阅读器基于检索到的段落和原始问题生成答案。在 (Wang et al., 2023g) 中，我们提出了一种简单且新颖的互学习框架，通过一个名为知识选择代理 (knowledge selector agent) 的中间模块来提升检索-阅读模型的性能，该模块通过强化学习进行训练。细粒度的知识选择器被引入检索-阅读范式，其目标是构建一个保留与问题相关信息的小段落子集。如图30所示，知识选择代理作为我们新颖互学习框架的组成部分进行训练，该框架迭代训练知识选择器和阅读器。我们采用一种简单且新颖的方法，利用策略梯度优化知识选择代理，使用阅读器的反馈训练其选择一小组信息丰富的段落。该方法避免了暴力搜索或手工设计的启发式规则，也不需要任何带注释的查询-文档对作为监督。我们展示了迭代训练阅读器和知识选择代理能够在一些公开的开放域问答基准上获得更好的预测性能。

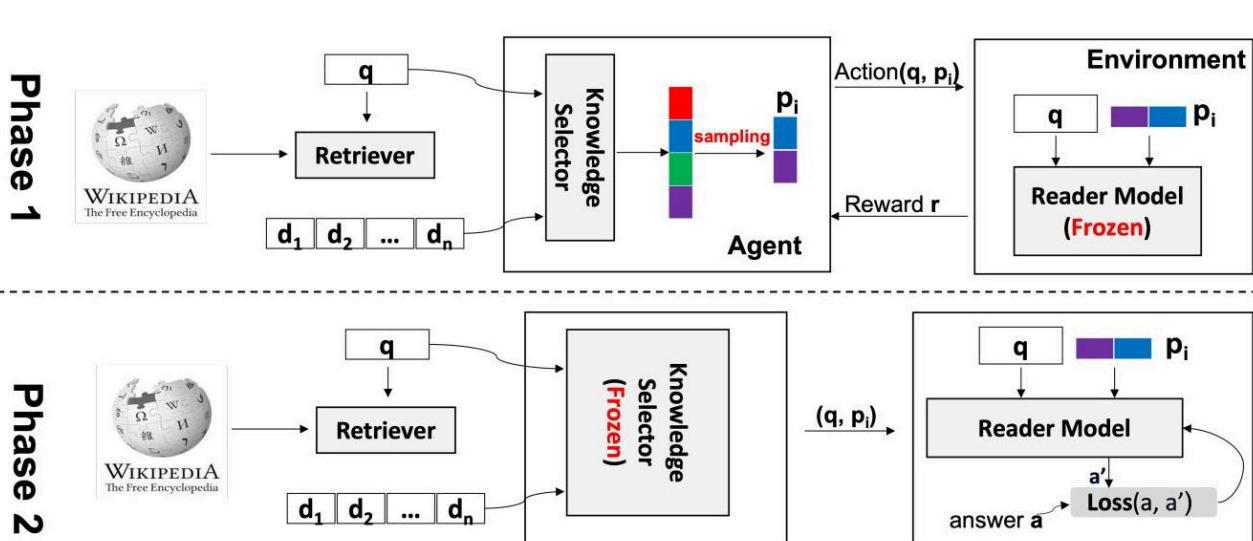


Figure 30: Architecture of one proposed NLP agent (Wang et al., 2023g) mutual learning framework. In each epoch, Phase 1 and Phase 2 are executed alternately. During Phase 1, the parameters of the reader model remain fixed, and only the weights of the knowledge selector are updated. Conversely, during Phase 2, the reader model's parameters are adjusted, while the knowledge selector's weights remain frozen.

图30：所提NLP代理（Wang et al., 2023g）互学习框架的架构。在每个训练周期中，阶段1和阶段2交替执行。阶段1期间，阅读模型的参数保持固定，仅更新知识选择器的权重。相反，阶段2期间，调整阅读模型的参数，而知识选择器的权重保持冻结。

23 7 Agent AI Across Modalities, Domains, and Realities

24 7 跨模态、跨领域与跨现实的智能代理

24.1 7.1 Agents for Cross-modal Understanding

24.2 7.1 跨模态理解的智能代理

Multi-modal understanding is a significant challenge for creating generalist AI agents due to the lack of large-scale datasets that contain vision, language, and agent behavior. More generally, training data for AI agents is often modality specific. This results in most modern multi-modal systems using a combination of frozen submodules. Some notable examples are Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023c), and LLaVA (Liu et al., 2023c), all of which utilize a frozen LLM and frozen visual encoder. These submodules are trained individually on separate datasets, and then adaptation layers are trained to encode the visual encoder into the LLM embedding space. In order to make further progress for cross-modal understanding for AI agents, it is likely that the strategy of using frozen LLMs and visual encoders will need to change. Indeed, RT-2, a recent visual-language model that is capable of taking actions within the domain of robotics showed significantly improved performance when jointly tuning the visual encoder and LLM for robotics and visual-language tasks (Brohan et al., 2023).

多模态理解是构建通用智能代理的一大挑战，原因在于缺乏包含视觉、语言和代理行为的大规模数据集。更普遍地说，AI代理的训练数据通常是特定模态的。这导致大多数现代多模态系统采用冻结子模块的组合。一些显著的例子包括Flamingo (Alayrac et al., 2022)、BLIP-2 (Li et al., 2023c) 和LLaVA (Liu et al., 2023c)，它们均使用冻结的大型语言模型 (LLM) 和冻结的视觉编码器。这些子模块分别在独立数据集上训练，然后训练适配层将视觉编码器映射到LLM的嵌入空间。为了在AI代理的跨模态理解上取得进一步进展，使用冻结LLM和视觉编码器的策略可能需要改变。事实上，RT-2是一种近期的视觉语言模型，能够在机器人领域执行动作，当对视觉编码器和LLM进行联合调优以适应机器人和视觉语言任务时，性能显著提升 (Brohan et al., 2023)。

24.3 7.2 Agents for Cross-domain Understanding

24.4 7.2 跨领域理解的智能代理

A key challenge for creating generalist agents is the distinctive visual appearance and disparate action spaces across different domains. Humans possess the capability to interpret images and videos from various sources, including the real world, video games, and specialized domains such as robotics and healthcare, once they become familiar with the specific details of these areas. However, existing LLMs and VLMs often demonstrate significant differences between the data they were trained on and the varied domains in which they are applied. And notably, training agent models to predict specific actions presents a considerable challenge when trying to develop a single policy that can effectively learn multiple control systems across domains. Generally, the approach most modern works take when applying systems within specific domains is to start from a pretrained foundation model and then finetune a separate model for each specific domain. This fails to capture any commonalities between domains and results in a smaller total set of data used for training instead of leveraging each domain's data.

构建通用智能代理的一个关键挑战是不同领域间视觉表现的差异和动作空间的多样性。人类具备解读来自不同来源的图像和视频的能力，包括现实世界、电子游戏以及机器人和医疗等专业领域，只要熟悉这些领域的具体细节。然而，现有的大型语言模型（LLM）和视觉语言模型（VLM）在训练数据与应用领域之间往往存在显著差异。尤其是，训练代理模型预测特定动作在尝试开发能够有效学习跨领域多控制系统的单一策略时面临巨大挑战。通常，现代工作在特定领域应用系统时，采用的做法是从预训练基础模型出发，然后为每个具体领域微调单独模型。这种方法未能捕捉领域间的共性，导致训练所用数据总量较小，未能充分利用各领域的数据。

24.5 7.3 Interactive agent for cross-modality and cross-reality

24.6 7.3 跨模态与跨现实的交互式智能代理

Developing AI agents that can successfully understand and perform tasks across different realities is an on-going challenge that has seen some recent success for image and scene generation (Huang et al., 2023a). In particular, it is challenging for agents to simultaneously understand real-world and virtual reality environments due to their visual dissimilarities and separate environment physics. Within the context of cross-reality, Sim to Real transfer is a particularly important problem when using simulation-trained policies for real-world data, which we discuss in the next section.

开发能够成功理解并执行跨不同现实任务的AI代理是一个持续的挑战，近期在图像和场景生成方面取得了一些进展 (Huang et al., 2023a)。特别是，由于视觉差异和环境物理的不同，代理同时理解现实世界和虚拟现实环境具有较大难度。在跨现实背景下，模拟到现实 (Sim to Real) 迁移是一个尤为重要的问题，涉及将模拟训练的策略应用于现实世界数据，我们将在下一节讨论该问题。

24.7 7.4 Sim to Real Transfer

24.8 7.4 模拟到现实的迁移

Techniques which enable models trained in simulation to be deployed in the real world. Embodied agents, especially one based on RL policies, are typically trained in simulated environments. These simulations do not fully replicate the characteristics of the real world (e.g., disturbances, light, gravity, and other physical properties). Due to this discrepancy between simulation and reality, models trained in simulation often struggle to perform well when applied in the real world. This issue is known as the "sim-to-real" problem. To solve this problem, several approaches can be taken:

使在仿真中训练的模型能够部署到现实世界的技术。具身智能体，尤其是基于强化学习（RL）策略的智能体，通常在仿真环境中训练。这些仿真无法完全复制现实世界的特性（例如，干扰、光照、重力及其他物理属性）。由于仿真与现实之间的差异，在仿真中训练的模型在应用于现实世界时往往表现不佳。这个问题被称为“仿真到现实”（sim-to-real）问题。为了解决该问题，可以采取以下几种方法：

- Domain randomization: domain randomization is a technique that trains a model while randomly varying parameters within a simulation environment (e.g., object appearance, sensor noise, and optical properties) in anticipation of the uncertainties and variations of the real world (Tobin et al., 2017). For instance, in the context of training a RL-based grasping skills, introducing randomness in the shapes of objects can lead to a policy capable of adapting to objects with somewhat different shapes (Saito et al., 2022).
- 域随机化：域随机化是一种在仿真环境中随机变化参数（例如，物体外观、传感器噪声和光学属性）来训练模型的技术，旨在应对现实世界中的不确定性和变化 (Tobin 等, 2017)。例如，在训练基于强化学习的抓取技能时，通过引入物体形状的随机性，可以获得能够适应形状略有不同物体的策略 (Saito 等, 2022)。
- Domain adaptation: Domain adaptation, or domain transfer is a technique that bridges the gap between simulated and real-world domains by training models with a large number of simulated images and a smaller set of real-world images. In practical settings, unpaired image-to-image translation methods such as CycleGAN (Zhu et al., 2017b) are employed due to the difficulty in preparing paired images across domains. Several enhanced versions exist for reinforcement learning, including RL-CycleGAN (Rao et al., 2020), and for imitation learning, such as RetinaGAN (Ho et al., 2021).

- 域适应：域适应或域迁移是一种通过使用大量仿真图像和少量现实图像训练模型，来弥合仿真域与现实域差距的技术。在实际应用中，由于难以准备跨域的配对图像，通常采用无配对图像到图像的转换方法，如CycleGAN（Zhu 等，2017b）。针对强化学习，有多种改进版本，如RL-CycleGAN（Rao 等，2020）；针对模仿学习，则有RetinaGAN（Ho 等，2021）。
- Improvement of simulation: Realistic simulation is a key for sim-to-real transfer. Part of this effort is achieved by a system identification techniques (Zhu et al., 2017c; Allevato et al., 2020), which aims to identify simulation parameters to mimic the real-world environments. Additionally, use of photorealistic simulators would be effective in image-based reinforcement learning (Martinez-Gonzalez et al., 2020; Müller et al., 2018; Shah et al., 2018; Sasabuchi et al., 2023).
- 仿真改进：逼真的仿真是实现仿真到现实迁移的关键。部分工作通过系统识别技术（Zhu 等，2017c; Allevato 等，2020）完成，旨在识别仿真参数以模拟现实环境。此外，使用光真实感仿真器在基于图像的强化学习中也非常有效（Martinez-Gonzalez 等，2020; Müller 等，2018; Shah 等，2018; Sasabuchi 等，2023）。

The sim-to-real transfer remains a central challenge in the study of Embodied Agents, as approaches keep evolving. Both theoretical and empirical research are essential to advance these technologies further.

仿真到现实的迁移仍然是具身智能体研究中的核心挑战，相关方法不断演进。理论与实证研究对于推动这些技术的进一步发展至关重要。

25 8 Continuous and Self-improvement for Agent AI

26 8 智能体AI的持续学习与自我提升

Currently, foundation model based AI agents have the capacity to learn from multiple different data sources, which allow for more flexible sources for data for training. Two key consequences of this are (1) user and human-based interaction data can be used to further refine and improve the agent and (2) existing foundation models and model artifacts can be used to generate training data. We discuss each of these in more detail in the following sections, but we note that since current AI Agents are largely tied to existing pretrained foundation models, they generally do not learn from continuous interaction with their environments. We think this is an exciting future direction, and initial work by Bousmalis et al. has shown that self-improving agents for robotic control are able to continuously learn and improve through environmental interactions without supervision (Bousmalis et al., 2023).

目前，基于基础模型的AI智能体具备从多种不同数据源学习的能力，这使得训练数据来源更加灵活。其两个关键影响是：（1）可以利用用户和人类交互数据进一步优化和提升智能体；（2）可以利用现有基础模型及其模型产物生成训练数据。以下章节将详细讨论这两点，但需要指出的是，由于当前AI智能体大多依赖于已有的预训练基础模型，它们通常不通过与环境的持续交互进行学习。我们认为这是一个令人兴奋的未来方向，Bousmalis 等人的初步工作表明，机器人控制的自我提升智能体能够通过环境交互在无监督条件下持续学习和改进（Bousmalis 等，2023）。

26.1 8.1 Human-based Interaction Data

26.2 8.1 基于人类的交互数据

The core idea behind using human-based interaction data is to leverage a large number of agent-human interactions to train and improve future iterations of the agent. There are several strategies used to improve agents from human-agent interactions.

利用基于人类的交互数据的核心思想是借助大量智能体与人类的交互，训练和改进未来版本的智能体。提升智能体的策略有多种。

- Additional training data Perhaps the simplest usage of human-agent interactions is to use the interaction examples themselves as training data for a future iteration of the agent. This generally requires filtering strategies to differentiate successful agent examples from unsuccessful interaction examples. Filtering can be rules-based (e.g., reaching some desired end goal state), model-based (e.g., classifying successful vs unsuccessful interactions), or manually selected after a posthoc inspection and/or modification of the interaction examples.

- 额外训练数据 人类与智能体的交互示例最简单的用法是将其作为未来智能体迭代的训练数据。这通常需要过滤策略，以区分成功的智能体示例和失败的交互示例。过滤可以基于规则（例如达到某个期望的终止状态）、基于模型（例如分类成功与失败的交互）或通过事后检查和/或修改交互示例后手动选择。
- Human preference learning During interaction with the user, the agent system can prompt the user with several different model outputs and allow for the user to select the best output. This is commonly used by LLMs like ChatGPT and GPT-4, whereby users can select one output (out of several) that aligns best with their preferences.
- 人类偏好学习 在与用户交互过程中，智能体系统可以向用户展示多个模型输出，允许用户选择最符合其偏好的输出。这种方法常见于大型语言模型（LLM）如ChatGPT和GPT-4，用户可以从多个输出中选出最符合其偏好的一个。
- Safety training (red-teaming) Red-teaming within the context of Agent AI refers to having a dedicated team of adversaries (either human or computer) that seek to exploit and expose weaknesses and vulnerabilities within the Agent AI system. Although adversarial in nature, red-teaming is commonly used as a means for understanding how to improve AI safety measures and reduce the occurrence of harmful outputs. The core principle is to discover consistent methods for inducing unwanted agent outputs so that the model can be trained on data that explicitly corrects this behavior.
- 安全训练（红队测试） 在智能体AI的语境中，红队测试指的是由专门的对抗团队（人类或计算机）试图利用并揭示智能体系统中的弱点和漏洞。尽管具有对抗性质，红队测试通常用于理解如何改进AI安全措施，减少有害输出的发生。其核心原则是发现诱发不良智能体输出的稳定方法，从而使模型能够基于明确纠正该行为的数据进行训练。

26.3 8.2 Foundation Model Generated Data

26.4 8.2 基础模型生成的数据

With the advent of powerful foundation model artifacts produced by academia and industry, there have been a variety of methods developed to extract and generate meaningful training data from these artifacts using a variety of prompting and data-pairing techniques.

随着学术界和工业界强大基础模型成果的出现，已经开发出多种方法，利用各种提示和数据配对技术，从这些成果中提取和生成有意义的训练数据。

- LLM Instruction-tuning Methods for generating instruction-following training data from LLMs have allowed for the finetuning of smaller, open-source models based on the outputs of larger proprietary LLMs (Wang et al., 2022b). For example, Alpaca (Taori et al., 2023) and Vicuna (Zheng et al., 2023) are LLMs based on the open-source LLaMA family (Touvron et al., 2023) that have been tuned on various outputs from ChatGPT and human participants. This method of instruction tuning can be viewed as a form of knowledge distillation, where the larger LLM serves as a teacher model to a smaller student model. Importantly, although LLM instruction-tuning has been shown to transfer the writing style and some instruction-following capabilities of the teacher model to the student model, significant gaps still exist between the factuality and capabilities of the teacher and student models (Gudibande et al., 2023).
- 大型语言模型（LLM）指令微调方法通过从大型专有LLM生成的指令遵循训练数据，实现了基于这些输出对较小的开源模型进行微调（Wang et al., 2022b）。例如，Alpaca (Taori et al., 2023) 和Vicuna (Zheng et al., 2023) 是基于开源LLaMA系列 (Touvron et al., 2023) 的LLM，通过对ChatGPT和人类参与者多样化输出的微调。该指令微调方法可视为一种知识蒸馏，其中大型LLM作为教师模型指导较小的学生模型。值得注意的是，尽管LLM指令微调已被证明能将教师模型的写作风格和部分指令遵循能力传递给学生模型，但教师与学生模型在事实准确性和能力上仍存在显著差距 (Gudibande et al., 2023)。
- Vision-language pairs A number of recent works have sought to increase the number of diversity of pretraining data available to visual-language models by automatically generating captions and other text for visual

content. For example, LLaVA (Liu et al., 2023c) uses 150,000 examples of instruction-following behavior from textual and visual inputs that are mainly LLM-generated. Other work has shown that using VLMs to re-caption images can improve the training data and subsequent quality of image generation models (Segalis et al., 2023). Within the realm of video understanding, using VLMs and LLMs to recaption videos has been shown to improve the performance and quality of subsequent VLMs trained on the recaptioned videos (Wang et al., 2023f; Zhao et al., 2022).

- 视觉-语言对近期多项研究致力于通过自动生成视觉内容的字幕及其他文本，增加视觉语言模型预训练数据的多样性。例如，LLaVA (Liu et al., 2023c) 利用15万个主要由LLM生成的文本与视觉输入的指令遵循示例。其他研究表明，使用视觉语言模型（VLM）重新为图像生成字幕，可以提升训练数据质量及后续图像生成模型的表现 (Segalis et al., 2023)。在视频理解领域，利用VLM和LLM为视频重新生成字幕，已被证明能提升基于这些重新字幕视频训练的VLM的性能和质量 (Wang et al., 2023f; Zhao et al., 2022)。

27 9 Agent Dataset and Leaderboard

28 9 代理数据集与排行榜

To accelerate research in this domain, we propose two benchmarks respectively for multi-agent gaming and agentic visual language tasks. We will release two new datasets - "CuisineWorld" and "VideoAnalytica" - and a set of baseline models, encouraging participants to explore new models, systems, and submit their results on the test set of our leaderboard.

为加速该领域研究，我们提出两个基准，分别针对多代理游戏和代理视觉语言任务。我们将发布两个新数据集——“CuisineWorld”和“VideoAnalytica”——以及一组基线模型，鼓励参与者探索新模型、新系统，并在我们的排行榜测试集上提交结果。

28.1 9.1 "CuisineWorld" Dataset for Multi-agent Gaming

28.2 9.1 面向多代理游戏的“CuisineWorld”数据集

CuisineWorld is a text-based game reminiscent of Overcooked! It offers a platform for AI-powered agents to cooperate and play in tandem. This dataset will test the collaboration efficiency of multi-agent systems, offering insights into how well LLMs and other systems can work together in dynamic scenarios. In particular, the dataset will focus on how well the agents understand goals, and how well the agents can coordinate among themselves. Two types of modes are supported in this dataset: a centralized dispatcher mode and a decentralized mode. Participants can choose a play mode and make a submission to our leaderboard.

CuisineWorld是一款类似《胡闹厨房！（Overcooked!）》的文本游戏，提供了一个AI代理协作并联动游戏的平台。该数据集旨在测试多代理系统的协作效率，深入了解大型语言模型（LLM）及其他系统在动态场景中的协同能力。特别地，数据集关注代理对目标的理解程度及代理间的协调能力。该数据集支持两种模式：集中式调度模式和去中心化模式。参与者可选择游戏模式并提交排行榜成绩。

28.2.1 9.1.1 Benchmark

28.2.2 9.1.1 基准

For our competition, we will release a benchmark, the CuisineWorld benchmark, which includes a text interface that includes extendable task definition files, and an interface for multi-agent interaction, and human-machine interactions. We introduce the gaming interaction task in which the goal is to generate relevant, appropriate, multi-agent collaboration strategies that can maximize collaboration efficiency. We evaluate the collaboration efficiency with the proposed evaluation metric: CoS.

本次竞赛将发布CuisineWorld基准，包含一个文本接口，支持可扩展的任务定义文件、多代理交互接口及人机交互。我们引入游戏交互任务，目标是生成相关且恰当的多代理协作策略，以最大化协作效率。协作效率通过提出的评估指标CoS进行评估。

The "CuisineWorld" dataset was collected by Microsoft, UCLA, and Stanford University. The goal of the competition is to explore how different, existing and novel, grounded-LLM and interactive techniques perform with this benchmark and establish strong baselines for the task of multi-agent gaming infrastructure.

“CuisineWorld”数据集由微软、加州大学洛杉矶分校（UCLA）和斯坦福大学联合收集。竞赛目标是探索现有及新颖的基于基础LLM的交互技术在该基准上的表现，并为多代理游戏基础设施任务建立强有力的基线。

The dataset of CuisineWorld includes:

CuisineWorld数据集包括：

- A selection of well-defined multi-agent collaboration tasks.
- 一系列定义明确的多代理协作任务。
- An API system to facilitate agent interactions.
- 便于代理交互的API系统。
- An automatic evaluation system.
- 自动评估系统。

(The link for downloading the dataset will soon be made available and this article will be updated to include it here.)

(数据集下载链接将很快提供，本文届时会更新包含该链接。)

28.2.3 9.1.2 Task

28.2.4 9.1.2 任务

- We provide a dataset and related the benchmark, called Microsoft MindAgent and and correspondingly release a dataset "CuisineWorld" to the to the research community.
- 我们向研究社区提供名为Microsoft MindAgent的基准及相应发布的“CuisineWorld”数据集。
- We will provide benchmarks to evaluate and rank the submitted "MindAgent" algorithms. We will also provide baseline results generated using popular infrastructures.
- 我们将提供基准测试以评估和排名提交的“MindAgent”算法。我们还将提供使用流行基础设施生成的基线结果。

28.2.5 9.1.3 Metrics and Judging

28.2.6 9.1.3 指标与评判

The quality of multi-agent collaboration efficiency is determined by the new "cos" auto-metric (from MindAgent (Gong et al., 2023a)). The final rating of out metric is calculated as an average over the evaluated collaboration efficiency metrics of the multi-agent system on all tasks. Human evaluators will be asked to rate individual responses as well as provide subjective judgement of the engagement, breadth and an overall quality of the users' interactions with the agents.

多智能体协作效率的质量由新的“cos”自动指标（来自MindAgent (Gong et al., 2023a)）决定。最终的指标评分是对多智能体系统在所有任务上评估的协作效率指标的平均值。人工评审员将被要求对单个响应进行评分，并对用户与智能体交互的参与度、广度及整体质量提供主观评价。

28.2.7 9.1.4 Evaluation

28.2.8 9.1.4 评估

- Automated Evaluation. We plan to release a leaderboard, starting on the release date (TBA), registered participants will be asked to submit their results on the task associated with the dataset "CuisineWorld" (our publicly released dataset for the leaderboard). Submission of results will be closed on the end date (TBA).

Each team will be required to submit their generated results on the testing set for automated evaluation of the "cos" metric.

- 自动评估。我们计划在发布日期（待定）启动排行榜，注册参与者需提交与数据集“CuisineWorld”（我们公开发布的排行榜数据集）相关任务的结果。结果提交截止日期为结束日期（待定）。每个团队需提交其在测试集上生成的结果，以便对“cos”指标进行自动评估。
- Human Evaluation on our leaderboard. The leaderboard participants will need to provide a submission file generated by evaluation scripts locally. We will use the evalAI system to check the submission file and optionally rerun the code for top challenge contenders. Therefore, teams must also submit their code with a Readme file on how to run their code. Human evaluation will be performed by the organization team.
- 排行榜上的人工评估。排行榜参与者需提供由本地评估脚本生成的提交文件。我们将使用evalAI系统检查提交文件，并可选择对顶级挑战者重新运行代码。因此，团队还必须提交其代码及包含运行说明的Readme文件。人工评估将由组织团队执行。
- Winner Announcement. We will make an announcement of the winners and post the final ratings of the submissions on our leaderboard.
- 获奖者公布。我们将公布获奖者名单，并在排行榜上发布提交结果的最终评分。

28.3 9.2 Audio-Video-Language Pre-training Dataset.

28.4 9.2 音视频语言预训练数据集

We introduce VideoAnalytica: a new benchmark for analytical video demonstration comprehension. VideoAnalytica focuses on leveraging video demonstrations as aids to better understand complex, high-level reasoning embedded within long-formed instructional videos. The objective is to evaluate the cognitive reasoning abilities of video language models, pushing them beyond mere recognition tasks and basic comprehension, towards a more sophisticated and nuanced understanding of videos. Crucially, VideoAnalytica emphasizes the integration of multiple modalities, such as audio, video, and language, as well as the ability of models to apply domain-specific knowledge, to contextualize and interpret the information presented in the videos. Specifically, VideoAnalytica involves two primary tasks:

我们介绍VideoAnalytica：一个用于分析视频演示理解的新基准。VideoAnalytica聚焦于利用视频演示辅助更好地理解长篇教学视频中蕴含的复杂高级推理。其目标是评估视频语言模型的认知推理能力，推动其超越简单识别和基础理解，达到对视频更复杂细致的理解。关键在于VideoAnalytica强调多模态融合，如音频、视频和语言，以及模型应用领域知识以语境化和解释视频信息的能力。具体而言，VideoAnalytica包含两个主要任务：

1. Video Text Retrieval: This task involves accurately retrieving relevant text from the instructional videos. The challenge lies in distinguishing between relevant and irrelevant information, thus requiring a deep understanding of the video content, and analysis of the demonstration to retrieve the correct query. To further increase the complexity of these tasks, we introduce hard negatives into our datasets generated by large language models. We run human validation on the generated negatives and remove instances that make the task invalid and unfair (e.g. negatives being valid).
2. 视频文本检索：该任务涉及准确检索教学视频中的相关文本。挑战在于区分相关与无关信息，因此需要对视频内容有深入理解，并分析演示以检索正确查询。为增加任务难度，我们在数据集中引入由大型语言模型生成的困难负样本。我们对生成的负样本进行人工验证，剔除使任务无效或不公平的实例（如负样本为有效内容）。
3. Video Assisted Informative Question Answering: This task requires the model to answer questions based on the information extracted from the videos. The focus is on complex questions that require analytical reasoning and a thorough comprehension of the video demonstration.
3. 视频辅助信息性问答：该任务要求模型基于视频中提取的信息回答问题。重点是需要分析推理和深入理解视频演示的复杂问题。

To facilitate the development of an audio-video-language agent for analytical video understanding, we introduce a benchmark leaderboard for the two tasks from VideoAnalytica.

为促进音视频语言智能体在分析视频理解方面的发展，我们为VideoAnalytica的两个任务引入了基准排行榜。

- The leaderboard participants will need to submit their solutions for evaluation. The evaluation will be based on the model's performance on the two tasks, and the results will be displayed on the leaderboard. Participants are required to submit their code, along with a detailed explanation of their approach and methodology.
- 排行榜参与者需提交其解决方案以供评估。评估将基于模型在两个任务上的表现，结果将在排行榜上展示。参与者需提交代码及其方法和策略的详细说明。
- Ethical considerations: The leaderboard focuses on understanding and interpreting video content, which could potentially be used in surveillance or other privacy-invasive applications. Therefore, it's crucial to consider the ethical implications and potential misuse of the technology. We encourage participants to consider these aspects in their submissions and promote the ethical use of AI.
- 伦理考量：排行榜聚焦于理解和解读视频内容，可能被用于监控或其他侵犯隐私的应用。因此，必须考虑技术的伦理影响及潜在滥用。我们鼓励参与者在提交中考虑这些方面，促进AI的伦理使用。

29 10 Broader Impact Statement

30 10 广泛影响声明

This article and our associated forum⁷ aim to be a catalyst for innovative research, fostering collaborations that will drive the next wave of AI applications. By focusing on multimodal agents, we emphasize the future direction of human-AI interactions, leader-board, and solutions. We detail three ways in which we make significant contributions to the broader community.

本文及其相关论坛⁷旨在成为创新研究的催化剂，促进推动下一波AI应用的合作。通过聚焦多模态智能体，我们强调人机交互、排行榜及解决方案的未来方向。我们详细阐述了三种对更广泛社区的重要贡献方式。

Firstly, we hope our forum grounds AI researchers to develop solutions motivated by real-world problems in gaming, robotics, healthcare, and long-video understanding. Specifically, the development of multimodal agents in gaming could lead to more immersive and personalized gaming experiences, thereby transforming the gaming industry. In robotics, the development of adaptive robotic systems could revolutionize industries ranging from manufacturing to agriculture, potentially addressing labor shortages and improving efficiency. In healthcare, the use of LLMs and VLMs as diagnostic agents or patient care assistants could lead to more accurate diagnoses, improved patient care, and increased accessibility to medical services, particularly in underserved areas. Furthermore, the ability of these models to interpret long-form videos could have far-reaching applications, from enhancing online learning to improving technical support services. In general, the topics covered in our forum will have significant downstream effects on a wide range of industries and humans across the world.

首先，我们希望我们的论坛能够让人工智能（AI）研究人员基于游戏、机器人技术、医疗保健和长视频理解等现实世界问题开发解决方案。具体来说，游戏中多模态智能体的发展可能带来更具沉浸感和个性化的游戏体验，从而改变游戏产业。在机器人领域，自适应机器人系统的发展有望革新从制造业到农业的多个行业，可能缓解劳动力短缺并提高效率。在医疗保健方面，利用大型语言模型（LLMs）和视觉语言模型（VLMs）作为诊断代理或患者护理助手，能够实现更准确的诊断、改善患者护理，并提升医疗服务的可及性，尤其是在服务不足的地区。此外，这些模型解读长视频的能力具有广泛应用前景，从增强在线学习到提升技术支持服务均有潜力。总体而言，我们论坛涵盖的主题将对全球众多行业和人类产生深远影响。

Secondly, we hope our forum stands as a valuable resource for AI practitioners and researchers alike, serving as a platform to explore and deeply comprehend the diverse and complex leader-board that come with implementing AI agents across a wide variety of environments and situations. This exploration includes, for instance, understanding the specific limitations and potential hazards linked to Agentic AI systems when they are developed for specialized sectors such as healthcare diagnostics. In this domain, issues like dangerous hallucinations in AI behavior can pose

significant risks, highlighting the critical need for meticulous design and testing. However, these specific leader-board may not be equally relevant or noticeable when considering AI agents crafted for the gaming industry. In such recreational fields, developers might instead prioritize tackling different hurdles, such as the need for AI to perform more open-ended generation and exhibit creativity, adapting dynamically to unpredictable gameplay scenarios and player interactions. By attending the forum, participants will gain insights into how these varied environments dictate the focus and direction of AI development, and how best to tailor AI solutions to meet these distinct needs and overcome the pertinent leader-board. Thirdly, the various elements of our event, including the expert presentations, informative posters, and notably the winners of our two leader-board, are set to offer a substantive yet succinct overview of the latest and significant trends, research directions, and innovative concepts in the realm of multimodal agents. These presentations will encapsulate pivotal findings and developments, shining a light on new systems, ideas, and technologies in the field of multimodal agent AI. This assortment of knowledge is not only beneficial for the attendees of our forum, who are looking to deepen their understanding and expertise in this domain, but it also serves as a dynamic and rich resource board. Those visiting our forum's website can tap into this reservoir of information to discover and understand the cutting-edge advancements and creative ideas steering the future of multimodal agent AI. We strive to serve as a useful knowledge base for both newcomers and veterans in the field. By engaging with these resources, we hope participants and online visitors alike can remain informed of the transformative changes and novel approaches that are shaping the exciting landscape surrounding multimodal agent AI.

其次，我们希望我们的论坛成为AI从业者和研究人员的重要资源平台，助力他们探索并深入理解在各种环境和情境中部署AI智能体所面临的多样且复杂的排行榜挑战。这种探索包括例如理解在医疗诊断等专业领域开发的智能体AI系统所特有的局限性和潜在风险。在该领域，AI行为中的危险幻觉问题可能带来重大风险，凸显了细致设计和测试的关键必要性。然而，当考虑为游戏产业打造的AI智能体时，这些特定的排行榜问题可能并不那么相关或显著。在此类娱乐领域，开发者可能更关注解决不同的难题，如AI需要进行更开放式的生成和展现创造力，动态适应不可预测的游戏场景和玩家互动。通过参与论坛，参会者将深入了解这些多样环境如何决定AI开发的重点和方向，以及如何最佳地定制AI解决方案以满足这些独特需求并克服相关排行榜挑战。第三，我们活动的各个环节，包括专家报告、信息海报，尤其是我们两个排行榜的获奖者，将提供一个内容丰富且简明扼要的多模态智能体领域最新重要趋势、研究方向和创新理念的概览。这些报告将总结关键发现和进展，聚焦多模态智能体AI领域的系统、新思想和新技术。这些知识不仅对希望深化该领域理解和专业技能的论坛参与者有益，也构成了一个动态且丰富的资源库。访问我们论坛网站的用户可以利用这一信息宝库，发现并理解引领多模态智能体AI未来的前沿进展和创新思路。我们致力于为该领域的新手和资深人士提供有价值的知识基础。通过利用这些资源，我们希望参与者和在线访客都能及时了解塑造多模态智能体AI激动人心前景的变革性变化和新颖方法。

⁷ <https://multimodalagentai.github.io>

⁷ <https://multimodalagentai.github.io>

31 11 Ethical Considerations

32 11 伦理考量

Multimodal Agent AI systems have many applications. In addition to interactive AI, grounded multimodal models could help drive content generation for bots and AI agents, and assist in productivity applications, helping to replay, paraphrase, action prediction or synthesize 3D or 2D scenario. Fundamental advances in agent AI help contribute towards these goals and many would benefit from a greater understanding of how to model embodied

and empathetic in a simulate reality or a real world. Arguably many of these applications could have positive benefits.

多模态智能体AI系统拥有广泛应用。除了交互式AI，基于现实的多模态模型还可推动机器人和AI智能体的内容生成，辅助生产力应用，帮助回放、释义、动作预测或合成3D或2D场景。智能体AI的基础性进展有助于实现这些目标，许多应用将受益于对如何在模拟现实或真实世界中建模具身性和共情能力的更深入理解。可以说，这些应用大多具有积极的效益。

However, this technology could also be used by bad actors. Agent AI systems that generate content can be used to manipulate or deceive people. Therefore, it is very important that this technology is developed in accordance with responsible AI guidelines. For example, explicitly communicating to users that content is generated by an AI system and providing the user with controls in order to customize such a system. It is possible the Agent AI could be used to develop new methods to detect manipulative content - partly because it is rich with hallucination performance of large foundation model - and thus help address another real world problem.

然而，该技术也可能被不良分子利用。生成内容的智能体AI系统可能被用来操纵或欺骗他人。因此，按照负责任的AI准则开发此技术极为重要。例如，明确告知用户内容由AI系统生成，并为用户提供定制该系统的控制选项。智能体AI也可能被用来开发检测操纵性内容的新方法——部分原因是大型基础模型丰富的幻觉表现——从而帮助解决另一个现实世界问题。

For examples, 1) in health topic, ethical deployment of LLM and VLM agents, especially in sensitive domains like healthcare, is paramount. AI agents trained on biased data could potentially worsen health disparities by providing inaccurate diagnoses for underrepresented groups. Moreover, the handling of sensitive patient data by AI agents raises significant privacy and confidentiality concerns. 2) In the gaming industry, AI agents could transform the role of developers, shifting their focus from scripting non-player characters to refining agent learning processes.

Similarly, adaptive robotic systems could redefine manufacturing roles, necessitating new skill sets rather than replacing human workers. Navigating these transitions responsibly is vital to minimize potential socio-economic disruptions.

例如，1）在健康领域，尤其是在医疗等敏感领域，LLM和VLM智能体的伦理部署至关重要。基于偏见数据训练的AI智能体可能通过为代表性不足群体提供不准确诊断而加剧健康差距。此外，AI智能体处理敏感患者数据也引发重大隐私和保密性问题。2）在游戏产业，AI智能体可能改变开发者的角色，使其从编写非玩家角色脚本转向优化智能体学习过程。同样，自适应机器人系统可能重新定义制造业岗位，要求新的技能而非替代人类工人。负责任地应对这些转变对于最大限度减少潜在的社会经济冲击至关重要。

Furthermore, the agent AI focuses on learning collaboration policy in simulation and there is some risk if directly applying the policy to the real world due to the distribution shift. Robust testing and continual safety monitoring mechanisms should be put in place to minimize risks of unpredictable behaviors in real-world scenarios. Our "VideoAnalytica" dataset is collected from the Internet and considering which is not a fully representative source, so we already go through-ed the ethical review and legal process from both Microsoft and University Washington. Be that as it may, we also need to understand biases that might exist in this corpus. Data distributions can be characterized in many ways. In this workshop, we have captured how the agent level distribution in our dataset is different from other existing datasets. However, there is much more than could be included in a single dataset or workshop. We would argue that there is a need for more approaches or discussion linked to real tasks or topics and that by making these data or system available.

此外，智能体AI专注于在仿真环境中学习协作策略，直接将该策略应用于现实世界存在一定风险，原因在于分布偏移。应建立稳健的测试和持续的安全监控机制，以最大限度地减少现实场景中不可预测行为的风险。我们的“VideoAnalytica”数据集来源于互联网，考虑到其并非完全具有代表性的来源，我们已通过微软和华盛顿大学的伦理审查和法律程序。尽管如此，我们仍需理解该语料库中可能存在的偏见。数据分布可以通过多种方式进行表征。在本次研讨会上，我们捕捉到了数据集中智能体层级分布与其他现有数据集的差异。然而，一个数据集或研讨会所涵盖的内容远不止于此。我们认为，需要更多与实际任务或主题相关的方法或讨论，并通过开放这些数据或系统来实现。

We will dedicate a segment of our project to discussing these ethical issues, exploring potential mitigation strategies, and deploying a responsible multi-modal AI agent. We hope to help more researchers answer these questions together via this paper.

我们将专门安排项目的一部分来讨论这些伦理问题，探索潜在的缓解策略，并部署负责任的多模态AI智能体。我们希望通过这篇论文帮助更多研究人员共同回答这些问题。

33 12 Diversity Statement

34 12 多样性声明

By examining the adaptability of AI agent models in various domains, we inherently embrace a diversity of leader-board, perspectives, and solutions. In this vein, our project aims to build a diverse community by exploring the wide array of subjects in multimodal and agentic AI. With these principles in mind, this project focuses on advanced multimodal systems that interact effectively within both physical and virtual environments and facilitate effective interaction with humans. As such, we intend to engage a broad range of experts and practitioners across a wide-range of technical specialities, cultures, countries, and scholarly fields to discuss important topics, including but not limited to:

通过考察AI智能体模型在各个领域的适应性，我们本质上拥抱了排行榜、观点和解决方案的多样性。基于此，我们的项目旨在通过探索多模态和智能体AI的广泛主题，构建一个多元化的社区。秉持这些原则，本项目聚焦于先进的多模态系统，这些系统能够在物理和虚拟环境中有效交互，并促进与人类的有效互动。因此，我们计划邀请来自广泛技术专长、文化、国家和学术领域的专家和从业者，共同讨论重要议题，包括但不限于：

- Application of foundation models: the development of agents with integrated modalities (audio, image, text, sensor inputs), aiming to enhance their recognition and response capabilities for a wide variety of applications.
• 基础模型的应用：开发集成多模态（音频、图像、文本、传感器输入）的智能体，旨在提升其识别和响应多样化应用的能力。
- General-purpose end-to-end systems: the development of end-to-end models that are trained with large-scale data, seeking to create versatile and adaptable AI solutions.
• 通用端到端系统：开发基于大规模数据训练的端到端模型，力求打造多功能且适应性强的AI解决方案。
- Methodologies for grounding modalities: integrating information across various modalities, enhancing the coherence and efficacy of data processing.
• 多模态基础方法：整合多种模态信息，提升数据处理的连贯性和效率。
- Intuitive human interface: the development of effective and meaningful interaction between humans and agents.
• 直观的人机界面：开发人与智能体之间有效且有意义的交互方式。
- Taming LLM/VLMs: exploring new approaches to address common issues in large-scale models, such as hallucinations and biases in their outputs.
• 驯服大型语言模型/视觉语言模型（LLM/VLMs）：探索解决大型模型常见问题的新方法，如幻觉和输出偏见。

We aspire to broaden our collective understanding of the potential and limitations of agentic AI by leveraging our unique and diverse perspectives. We strongly believe that this approach will not only enrich individual perspectives, but will also enhance the community's collective knowledge and promote a holistic view that is more inclusive of the wide-ranging leader-board faced by multimodal AI agents. References

我们希望通过利用独特且多样的视角，拓宽对智能体AI潜力与局限的集体理解。我们坚信，这种方法不仅能丰富个体视角，还将增强社区的集体知识，促进更包容多模态AI智能体所面临广泛排行榜的整体视野。参考文献

- M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do as i can and not as i say: Grounding language in robotic affordances," in arXiv preprint arXiv:2204.01691, 2022.
- M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do as i can and not as i say: Grounding language in robotic affordances," in arXiv preprint arXiv:2204.01691, 2022.
- M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog et al., "Do as i can, not as i say: Grounding language in robotic affordances," arXiv preprint arXiv:2204.01691, 2022.
- M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog et al., "Do as i can, not as i say: Grounding language in robotic affordances," arXiv preprint arXiv:2204.01691, 2022.
- J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds et al., "Flamingo: a visual language model for few-shot learning," Advances in Neural Information Processing Systems, vol. 35, pp. 23716-23736, 2022.
- J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds 等, “Flamingo: 一种用于少样本学习的视觉语言模型”, 《神经信息处理系统进展》(Advances in Neural Information Processing Systems), 第35卷, 第23716-23736页, 2022年。
- A. Allevato, E. S. Short, M. Pryor, and A. Thomaz, "Tunenet: One-shot residual tuning for system identification and sim-to-real robot task transfer," in Conference on Robot Learning. PMLR, 2020, pp. 445-455.
- A. Allevato, E. S. Short, M. Pryor, 和 A. Thomaz, “Tunenet：用于系统识别和仿真到现实机器人任务迁移的一次残差调优”，发表于机器人学习会议(Conference on Robot Learning)，PMLR, 2020年，第445-455页。
- A. Amjad, P. Kordel, and G. Fernandes, "A review on innovation in healthcare sector (telehealth) through artificial intelligence," Sustainability, vol. 15, no. 8, p. 6655, 2023.
- A. Amjad, P. Kordel, 和 G. Fernandes, “通过人工智能推动医疗保健领域（远程医疗）的创新综述”, 《可持续性》(Sustainability), 第15卷, 第8期, 6655页, 2023年。
- S. An, Z. Lin, Q. Fu, B. Chen, N. Zheng, J.-G. Lou, and D. Zhang, "How do in-context examples affect compositional generalization?" arXiv preprint arXiv:2305.04835, 2023.
- S. An, Z. Lin, Q. Fu, B. Chen, N. Zheng, J.-G. Lou, 和 D. Zhang, “上下文示例如何影响组合泛化? ”, arXiv预印本 arXiv:2305.04835, 2023年。
- P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva et al., "On evaluation of embodied navigation agents," arXiv preprint arXiv:1807.06757, 2018.
- P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva 等, “关于具身导航代理评估的研究”, arXiv预印本 arXiv:1807.06757, 2018年。
- P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3674-3683.
- P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, 和 A. Van Den Hengel,

“视觉与语言导航：在真实环境中解释视觉基础的导航指令”，发表于IEEE计算机视觉与模式识别会议论文集，2018年，第3674-3683页。

S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425-2433.

S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, 和 D. Parikh, “视觉问答（VQA）”，发表于IEEE国际计算机视觉会议论文集，2015年，第2425-2433页。

M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1728-1738.

M. Bain, A. Nagrani, G. Varol, 和 A. Zisserman, “Frozen in time: 一种用于端到端检索的联合视频与图像编码器”，发表于IEEE/CVF国际计算机视觉会议论文集，2021年，第1728-1738页。

B. Baker, I. Akkaya, P. Zhokov, J. Huizinga, J. Tang, A. Ecoffet, B. Houghton, R. Sampedro, and J. Clune, "Video pretraining (vpt): Learning to act by watching unlabeled online videos," Advances in Neural Information Processing Systems, vol. 35, pp. 24639-24654, 2022.

B. Baker, I. Akkaya, P. Zhokov, J. Huizinga, J. Tang, A. Ecoffet, B. Houghton, R. Sampedro, 和 J. Clune, “视频预训练（VPT）：通过观看无标签在线视频学习行动”，《神经信息处理系统进展》（Advances in Neural Information Processing Systems），第35卷，第24639-24654页，2022年。

D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, "Objectnav revisited: On evaluation of embodied agents navigating to objects," arXiv preprint arXiv:2006.13171, 2020.

D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, 和 E. Wijmans, “ObjectNav再探：关于具身代理导航至目标物体的评估”，arXiv预印本 arXiv:2006.13171, 2020年。

K. Black, M. Nakamoto, P. Atreya, H. Walke, C. Finn, A. Kumar, and S. Levine, "Zero-shot robotic manipulation with pretrained image-editing diffusion models," arXiv preprint arXiv:2310.10639, 2023.

K. Black, M. Nakamoto, P. Atreya, H. Walke, C. Finn, A. Kumar, 和 S. Levine, “利用预训练图像编辑扩散模型实现零样本机器人操作”，arXiv预印本 arXiv:2310.10639, 2023年。

A. Blair-Stanek, N. Holzenberger, and B. Van Durme, "Can gpt-3 perform statutory reasoning?" arXiv preprint arXiv:2302.06100, 2023.

A. Blair-Stanek, N. Holzenberger, 和 B. Van Durme, “GPT-3能否执行法定推理？”，arXiv预印本 arXiv:2302.06100, 2023年。

K. Bousmalis, G. Vezzani, D. Rao, C. Devin, A. X. Lee, M. Bauza, T. Davchev, Y. Zhou, A. Gupta, A. Raju et al., "Robocat: A self-improving foundation agent for robotic manipulation," arXiv preprint arXiv:2306.11706, 2023.

K. Bousmalis, G. Vezzani, D. Rao, C. Devin, A. X. Lee, M. Bauza, T. Davchev, Y. Zhou, A. Gupta, A. Raju 等, “RoboCat：一种自我改进的机器人操作基础代理”，arXiv预印本 arXiv:2306.11706, 2023年。

A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu et al., "Rt-1: Robotics transformer for real-world control at scale," arXiv preprint arXiv:2212.06817, 2022.

A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu 等, “RT-1：面向大规模现实世界控制的机器人变换器（Robotics Transformer）”，arXiv预印本 arXiv:2212.06817, 2022年。

A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn et al., "Rt-2: Vision-language-action models transfer web knowledge to robotic control," arXiv preprint arXiv:2307.15818, 2023.

A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn 等, “RT-2：视觉-语言-动作模型将网络知识迁移至机器人控制”，arXiv预印本 arXiv:2307.15818, 2023年。

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners," Advances in neural information processing systems, vol. 33, pp. 1877-1901, 2020.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell 等, “语言模型是少样本学习者”, 《神经信息处理系统进展》(Advances in Neural Information Processing Systems), 第33卷, 第1877-1901页, 2020年。

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg et al., "Sparks of artificial general intelligence: Early experiments with gpt-4," arXiv preprint arXiv:2303.12712, 2023.

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg 等, “通用人工智能的火花: GPT-4的早期实验”, arXiv预印本 arXiv:2303.12712, 2023年。

W. Cai, S. Huang, G. Cheng, Y. Long, P. Gao, C. Sun, and H. Dong, "Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill," arXiv preprint arXiv:2309.10309, 2023.

W. Cai, S. Huang, G. Cheng, Y. Long, P. Gao, C. Sun, 和 H. Dong, “通过像素引导的导航技能桥接零样本目标导航与基础模型”, arXiv预印本 arXiv:2309.10309, 2023年。

M. Carroll, R. Shah, M. K. Ho, T. Griffiths, S. Seshia, P. Abbeel, and A. Dragan, "On the utility of learning about humans for human-ai coordination," Advances in neural information processing systems, vol. 32, 2019.

M. Carroll, R. Shah, M. K. Ho, T. Griffiths, S. Seshia, P. Abbeel, 和 A. Dragan, “关于学习人类以促进人机协作的效用”, 《神经信息处理系统进展》, 第32卷, 2019年。

Y. Chang, M. Narang, H. Suzuki, G. Cao, J. Gao, and Y. Bisk, "WebQA: Multihop and Multimodal QA," arXiv preprint arXiv:2109.00590, 2021.

Y. Chang, M. Narang, H. Suzuki, G. Cao, J. Gao, 和 Y. Bisk, “WebQA: 多跳多模态问答”, arXiv预印本 arXiv:2109.00590, 2021年。

D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," Advances in Neural Information Processing Systems, vol. 33, pp. 4247-4258, 2020.

D. S. Chaplot, D. P. Gandhi, A. Gupta, 和 R. R. Salakhutdinov, “基于目标导向语义探索的目标导航”, 《神经信息处理系统进展》, 第33卷, 第4247-4258页, 2020年。

D. S. Chaplot, R. Salakhutdinov, A. Gupta, and S. Gupta, "Neural topological slam for visual navigation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12875-12884.

D. S. Chaplot, R. Salakhutdinov, A. Gupta, 和 S. Gupta, “用于视觉导航的神经拓扑SLAM”, 发表于IEEE/CVF计算机视觉与模式识别会议论文集, 2020年, 第12875-12884页。

G. Chen, Y.-D. Zheng, J. Wang, J. Xu, Y. Huang, J. Pan, Y. Wang, Y. Wang, Y. Qiao, T. Lu, and L. Wang, "Videollm: Modeling video sequence with large language models," 2023.

G. Chen, Y.-D. Zheng, J. Wang, J. Xu, Y. Huang, J. Pan, Y. Wang, Y. Wang, Y. Qiao, T. Lu, 和 L. Wang, “VideoLLM: 利用大型语言模型建模视频序列”, 2023年。

K. Chen, Q. Huang, H. Palangi, P. Smolensky, K. D. Forbus, and J. Gao, "Mapping natural-language problems to formal-language solutions using structured neural representations," in ICML 2020, July 2020.

K. Chen, Q. Huang, H. Palangi, P. Smolensky, K. D. Forbus, 和 J. Gao, “使用结构化神经表示将自然语言问题映射为形式语言解决方案”, ICML 2020, 2020年7月。

K. Chen, Q. Huang, D. McDuff, X. Gao, H. Palangi, J. Wang, K. Forbus, and J. Gao, "Nice: Neural image commenting with empathy," in EMNLP 2021, October 2021. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/nice-neural-image-commenting-with-empathy/>

K. Chen, Q. Huang, D. McDuff, X. Gao, H. Palangi, J. Wang, K. Forbus, 和 J. Gao, “NICE: 具有共情能力的神经

图像评论”，EMNLP 2021，2021年10月。[在线]。可访问：<https://www.microsoft.com/en-us/research/publication/nice-neural-image-commenting-with-empathy/>

J. H. Choi, K. E. Hickman, A. Monahan, and D. Schwarcz, "Chatgpt goes to law school," Available at SSRN, 2023.
J. H. Choi, K. E. Hickman, A. Monahan, 和 D. Schwarcz, “ChatGPT进军法学院”，SSRN可用，2023年。

H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma et al., "Scaling instruction-finetuned language models," arXiv preprint arXiv:2210.11416, 2022.

H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma 等，“扩展指令微调语言模型”，arXiv预印本 arXiv:2210.11416, 2022年。

N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 168-172.

N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, 和 A. Halpern, “皮肤病变分析以检测黑色素瘤：2017年国际生物医学成像研讨会（ISBI）上的挑战，由国际皮肤成像合作组织（ISIC）主办，”发表于2018年IEEE第15届国际生物医学成像研讨会（ISBI 2018），2018年，第168-172页。

A. Creswell, M. Shanahan, and I. Higgins, "Selection-inference: Exploiting large language models for interpretable logical reasoning," arXiv preprint arXiv:2205.09712, 2022.

A. Creswell, M. Shanahan, 和 I. Higgins, “选择-推理：利用大型语言模型进行可解释的逻辑推理，”arXiv预印本 arXiv:2205.09712, 2022年。

B. Cui, A. Lupu, S. Sokota, H. Hu, D. J. Wu, and J. N. Foerster, "Adversarial diversity in hanabi," in The Eleventh International Conference on Learning Representations, 2023. [Online]. Available: https://openreview.net/forum?id=uLE3WF3-H_5

B. Cui, A. Lupu, S. Sokota, H. Hu, D. J. Wu, 和 J. N. Foerster, “Hanabi中的对抗多样性，”发表于第十一届国际学习表征会议，2023年。[在线]。可访问：https://openreview.net/forum?id=uLE3WF3-H_5

G. Dagan, F. Keller, and A. Lascarides, "Dynamic planning with a llm," arXiv preprint arXiv:2308.06391, 2023.

G. Dagan, F. Keller, 和 A. Lascarides, “基于大型语言模型的动态规划，”arXiv预印本 arXiv:2308.06391, 2023年。

W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," 2023.

W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, 和 S. Hoi, “InstructBLIP：面向通用视觉-语言模型的指令调优，”2023年。

A. d'Avila Garcez and L. C. Lamb, "Neurosymbolic ai: The 3rd wave," 2020.

A. d'Avila Garcez 和 L. C. Lamb, “神经符号人工智能：第三波，”2020年。

M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M. Wallingford et al., "Robothor: An open simulation-to-real embodied ai platform," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3164-3174.

M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M. Wallingford 等，“RoboTHOR：一个开放的仿真到现实的具身人工智能平台，”发表于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第3164-3174页。

Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey for in-context learning," arXiv preprint arXiv:2301.00234, 2022.

Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, 和 Z. Sui, “上下文学习综述，”arXiv预印本 arXiv:2301.00234, 2022年。

V. S. Dorbala, G. Sigurdsson, R. Piramuthu, J. Thomason, and G. S. Sukhatme, "Clip-nav: Using clip for zero-shot vision-and-language navigation," arXiv preprint arXiv:2211.16649, 2022.

V. S. Dorbala, G. Sigurdsson, R. Piramuthu, J. Thomason, 和 G. S. Sukhatme, “CLIP-NAV：利用CLIP进行零样本视觉与语言导航，”arXiv预印本 arXiv:2211.16649, 2022年。

V. S. Dorbala, J. F. Mullen Jr, and D. Manocha, "Can an embodied agent find your" cat-shaped mug"? Ilm-based zero-shot object navigation," arXiv preprint arXiv:2303.03480, 2023.

V. S. Dorbala, J. F. Mullen Jr, 和 D. Manocha, “具身代理能找到你的‘猫形杯’吗？基于ILM的零样本目标导航，” arXiv预印本 arXiv:2303.03480, 2023年。

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," ICLR, 2021.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly 等, “一张图像相当于16x16个词：大规模图像识别的Transformer，”ICLR, 2021年。

D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu et al., "Palm-e: An embodied multimodal language model," arXiv preprint arXiv:2303.03378, 2023.

D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu 等, “PaLM-E：一个具身多模态语言模型，”arXiv预印本 arXiv:2303.03378, 2023年。

Y. Du, M. Yang, P. Florence, F. Xia, A. Wahid, B. Ichter, P. Sermanet, T. Yu, P. Abbeel, J. B. Tenenbaum et al., "Video language planning," arXiv preprint arXiv:2310.10625, 2023.

Y. Du, M. Yang, P. Florence, F. Xia, A. Wahid, B. Ichter, P. Sermanet, T. Yu, P. Abbeel, J. B. Tenenbaum 等, “视频语言规划，”arXiv预印本 arXiv:2310.10625, 2023年。

N. Dziri, A. Madotto, O. Zaiane, and A. J. Bose, "Neural path hunter: Reducing hallucination in dialogue systems via path grounding," arXiv preprint arXiv:2104.08455, 2021.

N. Dziri, A. Madotto, O. Zaiane, 和 A. J. Bose, “神经路径猎手：通过路径定位减少对话系统中的幻觉，”arXiv预印本 arXiv:2104.08455, 2021年。

K. Ehsani, W. Han, A. Herrasti, E. VanderBilt, L. Weihs, E. Kolve, A. Kembhavi, and R. Mottaghi, "Manipulathor: A framework for visual object manipulation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 4497-4506.

K. Ehsani, W. Han, A. Herrasti, E. VanderBilt, L. Weihs, E. Kolve, A. Kembhavi, 和 R. Mottaghi, “Manipulathor：一个用于视觉物体操作的框架，”发表于IEEE/CVF计算机视觉与模式识别会议论文集，2021年，第4497-4506页。

D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, and T. Darrell, "Speaker-follower models for vision-and-language navigation," in Advances in Neural Information Processing Systems (NIPS), 2018.

D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, 和 T. Darrell, “面向视觉与语言导航的说话者-跟随者模型，”发表于神经信息处理系统进展（NIPS），2018年。

T.-J. Fu, L. Li, Z. Gan, K. Lin, W. Y. Wang, L. Wang, and Z. Liu, "Violet : End-to-end video-language transformers with masked visual-token modeling," 2022.

T.-J. Fu, L. Li, Z. Gan, K. Lin, W. Y. Wang, L. Wang, 和 Z. Liu, “Violet：基于掩码视觉标记建模的端到端视频-语言变换器，”2022年。

—, "An empirical study of end-to-end video-language transformers with masked visual modeling," 2023.

—, “基于掩码视觉建模的端到端视频-语言变换器的实证研究，”2023年。

S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song, "Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23 171-23 181.

S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, 和 S. Song, “牧场上的奶牛：基于语言驱动的零样本对象导航的基线与基准,” 载于《IEEE/CVF计算机视觉与模式识别会议论文集》, 2023年, 第23171-23181页。

J. Gao, B. Peng, C. Li, J. Li, S. Shayandeh, L. Liden, and H.-Y. Shum, "Robust conversational ai with grounded text generation," arXiv preprint arXiv:2009.03457, 2020.

J. Gao, B. Peng, C. Li, J. Li, S. Shayandeh, L. Liden, 和 H.-Y. Shum, “具有基础文本生成的鲁棒对话式人工智能,” arXiv预印本 arXiv:2009.03457, 2020年。

J. Gao, C. Xiong, P. Bennett, and N. Craswell, "Neural approaches to conversational information retrieval," arXiv preprint arXiv:2201.05176, 2022.

J. Gao, C. Xiong, P. Bennett, 和 N. Craswell, “面向对话信息检索的神经方法,” arXiv预印本 arXiv:2201.05176, 2022年。

C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, and T. Lozano-Pérez, "Integrated task and motion planning," Annual review of control, robotics, and autonomous systems, vol. 4, pp. 265-293, 2021.

C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, 和 T. Lozano-Pérez, “集成任务与运动规划,” 《控制、机器人与自主系统年评》, 第4卷, 第265-293页, 2021年。

T. Gervet, S. Chintala, D. Batra, J. Malik, and D. S. Chaplot, "Navigating to objects in the real world," Science Robotics, vol. 8, no. 79, p. eadf6991, 2023.

T. Gervet, S. Chintala, D. Batra, J. Malik, 和 D. S. Chaplot, “现实世界中的对象导航,” 《科学机器人》, 第8卷, 第79期, 文章编号 eadf6991, 2023年。

R. Gong, J. Huang, Y. Zhao, H. Geng, X. Gao, Q. Wu, W. Ai, Z. Zhou, D. Terzopoulos, S.-C. Zhu et al., "Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

R. Gong, J. Huang, Y. Zhao, H. Geng, X. Gao, Q. Wu, W. Ai, Z. Zhou, D. Terzopoulos, S.-C. Zhu 等, “Arnold: 基于语言的任务学习基准, 涵盖现实3D场景中的连续状态,” 载于《IEEE/CVF国际计算机视觉大会 (ICCV) 论文集》, 2023年。

R. Gong, Q. Huang, X. Ma, H. Vo, Z. Durante, Y. Noda, Z. Zheng, S.-C. Zhu, D. Terzopoulos, L. Fei-Fei et al., "Mindagent: Emergent gaming interaction," arXiv preprint arXiv:2309.09971, 2023.

R. Gong, Q. Huang, X. Ma, H. Vo, Z. Durante, Y. Noda, Z. Zheng, S.-C. Zhu, D. Terzopoulos, L. Fei-Fei 等, “Mindagent: 新兴的游戏交互,” arXiv预印本 arXiv:2309.09971, 2023年。

A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox, "Rvt: Robotic view transformer for 3d object manipulation," arXiv preprint arXiv:2306.14896, 2023.

A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, 和 D. Fox, “RVT: 用于3D对象操作的机器人视图变换器,” arXiv 预印本 arXiv:2306.14896, 2023年。

M. Gramopadhye and D. Szafir, "Generating executable action plans with environmentally-aware language models," arXiv preprint arXiv:2210.04964, 2022.

M. Gramopadhye 和 D. Szafir, “利用环境感知语言模型生成可执行动作计划,” arXiv预印本 arXiv:2210.04964, 2022年。

A. Gudibande, E. Wallace, C. Snell, X. Geng, H. Liu, P. Abbeel, S. Levine, and D. Song, "The false promise of imitating proprietary llms," arXiv preprint arXiv:2305.15717, 2023.

A. Gudibande, E. Wallace, C. Snell, X. Geng, H. Liu, P. Abbeel, S. Levine, 和 D. Song, “模仿专有大型语言模型的虚假承诺,” arXiv预印本 arXiv:2305.15717, 2023年。

L. Gui, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, "Vlc: Training vision-language transformers from captions," May 2022.

L. Gui, Q. Huang, A. Hauptmann, Y. Bisk, 和 J. Gao, “VLC：基于字幕训练视觉-语言变换器,” 2022年5月。

L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, "Kat: A knowledge augmented transformer for vision-and-language," in NAACL 2022. Long paper, Oral., January 2022.

L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, 和 J. Gao, “KAT：一种知识增强的视觉与语言变换器,” 载于 NAACL 2022。长文，口头报告，2022年1月。

R. L. Guimaraes, A. S. de Oliveira, J. A. Fabro, T. Becker, and V. A. Brenner, "Ros navigation: Concepts and tutorial," Robot Operating System (ROS) The Complete Reference (Volume 1), pp. 121-160, 2016.

R. L. Guimaraes, A. S. de Oliveira, J. A. Fabro, T. Becker, 和 V. A. Brenner, “ROS导航：概念与教程,” 《机器人操作系统（ROS）完整参考（第1卷）》，第121-160页，2016年。

K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in International conference on machine learning. PMLR, 2020, pp. 3929-3938.

K. Guu, K. Lee, Z. Tung, P. Pasupat, 和 M. Chang, “检索增强语言模型预训练,” 载于国际机器学习大会。PMLR, 2020年，第3929-3938页。

H. Ha, P. Florence, and S. Song, "Scaling up and distilling down: Language-guided robot skill acquisition," arXiv preprint arXiv:2307.14535, 2023.

H. Ha, P. Florence, 和 S. Song, “规模化与蒸馏：语言引导的机器人技能获取,” arXiv预印本 arXiv:2307.14535, 2023年。

T. Haarnoja, B. Moran, G. Lever, S. H. Huang, D. Tirumala, M. Wulfmeier, J. Humplik, S. Tunyasuvunakool, N. Y. Siegel, R. Hafner et al., "Learning agile soccer skills for a bipedal robot with deep reinforcement learning," arXiv preprint arXiv:2304.13653, 2023.

T. Haarnoja, B. Moran, G. Lever, S. H. Huang, D. Tirumala, M. Wulfmeier, J. Humplik, S. Tunyasuvunakool, N. Y. Siegel, R. Hafner 等, “利用深度强化学习（deep reinforcement learning）学习双足机器人灵活的足球技能”, arXiv预印本 arXiv:2304.13653, 2023年。

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," CVPR, 2022.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, 和 R. Girshick, “掩码自编码器（masked autoencoders）是可扩展的视觉学习者”, CVPR, 2022年。

S. Hemachandra, F. Duvallet, T. M. Howard, N. Roy, A. Stentz, and M. R. Walter, "Learning models for following natural language directions in unknown environments," arXiv preprint arXiv:1503.05079, 2015.

S. Hemachandra, F. Duvallet, T. M. Howard, N. Roy, A. Stentz, 和 M. R. Walter, “在未知环境中学习遵循自然语言指令的模型”, arXiv预印本 arXiv:1503.05079, 2015年。

J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world?" Behavioral and Brain Sciences, vol. 33, no. 2-3, p. 61-83, 2010.

J. Henrich, S. J. Heine, 和 A. Norenzayan, “世界上最怪异的人群? ”《行为与脑科学》（Behavioral and Brain Sciences），第33卷，第2-3期，61-83页，2010年。

L. B. Hensel, N. Yongsatianchot, P. Torshizi, E. Minucci, and S. Marsella, "Large language models in textual analysis for gesture selection," in INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION, 2023, pp. 378-387.

L. B. Hensel, N. Yongsatianchot, P. Torshizi, E. Minucci, 和 S. Marsella, “用于手势选择的文本分析中的大型语言模型”, 发表于国际多模态交互会议（INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION），2023年，页378-387。

D. Ho, K. Rao, Z. Xu, E. Jang, M. Khansari, and Y. Bai, "Retinagan: An object-aware approach to sim-to-real transfer," in 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021, pp. 10920-10926.

D. Ho, K. Rao, Z. Xu, E. Jang, M. Khansari, 和 Y. Bai, “Retinagan：一种面向对象的仿真到现实转移方法”，发表于2021年IEEE国际机器人与自动化会议（ICRA），IEEE，2021年，页10920-10926。

C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023, pp. 10608-10615.

C. Huang, O. Mees, A. Zeng, 和 W. Burgard, “用于机器人导航的视觉语言地图”，发表于2023年IEEE国际机器人与自动化会议（ICRA），IEEE，2023年，页10608-10615。

Q. Huang, J. S. Park, A. Gupta, P. Bennett, R. Gong, S. Som, B. Peng, O. K. Mohammed, C. Pal, Y. Choi et al., "Ark: Augmented reality with knowledge interactive emergent ability," arXiv preprint arXiv:2305.00970, 2023.

Q. Huang, J. S. Park, A. Gupta, P. Bennett, R. Gong, S. Som, B. Peng, O. K. Mohammed, C. Pal, Y. Choi 等, “ARK：具备知识交互新兴能力的增强现实”，arXiv预印本 arXiv:2305.00970，2023年。

W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in Proceedings of the 39th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17-23 Jul 2022, pp. 9118-9147. [Online]. Available: <https://proceedings.mlr.press/v162/huang22a.html>

W. Huang, P. Abbeel, D. Pathak, 和 I. Mordatch, “语言模型作为零样本规划者：为具身智能体提取可执行知识”，发表于第39届国际机器学习大会论文集，机器学习研究系列，K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, 和 S. Sabato 编，第162卷，PMLR，2022年7月17-23日，页9118-9147。[在线] 可访问：<https://proceedings.mlr.press/v162/huang22a.html>

W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, P. Sermanet, N. Brown, T. Jackson, L. Luu, S. Levine, K. Hausman, and B. Ichter, "Inner monologue: Embodied reasoning through planning with language models," in arXiv preprint arXiv:2207.05608, 2022.

W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, P. Sermanet, N. Brown, T. Jackson, L. Luu, S. Levine, K. Hausman, 和 B. Ichter, “内心独白：通过语言模型进行规划的具身推理”，arXiv预印本 arXiv:2207.05608，2022年。

Z. Huang, H. Feng, Z. Chongzhi, L. Sheng, L. Ziwei, and J. Shao, "Dolphin: General video interaction platform based on llms," 2023, <https://github.com/kaleido-lab/dolphin>.

Z. Huang, H. Feng, Z. Chongzhi, L. Sheng, L. Ziwei, 和 J. Shao, “Dolphin：基于大型语言模型（LLMs）的通用视频交互平台”，2023年，<https://github.com/kaleido-lab/dolphin>。

K. Ikeuchi, N. Wake, K. Sasabuchi, and J. Takamatsu, "Semantic constraints to represent common sense required in household actions for multimodal learning-from-observation robot," The International Journal of Robotics Research, vol. 0, no. 0, p. 02783649231212929, 0.

K. Ikeuchi, N. Wake, K. Sasabuchi, 和 J. Takamatsu, “用于多模态观察学习机器人家庭动作中常识表示的语义约束”，《国际机器人研究杂志》（The International Journal of Robotics Research），第0卷，第0期，文章编号02783649231212929，年份未定。

K. Ikeuchi, J. Takamatsu, K. Sasabuchi, N. Wake, and A. Kanehiro, "Applying learning-from-observation to household service robots: three common-sense formulation," arXiv preprint arXiv:2304.09966, 2023.

K. Ikeuchi, J. Takamatsu, K. Sasabuchi, N. Wake, 和 A. Kanehiro, “将观察学习应用于家庭服务机器人：三种常识性表述”，arXiv预印本 arXiv:2304.09966，2023年。

S. Imani, L. Du, and H. Shrivastava, "Mathprompter: Mathematical reasoning using large language models," arXiv preprint arXiv:2303.05398, 2023.

S. Imani, L. Du, 和 H. Shrivastava, "Mathprompter: 利用大型语言模型进行数学推理," arXiv预印本 arXiv:2303.05398, 2023.

S. James and A. J. Davison, "Q-attention: Enabling efficient learning for vision-based robotic manipulation," IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 1612-1619, 2022.

S. James 和 A. J. Davison, "Q-attention: 实现基于视觉的机器人操作的高效学习," IEEE机器人与自动化快报, 第7卷, 第2期, 页1612-1619, 2022.

E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in Conference on Robot Learning. PMLR, 2022, pp. 991-1002.

E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, 和 C. Finn, "Bc-z: 通过机器人模仿学习实现零样本任务泛化," 机器人学习会议论文集, PMLR, 2022, 页991-1002.

Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1-38, 2023.

Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, 和 P. Fung, "自然语言生成中的幻觉现象综述," ACM计算机调查, 第55卷, 第12期, 页1-38, 2023.

Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "Vima: General robot manipulation with multimodal prompts," arXiv, 2022.

Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, 和 L. Fan, "Vima: 基于多模态提示的通用机器人操作," arXiv, 2022.

D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke et al., "Scalable deep reinforcement learning for vision-based robotic manipulation," in Conference on Robot Learning. PMLR, 2018, pp. 651-673.

D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke 等, "面向基于视觉的机器人操作的可扩展深度强化学习," 机器人学习会议论文集, PMLR, 2018, 页651-673.

A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," Advances in neural information processing systems, vol. 27, 2014.

A. Karpathy, A. Joulin, 和 L. F. Fei-Fei, "用于双向图像-句子映射的深度片段嵌入," 神经信息处理系统进展, 第27卷, 2014.

P. Katara, Z. Xian, and K. Fragkiadaki, "Gen2sim: Scaling up robot learning in simulation with generative models," arXiv preprint arXiv:2310.18308, 2023.

P. Katara, Z. Xian, 和 K. Fragkiadaki, "Gen2sim: 利用生成模型扩展机器人仿真学习," arXiv预印本 arXiv:2310.18308, 2023.

L. Ke, X. Li, B. Yonatan, A. Holtzman, Z. Gan, J. Liu, J. Gao, Y. Choi, and S. Srinivasa, "Tactical rewind: Self-correction via backtracking in vision-and-language navigation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

L. Ke, X. Li, B. Yonatan, A. Holtzman, Z. Gan, J. Liu, J. Gao, Y. Choi, 和 S. Srinivasa, "战术回溯: 视觉与语言导航中的自我纠正," IEEE计算机视觉与模式识别会议论文集 (CVPR) , 2019.

J. Kim, J. Kim, and S. Choi, "Flame: Free-form language-based motion synthesis & editing," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 7, 2023, pp. 8255-8263.

J. Kim, J. Kim, 和 S. Choi, "Flame: 基于自由形式语言的动作合成与编辑," AAAI人工智能会议论文集, 第37卷, 第7期, 2023, 页8255-8263.

P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, and J. B. Tenenbaum, "Learning to act from actionless videos through dense correspondences," arXiv preprint arXiv:2310.08576, 2023.

P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, 和 J. B. Tenenbaum, “通过密集对应关系从无动作视频中学习行动,” arXiv预印本 arXiv:2310.08576, 2023.

E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, M. Deitke, K. Ehsani, D. Gordon, Y. Zhu et al., "Ai2-thor: An interactive 3d environment for visual ai," arXiv preprint arXiv:1712.05474, 2017.

E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, M. Deitke, K. Ehsani, D. Gordon, Y. Zhu 等, "Ai2-thor: 用于视觉人工智能的交互式三维环境," arXiv预印本 arXiv:1712.05474, 2017.

R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," in arXiv:1602.07332, 2016.

R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, 和 L. Fei-Fei, “视觉基因组（Visual Genome）：利用众包密集图像标注连接语言与视觉,” arXiv:1602.07332, 2016.

K. N. Kumar, I. Essa, and S. Ha, "Words into action: Learning diverse humanoid robot behaviors using language guided iterative motion refinement," arXiv preprint arXiv:2310.06226, 2023.

K. N. Kumar, I. Essa, 和 S. Ha, “言语化为行动：利用语言引导的迭代动作优化学习多样化类人机器人行为,” arXiv 预印本 arXiv:2310.06226, 2023.

P. Lee, S. Bubeck, and J. Petro, "Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine," New England Journal of Medicine, vol. 388, no. 13, pp. 1233-1239, 2023.

P. Lee, S. Bubeck, 和 J. Petro, “作为医学领域AI聊天机器人的GPT-4的益处、局限与风险,”《新英格兰医学杂志》(New England Journal of Medicine), 第388卷, 第13期, 页1233-1239, 2023年。

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks," Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel 等, “面向知识密集型自然语言处理任务的检索增强生成,”《神经信息处理系统进展》(Advances in Neural Information Processing Systems), 第33卷, 页9459-9474, 2020年。

B. Li, P. Wu, P. Abbeel, and J. Malik, "Interactive task planning with language models," arXiv preprint arXiv:2310.10645, 2023.

B. Li, P. Wu, P. Abbeel, 和 J. Malik, “基于语言模型的交互式任务规划,” arXiv预印本 arXiv:2310.10645, 2023年。

C. Li, F. Xia, R. Martín-Martín, M. Lingelbach, S. Srivastava, B. Shen, K. Vainio, C. Gokmen, G. Dharan, T. Jain et al., "igibson 2.0: Object-centric simulation for robot learning of everyday household tasks," arXiv preprint arXiv:2108.03272, 2021.

C. Li, F. Xia, R. Martín-Martín, M. Lingelbach, S. Srivastava, B. Shen, K. Vainio, C. Gokmen, G. Dharan, T. Jain 等, “iGibson 2.0: 面向机器人学习日常家务任务的以对象为中心的仿真,” arXiv预印本 arXiv:2108.03272, 2021年。

C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," arXiv preprint arXiv:2306.00890, 2023.

C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, 和 J. Gao, “Llava-med: 一天内训练的大型语言与视觉助手, 用于生物医学,” arXiv预印本 arXiv:2306.00890, 2023年。

G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "Camel: Communicative agents for " mind" exploration of large scale language model society," arXiv preprint arXiv:2303.17760, 2023.

G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, 和 B. Ghanem, “Camel：用于大规模语言模型社会‘思维’探索的交流代理,” arXiv预印本 arXiv:2303.17760, 2023年。

J. Li, Q. Gao, M. Johnston, X. Gao, X. He, S. Shakiah, H. Shi, R. Ghanadan, and W. Y. Wang, "Mastering robot manipulation with multimodal prompts through pretraining and multi-task fine-tuning," arXiv preprint arXiv:2310.09676, 2023.

J. Li, Q. Gao, M. Johnston, X. Gao, X. He, S. Shakiah, H. Shi, R. Ghanadan, 和 W. Y. Wang, “通过预训练和多任务微调，利用多模态提示掌握机器人操作，” arXiv预印本 arXiv:2310.09676, 2023年。

J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," arXiv preprint arXiv:2301.12597, 2023.

J. Li, D. Li, S. Savarese, 和 S. Hoi, “BLIP-2：利用冻结的图像编码器和大型语言模型引导语言-图像预训练，” arXiv 预印本 arXiv:2301.12597, 2023年。

K. Li, Y. He, W. Yi, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," arXiv preprint arXiv:2305.06355, 2023.

K. Li, Y. He, W. Yi, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, 和 Y. Qiao, “VideoChat：以聊天为中心的视频理解，” arXiv预印本 arXiv:2305.06355, 2023年。

L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "Hero: Hierarchical encoder for video+language omni-representation pre-training," 2020.

L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, 和 J. Liu, “HERO：视频与语言全方位表示预训练的分层编码器，” 2020年。

L. Li, J. Lei, Z. Gan, L. Yu, Y.-C. Chen, R. Pillai, Y. Cheng, L. Zhou, X. E. Wang, W. Y. Wang, T. L. Berg, M. Bansal, J. Liu, L. Wang, and Z. Liu, "Value: A multi-task benchmark for video-and-language understanding evaluation," 2021.

L. Li, J. Lei, Z. Gan, L. Yu, Y.-C. Chen, R. Pillai, Y. Cheng, L. Zhou, X. E. Wang, W. Y. Wang, T. L. Berg, M. Bansal, J. Liu, L. Wang, 和 Z. Liu, “VALUE：视频与语言理解评估的多任务基准，” 2021年。

X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Liu et al., "Vision-language foundation models as effective robot imitators," arXiv preprint arXiv:2311.01378, 2023.

X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Liu 等, “视觉-语言基础模型作为高效的机器人模仿者，” arXiv预印本 arXiv:2311.01378, 2023年。

J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in arXiv preprint arXiv:2209.07753, 2022.

J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, 和 A. Zeng, “代码即策略：用于具身控制的语言模型程序，” arXiv预印本 arXiv:2209.07753, 2022年。

X. Liang, L. Ma, S. Guo, J. Han, H. Xu, S. Ma, and X. Liang, "Mo-vln: A multi-task benchmark for open-set zero-shot vision-and-language navigation," arXiv preprint arXiv:2306.10322, 2023.

X. Liang, L. Ma, S. Guo, J. Han, H. Xu, S. Ma, 和 X. Liang, “MO-VLN：开放集零样本视觉与语言导航的多任务基准，” arXiv预印本 arXiv:2306.10322, 2023年。

S. Lifshitz, K. Paster, H. Chan, J. Ba, and S. McIlraith, "Steve-1: A generative model for text-to-behavior in minecraft," arXiv preprint arXiv:2306.00937, 2023.

S. Lifshitz, K. Paster, H. Chan, J. Ba, 和 S. McIlraith, “Steve-1：Minecraft中从文本到行为的生成模型，” arXiv预印本 arXiv:2306.00937, 2023.

K. Lin, F. Ahmed, L. Li, C.-C. Lin, E. Azarnasab, Z. Yang, J. Wang, L. Liang, Z. Liu, Y. Lu, C. Liu, and L. Wang, "Mm-vid: Advancing video understanding with gpt-4v(ision)," 2023.

K. Lin, F. Ahmed, L. Li, C.-C. Lin, E. Azarnasab, Z. Yang, J. Wang, L. Liang, Z. Liu, Y. Lu, C. Liu, 和 L. Wang, “Mm-vid：利用GPT-4V(ision)推进视频理解，” 2023.

T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollar, "Microsoft coco: Common objects in context," Proceedings of ECCV, 2014.

T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, 和 P. Dollar, "Microsoft COCO: 上下文中的常见物体," ECCV会议论文集, 2014.

C. K. Liu and D. Negrut, "The role of physics-based simulators in robotics," Annual Review of Control, Robotics, and Autonomous Systems, vol. 4, pp. 35-58, 2021.

C. K. Liu 和 D. Negrut, "基于物理的仿真器在机器人学中的作用," 《控制、机器人与自主系统年评》, 第4卷, 第35-58页, 2021.

H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

H. Liu, C. Li, Q. Wu, 和 Y. J. Lee, "视觉指令调优," 2023.

H. Liu, A. Chen, Y. Zhu, A. Swaminathan, A. Kolobov, and C.-A. Cheng, "Interactive robot learning from verbal correction," arXiv preprint arXiv:2310.17555, 2023.

H. Liu, A. Chen, Y. Zhu, A. Swaminathan, A. Kolobov, 和 C.-A. Cheng, "通过口头纠正进行交互式机器人学习," arXiv预印本 arXiv:2310.17555, 2023.

Y. Liu, W. Held, and D. Yang, "Dada: Dialect adaptation via dynamic aggregation of linguistic rules," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

Y. Liu, W. Held, 和 D. Yang, "DADA: 通过语言规则的动态聚合进行方言适应," 2023年自然语言处理实证方法会议论文集, 2023.

P. Lu, B. Peng, H. Cheng, M. Galley, K.-W. Chang, Y. N. Wu, S.-C. Zhu, and J. Gao, "Chameleon: Plug-and-play compositional reasoning with large language models," 2023.

P. Lu, B. Peng, H. Cheng, M. Galley, K.-W. Chang, Y. N. Wu, S.-C. Zhu, 和 J. Gao, "Chameleon: 基于大型语言模型的即插即用组合推理," 2023.

Z. Luo, Z. Durante, L. Li, W. Xie, R. Liu, E. Jin, Z. Huang, L. Y. Li, J. Wu, J. C. Niebles et al., "Moma-Irg: Language-refined graphs for multi-object multi-actor activity parsing," Advances in Neural Information Processing Systems, vol. 35, pp. 5282-5298, 2022.

Z. Luo, Z. Durante, L. Li, W. Xie, R. Liu, E. Jin, Z. Huang, L. Y. Li, J. Wu, J. C. Niebles 等, "MOMA-IRG: 用于多对象多参与者活动解析的语言精炼图," 《神经信息处理系统进展》, 第35卷, 第5282-5298页, 2022.

C.-Y. Ma, J. Lu, Z. Wu, G. AlRegib, Z. Kira, R. Socher, and C. Xiong, "Self-monitoring navigation agent via auxiliary progress estimation," arXiv preprint arXiv:1901.03035, 2019.

C.-Y. Ma, J. Lu, Z. Wu, G. AlRegib, Z. Kira, R. Socher, 和 C. Xiong, "通过辅助进度估计的自我监控导航代理," arXiv预印本 arXiv:1901.03035, 2019.

C.-Y. Ma, Z. Wu, G. AlRegib, C. Xiong, and Z. Kira, "The regretful agent: Heuristic-aided navigation through progress estimation," arXiv preprint arXiv:1903.01602, 2019.

C.-Y. Ma, Z. Wu, G. AlRegib, C. Xiong, 和 Z. Kira, "后悔代理: 通过进度估计的启发式辅助导航," arXiv预印本 arXiv:1903.01602, 2019.

Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, "Eureka: Human-level reward design via coding large language models," arXiv preprint arXiv:2310.12931, 2023.

Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, 和 A. Anandkumar, "EUREKA: 通过编码大型语言模型实现人类水平的奖励设计," arXiv预印本 arXiv:2310.12931, 2023.

M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," 2023.

M. Maaz, H. Rasheed, S. Khan, 和 F. S. Khan, "Video-ChatGPT: 通过大型视觉与语言模型实现详细视频理解," 2023.

R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," IEEE Transactions on Affective Computing, 2022.

R. Mao, Q. Liu, K. He, W. Li, 和 E. Cambria, "预训练语言模型的偏见：基于提示的情感分析与情绪检测的实证研究," IEEE情感计算汇刊, 2022.

G. Marcus, "The next decade in ai: four steps towards robust artificial intelligence," arXiv preprint arXiv:2002.06177, 2020.

G. Marcus, "人工智能的下一个十年：迈向稳健人工智能的四个步骤," arXiv预印本 arXiv:2002.06177, 2020.

G. Marcus and E. Davis, Rebooting AI: Building artificial intelligence we can trust. Pantheon, 2019.

G. Marcus 和 E. Davis, 《重启人工智能：构建我们可以信赖的人工智能》. Pantheon出版社, 2019.

K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in CVPR, 2019.

K. Marino, M. Rastegari, A. Farhadi, 和 R. Mottaghi, "Ok-vqa: 一个需要外部知识的视觉问答基准," 发表在 CVPR, 2019。

P. Martinez-Gonzalez, S. Oprea, A. Garcia-Garcia, A. Jover-Alvarez, S. Orts-Escalano, and J. Garcia-Rodriguez, "Unrealrox: an extremely photorealistic virtual reality environment for robotics simulations and synthetic data generation," Virtual Reality, vol. 24, pp. 271-288, 2020.

P. Martinez-Gonzalez, S. Oprea, A. Garcia-Garcia, A. Jover-Alvarez, S. Orts-Escalano, 和 J. Garcia-Rodriguez, "Unrealrox: 一个极度逼真的虚拟现实环境，用于机器人仿真和合成数据生成," 《虚拟现实》，第24卷, 第271-288页, 2020。

J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 1906-1919. [Online]. Available: <https://aclanthology.org/2020.acl-main.173>

J. Maynez, S. Narayan, B. Bohnet, 和 R. McDonald, "关于抽象摘要中的忠实性和事实性," 发表在第58届计算语言学协会年会论文集, D. Jurafsky, J. Chai, N. Schluter, 和 J. Tetreault 编, 在线: 计算语言学协会, 2020年7月, 第1906-1919页。[在线]. 可获取: <https://aclanthology.org/2020.acl-main.173>

O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," IEEE Robotics and Automation Letters, vol. 7, no. 3, pp. 7327-7334, 2022.

O. Mees, L. Hermann, E. Rosete-Beas, 和 W. Burgard, "Calvin: 一个用于长时程机器人操作任务的语言条件策略学习基准," 《IEEE机器人与自动化快报》, 第7卷, 第3期, 第7327-7334页, 2022。

Meta Fundamental AI Research (FAIR) Diplomacy Team, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu et al., "Human-level play in the game of Diplomacy by combining language models with strategic reasoning," Science, vol. 378, no. 6624, pp. 1067-1074, 2022.

Meta基础人工智能研究(FAIR)外交团队, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu 等, "通过结合语言模型与战略推理实现外交游戏中的人类水平玩法," 《科学》, 第378卷, 第6624期, 第1067-1074页, 2022。

S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" arXiv preprint arXiv:2202.12837, 2022.

S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, 和 L. Zettlemoyer, "重新思考示范的作用：是什么使得上下文学习有效? " arXiv预印本 arXiv:2202.12837, 2022.

M. L. Minsky, "Minsky's frame system theory," in Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing, ser. TINLAP '75. USA: Association for Computational Linguistics, 1975, p. 104-116. [Online]. Available: <https://doi.org/10.3115/980190.980222>

M. L. Minsky, "明斯基的框架系统理论," 发表在1975年自然语言处理理论问题研讨会论文集, 系列 TINLAP '75. 美国: 计算语言学协会, 1975, 第104-116页。[在线]. 可获取: <https://doi.org/10.3115/980190.980222>

S. Mirchandani, F. Xia, P. Florence, B. Ichter, D. Driess, M. G. Arenas, K. Rao, D. Sadigh, and A. Zeng, "Large language models as general pattern machines," arXiv preprint arXiv:2307.04721, 2023.

S. Mirchandani, F. Xia, P. Florence, B. Ichter, D. Driess, M. G. Arenas, K. Rao, D. Sadigh, 和 A. Zeng, "大型语言模型作为通用模式机器," arXiv预印本 arXiv:2307.04721, 2023。

P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu et al., "Learning to navigate in complex environments," arXiv preprint arXiv:1611.03673, 2016.

P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu 等, "学习在复杂环境中导航," arXiv预印本 arXiv:1611.03673, 2016。

M. Mittal, C. Yu, Q. Yu, J. Liu, N. Rudin, D. Hoeller, J. L. Yuan, R. Singh, Y. Guo, H. Mazhar et al., "Orbit: A unified simulation framework for interactive robot learning environments," IEEE Robotics and Automation Letters, 2023.

M. Mittal, C. Yu, Q. Yu, J. Liu, N. Rudin, D. Hoeller, J. L. Yuan, R. Singh, Y. Guo, H. Mazhar 等, "Orbit: 一个用于交互式机器人学习环境的统一仿真框架," 《IEEE机器人与自动化快报》, 2023。

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski et al., "Human-level control through deep reinforcement learning," nature, vol. 518, no. 7540, pp. 529-533, 2015.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski 等, "通过深度强化学习实现人类水平控制," 《自然》, 第518卷, 第7540期, 第529-533页, 2015。

A. Mousavian, A. Toshev, M. Fiser, J. Kosecka, and J. Davidson, "Visual representations for semantic target driven navigation," arXiv preprint arXiv:1805.06066, 2018.

A. Mousavian, A. Toshev, M. Fiser, J. Kosecka, 和 J. Davidson, "用于语义目标驱动导航的视觉表示," arXiv预印本 arXiv:1805.06066, 2018。

T. Mu, Z. Ling, F. Xiang, D. Yang, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su, "Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations," arXiv preprint arXiv:2107.14483, 2021.

T. Mu, Z. Ling, F. Xiang, D. Yang, X. Li, S. Tao, Z. Huang, Z. Jia, 和 H. Su, "Maniskill: 具有大规模示范的可泛化操作技能基准, "arXiv预印本 arXiv:2107.14483, 2021。

M. Müller, V. Casser, J. Lahoud, N. Smith, and B. Ghanem, "Sim4cv: A photo-realistic simulator for computer vision applications," International Journal of Computer Vision, vol. 126, pp. 902-919, 2018.

M. Müller, V. Casser, J. Lahoud, N. Smith, 和 B. Ghanem, "Sim4cv: 用于计算机视觉应用的照片级真实感模拟器, " 《国际计算机视觉杂志》, 第126卷, 第902-919页, 2018。

J. J. Nay, "Law informs code: A legal informatics approach to aligning artificial intelligence with humans," Nw. J. Tech. & Intell. Prop., vol. 20, p. 309, 2022.

J. J. Nay, "法律指导代码: 一种将人工智能与人类对齐的法律信息学方法, " 《西北技术与知识产权杂志》, 第20卷, 第309页, 2022。

K. Nguyen, D. Dey, C. Brockett, and B. Dolan, "Vision-based navigation with language-based assistance via imitation learning with indirect intervention," arXiv preprint arXiv:1812.04155, 2018.

K. Nguyen, D. Dey, C. Brockett, 和 B. Dolan, "基于视觉的导航结合基于语言的辅助, 通过间接干预的模仿学习, " arXiv预印本 arXiv:1812.04155, 2018。

Z. Ni, X.-X. Deng, C. Tai, X.-Y. Zhu, X. Wu, Y.-J. Liu, and L. Zeng, "Grid: Scene-graph-based instruction-driven robotic task planning," arXiv preprint arXiv:2309.07726, 2023.

Z. Ni, X.-X. Deng, C. Tai, X.-Y. Zhu, X. Wu, Y.-J. Liu, 和 L. Zeng, "Grid: 基于场景图的指令驱动机器人任务规划, "arXiv预印本 arXiv:2309.07726, 2023。

OpenAI, "GPT-4 technical report," OpenAI, Tech. Rep., 2023.

OpenAI, "GPT-4技术报告, "OpenAI, 技术报告, 2023。

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray et al., "Training language models to follow instructions with human feedback," Advances in Neural Information Processing Systems, vol. 35, pp. 27730-27744, 2022.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray 等, "通过人类反馈训练语言模型以遵循指令, "《神经信息处理系统进展》, 第35卷, 第27730-27744页, 2022。

A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan et al., "Open x-embodiment: Robotic learning datasets and rt-x models," arXiv preprint arXiv:2310.08864,2023.

A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan 等, "开放X-具身: 机器人学习数据集与RT-X模型, "arXiv预印本 arXiv:2310.08864, 2023。

M. Parakh, A. Fong, A. Simeonov, A. Gupta, T. Chen, and P. Agrawal, "Human-assisted continual robot learning with foundation models," arXiv preprint arXiv:2309.14321, 2023.

M. Parakh, A. Fong, A. Simeonov, A. Gupta, T. Chen, 和 P. Agrawal, "基于基础模型的人类辅助持续机器人学习, "arXiv预印本 arXiv:2309.14321, 2023。

J. S. Park, J. Hessel, K. Chandu, P. P. Liang, X. Lu, P. West, Q. Huang, J. Gao, A. Farhadi, and Y. Choi, "Multimodal agent - localized symbolic knowledge distillation for visual commonsense models," in NeurIPS 2023, October 2023.

J. S. Park, J. Hessel, K. Chandu, P. P. Liang, X. Lu, P. West, Q. Huang, J. Gao, A. Farhadi, 和 Y. Choi, "多模态代理——用于视觉常识模型的局部符号知识蒸馏, "NeurIPS 2023, 2023年10月。

J. S. Park, J. Hessel, K. Chandu, P. P. Liang, X. Lu, P. West, Y. Yu, Q. Huang, J. Gao, A. Farhadi, and Y. Choi, "Localized symbolic knowledge distillation for visual commonsense models," in Thirty-seventh Conference on Neural Information Processing Systems, 2023. [Online]. Available: <https://openreview.net/forum?id=V5eG47pyVl>

J. S. Park, J. Hessel, K. Chandu, P. P. Liang, X. Lu, P. West, Y. Yu, Q. Huang, J. Gao, A. Farhadi, 和 Y. Choi, "用于视觉常识模型的局部符号知识蒸馏, "第三十七届神经信息处理系统会议, 2023。[在线]. 可获取: <https://openreview.net/forum?id=V5eG47pyVl>

J. Park, Q. Huang, Y. Bisk, J. Yang, S. Som, A. Farhadi, Y. Choi, and J. Gao, "Ink: Intensive neural knowledge," July 2022.

J. Park, Q. Huang, Y. Bisk, J. Yang, S. Som, A. Farhadi, Y. Choi, 和 J. Gao, "INK: 密集神经知识, "2022年7月。

J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," arXiv preprint arXiv:2304.03442, 2023.

J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, 和 M. S. Bernstein, "生成代理: 人类行为的交互模拟体," arXiv预印本 arXiv:2304.03442, 2023.

B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen et al., "Check your facts and try again: Improving large language models with external knowledge and automated feedback," arXiv preprint arXiv:2302.12813, 2023.

B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen 等, "核实事实并重试: 利用外部知识和自动反馈提升大型语言模型," arXiv预印本 arXiv:2302.12813, 2023.

X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in 2018 IEEE international conference on robotics and automation (ICRA). IEEE, 2018, pp. 3803-3810.

X. B. Peng, M. Andrychowicz, W. Zaremba, 和 P. Abbeel, “通过动力学随机化实现机器人控制的仿真到现实转移,”载于2018年IEEE国际机器人与自动化会议 (ICRA) , IEEE, 2018, 页3803-3810.

X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, and A. Torralba, "Virtualhome: Simulating household activities via programs," in 2018 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8494-8502.

X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, 和 A. Torralba, “Virtualhome：通过程序模拟家庭活动,” 载于2018年IEEE国际计算机视觉与模式识别会议 (CVPR) , 2018, 页8494-8502.

X. Puig, E. Undersander, A. Szot, M. D. Cote, T.-Y. Yang, R. Partsey, R. Desai, A. W. Clegg, M. Hlavac, S. Y. Min et al., "Habitat 3.0: A co-habitat for humans, avatars and robots," arXiv preprint arXiv:2310.13724, 2023.

X. Puig, E. Undersander, A. Szot, M. D. Cote, T.-Y. Yang, R. Partsey, R. Desai, A. W. Clegg, M. Hlavac, S. Y. Min 等, “Habitat 3.0：人类、虚拟形象与机器人大共居环境,” arXiv预印本 arXiv:2310.13724, 2023.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning. PMLR, 2021, pp. 8748-8763.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark 等, “从自然语言监督中学习可迁移的视觉模型,” 载于国际机器学习会议, PMLR, 2021, 页8748-8763.

S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "Poni: Potential functions for objectgoal navigation with interaction-free learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18890-18900.

S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, 和 K. Grauman, “Poni：基于势函数的无交互学习目标导航,” 载于IEEE/CVF计算机视觉与模式识别会议论文集, 2022, 页18890-18900.

S. S. Raman, V. Cohen, D. Paulius, I. Idrees, E. Rosen, R. Mooney, and S. Tellex, "Cape: Corrective actions from precondition errors using large language models," in 2nd Workshop on Language and Robot Learning: Language as Grounding, 2023.

S. S. Raman, V. Cohen, D. Paulius, I. Idrees, E. Rosen, R. Mooney, 和 S. Tellex, “Cape：利用大型语言模型纠正前置条件错误的修正动作,” 载于第二届语言与机器人学习研讨会：语言作为基础, 2023.

K. Rao, C. Harris, A. Irpan, S. Levine, J. Ibarz, and M. Khansari, "Rl-cyclegan: Reinforcement learning aware simulation-to-real," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11157-11166.

K. Rao, C. Harris, A. Irpan, S. Levine, J. Ibarz, 和 M. Khansari, “RL-CycleGAN：强化学习感知的仿真到现实,” 载于IEEE/CVF计算机视觉与模式识别会议论文集, 2020, 页11157-11166.

V. Raunak, A. Menezes, and M. Junczys-Dowmunt, "The curious case of hallucinations in neural machine translation," arXiv preprint arXiv:2104.06683, 2021.

V. Raunak, A. Menezes, 和 M. Junczys-Dowmunt, “神经机器翻译中的幻觉现象探析,” arXiv预印本 arXiv:2104.06683, 2021.

S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg et al., "A generalist agent," arXiv preprint arXiv:2205.06175, 2022.

S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg 等, “通用智能体,” arXiv预印本 arXiv:2205.06175, 2022.

M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," Advances in neural information processing systems, vol. 28, 2015.

M. Ren, R. Kiros, 和 R. Zemel, “图像问答的模型与数据探索,” 神经信息处理系统进展, 第28卷, 2015.

A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object hallucination in image captioning," arXiv preprint arXiv:1809.02156, 2018.

A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, 和 K. Saenko, “图像字幕中的对象幻觉,” arXiv预印本 arXiv:1809.02156, 2018.

A. Rosinol, J. J. Leonard, and L. Carlone, "Nerf-slam: Real-time dense monocular slam with neural radiance fields," arXiv preprint arXiv:2210.13641, 2022.

A. Rosinol, J. J. Leonard, 和 L. Carlone, “Nerf-SLAM: 基于神经辐射场的实时密集单目SLAM,” arXiv预印本 arXiv:2210.13641, 2022.

F. Sadeghi and S. Levine, "Cad2rl: Real single-image flight without a single real image," arXiv preprint arXiv:1611.04201, 2016.

F. Sadeghi 和 S. Levine, “Cad2rl: 无需真实图像的单张图像真实飞行”, arXiv预印本 arXiv:1611.04201, 2016年。

D. Saito, K. Sasabuchi, N. Wake, J. Takamatsu, H. Koike, and K. Ikeuchi, "Task-grasping from a demonstrated human strategy," in 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids), 2022, pp. 880-887.

D. Saito, K. Sasabuchi, N. Wake, J. Takamatsu, H. Koike 和 K. Ikeuchi, “基于示范的人类策略的任务抓取”, 载于2022年IEEE-RAS第21届国际仿人机器人会议 (Humanoids) , 2022年, 第880-887页。

D. Saito, K. Sasabuchi, N. Wake, A. Kanehira, J. Takamatsu, H. Koike, and K. Ikeuchi, "Constraint-aware policy for compliant manipulation," 2023.

D. Saito, K. Sasabuchi, N. Wake, A. Kanehira, J. Takamatsu, H. Koike 和 K. Ikeuchi, “考虑约束的柔顺操作策略”, 2023年。

B. Sarkar, A. Shih, and D. Sadigh, "Diverse conventions for human-AI collaboration," in Thirty-seventh Conference on Neural Information Processing Systems, 2023.

B. Sarkar, A. Shih 和 D. Sadigh, “人机协作的多样化约定”, 载于第三十七届神经信息处理系统会议, 2023年。

K. Sasabuchi, N. Wake, and K. Ikeuchi, "Task-oriented motion mapping on robots of various configuration using body role division," IEEE Robotics and Automation Letters, vol. 6, no. 2, pp. 413-420, 2021.

K. Sasabuchi, N. Wake 和 K. Ikeuchi, “基于身体角色划分的多配置机器人任务导向运动映射”, IEEE机器人与自动化快报, 卷6, 第2期, 第413-420页, 2021年。

K. Sasabuchi, D. Saito, A. Kanehira, N. Wake, J. Takamatsu, and K. Ikeuchi, "Task-sequencing simulator: Integrated machine learning to execution simulation for robot manipulation," arXiv preprint arXiv:2301.01382, 2023.

K. Sasabuchi, D. Saito, A. Kanehira, N. Wake, J. Takamatsu 和 K. Ikeuchi, “任务序列模拟器：集成机器学习与执行仿真的机器人操作”, arXiv预印本 arXiv:2301.01382, 2023年。

M. Savva, A. X. Chang, A. Dosovitskiy, T. Funkhouser, and V. Koltun, "Minos: Multimodal indoor simulator for navigation in complex environments," arXiv preprint arXiv:1712.03931, 2017.

M. Savva, A. X. Chang, A. Dosovitskiy, T. Funkhouser 和 V. Koltun, “Minos: 用于复杂环境导航的多模态室内模拟器”, arXiv预印本 arXiv:1712.03931, 2017年。

M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik et al., "Habitat: A platform for embodied ai research," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9339-9347.

M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik 等,

“Habitat：一个具身人工智能研究平台”，载于IEEE/CVF国际计算机视觉会议论文集，2019年，第9339-9347页。

T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," 2023.

T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda 和 T. Scialom, "Toolformer：语言模型自我学习使用工具”，2023年。

E. Segalis, D. Valevski, D. Lumen, Y. Matias, and Y. Leviathan, "A picture is worth a thousand words: Principled recaptioning improves image generation," arXiv preprint arXiv:2310.16656, 2023.

E. Segalis, D. Valevski, D. Lumen, Y. Matias 和 Y. Leviathan, “一图胜千言：原则性重写标题提升图像生成质量”，arXiv预印本 arXiv:2310.16656，2023年。

B. Shacklett, L. G. Rosenzweig, Z. Xie, B. Sarkar, A. Szot, E. Wijmans, V. Koltun, D. Batra, and K. Fatahalian, "An extensible, data-oriented architecture for high-performance, many-world simulation," ACM Trans. Graph., vol. 42, no. 4, 2023.

B. Shacklett, L. G. Rosenzweig, Z. Xie, B. Sarkar, A. Szot, E. Wijmans, V. Koltun, D. Batra 和 K. Fatahalian, “一种可扩展的数据导向架构用于高性能多世界仿真”，ACM图形学汇刊，卷42，第4期，2023年。

D. Shah, B. Osiński, S. Levine et al., "Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action," in Conference on Robot Learning. PMLR, 2023, pp. 492-504.

D. Shah, B. Osiński, S. Levine 等，“Lm-nav：基于大型预训练语言、视觉与动作模型的机器人导航”，载于机器人学习会议，PMLR，2023年，第492-504页。

R. Shah, R. Martín-Martín, and Y. Zhu, "Mutex: Learning unified policies from multimodal task specifications," arXiv preprint arXiv:2309.14320, 2023.

R. Shah, R. Martín-Martín 和 Y. Zhu, “Mutex：从多模态任务规范中学习统一策略”，arXiv预印本 arXiv:2309.14320，2023年。

S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in Field and Service Robotics: Results of the 11th International Conference. Springer, 2018, pp. 621-635.

S. Shah, D. Dey, C. Lovett 和 A. Kapoor, “Airsim：用于自动驾驶车辆的高保真视觉与物理仿真”，载于第11届国际现场与服务机器人会议成果集，Springer，2018年，第621-635页。

P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018.

P. Sharma, N. Ding, S. Goodman 和 R. Soricut, “概念字幕：一个清洗过的、带有上位词的图像替代文本数据集，用于自动图像字幕生成”，第56届计算语言学协会年会论文集，2018年。

R. Shi, Y. Liu, Y. Ze, S. S. Du, and H. Xu, "Unleashing the power of pre-trained language models for offline reinforcement learning," arXiv preprint arXiv:2310.20587, 2023.

R. Shi, Y. Liu, Y. Ze, S. S. Du 和 H. Xu, “释放预训练语言模型在离线强化学习中的潜力”，arXiv预印本 arXiv:2310.20587，2023年。

M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in Conference on Robot Learning. PMLR, 2023, pp. 785-799.

M. Shridhar, L. Manuelli, 和 D. Fox, "Perceiver-actor：一种用于机器人操作的多任务变换器（transformer）," 载于机器人学习会议（Conference on Robot Learning），PMLR，2023，页785-799。

K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," arXiv preprint arXiv:2104.07567, 2021.

K. Shuster, S. Poff, M. Chen, D. Kiela, 和 J. Weston, "检索增强减少对话中的幻觉现象," arXiv预印本 arXiv:2104.07567，2021。

A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, "Towards vqa models that can read," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 8317-8326.

A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, 和 M. Rohrbach, "迈向能够阅读的视觉问答 (VQA) 模型," 载于IEEE/CVF计算机视觉与模式识别会议论文集, 2019, 页8317-8326。

S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," IEEE Conference on Computer Vision and Pattern Recognition, 2017.

S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, 和 T. Funkhouser, "基于单幅深度图的语义场景补全," IEEE计算机视觉与模式识别会议, 2017。

S. A. Sontakke, J. Zhang, S. M. Arnold, K. Pertsch, E. Biyik, D. Sadigh, C. Finn, and L. Itti, "Roboclip: One demonstration is enough to learn robot policies," arXiv preprint arXiv:2310.07899, 2023.

S. A. Sontakke, J. Zhang, S. M. Arnold, K. Pertsch, E. Biyik, D. Sadigh, C. Finn, 和 L. Itti, "RoboCLIP: 一次示范即可学习机器人策略," arXiv预印本 arXiv:2310.07899, 2023。

S. Srivastava, C. Li, M. Lingelbach, R. Martín-Martín, F. Xia, K. E. Vainio, Z. Lian, C. Gokmen, S. Buch, K. Liu et al., "Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments," in Conference on Robot Learning. PMLR, 2022, pp. 477-490.

S. Srivastava, C. Li, M. Lingelbach, R. Martín-Martín, F. Xia, K. E. Vainio, Z. Lian, C. Gokmen, S. Buch, K. Liu 等, "Behavior: 虚拟、交互和生态环境中日常家务活动的基准测试," 载于机器人学习会议 (Conference on Robot Learning) , PMLR, 2022, 页477-490。

J. Sun, Q. Zhang, Y. Duan, X. Jiang, C. Cheng, and R. Xu, "Prompt, plan, perform: Llm-based humanoid control via quantized imitation learning," arXiv preprint arXiv:2309.11359, 2023.

J. Sun, Q. Zhang, Y. Duan, X. Jiang, C. Cheng, 和 R. Xu, "提示、规划、执行：基于大语言模型 (LLM) 的类人控制通过量化模仿学习," arXiv预印本 arXiv:2309.11359, 2023。

Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," arXiv preprint arXiv:2303.15389, 2023.

Q. Sun, Y. Fang, L. Wu, X. Wang, 和 Y. Cao, "EVA-CLIP: 大规模CLIP训练的改进技术," arXiv预印本 arXiv:2303.15389, 2023。

A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra, "Habitat 2.0: Training home assistants to rearrange their habitat," in Advances in Neural Information Processing Systems (NeurIPS), 2021.

A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, 和 D. Batra, "Habitat 2.0: 训练家庭助理重新布置其生活环境," 载于神经信息处理系统进展 (NeurIPS) , 2021。

J. Takamatsu, K. Sasabuchi, N. Wake, A. Kanehira, and K. Ikeuchi, "Learning-from-observation system considering hardware-level reusability," arXiv preprint arXiv:2212.09242, 2022.

J. Takamatsu, K. Sasabuchi, N. Wake, A. Kanehira, 和 K. Ikeuchi, "考虑硬件级可重用性的观察学习系统," arXiv预印本 arXiv:2212.09242, 2022。

C. Tang, D. Huang, W. Ge, W. Liu, and H. Zhang, "GraspGPT: Leveraging semantic knowledge from a large language model for task-oriented grasping," IEEE Robotics and Automation Letters, 2023.

C. Tang, D. Huang, W. Ge, W. Liu, 和 H. Zhang, "GraspGPT: 利用大型语言模型的语义知识进行面向任务的抓取," IEEE机器人与自动化快报, 2023。

R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model," https://github.com/tatsu-lab/stanford_alpaca, 2023.

R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, 和 T. B. Hashimoto, "斯坦福Alpaca: 一个遵循指令的LLaMA模型," https://github.com/tatsu-lab/stanford_alpaca, 2023。

H. Teshima, N. Wake, D. Thomas, Y. Nakashima, H. Kawasaki, and K. Ikeuchi, "Deep gesture generation for social robots using type-specific libraries," in 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2022, pp. 8286-8291.

H. Teshima, N. Wake, D. Thomas, Y. Nakashima, H. Kawasaki, 和 K. Ikeuchi, "基于类型特定库的社交机器人深度手势生成," 载于2022年IEEE/RSJ国际智能机器人与系统会议 (IROS) , IEEE, 2022, 页8286-8291。

G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," arXiv preprint arXiv:2209.14916, 2022.

G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, 和 A. H. Bermano, "人体动作扩散模型," arXiv预印本 arXiv:2209.14916, 2022。

J. Thomason, D. Gordan, and Y. Bisk, "Shifting the baseline: Single modality performance on visual navigation & qa," arXiv preprint arXiv:1811.00613, 2018.

J. Thomason, D. Gordan, 和 Y. Bisk, "基线转移: 视觉导航与问答中的单模态性能, "arXiv预印本 arXiv:1811.00613, 2018。

J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2017, pp. 23-30.

J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, 和 P. Abbeel, "领域随机化用于将深度神经网络从仿真转移到现实世界, "载于2017年IEEE/RSJ国际智能机器人与系统会议 (IROS) , IEEE, 2017, 页23-30。

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar et al., "Llama: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar 等, "Llama: 开放且高效的基础语言模型, "arXiv预印本 arXiv:2302.13971, 2023。

N. Tsoi, A. Xiang, P. Yu, S. S. Sohn, G. Schwartz, S. Ramesh, M. Hussein, A. W. Gupta, M. Kapadia, and M. Vázquez, "Sean 2.0: Formalizing and generating social situations for robot navigation," IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 11047-11054, 2022.

N. Tsoi, A. Xiang, P. Yu, S. S. Sohn, G. Schwartz, S. Ramesh, M. Hussein, A. W. Gupta, M. Kapadia, 和 M. Vázquez, "Sean 2.0: 用于机器人导航的社会情境形式化与生成, "IEEE机器人与自动化快报, 第7卷第4期, 页11047-11054, 2022。

N. Wake, R. Arakawa, I. Yanokura, T. Kiyokawa, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "A learning-from-observation framework: One-shot robot teaching for grasp-manipulation-release household operations," in 2021 IEEE/SICE International Symposium on System Integration (SII). IEEE, 2021.

N. Wake, R. Arakawa, I. Yanokura, T. Kiyokawa, K. Sasabuchi, J. Takamatsu, 和 K. Ikeuchi, "基于观察学习框架: 一次性机器人教学实现抓取-操作-释放的家务操作, "载于2021年IEEE/SICE系统集成国际研讨会 (SII) , IEEE, 2021。

N. Wake, I. Yanokura, K. Sasabuchi, and K. Ikeuchi, "Verbal focus-of-attention system for learning-from-observation," in 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021, pp. 10377-10384.

N. Wake, I. Yanokura, K. Sasabuchi, 和 K. Ikeuchi, "基于观察学习的口头注意焦点系统, "载于2021年IEEE国际机器人与自动化会议 (ICRA) , 2021, 页10377-10384。

N. Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "Interactive task encoding system for learning-from-observation," in 2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), 2023, pp. 1061-1066.

N. Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, 和 K. Ikeuchi, “基于观察学习的交互式任务编码系统,”载于2023年IEEE/ASME先进智能机电一体化国际会议（AIM），2023，页1061-1066。

—, "Bias in emotion recognition with chatgpt," arXiv preprint arXiv:2310.11753, 2023.

—, "ChatGPT情感识别中的偏见,"arXiv预印本 arXiv:2310.11753, 2023。

—, "Gpt models meet robotic applications: Co-speech gesturing chat system," arXiv preprint arXiv:2306.01741, 2023.

—, "GPT模型与机器人应用的结合：共语手势聊天系统,"arXiv预印本 arXiv:2306.01741, 2023。

—, "Gpt-4v(ision) for robotics: Multimodal task planning from human demonstration," arXiv preprint arXiv:2311.12015, 2023.

—, "GPT-4V(ision)在机器人领域的应用：基于人类示范的多模态任务规划,"arXiv预印本 arXiv:2311.12015, 2023。

—, "Chatgpt empowered long-step robot control in various environments: A case application," IEEE Access, vol. 11, pp. 95060-95 078, 2023.

—, "ChatGPT赋能的多步骤机器人控制在多环境中的应用案例,"IEEE Access, 第11卷, 页95060-95078, 2023。

N. Wake, D. Saito, K. Sasabuchi, H. Koike, and K. Ikeuchi, "Text-driven object affordance for guiding grasp-type recognition in multimodal robot teaching," Machine Vision and Applications, vol. 34, no. 4, p. 58, 2023.

N. Wake, D. Saito, K. Sasabuchi, H. Koike, 和 K. Ikeuchi, “基于文本驱动的物体可供性用于指导多模态机器人教学中的抓取类型识别,”机器视觉与应用, 第34卷第4期, 页58, 2023。

B. Wang, Q. Huang, B. Deb, A. L. Halfaker, L. Shao, D. McDuff, A. Awadallah, D. Radev, and J. Gao, "Logical transformers: Infusing logical structures into pre-trained language models," in Proceedings of ACL 2023, May 2023.

B. Wang, Q. Huang, B. Deb, A. L. Halfaker, L. Shao, D. McDuff, A. Awadallah, D. Radev, 和 J. Gao, “逻辑变换器：将逻辑结构注入预训练语言模型,”载于ACL 2023会议论文集, 2023年5月。

D. Wang, Q. Huang, M. Jackson, and J. Gao, "Retrieve what you need: A mutual learning framework for open-domain question answering," March 2023. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/retrieve-what-you-need-a-mutual-learning-framework-for-open-domain-question-answering/>

D. Wang, Q. Huang, M. Jackson, 和 J. Gao, “检索所需：开放域问答的互学框架,”2023年3月。[在线]. 可获取：<https://www.microsoft.com/en-us/research/publication/retrieve-what-you-need-a-mutual-learning-framework-for-open-domain-question-answering/>

G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar, "Voyager: An open-ended embodied agent with large language models," arXiv preprint arXiv:2305.16291, 2023.

G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, 和 A. Anandkumar, “Voyager：基于大型语言模型的开放式具身智能体,”arXiv预印本 arXiv:2305.16291, 2023。

L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin et al., "A survey on large language model based autonomous agents," arXiv preprint arXiv:2308.11432, 2023.

L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin 等, “基于大型语言模型的自主代理综述”, arXiv预印本 arXiv:2308.11432, 2023年。

P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick, "Explicit knowledge-based reasoning for visual question answering," arXiv preprint arXiv:1511.02570, 2015.

P. Wang, Q. Wu, C. Shen, A. v. d. Hengel 和 A. Dick, “基于显式知识推理的视觉问答”, arXiv预印本 arXiv:1511.02570, 2015年。

P. Wang, Q. Wu, C. Shen, A. Dick, and A. Van Den Hengel, "Fvqa: Fact-based visual question answering," TPAMI, vol. 40, no. 10, pp. 2413-2427, 2017.

P. Wang, Q. Wu, C. Shen, A. Dick 和 A. Van Den Hengel, “FVQA：基于事实的视觉问答”，TPAMI，第40卷，第10期，页2413-2427，2017年。

X. Wang, W. Xiong, H. Wang, and W. Y. Wang, "Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation," in The European Conference on Computer Vision (ECCV), September 2018.

X. Wang, W. Xiong, H. Wang 和 W. Y. Wang, “三思而后行：结合无模型与有模型强化学习实现前瞻性视觉与语言导航”，发表于欧洲计算机视觉大会（ECCV），2018年9月。

X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Weng, W. Y. Wang, and L. Zhang, "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," in CVPR 2019, June 2019.

X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Weng, W. Y. Wang 和 L. Zhang, “基于强化的跨模态匹配与自监督模仿学习用于视觉语言导航”，发表于CVPR 2019，2019年6月。

Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Chen, Y. Wang, P. Luo, Z. Liu, Y. Wang, L. Wang, and Y. Qiao, "Internvid: A large-scale video-text dataset for multimodal understanding and generation," 2023.

Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Chen, Y. Wang, P. Luo, Z. Liu, Y. Wang, L. Wang 和 Y. Qiao, “InternVid：用于多模态理解与生成的大规模视频文本数据集”，2023年。

Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language model with self generated instructions," arXiv preprint arXiv:2212.10560, 2022.

Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi 和 H. Hajishirzi, “Self-instruct：通过自生成指令对齐语言模型”，arXiv预印本 arXiv:2212.10560，2022年。

Y. Wang, Z. Xian, F. Chen, T.-H. Wang, Y. Wang, K. Fragkiadaki, Z. Erickson, D. Held, and C. Gan, "Robogen: Towards unleashing infinite data for automated robot learning via generative simulation," arXiv preprint arXiv:2311.01455, 2023.

Y. Wang, Z. Xian, F. Chen, T.-H. Wang, Y. Wang, K. Fragkiadaki, Z. Erickson, D. Held 和 C. Gan, “RoboGen：通过生成式仿真释放自动机器人学习的无限数据潜力”，arXiv预印本 arXiv:2311.01455，2023年。

Z. Wang, Y. Chen, T. Liu, Y. Zhu, W. Liang, and S. Huang, "Humanise: Language-conditioned human motion generation in 3d scenes," in Advances in Neural Information Processing Systems, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 14959-14971. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/6030db5195150ac86d942186f4abdad8-Paper-Conference.pdf

Z. Wang, Y. Chen, T. Liu, Y. Zhu, W. Liang 和 S. Huang, “Humanise：基于语言条件的三维场景人体动作生成”，发表于神经信息处理系统进展，S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho 和 A. Oh 编，第35卷，Curran Associates, Inc., 2022年，页14959-14971。[在线]. 可获取：https://proceedings.neurips.cc/paper_files/paper/2022/file/6030db5195150ac86d942186f4abdad8-Paper-Conference.pdf

Z. Wang, S. Cai, A. Liu, X. Ma, and Y. Liang, "Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents," arXiv preprint arXiv:2302.01560, 2023.

Z. Wang, S. Cai, A. Liu, X. Ma 和 Y. Liang, “描述、解释、规划与选择：利用大型语言模型实现开放世界多任务代理的交互式规划”，arXiv预印本 arXiv:2302.01560，2023年。

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou et al., "Chain-of-thought prompting elicits reasoning in large language models," Advances in Neural Information Processing Systems, vol. 35, pp. 24824-24837, 2022.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou 等，“链式思维提示激发大型语言模型的推理能力”，神经信息处理系统进展，第35卷，页24824-24837，2022年。

World Health Organization and World Bank, "Tracking universal health coverage: First global monitoring report," www.who.int/healthinfo/universal_health_coverage/report/2015/en, Jun 2015.

世界卫生组织与世界银行，“追踪全民健康覆盖：首份全球监测报告”，www.who.int/healthinfo/universal_health_coverage/report/2015/en，2015年6月。

Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. E. Zhu, B. Li, L. Jiang, X. Zhang, and C. Wang, "Autogen: Enabling next-gen Ilm applications via multi-agent conversation," Microsoft, Tech. Rep. MSR-TR-2023-33, August 2023. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/autogen-enabling-next-gen-ilm-applications-via-multi-agent-conversation-framework/>

Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. E. Zhu, B. Li, L. Jiang, X. Zhang 和 C. Wang, “AutoGen: 通过多代理对话实现下一代大型语言模型应用”，微软，技术报告 MSR-TR-2023-33, 2023年8月。[在线]. 可获取：<http://www.microsoft.com/en-us/research/publication/autogen-enabling-next-gen-ilm-applications-via-multi-agent-conversation-framework/>

Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou et al., "The rise and potential of large language model based agents: A survey," arXiv preprint arXiv:2309.07864, 2023.

Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou 等, “基于大型语言模型代理的兴起与潜力：综述”，arXiv预印本 arXiv:2309.07864, 2023年。

F. Xia, A. R. Zamir, Z.-Y. He, A. Sax, J. Malik, and S. Savarese, "Gibson Env: real-world perception for embodied agents," in Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on. IEEE, 2018.

F. Xia, A. R. Zamir, Z.-Y. He, A. Sax, J. Malik, 和 S. Savarese, “Gibson Env：面向具身智能体的真实世界感知”，发表于2018年IEEE计算机视觉与模式识别会议（CVPR），IEEE, 2018年。

M. Xu, P. Huang, W. Yu, S. Liu, X. Zhang, Y. Niu, T. Zhang, F. Xia, J. Tan, and D. Zhao, "Creative robot tool use with large language models," arXiv preprint arXiv:2310.13065, 2023.

M. Xu, P. Huang, W. Yu, S. Liu, X. Zhang, Y. Niu, T. Zhang, F. Xia, J. Tan, 和 D. Zhao, “利用大型语言模型实现机器人创造性工具使用”，arXiv预印本 arXiv:2310.13065, 2023年。

J. Yang, Y. Dong, S. Liu, B. Li, Z. Wang, C. Jiang, H. Tan, J. Kang, Y. Zhang, K. Zhou et al., "Octopus: Embodied vision-language programmer from environmental feedback," arXiv preprint arXiv:2310.08588, 2023.

J. Yang, Y. Dong, S. Liu, B. Li, Z. Wang, C. Jiang, H. Tan, J. Kang, Y. Zhang, K. Zhou 等, “Octopus：基于环境反馈的具身视觉-语言程序员”，arXiv预印本 arXiv:2310.08588, 2023年。

K. Yang, S. Ji, T. Zhang, Q. Xie, and S. Ananiadou, "On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis," arXiv preprint arXiv:2304.03347, 2023.

K. Yang, S. Ji, T. Zhang, Q. Xie, 和 S. Ananiadou, “关于ChatGPT及情感增强提示在心理健康分析中的评估”，arXiv预印本 arXiv:2304.03347, 2023年。

Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, and L. Wang, "Mm-react: Prompting chatgpt for multimodal reasoning and action," 2023.

Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, 和 L. Wang, “MM-React：利用ChatGPT进行多模态推理与行动提示”，2023年。

S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," 2023.

S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, 和 K. Narasimhan, “思维树（Tree of Thoughts）：利用大型语言模型进行深思熟虑的问题解决”，2023年。

S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," 2023.

S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, 和 Y. Cao, “React：语言模型中推理与行动的协同”，2023年。

Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, C. Li, Y. Xu, H. Chen, J. Tian, Q. Qi, J. Zhang, and F. Huang, "mplug-owl: Modularization empowers large language models with multimodality," 2023.

Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, C. Li, Y. Xu, H. Chen, J. Tian, Q. Qi, J. Zhang, 和 F. Huang, “MPlug-Owl：模块化赋能大型语言模型多模态能力”，2023年。

Y. Ye, H. You, and J. Du, "Improved trust in human-robot collaboration with chatgpt," IEEE Access, 2023.

Y. Ye, H. You, 和 J. Du, “利用ChatGPT提升人机协作中的信任度”，IEEE Access, 2023年。

P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2014.

P. Young, A. Lai, M. Hodosh, 和 J. Hockenmaier, “从图像描述到视觉指称：事件描述语义推理的新相似度度量”，计算语言学协会年会论文集，2014年。

J. Yu, X. Wang, S. Tu, S. Cao, D. Zhang-Li, X. Lv, H. Peng, Z. Yao, X. Zhang, H. Li et al., "Kola: Carefully benchmarking world knowledge of large language models," arXiv preprint arXiv:2306.09296, 2023.

J. Yu, X. Wang, S. Tu, S. Cao, D. Zhang-Li, X. Lv, H. Peng, Z. Yao, X. Zhang, H. Li 等, “KOLA：大型语言模型世界知识的精细基准测试”，arXiv预印本 arXiv:2306.09296, 2023年。

L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer, 2016, pp. 69-85.

L. Yu, P. Poirson, S. Yang, A. C. Berg, 和 T. L. Berg, “指称表达中的上下文建模”，发表于2016年欧洲计算机视觉会议（ECCV 2016），荷兰阿姆斯特丹，2016年10月11-14日，论文集第二部分14，施普林格，2016年，第69-85页。

W. Yu, N. Gileadi, C. Fu, S. Kirmani, K.-H. Lee, M. G. Arenas, H.-T. L. Chiang, T. Erez, L. Hasenclever, J. Humplik et al., "Language to rewards for robotic skill synthesis," arXiv preprint arXiv:2306.08647, 2023.

W. Yu, N. Gileadi, C. Fu, S. Kirmani, K.-H. Lee, M. G. Arenas, H.-T. L. Chiang, T. Erez, L. Hasenclever, J. Humplik 等, “从语言到奖励的机器人技能合成”，arXiv预印本 arXiv:2306.08647, 2023年。

R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "Merlot: Multimodal neural script knowledge models," 2021.

R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, 和 Y. Choi, “Merlot：多模态神经脚本知识模型”，2021年。

R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and Y. Choi, "Merlot reserve: Neural script knowledge through vision and language and sound," 2022.

R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, 和 Y. Choi, “Merlot reserve：通过视觉、语言和声音获取神经脚本知识，”2022年。

A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani et al., "Transporter networks: Rearranging the visual world for robotic manipulation," in Conference on Robot Learning. PMLR, 2021, pp. 726-747.

A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani 等, “Transporter网络：为机器人操作重新排列视觉世界，”发表于机器人学习会议。PMLR, 2021年，第726-747页。

A. Zeng, M. Liu, R. Lu, B. Wang, X. Liu, Y. Dong, and J. Tang, "Agenttuning: Enabling generalized agent abilities for llms," 2023.

A. Zeng, M. Liu, R. Lu, B. Wang, X. Liu, Y. Dong, 和 J. Tang, “Agenttuning：赋能大语言模型（LLMs）的通用代理能力，”2023年。

L. Zha, Y. Cui, L.-H. Lin, M. Kwon, M. G. Arenas, A. Zeng, F. Xia, and D. Sadigh, "Distilling and retrieving generalizable knowledge for robot manipulation via language corrections," arXiv preprint arXiv:2311.10678, 2023.
L. Zha, Y. Cui, L.-H. Lin, M. Kwon, M. G. Arenas, A. Zeng, F. Xia, 和 D. Sadigh, “通过语言纠正蒸馏和检索机器人操作的可泛化知识,”arXiv预印本 arXiv:2311.10678, 2023年。

M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "Motiondiffuse: Text-driven human motion generation with diffusion model," arXiv preprint arXiv:2208.15001, 2022.

M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, 和 Z. Liu, “Motiondiffuse：基于扩散模型的文本驱动人体动作生成，”arXiv预印本 arXiv:2208.15001, 2022年。

S. Zhang, X. Song, Y. Bai, W. Li, Y. Chu, and S. Jiang, "Hierarchical object-to-zone graph for object navigation," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 15130-15140.

S. Zhang, X. Song, Y. Bai, W. Li, Y. Chu, 和 S. Jiang, “用于物体导航的分层物体到区域图，”发表于IEEE/CVF国际计算机视觉会议论文集，2021年，第15130-15140页。

W. Zhao, Y. Zhao, X. Lu, S. Wang, Y. Tong, and B. Qin, "Is chatgpt equipped with emotional dialogue capabilities?" arXiv preprint arXiv:2304.09582, 2023.

W. Zhao, Y. Zhao, X. Lu, S. Wang, Y. Tong, 和 B. Qin, “ChatGPT具备情感对话能力吗？”arXiv预印本 arXiv:2304.09582, 2023年。

Y. Zhao, I. Misra, P. Krähenbühl, and R. Girdhar, "Learning video representations from large language models," in arXiv preprint arXiv:2212.04501, 2022.

Y. Zhao, I. Misra, P. Krähenbühl, 和 R. Girdhar, “从大型语言模型学习视频表示，”arXiv预印本 arXiv:2212.04501, 2022年。

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, 和 I. Stoica, “用MT-Bench和Chatbot Arena评测大语言模型作为评审的能力，”2023年。

Z. Zhong, J. Cao, S. Gu, S. Xie, W. Gao, L. Luo, Z. Yan, H. Zhao, and G. Zhou, "Assist: Interactive scene nodes for scalable and realistic indoor simulation," arXiv preprint arXiv:2311.06211, 2023.

Z. Zhong, J. Cao, S. Gu, S. Xie, W. Gao, L. Luo, Z. Yan, H. Zhao, 和 G. Zhou, “Assist: 用于可扩展且逼真的室内仿真的交互式场景节点，” arXiv预印本 arXiv:2311.06211, 2023.

G. Zhou, Y. Hong, and Q. Wu, "Navgpt: Explicit reasoning in vision-and-language navigation with large language models," arXiv preprint arXiv:2305.16986, 2023.

G. Zhou, Y. Hong, 和 Q. Wu, “Navgpt: 利用大型语言模型在视觉与语言导航中的显式推理，” arXiv预印本 arXiv:2305.16986, 2023.

H. Zhou, M. Ding, W. Peng, M. Tomizuka, L. Shao, and C. Gan, "Generalizable long-horizon manipulations with large language models," arXiv preprint arXiv:2310.02264, 2023.

H. Zhou, M. Ding, W. Peng, M. Tomizuka, L. Shao, 和 C. Gan, “基于大型语言模型的可泛化长时域操作，” arXiv预印本 arXiv:2310.02264, 2023.

X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in ECCV, 2022.

X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, 和 I. Misra, “利用图像级监督检测两万类目标，”载于ECCV, 2022.

Y. Zhou, C. Cui, J. Yoon, L. Zhang, Z. Deng, C. Finn, M. Bansal, and H. Yao, "Analyzing and mitigating object hallucination in large vision-language models," arXiv preprint arXiv:2310.00754, 2023.

Y. Zhou, C. Cui, J. Yoon, L. Zhang, Z. Deng, C. Finn, M. Bansal, 和 H. Yao, “分析与缓解大型视觉语言模型中的目标幻觉，” arXiv预印本 arXiv:2310.00754, 2023.

D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," 2023.

D. Zhu, J. Chen, X. Shen, X. Li, 和 M. Elhoseiny, “Minigpt-4: 利用先进大型语言模型增强视觉语言理解,” 2023.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223-2232.
J.-Y. Zhu, T. Park, P. Isola, 和 A. A. Efros, “基于循环一致性对抗网络的无配对图像到图像转换,” 载于IEEE国际计算机视觉会议论文集, 2017, 页2223-2232.

S. Zhu, A. Kimmel, K. E. Bekris, and A. Boualiaris, "Fast model identification via physics engines for data-efficient policy search," arXiv preprint arXiv:1710.08893, 2017.

S. Zhu, A. Kimmel, K. E. Bekris, 和 A. Boualiaris, “通过物理引擎快速模型识别以实现数据高效的策略搜索,” arXiv 预印本 arXiv:1710.08893, 2017.

X. Zhu, J. Wang, L. Zhang, Y. Zhang, R. Gan, J. Zhang, and Y. Yang, "Solving math word problem via cooperative reasoning induced language models," arXiv preprint arXiv:2210.16257, 2022.

X. Zhu, J. Wang, L. Zhang, Y. Zhang, R. Gan, J. Zhang, 和 Y. Yang, “通过协作推理引导的语言模型解决数学应用题,” arXiv预印本 arXiv:2210.16257, 2022.

Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in Robotics and Automation (ICRA), 2017 IEEE International Conference on. IEEE, 2017, pp. 3357-3364.

Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, 和 A. Farhadi, “基于深度强化学习的室内场景目标驱动视觉导航,” 载于2017年IEEE国际机器人与自动化会议(ICRA), IEEE, 2017, 页3357-3364.

Appendix for

附录

Agent AI

智能体AI

35 A GPT-4V Agent Prompt Details

36 A GPT-4V智能体提示细节

Unless specified, we use the default system prompt for GPT-4V. We show a detailed description of the process to prompt GPT-4V for Minecraft and how we generate GPT-4V's responses in Fig. 31.

除非另有说明，我们使用GPT-4V的默认系统提示。图31展示了针对Minecraft提示GPT-4V的详细过程及我们如何生成GPT-4V的响应。

37 B GPT-4V for Bleeding Edge

38 B GPT-4V在Bleeding Edge中的应用

Bleeding Edge is a third person team-based combat game where players attempt to capture objective points or collect more resources than the enemy team. We show an example input and output when prompting GPT-4V for the game Bleeding Edge in Fig. 32. Compared to Minecraft, we qualitatively find that GPT-4V has a less thorough understanding of the visual content and game rules. This is likely due to the (1) the amount of minecraft data present in GPT-4V's training data and (2) the visual complexity of Bleeding Edge compared to Minecraft.

Bleeding Edge是一款第三人称团队战斗游戏，玩家试图占领目标点或收集比敌队更多的资源。图32展示了针对Bleeding Edge游戏提示GPT-4V的示例输入和输出。与Minecraft相比，我们定性发现GPT-4V对视觉内容和游戏规

则的理解较为浅显。这可能是由于(1) GPT-4V训练数据中Minecraft数据量较大，(2) Bleeding Edge的视觉复杂度高于Minecraft。

39 C GPT-4V for Microsoft Fight Simulator

40 C GPT-4V 用于微软飞行模拟器

As shown in Fig. 33, a GPT-4V based agent can provide the high-level action description for the player in Microsoft Flight Simulator. It describes how the player is in the process of flying an aircraft, shown by the cockpit perspective and external views of the plane, managing various flight controls and instruments to maintain appropriate airspeed and altitude while navigating through the virtual airspace.

如图33所示，基于GPT-4V的代理可以为微软飞行模拟器中的玩家提供高级动作描述。它描述了玩家正在驾驶飞机的过程，通过驾驶舱视角和飞机外部视图展示，操作各种飞行控制和仪表，以维持适当的空速和高度，同时在虚拟空域中导航。

41 D GPT-4V for Assassin's Creed Odyssey

42 D GPT-4V 用于刺客信条：奥德赛

As shown in Fig. 34, the GPT-4V agent provides the high-level action description of the character in the images being engaged in nighttime combat during a historical battle, wielding a spear and utilizing special abilities, such as the Bull Rush, to fight against multiple enemies marked with level indicators, while managing their health and ability meters in a third-person action role-playing setting. GPT-4V also describes that the player is engaged in a large-scale melee combat scenario, controlling a character with red accents on their armor, and is prompted to use a "Bull Rush Ability" as indicated by the on-screen instructions. The environment is dark and fiery, suggesting a battlefield, and the player's character is facing multiple enemies marked with level indicators and health bars overhead. The action is part of a tutorial or combat sequence where the player learns and executes combat abilities.

如图34所示，GPT-4V代理提供了图中角色在历史战斗中夜间战斗的高级动作描述，角色挥舞长矛并使用特殊技能，如“猛牛冲锋”（Bull Rush），与多个带有等级标记的敌人作战，同时管理其生命值和技能值，场景为第三人称动作角色扮演。GPT-4V还描述玩家正参与一场大规模近战战斗，控制一名盔甲带红色装饰的角色，并根据屏幕指示使用“猛牛冲锋”技能。环境昏暗且充满火光，暗示战场，玩家角色面对多个带有等级标记和生命条的敌人。该动作是教程或战斗序列的一部分，玩家在此学习并执行战斗技能。

43 E GPT-4V for GEARS of WAR 4

44 E GPT-4V 用于战争机器4

As shown in Fig. 35, Based on the sequence of 10 frames of video provided, the GPT-4V agent can provide a general description of the player's actions in the video game:

如图35所示，基于提供的10帧视频序列，GPT-4V代理可以对视频游戏中玩家的动作做出总体描述：

"The player is navigating a character through an industrial environment at night, taking strategic cover behind various objects like crates and walls. The character is armed and engages in combat with enemy entities. The player uses firearms and possibly melee attacks, as indicated by close-quarters combat. They appear to manage their ammunition and health, indicated by visual cues such as a flashing red indicator on the screen, which often signifies damage to the character."

“玩家正在夜间的工业环境中操控角色，策略性地利用箱子和墙壁等物体作为掩护。角色持有武器并与敌方实体交战。玩家使用火器，可能还有近战攻击，表现为近距离战斗。通过屏幕上闪烁的红色指示器等视觉提示，显示角色正在管理弹药和生命值，红色闪烁通常表示角色受伤。”

Additionally, the player seems to be moving tactically, shifting between different cover positions to advance or defend against the enemy. There is evidence of a protective shield being activated at some point, which suggests the player is responding to incoming fire by using defensive abilities. The character also seems to be reloading their weapon, suggesting an element of resource management during combat. Blood and impact effects indicate successful hits on the enemies.

此外，玩家似乎在战术性移动，在不同掩护点之间切换以推进或防御敌人。有证据表明某时刻激活了防护盾，表明玩家通过防御技能应对来袭火力。角色还在重新装填武器，显示战斗中资源管理的元素。血迹和冲击效果表明对敌人造成了有效打击。

The player's character collaborates with allies, as other characters are seen in the vicinity, engaging in similar combat actions. The environment has interactive elements, such as exploding barrels, which the player might use to their advantage.

玩家的角色与盟友协同作战，附近可见其他角色参与类似战斗行动。环境中存在可互动元素，如爆炸桶，玩家可能利用这些元素取得优势。

Figure 31: We provide code snippets to show the complete prompting process for calling GPT-4V for Minecraft videos. First, we encode the frames of the video and then call the GPT-4V endpoint. The model's response is shown on the bottom of the screen.

图31：我们提供代码片段展示调用GPT-4V处理Minecraft视频的完整提示过程。首先对视频帧进行编码，然后调用

GPT-4V接口。模型的响应显示在屏幕底部。

45 GPT-4-Vision

46 GPT-4视觉

47 Prompt:

48 提示:

These are frames of a video of a Bleeding Edge player ordered from left to right and top to bottom as a grid. Give a simple, but precise description of what the player is doing in 1 sentence. Be specific about important items, entities, and actions. In your description do not mention specific frame numbers or the name of the game.

这些是Bleeding Edge玩家视频的帧，按从左到右、从上到下的网格顺序排列。请用一句话简洁而准确地描述玩家的动作。具体说明重要物品、实体和动作。描述中不要提及具体帧号或游戏名称。

49 Video input:

50 视频输入:



51 Output:

52 输出:

The player begins by running around the map, passing through different checkpoints and interacting with several capture points, then fights against an enemy player, and finally captures an objective while being attacked by another enemy.

玩家开始在地图上奔跑，经过多个检查点并与几个占领点互动，随后与敌方玩家战斗，最终在被另一名敌人攻击的同时占领目标。

Figure 32: GPT-4V is able to produce meaningful content for games like Bleeding Edge that have 3rd person viewpoints and visually complex scenes. In order to input a large number of frames (48) to GPT-4V, we input the frames as a grid with frame numbers overlaid on each frame (as shown above).

图32: GPT-4V能够为像《Bleeding Edge》这样具有第三人称视角和视觉复杂场景的游戏生成有意义的内容。为了向GPT-4V输入大量帧（48帧），我们将帧以网格形式输入，并在每帧上叠加帧编号（如上图所示）。

53 GPT-4-Vision

54 GPT-4视觉

55 Prompt:

56 提示:

These are frames of a video of a Flight Simulator gaming player ordered from left to right and top to bottom. Give a simple, but precise description of what the player is doing in 1 sentence. Be specific about important items, entities, and actions. In your description do not mention specific frame numbers or the name of the game.

这些是一个飞行模拟游戏玩家的视频帧，按从左到右、从上到下的顺序排列。请用一句简洁但准确的话描述玩家正在做什么。具体说明重要的物品、实体和动作。描述中不要提及具体帧号或游戏名称。

57 Video input:

58 视频输入:



Output: In the images from the flight simulator game, the player is piloting an aircraft, taking off from an airstrip over water, adjusting the cockpit controls, and navigating through the game's interface to manage the flight, while dealing with an alert about low airspeed and the need to increase it to avoid a stall.

输出：在飞行模拟游戏的图像中，玩家正在驾驶飞机，从水面上的跑道起飞，调整驾驶舱控制装置，并通过游戏界面管理飞行，同时应对低空速警报，需要增加速度以避免失速。

Figure 33: GPT-4V is able to produce meaningful content description for Microsoft Flight Simulator games. The GPT-4V agent can provide a high-level action description as the player in the flight simulator is in the process of flying an aircraft, shown by the cockpit perspective and external views of the plane, managing various flight controls and instruments to maintain appropriate airspeed and altitude while navigating through the virtual airspace. In order to input a large video, we select a number of key frames (6) to send to GPT-4V. We input each key frame individually and do not use a grid (as shown above).

图33：GPT-4V能够为微软飞行模拟游戏生成有意义的内容描述。基于GPT-4V的代理可以提供高级动作描述，玩家正在驾驶飞机，画面显示驾驶舱视角和飞机外部视图，管理各种飞行控制和仪表，以保持适当的空速和高度，同时在虚拟空域中导航。为了输入较长视频，我们选择若干关键帧（6帧）发送给GPT-4V。我们逐帧单独输入，不使用网格（如上图所示）。

59 GPT-4-Vision

60 GPT-4视觉

61 Prompt:

62 提示：

These are frames of a video of an ASSASSIN's CREED ODYSSEY gaming player ordered from left to right and top to bottom. Give a simple, but precise description of what the player is doing. Be specific about important items, entities, and actions. In your description do not mention specific frame numbers or the name of the game.

这些是《刺客信条：奥德赛》游戏玩家的视频帧，按从左到右、从上到下的顺序排列。请用一句简洁但准确的话描述玩家正在做什么。具体说明重要的物品、实体和动作。描述中不要提及具体帧号或游戏名称。

63 Video input:

64 视频输入：

Output:

输出：



tput:

The player is engaged in a large-scale melee combat scenario, controlling a character with red accents on their armor, and is prompted to use a "Bull Rush Ability" as indicated by the on-screen instructions. The environment is dark and fiery, suggesting a battlefield, and the player's character is facing multiple enemies marked with level indicators and health bars overhead. The action is part of a tutorial or combat sequence where the player learns and executes combat abilities.

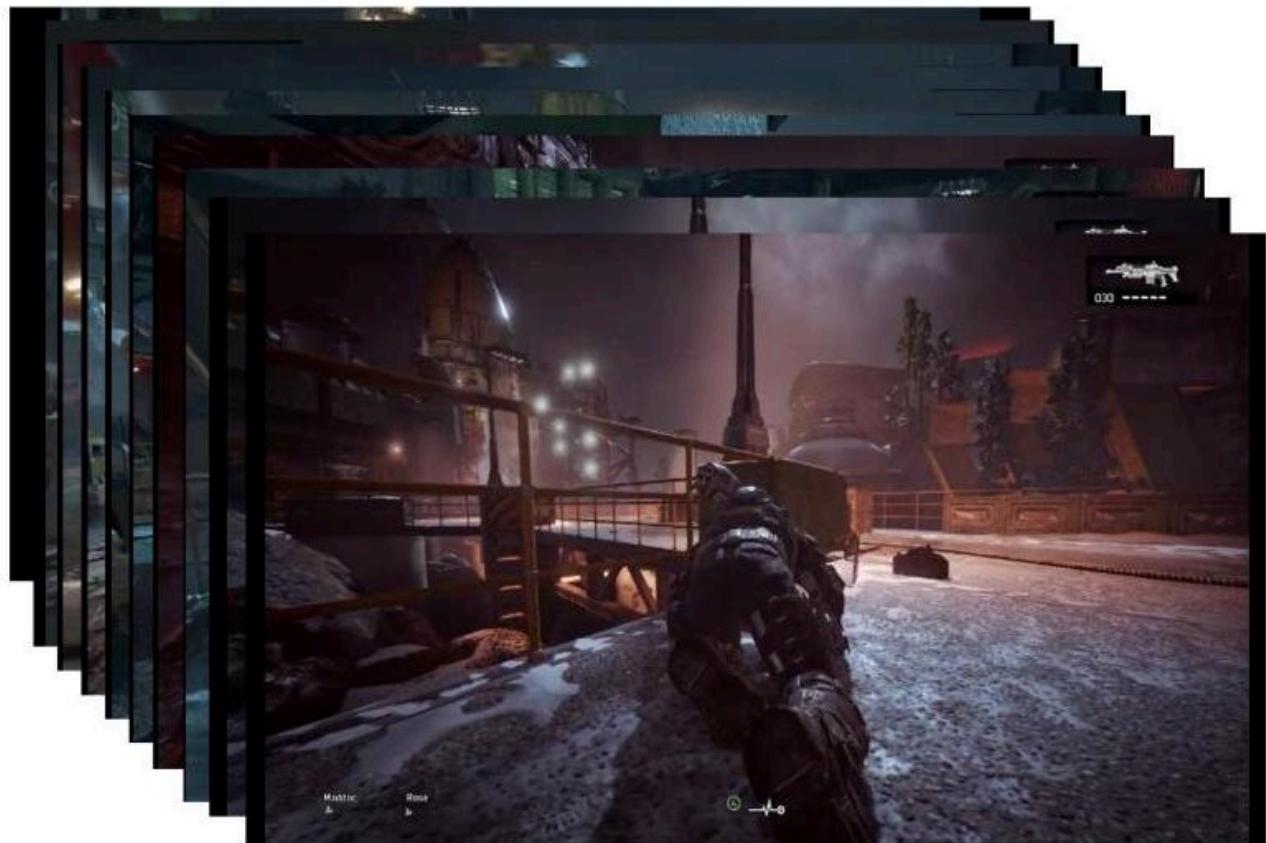
玩家正参与一场大规模近战战斗，控制一名盔甲带有红色装饰的角色，屏幕提示使用“冲锋技能（Bull Rush Ability）”。环境昏暗且充满火光，暗示战场场景，玩家角色面对多名带有等级标记和生命值条的敌人。该动作属于教程或战斗序列，玩家正在学习并执行战斗技能。

Figure 34: GPT-4V is able to produce meaningful content description for the game, Assassin's Creed Odyssey. The GPT-4V-based agent can also provide a high-level action description of "the character in the images [being] engaged in nighttime combat during a historical battle, wielding a spear and utilizing special abilities, such as the Bull Rush, to fight against multiple enemies marked with level indicators, while managing their health and ability meters in a third-person action role-playing setting." In order to input a large video, we select a number of key frames (8) to send to GPT-4V. We input each key frame individually and do not use a grid (as shown above). GPT-4-Vision Prompt: These are frames of a video of the GEARS of WAR 4 gaming player ordered from left to right and top to bottom. Give a simple, but precise description of what the player is doing. Be specific about important items, entities, and actions. In your description do not mention specific frame numbers or the name of the game.

图34：GPT-4V能够为游戏《刺客信条：奥德赛》生成有意义的内容描述。基于GPT-4V的代理还能提供高级动作描述：“图中角色正参与历史战役中的夜间战斗，挥舞长矛并使用诸如冲锋（Bull Rush）等特殊技能，对抗多名带有等级标记的敌人，同时管理其生命值和技能条，处于第三人称动作角色扮演环境中。”为了输入较长视频，我们选择若干关键帧（8帧）发送给GPT-4V。我们逐帧单独输入，不使用网格（如上图所示）。GPT-4视觉提示：这些是《战争机器4》游戏玩家的视频帧，按从左到右、从上到下的顺序排列。请用一句简洁但准确的话描述玩家正在做什么。具体说明重要的物品、实体和动作。描述中不要提及具体帧号或游戏名称。

1 | Video input:

2 | 视频输入：



1 | Based on the sequence of images provided, here is a general
2 | 基于所提供的图像序列，以下是一个总体描述

3 |

4 | description of the player's actions in a video game:
5 | 对玩家在视频游戏中动作的描述：

6 |

7 | The player is navigating a character through an industrial environment
8 | 玩家正在操控角色穿越工业环境

9 |

10 | at night, taking strategic cover behind various objects like crates
11 | 夜晚，利用箱子等各种物体作为战略掩护

12 |

13 | and walls. The character is armed and engages in combat with enemy
14 | 和墙壁。角色装备武器，与敌人交战

15 |

16 | entities. The player uses firearms and possibly melee attacks, as
17 | 实体。玩家使用火器，可能还会进行近战攻击，正如

18 |

19 indicated by close-quarters combat. They appear to manage their
20 通过近身格斗表现出来。他们似乎能够控制他们的
21
22 ammunition and health, indicated by visual cues such as a flashing red
23 弹药和生命值，通过闪烁的红色等视觉提示表示
24
25 indicator on the screen, which often signifies damage to the
26 屏幕上的指示灯，通常表示损坏
27
28 character.
29 字符。
30
31 Additionally, the player seems to be moving tactically, shifting
32 此外，玩家似乎在战术性移动，正在转移
33
34 between different cover positions to advance or defend against the
35 在不同掩护位置之间移动以推进或防御
36
37 enemy. There is evidence of a protective shield being activated at
38 敌人。有证据表明保护盾已被激活于
39
40 some point, which suggests the player is responding to incoming fire
41 某个点，表明玩家正在应对来袭火力
42
43 by using defensive abilities. The character also seems to be reloading
44 通过使用防御技能。角色似乎也在重新装填
45
46 their weapon, suggesting an element of resource management during
47 他们的武器，暗示了在过程中存在资源管理的因素
48
49 combat. Blood and impact effects indicate successful hits on the
50 战斗。血液和冲击效果表明击中了目标
51
52 enemies.
53 敌人。
54
55 The player's character collaborates with allies, as other characters
56 玩家的角色与盟友合作，就像其他角色一样
57
58 are seen in the vicinity, engaging in similar combat actions. The
59 在附近可以看到，正在进行类似的战斗动作。
60
61 environment has interactive elements, such as exploding barrels, which
62 环境中有交互元素，例如会爆炸的桶，
63
64 the player might use to their advantage.
65 玩家可能会利用的优势。
66
67 The scenes depict fast-paced action with a focus on strategic
68 这些场景描绘了节奏快速且注重策略的动作
69
70 movement, shooting, and situational awareness to overcome adversaries
71 通过移动、射击和态势感知来战胜对手
72
73 and possibly achieve a specific objective within the game.
74 并可能在游戏中实现特定目标。

Figure 35: GPT-4V is able to produce meaningful content description for GEARS of WAR 4 games. The GPT-4V agent provides the high-level action description of the player being engaged in a combat scenario within a visually rich, futuristic environment, reminiscent of industrial or military complexes. The character, armored and equipped with a firearm, is navigating through this area, using cover and aiming systems to engage with enemies. Actions captured include the player character aiming from behind cover, firing upon enemies, taking damage as indicated by visual cues on the screen, and performing close-quarters combat. The environment is marked by night-time lighting with artificial sources highlighting areas of interest or action. There are also moments where the player's character is seen in a transparent, outlined form, which could denote a special ability or a temporary state like invincibility or stealth. In order to input a large video, we select a number of key frames (10) to send to GPT-4V. We input each key frame individually and do not use a grid (as shown above).

图35：GPT-4V能够为《战争机器4》（GEARS of WAR 4）游戏生成有意义的内容描述。GPT-4V代理提供了玩家在视觉丰富的未来环境中参与战斗场景的高级动作描述，该环境类似于工业或军事综合体。角色身穿盔甲，配备火器，正在穿越该区域，利用掩体和瞄准系统与敌人交战。捕捉到的动作包括玩家角色在掩体后瞄准、向敌人开火、屏幕上的视觉提示显示受到伤害，以及进行近身格斗。环境以夜间照明为特征，人工光源突出显示关注或行动区域。还有一些时刻玩家角色以透明轮廓形式出现，可能表示特殊能力或临时状态，如无敌或隐身。为了输入大型视频，我们选择若干关键帧（10帧）发送给GPT-4V。我们逐帧输入每个关键帧，且不使用网格（如上所示）。

The scenes depict fast-paced action with a focus on strategic movement, shooting, and situational awareness to overcome adversaries and possibly achieve a specific objective within the game."

这些场景描绘了快节奏的动作，强调战略性移动、射击和情境意识，以击败敌人并可能在游戏中达成特定目标。

65 F GPT-4V for Starfield

66 F GPT-4V 用于星域（Starfield）

As showed in Fig. 36, based on the sequence of 10 frames of video provided for Starfield, the GPT-4V agent provides a general description of the player's actions in a video game. According to GPT-4V: "The images depict a series of events within the game Starfield, showcasing the player's interactions within a mining environment. Here's a description of the sequence:

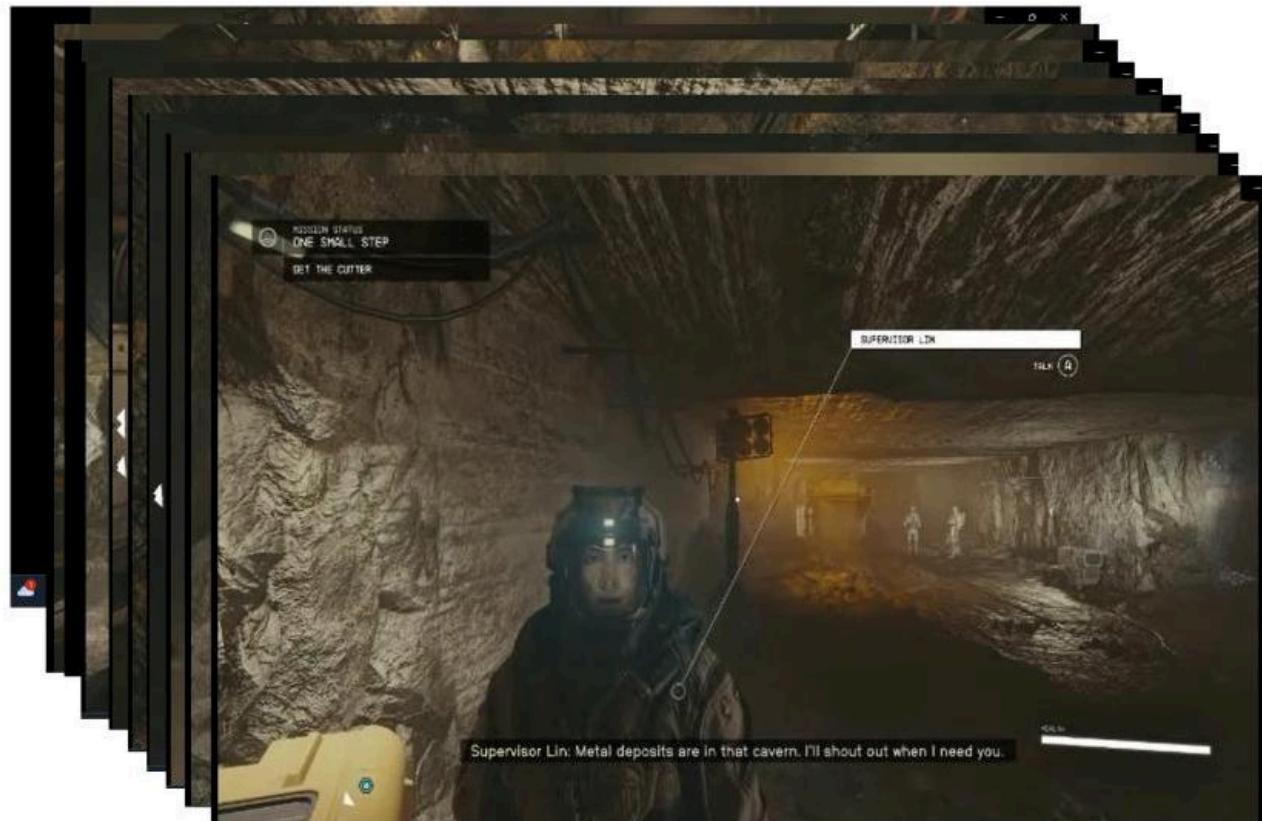
如图36所示，基于提供的Starfield游戏的10帧视频序列，GPT-4V代理对玩家在视频游戏中的动作进行了总体描述。根据GPT-4V的说法：“这些图像展示了游戏Starfield中的一系列事件，呈现了玩家在采矿环境中的互动。以下是该序列的描述：

1. The player stands in a dimly lit underground cavern with stalactites, accompanied by another character in similar space gear. They face a highlighted waypoint, indicating an objective or location of interest.
2. 玩家站在一个昏暗的地下洞穴中，洞顶垂挂着钟乳石，身边有另一位穿着类似太空装备的角色。他们面向一个高亮的航点，指示着一个目标或感兴趣的位置。
2. Moving through an industrial interior, likely a mining station, the player approaches another character. The HUD (Heads-Up Display) shows a dialogue line from "Supervisor Lin," which suggests NPC (non-player character) interaction and possibly a mission briefing or update.
3. 穿过一个工业内部空间，可能是一个采矿站，玩家接近另一个角色。HUD（抬头显示器）显示了“林主管”的对话内容，暗示了NPC（非玩家角色）互动，可能是任务简报或更新。
3. The player seems to be following the NPC, Supervisor Lin, through the cavern, hinting at a guided segment of gameplay where the NPC leads the player to a specific location or task.

4. 玩家似乎正在跟随NPC林主管穿过洞穴，暗示这是一个由NPC引导玩家前往特定地点或完成任务的游戏环节。
 4. A dialogue from Supervisor Lin suggests a decision point or mission update, possibly directing the player to a new objective related to a "big vein" of resources.
 - 4) 林主管的对话暗示了一个决策点或任务更新，可能指引玩家前往与“大矿脉”相关的新目标。
5. The player encounters Supervisor Lin again, who provides instruction or information about metal deposits, implying the start of a mining task or exploration of a new area.
6. 玩家再次遇到林主管，他提供有关金属矿床的指导或信息，暗示采矿任务或新区域探索的开始。
6. In this frame, the player appears to be observing or interacting with a piece of machinery, with another character, Isabelle, indicating an action or dialogue related to equipment operation or technical work.
7. 在这一画面中，玩家似乎正在观察或操作一台机械设备，另一角色伊莎贝尔（Isabelle）则指示与设备操作或技术工作相关的动作或对话。
7. The player looks down at a rough terrain with two glowing objects, which could represent resources to be collected or points of interest for the mission.
 8. 玩家俯视一片崎岖地形，上面有两个发光物体，可能代表可收集的资源或任务中的兴趣点。
8. The player examines a specific resource, indicated by a HUD marker as "Beryllium Deposit," suggesting a resource gathering or mining activity.
 9. 玩家检查一个特定资源，HUD标记为“铍矿床（Beryllium Deposit）”，表明正在进行资源采集或采矿活动。
9. The player traverses through the caverns, potentially searching for additional resources or navigating to the next objective, as indicated by another character in space gear ahead.
 10. 玩家穿越洞穴，可能在寻找更多资源或前往下一个目标，前方有另一位穿着太空装备的角色指引方向。
10. The player is positioned in a darker area of the cavern, with two glowing markers indicating a path or objective points to be reached.
 11. 玩家位于洞穴较暗区域，两个发光标记指示需到达的路径或目标点。

Throughout these images, the player seems to be engaged in a mission involving exploration, NPC interaction, resource collection, and possibly machinery operation within a mining environment. The HUD elements like waypoints, dialogue text, and mission updates provide context to the player's actions and the game's objectives."在这些画面中，玩家似乎正执行一项涉及探索、NPC互动、资源收集及可能的机械操作的任务，HUD元素如路径点、对话文本和任务更新为玩家的行动和游戏目标提供了背景信息。

1 GPT-4-Vision
2 GPT-4-Vision
3
4 These are frames of a video of the Starfield Console game player ordered from left
to right and top to bottom. Give a simple, but
5 这些是按从左到右、从上到下顺序排列的Starfield游戏玩家视频帧。请给出简洁但
6
7 precise description of what the player is doing. Be specific about important items,
entities, and actions. In your description do not
8 准确的描述，说明玩家正在做什么。具体指出重要物品、实体和动作。描述中不要
9
10 mention specific frame numbers or the name of the game.
11 提及具体帧号或游戏名称。



1 Output:
2 输出：
3
4 The images depict a series of events within the game Starfield, showcasing the
5 这些图像展示了Starfield游戏中一系列事件，体现了
6
7 player's interactions within a mining environment. Here's a description of the
8 玩家在采矿环境中的互动。以下是该
9
10 sequence:
11 序列的描述：

12
13 1.The player stands in a dimly lit underground cavern with stalactites,
14 accompanied
15 1.玩家站在一个昏暗的地下洞穴中，洞顶挂满钟乳石，身边有同伴
16 by another character in similar space gear. They face a highlighted waypoint,
17 由另一位穿着类似太空服的角色。他们面对一个高亮的航点，
18
19 indicating an objective or location of interest.
20 表示一个目标或感兴趣的位置。
21
22 2.Moving through an industrial interior, likely a mining station, the player
23 穿过一个工业内部空间，可能是一个采矿站，玩家
24
25 approaches another character. The HUD (Heads-Up Display) shows a dialogue line
from
26 接近另一个角色。HUD（抬头显示器）显示一行对话
27
28 "Supervisor Lin," which suggests NPC (non-player character) interaction and
possibly
29 "林主管"，这暗示了NPC（非玩家角色）的互动，可能还包括
30
31 a mission briefing or update.
32 任务简报或更新。
33
34 3. The player seems to be following the NPC, Supervisor Lin, through the cavern,
35 3. 玩家似乎正在跟随NPC林主管穿过洞穴，
36
37 hinting at a guided segment of gameplay where the NPC leads the player to a
specific
38 暗示一个由NPC引导玩家到特定位置的游戏环节
39
40 location or task.
41 位置或任务。
42
43 4.A dialogue from Supervisor Lin suggests a decision point or mission update,
44 4.林主管的对话暗示了一个决策点或任务更新，
45
46 possibly directing the player to a new objective related to a "big vein" of
47 可能引导玩家前往与"大矿脉"(big vein)相关的新目标
48
49 resources.
50 资源。
51
52 5.The player encounters Supervisor Lin again, who provides instruction or
53 5.玩家再次遇到林主管，他提供指导或
54
55 information about metal deposits, implying the start of a mining task or
exploration
56 关于金属矿床的信息，暗示采矿任务或勘探的开始
57
58 of a new area.
59 新领域的。
60
61 6.In this frame, the player appears to be observing or interacting with a piece of
62 6.在这一帧中，玩家似乎正在观察或与一件物品互动

63
64 machinery, with another character, Isabelle, indicating an action or dialogue
65 机械装置，另有一名角色伊莎贝尔（Isabelle）表示一个动作或对话
66
67 related to equipment operation or technical work.
68 与设备操作或技术工作相关。
69
70 7. The player looks down at a rough terrain with two glowing objects, which could
71 7. 玩家俯视一片崎岖地形，看到两个发光物体，可能代表可收集的资源或任务关注点。
72
73 represent resources to be collected or points of interest for the mission.
74 代表需要收集的资源或任务的兴趣点。
75
76 8. The player examines a specific resource, indicated by a HUD marker as
77 "Beryllium
78 8. 玩家检查一个特定资源，HUD标记为"铍矿床（Beryllium Deposit）"，暗示资源采集或采矿活
动。
79
80 Deposit," suggesting a resource gathering or mining activity.
81 暗示资源采集或采矿活动。
82
83 9. The player traverses through the caverns, potentially searching for additional
9. 玩家穿越洞穴，可能在寻找更多资源或前往下一个目标，前方有另一名穿着太空装备的角色指
引。
84
85 resources or navigating to the next objective, as indicated by another character
in 资源或导航至下一个目标，前方有另一名穿着太空装备的角色。
86
87 space gear ahead.
88 太空装备的角色在前方。
89
90 10. The player is positioned in a darker area of the cavern, with two glowing
91 markers
92 10. 玩家位于洞穴较暗区域，两个发光标记指示路径或需到达的目标点。
93
94 indicating a path or objective points to be reached.
95 指示路径或需到达的目标点。
96
97 Throughout these images, the player seems to be engaged in a mission involving
98 在这些画面中，玩家似乎正在执行一项涉及探索、NPC互动、资源收集及可能的机械操作的任务，场
景为采矿环境。
99
100 exploration, NPC interaction, resource collection, and possibly machinery
operation
101 探索、NPC互动、资源收集及可能的机械操作的任务，场景为采矿环境。
102
103 within a mining environment. The HUD elements like waypoints, dialogue text, and
104 HUD元素如航点、对话文本和任务更新为玩家的行动及游戏目标提供了背景信息。
105
106 mission updates provide context to the player's actions and the game's objectives.
107 为玩家的行动及游戏目标提供了背景信息。
108
109 Figure 36: GPT-4V is able to produce meaningful content description for the X-Box
110 console game Starfield. The
110 图36：GPT-4V能够为X-Box游戏主机游戏《星空》（Starfield）生成有意义的内容描述。

111
112 GPT-4V agent provides a high-level action description for the space-themed exploration
113 game. In order to input a large
114 GPT-4V代理为太空主题的探索游戏提供高级动作描述。为了输入大型
115 video, we select a number of key frames (10) to send to GPT-4V. We input each key
116 frame individually and do not use a
117 视频, 我们选择若干关键帧 (10帧) 发送给GPT-4V。我们逐帧输入每个关键帧, 不使用
118 grid (as shown above).
119 网格 (如上所示)。

67 Author Biographies

68 作者简介

Zane Durante PhD Student in Computer Science at Stanford University
Zane Durante 斯坦福大学计算机科学博士生

Email: durante@stanford.edu

邮箱: durante@stanford.edu

Web: <https://zanedurante.github.io/>

主页: <https://zanedurante.github.io/>

Zane Durante is a PhD student at Stanford Vision Lab advised by Dr. Fei-Fei Li. His research interests include video understanding, connecting computer vision with natural language, and AI applications in hospital care. He is leading an ongoing collaboration between Stanford Medical School and Stanford Vision Lab to develop a dataset of natural language descriptions of video clips of hospital scenes. His PhD work is supported by NSF's Graduate Research Fellowship.

Zane Durante是斯坦福视觉实验室的博士生, 导师为李飞飞博士。他的研究兴趣包括视频理解、将计算机视觉与自然语言连接, 以及人工智能在医院护理中的应用。他正领导斯坦福医学院与斯坦福视觉实验室之间的合作, 开发医院场景视频片段的自然语言描述数据集。他的博士研究由美国国家科学基金会 (NSF) 研究生奖学金支持。

Qiuyuan Huang Principal Researcher at Microsoft Research, Redmond, WA, USA.

Qiuyuan Huang 微软研究院首席研究员, 美国华盛顿州雷德蒙德

Email: qihua@microsoft.com

邮箱: qihua@microsoft.com

Web: <https://www.microsoft.com/en-us/research/people/qihua/>

主页: <https://www.microsoft.com/en-us/research/people/qihua/>

Qiuyuan Huang is a principal researcher in the deep learning group at Microsoft Research (MSR), Redmond, WA. Her current research interests are mainly in the deep learning, multi-modality, and natural language processing areas, specifically on Agent AI for Gaming, Robotics and Healthcare; Knowledge-reasoning Intelligence for Interactive AI; Neuro-symbolic Computation for Inference Reasoning; and Large Foundation models for NLP and Multi-modality.

Qiuyuan Huang是微软研究院 (MSR) 雷德蒙德深度学习组的首席研究员。她目前的研究兴趣主要集中在深度学

习、多模态和自然语言处理领域，具体包括游戏、机器人和医疗保健领域的智能代理AI；交互式AI的知识推理智能；用于推理的神经符号计算；以及面向自然语言处理和多模态的大型基础模型。

Naoki Wake Researcher at Microsoft, Redmond, WA, USA.

Naoki Wake 微软研究员，美国华盛顿州雷德蒙德

Email: naoki.wake@microsoft.com

邮箱: naoki.wake@microsoft.com

Web: <https://www.microsoft.com/en-us/research/people/nawake/>

主页: <https://www.microsoft.com/en-us/research/people/nawake/>

Naoki Wake is a researcher in the Applied Robotics Research group at Microsoft. His current research involves the development of multimodal perception systems for robots and co-speech gesturing systems. His past research has spanned auditory neuroscience, neuro-rehabilitation, and speech processing. Naoki received his B.S. degree in Engineering in 2014, and his Ph.D. in Information Science and Technology in 2019, both from the University of Tokyo.

Naoki Wake是微软应用机器人研究组的研究员。他目前的研究涉及机器人多模态感知系统和共语手势系统的开发。其过去的研究涵盖听觉神经科学、神经康复和语音处理。Naoki于2014年获得东京大学工程学士学位，2019年获得信息科学与技术博士学位。

Ran Gong PhD Student in Computer Science at University of California, Los Angeles.

Ran Gong, 加州大学洛杉矶分校计算机科学博士生。

Email: nikepupu@ucla.edu

电子邮件: nikepupu@ucla.edu

Web: <https://nikepupu.github.io>

网站: <https://nikepupu.github.io>

Ran Gong, is a PhD student at the UCLA VCLA Lab. His research lies in the intersection of Robotics, Computer Vision, Computer Graphics, and Machine Learning. His research focuses on embodied simulation and interaction with a goal of creating intelligent behaviors that can solve diverse tasks in diverse environments as well as the capability of collaborating with humans. He received his B.S. degree in Computer Science and Engineering at the University of California, Los Angeles.

Ran Gong是UCLA VCLA实验室的博士生。他的研究位于机器人学、计算机视觉、计算机图形学和机器学习的交叉领域。研究重点是具身仿真与交互，目标是创造能够在多样环境中解决多样任务的智能行为，以及具备与人类协作的能力。他获得了加州大学洛杉矶分校的计算机科学与工程学士学位。

Jae Sung Park PhD Student at University of Washington

Jae Sung Park, 华盛顿大学博士生

Email: jspark96@cs.washington.edu

电子邮件: jspark96@cs.washington.edu

Web: <https://homes.cs.washington.edu/~jspark96/>

网站: <https://homes.cs.washington.edu/~jspark96/>

Jae Sung is a PhD student advised by Yejin Choi and Ali Farhadi. His research focuses on developing models with multimodal commonsense reasoning. He is interested in equipping models with grounding linguistic concepts to visual modalities, and having them understand multimedia content in a way that humans process the visual information. Jae Sung received his B.S. degree in Computer Science at University of California, Berkeley.

Jae Sung是由Yejin Choi和Ali Farhadi指导的博士生。他的研究聚焦于开发具备多模态常识推理的模型。感兴趣于使

模型将语言概念与视觉模态相结合，并让它们以人类处理视觉信息的方式理解多媒体内容。Jae Sung获得了加州大学伯克利分校的计算机科学学士学位。

Bidipta Sarkar Undergraduate Student at Stanford University
Bidipta Sarkar, 斯坦福大学本科生

Email: bidiptas@stanford.edu
电子邮件: bidiptas@stanford.edu

Web: <https://bsarkar321.github.io/>
网站: <https://bsarkar321.github.io/>

Bidipta Sarkar is a senior undergraduate student at Stanford University and a member of Stanford's ILIAD lab. His research focuses on creating AI agents that can interact with their environment and safely work alongside humans and other autonomous agents.

Bidipta Sarkar是斯坦福大学的大四本科生，斯坦福ILIAD实验室成员。他的研究专注于创建能够与环境交互并能安全地与人类及其他自主代理协作的人工智能代理。

Rohan Taori PhD Student in Computer Science at Stanford University
Rohan Taori, 斯坦福大学计算机科学博士生

Email: rtaori@cs.stanford.edu
电子邮件: rtaori@cs.stanford.edu

Web: <https://www.rohantaori.com/> Rohan Taori is a PhD student at the Stanford AI Lab. His research focuses on studying the foundations of machine learning in the context of real-world systems. Most recently, Rohan has pushed forward the frontier of open-source large language models, finetuning them to be helpful, general-purpose assistants. Rohan is also very interested in augmenting language models with multi-modality capability to allow them to reason over images and videos. Rohan received his Bachelor's degree in Computer Science at UC Berkeley.

网站: <https://www.rohantaori.com/> Rohan Taori是斯坦福人工智能实验室的博士生。他的研究聚焦于在现实系统背景下研究机器学习的基础。最近，Rohan推动了开源大型语言模型的前沿，微调它们成为有用的通用助手。Rohan还非常感兴趣于增强语言模型的多模态能力，使其能够对图像和视频进行推理。Rohan获得了加州大学伯克利分校的计算机科学学士学位。

69 Yusuke Noda Principal Software Engineer. Microsoft Gaming, Redmond, WA.

Yusuke Noda, 微软游戏部门首席软件工程师，雷德蒙德，华盛顿州。

Email: yusuke.noda@microsoft.com
电子邮件: yusuke.noda@microsoft.com

Web: <https://www.linkedin.com/in/yusuke-noda-908797/>
网站: <https://www.linkedin.com/in/yusuke-noda-908797/>

Yusuke Noda is a principal software engineer at Microsoft Gaming Platform. He has led the development of gaming infrastructure for Xbox One and Xbox Cloud Gaming and has over 16 years of experience developing efficient infrastructure for cloud and gaming technologies.

Yusuke Noda 是微软游戏平台的首席软件工程师。他领导了Xbox One和Xbox云游戏的游戏基础设施开发，拥有超过16年云计算和游戏技术高效基础设施开发经验。

Demetri Terzopoulos UCLA Distinguished Professor, Academy Award winning computer scientist and entrepreneur.

Email: dt@cs.ucla.edu

Demetri Terzopoulos 加州大学洛杉矶分校（UCLA）杰出教授，获得奥斯卡奖的计算机科学家和企业家。电子邮件：dt@cs.ucla.edu

Web: <https://web.cs.ucla.edu/~dt/>

网站: <https://web.cs.ucla.edu/~dt/>

Dr. Demetri Terzopoulos is a Distinguished Professor and Chancellor's Professor of Computer Science at the University of California, Los Angeles, where he directs the UCLA Computer Graphics & Vision Laboratory. He is also Co-Founder and Chief Scientist of VoxelCloud, Inc., a multinational healthcare AI company. He is or was a Guggenheim Fellow, a Fellow of the Royal Society (FRS) of London and the Royal Society of Canada (FRSC), a Fellow of the Association for Computing Machinery (ACM), a Life Fellow of the Institute of Electrical and Electronics Engineers (IEEE), a Fellow of the Asia-Pacific Artificial Intelligence Association (AAIA), a Distinguished Fellow of the International Engineering and Technology Institute (IETI), a Member of the European Academy of Sciences (EAS) and the New York Academy of Sciences (NYAS), and a Life Member of Sigma Xi. Demetri Terzopoulos博士是加州大学洛杉矶分校计算机科学系的杰出教授及校长教授，主管UCLA计算机图形与视觉实验室。他还是跨国医疗人工智能公司VoxelCloud, Inc.的联合创始人兼首席科学家。他曾是古根海姆学者，伦敦皇家学会（FRS）和加拿大皇家学会（FRSC）院士，计算机协会（ACM）院士，电气电子工程师学会（IEEE）终身院士，亚太人工智能协会（AAIA）院士，国际工程与技术研究院（IETI）杰出院士，欧洲科学院（EAS）和纽约科学院（NYAS）成员，以及Sigma Xi终身会员。

Yejin Choi Wisnner-Slivaka Chair and Brett Helsel Professor at University of Washington, Senior Research Manager at Allen Institute for Artificial Intelligence, and MacArthur Fellow.

Yejin Choi 华盛顿大学Wissner-Slivaka讲席教授兼Brett Helsel教授，艾伦人工智能研究所高级研究经理，麦克阿瑟奖学金获得者。

Email: yejin@cs.washington.edu

电子邮件: yejin@cs.washington.edu

Web: <https://homes.cs.washington.edu/~yejin/>

网站: <https://homes.cs.washington.edu/~yejin/>

Yejin is a Wisnner-Slivaka Chair and Brett Helsel Professor at University of Washington and Senior Research Manager at Allen Institute of Artificial Intelligence. She has won the Anita Borg Early Career Award in 2018. She was the recipient of MacArthur Fellow foundation fellowship in 2020. She has received outstanding paper award in AAAI 2020, Neurips 2021, ICML 2022, and ACL 2023, and the best paper award in NAACL 2022 and ACL 2023. She is one of the main organizers of COLM 2024, an academic venue focused on the study of language modeling. Her main research interests are commonsense reasoning in the fields of Natural Language Processing, Machine Learning, Artificial Intelligence, with broader interests in Computer Vision and Digital Humanities.

Yejin是华盛顿大学Wissner-Slivaka讲席教授和Brett Helsel教授，同时担任艾伦人工智能研究所高级研究经理。她于2018年获得Anita Borg早期职业奖，2020年获得麦克阿瑟基金会奖学金。她曾获得AAAI 2020、NeurIPS 2021、ICML 2022和ACL 2023的优秀论文奖，以及NAACL 2022和ACL 2023的最佳论文奖。她是2024年COLM会议的主要组织者之一，该会议专注于语言建模研究。她的主要研究兴趣包括自然语言处理、机器学习和人工智能领域的常识推理，同时对计算机视觉和数字人文学科也有广泛兴趣。

Fei-Fei Li Professor of Computer Science at Stanford University

Fei-Fei Li 斯坦福大学计算机科学教授

Email: feifeili@stanford.edu

电子邮件: feifeili@stanford.edu

Web: <https://profiles.stanford.edu/fei-fei-li/>

网站: <https://profiles.stanford.edu/fei-fei-li/>

Fei-Fei is the inaugural Sequoia Professor in the Computer Science Department at Stanford University, and Co-Director of Stanford's Human-Centered AI Institute. She served as the Director of Stanford's AI Lab from 2013 to 2018. And during her sabbatical from Stanford from January 2017 to September 2018, Dr. Li was Vice President at Google and served as Chief Scientist of AI/ML at Google Cloud. Since then she has served as a Board member or advisor in various public or private companies.

Fei-Fei是斯坦福大学计算机科学系首任Sequoia讲席教授，斯坦福以人本人工智能研究所联合主任。她曾于2013年至2018年担任斯坦福人工智能实验室主任。2017年1月至2018年9月休假期间，李博士任谷歌副总裁兼谷歌云AI/机器学习首席科学家。此后，她在多家公私企业担任董事或顾问。

Katsushi Ikeuchi Senior Principal Research Manager at Microsoft and an IEEE Life Fellow. Microsoft Research, Redmond, WA. IEEE Life Fellow.

Katsushi Ikeuchi 微软高级首席研究经理，IEEE终身院士。微软研究院，雷德蒙德，华盛顿州。IEEE终身院士。

70 Email: katsuike@microsoft.com

71 电子邮件: katsuike@microsoft.com

Web: <https://www.microsoft.com/en-us/research/people/katsuike/>

网页: <https://www.microsoft.com/en-us/research/people/katsuike/>

Dr. Katsushi Ikeuchi received the BE degree in Mechanical Engineering from Kyoto University in 1973 and the PhD degree in Information Engineering from the University of Tokyo in 1978. After working at the Artificial Intelligence Laboratory, Massachusetts Institute of Technology for three years, Electrotechnical Laboratory, Ministry of International Trade and Industry, Japanese government for five years, and Robotics Institute, Carnegie Mellon University for 10 years, Institute of Industrial Science, the University of Tokyo for 19 years, he joined Microsoft as a Principal Researcher in 2015. During this tenure of CMU and UTokyo, he supervised more than 50PhD students.

池内克志博士于1973年毕业于京都大学机械工程专业，1978年获得东京大学信息工程博士学位。曾先后在麻省理工学院人工智能实验室工作三年，日本通商产业省电气技术实验室工作五年，卡内基梅隆大学机器人研究所工作十年，东京大学产业科学研究所工作十九年，2015年加入微软担任首席研究员。在卡内基梅隆大学和东京大学任职期间，他指导了超过50PhD名学生。

His research interest spans computer vision, robotics, and computer graphics. In these research fields, he has received several best paper awards, including the David Marr Prize in computational vision and IEEE Robotics and Automation Society K. S. Fu memorial best transaction paper.

他的研究兴趣涵盖计算机视觉、机器人学和计算机图形学。在这些领域，他获得了多项最佳论文奖，包括计算视觉领域的David Marr奖和IEEE机器人与自动化学会K. S. Fu纪念最佳论文奖。

His community service includes: general chair of IROS95, ITSC99, IV01, ICCV05, ACCV07, ICCV17; program chair of CVPR96, ICCV03, ICRA09, ICPR12, ICCV15; EIC of IJCV (2000-2017), EIC of IJ ITS (2012-2014), associate editor of IEEE Trans. RA, IEEE Trans. PAMI; and a distinguished lecturer of IEEE Signal Processing Society in 2000-2002, IEEE Robotics and Automation Society in 2004-2006 and IEEE Computer Society in 2008-2010.

他的学术服务包括：IROS95、ITSC99、IV01、ICCV05、ACCV07、ICCV17大会的总主席；CVPR96、ICCV03、ICRA09、ICPR12、ICCV15大会的程序主席；《国际计算机视觉杂志》(IJCV)主编（2000-2017）、《国际智能交通系统杂志》(IJ ITS)主编（2012-2014）、IEEE机器人与自动化汇刊和IEEE模式分析与机器智能汇刊副编辑；以及IEEE信号处理学会（2000-2002）、IEEE机器人与自动化学会（2004-2006）和IEEE计算机学会（2008-2010）的杰出讲师。

Through these research and society service, he was awarded a (life) fellow from IEEE, IEICE, IPSJ and RSJ. He received the Distinguished Researcher Award from IEEE-PAMI, the Medal of Honor with Purple Ribbon from Japanese Emperor, the Okawa prize from Okawa foundation as well as a couple of research achievement awards from Japanese professional societies.

通过这些研究和学术服务，他被授予IEEE、IEICE、IPSJ和RSJ的终身会士称号。他获得了IEEE-PAMI杰出研究奖、日本天皇颁发的紫绶褒章、冈和基金会冈和奖，以及日本多个专业学会的若干研究成就奖。

Hoi Vo Technical Fellow. Microsoft Gaming and X-Box Emerging Technologies, Redmond, WA.

Hoi Vo 技术院士。微软游戏与Xbox新兴技术部门，华盛顿雷德蒙德。

Email: hoiv@microsoft.com

电子邮件：hoiv@microsoft.com

Web: <https://www.linkedin.com/in/hoi-vo-193420/>

网页：<https://www.linkedin.com/in/hoi-vo-193420/>

Hoi Vo is a Technical Fellow from MS Gaming Division focusing on bridging AGI with new Gaming IP. He has played a pivotal role in establishing the Xbox Cloud Gaming service has led cloud and gaming efforts across teams in Microsoft, including Microsoft's Windows Azure platform. His focus is on the optimization of AI models to run efficiently at the edge while leveraging the cloud for scenarios that exceed the hardware capabilities, including fine-tuning models to be more engaging with various playing experiences.

Hoi Vo是微软游戏部门的技术院士，专注于将通用人工智能（AGI）与新游戏知识产权（IP）相结合。他在建立Xbox云游戏服务方面发挥了关键作用，领导了微软内部跨团队的云计算和游戏工作，包括微软的Windows Azure平台。他的重点是优化人工智能模型以高效运行于边缘设备，同时利用云计算处理超出硬件能力的场景，包括微调模型以提升各种游戏体验的互动性。

Jianfeng Gao Distinguished Scientist and Vice President at Microsoft and an IEEE Fellow. Microsoft Research, Redmond, WA. IEEE Fellow.

高建峰，微软杰出科学家兼副总裁，IEEE会士。微软研究院，华盛顿雷德蒙德。IEEE会士。

Email: jfgao@microsoft.com

电子邮件：jfgao@microsoft.com

Web: <https://www.microsoft.com/en-us/research/people/jfgao/>

网页：<https://www.microsoft.com/en-us/research/people/jfgao/>

Dr. Jianfeng Gao is Distinguished Scientist and Vice President at Microsoft Research, IEEE Fellow, ACM Distinguished Member, and the current head of the Deep Learning Group at Microsoft Research. From 2014 to 2017, he was Partner Research Manager in Business AI at Microsoft AI Research and at Deep Learning Technology Center (DLTC) at Microsoft Research, Redmond. He lead the development of AI solutions to Predictive Sales and Marketing. He also works on deep learning for text and image processing (see ACL/SIGIR 2018 Tutorial, Deep Learning 2017 Tutorial and IJCAI 2016 Tutorial or MS internal site) and lead the development of AI systems for dialogue, machine reading comprehension (MRC), and question answering (QA). From 2022, he leads the research of self-improving AI where LLMs (e.g., ChatGPT/GPT4) are augmented and adapted for the development of commercial AI systems.

高建峰博士是微软研究院的杰出科学家兼副总裁，IEEE会士，ACM杰出会员，现任微软研究院深度学习组负责人。2014年至2017年，他曾任微软人工智能研究业务AI合伙人研究经理及微软研究院深度学习技术中心（DLTC）负责

人，领导开发面向预测销售和市场营销的AI解决方案。他还从事文本和图像处理的深度学习研究（参见ACL/SIGIR 2018教程、Deep Learning 2017教程及IJCAI 2016教程或微软内部网站），领导对话系统、机器阅读理解（MRC）和问答系统（QA）的AI系统开发。自2022年起，他领导自我改进型AI的研究，致力于将大型语言模型（如ChatGPT/GPT4）增强和适配于商业AI系统的开发。

72 Acknowledgements

73 致谢

We are especially grateful to Peter Lee, Doug Burger, Desney Tan, Johannes Gehrke, Ryen White, Ece Kamar, Subhojit Som, and Kareem Choudhry for their advices, enormous support, and encouragement. We thank to Haiyan Zhang, Spencer Perreault, Dave Bignell, Katja Hofmann, Sam Devlin, Shanzheng Tan, Raluca Georgescu, Bill Dolan, Nebojsa Jojic, Sudha Rao, Adrian Brown, Andrzej Banburski-Fahey, Jianwei Yang for the early insightful discussions and helps of Gaming. We appreciate Kiran Muthabatulla, Antonio Criminisi, Tom Cashman, Nguyen Bach, Jennifer Marsman, Jaron Lanier from Mesh team, OCTO of Microsoft, and Microsoft office team for their mix-reality work, dataset work, and their generous helps and feedback for the project. We special thanks to Paul Bennett, Corby Rosset, Michel Galley, Chenglong Wang, Baolin Peng, Hao Chen, Silviu Cucerzan, Ahmed Awadallah, Saleema Amershi for their suggestion and comments for the NLP part. The authors gratefully acknowledge Paul Smolensky, Yonatan Bisk, Kezhen Chen, Borui Wang, Liangke Gui, Dingmin Wang, Xin (Eric) Wang, Zhe Gan, Xiaojian Ma, Zilong Zheng, Song-chun Zhu, Dragomir R. Radev, Daniel McDuff, Harry Shum for the related previous works, comments, suggestions, painstaking multiple reviews of this paper, and their pointers to the literature. Finally, we would like to really appreciate Microsoft Hololens team, Microsoft X-box team, and Meta Quest team for their generous provision of the equipment; MSR Central Engineering (CE) team, Microsoft 343 team for the data collection and sharing; Microsoft AOAI and GCR team for their Azure-OpenAI endpoint supporting.

我们特别感谢Peter Lee、Doug Burger、Desney Tan、Johannes Gehrke、Ryen White、Ece Kamar、Subhojit Som和Kareem Choudhry的建议、巨大支持和鼓励。感谢Haiyan Zhang、Spencer Perreault、Dave Bignell、Katja Hofmann、Sam Devlin、Shanzheng Tan、Raluca Georgescu、Bill Dolan、Nebojsa Jojic、Sudha Rao、Adrian Brown、Andrzej Banburski-Fahey、Jianwei Yang在游戏领域早期富有洞见的讨论和帮助。感谢Mesh团队的Kiran Muthabatulla、Antonio Criminisi、Tom Cashman、Nguyen Bach、Jennifer Marsman、Jaron Lanier，微软OCTO团队及微软办公团队在混合现实工作、数据集工作以及对本项目的慷慨帮助和反馈。特别感谢Paul Bennett、Corby Rosset、Michel Galley、Chenglong Wang、Baolin Peng、Hao Chen、Silviu Cucerzan、Ahmed Awadallah、Saleema Amershi对自然语言处理（NLP）部分的建议和评论。作者衷心感谢Paul Smolensky、Yonatan Bisk、Kezhen Chen、Borui Wang、Liangke Gui、Dingmin Wang、Xin (Eric) Wang、Zhe Gan、Xiaojian Ma、Zilong Zheng、Song-chun Zhu、Dragomir R. Radev、Daniel McDuff、Harry Shum对相关前期工作、评论、建议、对本文的多次细致审阅及文献指引。最后，我们衷心感谢微软Hololens团队、微软Xbox团队和Meta Quest团队慷慨提供设备；微软MSR中央工程（CE）团队、微软343团队负责数据收集与共享；微软AOAI和GCR团队对Azure-OpenAI端点的支持。

We would like to thank our colleagues from Stanford's Partnership in AI-assisted Care, who helped inform the medical applications explored in this work. In particular, we would like to thank Ehsan Adeli, Paul Tang, Amit Kaushal, Roger Bohn, Kevin Schulman, and Arnold Milstein for their clinical expertise and guidance.

我们感谢斯坦福人工智能辅助护理合作项目的同事们，他们为本研究中探索的医疗应用提供了宝贵信息。特别感谢Ehsan Adeli、Paul Tang、Amit Kaushal、Roger Bohn、Kevin Schulman和Arnold Milstein的临床专业知识和指导。

This research was supported by Microsoft Research project Fair 2023, Microsoft HackBox 2023, OCTO team.
本研究得到了微软研究院Fair 2023项目、微软HackBox 2023项目及OCTO团队的支持。