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

Introduction

What is it to understand human language, from a task-oriented machine perspective?

Image from: https://www.goodfreephotos.com

Human Language Understanding (HLU)

Example applications in machine learning:

- **Machine translation**:
  English: "The car on my right was going too fast."
  ↓
  French: "La voiture à ma droite allait trop vite."
- **Image description generation**:

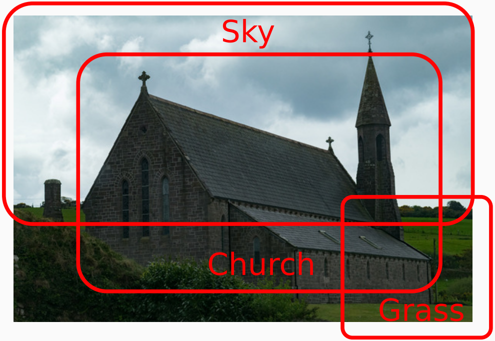
➔ "A small church on a cloudy day with a green grass field in the background."

This requires language to be related to an actual visual scene!

Human Language Understanding (HLU)

But it is impractical in real life to have all annotations:

- Which objects are visible in the scene?
- What are the size and material properties of the objects?
- What are the spatial relationships between objects?



Motivation

Use realistic but virtual environments for a situated agent to learn to ground language!

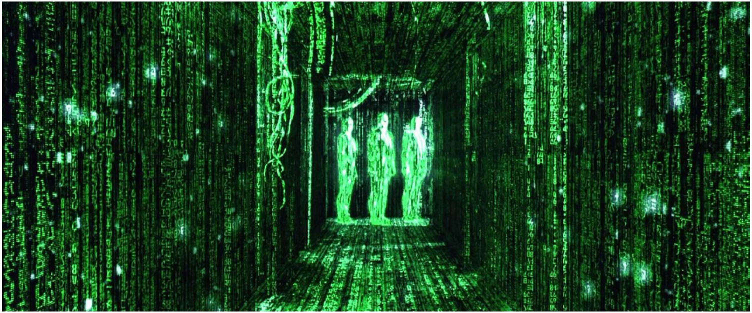
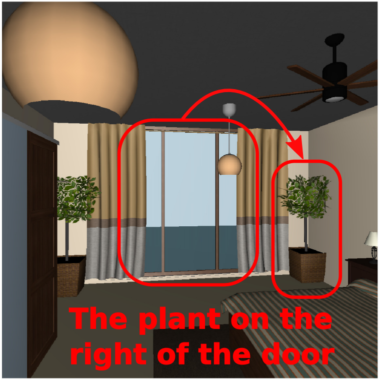
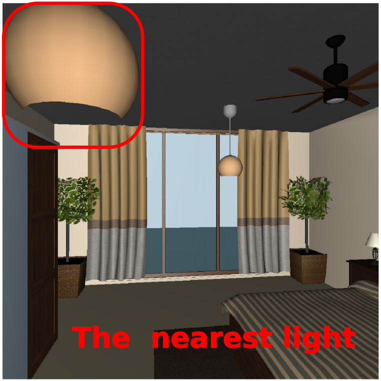


Image from: https://www.walldevil.com

Language Grounding

Examples of language grounding in virtual visual scenes:





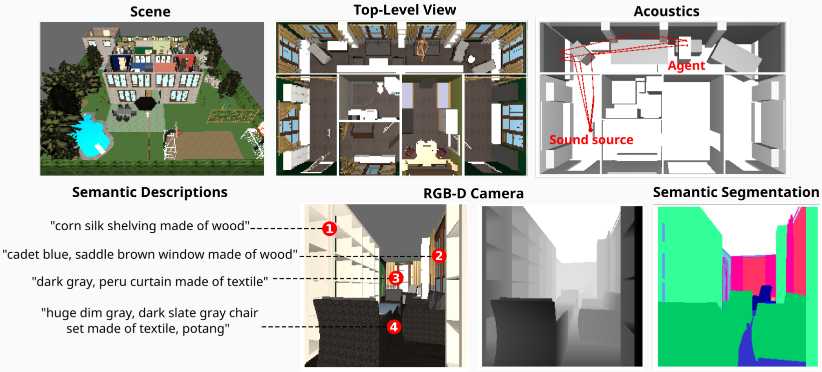
Language Grounding

Examples of language grounding in virtual visual scenes:

HoME overview

HoME in a nutshell

- Using SUNCG (Song et al., 2017) data in the Panda3d game engine.
- Integrating language, vision, physics and acoustics.



SUNCG Dataset

A large-scale synthetic scene dataset of 3D houses:

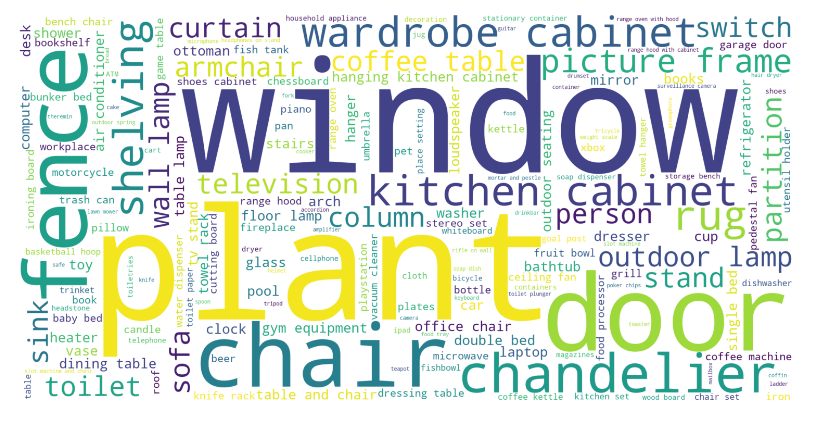
- Over 45,000 different human-designed houses.

Image adapted from http://suncg.cs.princeton.edu/

SUNCG Dataset

A large-scale synthetic scene dataset of 3D houses:

- Over 2500 different objects (e.g. table, closet).

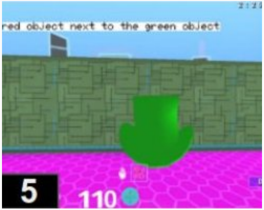


Comparison with existing frameworks

Grounded language learning for navigation
(Can and Yuret, 2017)

Image from: www.denizyuret.com

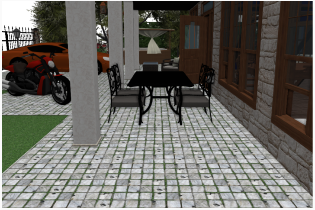
Comparison with existing frameworks



(a) DeepMind Lab
(Beattie et al., 2016)

(b) Malmo
(Johnson et al., 2016)

(c) ViZDoom
(Kempka et al., 2016)



(d) SUNCG
(Song et al., 2017)

Comparison with existing frameworks

Many recent frameworks on complex and navigable 3D indoor environments:

- **House3D** (Wu et al., 2017)
- **AI2-THOR** (Kolve et al., 2017)
- **CHALET** (Yan et al., 2018)
- **MINOS** (Savva et al., 2017)
- **Matterport3D** (Anderson et al., 2017)

Overview of the Panda3D engine

A maintained, mature engine with full Python integration:

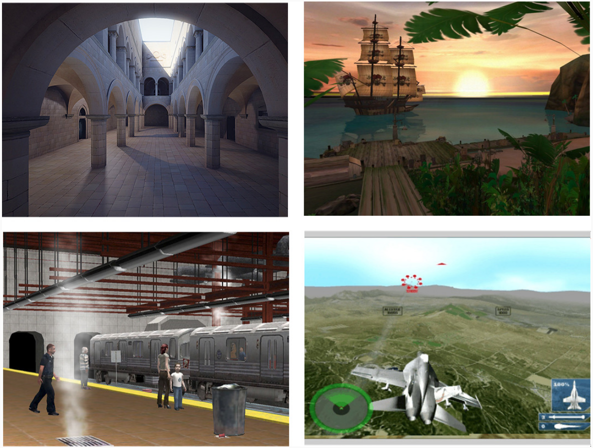


Image adapted from: https://www.panda3d.org/

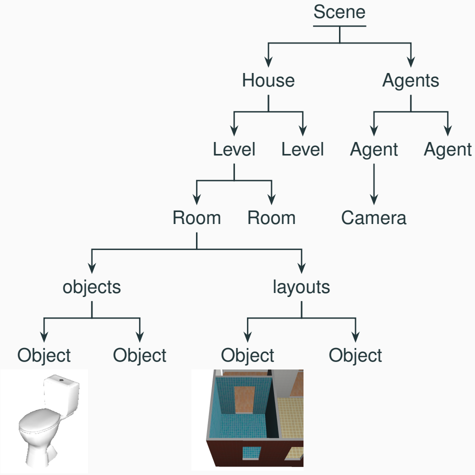
Overview of the Panda3D engine

A maintained, mature engine with full Python integration:

Video adapted from: https://www.youtube.com/watch?v=SVRshBffzM4&t=17s

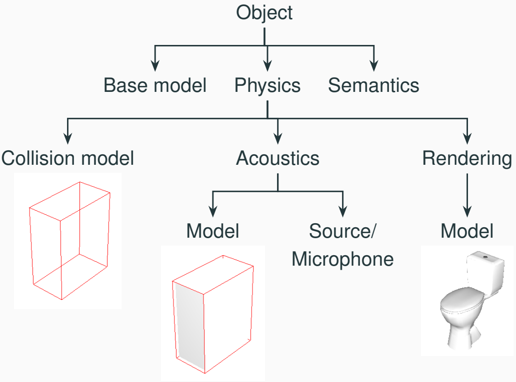
Overview of the Panda3D engine

Hierarchical scene graph: store objects, relative positions and annotations.



Overview of the Panda3D engine

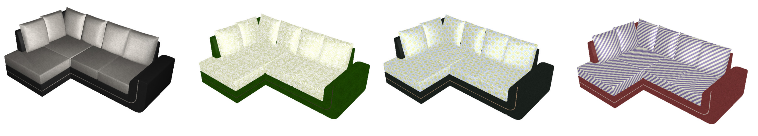
Hierarchical scene graph: joint graph for rendering, physics and acoustics.



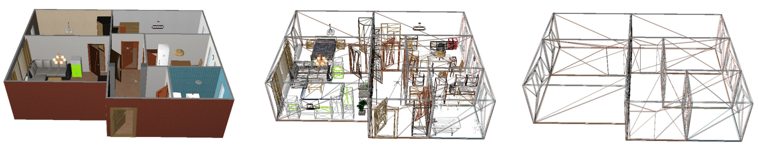
Overview of the Panda3D engine

Low-level access to objects in the graph allows:

- Easy data augmentation (e.g. override colors/textures)



- Geometry-dependent processing (e.g. acoustics)



Overview of the Panda3D engine

Low-level access to objects in the graph allows:

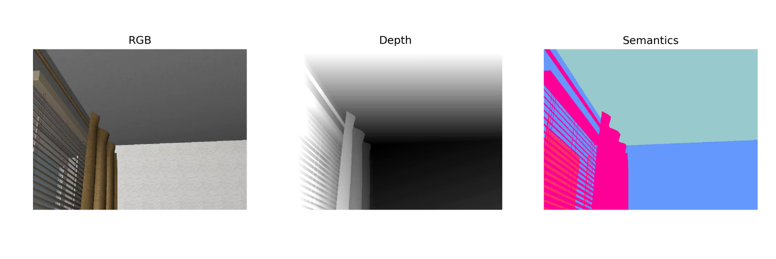
- Geometry-dependent processing (e.g. navigation)

HoME core components

Rendering

RGB-D and dense semantic segmentation (GPU-accelerated)

- Framerate > 300 FPS on a high-end GPU (120x90 pixels).



Physics

Collision and (simple) physical interactions with objects

- Based on Bullet Physics, integrated in Panda3d.

Acoustics

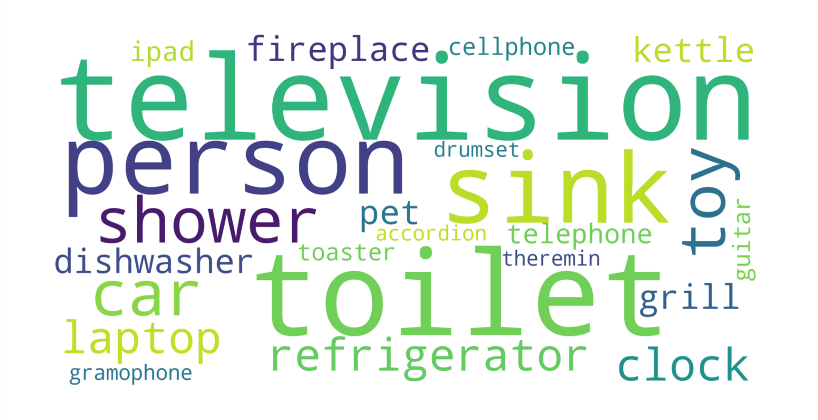
Real-time acoustic ray tracing with multiple sources/microphones

- Based on EVERT (Laine et al., 2009), using 3D geometry of the scene.
- Frequency-dependent absorption using object materials.

Acoustics

There are about 30 audible objects in SUNCG

- Thousands of sound samples available on Freesound.org



Semantics

What semantic information is provided:

- **Category** (86): from SUNCG object metadata (e.g. “air conditioner,” “mirror,” or “window”).
- **Location** (24): from ground-truth object coordinates and SUNCG room metadata (e.g. “in the kitchen”).
- **Color** (16-950): from object textures and discretized from basic colors to detailed colors (e.g. “brownish orange”).
- **Material** (15): from object textures (e.g. “wood,” “textile,” “leather”).

Semantics

In comparison with CLEVR dataset (Johnson et al., 2016):
3 objects x 2 sizes x 8 colors x 2 materials

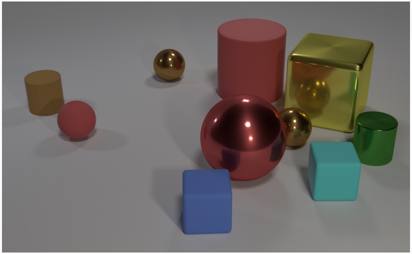


Image from: https://cs.stanford.edu/people/jcjohns/clevr/

HoME is an order of magnitude more diverse!

Applications

General applications

- **Instruction following**: An agent is given a description of how to achieve a reward (e.g. “Go to the kitchen.” or “Find the red sofa.”).
- **Visual question answering**: An agent must answer an environment-based question which might require exploration (e.g. “How many rooms have a wooden table?”).
- **Dialogue and multi-agent communication**: An agent converses with a human or another agent to solve a task (e.g. “Where can I find the kitchen?.”).

Dialogue-related applications

Goal-oriented cooperative dialogue game in HoME:

- GuessWhat?! (H. de Vries et al., 2017):

Questioner	**Oracle**
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

Conclusion

Conclusion

- ✓ Learn to ground language from vision, acoustics, physics and interaction with objects and other agents.
- ✓ Realistic context in house environments and large-scale.
- ✓ Objective evaluation metrics, leading to controllable and reproducible research.
- ✗ Not photorealistic, still synthetic environments.
- ✗ No human data gathered yet.

References

References

S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, “Semantic scene completion from a single depth image,” *CVPR*, 2017.

S. Laine, S. Siltanen, T. Lokki, and L. Savioja, “Accelerated beam tracing algorithm,” *Applied Acoustics*, vol. 70, no. 1, pp. 172–181, 2009.

M. Johnson, K. Hofmann, T. Hutton, and D. Bignell, “The malmo platform for artificial intelligence experimentation,” in *IJCAI*, 2016.

C. Beattie, J. Z. Leibo, D. Teplyashin, T. Ward, M. Wainwright, H. Küttler, A. Lefrancq, S. Green, V. Valdés, A. Sadik *et al.*, “Deepmind lab,” *arXiv preprint arXiv:1612.03801*, 2016.

M. Kempka, M. Wydmuch, G. Runc, J. Toczek, and W. Jaśkowski, “ViZDoom: A Doom-based AI research platform for visual reinforcement learning,” in *Computational Intelligence and Games*, 2016.

Y. Wu, Y. Wu, G. Gkioxari, and Y. Tian, “Building generalizable agents with a realistic and rich 3d environment,” *arXiv preprint arXiv:1801.02209*, 2018.

M. Savva, A. X. Chang, A. Dosovitskiy, T. Funkhouser, and V. Koltun, “MINOS: Multimodal indoor simulator for navigation in complex environments,” *arXiv:1712.03931*, 2017.

A. Dries, A. Kimmig, W. Meert, J. Renkens, G. Van den Broeck, J. Vlasselaer and L. De Raedt. “ProbLog2: Probabilistic logic programming,” *Lecture Notes in Computer Science, 9286*, pp. 312–315, Springer, 2015.

S. Brodeur, E. Perez, A. Anand, F. Golemo, L. Celotti, F. Strub, J. Rouat, H. Larochelle, and A. C. Courville, “HoME: a household multimodal environment,” *arXiv:1711.11017*, 2017.

H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. C. Courville, “Guesswhat?! visual object discovery through multi-modal dialogue,” in *CVPR*, 2017.

Thank you!

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