HoME: A HOUSEHOLD MULTIMODAL ENVIRONMENT

Simon Brodeur[1], **Ethan Perez**[2,3]*, **Ankesh Anand**[2]*, **Florian Golemo**[2,4]*,
Luca Celotti[1], **Florian Strub**[2,5], **Jean Rouat**[1], **Hugo Larochelle**[6,7], **Aaron Courville**[2,7]
[1]Université de Sherbrooke, [2]MILA, Université de Montréal, [3]Rice University, [4]INRIA Bordeaux,
[5]Univ. Lille, Inria, UMR 9189 - CRIStAL, [6]Google Brain, [7]CIFAR Fellow

ABSTRACT

We introduce HoME: a **Ho**usehold **M**ultimodal **E**nvironment for artificial agents to learn from vision, audio, semantics, physics, and interaction with objects and other agents, all within a realistic context. HoME integrates over 45,000 diverse 3D house layouts based on the SUNCG dataset, a scale which may facilitate learning, generalization, and transfer. HoME is open-source, OpenAI Gym-compatible platform extensible to tasks in reinforcement learning, language grounding, sound-based navigation, robotics, multi-agent learning, and more. We hope HoME better enables artificial agents to learn as humans do: in an interactive, multimodal, and richly contextualized setting.

1 INTRODUCTION

Human learning occurs through interaction (Fisher et al., 2012a) and multimodal experience (Landau et al., 1998; Smith & Yu, 2008). Prior work has argued that machine learning may also benefit from interactive, multimodal learning (Hermann et al., 2017; Oh et al., 2017; de Vries et al., 2017; Gauthier & Mordatch, 2016), termed *virtual embodiment* (Kiela et al., 2016). Driven by breakthroughs in static, unimodal tasks such as image classification (Krizhevsky et al., 2012) and language processing (Mikolov et al., 2013), machine learning has moved in this direction. Recent tasks such as visual question answering (Antol et al., 2015), image captioning (Vinyals et al., 2017), and audio-video classification (Dhall et al., 2015) make steps towards learning from multiple modalities but lack the dynamic, responsive signal from exploratory learning. Modern, challenging tasks incorporating interaction, such as Atari (Bellemare et al., 2013) and Go (Silver et al., 2016), push agents to learn complex strategies through trial-and-error but miss information-rich connections across vision, language, sounds, and actions. To remedy these shortcomings, subsequent work introduces tasks that are both multimodal and interactive, successfully training virtually embodied agents that, for example, ground language in actions and visual percepts in 3D worlds (Hermann et al., 2017; Oh et al., 2017; Chaplot et al., 2018).

For virtual embodiment to reach its full potential, though, agents should be immersed in a rich, lifelike context as humans are. Agents may then learn to ground concepts not only in various modalities but also in relationships to other concepts, i.e. that forks are often in kitchens, which are near living rooms, which contain sofas, etc. Humans learn by concept-to-concept association, as shown in child learning psychology (Landau et al., 1998; Smith & Yu, 2008), cognitive science (Barsalou, 2008), neuroscience (Nakazawa et al., 2002), and linguistics (Quine et al., 2013). Even in machine learning, contextual information has given rise to effective word representations (Mikolov et al., 2013), improvements in recommendation systems (Adomavicius & Tuzhilin, 2011), and increased reward quality in robotics (Jaderberg et al., 2017). Importantly, large datasets have proven key in algorithms learning from context (Mikolov et al., 2013) and in general (Russakovsky et al., 2015; Bojar et al., 2015; Tobin et al., 2017).

To this end, we present HoME: the **Ho**usehold **M**ultimodal **E**nvironment (Figure 1). HoME is a large-scale platform[1] for agents to navigate and interact within over 45,000 hand-designed houses from the SUNCG dataset (Song et al., 2017).

[*] These authors contributed equally.

[1] Available at `https://home-platform.github.io/`

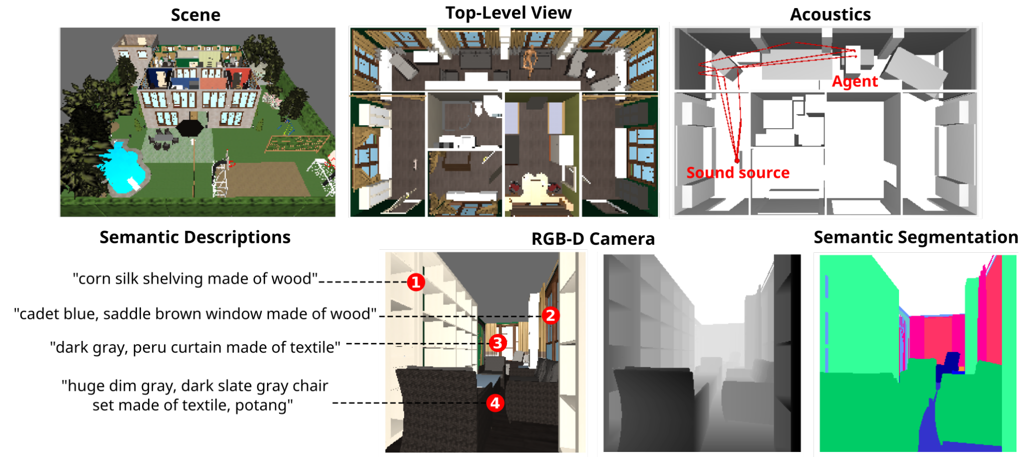
Scene
Top-Level View
Acoustics
Agent
Sound source
Semantic Descriptions
RGB-D Camera
Semantic Segmentation
"corn silk shelving made of wood" 1
"cadet blue, saddle brown window made of wood" 2
"dark gray, peru curtain made of textile" 3
"huge dim gray, dark slate gray chair set made of textile, potang" 4

Figure 1: A single example that demonstrates HoME's features.

Specifically, HoME provides:

- 3D visual renderings based on Panda3D.
- 3D acoustic renderings based on EVERT (Laine et al., 2009), using ray-tracing for high fidelity.
- Semantic image segmentations and language descriptions of objects.
- Physics simulation based on Bullet, handling collisions, gravity, agent-object interaction, and more.
- Multi-agent support.
- A Python framework integrated with OpenAI Gym (Brockman et al., 2016).

HoME is a general platform extensible to many specific tasks, from reinforcement learning to language grounding to blind navigation, in a real-world context. HoME is also the first major interactive platform to support high fidelity audio, allowing researchers to better experiment across modalities and develop new tasks. While HoME is not the first platform to provide realistic context, we highlight in the related work section in the appendix that HoME provides a more large-scale and/or multimodal testbed than existing environments, making it more conducive to virtually embodied learning in many scenarios.

2 HoME

Overviewed in Figure 1, HoME is an interactive extension of the SUNCG dataset (Song et al., 2017). SUNCG provides over 45,000 house layouts containing over 750,000 rooms and sometimes multiple floors. Within these rooms, of which there are 24 kinds, there are objects from among 84 categories and on average over 14 objects per room. HoME consists of several, distinct components built on SUNCG that can be utilized individually. The platform runs faster than real-time on a single-core CPU, enables GPU acceleration, and allows users to run multiple environment instances in parallel. These features facilitate faster algorithmic development and learning with more data. HoME provides an OpenAI Gym-compatible environment which loads agents into randomly selected houses and lets it explore via actions such as moving, looking, and interacting with objects (i.e. pick up, drop, push). HoME also enables multiple agents to be spawned at once. The following sections detail HoME's core components.

2.1 Rendering Engine

The rendering engine uses Panda3D (Goslin & Mine, 2004), an open-source 3D game engine which ships with complete Python bindings. For each SUNCG house, HoME renders RGB+depth scenes based on house and object textures (wooden, metal, rubber, etc.), multi-source lighting, and shadows. The rendering engine enables tasks such as vision-based navigation, imitation learning, and planning. **This module provides:** RGB image (with different shader presets), depth image.

2.2 Acoustic engine

The acoustic engine is implemented using EVERT[2], which handles real-time acoustic ray-tracing based on the house and object 3D geometry. EVERT also supports multiple microphones and sound sources, distance-dependent sound attenuation, frequency-dependent material absorption and reflection (walls muffle sounds, metallic surfaces reflect acoustics, etc.), and air-absorption based on atmospheric conditions (temperature, pressure, humidity, etc.). Sounds may be instantiated artificially or based on the environment (i.e. a TV with static noise or an agent's surface-dependent footsteps).
This module provides: stereo sound frames for agents w.r.t. environmental sound sources.

2.3 Semantic engine

HoME provides a short text description for each object, as well as the following semantic information:

- **Color**, calculated from object textures and discretized into 16 basic colors, ~130 intermediate colors, and ~950 detailed colors.
- **Category**, extracted from SUNCG object metadata. HoME provides both generic object categories (i.e. "air conditioner," "mirror," or "window") as well as more detailed categories (i.e. "accordion," "mortar and pestle," or "xbox").
- **Material**, calculated to be the texture, out of 20 possible categories ("wood," "textile," etc.), covering the largest object surface area.
- **Size** ("small," "medium," or "large") calculated by comparing an object's mesh volume to a histogram of other objects of the same category.
- **Location**, based on ground-truth object coordinates from SUNCG.

With these semantics, HoME may be extended to generate language instructions, scene descriptions, or questions, as in (Hermann et al., 2017; Oh et al., 2017; Chaplot et al., 2018). HoME can also provide agents dense, ground-truth, semantically-annotated images based on SUNCG's 187 fine-grained categories (e.g. bathtub, wall, armchair).
This module provides: image segmentations, object semantic attributes and text descriptions.

2.4 Physics engine

The physics engine is implemented using the Bullet engine[3]. For objects, HoME provides two rigid body representations: (a) fast minimal bounding box approximation and (b) exact mesh-based body. Objects are subject to external forces such as gravity, based on volume-based weight approximations. The physics engine also allows agents to interact with objects via picking up, dropping, pushing, etc. These features are useful for applications in robotics and language grounding, for instance.
This module provides: agent and object positions, velocities, physical interaction, collision.

3 Applications

Using HoME's various features and/or external data collection, HoME can facilitate tasks such as:

- Instruction Following: An agent is given a description of how to achieve a reward (i.e. "Go to the kitchen." or "Find the red sofa.").
- Visual Question Answering: An agent must answer an environment-based question which might require exploration (i.e. "How many rooms have a wooden table?").
- Dialogue: An agent converses with an oracle with full scene knowledge to solve a difficult task.
- Pied Piper: An agent has to locate another agent by sound alone.
- Multi-agent communication: Multiple agents communicate to solve a task and maximize a shared reward. For example, one agent might know reward locations to which it must guide other agents.

[2] https://github.com/sbrodeur/evert

[3] http://bulletphysics.org/

ACKNOWLEDGMENTS

We acknowledge the following agencies for research funding and computing support: CIFAR, CHISTERA IGLU, Collège Doctoral Lille Nord de France and CPER Nord-Pas de Calais/FEDER DATA Advanced data science and technologies 2015-2020, Calcul Québec, Compute Canada, and Google. We further thank NVIDIA for donating a DGX-1, Titan Xp, and Tesla K40 used in this work.

REFERENCES

Openai universe. `https://universe.openai.com/`, 2016.

Gediminas Adomavicius and Alexander Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pp. 217–253. 2011.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. *arXiv preprint arXiv:1711.07280*, 2017.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *CVPR*, 2015.

Lawrence W Barsalou. Grounded cognition. *Annual Review of Psychology*, 59:617–645, 2008.

Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. Deepmind lab. *arXiv preprint arXiv:1612.03801*, 2016.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *JAIR*, 47:253–279, 2013.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the 2015 workshop on statistical machine translation. In *Workshop on Statistical Machine Translation*, 2015.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*, 2017.

Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, and Ruslan Salakhutdinov. Gated-attention architectures for task-oriented language grounding. In *AAAI*, 2018.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017.

Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *International Conference on Multimodal Interaction*, 2015.

Kelly Fisher, Kathy Hirsh-Pasek, and Roberta M Golinkoff. Fostering mathematical thinking through playful learning. *Contemporary debates on child development and education*, pp. 81–92, 2012a.

Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas Funkhouser, and Pat Hanrahan. Example-based synthesis of 3d object arrangements. *ACM Transactions on Graphics*, 2012b.

Jon Gauthier and Igor Mordatch. A paradigm for situated and goal-driven language learning. *NIPS Machine Intelligence Workshop*, 2016.

Mike Goslin and Mark R Mine. The panda3d graphics engine. *Computer*, 37(10):112–114, 2004.

Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *CVPR*, 2017.

Ankur Handa, Viorica Pătrăucean, Simon Stent, and Roberto Cipolla. Scenenet: An annotated model generator for indoor scene understanding. In *ICRA*, 2016.

Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojtek Czarnecki, Max Jaderberg, Denis Teplyashin, et al. Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*, 2017.

Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *3DV*, pp. 92–101. IEEE, 2016.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *ICLR*, 2017.

Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. The malmo platform for artificial intelligence experimentation. In *IJCAI*, 2016.

Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. ViZDoom: A Doom-based AI research platform for visual reinforcement learning. In *Computational Intelligence and Games*, 2016.

Douwe Kiela, Luana Bulat, Anita L Vero, and Stephen Clark. Virtual embodiment: A scalable long-term strategy for artificial intelligence research. In *Machine intelligence workshop at NIPS*, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

Samuli Laine, Samuel Siltanen, Tapio Lokki, and Lauri Savioja. Accelerated beam tracing algorithm. *Applied Acoustics*, 70(1):172–181, 2009.

Barbara Landau, Linda Smith, and Susan Jones. Object perception and object naming in early development. *Trends in cognitive sciences*, 2(1):19–24, 1998.

John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *CVPR*, 2017.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

Kazu Nakazawa, Michael C Quirk, Raymond A Chitwood, Masahiko Watanabe, Mark F Yeckel, Linus D Sun, Akira Kato, Candice A Carr, Daniel Johnston, Matthew A Wilson, et al. Requirement for hippocampal ca3 nmda receptors in associative memory recall. *Science*, 297(5579):211–218, 2002.

Junhyuk Oh, Satinder Singh, Honglak Lee, and Pushmeet Kohli. Zero-shot task generalization with multi-task deep reinforcement learning. In *ICML*, 2017.

Willard Van Orman Quine, Patricia S Churchland, and Dagfinn Føllesdal. *Word and object*. MIT press, 2013.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

Manolis Savva, Angel X. Chang, Alexey Dosovitskiy, Thomas Funkhouser, and Vladlen Koltun. MINOS: Multimodal indoor simulator for navigation in complex environments. *arXiv:1712.03931*, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Linda Smith and Chen Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008.

Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *CVPR*, 2017.

Joshua Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IROS*, 2017.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *TPAMI*, 39(4):652–663, 2017.

Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209*, 2018.

Claudia Yan, Dipendra Misra, Andrew Bennett, Aaron Walsman, Yonatan Bisk, and Yoav Artzi. CHALET: Cornell house agent learning environment. *arXiv:1801.07357*, 2018.

Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven Visual Navigation in Indoor Scenes using Deep Reinforcement Learning. In *ICRA*, 2017.

Table 1: A comparison of modern environments with HoME. **3D**: Supports 3D settings. **Context**: Provides a realistic context. **Large-scale**: Thousands of readily available environments. **Fast**: >100 FPS rendering speed. **Flexible**: Adaptable towards various, specific tasks. **Physics**: Supports rigid body dynamics and external forces (gravity, collisions, etc.) on agents and objects. **Acoustics**: Renders audio. **Photorealism**: Lifelike visual rendering. **Actionable**: Some objects can be toggled (e.g. open/close)

Environment	3D	Context	Large-scale	Fast	Flexible	Physics	Acoustics	Photorealistic	Actionable
Atari (Bellemare et al., 2013)				•					
OpenAI Universe (Uni, 2016)	•	•	•		•				
Malmo (Johnson et al., 2016)	•		•		•				
DeepMind Lab (Beattie et al., 2016)	•			•	•				
VizDoom (Kempka et al., 2016)	•			•	•				
AI2-THOR (Zhu et al., 2017)	•	•				•		•	•
CMP (Gupta et al., 2017)	•	•		•				•	
Matterport3D (Anderson et al., 2017)	•	•		•				•	
House3D (Wu et al., 2018)	•	•	•	•	•				
Minos (Savva et al., 2017)	•	•	•	•	•				
CHALET (Yan et al., 2018)	•	•			•	•			•
HoME	•	•	•	•	•	•	•		

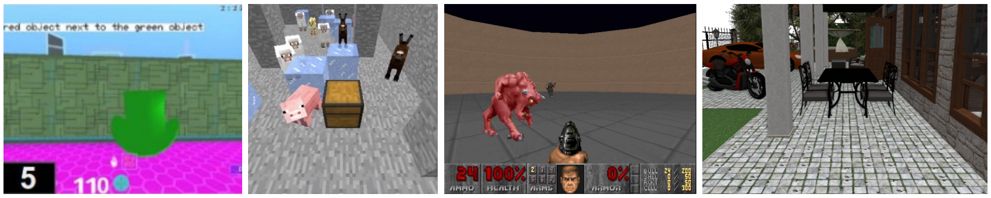

Figure 2: 3D environments (Left to right): DeepMind Lab, Malmo, VizDoom, HoME. Prior environments focus on algorithmic challenges; HoME adds a full context in which to learn concepts.

A Related work

The AI community has built numerous platforms to drive algorithmic advances by tackling video games (Bellemare et al., 2013; Uni, 2016; Johnson et al., 2016; Beattie et al., 2016; Kempka et al., 2016), 3D navigation in real environments reconstructed from photos (Gupta et al., 2017; Anderson et al., 2017), and language-related tasks in simulated 3D households (Zhu et al., 2017; Savva et al., 2017; Yan et al., 2018; Wu et al., 2018).

Earlier environments were created to be powerful 3D sandboxes for developing learning algorithms, while HoME additionally aims to provide a unified platform for multimodal learning in a realistic context (Figure 2). Table 1 compares these environments to HoME.

The most closely related environments to HoME are AI2-THOR (Zhu et al., 2017), House3D (Wu et al., 2018), Minos (Savva et al., 2017), and CHALET (Yan et al., 2018), four other household environments, of which the last three were concurrently developed and released after HoME. House3D and Minos are also based on SUNCG, but both lack audio support, true physics simulation, and the ability for agents to interact with objects — key aspects of multimodal, interactive learning. AI2-THOR and Chalet focus specifically on visual navigation and fine-grained physical interaction with objects (like opening/closing a fridge), but are limited to 32 and 10 houses, respectively. HoME instead aims to provide an extensive number of houses and easy integration with multiple modalities and new tasks.

Other 3D house datasets could also be turned into interactive platforms, but these datasets are not as large-scale as SUNCG, which consists of 45622 house layouts. These datasets include Stanford Scenes (1723 layouts) (Fisher et al., 2012b), Matterport3D (Chang et al., 2017) (90 layouts), sceneNN (100 layouts) (Hua et al., 2016), SceneNet (57 layouts) (Handa et al., 2016), and SceneNet RGB-D (57 layouts) (McCormac et al., 2017). We used SUNCG, as scale and diversity in data have proven critical for machine learning algorithms to generalize (Russakovsky et al., 2015; Bojar et al., 2015) and transfer, such as from simulation to real (Tobin et al., 2017). SUNCG's simpler graphics also allow for faster rendering.