Gesturing Toward Abstraction: Multimodal Convention Formation in Collaborative Physical Tasks

Kiyosu Maeda
km9567@princeton.edu
Princeton University
Princeton, NJ, USA

William P. McCarthy
wmccarthy@ucsd.edu
UC San Diego
La Jolla, CA, USA

Ching-Yi Tsai
ching-yi@princeton.edu
Princeton University
Princeton, NJ, USA

Jeffrey Mu
jeffrey_mu@brown.edu
Brown University
Providence, RI, USA

Haoliang Wang
hlwang@mit.edu
MIT
Cambridge, MA, USA

Robert D. Hawkins
rdhawkins@stanford.edu
Stanford University
Stanford, CA, USA

Judith E. Fan
jefan@stanford.edu
Stanford University
Stanford, CA, USA

Parastoo Abtahi
parastoo@princeton.edu
Princeton University
Princeton, NJ, USA

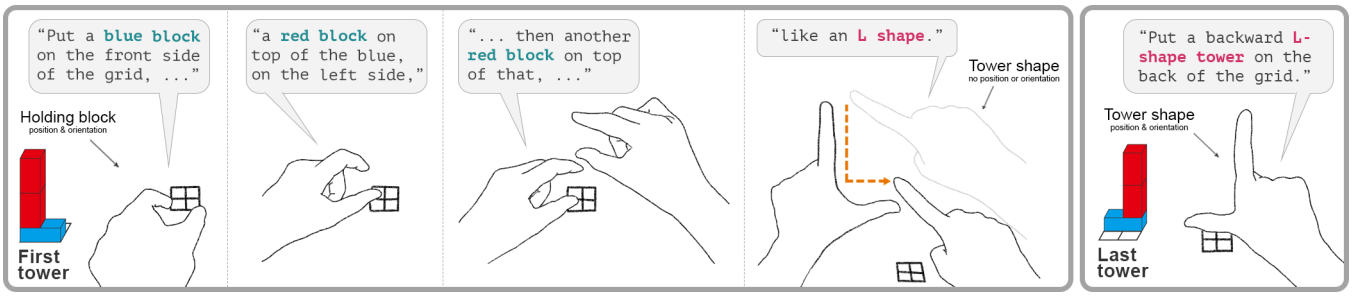

Figure 1: We study how people instruct a partner in a physical assembly task using speech and gesture. Participants begin with block-level instructions (low abstraction), then establish linguistic and gestural conventions and use them to describe the overall tower shape (high abstraction), while employing cross-modal redundancy to emphasize changes in position and orientation.

Abstract

A quintessential feature of human intelligence is the ability to create ad hoc conventions over time to achieve shared goals efficiently. We investigate how communication strategies evolve through repeated collaboration as people coordinate on shared procedural abstractions. To this end, we conducted an online unimodal study (n = 98) using natural language to probe abstraction hierarchies. In a follow-up lab study (n = 40), we examined how multimodal communication (speech and gestures) changed during physical collaboration. Pairs used augmented reality to isolate their partner's hand and voice; one participant viewed a 3D virtual tower and sent instructions to the other, who built the physical tower. Participants became faster and more accurate by establishing linguistic and gestural abstractions and using cross-modal redundancy to emphasize key changes from previous interactions. Based on these findings, we extend probabilistic models of convention formation to multimodal settings, capturing shifts in modality preferences. Our findings and model provide building blocks for designing convention-aware intelligent agents situated in the physical world.

This work is licensed under a Creative Commons Attribution 4.0 International License.

CHI '26, Barcelona, Spain

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790618

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**; *HCI theory, concepts and models.*

Keywords

Multimodal conventions, hand gestures, abstraction, modality preference, complementary, augmented reality, rational speech act

ACM Reference Format:
Kiyosu Maeda, William P. McCarthy, Ching-Yi Tsai, Jeffrey Mu, Haoliang Wang, Robert D. Hawkins, Judith E. Fan, and Parastoo Abtahi. 2026. Gesturing Toward Abstraction: Multimodal Convention Formation in Collaborative Physical Tasks. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3772318.3790618

1 Introduction

People leverage multimodal signals to convey their intentions and understand others' intentions. While people can naturally use these cues in human-human communication, supporting multimodal communication with agents (e.g., AI-powered wearables and robots) is a challenging problem. Prior works have proposed AI systems that interpret multimodal cues [82, 98] and communicate their intentions in predictable and legible ways during one-time interactions [28]. However, across repeated interactions, people form ad hoc conventions with their collaborators, using more concise expressions to refer to objects or concepts [73] and convey more abstract or chunked

information to others. While these signals become increasingly ambiguous, sacrificing some informativeness [119], partners still communicate accurately and efficiently based on their shared history.

To date, most controlled studies have been focused on unimodal scenarios [29, 35, 51] or 2D collaborative tasks [60, 91]. Yet, many real-world tasks require multiple agents to coordinate their behavior over time in physical settings, using their hands to convey information that is not expressed in speech. For example, people frequently say ambiguous words (e.g., "this") while referring to an object with a deictic gesture [12]. Hands also communicate the shape of an object, describe the distance, and demonstrate actions [10, 34]. Understanding how such multimodal signals *evolve* during iterated collaboration in physical settings is crucial for designing agents that can form and use multimodal conventions.

Our study examined how people use voice and gestures to establish conventions and adapt multimodal communication over repeated physical collaboration. *First*, we ran an online *unimodal* study to probe how partners coordinate on linguistic abstractions; dyads shifted from block-based descriptions to concise tower-level expressions, reducing instruction length over repetitions. *Then*, we conducted a controlled AR lab study that isolated speech and gestures while participants built physical towers. Results indicated that participants' instructions were shortened by forming gestural and linguistic abstractions and shifting from block-based to tower-based instructions. We also found that shared abstractions were introduced near the end of the first exchange and later used at the beginning of instructions. Furthermore, redundancy in multimodal signals increased for tower-based instructions; participants used both modalities to describe changes in the position and orientation of target towers, emphasizing variations across repetitions.

Based on our observations, we built multimodal computational agents that probabilistically reason about each other's beliefs and continuously learn abstract tower representations in our task setting. We extended the Rational Speech Act (RSA) framework [31, 44] to develop a model of convention formation [49] that can operate over a multimodal lexicon (i.e., a mapping between symbols and utterances/gestures) and consider different combinations of multimodal signals (i.e., redundant, complementary, and language-only) by assuming that the lexicon returns continuous real values [26].

We conducted simulations to answer two research questions: **(RQ1)** Can agents acquire abstract tower representations over repeated interaction? and **(RQ2)** can the model capture how modality use shifts for different participants? The model successfully acquired abstract programs and captured modality-dependent behavioral shifts across repetitions. The proposed computational models and the findings from the study will pave the way to design multimodal agents that form conventions with humans for efficient and successful iterated collaboration in physical tasks.

Contributions:

- Large-scale **online unimodal study** investigating natural language communication over repeated interactions.
- AR-mediated **multimodal lab study** examining gesture and speech communication, including a multimodal dataset to support future research in repeated physical collaboration.
- **Computational model** and simulation, exhibiting behaviors aligned with study findings, with a flexible design extendable to more synchronous and symmetric communication.

2 Related Work

2.1 Gestures in Thought and Communication

Gestures play a critical role in cognitive processes when thinking and speaking and have implications for both learning and communication [41, 64, 67, 69, 92]. Gestures have been shown to enhance spatial reasoning [21, 43] and improve problem solving and concept learning [9]. They can promote knowledge transfer [94] and convey visual concepts difficult to express in speech [2]. Gestures also facilitate language comprehension by adding information to speech. People use *deictic* gestures to refer to objects or concepts [18, 61], *symbolic* gestures for abstract meanings [87], and *iconic* gestures to disambiguate utterances [52].

Gestures change over extended interaction. In mental rotation tasks, people initially use gestures that simulate direct manipulation of objects and later represent abstract object motions [20], and gesture frequency also changes over time with changes in strategy [107]. In iterated communication, gestures become more efficient and less kinematically complex [97]. Between partners, repeated interaction leads to convergence on shared gestural conventions [30]. People adapt to each other and synchronizing their motion through *gesture entrainment* [4, 68, 95, 112], and systematic structures can emerge over time within a community [104].

However, gestures rarely occur in isolation and typically co-occur with speech, distributing information across modalities [42, 92, 100]. We investigate how people use speech and gesture to communicate during physical assembly [123], examining how information is allocated across channels and how it changes over time.

2.2 Multimodal Cues in Collaborative Tasks

Multimodal signals are fundamental to human communication. As computing systems become more integrated into physical environments, it is crucial to interpret human intent from multimodal cues [12, 98] and to generate information, not only through text but also with embodied output [108]. Extensive research has focused on enabling agents to interpret and present multimodal information, thereby enhancing human-computer communication. On the input side, understanding co-speech gestures has been used to specify spatiotemporal references in VR [53], disambiguate queries in AR [82], generate live visual effects [105], and provide robot assistance [84, 102]. On the output side, systems have presented multimodal robot intent [39] and embodied instructions for navigation [85], music learning [65], and cooking [121].

Analyzing human-human communication provides insights for these applications [28, 33]. Narayana et al. [93] analyzed gesture use in collaborative tasks, identifying common gesture types and how their use varies with speech availability. Gergle et al. [36, 37] investigated how visual information affects collaboration in 2D puzzle tasks, building on the conversational grounding framework [22] and the principle of least collaborative effort [23]. They found that with shared visual context, collaborators use visual cues (e.g., pointing, moving pieces) rather than explicit verbal grounding. In assembly tasks, multimodal communication significantly reduced completion time compared to speech alone [115, 116]. Gleeson et al. [40] analyzed these interactions to derive a gesture vocabulary and develop a robotic arm that communicates through gestures.

Recent studies have captured datasets of human–human communication (e.g., EGGNOG [114], HoloAssist [118]). We build on these by conducting a more controlled experiment that carefully isolates speech and gesture using AR and limits feedback between collaborators; this is needed for analysis of information across modalities and to support computational modeling. Moreover, representations have been introduced to encode gesture semantics and support multimodal understanding [17]. Prior simulations have also interpreted gesture and language instructions using predefined symbolic mappings to update beliefs about the physical world [76–78, 98]. However, there is limited work on modeling or simulating multimodal human–human communication in *repeated interactions* where collaborators adapt and coordinate their behaviors over time. We study how multimodal communication signals change during repeated collaborative physical tasks and model how beliefs about semantic conventions are updated.

2.3 Conventions in Repeated Collaboration

In repeated interactions, people form ad hoc conventions—solutions to recurring coordination problems with others [83]. For instance, people often shorten referring expressions to reduce the effort required to identify objects [13, 111] or align their word choices with those of their conversational partner when referring to the same object repeatedly, known as *lexical entrainment* [16, 56, 96]. People also chunk complicated steps into a single instruction to share their goals efficiently [50]. While those in different groups may struggle to understand others' conventions due to ambiguity [119], members within a group can readily disambiguate intentions by updating their expectations based on past interactions [23]. Thus, conventions are essential for successful and efficient collaboration.

Prior work in cognitive science has explored how conventions form [32, 35, 117] across a range of collaborative tasks, including reference games [14]. Over repeated interactions, people develop increasingly concise linguistic expressions [49, 73] and graphical depictions [29, 51, 91], enabling more efficient communication. This understanding has motivated computational approaches that model how AI agents can establish conventions [80, 99]. For example, Hua et al. [54] show that current vision–language models rarely form conventions with users and propose methods to support such linguistic adaptation [55]. Shih et al. [110] separate convention-dependent and rule-based behaviors during iterated collaboration to acquire conventions. However, most of these efforts focus on unimodal, text-based communication. There is limited work on multimodal conventions in multi-step physical tasks.

To address this, we first establish our experimental paradigm in a unimodal online study [88] and then conduct a lab study to investigate multimodal communication during repeated physical assembly [86]. Our findings and computational model aim to explain how people form multimodal abstractions when giving instructions on physical tasks and how they adjust the amounts of information across modalities (speech and gesture) over time.

3 Online Unimodal Study

The goal of our first, online unimodal study was to establish a paradigm for investigating how people simultaneously coordinate on a shared set of concepts and on a way of communicating about them. We explored this phenomenon in an assembly domain [7, 15, 89, 113] where participants encountered visual scenes populated by a recurring set of block towers. The scenes were hierarchically organized and could be represented at multiple levels of abstraction—for instance, either as whole structures or as combinations of simpler units. As participants viewed multiple scenes, we hypothesized that certain "chunks" would be preferred, grouping primitive elements (individual blocks) into more complex configurations [5, 6, 19]. However, these newly formed abstractions are only useful in this task if they can be successfully communicated to others, which requires overcoming the inherent risk of miscommunication that accompanies the use of new linguistic terms.

3.1 Method

3.1.1 Participants. We recruited 146 participants from Amazon Mechanical Turk and paired them into 73 dyads for our IRB-approved study. We excluded 24 dyads who failed to meet preregistered criteria ($\geq$ 75% reconstruction accuracy on $\geq$ 75% of trials, or self-reported confusion or non-fluency in English). The session lasted 30–50 minutes. Participants received a minimum compensation of $5.00 plus a performance bonus of up to $3.00.

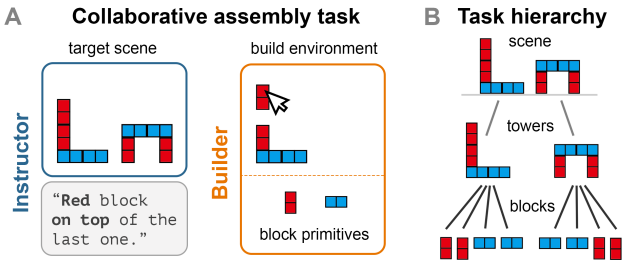

Figure 2: (A) Instructor viewed a target scene and gave assembly instructions to the Builder. (B) Scenes with two towers.

3.1.2 Stimuli. Each scene consisted of two towers built hierarchically from four domino-shaped blocks—two vertical and two horizontal (Figure 2B). We created three unique towers and used a *repeated* design where each tower appeared multiple times. All three tower pairs appeared in randomized order across four *repetitions*, yielding twelve trials. Each tower appeared equally on the left and right, ensuring no association between towers and their position.

Each participant was assigned a fixed role of *Instructor* or *Builder* and completed twelve trials. In each trial, the Instructor saw a target scene with block towers (Figure 2A), while the Builder viewed an empty grid for placing domino-like blocks. The Instructor provided step-by-step assembly instructions through a free-response text box, and the Builder used these to reconstruct the scene. They took turns as needed: on each Instructor turn, they sent one message (up to 100 characters), and on each Builder turn, they placed any number ($\geq$ 0) of blocks before selecting "done." Blocks had to be supported from below and could not be moved once placed. The Instructor saw the Builder's block placements in real time, but communication was otherwise unidirectional. A 30-second countdown appeared on each turn to encourage quick progress; exceeding the limit had no penalty. After all eight blocks were placed, participants received

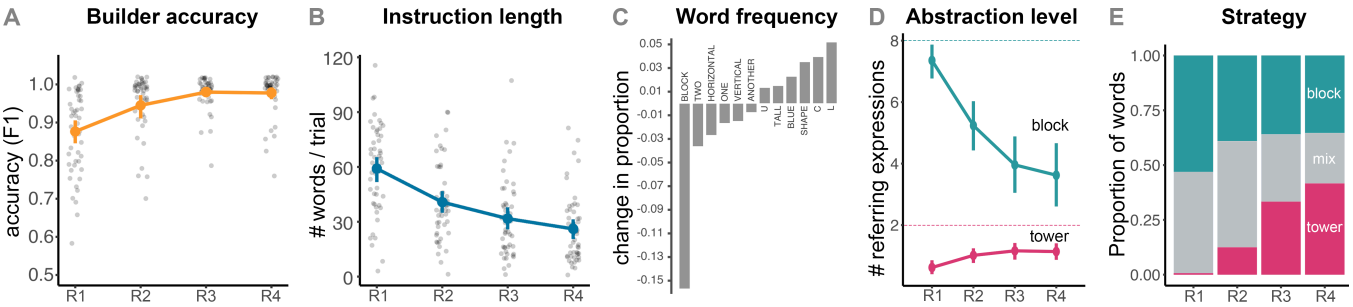

Figure 3: (A) Mean reconstruction accuracy improved across repetitions. (B) Mean instruction length per trial decreased across repetitions as dyads became more effective at collaborating. (C) Words with the largest positive or negative changes in frequency from R1 to R4. (D) Change in the number of block- and tower-level references. Dashed lines are the maximum possible number of blocks and towers. (E) The proportion of expressions exclusively referring to blocks or towers. Error bars: 95% CIs.

feedback on the mismatch between the target and reconstructed scenes before advancing to the next trial.

3.2 Results

3.2.1 Reconstruction accuracy improves across repetitions. Although each interaction spanned only twelve trials, **(H1)** we hypothesized that dyads would rapidly develop shared task representations, leading to more successful collaboration. We first verified that dyads could perform the assembly task. We measured performance using reconstruction accuracy, quantified as the F_1 overlap between the reconstructed tower and the target silhouette, which captures both missing blocks (recall) and extraneous ones (precision). F_1 ranges from 0 (no overlap) to 1 (perfect). Initial reconstructions were accurate (mean $F_1 = 0.88$, 95% CI = $[0.85, 0.90]$), roughly corresponding to one misplaced block, and final reconstructions were near ceiling ($F_1 = 0.98$, 95% CI = $[0.96, 0.99]$). A linear mixed-effects model predicting F_1 from repetition number, with random intercepts and slopes by dyad, revealed a significant improvement across repetitions ($\beta = 0.92$, $t(54.84) = 6.22$, $p < .001$; Figure 3A).

3.2.2 Communicative efficiency improves across repetitions. Having shown that Builders could reconstruct the towers from the instructions, we next examined a basic signature of increasing abstraction in language (Figure 4). Because the same towers recurred, **(H2)** we hypothesized that Instructors would exploit these regularities and provide more concise instructions over time. We analyzed both the number of words produced by the Instructor in each trial and the number of messages sent. A mixed-effects model with a fixed effect of repetition and maximal random effects for items and participants revealed that Instructors used significantly fewer words ($\beta = -8.53$, $t(36.9) = -9.58$, $p < .001$; Figure 3B) and sent fewer messages ($\beta = -18.1$, $t(24) = -7.11$, $p < .001$) across repetitions.

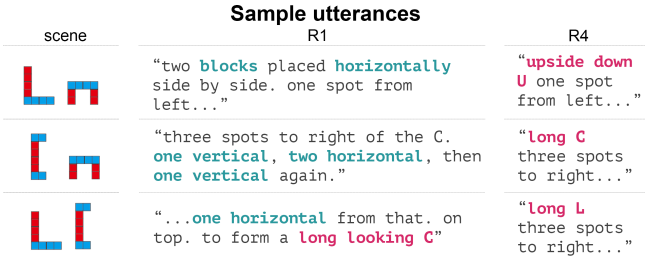

Figure 4: Example messages showing the emergence of tower-level expressions: *upside down U, long C,* and *long L.*

3.2.3 Level of referential abstraction increases across repetitions. What allowed dyads to perform better while also using fewer words? **(H3)** We hypothesized that the increase in communicative effectiveness reflects a gradual shift toward higher-level, more abstract instructions. We first conducted a qualitative analysis to explore this possibility. We tokenized all of the Instructor's messages into individual words and examined, across our entire dataset, which words changed the most in frequency from the beginning to the end of the experiment (Figure 3C). We observed that the frequency of low-level nouns like "block" and block-level modifiers like "horizontal" or "red" decreased the most, while that of high-level nouns like "L" or "C" and adjectives like "tall" increased the most.

To assess how strongly dyads converged on a common vocabulary for tower-level abstractions, we computed the Jensen–Shannon divergence (JSD) between word-frequency distributions of dyads per repetition. The mean JSD increased from R1 to R4 (0.080, 95% CI = [0.041, 0.118], $p = .004$), indicating that dyads' vocabularies became *more* divergent over time, and they developed distinct linguistic mappings from words to scene components.

We next conducted a more systematic analysis of message content. Four annotators, unaware of the study design and hypotheses, tagged each referring expression for block-level vs. tower-level references, yielding high agreement (intraclass correlation ICC = 0.83, 95% CI = $[0.82, 0.84]$). We fit a mixed-effects model with fixed effects of repetition, expression type (tower vs. block), and their interaction, as well as maximal random effects for dyad. The significant interaction ($b = 0.53$, $t(47.5) = 4.8$, $p < .001$; Figure 3D) indicated that block-level references decreased while tower-level references increased across repetitions. Mean block-level references dropped from 7.3 to 3.6, whereas tower-level references rose from 0.6 to 1.1. The shift was primarily driven by an increase in tower-level references and a decrease in block-level references, as well as messages containing a mixture of both (Figure 3E).

4 Physical Multimodal Study

In the second study, we examined a physical assembly task and changes in multimodal instructions (i.e., gesture and speech) over repeated interactions. Findings from the unimodal online study informed the design, including the selection of the task hierarchy, the number and shape of towers, and the number of repetitions to converge on task success.

To balance internal validity and ecological validity, we chose an experimental paradigm that tightly controlled participant communication rather than co-located, unconstrained settings. We used augmented reality (AR) to constrain communication channels while multimodal messages were sent between Instructors and Builders. Builders received embodied, spatially aligned 4D gestures with synchronous audio, and Instructors received images of built structures indicating the extent to which instructions were correctly interpreted. Factorizing study variables enabled analysis of how information was adjusted across modalities and how conventions emerged, from which we built a computational model for future convention-aware agents that have access to these multimodal signals (i.e., audio and hand tracking).

4.1 Method

4.1.1 Participants. We recruited 40 participants[1] (23 women, 14 men, 3 non-binary; ages 18-44, $M = 23.68$, $SD = 5.05$) for our IRB-approved study. All self-reported color vision (corrected as needed), sufficient hearing and manual dexterity for the study tasks, and fluency in English. Each session lasted ~1 hr, with $25 compensation.

4.1.2 Stimuli. The study included three block towers: *C*, *L*, and *TREE*, each consisting of three physical LEGO blocks (Figure 5). In all towers, the orientation and size of each colored block were fixed: the blue block (2x4x1) aligned with the x-axis, red (2x2x4) with the y-axis, and green (2x4x1) with the z-axis. We chose LEGO building as an example of a spatial cognitive task [109], with two simple 2D alphabetic towers (*L* and *C*) and one complex 3D tower (*TREE*) that lacked an obvious linguistic convention and used all three block primitives, making it difficult to represent with two hands.

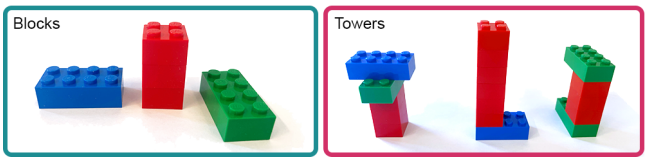

Figure 5: Physical LEGO blocks: blue (x-axis), red (y-axis), green (z-axis); and three-block towers: *TREE* (unknown mapping, 3D), *L* (alphabetic, 2D x-y), and *C* (alphabetic, 2D z-y).

We chose to have four repetitions (R1–R4), based on the first study and prior work showing that conventions are typically formed within the first four repetitions [73]. Because scene-level abstractions were not formed or used in the unimodal study, we used only one tower per trial. Each tower appeared once per repetition with a unique combination of position and orientation.[2] The order of the twelve towers was randomized within and across repetitions.

[1]We use P1-P20 to refer to the 20 pairs. See Appendix A.1.1 for demographics.
[2]See Appendix A.1.2 for an example sequence.

4.1.3 Apparatus. We examined how Instructors communicate to Builders using speech and gesture in a physically grounded task [74], and how these modalities evolve. Prior work shows that Builders' verbal (e.g., questions) and implicit non-verbal (e.g., pause) feedback influences Instructor behavior [74]. Because this feedback can vary in amount and timing, we used an AR setup to tightly control the feedback from Builders and eliminate confounding variables.

Two participants in separate rooms wore Meta Quest 3 headsets while seeing their physical surroundings (Figure 6). The Instructor's voice and hand movements were tracked, and a webcam captured images of the Builder's environment after each step. The system transmitted audio and hand keypoints to the Builder and tower images to the Instructor. This setup isolated the Instructor's speech and gestures, eliminated other non-verbal cues (e.g., eye gaze, facial expressions), and controlled the Builder's feedback, enabling analysis of information transfer and modality use over time.

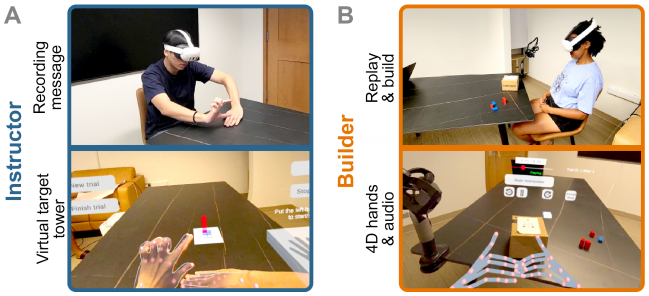

Figure 6: (A) The Instructor recorded a multimodal message (speech and gesture) in augmented reality (AR) describing how to build a virtual target tower with a specific pose. (B) The Builder replayed the message (audio and overlaid AR hands) and built the tower with physical blocks.

While prior work has explored gesture presentation methods, including 2D visualizations [11, 33], we presented 3D hand avatars with 24 joint keypoints and bones to provide high-resolution information about hand position and depth. We displayed the hands from an egocentric perspective because pilot testing showed that third-person views introduced perspective-taking ambiguity (e.g., "my left or your left?"), consistent with prior work [59, 70, 115]. This choice also supports anticipated applications (e.g., AI-powered glasses) that infer intent from egocentric inputs and present instructions from this perspective [3, 81, 122]. We used the Unity Depth API to align virtual hands with the physical environment.

For Builder feedback, pilot testing confirmed that 2D images were sufficient for this task. However, more complex towers or physical tasks may require live, photorealistic 3D renderings [57, 62, 79, 103].

4.1.4 Procedure. Participants were randomly assigned to the role of Instructor or Builder and worked in pairs in separate rooms. At the start of each trial, the Instructor viewed a 3D virtual twin of the target tower in AR, hidden from the Builder, and recorded multimodal messages (up to 2 minutes). Instructors could send multiple messages or all instructions at once. The Builder replayed each message using basic controls (i.e., play/stop, forward/backward, and seek bar), then assembled the tower with physical LEGO blocks on a 2×2 grid. Upon completing a step, a camera automatically captured

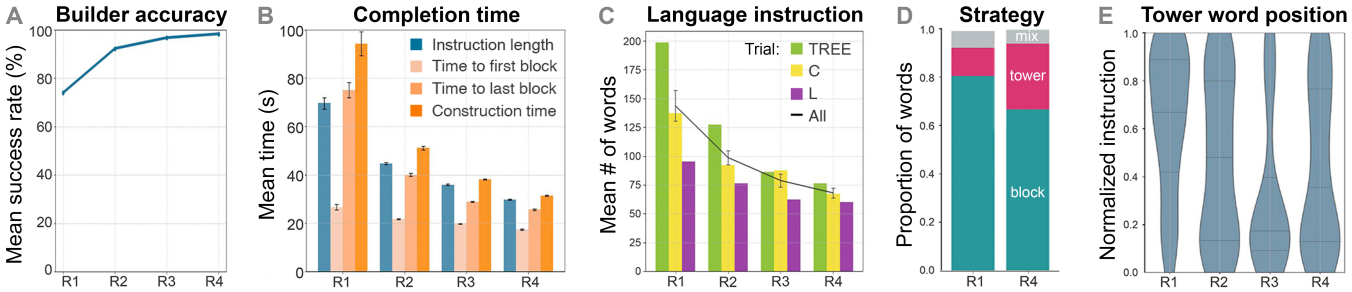

Figure 7: (A) Task performance in each repetition (R1-R4). (B) Instruction message length and builder time to place the first/last block, and complete the trial. (C) Number of words in instructions split by tower type. (D) Time proportion of block and tower words in R1 and R4. (E) Relative position of tower words during instructions. Error bars: standard error.

a photo of the reconstruction, which appeared in the Instructor's AR view. The Instructor could then send a new message with the next steps, including corrections, or end the trial if the reconstruction was complete. After each trial, the experimenter evaluated the final tower as correct or incorrect, and the result was displayed in AR for both participants. Participants completed 12 trials and were instructed to construct towers as quickly and accurately as possible.

4.1.5 Analysis. All Instructors' egocentric multimodal messages, 2D images of Builders' assemblies, and the experimenter's evaluations were saved for analysis. The entire session was also recorded by two external cameras to capture the Builders' and Instructors' actions. The footage of the construction was annotated with timestamps for each step. For analyzing speech, we first transcribed the audio in the messages with word-level timestamps using Whisper (large-v2) [101] and verified them manually for accuracy (11 of 256 sentences were corrected). Two authors annotated references in R1 and R4 for level of abstraction (block, tower), clarity (clear, ambiguous), and information (shape, position, orientation). For analyzing gestures, we built a custom Unity-based desktop application for visualizing hand keypoints and bones in 3D space from different views, synchronized with the corresponding audio and transcript.[3] Two authors manually annotated right- and left-hand gestures in R1 and R4 based on their level of abstraction (block, tower), type (static, dynamic), and information (shape, position, orientation). Given the complexity of the high-dimensional, time-series data, instead of computing inter-rater reliability, we took a collaborative coding approach, where two authors discussed until they agreed on the coding or consulted a third author to resolve disagreements.

5 Multimodal Study Results

5.1 Task Performance

5.1.1 Success rate. Task success rate was defined as the proportion of valid messages in which Builders correctly constructed the towers. **(H1)** We hypothesized that task success would improve across repetitions. The success rate improved from 74.03% in R1 to 98.36% in R4 with only one unsuccessful trial (Figure 7A).

5.1.2 Completion time. Instruction length was the duration of the multimodal message that Instructors recorded. *Construction time* was measured from when the Builders received an instruction to when they pressed the "Finished" button. The results are shown in Figure 7B, along with Builders' *Time to first block* (from receiving an instruction to placing the first block) and *Time to last block* (from receiving the instruction to placing the last block) in each trial. **(H2)** We hypothesized that Instructors and Builders would become more efficient across repetitions. To test this, we used a linear mixed-effects model with one within-subject predictor (repetition) as a fixed effect and individual pairs as a random effect. There were significant effects of repetition on both instruction length ($\beta = -12.80, SE = 1.45, p < .001$), time to first block ($\beta = -2.99, SE = 1.10, p = .007$), time to last block ($\beta = -15.96, SE = 1.99, p < .001$), and construction time ($\beta = -20.19, SE = 2.73, p < .001$), indicating that instructions became shorter, and construction became faster.

5.2 Convention Formation

5.2.1 Linguistic abstractions. We analyzed changes in the total number of words per trial and found that Instructors used fewer words to provide more concise instructions over time (Figure 7C). We used a linear mixed-effects model with two within-subject predictors (repetition and tower shape) as fixed effects and individual pairs as a random effect to evaluate their impact on word count. The effect of repetition was significant ($\beta = -24.64, SE = 2.81, p < .001$). Moreover, instructions for *L* had significantly fewer words ($\beta = -22.70, SE = 7.70, p = .003$), while instructions for *TREE* had significantly more words ($\beta = 26.05, SE = 7.70, p = .001$). Although this gap narrowed across repetitions.

Based on the unimodal study results, **(H3)** we hypothesized that one explanation for this gain in linguistic efficiency is Instructors' utilization of abstractions that describe the entire tower shape with few words, which we refer to as "tower words," chunking a sequence of block-based instructions. We found evidence for the use of tower words across repetitions, including "C" (82 times) and "L" (29 times). The *TREE* tower was less commonly referred to as "Tree" (4 times), and more frequently as "T" (36) or "Cross" (21), likely due to the ambiguity of the mapping.[4] Instructors establish conventions during the very first repetition, with 86.7% of Instructors using at least one tower word in R1. To account for the reduction in message length from R1 to R4, we calculated the proportion of time spent describing block instructions and tower instructions (Figure 7D).

[3]The multimodal dataset (audio transcripts and 4D gestures) and the custom desktop viewing app are publicly available at: https://multimodal-conventions.github.io

[4]See Appendix A.1.3 for the list of all tower words.

We found that the proportion of tower instructions increased from 11.93% in R1 to 27.29% in R4. Using a permutation test (10,000 data shuffles) based on Euclidean distance (3D vector: block, tower, and other) for evaluating the difference between time proportions, we found this increase in proportion to be significant ($p = .015$).

We calculated the relative position of tower words (w) in messages (M) across all repetitions by dividing the index of w by the number of words in M (Figure 7E). We found that in R1, Instructors began by providing block instructions and then introduced tower words near the end of the trial. In later repetitions, they utilized these abstract tower words early in the trial. We used a linear mixed-effects model with one within-subject predictor (repetition) as a fixed effect and individual pairs as a random effect. We also included an orthogonalized quadratic effect of repetition to consider non-linearities. There were significant effects of repetition ($\beta = -0.44, SE = 0.096, p < .001$) and the orthogonalized quadratic term of repetition ($\beta = 0.071, SE = 0.019, p < .001$) on the relative positions of tower words. The results suggest that the position of tower words decreased, while the rate slowed across repetitions.

As Instructors shifted their strategy to abstract-first instructions, we also observed changes in Builders' construction strategy. In R1, most Builders (19 out of 20) placed blocks one by one on the grid, while in R4, 11 Builders constructed the entire tower, then placed it on the grid, reducing the gap between *time to last block* and *time to first block* in Figure 7B. This indicates that Builders also chunked their action execution, assembling towers during construction.

5.2.2 Gestural abstractions. Gestures play an important role in multimodal communication. Similar to speech, we found evidence of block-level gestural instructions, with Instructors manipulating imaginary block pieces (see example in Figure 1), pointing to indicate a specific block position, or using iconic gestures to describe the entire tower shape. Overall, there were 78 instances of tower gestures (41 static and 37 stroke-based), 36 of which were bimanual (Figure 8A). We calculated the proportion of time participants spent performing block gestures and tower gestures in R1 and R4 (Figure 8B); however, we found no significant change in the proportions across repetitions. This suggests that participants used tower gestures in R1 to establish a tower-shape convention. In later repetitions, while they continued to use tower gestures, they increasingly referenced tower shapes with words alone.

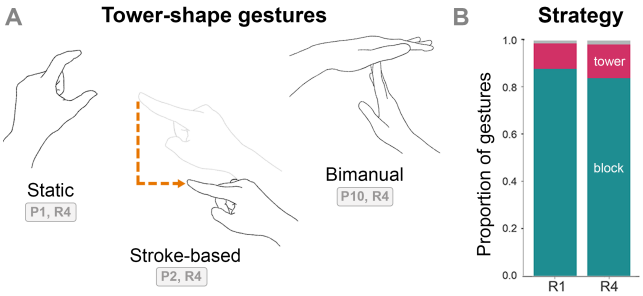

Figure 8: (A) Tower-shape gestures were static and stroke-based, with bimanual gestures across both. (B) Time proportion of block and tower gestures in first and last repetitions.

5.3 Multimodal Communication Dynamics

5.3.1 Informativeness of multimodal signals. We analyzed speech and gesture further to assess how much task-relevant information each modality encoded—what we refer to as the signals' "informativeness"—and how this changed over repeated interactions. We focused on shape, position, and orientation, the spatial information necessary for the task. For each gesture and utterance, we coded whether they contained information about these parameters at the block or tower abstraction level. For those that contained information, we examined the interplay between the two modalities [24]. We coded each multimodal reference as one of three categories: duplication of information across speech and gesture (redundant) to enhance comprehension [66], use of hand gestures to disambiguate unclear utterances (complementary), such as "like this" or "here" [58], or speech without gestures (language-only).

We calculated time proportions for each category (Figure 9). For segments containing task information, we normalized the proportions so that *redundant* + *complementary* + *language-only* = 1.0.

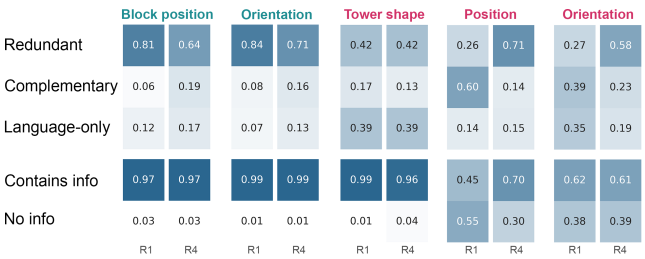

Figure 9: Mean time proportion of redundant, complementary, and language-only instructions for position and orientation information of blocks and shape, position, and orientation of towers in the first and last repetitions.

5.3.2 Shifts in block-level instructions. We found that in R1, signals were predominantly redundant for both position (81%) and orientation (84%). However, redundancy decreased by 17% for block position and 13% for block orientation from R1 to R4. This shift indicates that, over time, Instructors favored concise deictic phrasing (e.g., "this way") with co-speech gestures for block-level instructions, rather than lengthy, precise utterances [8] (Figure 10A).

5.3.3 Shifts in tower-level instructions. In R1, more than half of tower instructions lacked position information, and more than a third lacked orientation information. Instructors were likely establishing a shared convention for tower shapes after presenting detailed block-level instructions (Figure 10B). In R1, only 26% of tower position and 27% of tower orientation information were redundant. However, redundancy increased by 45% for tower position and 31% for tower orientation. In R4, half of the Instructors used multimodal redundancy, providing more information than necessary [46], to emphasize variations from the previous repetition, which in our study was the randomized orientation and position of towers. To analyze these trends, we ran permutation tests (10,000 data shuffles). We found a significant shift in the distribution of time proportions for tower position from R1 to R4 ($p = .0009$). While time proportions for other categories (block position, block

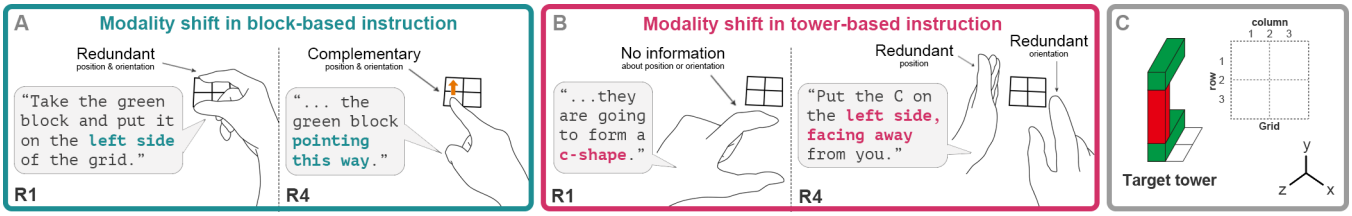

Figure 10: (A) Multimodal signals for position and orientation of blocks shift from redundant in R1 to complementary in R4. (B) For towers, no position or orientation information is provided when establishing an abstraction in R1, but redundancy is introduced in R4 to emphasize position and orientation changes. (C) Virtual target tower placed on the 2×2 grid.

orientation, and tower orientation) also changed from R1 to R4, these shifts in distribution were not significant (p = .085, p = .20, and p = .16, respectively).

When describing abstract tower shapes, gestures (redundant and complementary) typically followed tower words (Figure 11). Co-speech gestures were used near the end of R1 to establish conventions, and in R4 were spread across shorter messages. Once conventions were established, in some trials, participants used tower words early on without relying on gestures. The cross-modal distribution of tower-shape information did not change across repetitions.

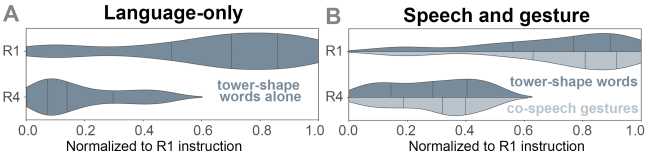

Figure 11: A: Relative position of tower words that appeared without gestures. B: Relative position of tower words and the corresponding redundant/complementary gestures.

5.3.4 User preferences and variability in modality shift. A more in-depth analysis of the shift in informativeness of block instructions revealed distinct participant groups that over time diverged in their modality preferences, consistent with prior work highlighting such individual differences [1]. For example, for block position, the first group, which we call the *Prefer H* group, gravitated toward shorter, ambiguous utterances accompanied by hand gestures (Figure 12A). This *Prefer H* group (P1, P2, P4, P8, P12, P17, P18, P19) was the largest subgroup, with 40% of Instructors shifting from redundant to complementary use of multimodal signals. The smaller *Prefer U* group (P7, P11, P14, P15) had the opposite preference and favored language-only instructions over time, eventually ceasing gesturing altogether. This reduced reliance on gestures may reflect gestural effort (e.g., arm fatigue) or growing familiarity (e.g., accumulated common ground on grid references and block orientation), making verbal instructions sufficient.

We saw a similar trend for block orientation, with the *Prefer H* group (P1, P2, P4, P5, P6, P11, P12, P19) being the largest subgroup at 40%. Unlike block position, however, 25% of the dyads shifted toward more redundancy when describing the block orientation in the final repetition; moving downward in Figure 12B. We hypothesize that this difference reflects differential gesturing costs: orientation can be conveyed with in-place hand rotations, whereas position often

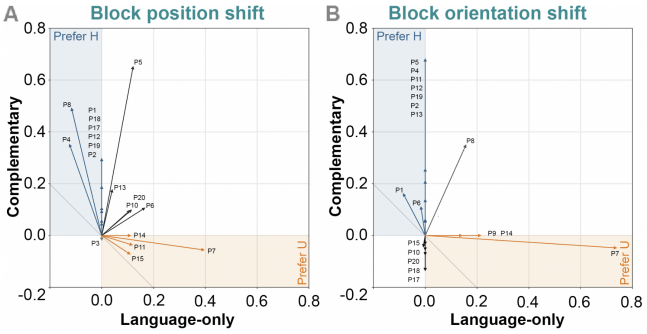

Figure 12: Shift in informativeness of block instructions from R1 to R4. Two participant groups diverged in their preferences: *Prefer U* group shifted toward language-only instructions over time, and *Prefer H* group shifted toward ambiguous language complemented with gestures.

requires reaching to the grid. This lower effort may explain the increase in redundancy observed for orientation, although this requires further investigation.

6 Computational Model

Toward developing convention-aware multimodal agents, we designed a unified computational model—a probabilistic (symbolic) system extending the Rational Speech Act (RSA) framework [31, 45]—to explain behavior in both studies. To allow for complementarity between modalities, we adapted previous unimodal approaches [26] to multimodal settings with both gesture and language. Our model allows agents to choose from redundant, complementary, or language-only messages and adjust their informativeness over time based on uncertainty about which communicative signals map to which abstractions. This enabled us to capture how participants adjusted their multimodal signals and explain the variability in how their communication preferences change over time.

6.1 Domain-Specific Language

We designed a domain-specific language (DSL) to describe agents' procedural knowledge about the assembly domain. This DSL $\mathcal{D}$ contains primitives for 3 colored blocks `R G B` and 9 positions $\mathsf{P}_{r,c}$ across 3 rows and 3 columns on the grid. It also includes sub-towers (`T_chunk`, `PL_chunk`) and towers (`C_chunk`, `L_chunk`, `TR_chunk`). For example, a C-shaped tower on the left can be expressed either as a sequence of primitives $\mathsf{G}\ \mathsf{P}_{2,1}\ \mathsf{R}\ \mathsf{P}_{3,1}\ \mathsf{G}\ \mathsf{P}_{2,1}$ or abstractly as $\mathsf{C_chunk}\ \mathsf{P}_{2,1}$.

6.2 Multimodal Lexicon

We assume each agent has a multimodal lexicon l that maps utterances and gestures to symbols in the DSL, and a belief state ($\mathcal{L}$) over their collaborator's lexicon. We also assume that agents initially agree on the mappings of primitive block and position symbols. Paired agents must then resolve two sources of uncertainty over time: abstraction mapping ambiguity and utterance ambiguity.

6.2.1 Abstraction mapping ambiguity. Abstractions improve efficiency despite introducing uncertainty [33]. To form ad hoc conventions, agents must coordinate on a mapping between utterances/gestures and the five chunked symbols at two levels (tower and sub-tower). This creates uncertainty over the 120 possible lexicons ($|\mathcal{L}| = 5!$), each initialized with prior probability $1/120$.

6.2.2 Utterance ambiguity. The second source of uncertainty arises from agents' ability to combine modalities to produce complementary (ambiguous language + gestures) instructions. Using complementary gestures to specify position shortens instructions, despite adding linguistic ambiguity [33]. To account for ambiguities in language that arise from using gestures in a multimodal setting, we define two possible utterance states: *clear* or *ambiguous*. For example, *"on the bottom half of the grid"* is *clear* and can be decoded as $\mathsf{P}_{3,2}$, while *"here"* is *ambiguous* and requires gestures to disambiguate the utterance. In this case, we assign a uniform probability of $1/9$ to each of the 9 possible positions on the grid.

6.3 Modality Preference in Instructions

For each trial, the Instructor chooses a multimodal message to communicate their intention to the Builder. In the traditional RSA framework, a pragmatic speaker S_1 acts according to the speaker's utility function U_s, defined as:

$$U_s(u;t) = \log P_{L_0}(t|u) - C(u) \tag{1}$$

where u is an utterance, t is a target state, L_0 is a literal listener, who infers the target state t given the literal meaning of u, and $C(u)$ is cost of u. The probability of generating u is then proportional to $e^{U_s(u;t)}$, i.e., softmax selection over the message space.

We extend this utility function to consider more than one modality, specifically hand gestures in addition to language, and adapt to sequential signals. Given a target tower in trial k, the Instructor considers candidate sequential tower programs $(T_1^k, T_2^k, ...)$ acquired up to that point. Each T_t^k consists of a set of steps $(t_{t,1}^k, t_{t,2}^k, ..., t_{t,|T_t^k|}^k)$ where $|T_t^k|$ is the number of steps (s) in the tower program T_t^k and $t_{t,i}^k \in \mathcal{D}$. Then, the Instructor agent chooses a multimodal message $M_j^k = (m_{j,1}^k, m_{j,2}^k, ..., m_{j,|T_j^k|}^k)$ considering the literal Builder B_0 and the following utility function:

$$\begin{aligned} U(M_j^k; T_j^k) =& \beta_i \sum_s \ln P_{B_0}(t_{j,s}^k | m_{j,s}^k) \\ &- \beta_u(r) \sum_s C_u(u_{j,s}) - \beta_h(r) \sum_s C_h(h_{j,s}) \end{aligned} \tag{2}$$

where β_i controls how much the Instructor considers the informativeness relative to the cost. Each sub-message $m_{j,s}^k$ corresponds to a step $t_{j,s}^k$ and consists of an utterance $u_{j,s}$ and a gesture $h_{j,s}$. The assumption in RSA is that prior probability (P_{B_0}) is proportional to the literal meaning $l(m_{j,s}^k, t_{j,s}^k) = l(u_{j,s}, t_{j,s}^k) \cdot l(h_{j,s}, t_{j,s}^k)$.

While the original RSA formulation assumes that $l(m_{j,s}^k, t_{j,s}^k)$ is a binary value, we relax this assumption by introducing two continuous semantics $x_u, x_h \in [0, 1]$ as proposed by Degen et al. [26], allowing generation of redundant messages.

$$l(u_{j,s}, t_{j,s}^k) = \begin{cases} x_u & \text{if } u \text{ is true of } t \\ 1 - x_u & \text{otherwise} \end{cases} \tag{3}$$

$$l(h_{j,s}, t_{j,s}^k) = \begin{cases} x_h & \text{if } h \text{ is true of } t \\ 1 - x_h & \text{otherwise} \end{cases} \tag{4}$$

In Equation 2, the second and third terms on the right-hand side are cost functions where $\beta_u(r)$ and $\beta_h(r)$ are parameters that control the effect of utterances or hand gestures. These two parameters change depending on the repetition r so that the model can simulate different combinations of utterances and gestures over time. High $\beta_u(r)$ leads to shorter, more ambiguous utterances, likely accompanied by gestures; and high $\beta_h(r)$ makes the agent prefer language-only instructions without gestures (i.e., $C_h(h_{j,s}) \to 0$).

After observing the message M_j^k, the pragmatic Builder B_1 assigns a symbol $t_{j,s}^k \in \mathcal{D}$ to $m_{j,s}^k$ with probability:

$$P_{B_1}(t_{j,s}^k | m_{j,s}^k) = \sum_{l \in \mathcal{L}} P_{I_1}(m_{j,s}^k | t_{j,s}^k, l) \cdot P_{I_1}(l) \tag{5}$$

where P_{I_1} refers to the Builder's belief about the pragmatic Instructor I_1 and is written with two terms using the utterance $u_{j,s}$ and the hand gesture $h_{j,s}$:

$$P_{I_1}(m_{j,s}^k | t_{j,s}^k, l) = \gamma \cdot P_{I_1}(u_{j,s} | t_{j,s}^k, l) + (1 - \gamma) \cdot P_{I_1}(h_{j,s} | t_{j,s}^k, l) \tag{6}$$

where γ determines how the model weighs utterances and gestures during inference. We assume that when a *clear* utterance is used in a multimodal setting, the meanings of the utterance and hand gesture are aligned, producing the same probability distribution across the two terms above.

6.4 Multimodal Abstractions

If the Builder agent can infer the correct tower program, the Instructor agent may propose an abstract tower program T^* with chunked symbols and send a new message M^*. Given that they have reached consensus, they update their prior beliefs following Bayes' rule:

$$\log P_{I_1}(l|M^*) = \log P_{I_1}(l) + \sum_{m^* \in M^*} \log P_{I_1}(m^*|t^*, l) \qquad \forall l \in \mathcal{L} \tag{7}$$

$$\log P_{B_1}(l|T^*) = \log P_{B_1}(l) + \sum_{t^* \in T^*} \log P_{B_1}(t^*|m^*, l) \qquad \forall l \in \mathcal{L} \tag{8}$$

where t^* and m^* are the corresponding symbol in $\mathcal{D}$ and sub-message, respectively. To correctly update beliefs over time, agents' $P(l)$ is normalized after each update, such that $\sum_{l \in \mathcal{L}} P(l) = 1$.

6.5 Simulations

6.5.1 Simulating abstractions. In the first simulation, we asked whether the proposed model can acquire abstract tower representations over repetitions. For this, we set $\beta_i = 0.3$ to prioritize minimizing cost over informativeness. To examine the effect of the cost of learning a new abstract program, we set $\beta_h = 0$ and varied β_u to 0.1, 0.5, and 1.0. The simulation setup was similar to the multimodal study, but tower positions and orientations were fixed across trials. The program length from 100 simulations was recorded and aggregated for each repetition: a length of 6 corresponds to no abstractions and 2 to the agent using a tower program and a position symbol. The results, shown in Figure 13, align with trends in both studies where program length decreased over time.

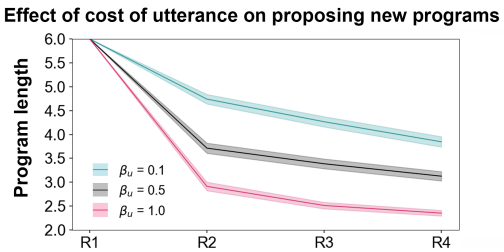

Figure 13: Simulation 1 results showing program length decreasing from R1 to R4 for β_u values of 0.1, 0.5, and 1.0.

6.5.2 Simulating modality preferences. We also asked whether the proposed model captures divergent modality preferences across participant groups. More specifically, we examined whether the model could explain changes in Instructors' preferences among redundant, language-only, and complementary instructions over time. As an example, we focused on block position information. In the multimodal study, we observed two distinct participant groups: the *Prefer U* group increasingly used language-only instructions from R1 to R4; the *Prefer H* group increasingly used complementary instructions.

Let p_r^{obs}, p_u^{obs}, and p_c^{obs} denote the empirical probabilities of redundant, language-only, and complementary instructions ($p_r^{obs} + p_u^{obs} + p_c^{obs} = 1$). In the model, the agent Instructor selects each type with probability p_r^{pred}, p_u^{pred}, and p_c^{pred}, proportional to e^{U_r}, e^{U_u}, and e^{U_c}, where utilities (U_r, U_u, and U_c) depend on parameter set $\theta = (\beta_i, \beta_u, \beta_h, x_u, x_h)$. We fit θ to match the predicted distribution ($p_r^{\text{pred}}, p_u^{\text{pred}}, p_c^{\text{pred}}$) to the observed distribution for three conditions: R1, *Prefer U* in R4, and *Prefer H* in R4. This yields parameter sets θ_{r1}, θ_{r4_u}, and θ_{r4_h}, which are then used for the simulation.

We fit the model parameters via Bayesian optimization to minimize the cross-entropy between predicted and observed distributions. The search ranges were: $\beta_u, \beta_h \in [0, 40]$ and $x_u, x_h \in [0.5, 1]$. We fixed β_i at 10, and held x_u and x_h constant at values found for θ_{r1} when estimating θ_{r4_u} and θ_{r4_h}, so we could examine how β_u and β_h contribute to the agent's behavioral change compared to β_i. We performed 200 optimization iterations with 40 random initializations.

Bayesian optimization yielded fitted parameters for the R1: $\beta_u = 20.25$, $\beta_h = 9.23$, $x_u = 0.87$, and $x_h = 0.62$. We held x_u and x_h constant when fitting R4 parameters. For the *Prefer U* group in R4, we obtained $\beta_u = 6.17$ and $\beta_h = 10.15$; for the *Prefer H* group, $\beta_u = 21.01$ and $\beta_h = 3.09$. These parameter shifts align with the observed behavioral changes. In *Prefer U*, the decrease in β_u and increase in β_h reflect reduced concern for utterance cost and increased sensitivity to gesture cost, leading to language-only instructions. Conversely, in *Prefer H*, the decrease in β_h reflects reduced sensitivity to gesture cost, leading to complementary instructions. We ran 200 simulations for each parameter set and computed the mean proportions of complementary and language-only messages per repetition. Figure 14 shows how the probability of the agent choosing language-only or complementary instructions changes from R1 to R4 in the *Prefer U* and *Prefer H* conditions. The resulting probabilities closely match the observed data from Study 2.

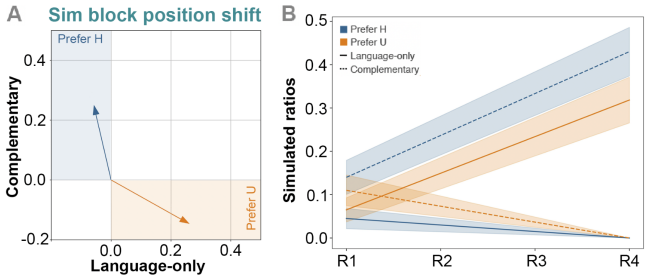

Figure 14: Simulation 2 results showing changes in language-only and complementary messages across repetitions.

7 Discussion

Effective collaboration in physical assembly tasks requires balancing communication cost and informativeness. In our iterated user study, we analyzed how Instructors' multimodal signals changed across repetitions to convey intent. Pairs became faster and more accurate over time as Instructors reduced words per trial by introducing tower-level abstractions. We found that Instructors formed linguistic and gestural conventions near the end of the first exchange without specifying the position or orientation of abstract towers. In later trials, they named the tower words early to speed up construction time, and then added redundant speech and gesture to mark changes in position and orientation of the towers.

While people generally adopted task abstractions, conventions varied widely across dyads. At times, conventions were arbitrary (e.g., "alligator" or "crocodile" for *C*; "J" for rotated *L*), and even three-block towers were chunked differently (e.g., "cross" vs. "T" for *TREE*). Modality strategies also diverged: some combined gestures with ambiguous language, while others minimized gesturing over time. To model these behaviors, we introduce a computational model that captures the formation of abstractions and balances informativeness with the costs of speech and gesture, and we show that it explains both the reduction in instruction length and participants' diverging modality preferences.

Based on our findings, convention-aware agents for assembly tasks should learn users' conventions for chunked instructions as they arise, over time default to abstract-first prompts when giving instructions, adapt modality to user preferences, and leverage redundancy to highlight changes from prior interactions.

7.1 Limitations and Future Work

7.1.1 Towards more complex physical structures. We focused on establishing an empirical foundation for multimodal convention formation using a limited set of block towers. However, people collaborate on far more diverse and complex physical structures in real-world settings. Future work should apply our approach to richer assembly domains (e.g., furniture assembly) where parts vary more in shape and in how they can be combined. This would reveal how well the principles identified here generalize to more complex tasks, such as those in EGGNOG [114] and HoloAssist [118].

7.1.2 From turn-based to real-time communication. We assigned consistent roles and controlled the turn-taking dynamics between dyads to draw precise links between the Instructor's communication and the Builder's building actions. However, in many real-world collaborative settings, the same individual might play multiple roles, sometimes giving instructions and other times following them. In addition, people collaborating are often able to communicate synchronously, which makes it possible for both communication and building actions to co-occur [123], such as when a collaborator chooses to build before receiving a "full" set of instructions, or when someone interrupts their partner to ask a clarification question. Future work should investigate these more naturalistic settings where multimodal communication can be synchronous.

Prior work shows that immediate feedback enhances conversational grounding [38, 74], whereas delayed feedback attenuates language adaptation [72]. Accordingly, the changes we observed in our turn-based protocol may be less dynamic than in-person collaboration. Although our study was turn-based, our model is flexible and can be extended to synchronous settings by adopting finer-grained message units (e.g., word- or gesture-stroke level) and using contrastive inference to infer segment boundaries [27, 75]. The framework can also incorporate lightweight, non-action feedback from Builders (e.g., uncertainty or confirmation signals) to influence instruction generation in real time. Evaluating these extensions is an important direction for future work.

7.1.3 Extensions of the computational model. Our model extends the RSA framework to generate and interpret abstract multimodal instructions and capture preferences in speech and gesture use. Several constants are currently manually set, but could be estimated online. For example, γ can be derived from eye-tracking to determine whether Builders attend to gestures [47, 48]. Additionally, modality costs likely depend on task environment (e.g., grid size) and communication duration (e.g., arm fatigue), requiring further investigation.

We also made simplifying assumptions, including perfect semantic and temporal alignment between gestures and utterances [120] and a uniform prior over lexicons. Misalignment can make Instructors appear "irrational," and cause pragmatic Builders to misinterpret intentions [66]; real-world agents would need to detect such misalignment using emerging conventions or accumulated common ground. Although some chunked instructions in our task are visually grounded (e.g., saying "T" while making a T-shaped gesture), we used a uniform prior to probe behavior in extreme cases.

Our model considers 120 possible lexicons; scaling to more complex tasks will expand the set of possible abstractions and make belief updates more costly. Larger vocabularies and multimodal symbol spaces may require language-model approximations to pragmatic inference [25, 106]. The model also does not explain how people *acquire* task abstractions or gestural conventions. Work on decomposing towers [90, 109] and learning abstractions for complex artifacts [63, 71] points toward ways to learn linguistic abstractions, but learning *gestural* abstractions remains an open challenge.

8 Conclusion

We examined how conventions emerge in collaborative assembly across two settings. An online unimodal (text-only) study showed dyads shifting from block-by-block descriptions to concise tower-level abstractions, reducing instruction length over repetitions. A follow-up AR study in a physically grounded setting isolated speech and gesture: Instructors introduced tower-level abstractions near the end of the first exchange, then shifted to abstract-first instructions in later trials, using redundancy to mark changes in position and orientation. Pairs became faster and more accurate over time. To explain these behaviors, we introduced a computational model that captures abstraction formation and balances informativeness with the production costs of speech and gesture. Simulations reproduced shorter instructions and diverging modality preferences. Together, these results inform convention-aware agents for physical assembly tasks that learn users' multimodal abstractions as they emerge, default to abstract-first instructions over time, add redundancy when conventions change or uncertainty is high, and personalize modality choices while adapting to changes in user preferences.

Acknowledgments

We thank the members of the Princeton HCI Group and the Cognitive Tools Lab at Stanford for fruitful discussions. Support for this work came from NSF CAREER #2436199, an ONR Science of Autonomy Award, and a Hoffman-Yee Grant from the Stanford Center for Human-Centered AI awarded to J.E.F. Partial support was provided by Toyota RIKEN Fellowship to K.M.

References

[1] Olga Abramov, Friederike Kern, Sofia Koutalidis, Ulrich Mertens, Katharina Rohlfing, and Stefan Kopp. 2021. The relation between cognitive abilities and the distribution of semantic features across speech and gesture in 4-year-olds. *Cognitive Science* 45, 7 (2021).

[2] Martha W. Alibali, Miriam Bassok, Karen Olseth Solomon, Sharon E. Syc, and Susan Goldin-Meadow. 1999. Illuminating mental representations through speech and gesture. *Psychological Science* 10, 4 (1999), 327–333.

[3] Judith Amores, Xavier Benavides, and Pattie Maes. 2015. ShowMe: A Remote Collaboration System that Supports Immersive Gestural Communication. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, Seoul, CHI 2015 Extended Abstracts, Republic of Korea, April 18 - 23, 2015.* ACM, New York, NY, 1343–1348. https://doi.org/10.1145/2702613.2732927

[4] Eva Ansermin, Ghilès Mostafaoui, Nils Beaussé, and Philippe Gaussier. 2016. Learning to Synchronously Imitate Gestures Using Entrainment Effect. In *From Animals to Animats 14 - 14th International Conference on Simulation of Adaptive Behavior, SAB 2016, Aberystwyth, UK, August 23-26, 2016, Proceedings (Lecture Notes in Computer Science, Vol. 9825).* Springer, 219–231. https://doi.org/10.1007/978-3-319-43488-9_20

[5] Richard N. Aslin, Jenny R. Saffran, and Elissa L. Newport. 1998. Computation of conditional probability statistics by 8-month-old infants. *Psychological Science* 9, 4 (1998), 321–324.

[6] Joseph L. Austerweil and Thomas L. Griffiths. 2013. A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review* 120, 4 (2013), 817.

[7] Victor Bapst, Alvaro Sanchez-Gonzalez, Carl Doersch, Kimberly Stachenfeld, Pushmeet Kohli, Peter Battaglia, and Jessica Hamrick. 2019. Structured agents for

physical construction. In *International Conference on Machine Learning*. PMLR, 464–474.

[8] Martin Bauer, Gerd Kortuem, and Zary Segall. 1999. "Where Are You Pointing At?" A Study of Remote Collaboration in a Wearable Videoconference System. In *Third International Symposium on Wearable Computers (ISWC 1999), San Francisco, California, USA, 18-19 October 1999, Proceedings*. IEEE Computer Society, New York, USA, 151–158. https://doi.org/10.1109/ISWC.1999.806696

[9] Sian L. Beilock and Susan Goldin-Meadow. 2010. Gesture changes thought by grounding it in action. *Psychological science* 21, 11 (2010), 1605–1610.

[10] Mathilde M. Bekker, Judith S. Olson, and Gary M. Olson. 1995. Analysis of Gestures in Face-to-Face Design Teams Provides Guidance for How to Use Groupware in Design. In *Proceedings of the 1st Conference on Designing Interactive Systems: Processes, Practices, Methods and Techniques, DIS '95, Ann Arbor, MI, USA, August 23-25, 1995*. ACM, New York, NY, 157–166. https://doi.org/10.1145/225434.225452

[11] Hrvoje Benko, Ricardo Jota, and Andrew Wilson. 2012. MirageTable: freehand interaction on a projected augmented reality tabletop. In *CHI Conference on Human Factors in Computing Systems, CHI '12, Austin, TX, USA - May 05 - 10, 2012*. ACM, New York, NY, 199–208. https://doi.org/10.1145/2207676.2207704

[12] Richard A. Bolt. 1980. "Put-that-there": Voice and gesture at the graphics interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1980, Seattle, Washington, USA, July 14-18, 1980*. ACM, New York, NY, 262–270. https://doi.org/10.1145/800250.807503

[13] Veronica Boyce and Michael C. Frank. 2023. Communicative reduction in referring expressions within a multi-player negotiation game. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 45. Wiley-Blackwell, USA.

[14] Veronica Boyce, Robert D. Hawkins, Noah D. Goodman, and Michael C. Frank. 2024. Interaction structure constrains the emergence of conventions in group communication. *Proceedings of the National Academy of Sciences* 121, 28 (2024), e2403888121. https://doi.org/10.1073/pnas.2403888121

[15] Neil R. Bramley and Fei Xu. 2023. Active inductive inference in children and adults: A constructivist perspective. *Cognition* 238 (2023), 105471.

[16] Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition* 22, 6 (1996), 1482.

[17] Richard Brutti, Lucia Donatelli, Kenneth Lai, and James Pustejovsky. 2022. Abstract meaning representation for gesture. In *Proceedings of the thirteenth language resources and evaluation conference*. https://par.nsf.gov/servlets/purl/10409402

[18] Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. 2021. YouRefIt: Embodied Reference Understanding with Language and Gesture. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, New York, NY, 1365–1375. https://doi.org/10.1109/ICCV48922.2021.00142

[19] Morten H. Christiansen and Nick Chater. 2016. The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences* 39 (2016).

[20] Mingyuan Chu and Sotaro Kita. 2008. Spontaneous gestures during mental rotation tasks: insights into the microdevelopment of the motor strategy. *Journal of Experimental Psychology: General* 137, 4 (2008), 706.

[21] Mingyuan Chu and Sotaro Kita. 2011. The nature of gestures' beneficial role in spatial problem solving. *Journal of experimental psychology: General* 140, 1 (2011), 102.

[22] Herbert H. Clark. 1996. *Using language*. Cambridge University Press.

[23] Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22, 1 (1986), 1–39.

[24] Jan Peter De Ruiter. 2006. Can gesticulation help aphasic people speak, or rather, communicate? *Advances in Speech Language Pathology* 8, 2 (2006), 124–127.

[25] Judith Degen. 2023. The rational speech act framework. *Annual Review of Linguistics* 9, 1 (2023), 519–540.

[26] Judith Degen, Robert D. Hawkins, Caroline Graf, Elisa Kreiss, and Noah D. Goodman. 2020. When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review* 127, 4 (2020), 591.

[27] Judith Degen, Leyla Kursat, and Daisy Dorothy Leigh. 2021. Seeing is believing: testing an explicit linking assumption for visual world eye-tracking in psycholinguistics. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society, CogSci 2021, virtual, July 26-29, 2021*.

[28] Tobias Ende, Sami Haddadin, Sven Parusel, Tilo Wüsthoff, Marc Hassenzahl, and Alin Albu-Schäffer. 2011. A human-centered approach to robot gesture based communication within collaborative working processes. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2011, San Francisco, CA, USA, September 25-30, 2011*. IEEE, USA, 3367–3374. https://doi.org/10.1109/IROS.2011.6094592

[29] Nicolas Fay, Simon Garrod, Leo Roberts, and Nik Swoboda. 2010. The interactive evolution of human communication systems. *Cognitive Science* 34, 3 (2010), 351–386.

[30] Nicolas Fay, Casey J. Lister, T. Mark Ellison, and Susan Goldin-Meadow. 2014. Creating a communication system from scratch: gesture beats vocalization hands down. *Frontiers in Psychology* 5 (2014), 354.

[31] Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336, 6084 (2012), 998–998.

[32] Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2022. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. *arXiv preprint arXiv:2211.08371* (2022).

[33] Susan R. Fussell, Robert E. Kraut, and Jane Siegel. 2000. Coordination of communication: effects of shared visual context on collaborative work. In *CSCW 2000, Proceeding on the ACM 2000 Conference on Computer Supported Cooperative Work, Philadelphia, PA, USA, December 2-6, 2000*. ACM, New York, NY, 21–30.

[34] Susan R. Fussell, Leslie D. Setlock, Jie Yang, Jiazhi Ou, Elizabeth Mauer, and Adam D. I. Kramer. 2004. Gestures Over Video Streams to Support Remote Collaboration on Physical Tasks. *Hum. Comput. Interact.* 19, 3 (2004), 273–309. https://doi.org/10.1207/S15327051HCI1903_3

[35] Simon Garrod and Gwyneth Doherty. 1994. Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition* 53, 3 (1994), 181–215.

[36] Darren Gergle, Robert E. Kraut, and Susan R. Fussell. 2004. Action as language in a shared visual space. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work, CSCW 2004, Chicago, Illinois, USA, November 6-10, 2004*. ACM, New York, NY, 487–496. https://doi.org/10.1145/1031607.1031687

[37] Darren Gergle, Robert E. Kraut, and Susan R. Fussell. 2004. Language efficiency and visual technology: Minimizing collaborative effort with visual information. *Journal of language and social psychology* 23, 4 (2004), 491–517.

[38] Darren Gergle, Robert E. Kraut, and Susan R. Fussell. 2006. The impact of delayed visual feedback on collaborative performance. In *Proceedings of the 2006 Conference on Human Factors in Computing Systems, CHI 2006, Montréal, Québec, Canada, April 22-27, 2006*. ACM, New York, NY, 1303–1312. https://doi.org/10.1145/1124772.1124968

[39] Michael J. Gielniak and Andrea Lockerd Thomaz. 2011. Generating anticipation in robot motion. In *20th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2011, Atlanta, Georgia, USA, July 31 - August 3, 2011*. IEEE, USA, 449–454. https://doi.org/10.1109/ROMAN.2011.6005255

[40] Brian T. Gleeson, Karon E. MacLean, Amir Haddadi, Elizabeth A. Croft, and Javier Adolfo Alcazar. 2013. Gestures for industry: intuitive human-robot communication from human observation. In *ACM/IEEE International Conference on Human-Robot Interaction, HRI 2013, Tokyo, Japan, March 3-6, 2013*. IEEE/ACM, USA, 349–356. http://dl.acm.org/citation.cfm?id=2447679

[41] Susan Goldin-Meadow. 2005. *Hearing gesture: How our hands help us think*. Harvard University Press.

[42] Susan Goldin-Meadow. 2006. Talking and thinking with our hands. *Current directions in psychological science* 15, 1 (2006), 34–39.

[43] Susan Goldin-Meadow and Susan M Wagner. 2005. How our hands help us learn. *Trends in cognitive sciences* 9, 5 (2005), 234–241.

[44] Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences* 20, 11 (2016), 818–829.

[45] Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences* 20, 11 (2016), 818 – 829.

[46] Herbert P. Grice. 1975. *Logic and Conversation*. Brill, Leiden, The Netherlands, 41 – 58. https://doi.org/10.1163/9789004368811_003

[47] Marianne Gullberg and Kenneth Holmqvist. 2006. What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. *Pragmatics & Cognition* 14, 1 (2006), 53–82.

[48] Marianne Gullberg and Sotaro Kita. 2009. Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of nonverbal behavior* 33, 4 (2009), 251–277.

[49] Robert D. Hawkins, Michael Franke, Michael C. Frank, Adele E. Goldberg, Kenny Smith, Thomas L. Griffiths, and Noah D. Goodman. 2023. From partners to populations: A hierarchical Bayesian account of coordination and convention. *Psychological Review* 130, 4 (2023), 977.

[50] Robert D. Hawkins, Minae Kwon, Dorsa Sadigh, and Noah D. Goodman. 2020. Continual Adaptation for Efficient Machine Communication. In *Proceedings of the 24th Conference on Computational Natural Language Learning, CoNLL 2020, Online, November 19-20, 2020*. Association for Computational Linguistics, USA, 408–419. https://doi.org/10.18653/V1/2020.CONLL-1.33

[51] Robert D. Hawkins, Megumi Sano, Noah D. Goodman, and Judith E. Fan. 2023. Visual resemblance and interaction history jointly constrain pictorial meaning. *Nature Communications* 14, 1 (2023), 2199.

[52] Henning Holle and Thomas C. Gunter. 2007. The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of cognitive neuroscience* 19, 7 (2007), 1175–1192. https://doi.org/10.1162/jocn.2007.19.7.1175

[53] Xiyun Hu, Dizhi Ma, Fengming He, Zhengzhe Zhu, Shao-Kang Hsia, Chenfei Zhu, Ziyi Liu, and Karthik Ramani. 2025. GesPrompt: Leveraging Co-Speech Gestures to Augment LLM-Based Interaction in Virtual Reality. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. 59–80.

[54] Yilun Hua and Yoav Artzi. 2024. Talk Less, Interact Better: Evaluating In-context Conversational Adaptation in Multimodal LLMs. In *First Conference on Language Modeling*. https://openreview.net/forum?id=lVOw78nYXS

[55] Yilun Hua, Evan Wang, and Yoav Artzi. 2025. Post-training for Efficient Communication via Convention Formation. In *Second Conference on Language Modeling*. https://openreview.net/forum?id=jRGGmbhX2s

[56] Takamasa Iio, Masahiro Shiomi, Kazuhiko Shinozawa, Takahiro Miyashita, Takaaki Akimoto, and Norihiro Hagita. 2009. Lexical entrainment in human-robot interaction: Can robots entrain human vocabulary?. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 11-15, 2009, St. Louis, MO, USA*. IEEE, 3727–3734. https://doi.org/10.1109/IROS.2009.5354149

[57] Andrew Irlitti, Mesut Latifoglu, Qiushi Zhou, Martin N. Reinoso, Thuong N. Hoang, Eduardo Velloso, and Frank Vetere. 2023. Volumetric Mixed Reality Telepresence for Real-time Cross Modality Collaboration. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*. ACM, New York, NY, 101:1–101:14. https://doi.org/10.1145/3544548.3581277

[58] Jana M. Iverson and Susan Goldin-Meadow. 2005. Gesture Paves the Way for Language Development. *Psychological Science* 16, 5 (2005), 367–371. https://doi.org/10.1111/j.0956-7976.2005.01542.x PMID: 15869695.

[59] Philip L. Jackson, Andrew N. Meltzoff, and Jean Decety. 2006. Neural circuits involved in imitation and perspective-taking. *Neuroimage* 31, 1 (2006), 429–439.

[60] Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. 2020. Learning to execute instructions in a Minecraft dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Association for Computational Linguistics, USA, 2589–2602. https://doi.org/10.18653/V1/2020.ACL-MAIN.232

[61] Guangyuan Jiang, Manjie Xu, Shiji Xin, Wei Liang, Yujia Peng, Chi Zhang, and Yixin Zhu. 2023. MEWL: Few-shot multimodal word learning with referential uncertainty. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, New York, NY, 15144–15169. https://proceedings.mlr.press/v202/jiang23i.html

[62] Janet G. Johnson, Tommy Sharkey, Iramuali Cynthia Butarbutar, Danica Xiong, Ruijie Huang, Lauren Sy, and Nadir Weibel. 2023. UnMapped: Leveraging Experts' Situated Experiences to Ease Remote Guidance in Collaborative Mixed Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*. ACM, New York, NY, 878:1–878:20. https://doi.org/10.1145/3544548.3581444

[63] R. Kenny Jones, Paul Guerrero, Niloy J. Mitra, and Daniel Ritchie. 2023. ShapeCoder: Discovering Abstractions for Visual Programs from Unstructured Primitives. *ACM Trans. Graph.* 42, 4 (2023). https://doi.org/10.1145/3592416

[64] Seokmin Kang and Barbara Tversky. 2016. From hands to minds: Gestures promote understanding. *Cognitive Research: Principles and Implications* 1 (2016), 1–15.

[65] Jakob Karolus, Johannes Sylupp, Albrecht Schmidt, and Pawel W. Wozniak. 2023. EyePiano: Leveraging Gaze For Reflective Piano Learning. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference, DIS 2023, Pittsburgh, PA, USA, July 10-14, 2023*. ACM, New York, NY, 1209–1223. https://doi.org/10.1145/3563657.3596065

[66] Spencer D. Kelly, Aslı Özyürek, and Eric Maris. 2010. Two Sides of the Same Coin: Speech and Gesture Mutually Interact to Enhance Comprehension. *Psychological Science* 21, 2 (2010), 260–267. https://doi.org/10.1177/0956797609357327 PMID: 20424055.

[67] Angela M. Kessell and Barbara Tversky. 2006. Using gestures and diagrams to think and talk about insight problems. In *Proceedings of the Meetings of the Cognitive Science Society*.

[68] Mitsuhiko Kimoto, Takamasa Iio, Masahiro Shiomi, Ivan Tanev, Katsunori Shimohara, and Norihiro Hagita. 2016. Alignment Approach Comparison between Implicit and Explicit Suggestions in Object Reference Conversations. In *Proceedings of the Fourth International Conference on Human Agent Interaction, HAI 2016, Biopolis, Singapore, October 4-7, 2016*. ACM, New York, NY, 193–200. https://doi.org/10.1145/2974804.2974814

[69] Sotaro Kita, Martha W. Alibali, and Mingyuan Chu. 2017. How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological review* 124, 3 (2017), 245.

[70] Bjorn B. de Koning, Katrina Mok, Nadine Marcus, and Paul Ayres. 2023. Investigating the role of hand perspective in learning from procedural animations. *British Journal of Educational Psychology* 93 (2023), 251–269.

[71] Juil Koo, Ian Huang, Panos Achlioptas, Leonidas J. Guibas, and Minhyuk Sung. 2022. PartGlot: Learning Shape Part Segmentation From Language Reference Games. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, USA, 16505–16514.

[72] Robert M. Krauss and Chi-Yue Chiu. 1998. Language and social behavior. (1998).

[73] Robert M. Krauss and Sidney Weinheimer. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science* 1, 1-12 (1964), 113–114.

[74] Robert E. Kraut, Steven H. Lewis, and Lawrence W. Swezey. 1982. Listener responsiveness and the coordination of conversation. *Journal of personality and social psychology* 43, 4 (1982), 718.

[75] Elisa Kreiss and Judith Degen. 2020. Production expectations modulate contrastive inference. In *Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual, July 29 - August 1, 2020*.

[76] Nikhil Krishnaswamy, Pradyumna Narayana, Rahul Bangar, Kyeongmin Rim, Dhruva Patil, David McNeely-White, Jaime Ruiz, Bruce Draper, Ross Beveridge, and James Pustejovsky. 2020. Diana's World: A Situated Multimodal Interactive Agent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13618–13619.

[77] Nikhil Krishnaswamy, Pradyumna Narayana, Isaac Wang, Kyeongmin Rim, Rahul Bangar, Dhruva Patil, Gururaj Mulay, Ross Beveridge, Jaime Ruiz, Bruce Draper, and James Pustejovsky. 2017. Communicating and acting: Understanding gesture in simulation semantics. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)—Short papers*.

[78] Nikhil Krishnaswamy, William Pickard, Brittany Cates, Nathaniel Blanchard, and James Pustejovsky. 2022. The voxworld platform for multimodal embodied agents. In *Proceedings of the thirteenth language resources and evaluation conference*.

[79] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George W. Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 2019. Loki: Facilitating Remote Instruction of Physical Tasks Using Bi-Directional Mixed-Reality Telepresence. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology, UIST 2019, New Orleans, LA, USA, October 20-23, 2019*. ACM, New York, NY, 161–174. https://doi.org/10.1145/3332165.3347872

[80] Angeliki Lazaridou and Marco Baroni. 2020. Emergent Multi-Agent Communication in the Deep Learning Era. *CoRR* abs/2006.02419 (2020). arXiv:2006.02419 https://arxiv.org/abs/2006.02419

[81] Gun A. Lee, Theophilus Teo, Seungwon Kim, and Mark Billinghurst. 2018. A User Study on MR Remote Collaboration Using Live 360 Video. In *IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2018, Munich, Germany, October 16-20, 2018*. IEEE, 153–164. https://doi.org/10.1109/ISMAR.2018.00051

[82] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S. Rodriguez, and Jon E. Froehlich. 2024. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*. ACM, USA, 408:1–408:20. https://doi.org/10.1145/3613904.3642230

[83] David Lewis. 2008. *Convention: A philosophical study*. John Wiley & Sons, USA.

[84] Li-Heng Lin, Yuchen Cui, Yilun Hao, Fei Xia, and Dorsa Sadigh. 2023. Gesture-Informed Robot Assistance via Foundation Models. In *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA (Proceedings of Machine Learning Research, Vol. 229)*. PMLR, USA, 3061–3082. https://proceedings.mlr.press/v229/lin23a.html

[85] Yung-Ching Liu. 2001. Comparative study of the effects of auditory, visual and multimodality displays on drivers' performance in advanced traveller information systems. *Ergonomics* 44, 4 (2001), 425–442.

[86] Kiyosu Maeda, Ching-Yi Tsai, Judith E. Fan, and Parastoo Abtahi. 2025. Using Gesture and Language to Establish Multimodal Conventions in Collaborative Physical Tasks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

[87] Ilias El Makrini, Kelly Merckaert, Dirk Lefeber, and Bram Vanderborght. 2017. Design of a collaborative architecture for human-robot assembly tasks. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017*. IEEE, USA, 1624–1629. https://doi.org/10.1109/IROS.2017.8205971

[88] William P. McCarthy, Robert D. Hawkins, Haoliang Wang, Cameron Holdaway, and Judith E. Fan. 2021. Learning to communicate about shared procedural abstractions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

[89] William P. McCarthy, David Kirsh, and Judith E. Fan. 2020. Learning to build physical structures better over time. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 42.

[90] William P. McCarthy, David Kirsh, and Judith E. Fan. 2023. Consistency and Variation in Reasoning About Physical Assembly. *Cognitive Science* 47, 12 (2023), e13397. https://doi.org/10.1111/cogs.13397

[91] William P. McCarthy, Justin Matejka, Karl D. Willis, Judith E. Fan, and Yewen Pu. 2024. Communicating Design Intent Using Drawing and Text. In *Proceedings of the 16th Conference on Creativity & Cognition* (Chicago, IL, USA) *(C&C '24)*. Association for Computing Machinery, New York, NY, USA, 512–519. https://doi.org/10.1145/3635636.3664261

[92] David McNeill. 1992. *Hand and Mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

[93] Pradyumna Narayana, Nikhil Krishnaswamy, Isaac Wang, Rahul Bangar, Dhruva Patil, Gururaj Mulay, Kyeongmin Rim, Ross Beveridge, Jaime Ruiz, James Pustejovsky, and Bruce Draper. 2018. Cooperating with avatars through gesture, language and action. In *Proceedings of SAI Intelligent Systems Conference*. Springer.

[94] Miriam A. Novack, Eliza L. Congdon, Naureen Hemani-Lopez, and Susan Goldin-Meadow. 2014. From action to abstraction: Using the hands to learn math. *Psychological science* 25, 4 (2014), 903–910.

[95] Tetsuo Ono, Michita Imai, and Hiroshi Ishiguro. 2001. A model of embodied communications with gestures between human and robots. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 23.

[96] Mridumoni Phukon and Abhishek Shrivastava. 2023. Effect of Speech Entrainment in Human-Computer Conversation: A Review. In *International Conference on Intelligent Human Computer Interaction*. Springer, 32–43.

[97] Wim Pouw, Mark Dingemanse, Yasamin Motamedi, and Aslı Özyürek. 2021. A systematic investigation of gesture kinematics in evolving manual languages in the lab. *Cognitive science* 45, 7 (2021), e13014.

[98] James Pustejovsky, Nikhil Krishnaswamy, Bruce Draper, Pradyumna Narayana, and Rahul Bangar. 2017. Creating common ground through multimodal simulations. In *Proceedings of the IWCS workshop on Foundations of Situated and Multimodal Communication*.

[99] Shuwen Qiu, Sirui Xie, Lifeng Fan, Tao Gao, Jungseock Joo, Song-Chun Zhu, and Yixin Zhu. 2022. Emergent Graphical Conventions in a Visual Communication Game. In *Advances in Neural Information Processing Systems*.

[100] Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E. McCullough, and Rashid Ansari. 2002. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)* 9, 3 (2002), 171–193.

[101] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 28492–28518.

[102] Nicole Robinson, Brendan Tidd, Dylan Campbell, Dana Kulić, and Peter Corke. 2023. Robotic vision for human-robot interaction and collaboration: A survey and systematic review. *ACM Transactions on Human-Robot Interaction* 12, 1 (2023), 1–66. https://doi.org/10.48550/ARXIV.2307.15363

[103] Mose Sakashita, Balasaravanan Thoravi Kumaravel, Nicolai Marquardt, and Andrew D. Wilson. 2024. SharedNeRF: Leveraging Photorealistic and View-dependent Rendering for Real-time and Remote Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*. ACM, New York, NY, 675:1–675:14. https://doi.org/10.1145/3613904.3642945

[104] Wendy Sandler, Irit Meir, Carol Padden, and Mark Aronoff. 2005. The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences* 102, 7 (2005), 2661–2665.

[105] Nazmus Saquib, Rubaiat Habib Kazi, Li-Yi Wei, and Wilmot Li. 2019. Interactive Body-Driven Graphics for Augmented Video Performance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*. ACM, New York, NY, 622. https://doi.org/10.1145/3290605.3300852

[106] Sebastian Schuster, Ayesha Ansar, Om Agarwal, and Vera Demberg. 2024. SpreadNaLa: A Naturalistic Code Generation Evaluation Dataset of Spreadsheet Formulas. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*. ELRA and ICCL, 15216–15225.

[107] Daniel L. Schwartz and John B. Black. 1996. Shuttling between depictive models and abstract rules: Induction and fallback. *Cognitive science* 20, 4 (1996), 457–497.

[108] Rajeev Sharma, Vladimir I. Pavlovic, and Thomas S. Huang. 1998. Toward multimodal human-computer interface. *Proc. IEEE* 86, 5 (1998), 853–869. https://doi.org/10.1109/5.664275

[109] Amy Lynne Shelton, E. Emory Davis, Cathryn S. Cortesa, Jonathan D. Jones, Gregory D. Hager, Sanjeev Khudanpur, and Barbara Landau. 2022. Characterizing the Details of Spatial Construction: Cognitive Constraints and Variability. *Cogn. Sci.* 46, 1 (2022). https://doi.org/10.1111/COGS.13081

[110] Andy Shih, Arjun Sawhney, Jovana Kondic, Stefano Ermon, and Dorsa Sadigh. 2021. On the Critical Role of Conventions in Adaptive Human-AI Collaboration. In *International Conference on Learning Representations*.

[111] Ian Stewart, Diyi Yang, and Jacob Eisenstein. 2020. Characterizing Collective Attention via Descriptor Context: A Case Study of Public Discussions of Crisis Events. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*. AAAI Press, 650–660. https://ojs.aaai.org/index.php/ICWSM/article/view/7331

[112] Darja Stoeva, Andreas Kriegler, and Margrit Gelautz. 2024. Body Movement Mirroring and Synchrony in Human–Robot Interaction. *ACM Transactions on Human-Robot Interaction* 13, 4 (2024), 1–26.

[113] Aaron Walsman, Muru Zhang, Klemen Kotar, Karthik Desingh, Ali Farhadi, and Dieter Fox. 2022. Break and make: Interactive structural understanding using lego bricks. In *European Conference on Computer Vision*. Springer, 90–107.

[114] Isaac Wang, Mohtadi Ben Fraj, Pradyumna Narayana, Dhruva Patil, and Gururaj Mulay. 2017. EGGNOG: A Continuous, Multi-modal Data Set of Naturally Occurring Gestures with Ground Truth Labels. In *12th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2017, Washington, DC, USA, May 30 - June 3, 2017*. IEEE Computer Society, 414–421. https://doi.org/10.1109/FG.2017.145

[115] Isaac Wang, Pradyumna Narayana, Dhruva Patil, Rahul Bangar, Bruce Draper, Ross Beveridge, and Jaime Ruiz. 2021. It's a Joint Effort: Understanding Speech and Gesture in Collaborative Tasks. In *International Conference on Human-Computer Interaction*. Springer, 159–178.

[116] Isaac Wang, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, Bruce Draper, Ross Beveridge, and Jaime Ruiz. 2017. Exploring the use of gesture in collaborative tasks. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 2990–2997.

[117] Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. Learning Language Games through Interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics. https://doi.org/10.18653/V1/P16-1224

[118] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. 2023. HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 20213–20224. https://doi.org/10.1109/ICCV51070.2023.01854

[119] Deanna Wilkes-Gibbs and Herbert H Clark. 1992. Coordinating beliefs in conversation. *Journal of memory and language* 31, 2 (1992), 183–194.

[120] Adam S. Williams and Francisco Raul Ortega. 2020. Understanding Gesture and Speech Multimodal Interactions for Manipulation Tasks in Augmented Reality Using Unconstrained Elicitation. *Proc. ACM Hum. Comput. Interact.* 4, ISS (2020), 202:1–202:21. https://doi.org/10.1145/3427330

[121] Jackie (Junrui) Yang, Leping Qiu, Emmanuel Angel Corona-Moreno, Louisa Shi, Hung Bui, Monica S. Lam, and James A. Landay. 2024. AMMA: Adaptive Multimodal Assistants Through Automated State Tracking and User Model-Directed Guidance Planning. In *IEEE Conference Virtual Reality and 3D User Interfaces, VR 2024, Orlando, FL, USA, March 16-21, 2024*. IEEE, 892–902. https://doi.org/10.1109/VR58804.2024.00108

[122] Jacob Young, Tobias Langlotz, Matthew Cook, Steven Mills, and Holger Regenbrecht. 2019. Immersive Telepresence and Remote Collaboration using Mobile and Wearable Devices. *IEEE Trans. Vis. Comput. Graph.* 25, 5 (2019), 1908–1918. https://doi.org/10.1109/TVCG.2019.2898737

[123] Chen Zheng and Barbara Tversky. 2024. Putting it together, together. *Cognitive Science* 48, 2 (2024), e13405. https://doi.org/10.1111/cogs.13405

A Appendix

A.1 Physical Multimodal Study

A.1.1 Participants. Demographics information on Instructors and Builders in the multimodal study.

Table 1: Participant demographics.

	Instructors	Builders
Gender	6 M, 12 W, 2 NB	8 M, 11 W, 1 NB
Age	M = 22.1, SD = 3.51	M = 25.2, SD = 5.90
Fluency	20 Native	9 Native, 11 Fluent
Handedness	18 Right, 2 Left	18 Right, 2 Left
AR/VR	6 No experience	8 No experience
LEGO	1 No experience	1 No experience

A.1.2 Target Towers. Figure 15 shows an example of the target towers. The study consisted of four repetitions (R1, R2, R3, R4), each with three trials (T), where the goal was to construct a tower. Each tower type (*C*, *L*, *TREE*) appeared once in a repetition in a fully randomized order and with a different position and orientation.

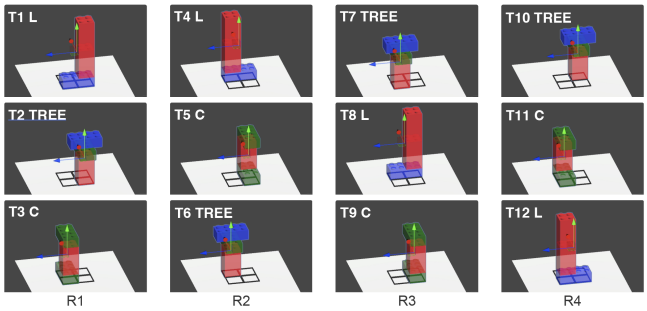

Figure 15: An example order of the twelve target towers. Each tower type (*C, L, TREE*) appears once in each repetition.

A.1.3 Analysis Tower Words. We extracted these words from the transcripts that refer to a tower/sub-tower shape.

- *C*: c, c-shaped, c-shape, crocodile, alligator, alligators
- *L*: l, l-shaped, l-shape, j
- *TREE*: tree
  - *T*: t, t-shaped, t-shape
  - *CROSS*: plus, cross, cross-like, cross-shaped
- OTHER: structure, shape, construction

A.2 Computational Model and Simulations

A.2.1 DSL and Multimodal Signals. The following table shows the mapping between symbols in $\mathcal{D}$ and the corresponding multimodal signals. Each utterance has a cost c_u, which is proportional to the number of words. Hand gestures also have costs c_h, which are manually set. When the Instructor does not use gestures (utterance-only instructions), c_h is set to 0. Since we assume that the Instructor and Builder agents have a uniform prior, one of five random symbols ($\alpha, \beta, \gamma, \delta, \epsilon$) is assigned to each chunked sub-tower/tower. The longer utterances for positions ($P_{i,j}$; cost: 0.6 or 0.7) refer to the statement *"on the x of the grid"*, where x is listed in the table below.

Table 2: Mapping between symbols and multimodal signals.

	Utterance (cost, c_u)	Gesture (cost, c_h)
R	place a red block (0.4)	- (-)
G	place a green block (0.4)	- (-)
B	place a blue block (0.4)	- (-)
$P_{1,1}$	top left (0.7), here (0.1)	point: top left (0.6)
$P_{1,2}$	top half (0.7), here (0.1)	point: top half (0.6)
$P_{1,3}$	top right (0.7), here (0.1)	point: top right (0.6)
$P_{2,1}$	left half (0.7), here (0.1)	point: left half (0.6)
$P_{2,2}$	middle (0.6), here (0.1)	point: middle (0.6)
$P_{2,3}$	right half (0.7), here (0.1)	point: right half (0.6)
$P_{3,1}$	bottom left (0.7), here (0.1)	point: bottom left (0.6)
$P_{3,2}$	bottom half (0.7), here (0.1)	point: bottom half (0.6)
$P_{3,3}$	bottom right (0.7), here (0.1)	point: bottom right (0.6)
C	place a $\{\alpha, \beta, \gamma, \delta, \epsilon\}$ tower (0.4)	shape: $\{\alpha, \beta, \gamma, \delta, \epsilon\}$ (0.6)
L	place a $\{\alpha, \beta, \gamma, \delta, \epsilon\}$ tower (0.4)	shape: $\{\alpha, \beta, \gamma, \delta, \epsilon\}$ (0.6)
TR	place a $\{\alpha, \beta, \gamma, \delta, \epsilon\}$ tower (0.4)	shape: $\{\alpha, \beta, \gamma, \delta, \epsilon\}$ (0.6)
T	place a $\{\alpha, \beta, \gamma, \delta, \epsilon\}$ tower (0.4)	shape: $\{\alpha, \beta, \gamma, \delta, \epsilon\}$ (0.6)
PL	place a $\{\alpha, \beta, \gamma, \delta, \epsilon\}$ tower (0.4)	shape: $\{\alpha, \beta, \gamma, \delta, \epsilon\}$ (0.6)

A.2.2 Tower Programs. Figure 16 shows target towers, tower programs, and their program length. In the simulation, the Instructor proposes a new tower program, from top to bottom, on each list.

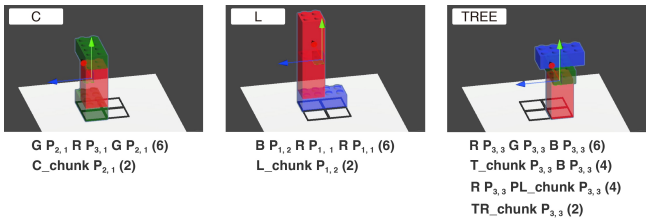

Figure 16: Example target towers, tower programs, and their program length in parentheses, as used by the model.