Language Grounding

Alane Suhr
March 3, 2023
CSE 447

World Scopes

(Bisk et al., 2020)

1. Corpora and representations: curated resources used for parsing, lexical semantics
2. The written world: large unstructured collections of texts used for language modeling, text understanding
3. The world of sight and sounds: multimodal resources pairing language and vision, speech, etc.
4. Embodiment and action: language in dynamic environments
5. The social world: language as it is used and learned in interaction with people

Experience grounds language
(Bisk et al. 2020, EMNLP)

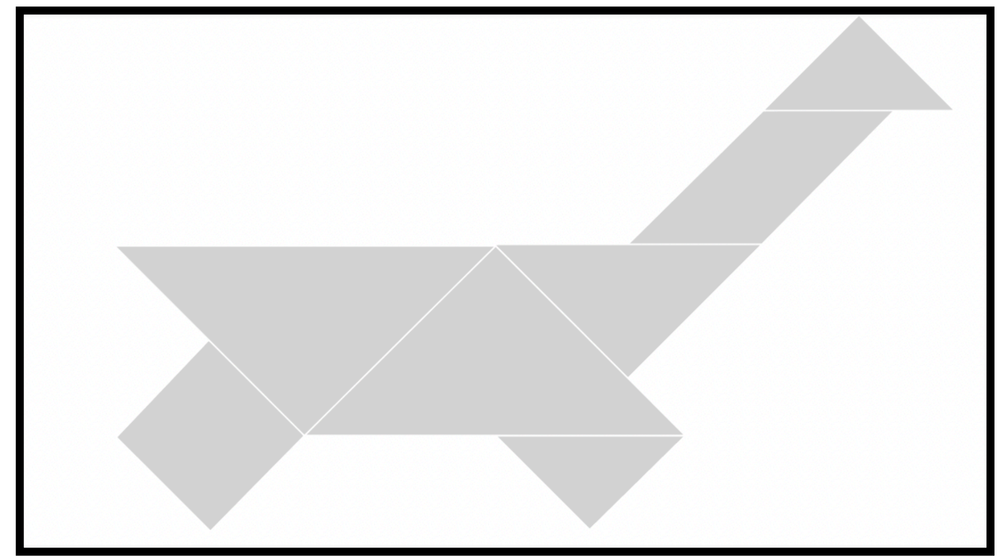
WS3: Multimodal Corpora

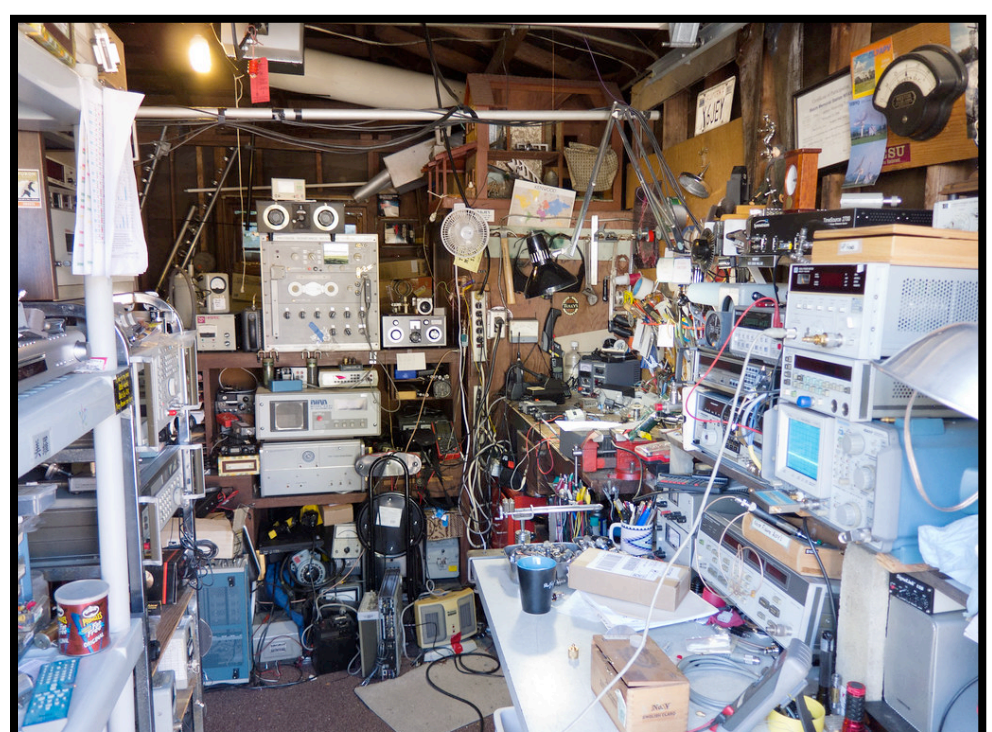
(Focusing on language and images)

What do we want our systems to do?

- Identify concepts in images
- Describe images
- Jointly reason about text and images
- Generate images given a description

Identifying Concepts in Images





Identifying Concepts in Images

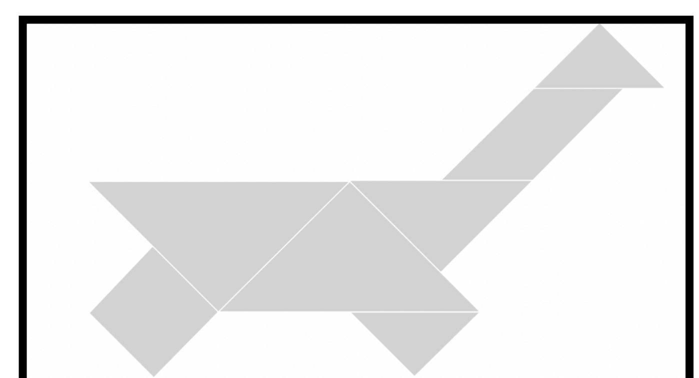
Image classification:
What object(s) is/are in the image? What is the most salient feature?

- Pascal VOC (Everingham et al. 2010, IJCV)
- ImageNet (Deng et al. 2009, CVPR)
- Microsoft COCO (Lin et al. 2014)
- KiloGram (Ji et al. 2022, EMNLP)

cat

workshop

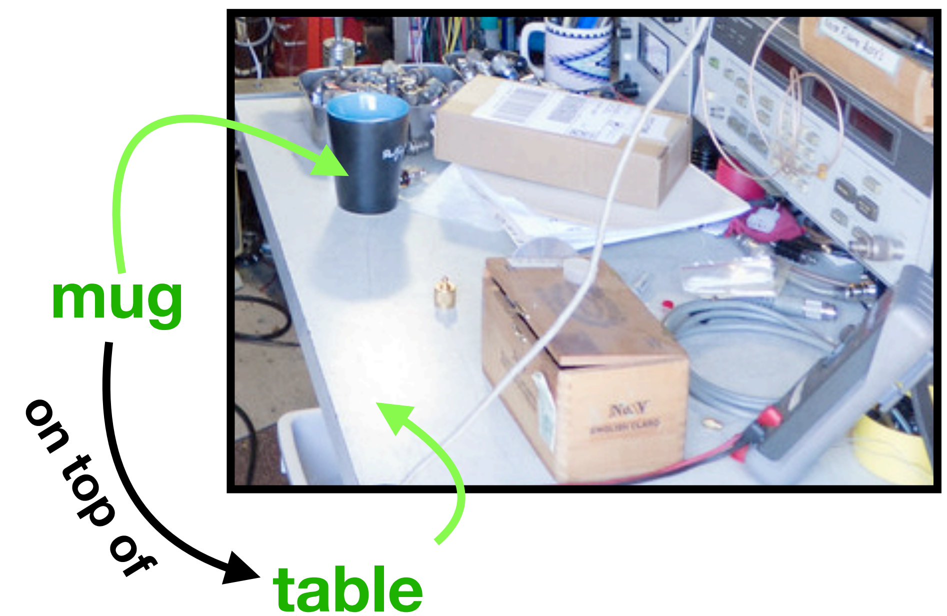
dinosaur

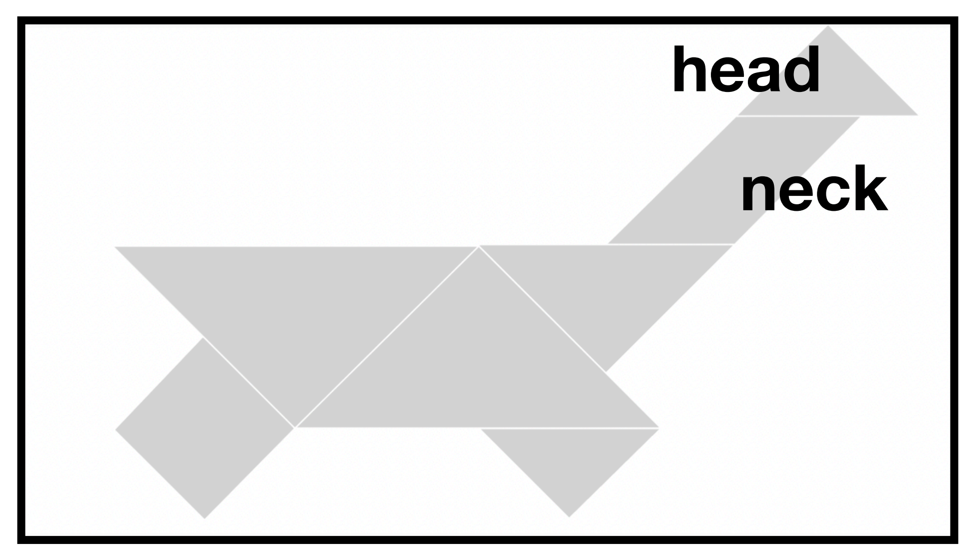


Identifying Concepts in Images

Scene graph generation:
What's the relationship between objects in the image? What are their parts?

- Visual Genome (Krishna et al. 2016)
- KiloGram (Ji et al. 2022, EMNLP)

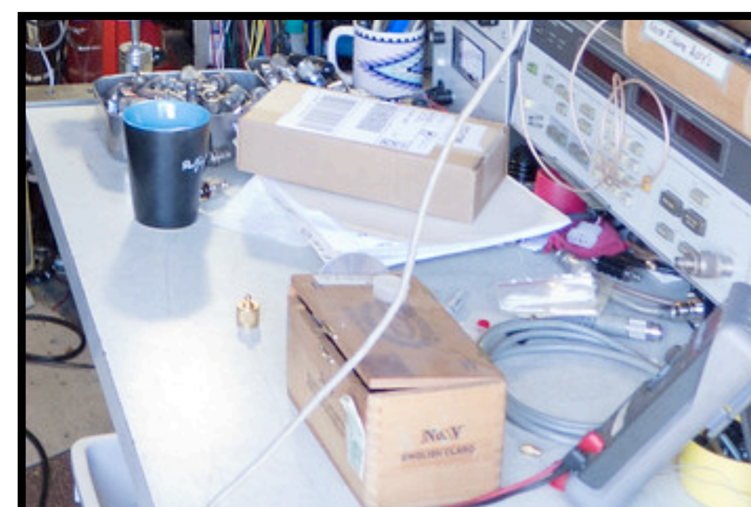




Describing Images

- **General-purpose captioning:** evaluation is difficult!
  - Microsoft COCO Captions (Chen et al. 2015)
  - Conceptual Captions (Sharma et al. 2018, ACL)
- **Context-dependent descriptions:** text has a purpose — easier to evaluate
  - VizWiz (Bigham et al. 2010, UIST)
  - ReferItGame (Kazemzadeh et al. 2014, EMNLP)
  - CapWAP (Fisch et al. 2020, EMNLP)

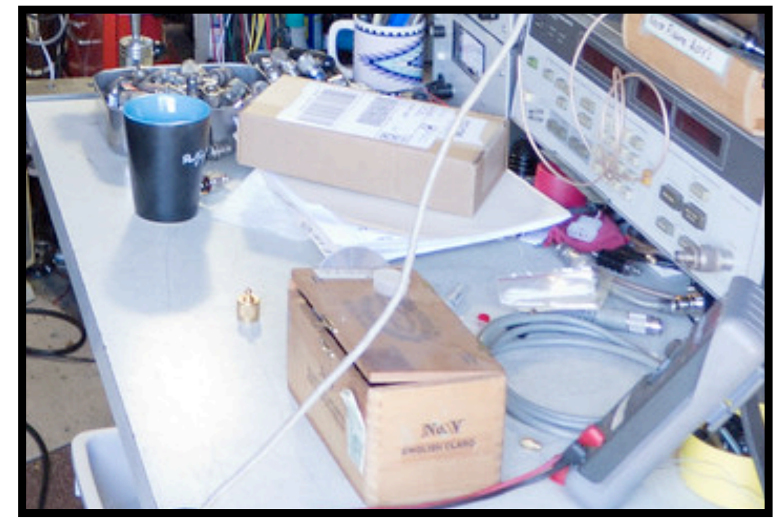
The cat is sleeping.



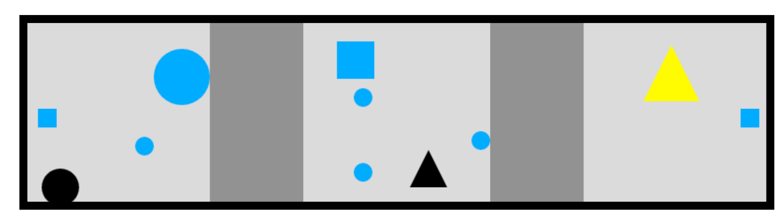
Your blue mug is on the table.

Jointly Reasoning about Text and Images

- **Visual question answering:** map text and image to natural language answer
  - VQA (Agrawal et al. 2015, ICCV)
  - CLEVR (Johnson et al. 2017, CVPR), GQA (Hudson and Manning 2019, NeurIPS)
  - VCR (Zellers et al. 2019, CVPR), Sherlock (Hessel et al. 2022, ECCV)
- **Image-text entailment:** determine whether text describes image
  - NLVR (Suhr et al. 2017, ACL), NLVR2 (Suhr et al. 2019, ACL)
  - MaRVL (Liu et al. 2021, EMNLP)
  - CLIP (Radford et al. 2021)



Q: How many mugs are there?
A: two



There is exactly one black triangle not touching any edge:
True

Generating Images from Text

Tools that allow image generation / editing conditioned on text

- DALL-E (OpenAI)
- Stable diffusion (Stability.AI)

Someone giving a virtual talk in a natural language processing class

Modeling Methods

- **Joint vs. separate image and text representations:** whether to learn text/image features independently, or jointly — LXMERT (Tan and Bansal 2019, EMNLP) vs. CLIP (Radford et al. 2021)
- **Masked autoencoding:** learn to reconstruct training data that has been perturbed, e.g., by "masking" words or image patches (Wang et al. 2022)
- **Diffusion models:** latent variable model trained using variational inference that iteratively generate images by denoising step-by-step (Ho et al. 2020, NeurIPS)
- **Modalities beyond vision:** video, speech, databases, etc.

WS4: Embodiment and Action

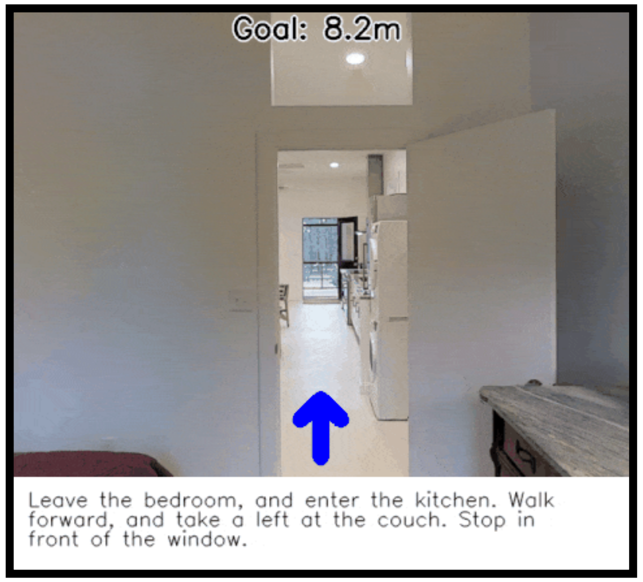
Why embodiment?

- **Embodiment:** an agent is able to manipulate its environment by taking action
- With static environments, agents are not evaluated on their ability to generalize to new environment states due to *world dynamics*
- Our language-using agents should be able to act in the world they share with us
- This requires them to take into account both **perception** and how their **actions influence the world state**

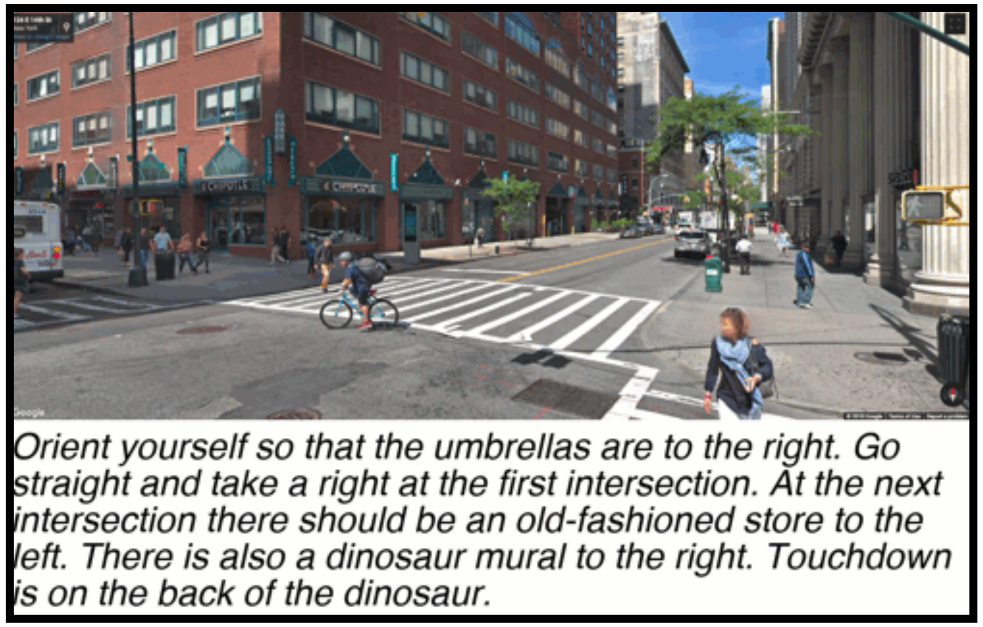
Vision-Language Navigation

- **Task:** navigate a *static* environment given a natural language instruction
- **Evaluation:** did agent end up in the correct location? Did it follow the correct path?
- **Datasets:** instructions and gold-standard action sequences or stopping positions
    - SAIL (MacMahon et al. 2006, AAAI)
    - Room2Room (Anderson et al. 2018, CVPR)
    - Touchdown (Chen et al. 2018, CVPR)

SAIL



Room2Room



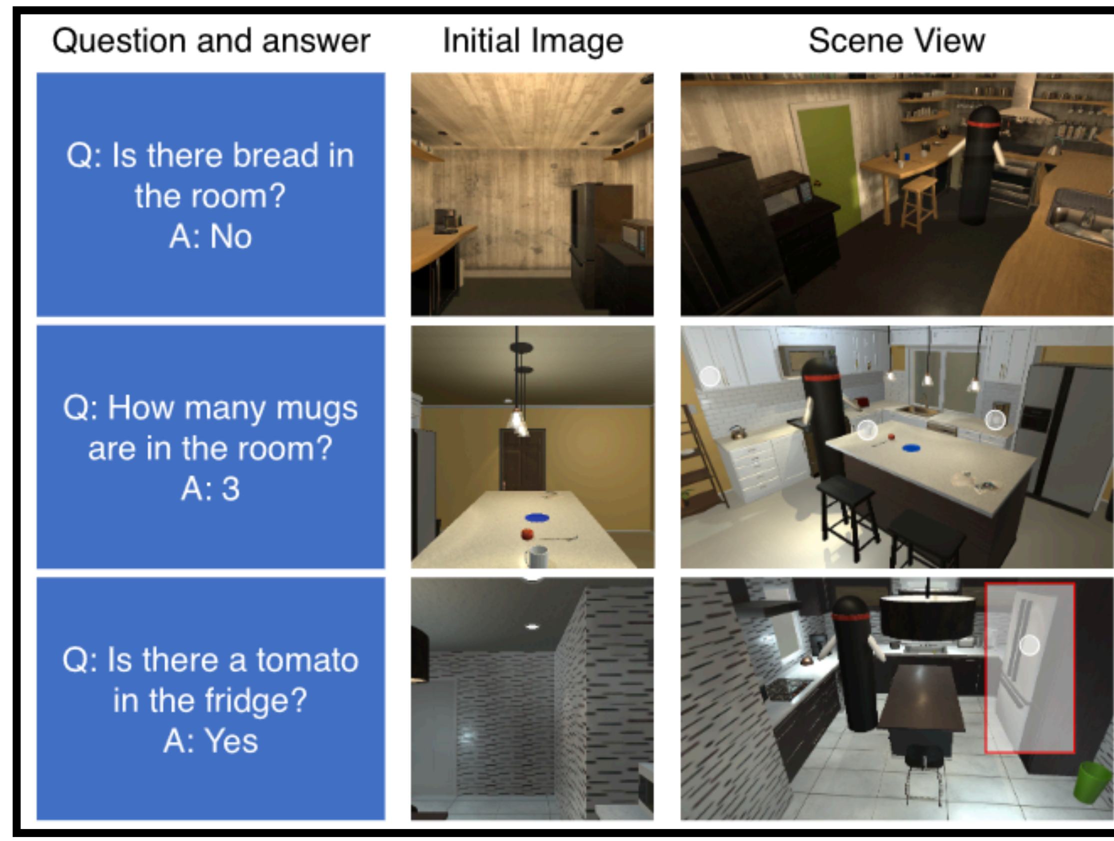
Touchdown

Embodied Question Answering

- **Task:** navigate a *static* environment until a question can be answered by the agent

- **Evaluation:** did agent give the correct answer?

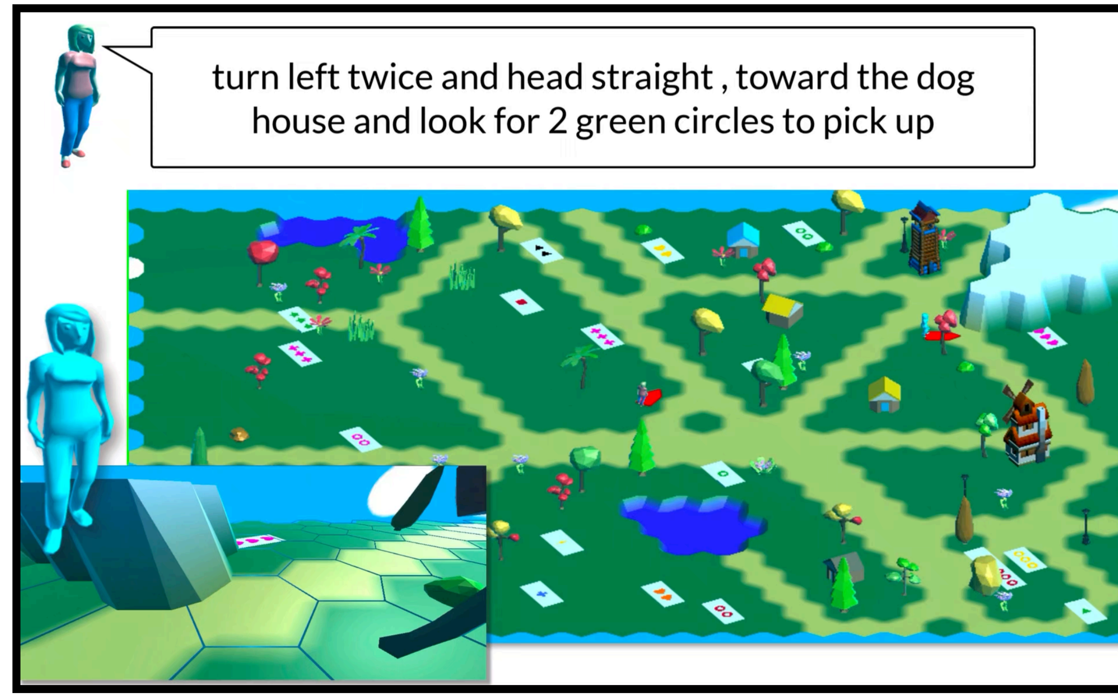
- **Datasets:** questions and correct answers
  - EQA (Das et al. 2018, CVPR)
  - IQA (Gordon et al. 2018, CVPR)



IQA

Manipulable Environments

- **Task:** act in a *dynamic* environment to execute a natural language instruction
- **Evaluation:** are we in the correct final state?
- **Datasets:** instructions and gold-standard action sequences or stopping positions
  - SCONE (Long et al. 2016, ACL)
  - CerealBar (Suhr et al. 2019)
  - ALFRED (Shridhar et al. 2020)
  - MindCraft (Bara et al. 2021)



↑ **CerealBar**

Alfred ↓

Put the cup with the knife on the table.

Learning Methods: Imitation Learning

- Goal: match what a human would do as closely as possible (hence “imitation”)
- Training data: either exact sequences the instruction-follower should take, or an “oracle” that tells it what it should do in specific situations
- Learning style: supervised learning; given some environment state, model gets direct supervision on the action to take

Learning Methods: Reinforcement Learning

- Goal: optimize some external reward
- Training data: current version of the agent takes actions the environment given a training instruction, and receives a scalar reward (e.g., in [-1, 1])
- Rewards can be derived from the training dataset, e.g., how close the agent is getting to the goal state
- Learning style: lots of options coming from RL; policy gradient, PPO, etc.

Instruction-Following in the Real World

SayCan (Ahn et al. 2022)

WS5: Human-Agent Language-Based Interaction

Why interaction?

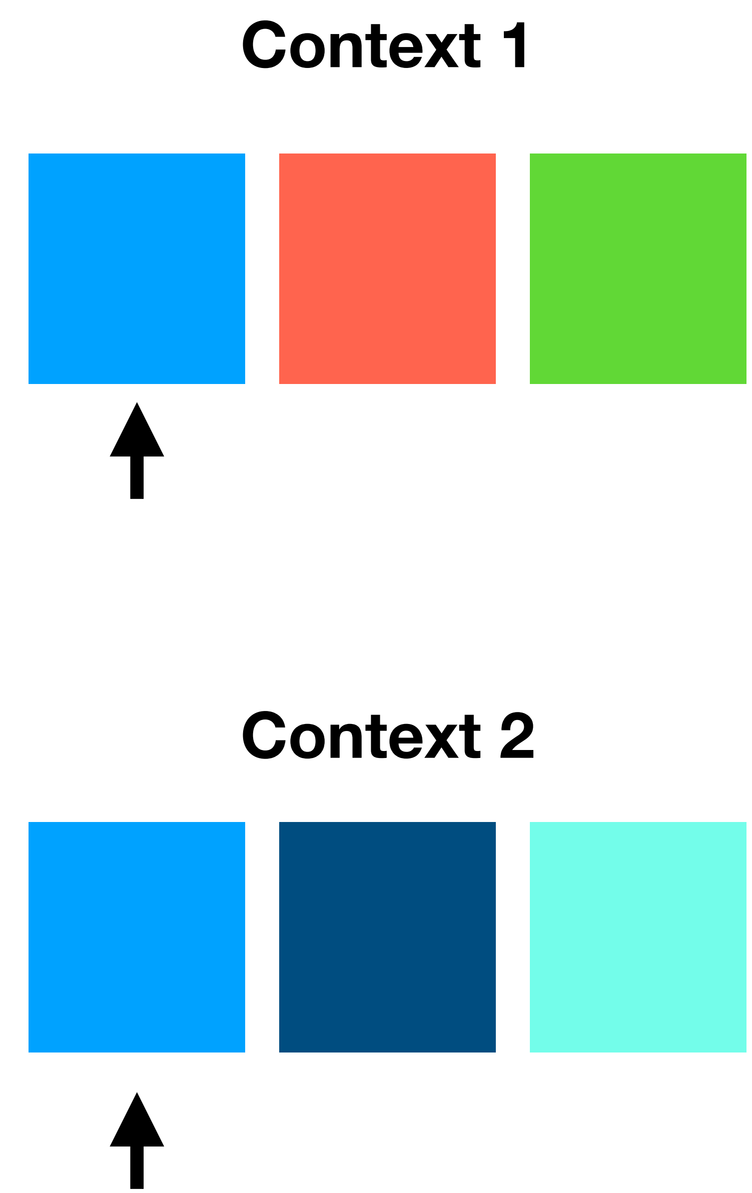
- **Interaction:** two or more agents act in an environment, and observe each others' actions
- Without interaction, agents are not exposed to *dynamics* that arise as agents adapt to one another
- Our language-using agents should be able to coordinate with us, and learn from us, through language
- This requires them to also take into account behavior of the other agents

Collaborative Interactions

- Two agents act in a shared world towards a common goal
- Coordinate their actions using natural language
- Tasks to study: **language understanding** and **language generation**
- Opportunities within collaborative interactions
  - Dynamics: convention formation, adaptation to mistakes
  - Continual learning through explicit and implicit feedback

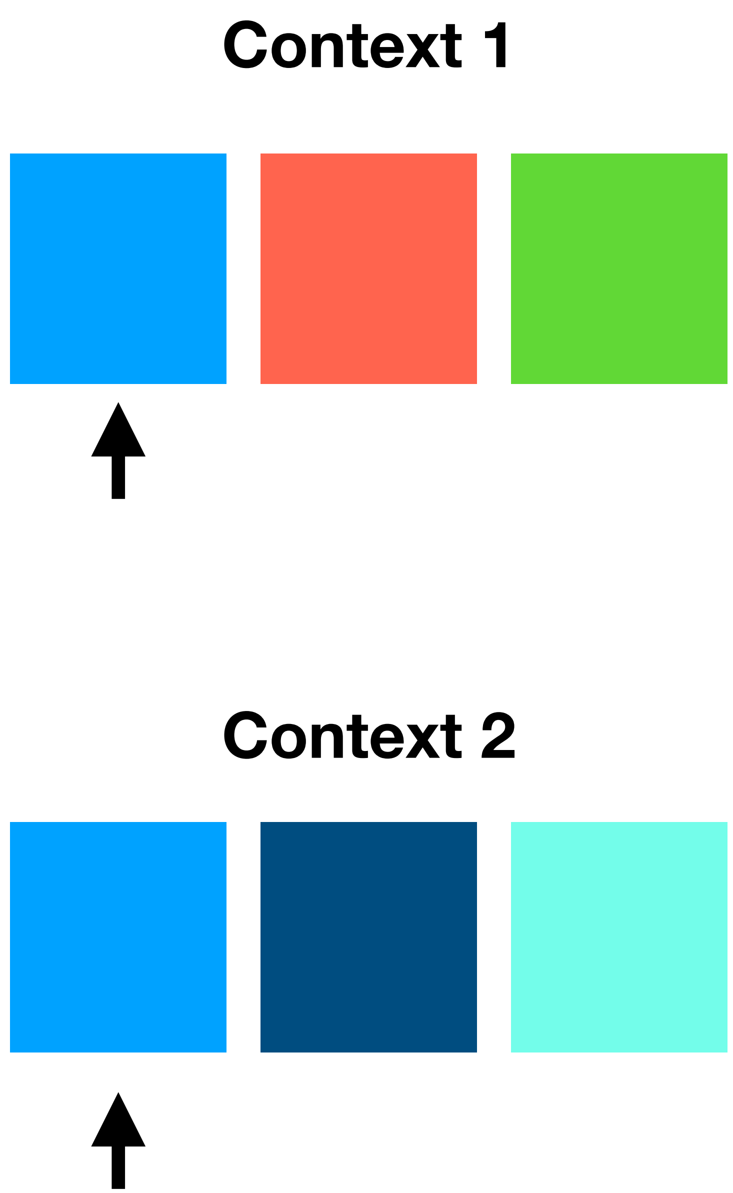
Reference Games

- Task: two agents view identical or similar environments
  - Agent 1 chooses something in the environment
  - Agent 1 writes a referring expression for that thing
  - Agent 2 should try their best to identify what Agent 1 is referring to
- E.g.: sets of colors (Monroe et al. 2017, TACL)
- The expressions generated depend heavily on:
  - Surrounding context in environment
  - What you know about the other agent



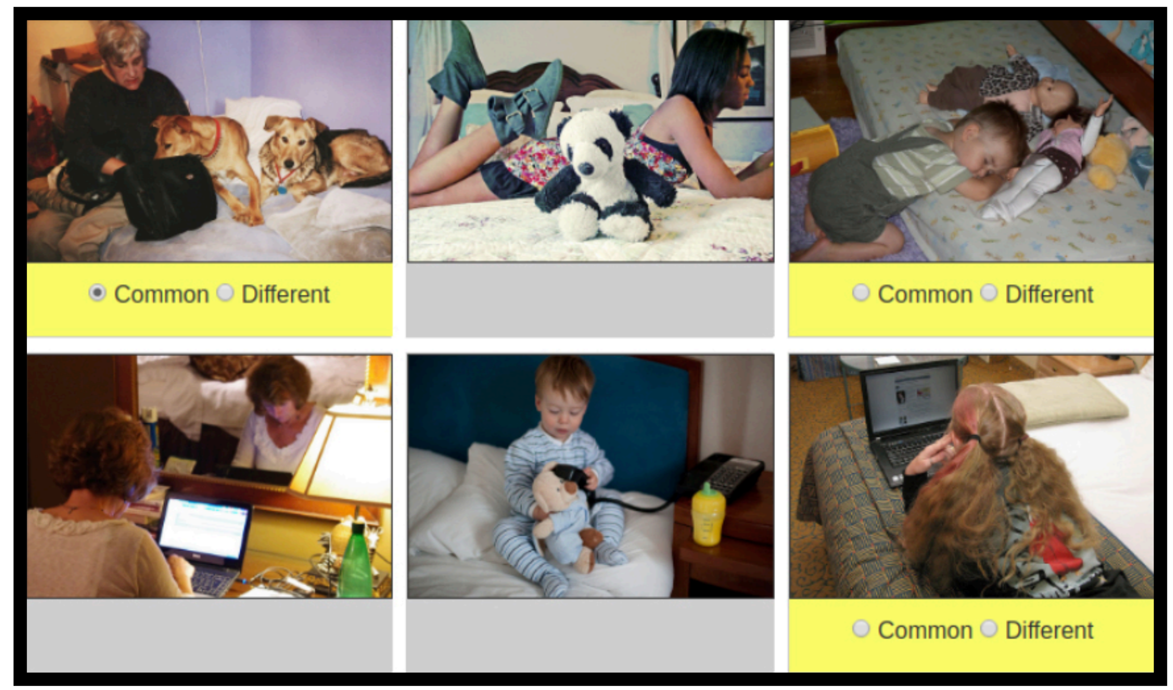
Reference Games

- This requires agents to maintain a model of the other (a.k.a. Theory of Mind)
- In linguistics, this the subject of **pragmatics**
- Formal models describe how we might consider each others' state of mind (e.g., RSA; Frank and Goodman 2012, Science)

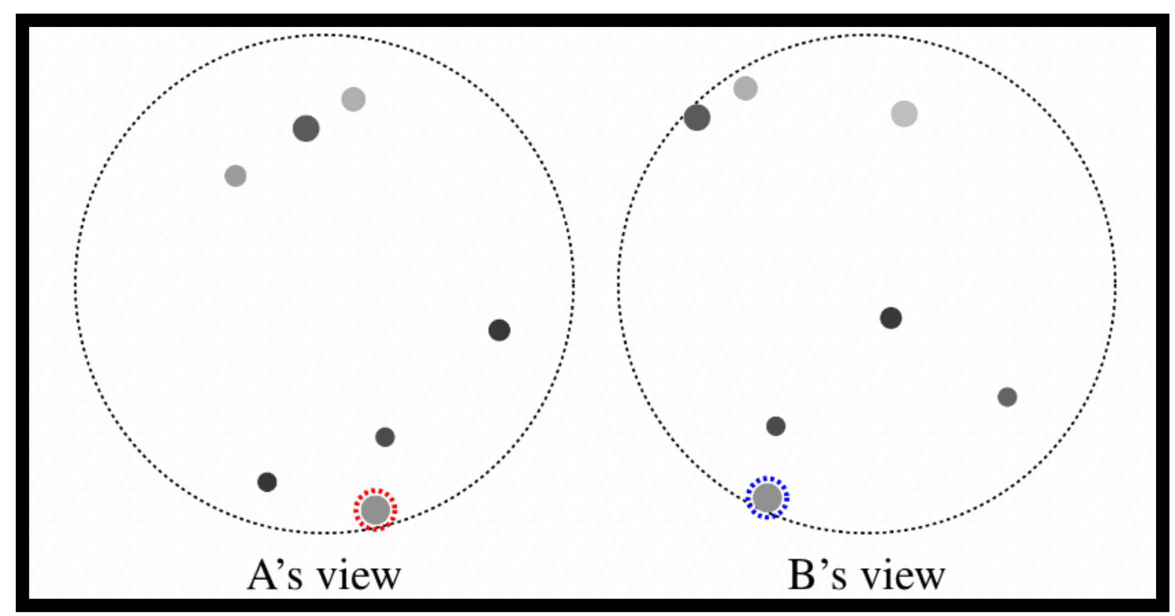


Reference Games

PhotoBook
(Haber et al. 2019, ACL)



OneCommon
(Udagawa and Aizawa 2019, AAAI)



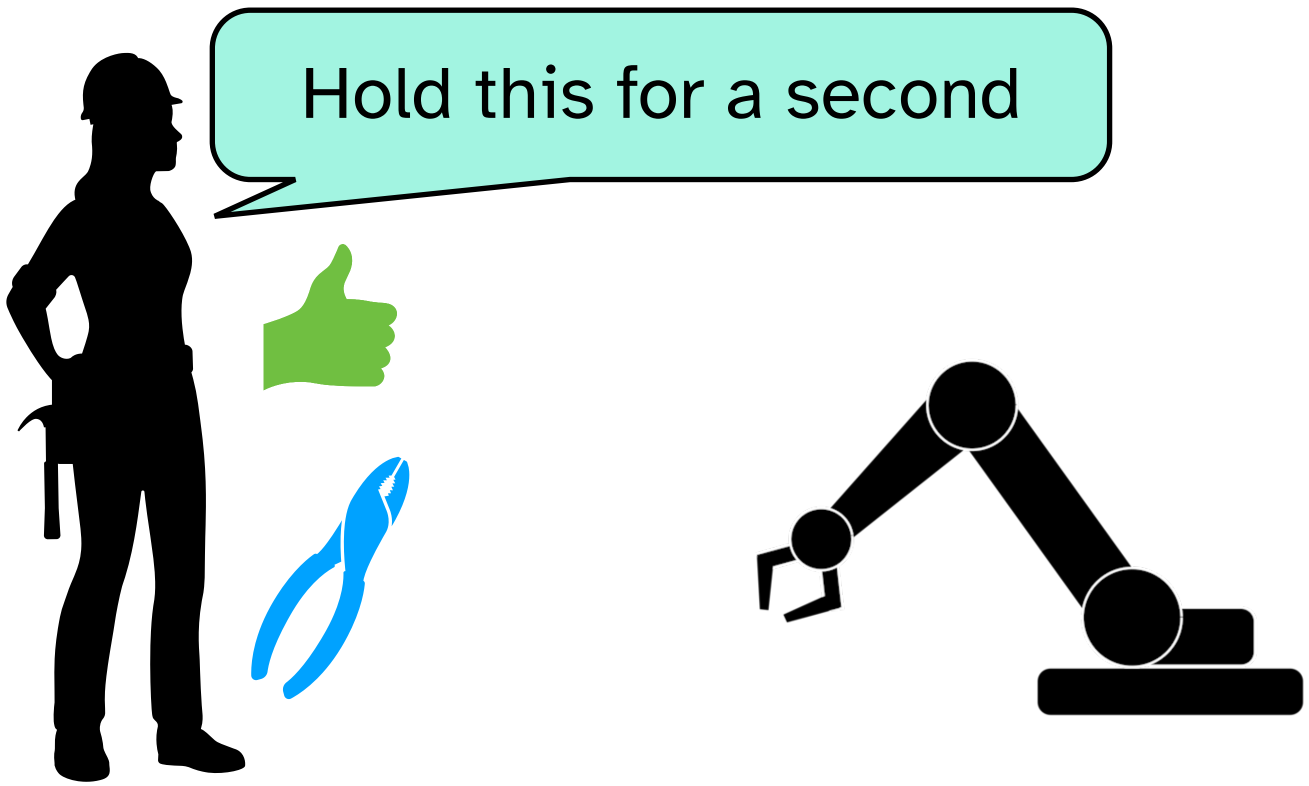
Continual Learning through Interaction

Why continual learning?

- **Continual learning:** agent adapts constantly within and across interactions given explicit/implicit user feedback
- This allows the models to adjust *on the fly* to the user's actual behavior
- Very natural way of learning:
  - We use feedback to drive our own language learning and change
  - We expect our interlocutors to adapt their language via feedback we give
- Our systems should be able to adjust to feedback we provide!

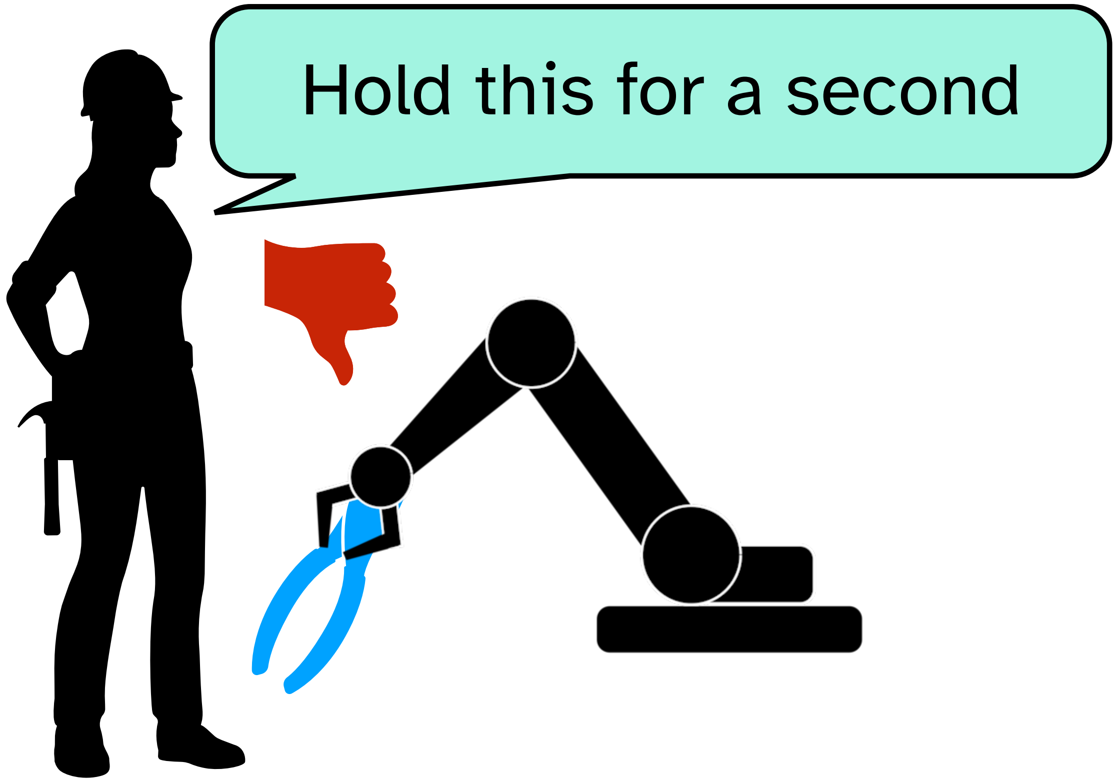
Continual Learning through Interaction

Explicit feedback



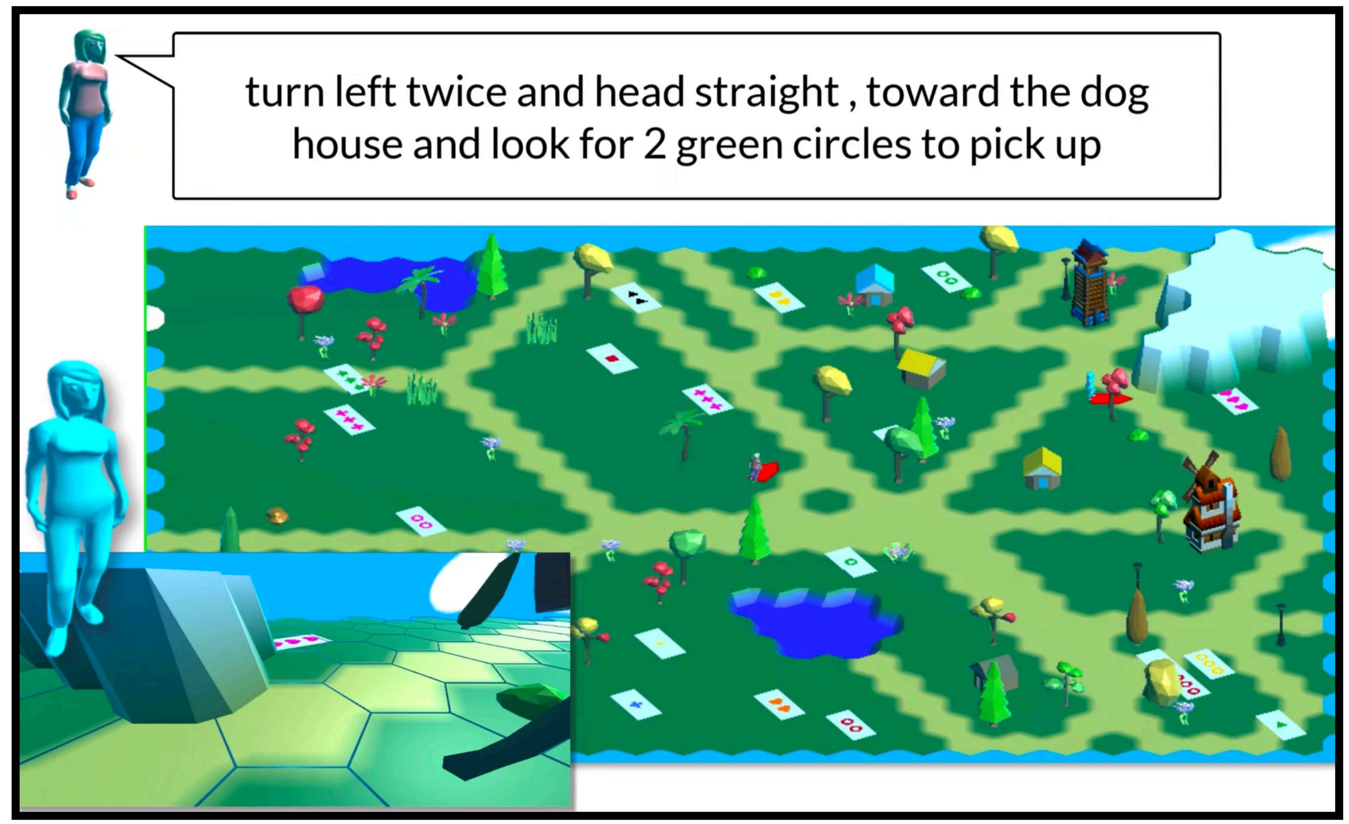
Continual Learning through Interaction

Explicit feedback



Learning from Explicit Feedback

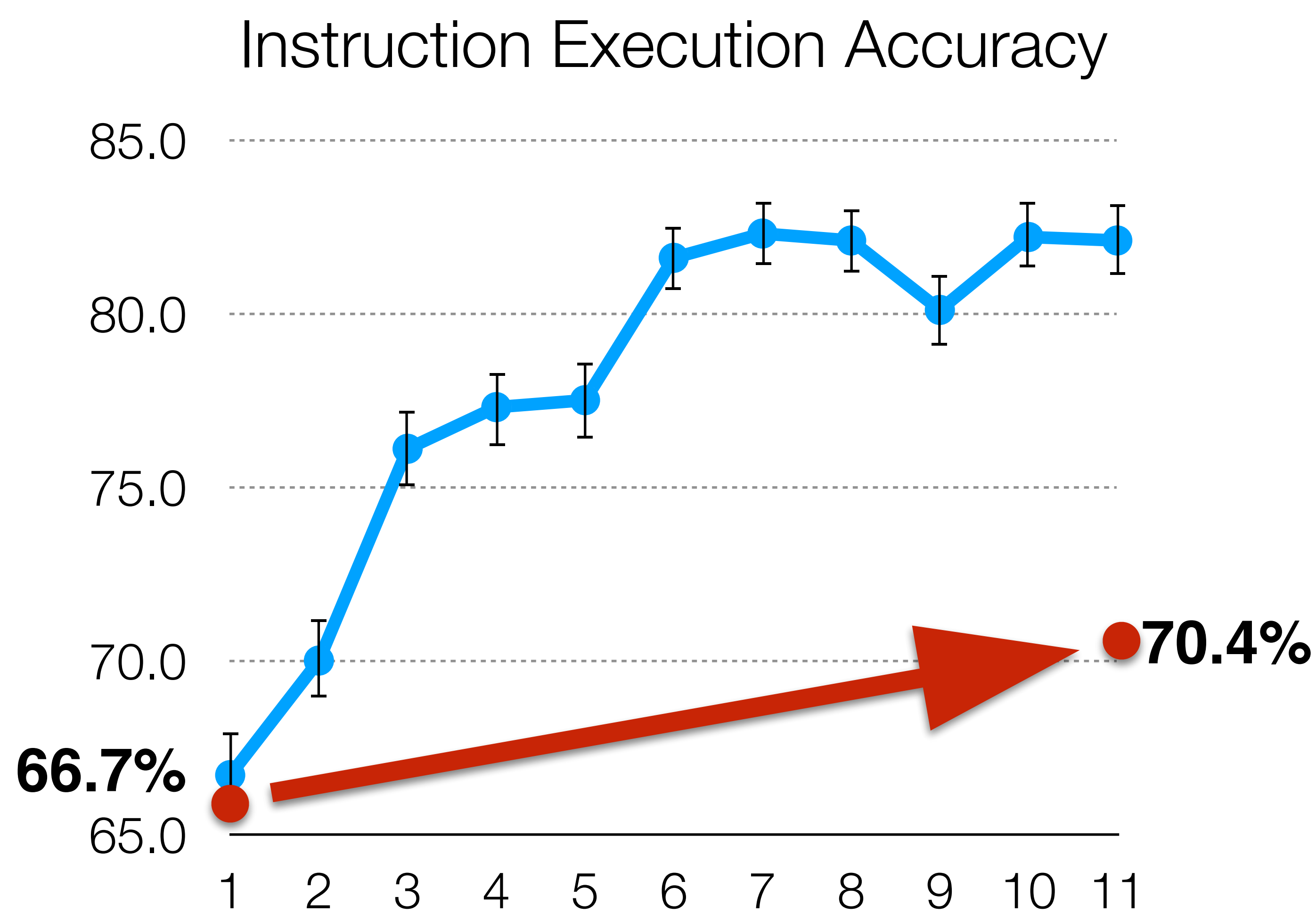
- **Task:** follow instructions
- In live interactions, get users to write new instructions
- Agent maps user-written instructions to actions
- Users provide binary feedback as the agent moves



Suhr and Artzi 2022

Learning from Explicit Feedback

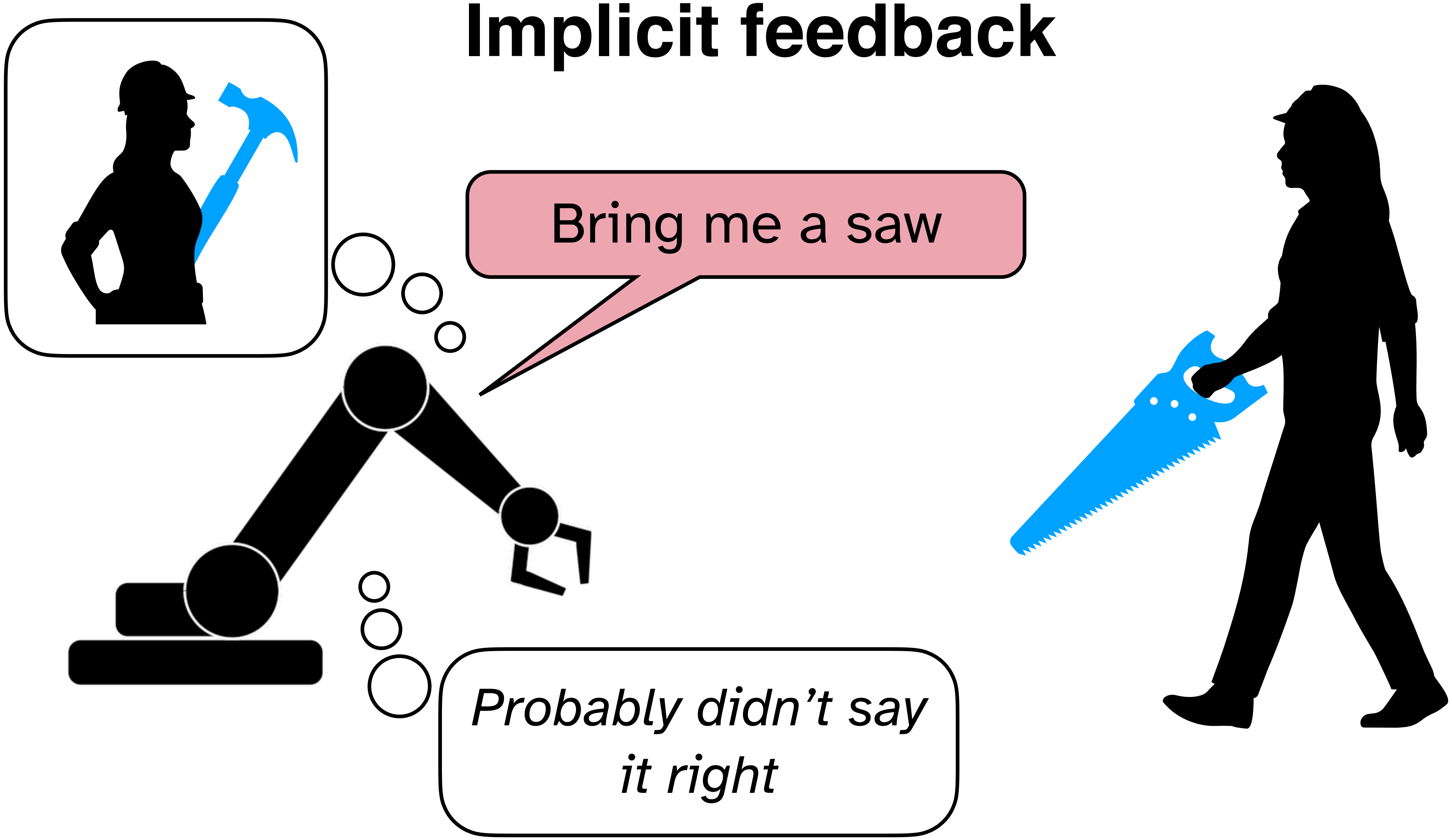
- Over many rounds of human-agent games, rate at which it follows instructions correctly increases!
- Confounding factor: user adaptation
- User adaptation produces an effect, but agent is still improving from the learning process



Suhr and Artzi 2022

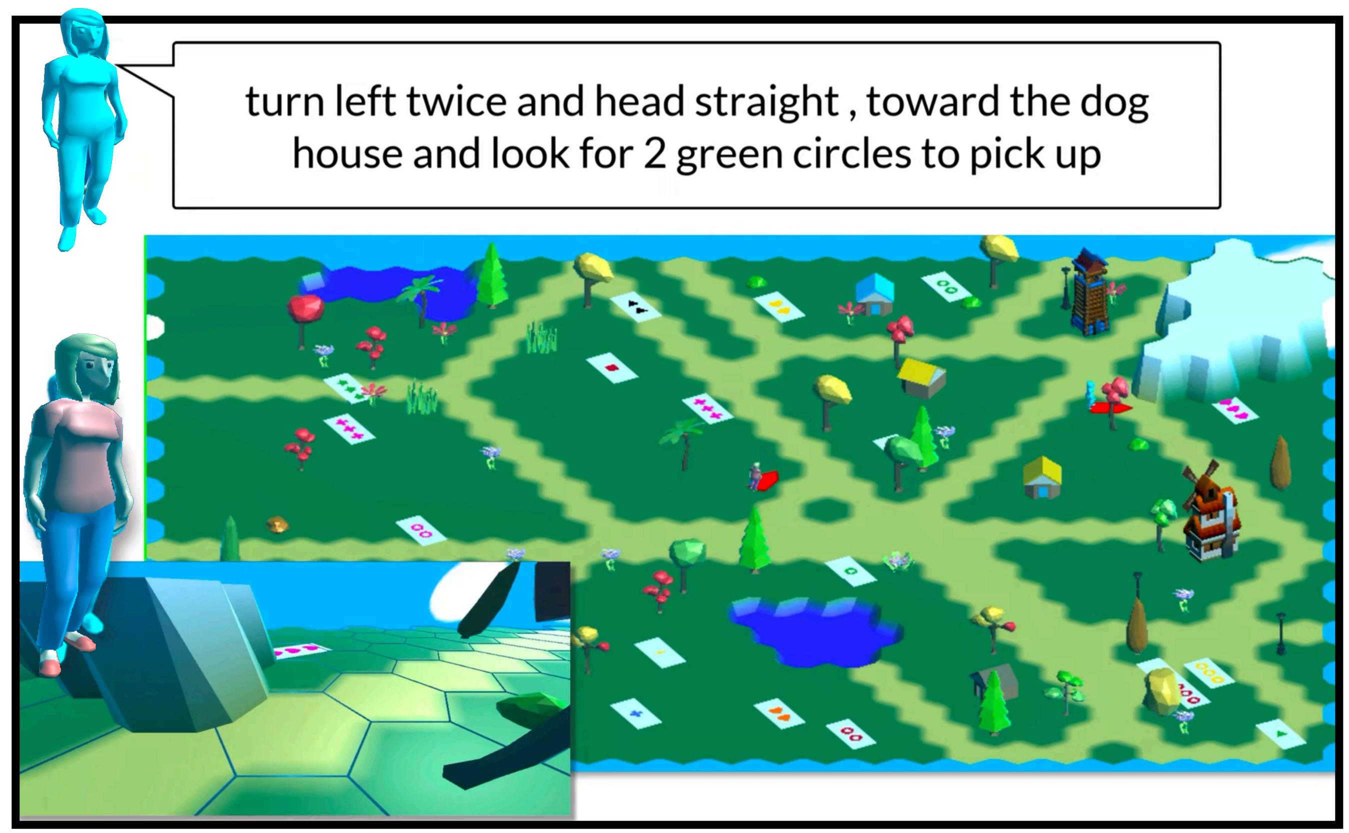
Continual Learning through Interaction

Implicit feedback



Learning from Implicit Feedback

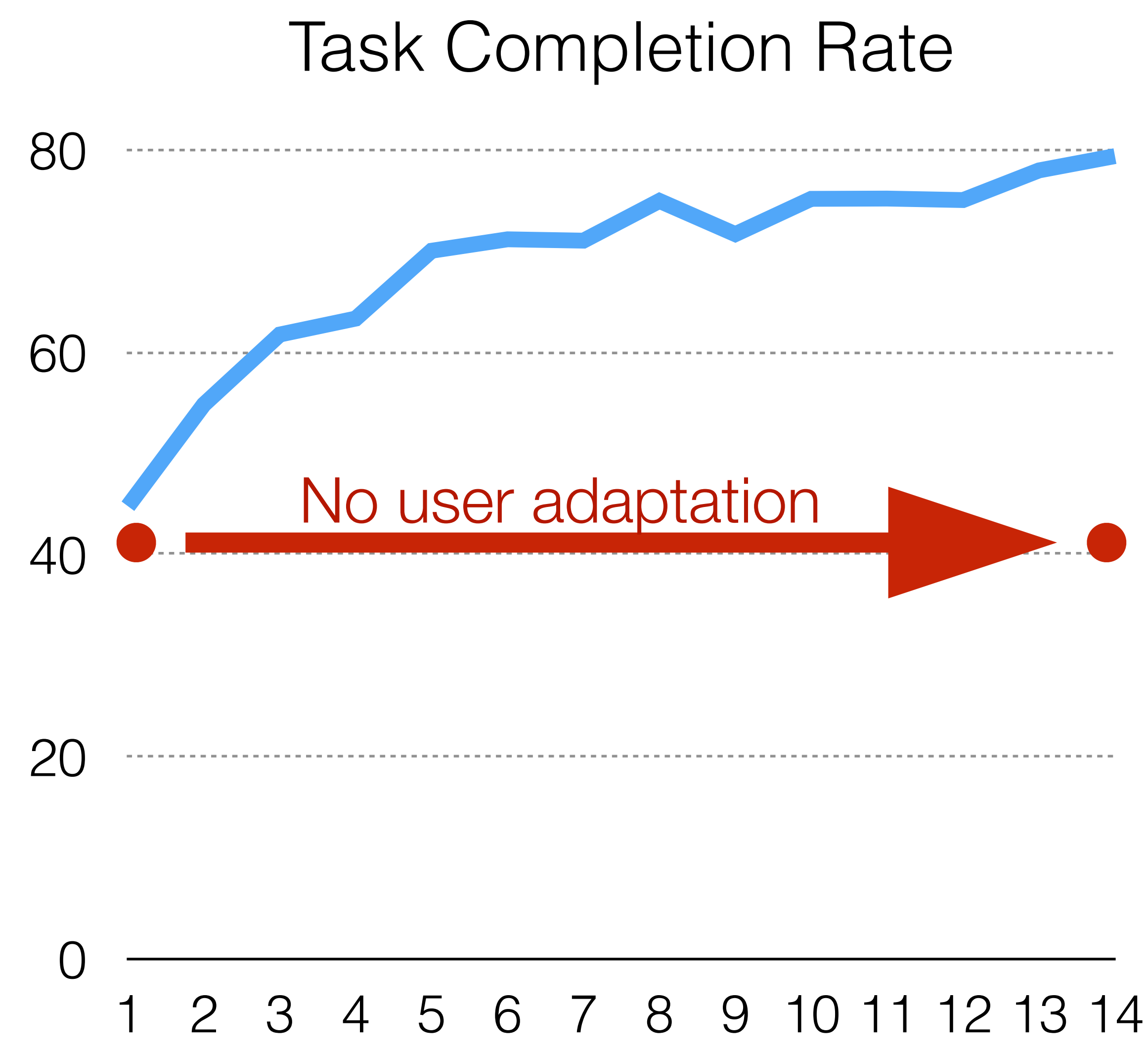
- **Task:** generate instructions
- In live interactions, agent generates an **intent** given game objective
- Agent maps intent to an instruction
- User's response to instruction provides implicit feedback on how correct the instruction was wrt. intent



Kojima et al. 2021, TACL

Learning from Implicit Feedback

- Model continually improves its ability to convey its intent via natural language
- We did not observe any user adaptation over time to agent-generated instructions

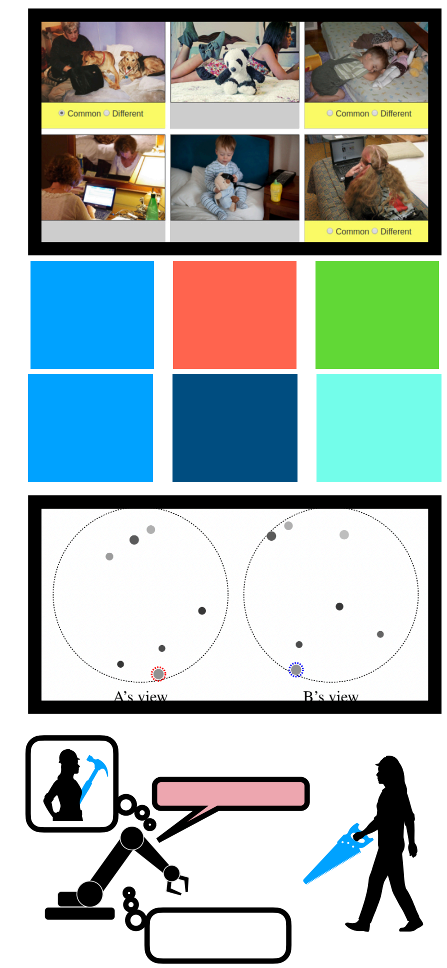


Kojima et al. 2021, TACL

Summary

Multimodal Corpora
- Static datasets
- Static environments

Embodied Corpora
- Static datasets
- Dynamic environments



Interaction
- Dynamic datasets
- Dynamic environments
- Continual learning