中山大學 软件工程学院
SUN YAT-SEN UNIVERSITY SCHOOL OF SOFTWARE ENGINEERING

AI2-THOR: An Interactive 3D Environment for Visual AI
写作技巧分析

论文发表：arXiv(2022)，引用超过1200+

论文团队：*Allen Institute for AI, University of Washington, Stanford University, Carnegie Mellon University*

汇报人：25111639 朱正阳

目录

- 优点
  - 明确贡献与消除歧义 (Who did what)
  - 突出核心贡献、结构醒目吸引人 (Highlighting Your Key findings)
  - 使用 Hedging， 学术语气稳健不夸大（Hedging）
  - 结论组织清晰 、内容完整(Conclusion Constructure)
  - 合理转写/引用 (Plagiarism and Paraphrasing)
- 缺点
  - 缺少局限性讨论 (Discussion Limitation)

优点1： 明确贡献与消除歧义 （Who did what)

- 使用第一人称（we 和 our）明确作者贡献
  - *We* introduce The House Of inteRactions (THOR), ….
  - *We* also provide support for many scenes designed manually by professional 3D artists, …
  - *we* set up two profiling experiments,
  - 使用z主动语态，区分作者贡献与其他人的贡献
  - 这避免了被动语态可能的歧义问题
    - *We found* that Habitat instances used slightly less GPU memory than ProcTHOR instances
    - 而不是写 *it is found that*

优点2: 突出核心贡献(Highlighting Key findings)

- 介绍工具用途时，使用小标题、粗体、标号等技巧 *(subheadings / bullet / bold)*
- 使用短句吸引读者 *(shorter sentences, easy to follow)*

3 What has AI2-THOR been used for?

Since the initial release of AI2-THOR in 2017, it has been used for experimentation in over 150 publications and downloaded over 500k times. Some areas of work that we found particularly interesting include:

- **Visual Navigation.** Visual navigation was the first use case of AI2-THOR [45], which trains an agent to perform ImageNav (*i.e.* navigating to an image where the target object is described with a picture of it). Here, the agent executes a sequence of move or rotate commands to reach the target from egocentric camera inputs at each time step. ObjectNav is another common navigation task, where the agent is tasked with navigating to a given semantic category, such as a bed. Follow-up work from [40, 3, 44] uses semantic priors about where objects typically occur to improve navigation efficiency; [37] used meta-learning to try and better adapt to unseen scenes; [22] uses a Markov network to build a map of the environment; [15] found that using CLIP as a pre-trained visual encoder helps significantly boost generalization performance; and [2] found that training on many procedurally generated scenes strongly generalizes to RoboTHOR, iTHOR, and ArchitecTHOR in a 0-shot setting.
- **Audio-Visual Navigation.** [8] proposes the task of audio-visual navigation in which the agent is tasked with navigating to find where the sound is coming from in the scene.
- **Vision-and-Language.** AI2-THOR has been used extensively for embodied vision-and-language research. Noteable datasets include ALFRED [31], for interactive instruction following from natural language; TEACh [27], for interactive instruction following from human-robot dialog; and DialFRED [9] and IQA [10] for interactive question-answering. Some other interesting work includes [28], which proposes the Episodic Transformer to encode the full history of vision and language inputs with each ALFRED task; [13], which uses grammar-based methods to learn high-level abstractions through decompositions of tasks; FILM [23], which builds a semantic map to perform exploration for instruction following; and PIGLeT [41], which learns natural language grounding through interaction.
- **Human-Robot Interaction.** [38] inserts a human into AI2-THOR and uses virtual reality to control its gestures in simulation. By controlling the human's gestures, it can communicate different tasks it wants the robot to achieve, such as pointing to an object to encode moving to that object.
- **Sim2Real Transfer.** RoboTHOR [1] studies sim2real transfer for robotics. Here, the goal is to train in simulation because it is faster, cheaper, and more scalable, and then to deploy the trained agent in the real-world. Agents train on 75 scenes in simulation and evaluate on unseen real-world scenes that come from a similar distribution. Initial work analyzed sim2real transfer for agents trained to perform ObjectNav.
- **Multi-Agent Interaction.** [12] proposes the collaborative task of having 2 agents move to lift up furniture in a scene. For example, both agents might have to navigate to find the television in the scene, and work together to lift it up. Follow-up work from [11] takes the task a step further, where the agents not only have to lift up the furniture, but also work together to move it. Both tasks require visual navigation from the agents, and for them to communicate and coordinate together. Some other notable multi-agent work includes [35], which tasks agents with playing Cache, a variant of hide-and-seek where one agent hides an object and the other agent is tasked with finding that object; [33], which uses multiple agents for interactive question answering; [20], which proposes using multiple agents to more efficiently find multiple target objects in a scene; and

优点2: 突出核心贡献(Highlighting Key findings)

- 介绍工具用途时，还加入图表，对应每一个小标题，视觉上吸引读者、结构清晰。*(Placing tables/figures strategically)*

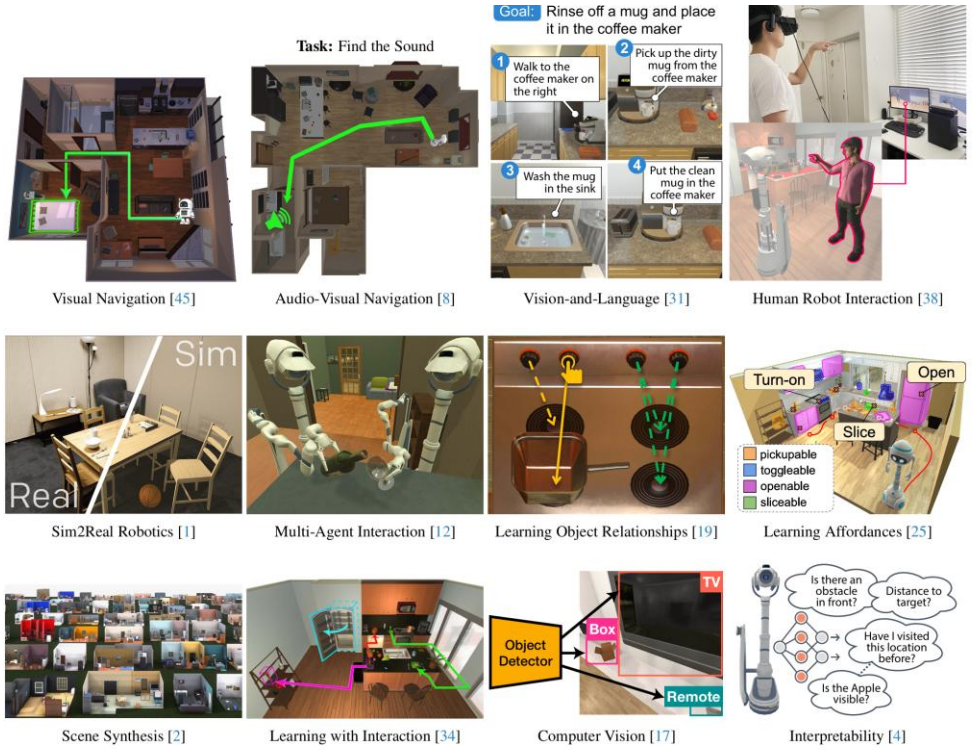

Figure 10: AI2-THOR has enabled research in a wide range of fields. Here, we highlight some examples of how it has been used.

优点3： 使用 Hedging， 学术语气稳健

- 使用“弱化语气”， 避免绝对化判断
  - …., which *may be* used to …
  - …, where one *may be* interested in studying high-level planning rather than low-level control

(a) Scene Bounds (b) 3D Bounding Box of Objects (c) Reachable Grid Positions

Figure 9: Examples of environment metadata, including the dimensions of the scene, the 3D bounding box of each object, and the reachable grid positions, which may be used to randomize the agent's starting position or to build a heuristic search agent.

优点4：结论组织清晰 (Conclusion Constructure)

- begin清晰，没有多余的如“In this paper”的冗余 *(how to begin)*
- 时态清晰，完成时和过去时陈述工作的贡献 *(appropriate tense)*
- 展望未来，说明方法能够应用到其他场景以及能够进行扩展 *(recommendations & future work)*

5 Conclusion

We present AI2-THOR, a large-scale interactive simulation platform for Embodied AI. It has been used for experimentation in over 150 publications, spanning a wide variety of tasks and research areas. It is highly customizable, and provides first-class support for many different types of scenes, agent embodiments, actions, and metadata. The capabilities of AI2-THOR are rapidly evolving, and we are excited to support new improvements and use cases to come. For the latest information, please visit our website: https://ai2thor.allenai.org/.

优点5：合理转写/引用 (Plagiarism and Paraphrasing)

- 作者对其他人的贡献进行转述/总结，清晰有力的总结 related work
- *[22] uses* a Markov network to build a map of the environment;
- *[15] found* that using CLIP as a pre-trained visual encoder helps significantly boost generalization performance;
- *and [2] found* that training on many procedurally generated scenes strongly generalizes to RoboTHOR, iTHOR, and ArchitecTHOR in a 0-shot setting.

缺点： 缺少Limitations

- 论文中没有说明平台的局限性、不足之处 *(Discussion Limitation)*
  - 对比时只说明了比其他平台好的方面，缺少对Limitation的讨论

4 Why use AI2-THOR?

Simulator	Scale		Interaction					Simulator	
	# of Scenes	# of Objects	Object States	Arm Manipulation	Multi-Agent	Sound	VR	Engine	Interactive Editor
AI2-THOR	∞ [2]	3578	✓	✓	✓	✓	✓	Unity	✓
iGibson 2.0	15	1217	✓	✓	✓	✗	✓	PyBullet	✗
Habitat 1.0	1000	–	✗	✗	✗	✓	✗	Magnum	✗
Habitat 2.0	105	92	✓	✓	✗	✗	✗	Magnum	✗
ThreeDWorld	15	200	✗	✓	✗	✓	✓	Unity	✓
SAPIEN	0	2346	✗	✓	✗	✗	✗	PhysX	✗

Table 1: A comparison table between Embodied AI simulators.

Following AI2-THOR's first release in 2017, a number of simulators have been developed, including iGibson 2.0 [18], Habitat 1.0 [30], Habitat 2.0 [32], ThreeDWorld [7], and SAPIEN [39]. Table 1 shows a comparison table between the simulators. AI2-THOR is significantly larger in scale than other simulators, while providing first-class support for interaction, and, by leveraging Unity, makes it easy to add new capabilities.

Thanks