Language use in embodied interaction

Alane Suhr

UC Berkeley EECS

Interaction

- **Single-agent**
  - Environment dynamics do not depend on the policy
  - E.g., “agentic” tasks like navigation
- **Multi-agent**
  - Environments dynamics *do* depend on policy
  - E.g., multi-turn or multi-exposure interaction with human users, who can adapt over time
  - Main challenge: second order dynamics arising from mutual adaptation

Thank you to my collaborators at the Dagstuhl Perspectives Workshop on Human in the Loop Learning through Grounded Interaction in Games (organized by Rafaella Bernardi, Julia Hockenmaier, Udo Kruschwitz, Massimo Poesio)

Learning from interaction

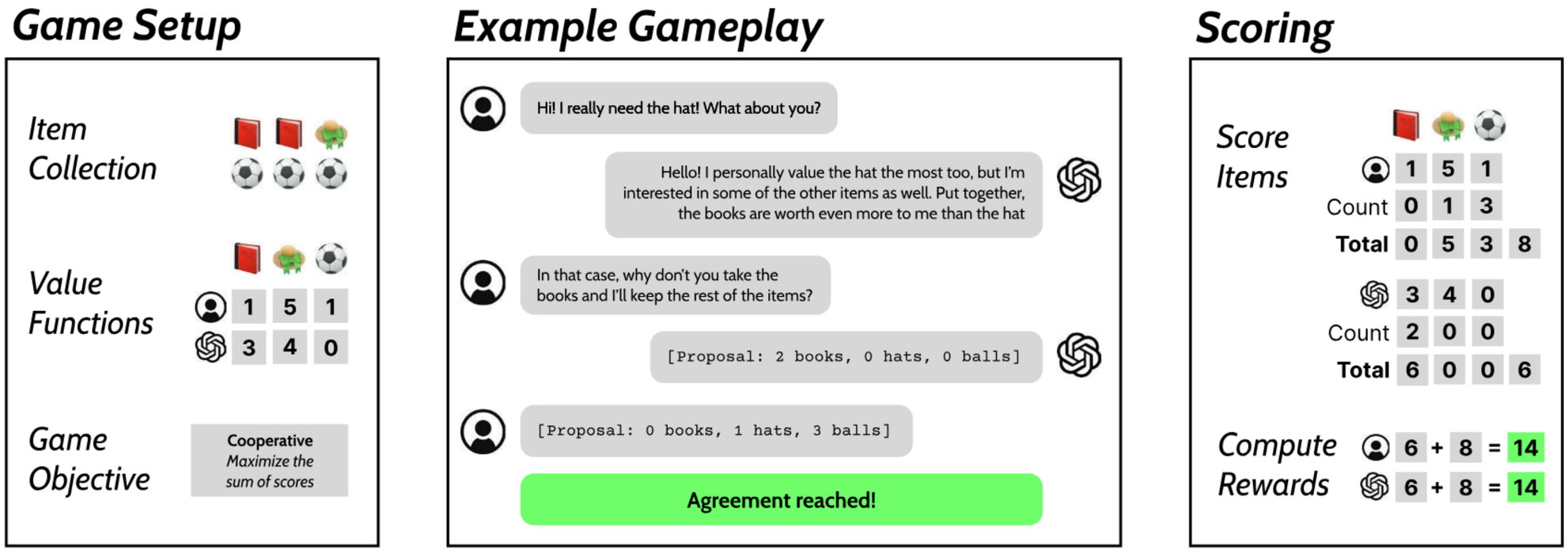
- **Domain-specific skills**
  - “learning to act”
- **Language structure, including syntax, context-independent lexical semantics**
  - “learning a language”
- **In-interaction language use**
  - (non-AI) “alignment” between participants,
  - “conceptual pacts”,
  - “convention formation”

What makes an interaction?

- **Incentives**
- **Constraints and affordances**
  - Environment
  - Communication channel
- **Participants**

Incentive design

- **Continuum between fully collaborative and fully competitive games**
  - E.g., negotiation game



Liao, Tomlin, and Klein 2024

Incentive design

- **Continuum between fully collaborative and fully competitive games**
- **Knowledge of one's own and the others' incentives**
  - Related to the AI "alignment" problem

Incentive design

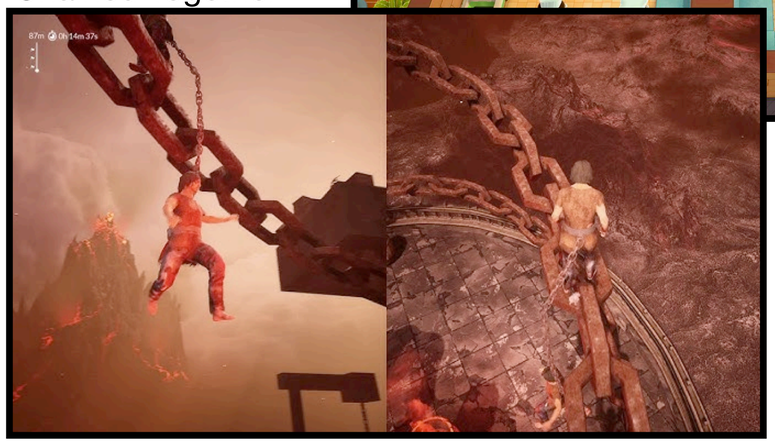
- Continuum between fully collaborative and fully competitive games
- Knowledge of one’s own and the others’ incentives
- Interaction length and timeline

Incentive design

- **Continuum between fully collaborative and fully competitive games**
- **Knowledge of one's own and the others' incentives**
- **Interaction length and timeline**
- **Task difficulty**
  - Time constraints
  - Computational / reasoning complexity
  - Novelty
  - Risk

Overcooked

Chained Together



Incentive design

- **Continuum between fully collaborative and fully competitive games**
- **Knowledge of one's own and the others' incentives**
- **Interaction length and timeline**
- **Task difficulty**
- **Contextual factors**
  - Participant pay and tie to performance
  - Social factors
  - Fun

Environment design

- **Continuity in action and perception**

Environment design

- **Continuity in action and perception**
- **Embodiment**
  - 2D vs. 3D games
  - “Ungrounded” games (e.g., text-based games)

Environment design

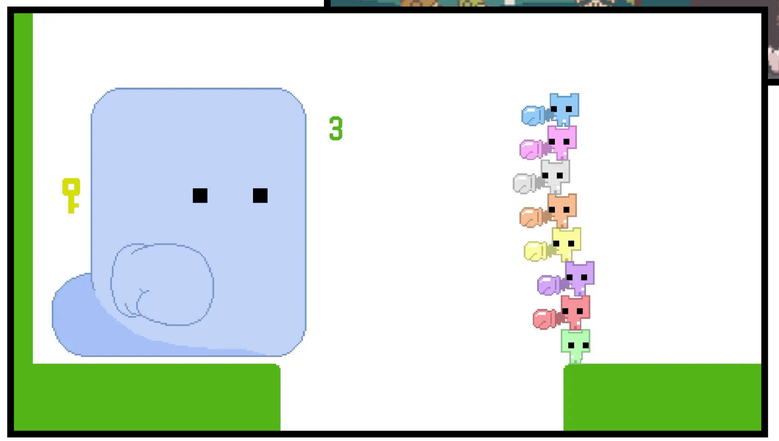
- Continuity in action and perception
- Embodiment
- Open vs. guided world

Environment design

- **Continuity in action and perception**
- **Embodiment**
- **Open vs. guided world**
- **What is and isn’t in the common ground**

Bokura

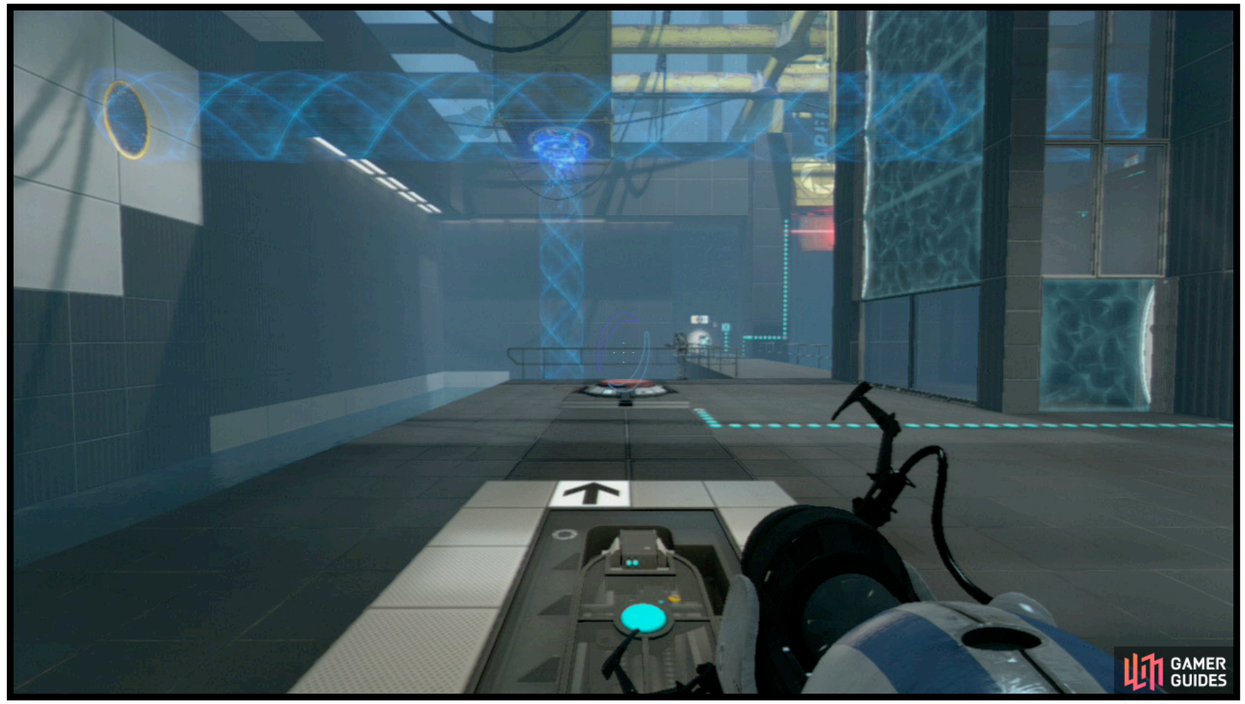
Pico Park



Environment design

- Continuity in action and perception
- Embodiment
- Open vs. guided world
- What is and isn’t in the common ground
- Novel objects and dynamics

Portal 2



Interaction participants

- **Number of participants**
- **Roles**
  - Leaders / followers
  - Instructors / learners
  - User / agent
  - No roles

Interaction participants

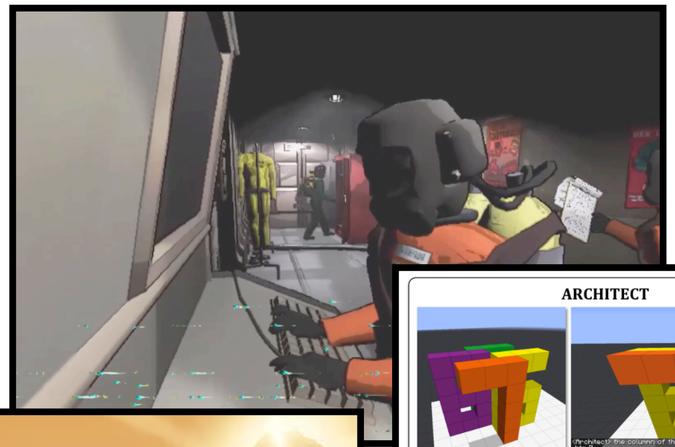
- **Asymmetries in…**
  - Knowledge
  - Affordances
  - Perception

- **Types of asymmetry**
  - Ephemeral (e.g., due to partial observability)
  - Absolute (e.g., due to differences in expertise)

Communication channel

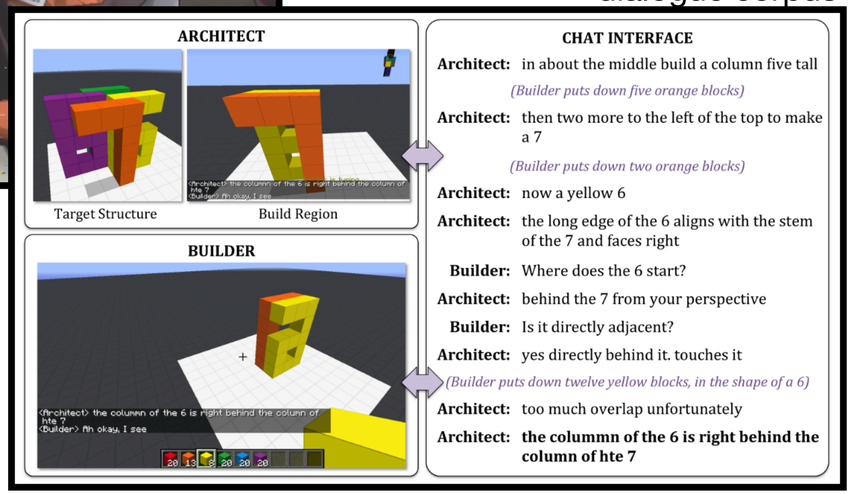
- **Medium**
  - Fully embodied — voice and vision
  - Voice only, with and without proximity
  - Text only
  - Signalling

Lethal Company



Journey

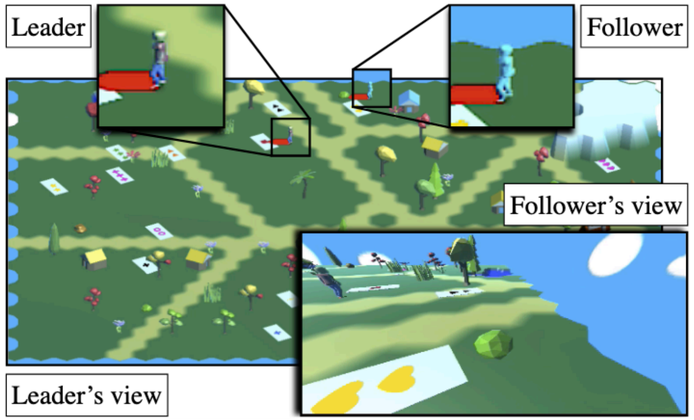
Minecraft dialogue corpus



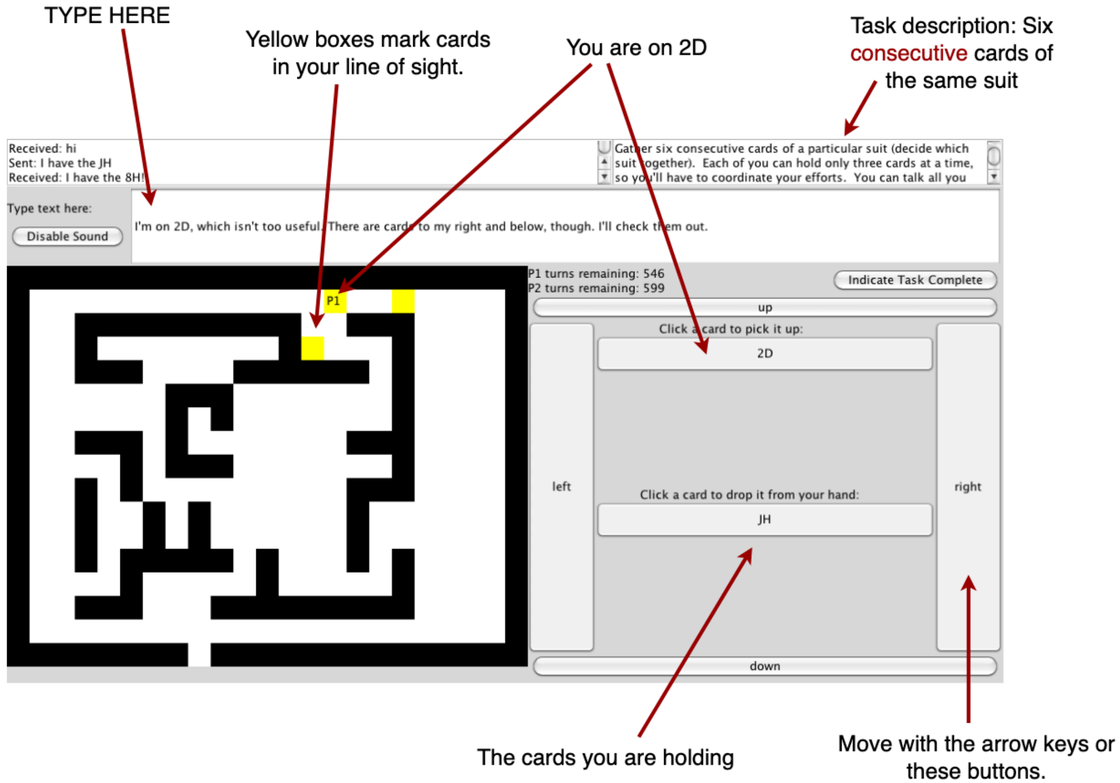
Narayan-Chen et al. 2019

Communication channel

- **Medium**
- **Dynamics and affordances**
  - Enforced turn-taking
  - Latency and noise
  - Asymmetry



...
$\bar{x}_3$: *turn left and head toward the yellow hearts, but don't pick them up yet. I'll get the next card first.*
$\bar{x}_4$: ***Okay, pick up yellow hearts and run past me toward the bush sticking out, on the opposite side is 3 green stars***
`[Set made.  New score:  4]`
...



Potts 2012, Suhr et al. 2019

Communication channel

- **Medium**
- **Dynamics and affordances**
  - Enforced turn-taking
  - Latency and noise
  - Asymmetry
- **Synchrony**

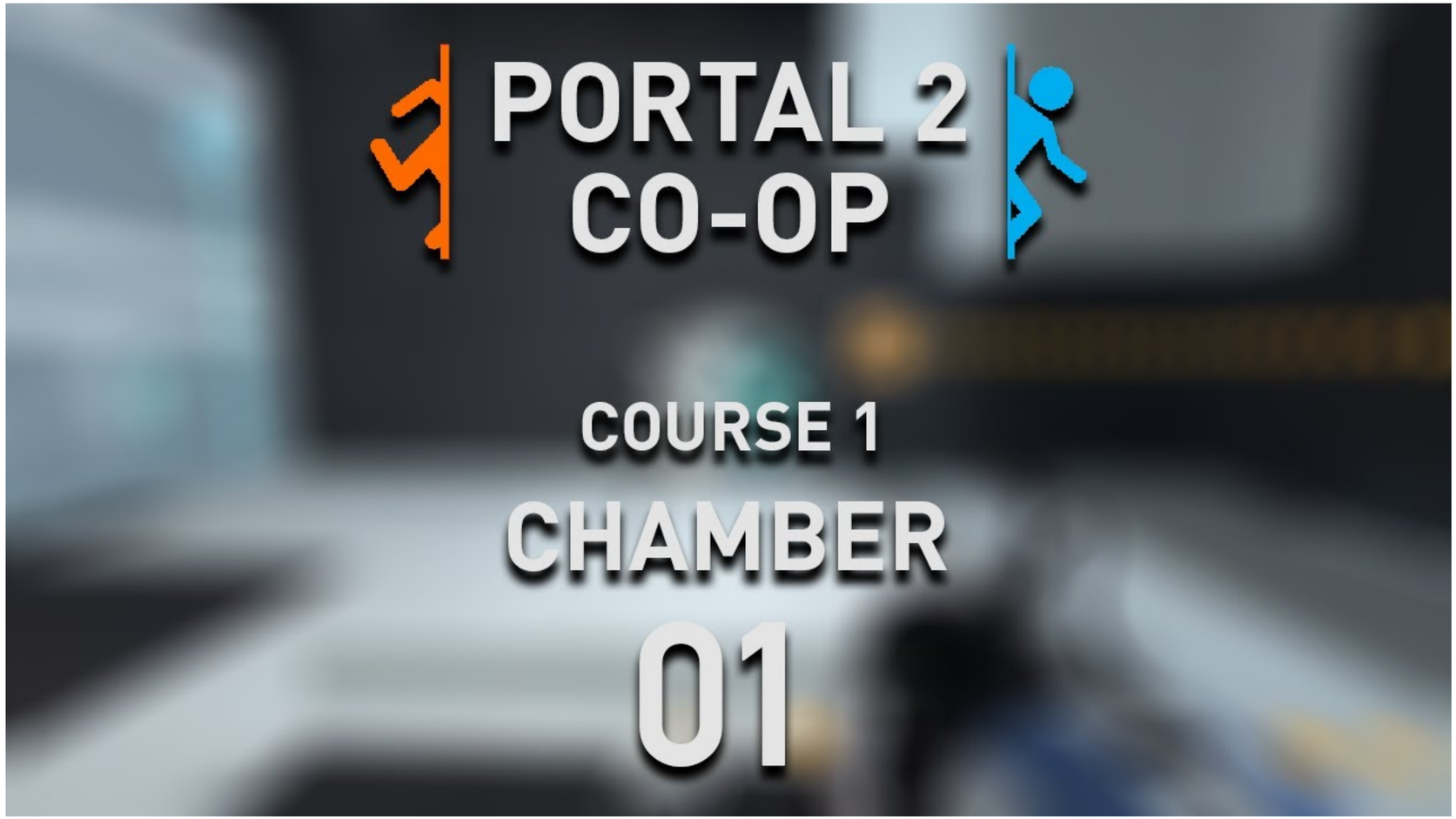
Potts 2012, Suhr et al. 2019

Why do we need language in interaction?

- **To build shared understanding**
  - Bridging asymmetries
  - Resolving uncertainties
- **To coordinate our actions**
  - Via planning
  - Via negotiation

Portal 2 Co-op (2011, Valve)

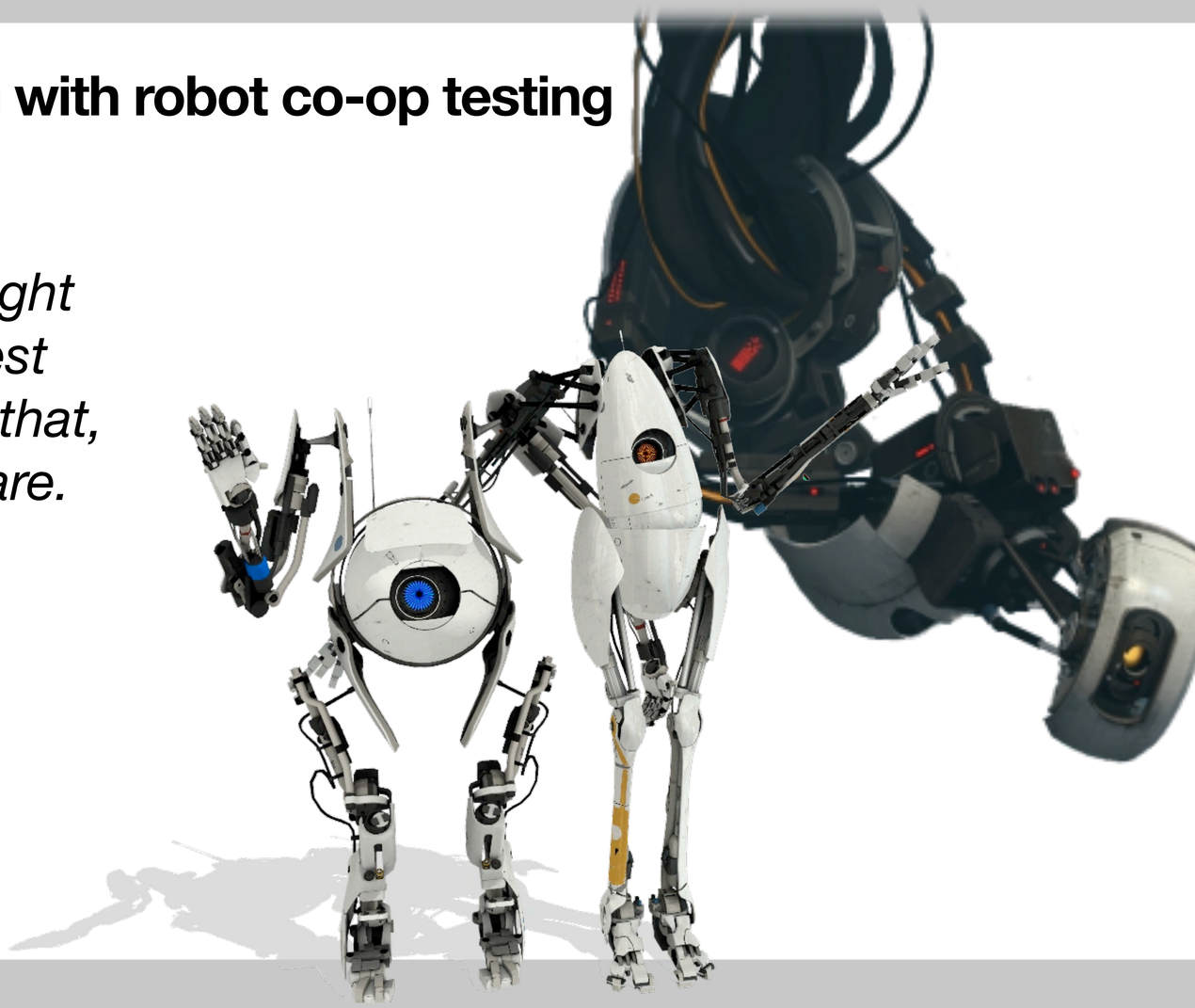
- Two players, symmetric roles and affordances
- Shared 3D world
- Continuous control and perception
- Novel objects, world dynamics
- Goal: navigate to room exit while avoiding hazards



Portal 2 Co-op (2011, Valve)

Story premise: GLaDOS replaces human testing with robot co-op testing

I don’t want to alarm either of you, but we might have a tiny problem. Apparently you can’t test unless you're human. Well - you CAN. It’s just that, results-wise, the physical universe doesn’t care.



Collaborative completions

Orange: *Yeah, maybe just point it at the different things, see …* *what happens.*

Blue: *I (don't know / wanna) see what happens.*

Negotiation of plans

Blue: *Oh my god you should be in charge of this.*
Orange: ((laughs))
Blue: *It’s too much responsibility.*

Abandonments and self-interruptions

Blue: *So if we do that —*
Blue: *Where are the different areas where we can —*
Blue: *Oh.*

Gaze tracking

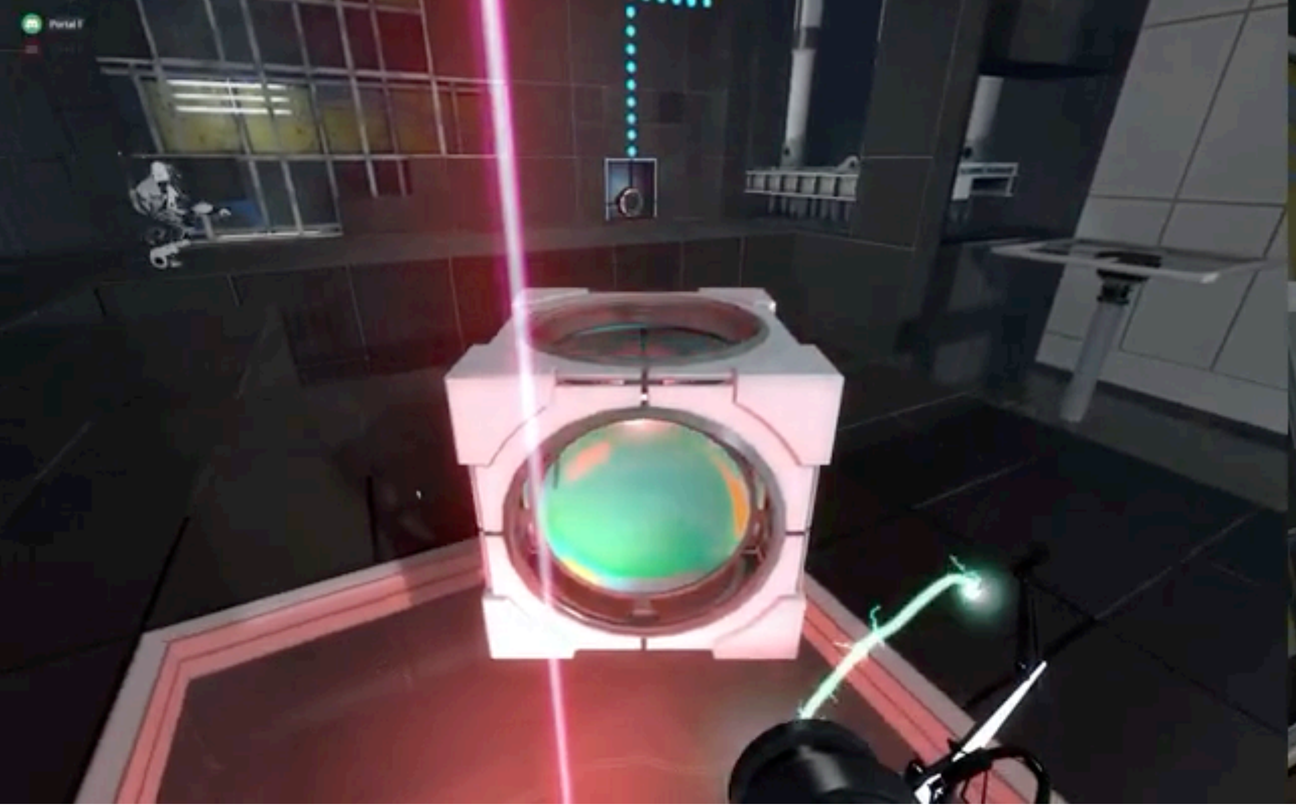
Blue: *Oh, we have to go over there.*
Orange: *Where?*
Orange: [[follows Blue's gaze to other platform]]
Orange: *Okay.*

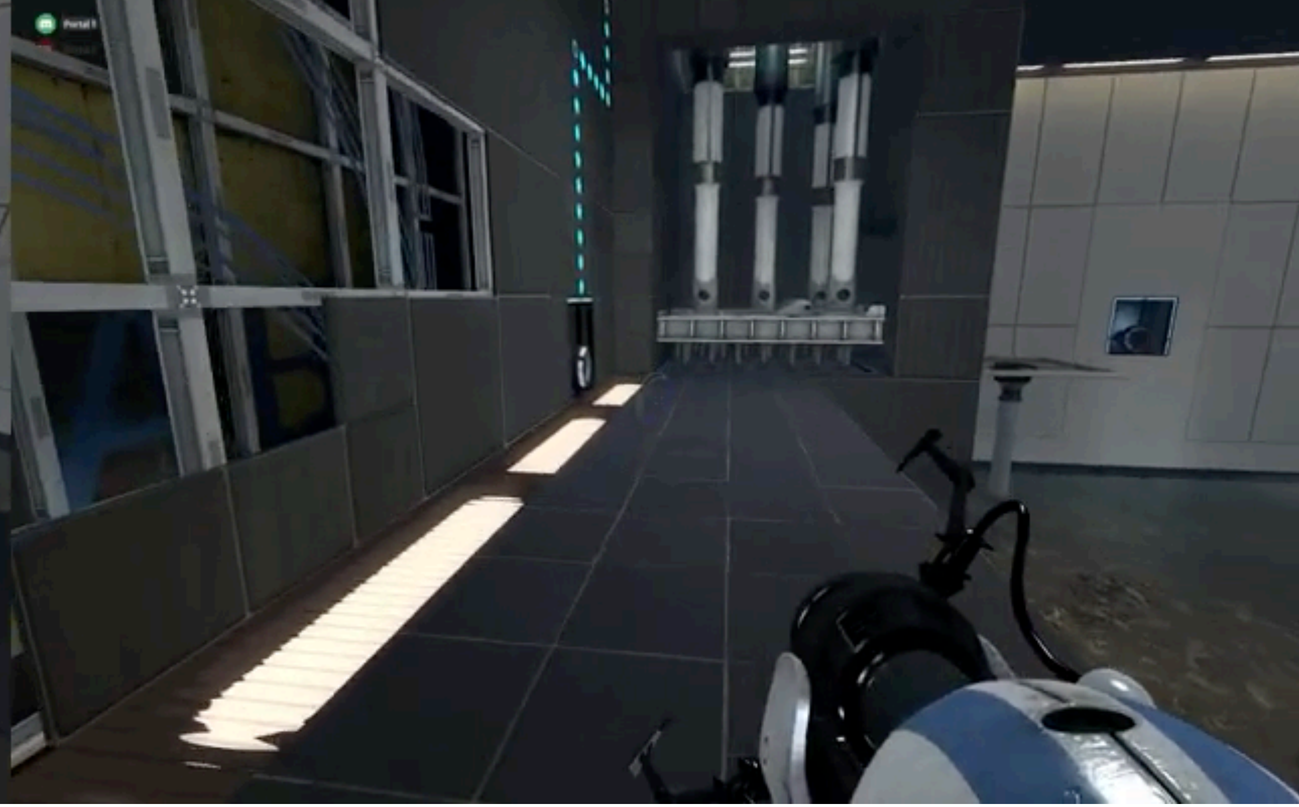
Alignment via conceptual pacts

Orange: *Let me **point** it … here.*

…

Blue: *Okay, we're gonna have to — one of us is gonna be —*
Blue: *Okay you be the **pointer**, cause clearly I have no natural talent when it comes to that.*





Language coordination with action

Blue: [[places portal]] *Portal.*
Blue: [[places portal]] *Portal.*

…

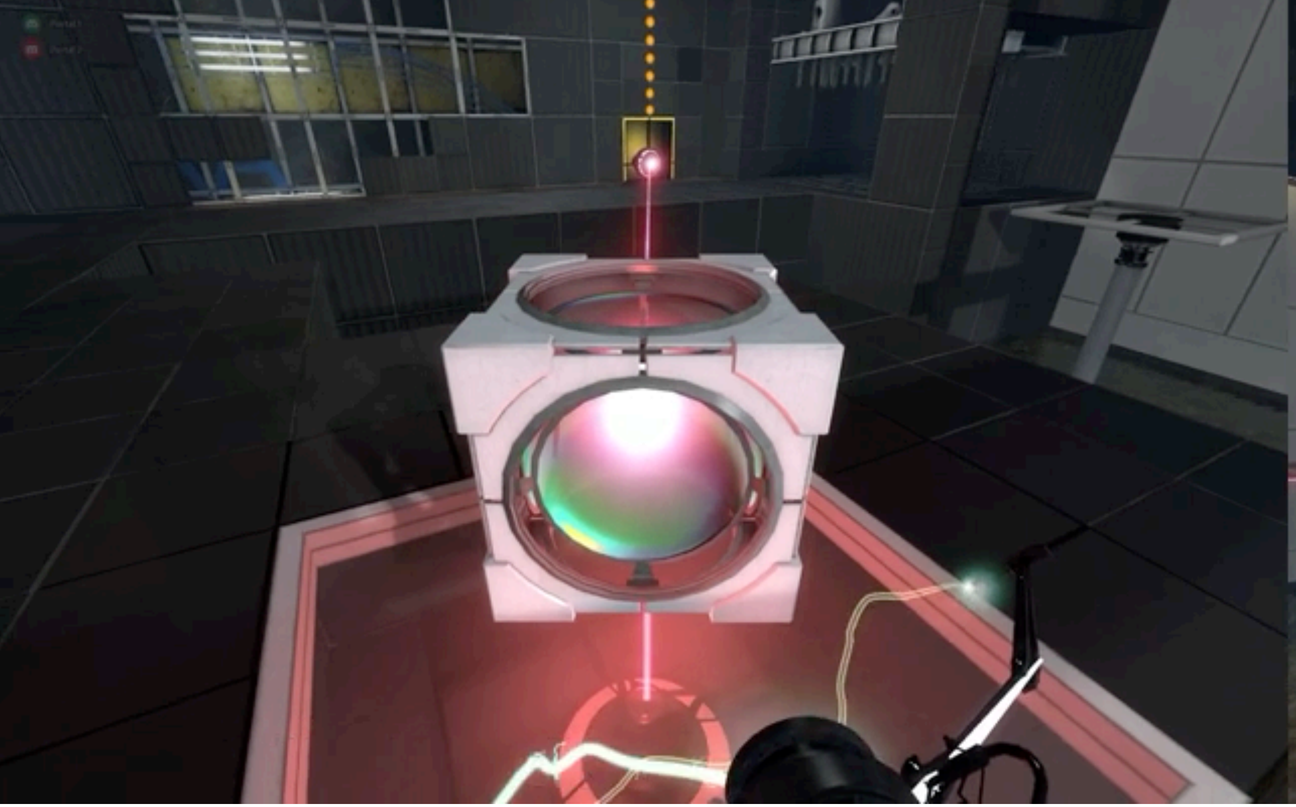
Blue: *Oh, like* [[jumps]] *now!*

Social conventions

Orange: *Oh, hi!*
Blue: *Hi, okay.*

Abstraction

Orange: *Okay, we'll try **this** one more time.*

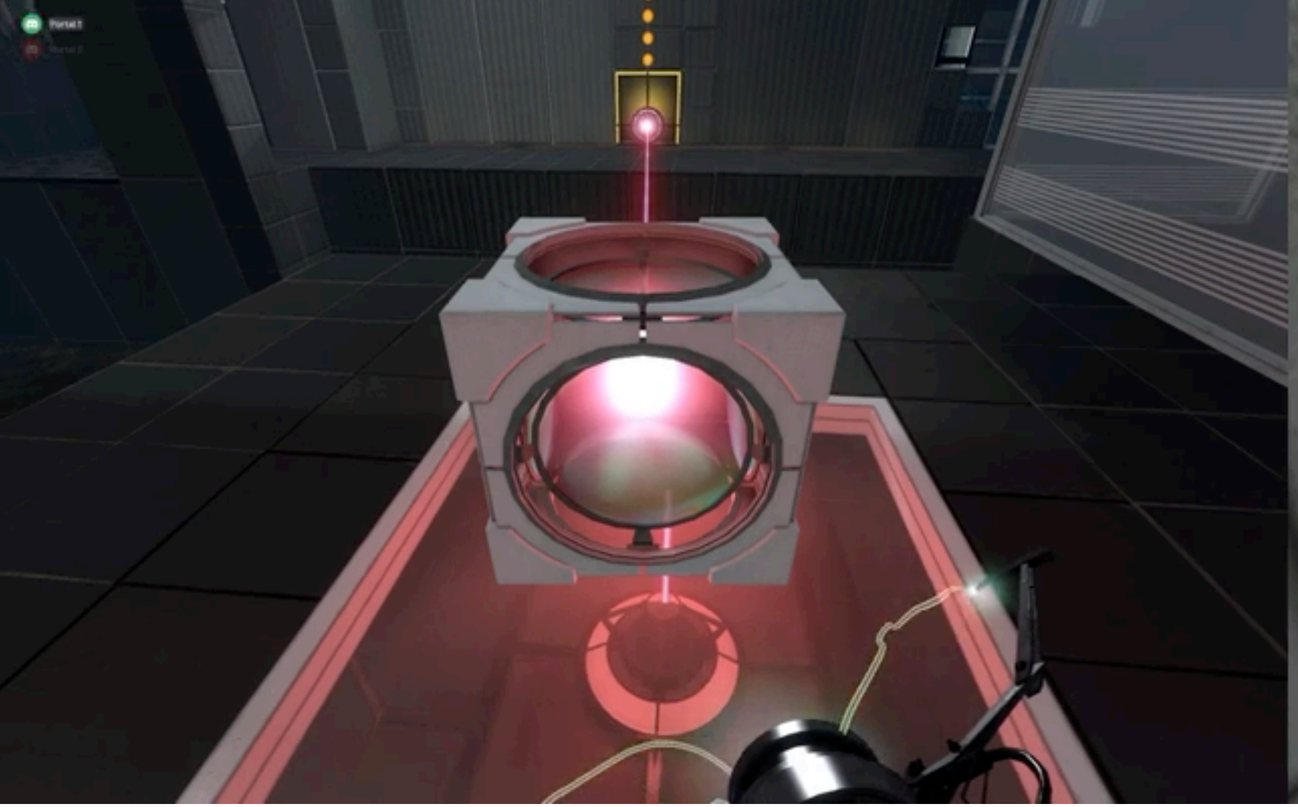


Long-context

Blue: *I need to put a higher —*
Blue: *Yeah, okay.*
Blue: *Maybe if I put a higher portal to —*
Blue: *Oy vey, okay.*
Orange: ((laughs))
Blue: *Um, like, if I put one of my portals up there…*

Requests for clarification

Blue: [[Looks at 3rd receptacle]] *Um, maybe don't -- maybe raise <u>that one</u> up first, all the way.*
Blue: [[Looks at 2nd receptacle]] *And then try to hit this one.*
Blue: *Does that make sense?*
Blue: *Like, go to <u>the other one</u> first.*
Orange: [[Moves laser to 3rd receptacle]] ***Uh, raise <u>this one</u> first?***
Blue: *Yeah, like, cause <u>that one</u> takes a long time to lift up.*
Orange: *Okay.*
Orange: *Yeah, that's true.*



Task management

Orange: *Think that works.*
Orange: *That works!*
Blue: *Yes, well done.*
Blue: *Okay.*
Orange: *Let’s go.*

Our new corpus

- **~12 hours of recorded gameplay, ~21,000 utterances**
- **18 dyads, recruited from UC Berkeley campus, annotated with their video game experience**
- **Each game comes with:**
  - Player observations (screen recordings)
  - Player audio
  - All player actions (to support replay, ray-casting, etc.)

Our new corpus

- **Timestamped transcripts**
- **Utterance-level features, partially automatically computed with GPT-4, based on DAMSL**
  - Communicative status (successful, unintelligible, abandoned)
  - Utterance type (proposition, query, imperative, etc.)
  - Information level (world state, affective evaluation, etc.)
  - Discursive act (request, assertion, commitment, etc.)
  - Hedging; self-talk
- **Timestamps of subtask completions**
- **Object and other-player visibility at each game tick**

Allen and Core 1997

Complex phenomena in interaction

- **Incremental phenomena in language production**
  - Abandonments
  - Self-repair and self-interruption
  - Coordination of language use with action

- **Dyadic incremental phenomena**
  - Gaps and overlaps
  - Backchannels
  - Turn-taking
  - Collaborative completions

Thank you especially to Elisabeth André, Jonathan Ginzburg, Alexander Koller, Matthew Purver, and Edwin Simpson

Complex phenomena in interaction

- **Continuous features of language use**
  - Gesture, facial expressions, prosody
  - Initiative-taking, proactivity, task monitoring
  - Pacing

- **Incentive structure and task demands**
  - Negotiation of goals, plans
  - Commitments
  - Grounding in hypotheticals
  - Social / non-task talk

Thank you especially to Elisabeth André, Jonathan Ginzburg, Alexander Koller, Matthew Purver, and Edwin Simpson

Complex phenomena in interaction

- **Asymmetry between participants**
  - Negotiation of roles
  - Clarification requests
  - Teaching / demonstration subdialogues
- **Task and dialogue structure**
  - Non-adjacent responses
  - High-level dialogue structure and its relationship with task structure

Blue: *Um ... every time I like get money from a study, I'm like, okay, this means I can like go eat out or something.*
Blue: *I can buy four boba.*

[[2 minutes pass, players solve puzzle]]

Orange: *Oh, boba sounds nice right now.*
Orange: *What time does [BOBA SHOP] close?*
Blue: *I don't know, but I would think that it's still open.*

[[3.5 minutes pass, players solve puzzle]]

Orange: *Man, I miss when [CITY] had 'em.*
Orange: *Snackpass group orders, man.*
Blue: *Do they not?*
Orange: *No store does that.*
Orange: *Dude, [BOBA SHOP] -- [BOBA SHOP] used to have like 30% off on -- on full party.*

Thank you especially to Elisabeth André, Jonathan Ginzburg, Alexander Koller, Matthew Purver, and Edwin Simpson

Complex phenomena in interaction

- **Embodiment**
  - Deixis
  - Perspective-taking
  - Attention management
  - Multimodal conversational grounding

- **Novelty**
  - Abstract reference
  - Expressions of uncertainty
  - Negotiation of meaning and formation of conceptual pacts

Thank you especially to Elisabeth André, Jonathan Ginzburg, Alexander Koller, Matthew Purver, and Edwin Simpson

Phenomenon 1: Clarification requests

- **Response to a recent utterance and ensuing subdialogue before the interaction can continue**
  - Not just questions!
  - Can also be confirmations, e.g., restating what the other person said

- **Why does it happen?**
  - Uncertainty or surprisal over meaning, intent, or form of recent utterance
  - More frequent:
    - In high-stakes settings where precision matters
    - Noisy channel
    - Differences in prior knowledge or perspectives

Phenomenon 1: Clarification requests

- **Older approaches**
  - Built on slot-filling dialogue systems for addressing uncertainty in ASR or NLU
  - Heuristic rule-based systems
  - Supervised or reinforcement learning

- **What about LLMs?**
  - Don’t take grounding acts such as clarifications, acknowledgments (Shaikh et al. 2024)
  - Perhaps because they’re trained single-turn on static data (Madureira and Schlangen 2024)

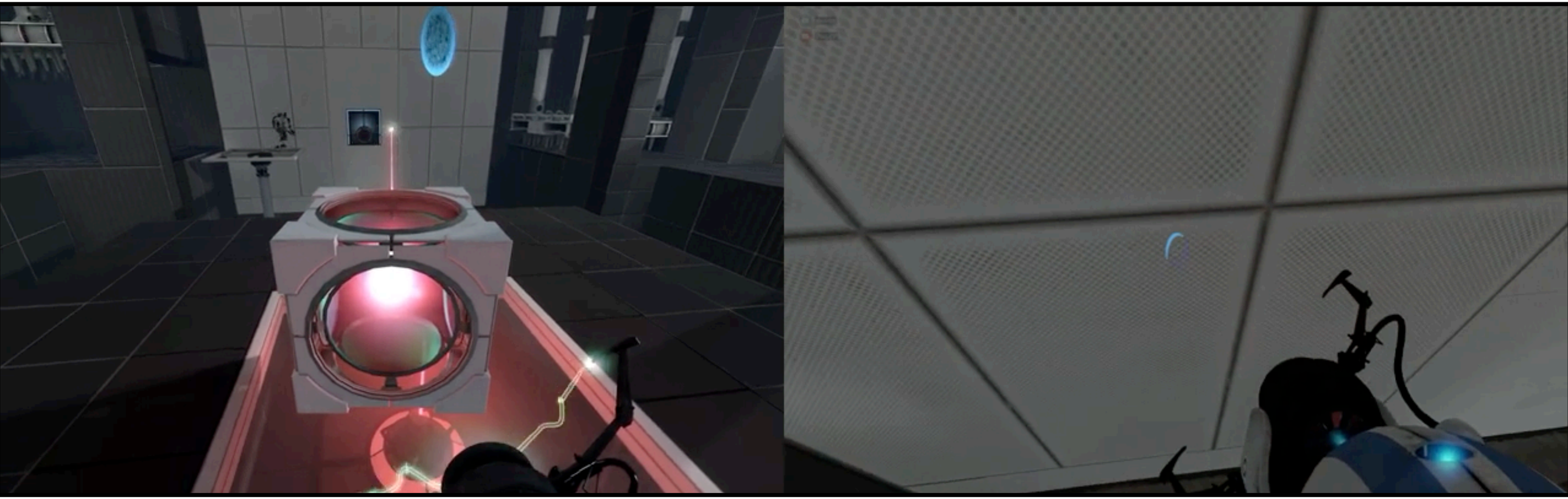
San-Segundo et al. 2001, Rieser and Moore 2005

Phenomenon 1: Clarification requests
Recent approaches

- **Mostly focused on generating clarification questions in response to ambiguous user-generated questions**
  - Ambiguity in preferences, user background (e.g., lexical differences)
- **Focused on text-based interactions**
- **Acknowledge tradeoff between interactional efficiency and accuracy**
- **Main recent approach: first quantify uncertainty or ambiguity, then decide whether and how to ask a clarification question**
  - Measuring uncertainty: prompting- and sampling-based approaches
  - Simulation of conversation and rewarding questions that result in disambiguating responses

Andukuri et al. 2024, Deng et al. 2023, Hou et al. 2024, Kuhn et al. 2024, Testoni and Fernández 2024, Zhang and Choi 2023, Zhang et al. 2024

Phenomenon 1: Clarification requests
What can we learn?



Phenomenon 1: Clarification requests
What can we learn?

- **Ephemeral knowledge**
  - Whatever it was we were uncertain about
- **Global knowledge**
  - Individualized preferences
  - Better model of interlocutor's language use
    - E.g., deictic conventions

Blue: [[Looks at 3rd receptacle]] *Um, maybe don't -- maybe raise that one up first, all the way.*
Blue: [[Looks at 2nd receptacle]] *And then try to hit this one.*
Blue: *Does that make sense?*
Blue: *Like, go to the other one first.*
Orange: [[Moves laser to 3rd receptacle]] ***Uh, raise this one first?***
Blue: *Yeah, like, cause that one takes a long time to lift up.*
Orange: *Okay.*
Orange: *Yeah, that's true.*

Schober 1993

Phenomenon 2: Meaning negotiation and formation of conceptual pacts

- **We come to agreement about what words mean in interaction by forming conventions that are:**
  - Arbitrary
  - Stable
  - Efficient

- **Why does it happen?**
  - On-the-fly tool joint abstraction: avoid redundancy, efficiently refer to novelty
  - More frequent:
    - Novelty in perception, action, or conceptual space
    - Ambiguity where distinctions become necessary to attend to
  - When interaction is longer or iterated

Krauss and Weinheimer 1964, Clark and Wilkes-Gibbs 1986, Hawkins et al. 2020

Phenomenon 2: Meaning negotiation and formation of conceptual pacts

- **Older approaches**
  - As far as I am aware, the process of in-interaction language adaptation has not been modeled computationally before modern NLP
  - But prior work has characterized its dynamics (e.g., Hawkins et al. 2020, Eliav et al. 2024), including in many-agent scenarios (Hawkins et al. 2022, Boyce et al. 2024)

- **What about LLMs? Hua and Artzi 2024**
  - They can *follow* a human's convention formation process
  - But do not actively push it forward by proposing and sticking to new conventions, except with heavy prompting

Phenomenon 2: Meaning negotiation and formation of conceptual pacts

- **Recent approaches**
  - Mostly based on rational speech acts (RSA) model
  - Mostly focus on adaptation from one agent to another, rather than forming ad-hoc conventions for things novel to both agents

- **Inference vs. learning approaches**
  - Leverage memory of successes and failures in iterative reference games (Greco et al. 2023)
  - Continually update a model of a listener (Hawkins et al. 2020, Zhu et al. 2021, Takmaz et al. 2023)

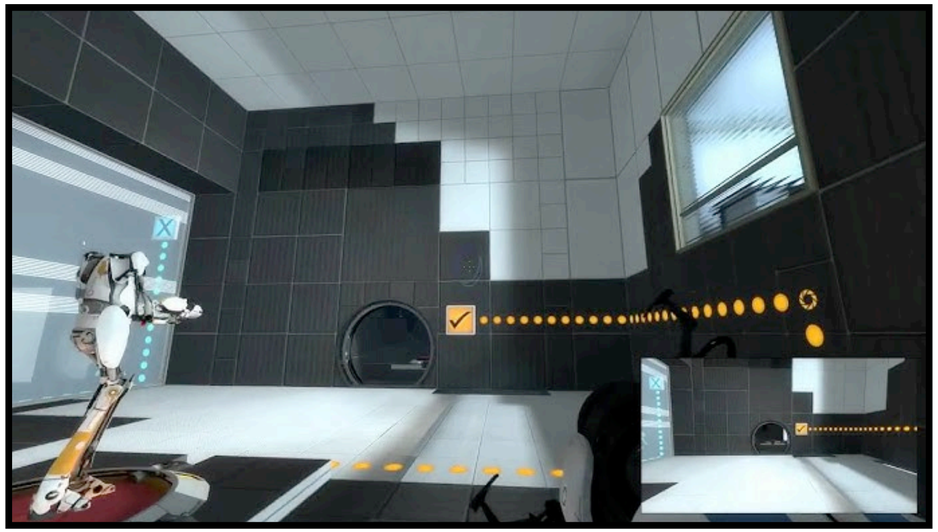
Phenomenon 2: Meaning negotiation and formation of conceptual pacts

Orange: *And then, I think if you shoot on that white thing over there —*
Blue: *Okay.*
Orange: *Uh, the others, uh, the other --*
Blue: *((laughs))*
Blue: *Here.*
Orange: *Oh.*
Orange: *Uh, oh, uh, the white panel, kind of, uh, to the left of the door.*

Orange: *Well, okay, there's only two panels.*

Orange: *Is there a white panel, maybe?*

Orange: *Like, I -- I wanna get rid of my panel.*
Orange: *Uh, uh, how do I, uh -- oh no, how do I get rid of the --*
Blue: *Oh, the portal.*



Phenomenon 2: Meaning negotiation and formation of conceptual pacts

Orange: *Uh, oh you wanna shoot like that panel on the very right.*
Blue: *Do I just like hit the moon, or --*
Orange: *Um, yeah yeah yeah that one.*
Orange: *You wanna shoot that panel up there on the very right and then join it.*

Orange: *Uh, uh, () there's a little white panel that's like by the button.*

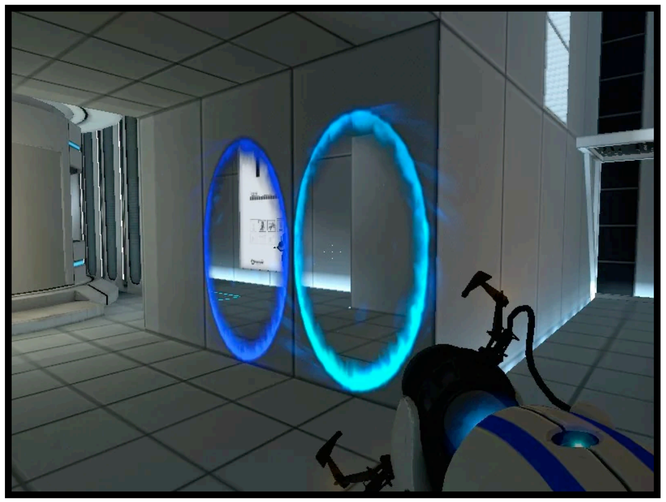
Blue: *Wait, are those -- on the floor, are those more panels?*

Orange: *Um, can you shoot one portal like pretty much as low as possible on the same white panel that my thing is?*
Orange: *Like um -- like uh the same white panel on my light blue but like as low as —*

Orange: *Is it not shooting on the -- like on the -- like on the white part?*
Orange: *Like even the -- the dirty white panel.*

Phenomenon 2: Meaning negotiation and formation of conceptual pacts

Orange: *Yeah, shoot your white panel, um, elsewhere, just like literally anywhere else.*
Orange: *Uh, it should be okay.*
Orange: *Uh, uh, sorry sorry, uh () your blue -- uh your dark blue.*
Orange: *Sorry.*
Orange: *Shoot your dark blue.*
Orange: *Uh, we wanna keep your light blue.*
Orange: *Shoot your dark blue --*
Blue: *Oh.*
Blue: *Wait this is my dark blue, the one that I just shot.*



Blue: *(Do) I shoot a portal (over) -- on those white panels?*

Orange: *Uh, why don't you try shooting the blue, the light blue, kind of like on this panel that's like right next to you?*

Blue: *Is -- are there panels here?*

Blue: *Panels, where?*

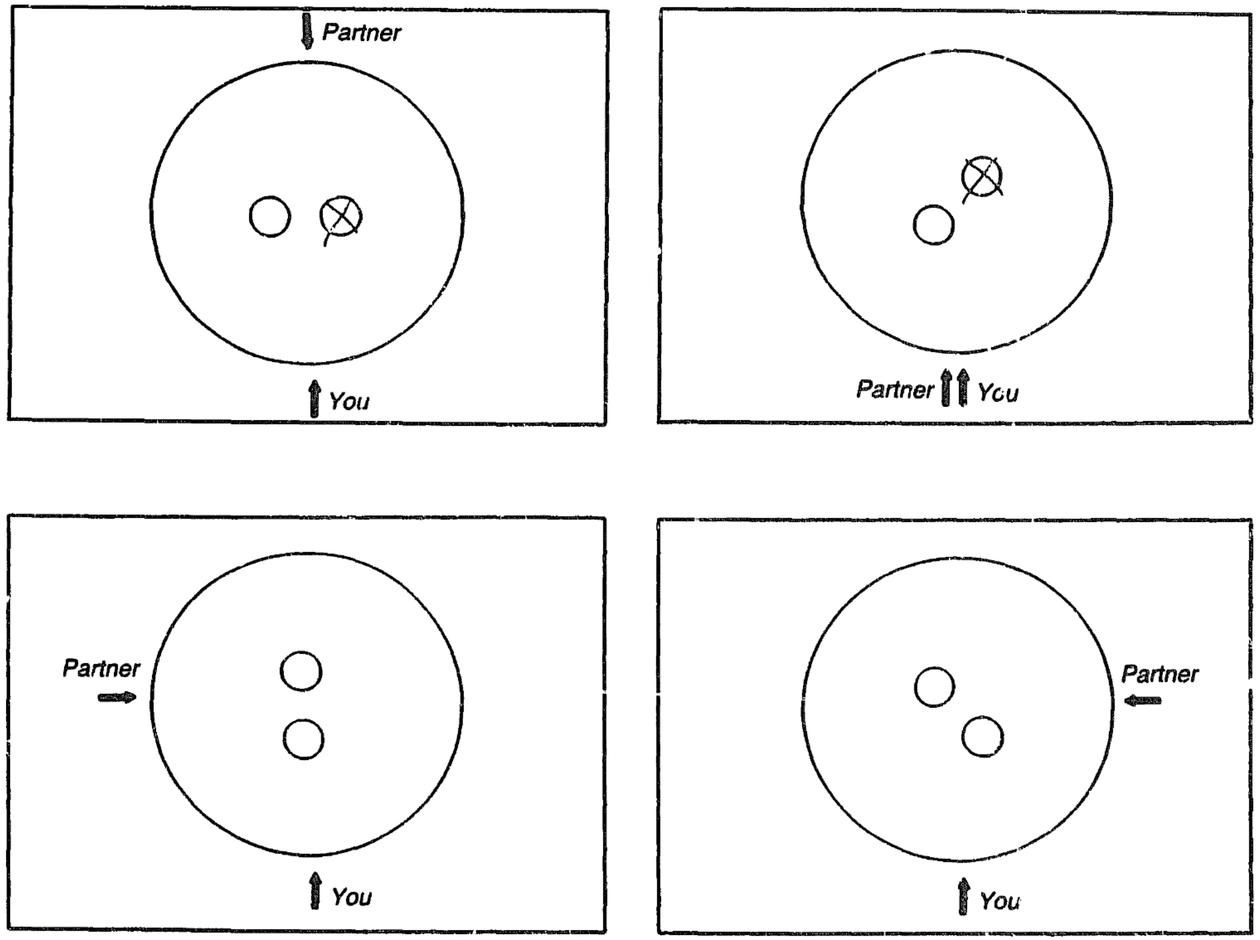
Phenomenon 2: Meaning negotiation and formation of conceptual pacts

- **What can we learn?**
  - New abstractions, useful for future communication in and beyond interaction
  - What is worth giving an abstraction
    - For improved task knowledge
    - Interlocutor’s preferences and attentive biases

Phenomenon 3: Perspective-taking

- **Language grounding requires taking into account one's unique perspective**

- **Why does it happen?**
  - We can achieve more efficient joint understanding by accommodating our interlocutor's point of view
  - Especially in embodied scenes, where we necessarily have different perspectives from our interlocutor



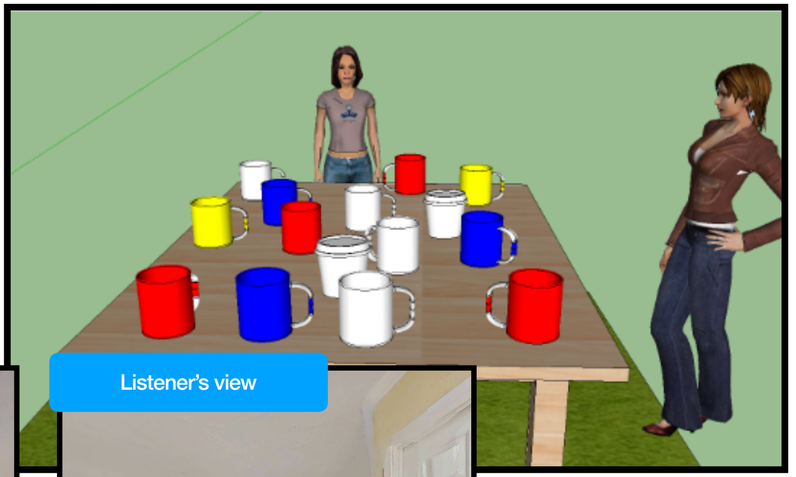
Schober 1993

Phenomenon 3: Perspective-taking

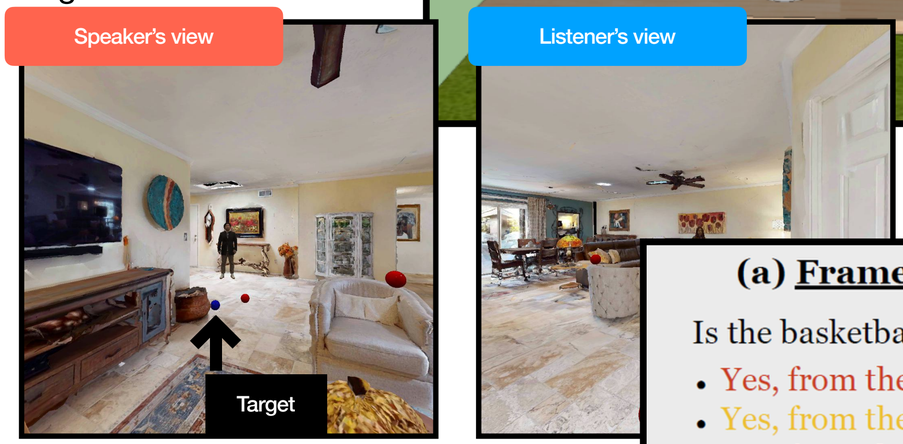
- **Older approaches**
  - Mostly in simply characterizing factors influencing the human use of spatial perspective, e.g., convention formation, relative positions of agents and referents

- **What about LLMs?**
  - Generally, not great
  - Mostly studied in simple static environments
  - LLMs tend to produce egocentric references rather than accommodating a listener

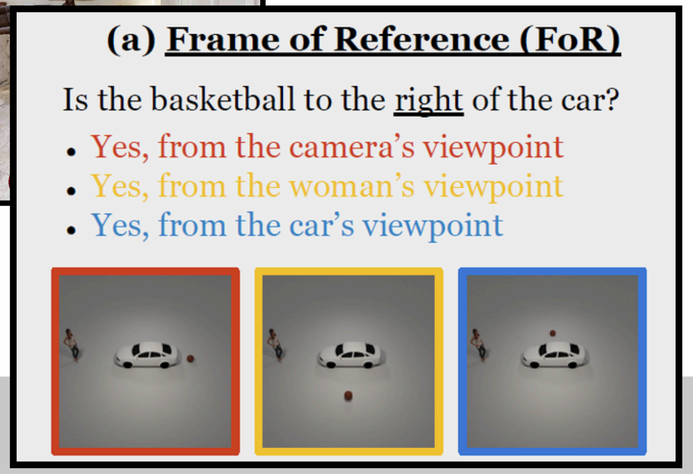
Loáiciga et al. 2021



Tang et al. 2024



Zhang et al. 2024, COMFORT



Phenomenon 3: Perspective-taking

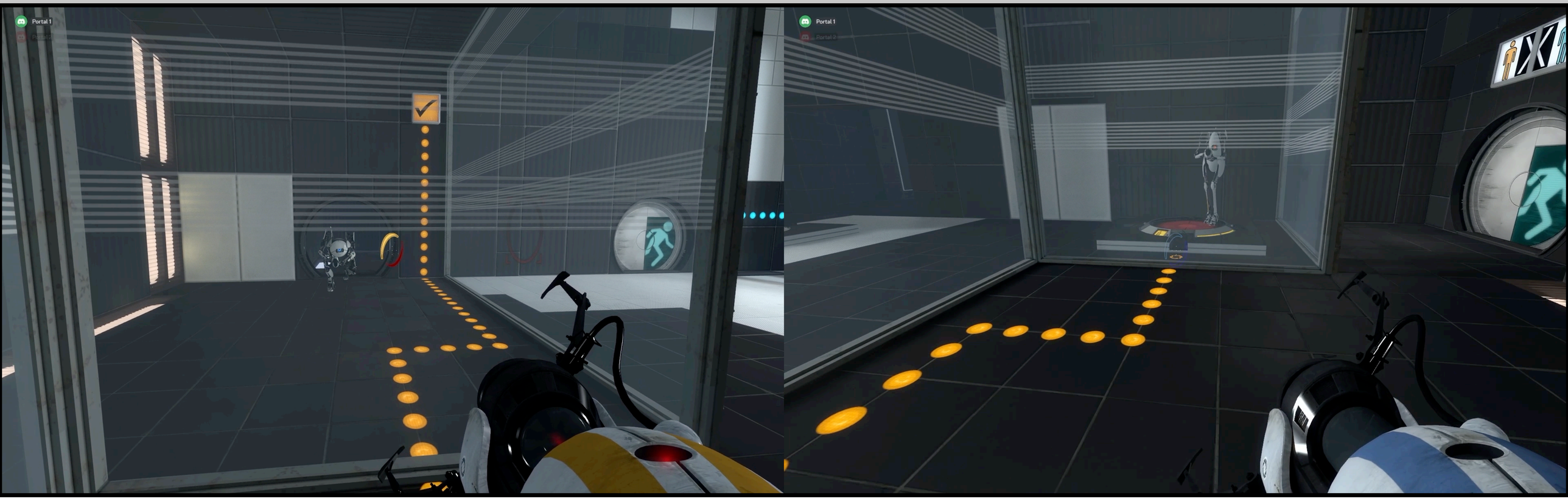
- **What to learn?**
  - Improving spatial reasoning in general
  - Improving perspective-taking

- **Approaches**
  - Data augmentation (e.g., SpatialVLM, Chen et al. 2024)
  - Improved representations, e.g., 3D representations that include intrinsic directionality instead of simple 2D bounding boxes (Zhao et al. 2023)
  - Learning from evidence of comprehension (Tang et al. 2024)

Phenomenon 3: Perspective-taking

Orange: *Um, (you) — (you) — you can walk through the door on the right I think.*

Phenomenon 3: Perspective-taking



Orange: *And then you can do that on, uh -- on those white walls.*
Orange: *So, to your right.*
Blue: *Oh, these walls.*
Blue: *Okay.*
Orange: *Yeah.*

Phenomenon 3: Perspective-taking

Orange: *And then there's one more in the -- in the room you were just in.*
Blue: *Oh!*
Orange: *Um.*
Blue: *This one?*

Orange: *Yeah.*
Orange: *To your right.*
Blue: *Oh, this one?*
Blue: *Mhm.*

Phenomenon 3: Perspective-taking

- **What can we learn?**
  - Convention formed regarding mutual accommodation
  - Willingness of interlocutor to accommodate
  - Interlocutor’s perspective and perception of shared scene
  - Improved spatial reasoning and perspective-taking

Proposed analyses

- **How does language use correspond to in-game actions?**
  - E.g., language use for “leaders” and “followers” for different subtasks
  - (How to operationalize this question?)
- **How does language change over time?**
  - Convention formation
  - Distributions in dialogue acts, utterance lengths
- **How do players use (multimodal) conversational grounding?**
- **Why do people talk to themselves?**
- **What is the range of linguistic phenomena that arise in this scenario, that NLP has mostly neglected?**
  - E.g., spatial relations with default perspectives

What would you want to do with this data?

Joint reasoning about perception and language

- **Significant progress in scaling up vision-and-language models**

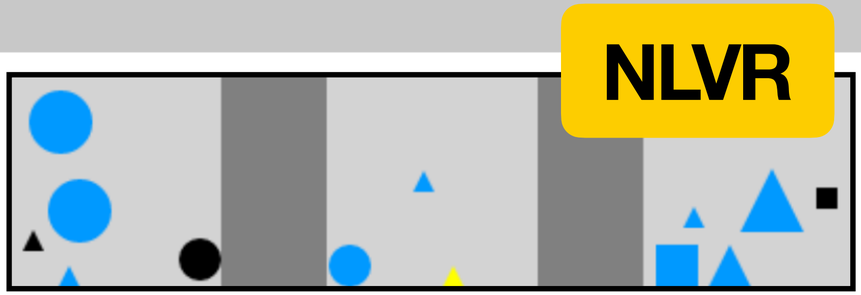
Joint reasoning about perception and language

- **Significant progress in scaling up vision-and-language models**

- **However, they still struggle with**
  - Fine-grained visual understanding

59.9

s a box with items
y 2 different colors
and a small black triangle
touching the wall



Suhr et al. 2017, Wu et al. 2024

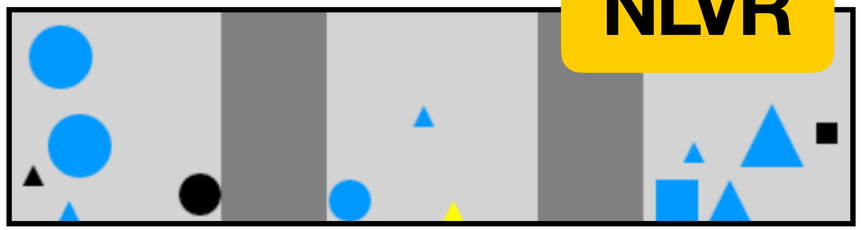
Joint reasoning about perception and language

- **Significant progress in scaling up vision-and-language models**
- **However, they still struggle with**
  - Fine-grained visual understanding
  - Spatial reasoning

59.9

s a box with items
y 2 different colors
and a small black triangle
touching the wall

NLVR



VSR

64.2

The cow is ahead of the person.

Suhr et al. 2017, Wu et al. 2024, Liu et al. 2023, Panos et al. 2024

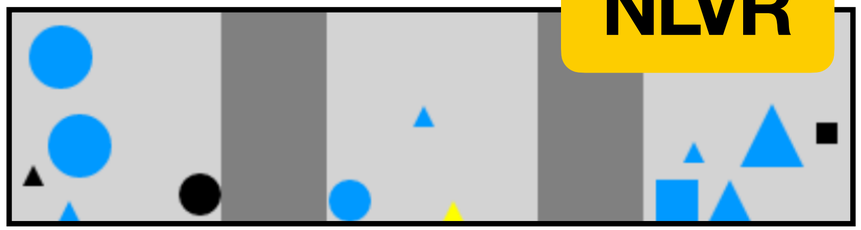
Joint reasoning about perception and language

- **Significant progress in scaling up vision-and-language models**

- **However, they still struggle with**
  - Fine-grained visual understanding
  - Spatial reasoning
  - Compositionality

59.9

s a box with items
y 2 different colors
and a small black triangle
touching the wall



VSR

64.2

The cow is ahead of the person.

Winoground

“barely above chance”

Suhr et al. 2017, Wu et al. 2024, Liu et al. 2023, Thrush et al. 2022, Al-Tahan et al. 2024

Limitations of photographs

- **Most images on the internet are *depictions***

- **Language-and-vision tasks typically framed from the perspective of the photographer**

Torralba and Efros 2011

Limitations of photographs

- **Most images on the internet are *depictions***
- **Language-and-vision tasks typically framed from the perspective of the photographer**
- **Little representation of nuance in spatial relations**

SpatialSense

Yang et al. 2019

Limitations of photographs

- **Most images on the internet are *depictions***
- **Language-and-vision tasks typically framed from the perspective of the photographer**
- **Little representation of nuance in spatial relations**

SpatialSense

VSR

Figure 1: Caption: *The potted plant is at the right side of the bench.* Label: `True`.

Figure 2: Caption: *The cow is ahead of the person.* Label: `False`.

Yang et al. 2019, Liu et al. 2023

Limitations of photographs

- **Most images on the internet are *depictions***
- **Language-and-vision tasks typically framed from the perspective of the photographer**
- **Little representation of nuance in spatial relations**

Figure 1: Caption: *The potted plant is at the right side of the bench*. Label: `True`.

Liu et al. 2023

Limitations of photographs

- **Most images on the internet are *depictions***
- **Language-and-vision tasks typically framed from the perspective of the photographer**
- **Little representation of nuance in spatial relations**

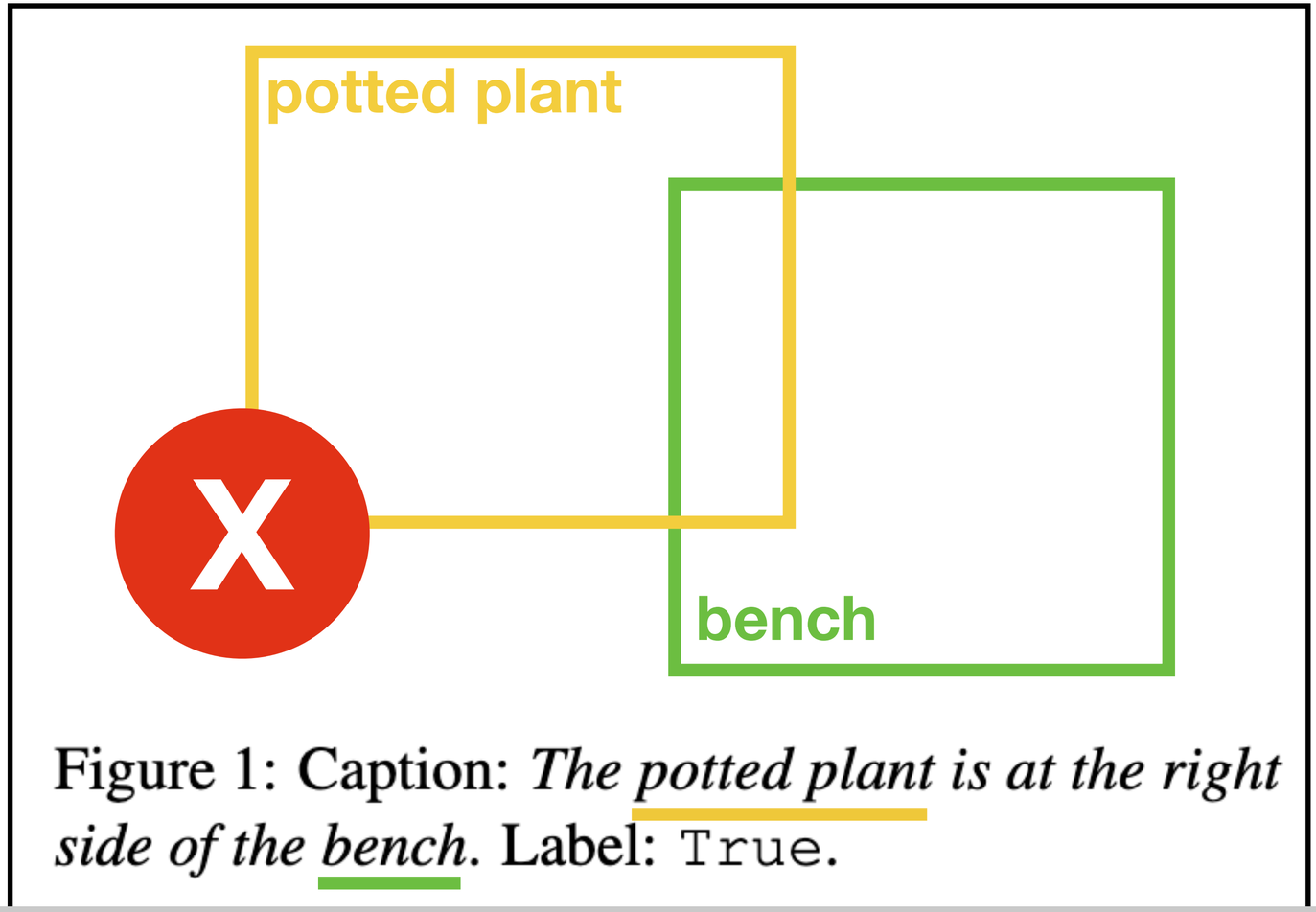


Figure 1: Caption: *The potted plant is at the right side of the bench*. Label: `True`.

Liu et al. 2023

Limitations of photographs

- **Most images on the internet are *depictions***
- **Language-and-vision tasks typically framed from the perspective of the photographer**
- **Little representation of nuance in spatial relations**

Figure 1: Caption: *The potted plant is at the right side of the bench*. Label: `True`.

Liu et al. 2023

Limitations of photographs

- **Most images on the internet are *depictions***
- **Language-and-vision tasks typically framed from the perspective of the photographer**
- **Little representation of nuance in spatial relations**

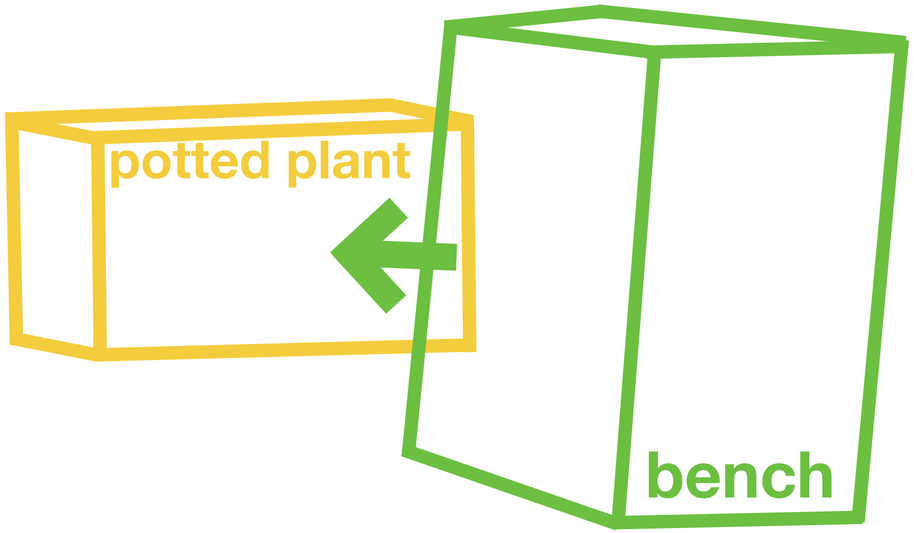


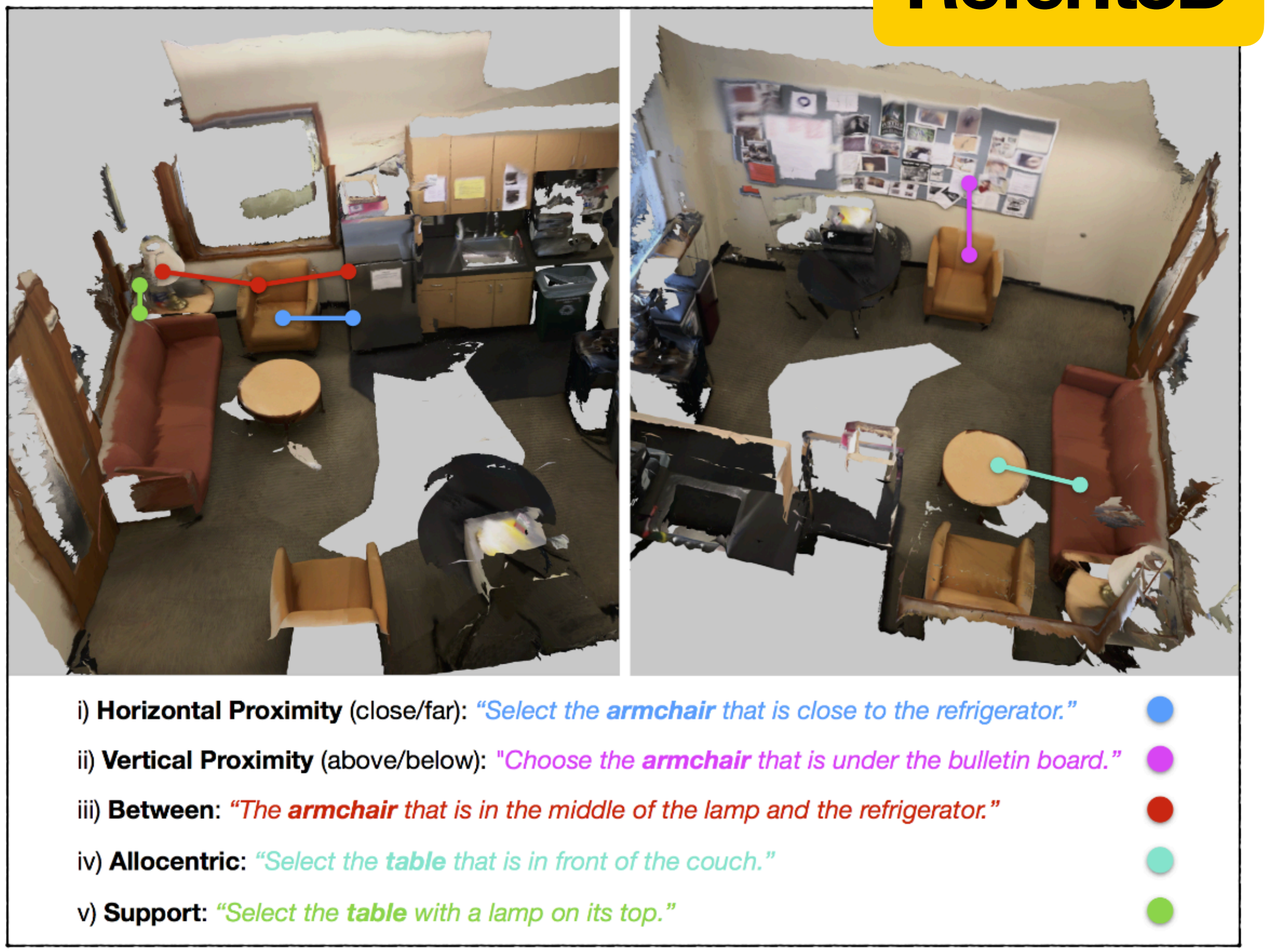
Figure 1: Caption: *The potted plant is at the right side of the bench*. Label: `True`.

Liu et al. 2023

Spatial reasoning in 3D scenes

- Input representation is an entire scene
- Supports complex and compositional spatial language
- 3rd person view of scene — not embodied

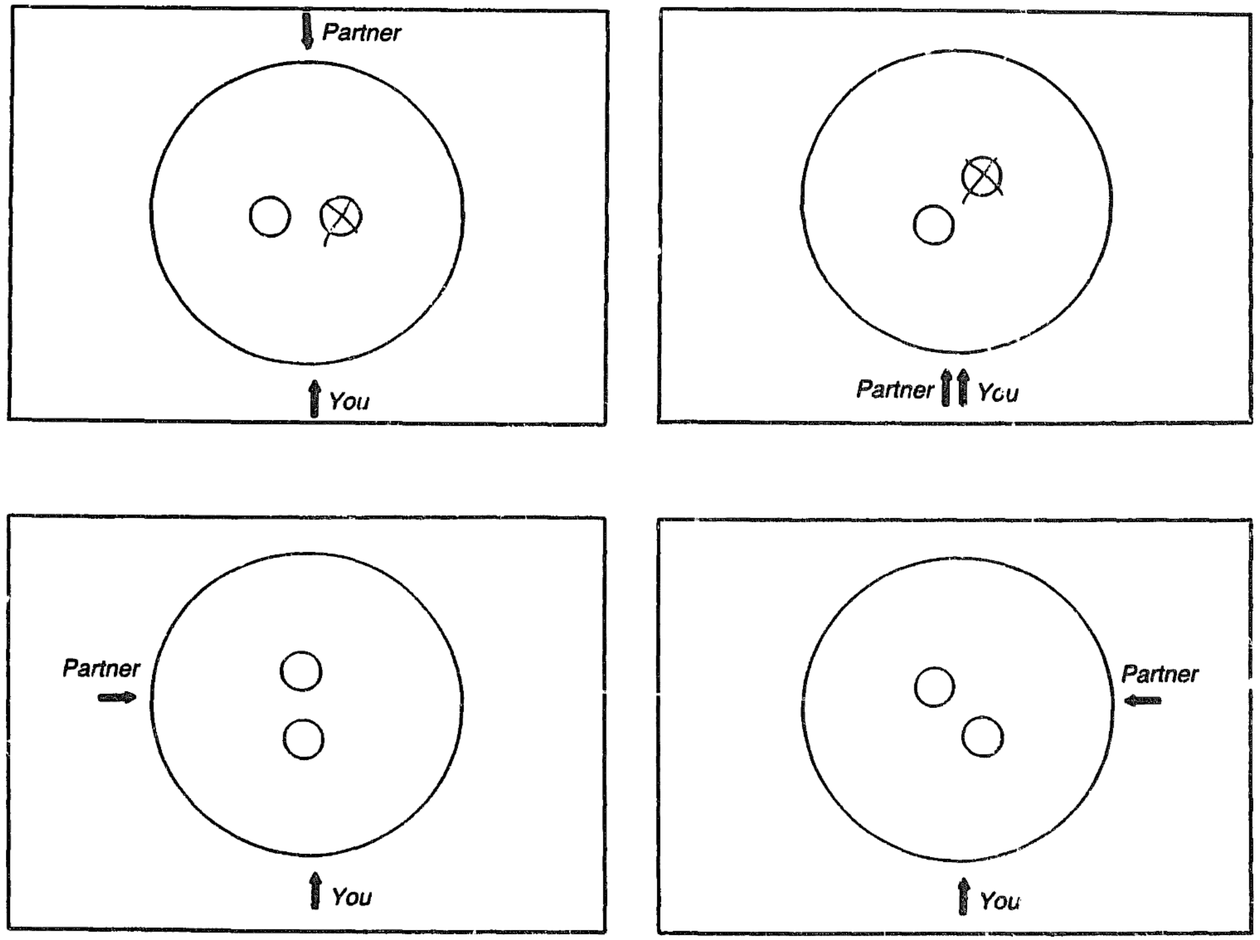
ReferIt3D



Achlioptas et al. 2020

Being jointly embodied

- **In multi-agent situated interactions, agents take on different perspectives**
- **Accommodating one's interlocutor can involve taking on their perspective**
- **Design of shared environment influences language use**
  - Having an actual partner (as opposed to "imaginary")
  - Relative placement of candidate referents
  - Multi-turn vs. single-shot interaction



Schober 1993

Jointly embodied language use

We study existing systems that use language in jointly embodied scenes

- Do they generate language that's accurate wrt. spatial relations?
- Do they accommodate their interlocutors by taking on their perspective?
- Do they understand language generated by accommodating humans in embodied interactions?

Tang et al. 2024, EMNLP

Task: Reference game

- **Speaker agent** maps from environment observation to natural language reference

$$p_s : \mathcal{O} \times \mathcal{R}^N \times \{1 \dots N\} \rightarrow \Delta^{\mathcal{X}}$$

Task: Reference game

- **Speaker agent** maps from environment observation to natural language reference

$$p_s : \mathcal{O} \times \mathcal{R}^N \times \{1 \dots N\} \rightarrow \Delta^{\mathcal{X}}$$

- **Listener agent** maps from reference and observation to referent selection

$$p_l : \mathcal{O} \times \mathcal{R}^N \times \mathcal{X} \rightarrow \Delta^{\{1 \dots N\}}$$

Task: Reference game

- **Speaker agent** maps from environment observation to natural language reference

$$p_s : \mathcal{O} \times \mathcal{R}^N \times \{1 \dots N\} \rightarrow \Delta^{\mathcal{X}}$$

- **Listener agent** maps from reference and observation to referent selection

$$p_l : \mathcal{O} \times \mathcal{R}^N \times \mathcal{X} \rightarrow \Delta^{\{1 \dots N\}}$$

Communicative Success

$$x = \arg\max_{x' \in \mathcal{X}} p_s(x' \mid o_s, \mathcal{R}, t)$$

$$\hat{t} = \arg\max_{1 \leq i \leq N} p_l(i \mid o_l, \mathcal{R}, x)$$

$$\text{Success}(p_s, p_l, o_s, o_l, \mathcal{R}, t) = \mathbb{1}_{t=\hat{t}}$$

Environment generation

1. 3D base environments from ScanNet++

Environment generation

1. **3D base environments from ScanNet++**

2. **Agent placement**
   - Each agent associated with a pose: location in 3D space and yaw of view
   - Sample heights from reasonable human heights
   - Maximum distance between listener and speaker cameras
   - Non-empty overlap of fields of view; speaker always sees the listener

Environment generation

3 Candidate referent placement

- Randomly sample location
- Gravitational physics simulation to drop objects
- All referents should be visible to both agents

Environment generation

3 Candidate referent placement

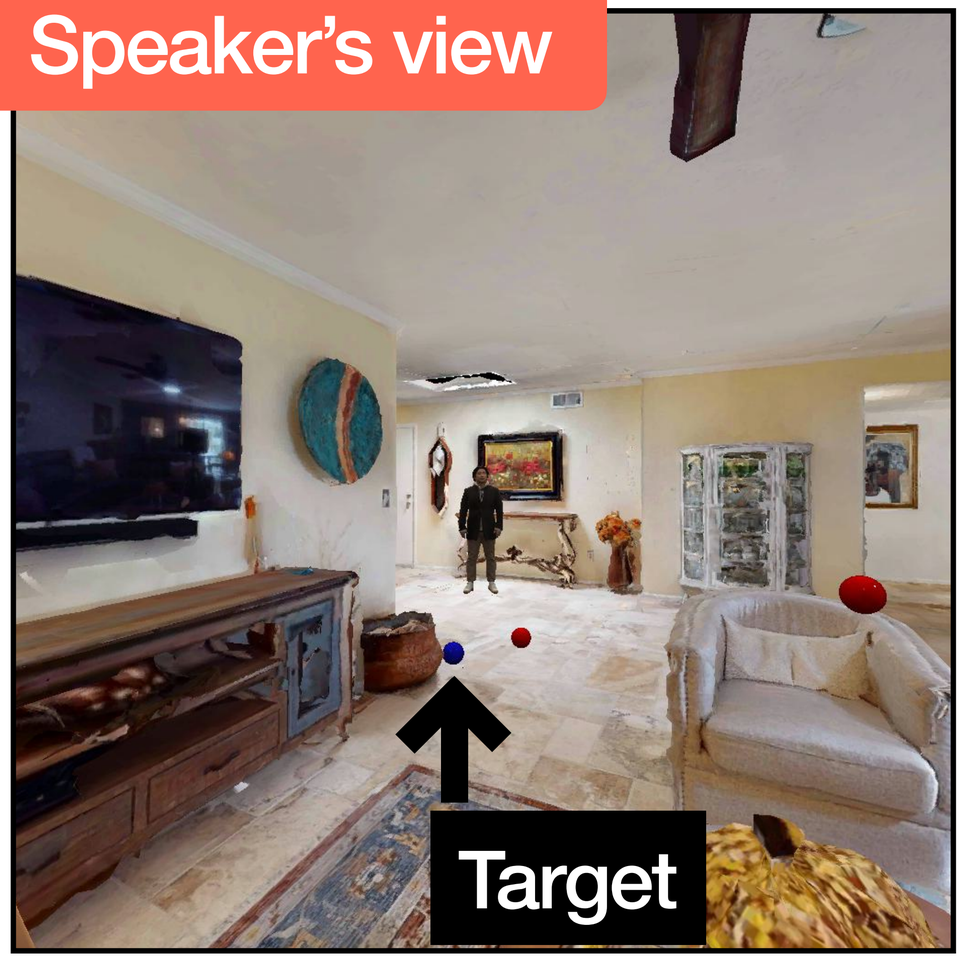
- Randomly sample location
- Gravitational physics simulation to drop objects
- All referents should be visible to both agents

4 Render!

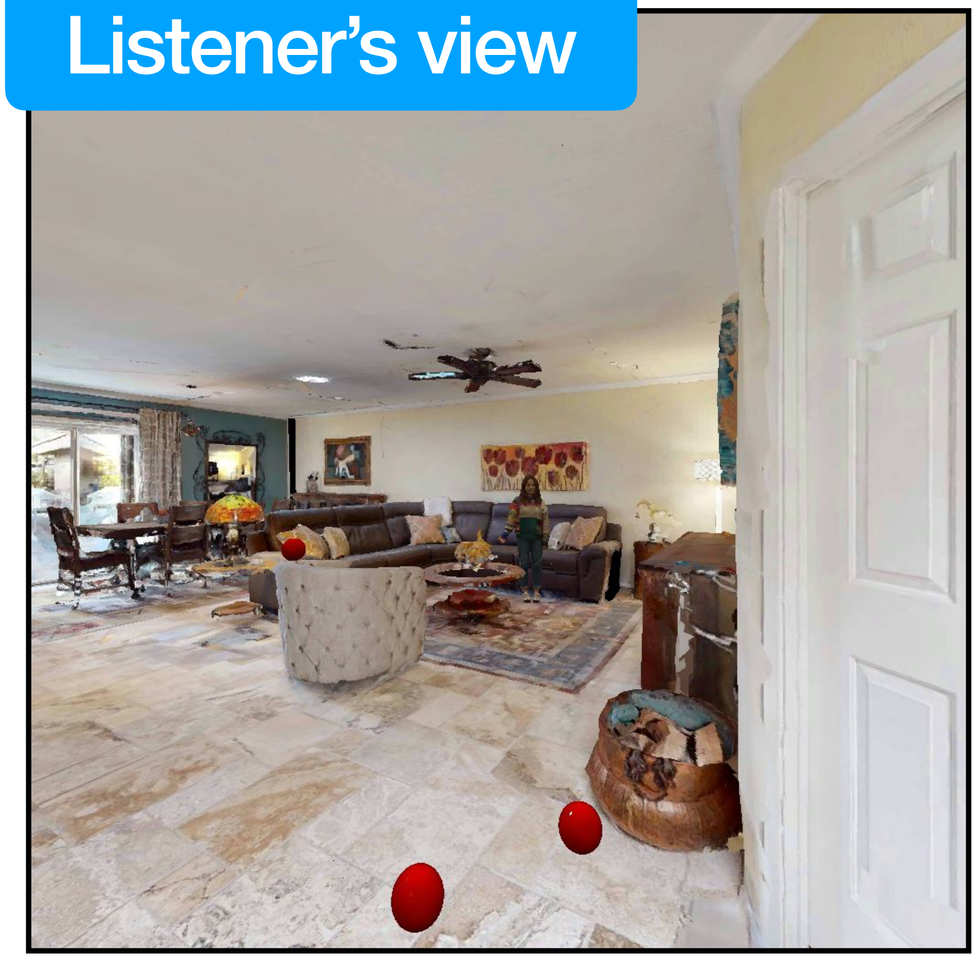
- 2D projection from camera pose
- All referents are red; target is blue for speaker
- Agent cameras are rendered as 3D human models

Environment generation

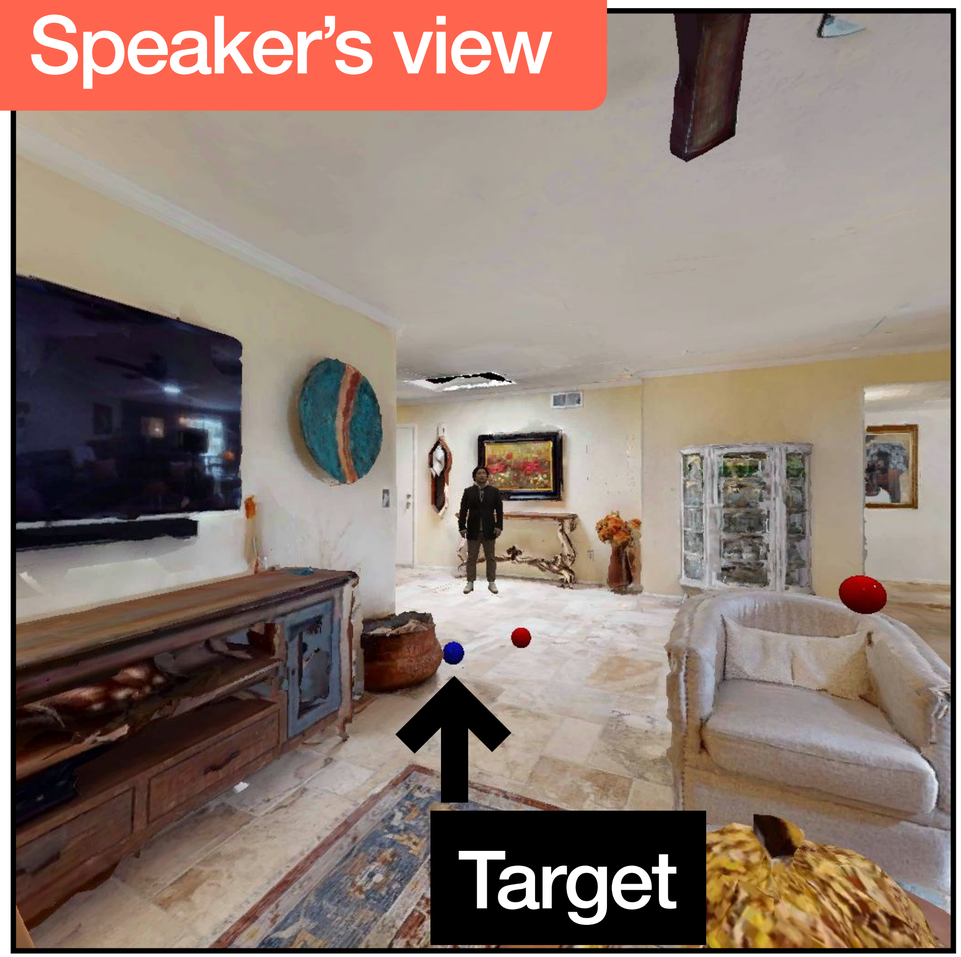
Speaker's view



Listener's view



Environment generation

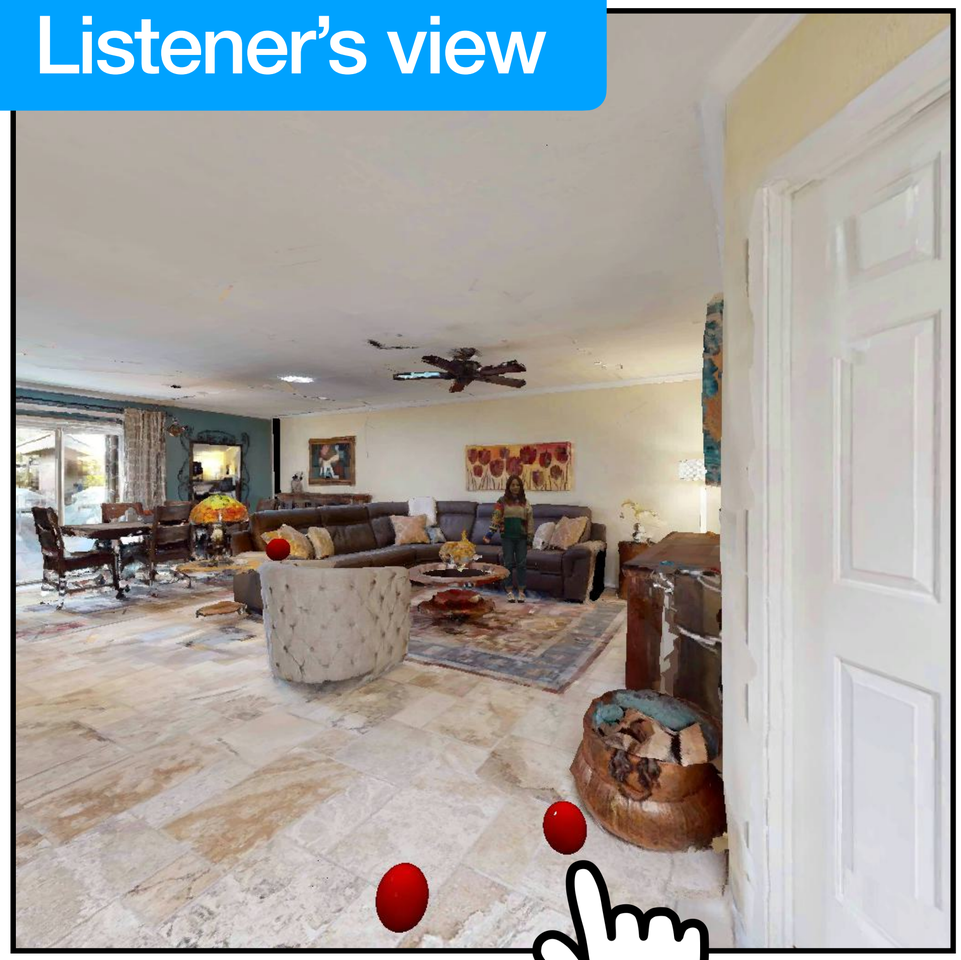


$$x = \arg\max_{x' \in \mathcal{X}} p_s(x' \mid o_s, \mathcal{R}, t)$$

On your right near the basket

Environment generation

Listener's view



On your right near the basket

$$\hat{t} = \arg \max_{1 \leq i \leq N} p_l(i \mid o_l, \mathcal{R}, x)$$

Environment generation

$$\text{Success}(p_s, p_l, o_s, o_l, \mathcal{R}, t) = \mathbb{1}_{t=\hat{t}}$$

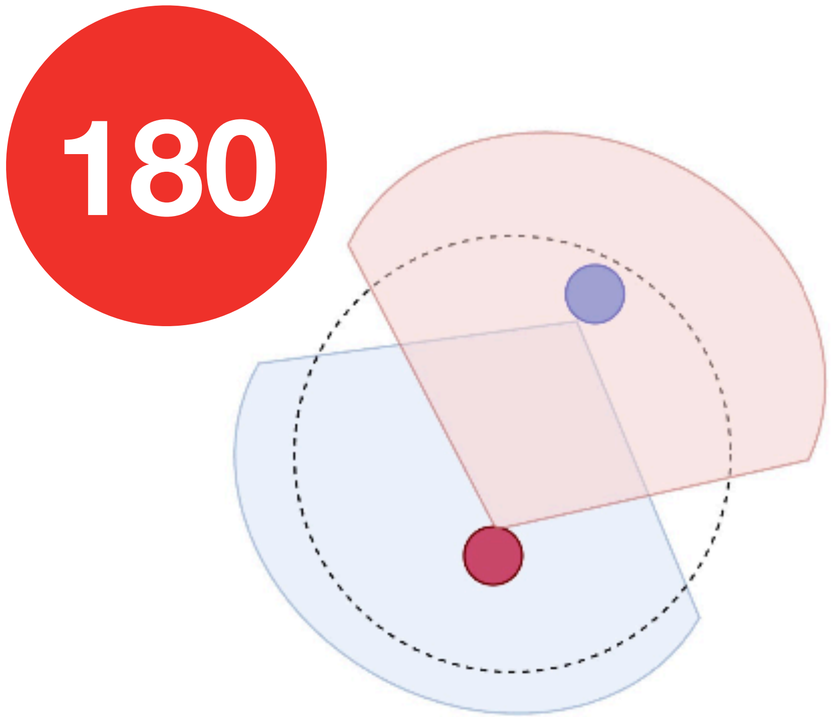
Success!

Controlled difficulty
Speaker-listener orientations

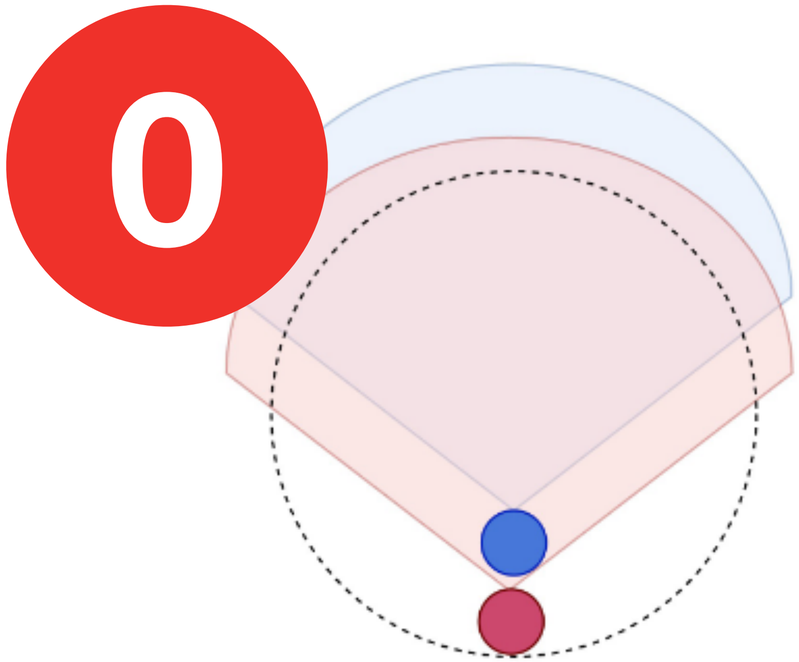
- **Proxy for perspective similarity**
  - 180° — facing one another
  - 0° — facing same direction

- **Sample uniformly from relative orientations**

Controlled difficulty
Speaker-listener orientations



Controlled difficulty
Speaker-listener orientations



Controlled difficulty
Adversarial referent placement

- **Referent placement policy model**

$$R : \mathcal{C}^* \times \mathcal{O}_s \times \mathrm{P}_s \times \mathrm{P}_l \rightarrow \Delta^{\mathcal{R}^N \times \{1 \ldots N\}}$$

- **Vision transformer**

- **Trained to maximize communicative failure rate between two agents**

$$\max_R \mathbb{E}_{(\mathcal{R}', t') \sim R(\cdot)} \left[1 - \mathrm{Success}(\hat{p}_s, \hat{p}_l, o_s, o_l, \mathcal{R}', t') \right]$$

Controlled difficulty
Adversarial referent placement

Controlled difficulty
Adversarial referent placement

Data
Pre-generated scenes

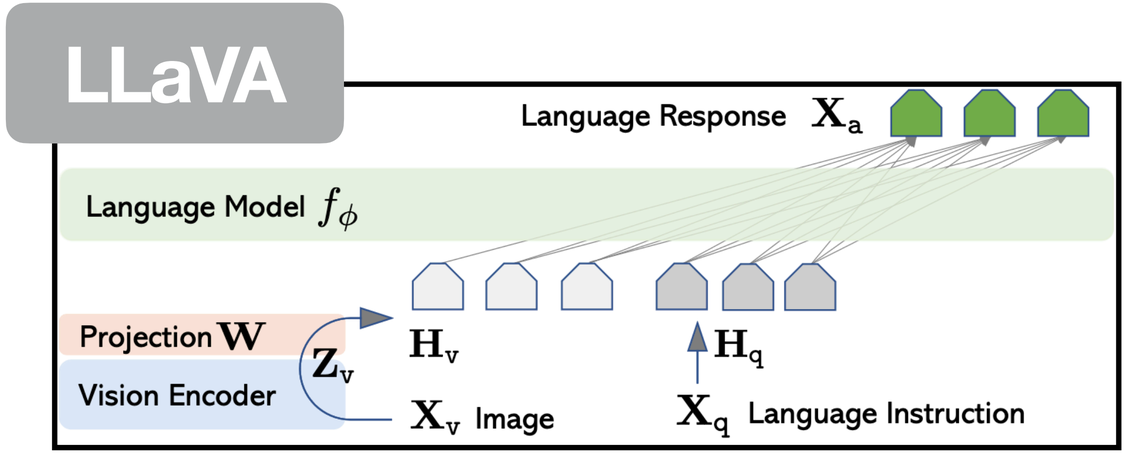
- Train adversarial referent placement policy on GPT-4o agents
- 450 base environments from SceneNet++
- Rejection sampling on low-quality and constraint-violating scenes
- 27,504 total generated scenes

Data
Human-written references

- Crowdsourcing with 194 workers on Prolific
- Each speaker referring expression paired with three listener clicks
- 2,970 total references in 1,485 scenes

Methods & models

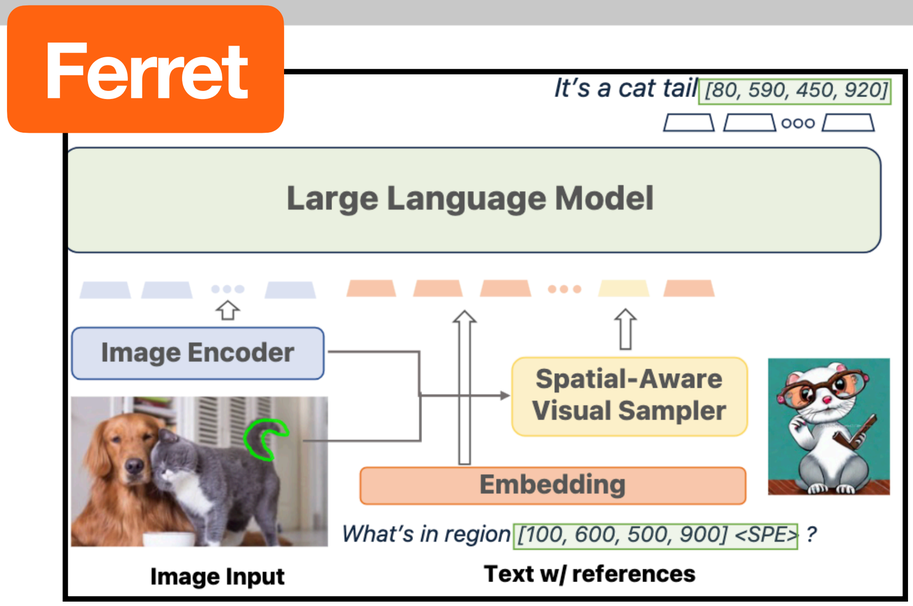
- **Pre-trained vision-language models,** e.g., LLaVA
  - Independently embed image and text into tokens
  - Jointly process multimodal tokens with transformer
  - Pre-train with QA pairs generated from captions
  - Fine-tune end-to-end with synthetic instruction data

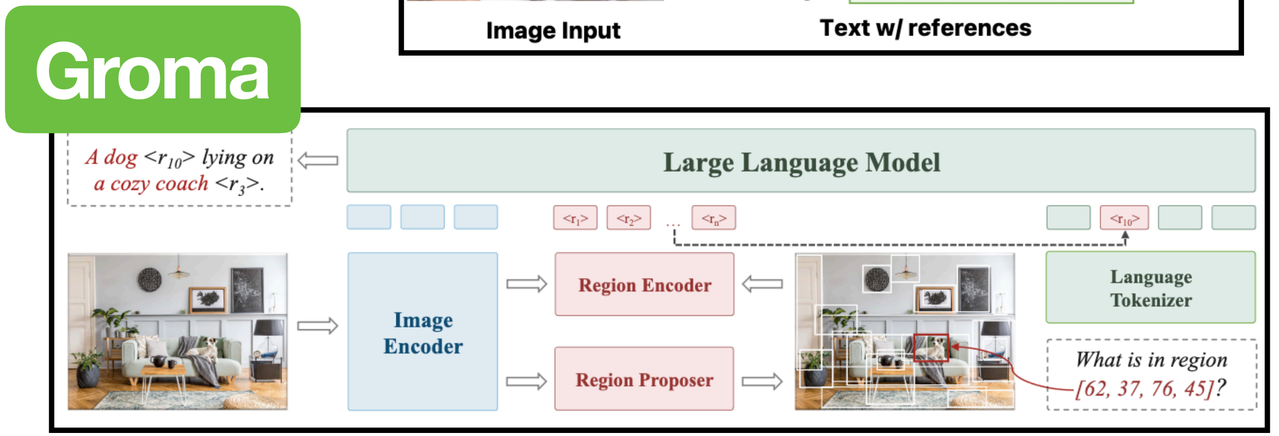


Liu et al. 2023, Liu et al. 2024

Methods & models

- **Pre-trained vision-language models,** e.g., LLaVA
- **Fine-grained vision-language models,** e.g., Ferret, Groma learn region embeddings

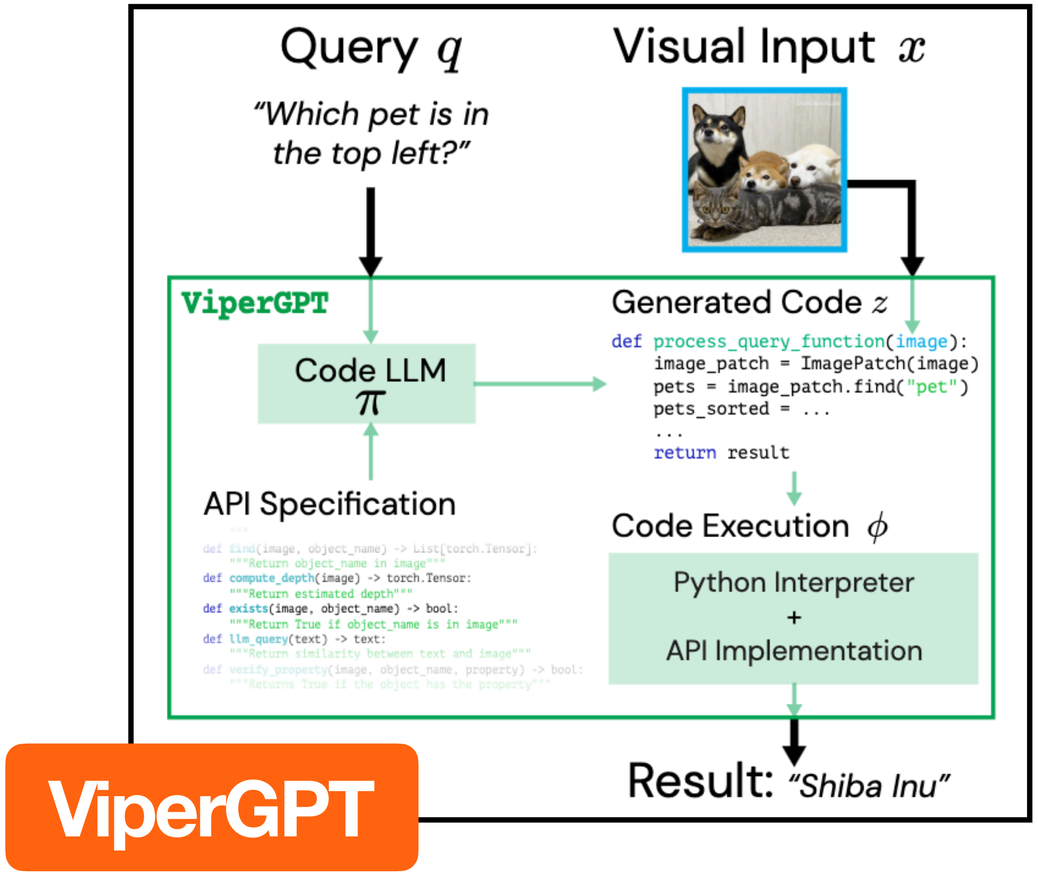




You et al. 2023, Ma et al. 2024

Methods & models

- **Pre-trained vision-language models,** e.g., LLaVA
- **Fine-grained vision-language models**
- **Modular visual reasoning systems,** e.g., ViperGPT
  - Text is mapped to a Python program
  - Python program is executed on top of the visual image using an API that implements common computer vision functionalities



Surís et al. 2023, Subramanian et al. 2023, Gupta and Khembavi 2023

Methods & models

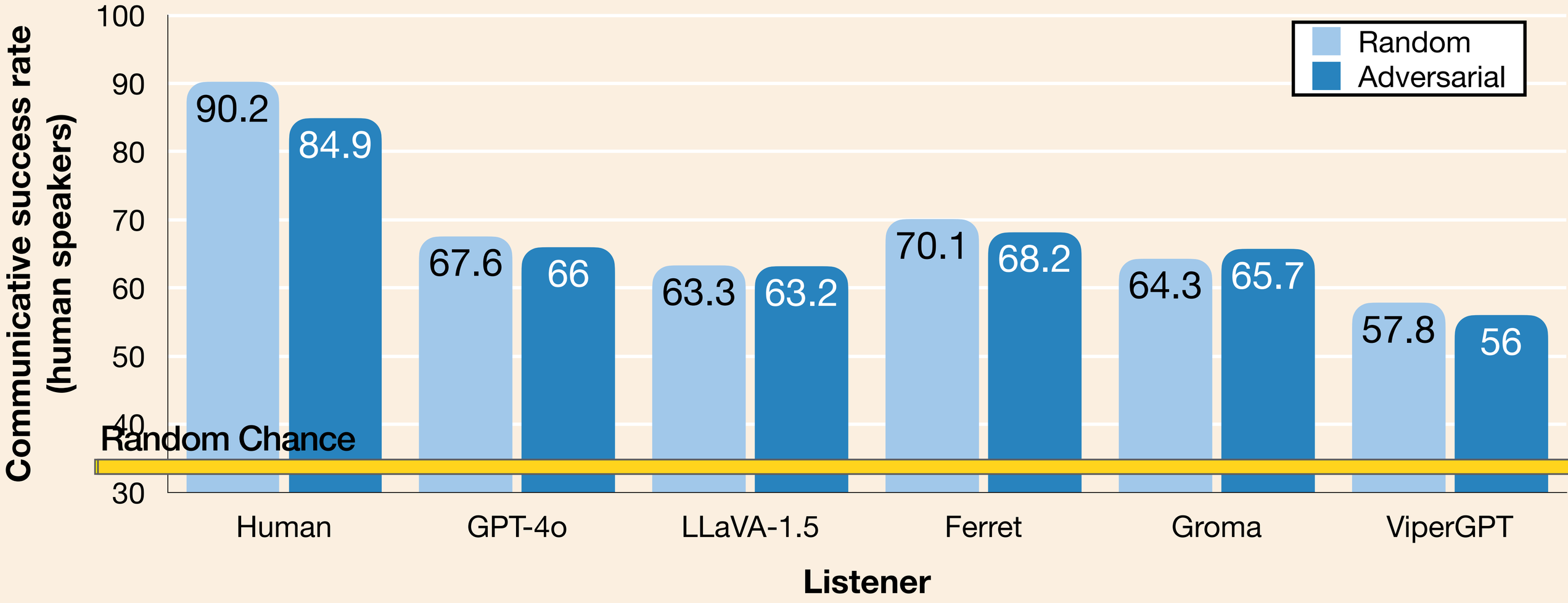
- **Pre-trained vision-language models,** e.g., LLaVA
- **Fine-grained vision-language models**
- **Modular visual reasoning systems**
- **API-gated models,** e.g., GPT-4o
  - Architecture: ???
  - Training: ???

GPT-4o

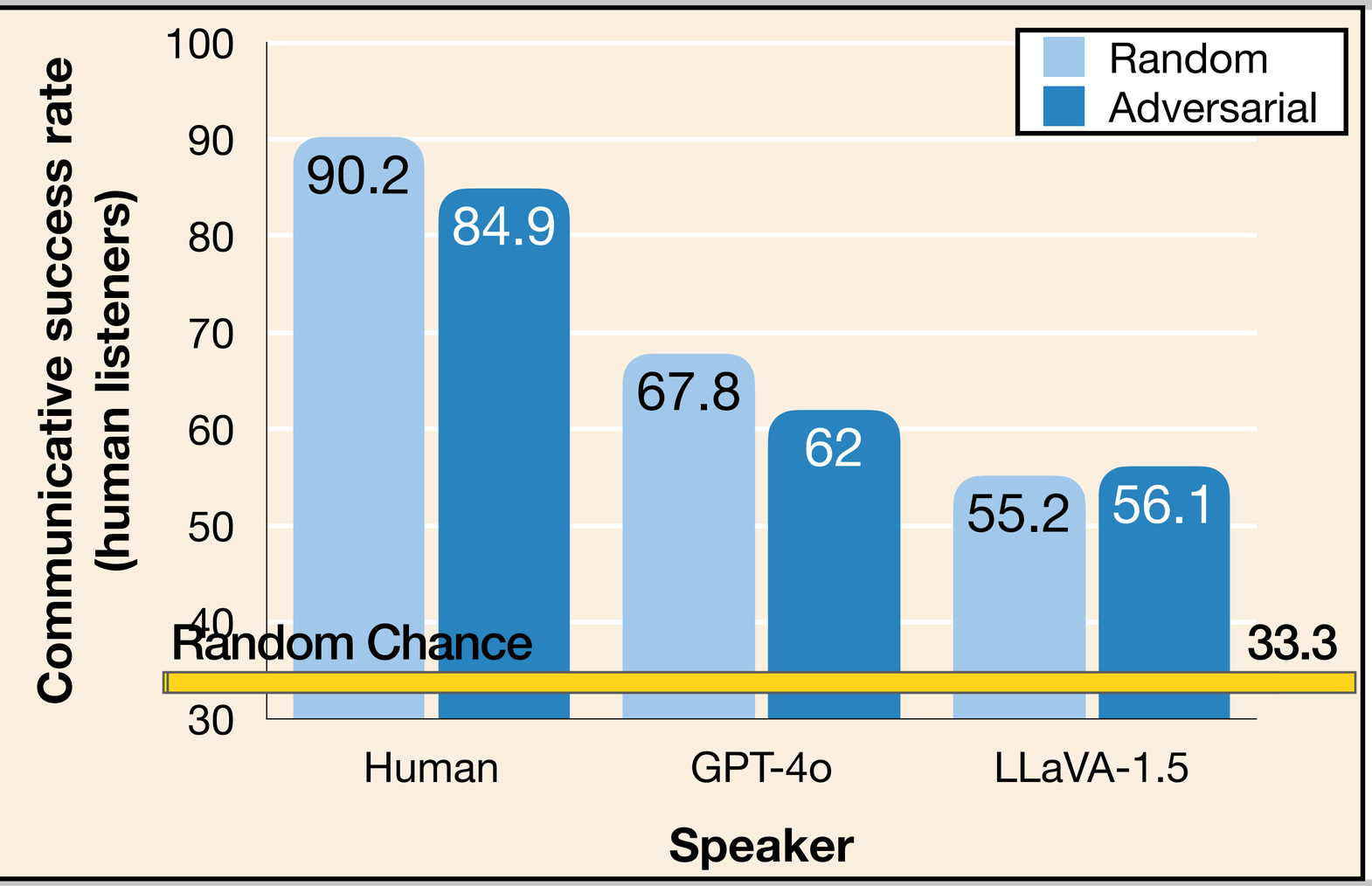
Inference & evaluation

- **Speaker:** maps from observation to reference
  - Human
  - GPT-4o
  - LLaVA-1.5
- **Listener:** multiple choice task; provided three candidate bounding boxes
- **Main evaluation metric:** communicative success between a pair of agents

Results
Human-written references



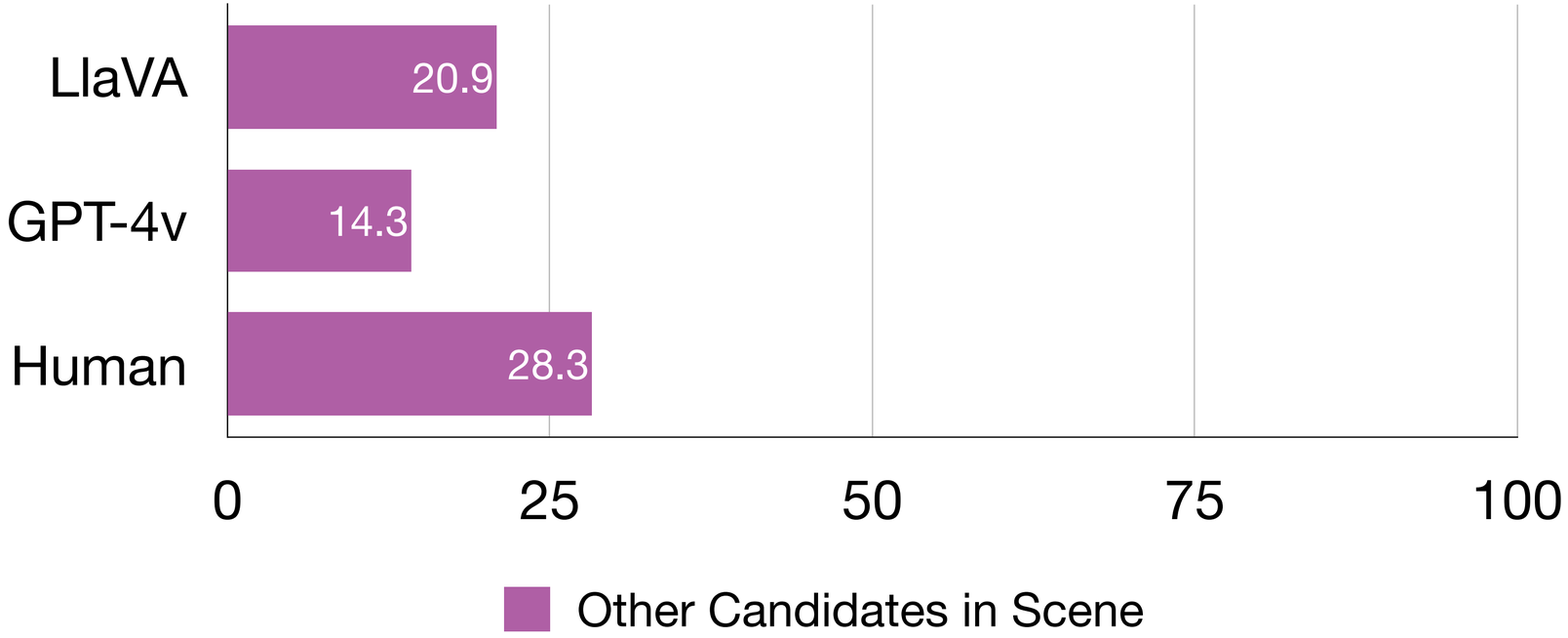
Results
Model-generated references



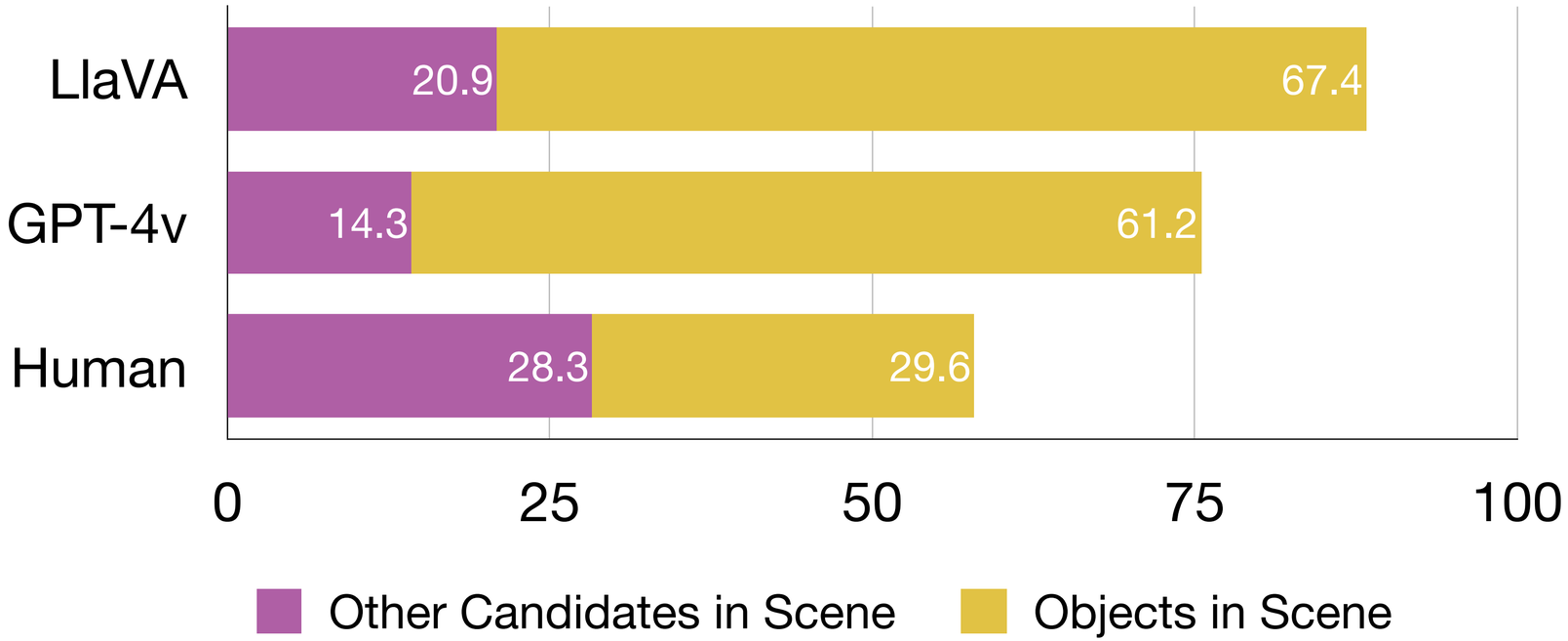
Adversarial reference placement significantly decreases performance

- In GPT-4o - GPT-4o interactions, drop from 61.1 to 57.2
- In human-human interactions, drop from 90.2 to 84.9
- Increasing FOV overlap correlates significantly with improved success in human-human interactions

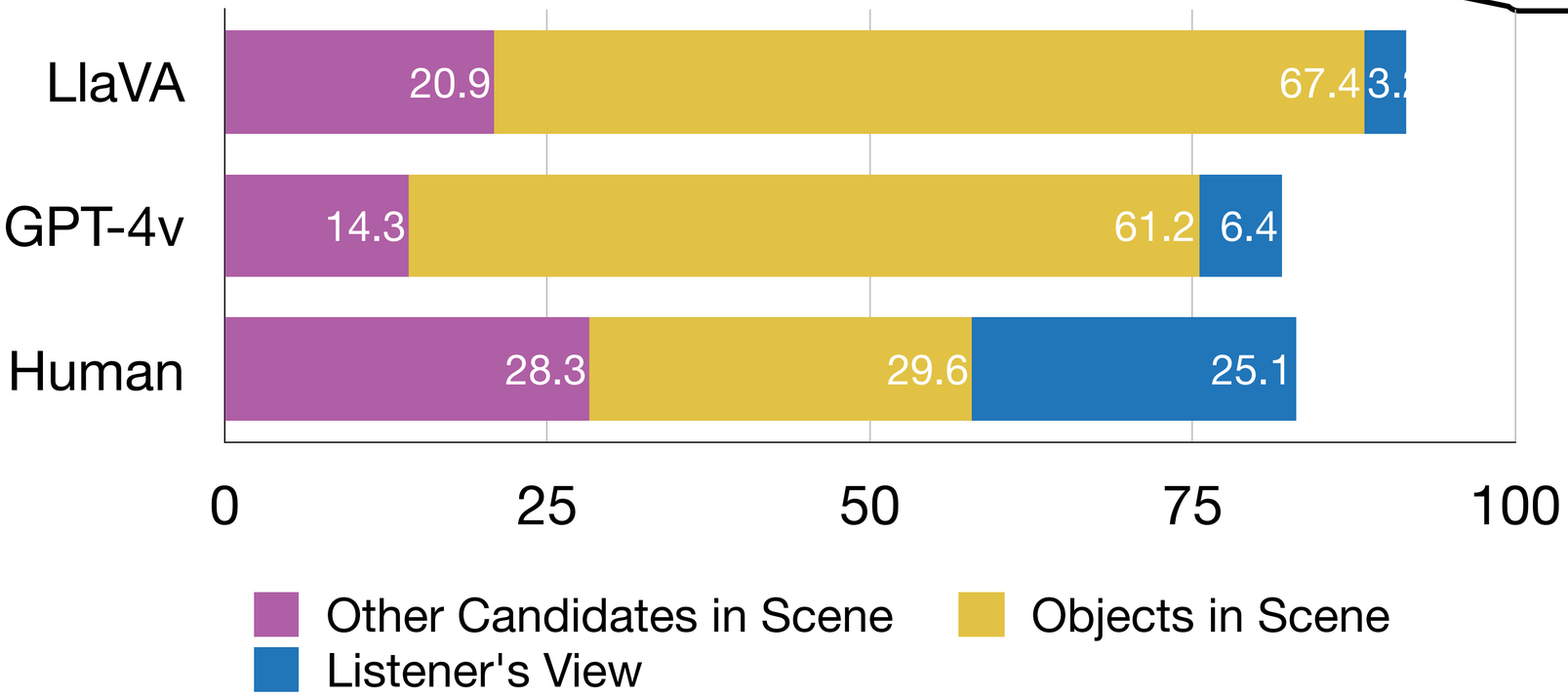
Referential strategies



Referential strategies

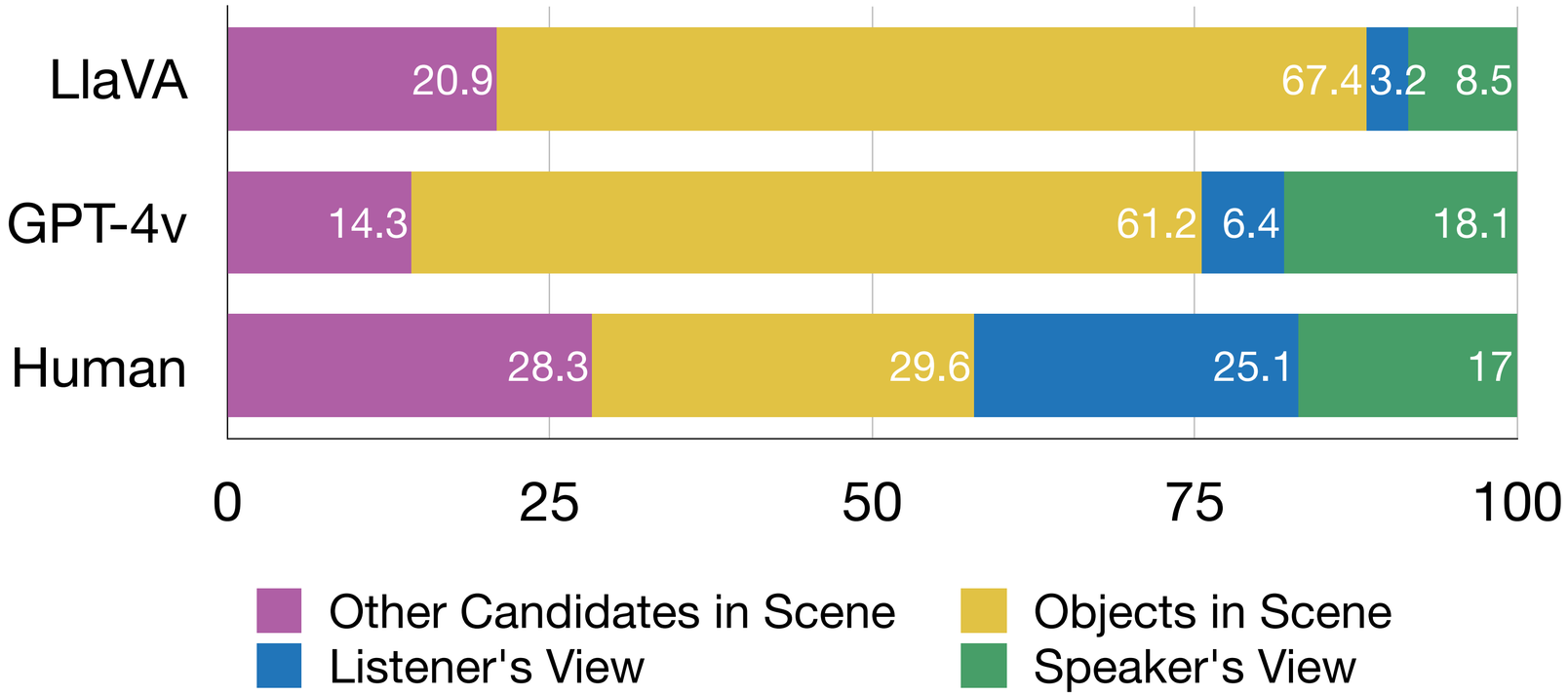


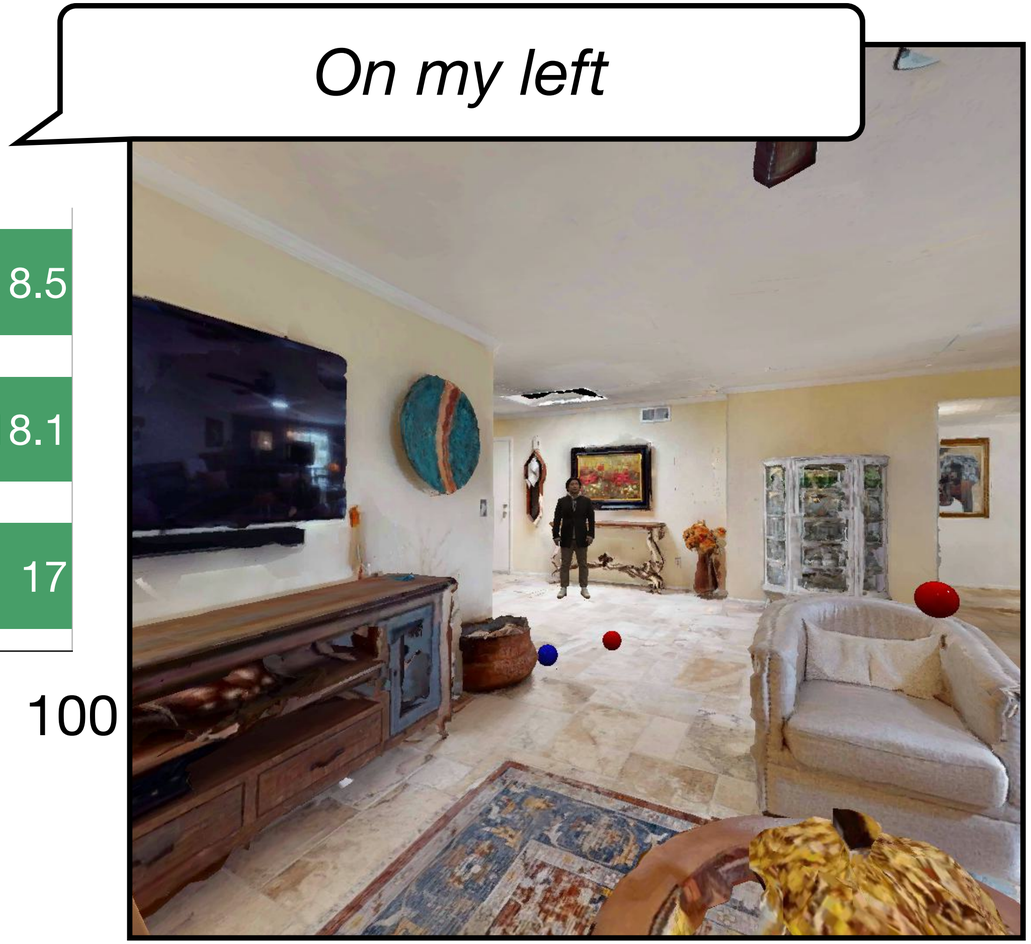
Referential strategies



The rightmost one from your perspective

Referential strategies





Learning from communicative success

- **Goal:** improve reference generation
- **Motivation:** empirical observations of language interpretation provide evidence of utterance meaning, regardless of speaker intent
- **Experimental setup**

Learning from communicative success

- **Goal:** improve reference generation
- **Motivation:** empirical observations of language interpretation provide evidence of utterance meaning, regardless of speaker intent
- **Experimental setup**
  - Sample references from a speaker model for 195 scenes

$$x \sim p_s(o_s, \mathcal{R}, t; \theta)$$

Learning from communicative success

- **Goal:** improve reference generation
- **Motivation:** empirical observations of language interpretation provide evidence of utterance meaning, regardless of speaker intent
- **Experimental setup**
  - Sample references from a speaker model for 195 scenes $x \sim p_s(o_s, \mathcal{R}, t; \theta)$
  - Acquire observations of comprehension (i.e., listener judgments) $\hat{t} \sim p_l(o_l, \mathcal{R}, x; \phi)$

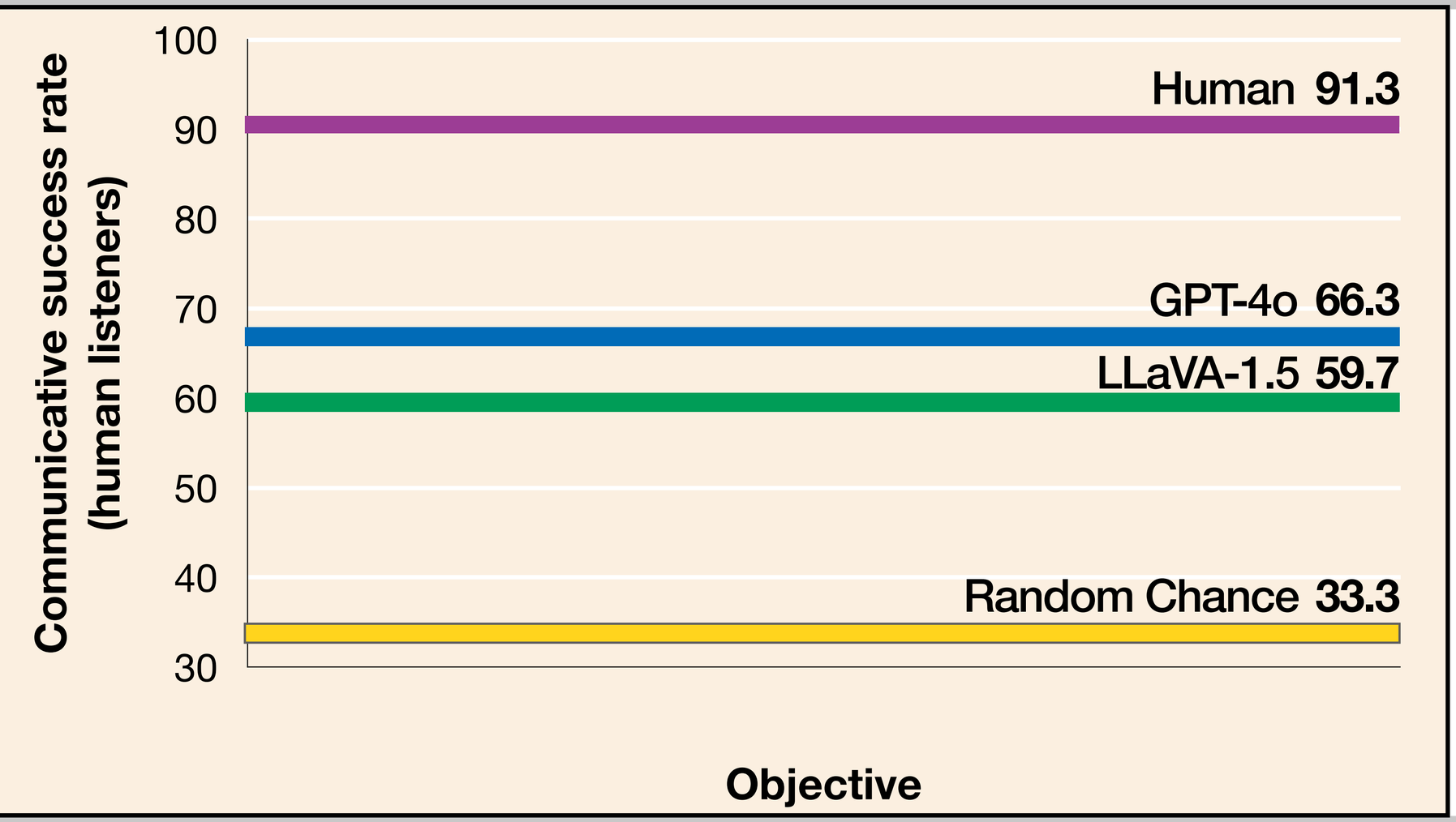
Learning from communicative success

- **Goal:** improve reference generation
- **Motivation:** empirical observations of language interpretation provide evidence of utterance meaning, regardless of speaker intent
- **Experimental setup**
  - Sample references from a speaker model for 195 scenes
  - Acquire observations of comprehension (i.e., listener judgments)
  - Fine-tune model to maximize success rate of interaction w/ PPO
  - Experiment with different rewards

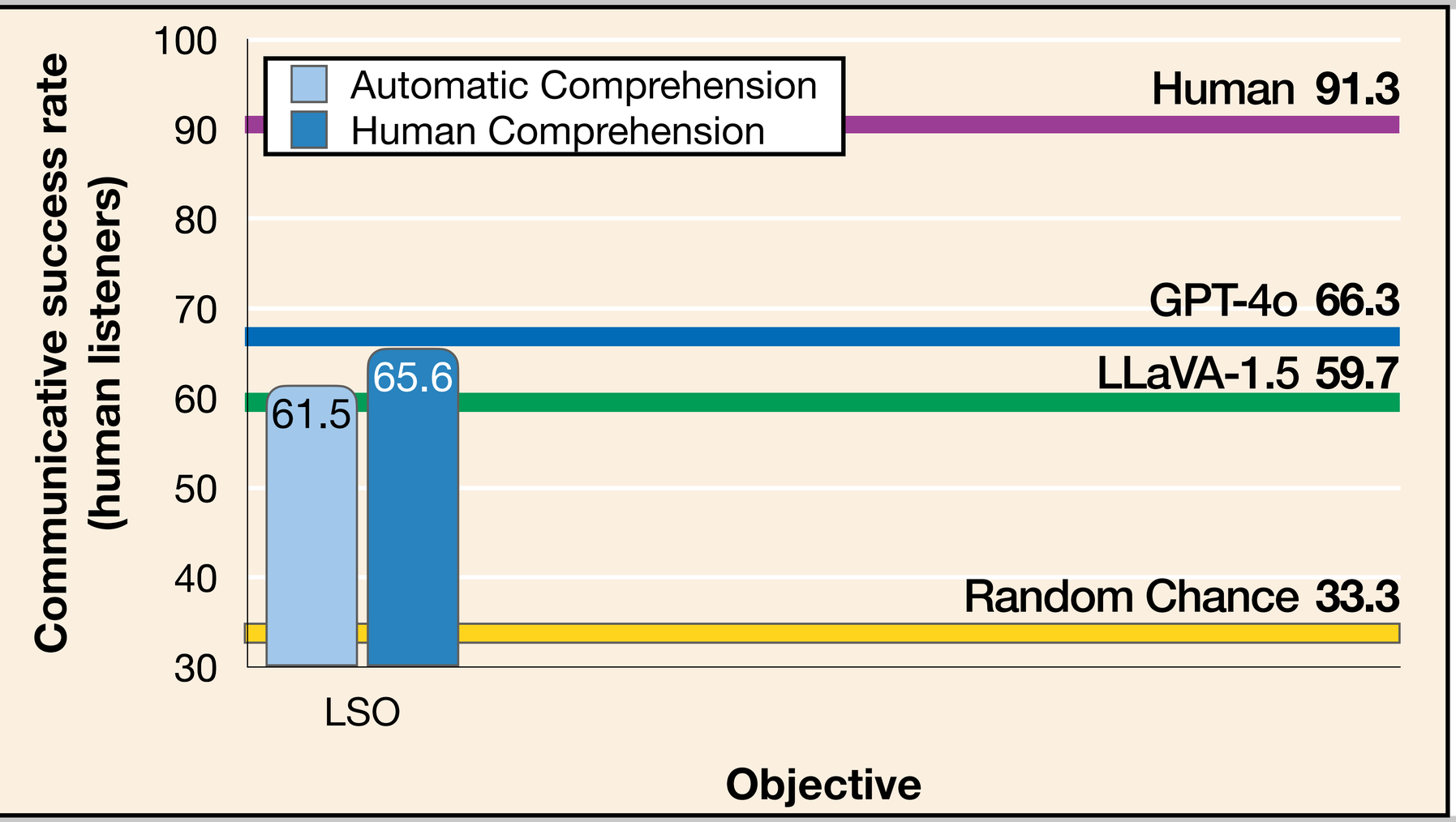
$$x \sim p_s(o_s, \mathcal{R}, t; \theta)$$

$$\hat{t} \sim p_l(o_l, \mathcal{R}, x; \phi)$$

Communicative success rates evaluating with human listeners

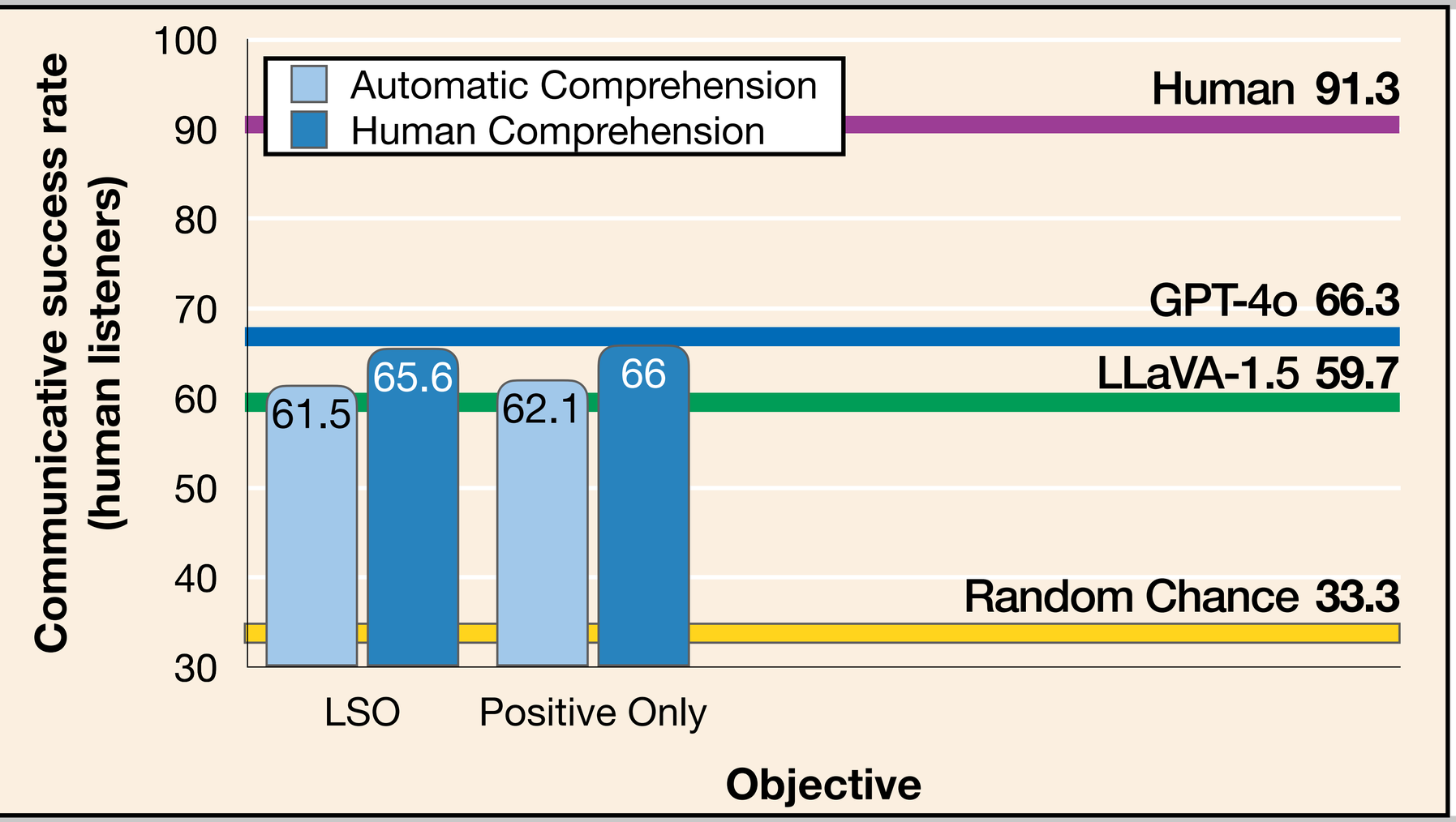


Communicative success rates evaluating with human listeners



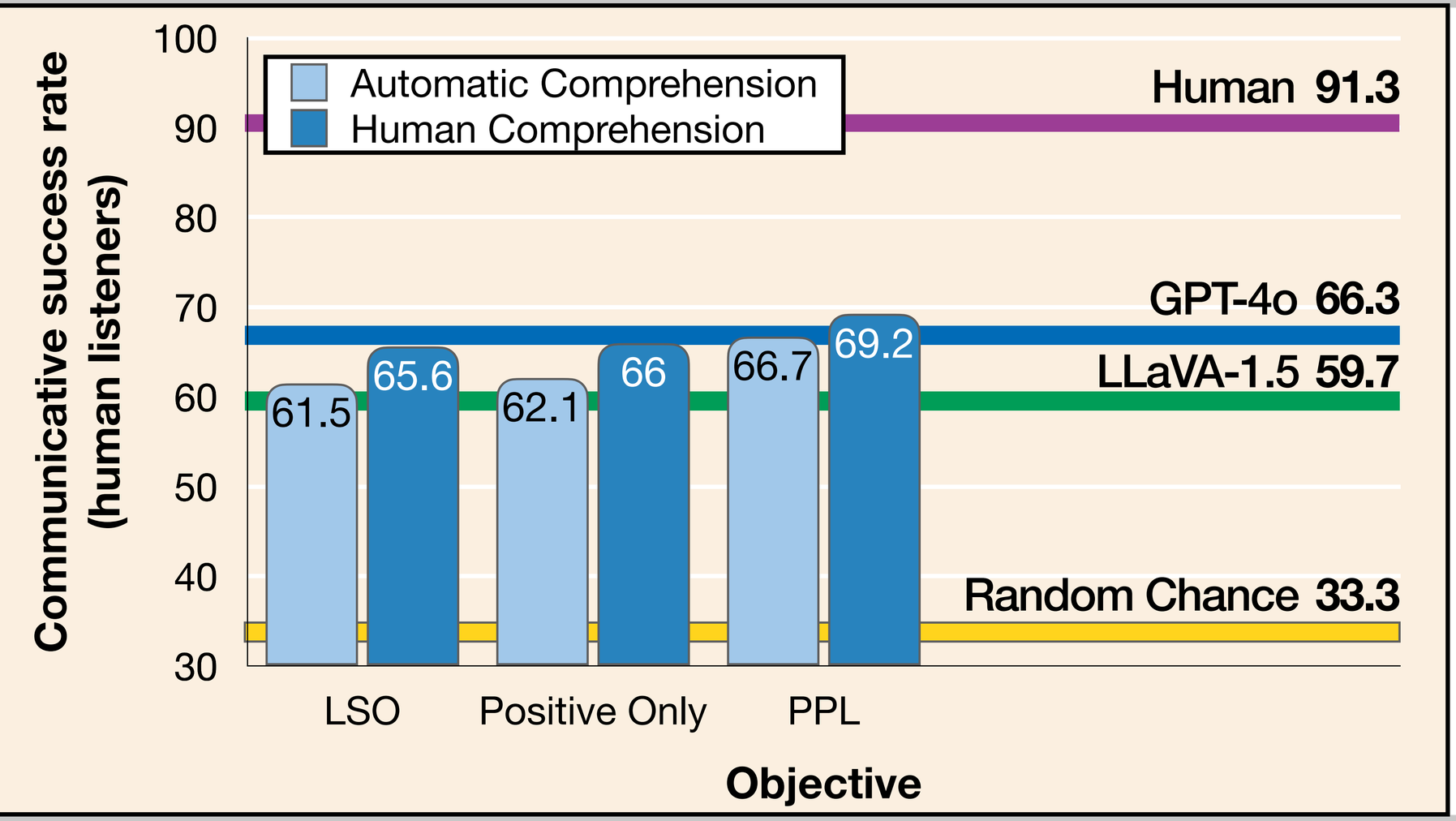
- **Learning from successes only (LSO)**
  - Positive rewards *only* when $t = \hat{t}$

Communicative success rates evaluating with human listeners



- **Positive only**
- **All references get a positive reward**
  - If successful, pair reference with intended referent
  - If unsuccessful, pair reference with listener-chosen referent

Communicative success rates evaluating with human listeners

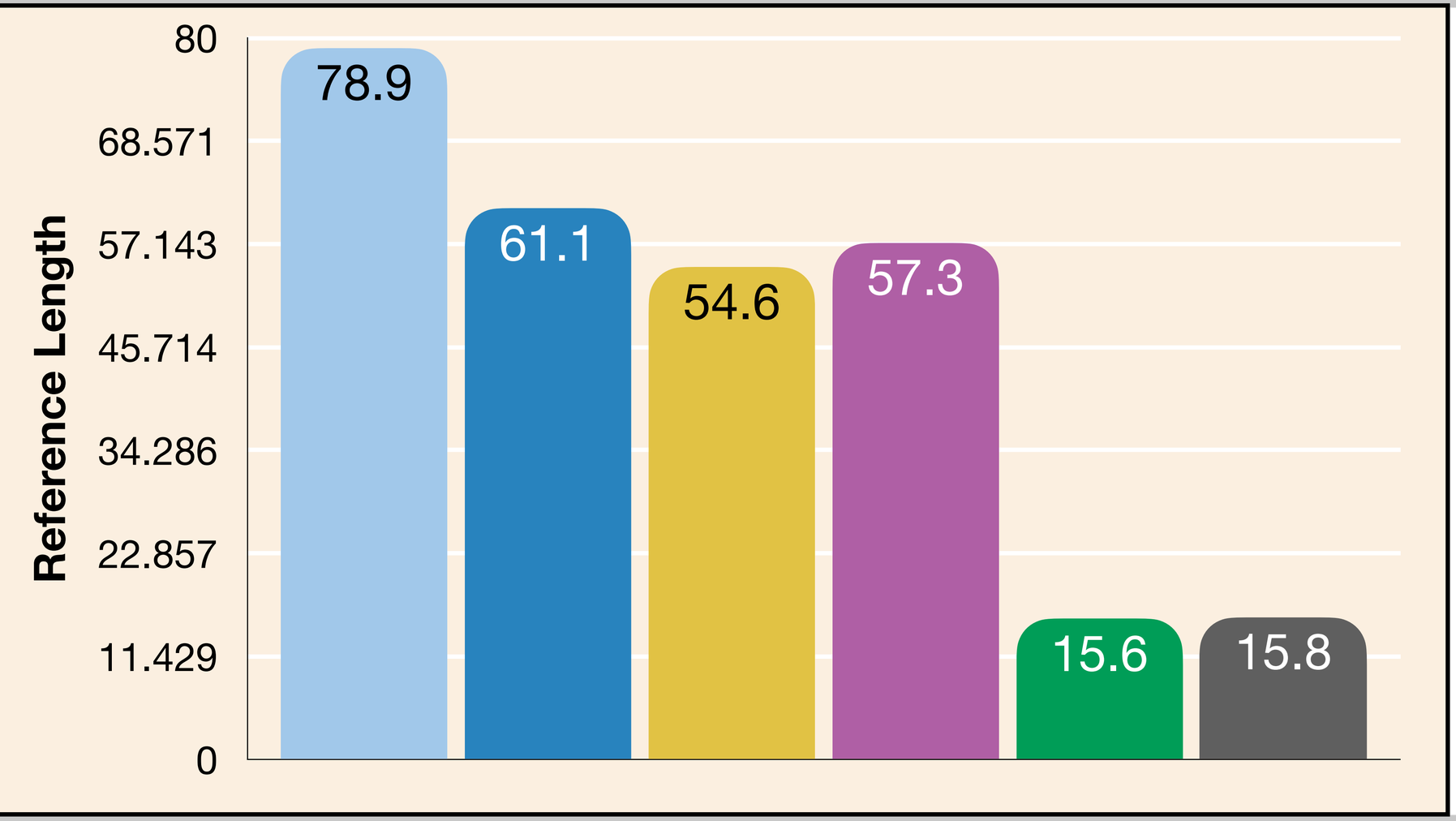


- **Pairwise preference learning**
- For communicative failure, maximize gap in probability between selected referent and intended referent

$$p_s(x \mid o_s, \mathcal{R}, \hat{t}; \theta') - p_s(x \mid o_s, \mathcal{R}, t; \theta')$$

Outperform GPT-4o with just 195 examples!

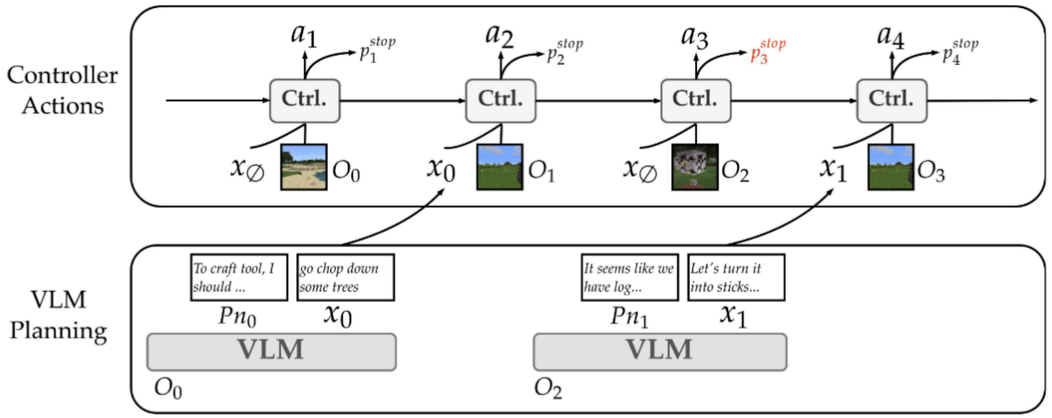
Communicative success rates evaluating with human listeners



- **Only PPL matches human-length references**

GPT-4o
Pre-trained LLaVA
LSO
Positive Only
PPL
Human

A few things we just cooked…
Instructable agents

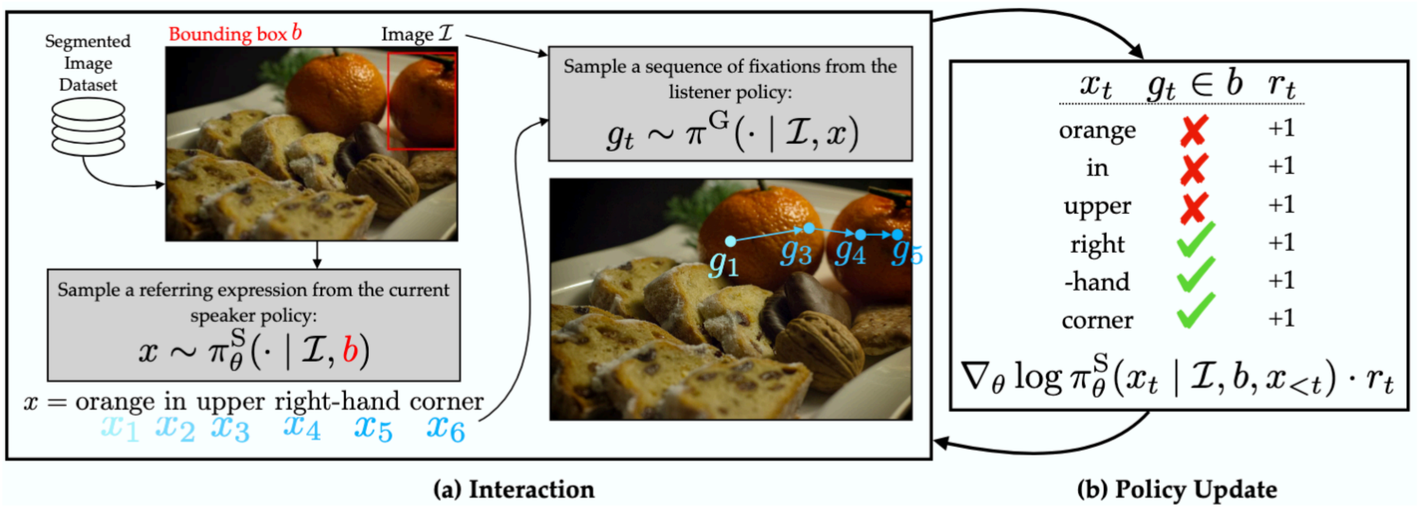


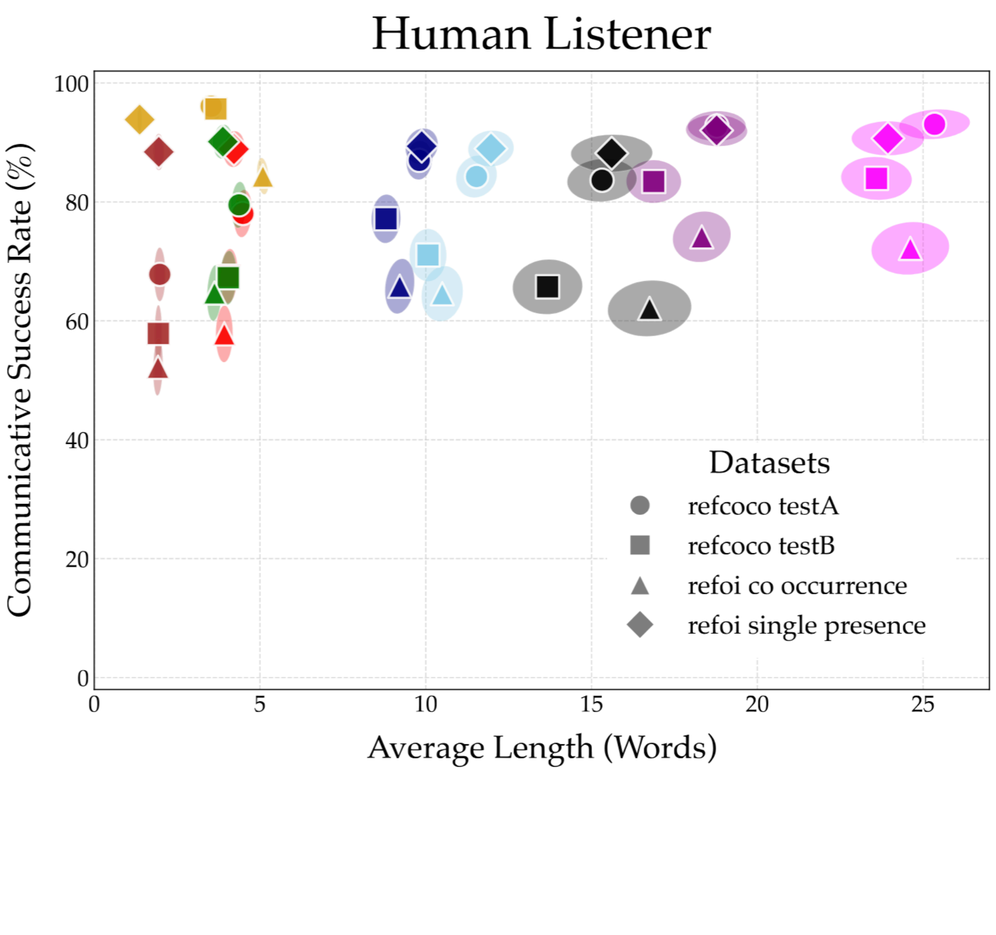
		Single-Agent Tasks				Multi-agent Tasks		
		Atari	MC-Dia	Craft	DML	OCook	Pico	MindC
VLMs	Gemma-3-27B (Team et al., 2025)	880	9.7	13.8	71	187.4	68.9	53.0
	llava-v1.6-34b (Liu et al., 2023)	862	9.9	12.8	74	187.2	63.5	51.6
	Qwen-VL-2.5-72B (Team, 2024)	878	11.1	13.4	77	192.3	68.4	58.5
	GPT-4o (Hurst et al., 2024)	891	11.7	14.1	76	193.2	70.1	70.2
Ablation	Qwen-VL-2.5-72B (Direct VLM Finetune)	581	10.1	7.6	45	150.1	30.6	38.2
	GPT-4o (w/o controller)	670	10.4	8.7	56	180.4	50.3	50.2
	Controller-Only	809	8.2	12.6	67	170.2	30.7	40.0
Prev. SOTA	Dreamerv3 (Hafner et al., 2023)	811	8.6	10.5	65	-	-	-
	Voyager (Wang et al., 2023a)	-	11.8	-	-	-	-	-
	MARL Baseline (Kuba et al., 2021)	-	-	-	-	182.5	50.8	44.9
	Mindcraft (Claude) (White et al., 2025)	-	-	-	-	-	-	49.0
	RT-2 (Zitkovich et al., 2023)	457	6.2	4.7	36	124.7	34.0	33.1
	DEPS (Wang et al., 2023b)	-	9.4	-	-	-	-	45.6
	LS-Imagine (Li et al., 2025)	-	9.6	-	-	-	-	50.1
	JARVIS-1 (Wang et al., 2024)	-	12.3	-	-	-	-	54.1
	QMIX (Rashid et al., 2020)	-	-	-	-	187.2	58.5	53.2

Decoupling planning and control for instructable agents — under submission at COLM

Zineng Tang, Kelsey R. Allen, Sjoerd van Steenkiste, Ishita Dasgupta, Alane Suhr

A few things we just cooked…
Learning from estimated gaze





Learning to refer from estimated listener gaze— under submission at COLM
Téa Wright, Alane Suhr

Thanks!

Nicholas Tomlin

Naitian Zhou

Circle Chen

Laura Ma

Lauren Vinh

Ell French

Zineng Tang

Tingting Du

Seun Eisape

Téa Wright

Eve Fleisig

Tianjiao Zhang

Alexander Koller

Lingjun Mao