Situated Grounding and Spatial Reasoning

Situated Grounding and Spatial Reasoning

- Multimodal Situated Grounding – co-perception and co-attention are necessary to understand deixis and relative spatial expressions
  - *Put the big one right here.*
  - *I want the cookie on the left behind the donut.*
  - *Show me a coffee shop around … here.*
- Understanding Events and their Results – actions change the spatial nature of the environment
  - *Mary opened the door and left the room.*
  - *Put the book in the bag. Take the bag to the car.*
  - *Remove the seeds and cut into thin strips, then brown in oil.*
- Appreciation of spatial properties of objects - intrinsic vs. relative Frame of Reference
  - *The tree behind the bench*
  - *The bench in front of the tree*

Approaches and Tools

- Spatial Reasoning and Situated Grounding

  Spatial AMR
- Human-Object Interactions and affordance reasoning

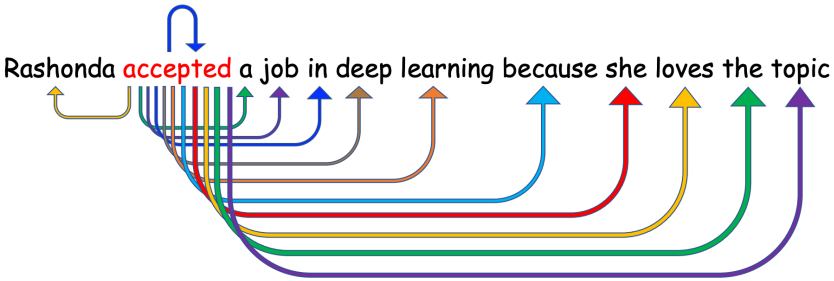
  multimodal dialogue and interactions; understanding events and their results.
- Dense paraphrasing - Data augmentation:

  GLAMR (Object Change-tracking)

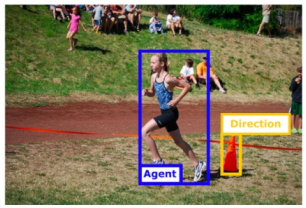
  Converting any modality into textual representations
- Vision Language Action Models (VLA)

Levels of Grounding

1. Self-grounding (unimodal)



2. Cross-grounding (multimodal)



Ground Truth/Retrieved:
[A young lady wearing blue and black]_Agent is running [past an orange cone]_Direction.

Ground Truth:
[A fashionable young woman seated on a bench]_Agent gazes [into a makeup mirror]_Direction.
Retrieved:
[An elderly man]_Agent sitting on [a bench]_Instrument [while reading a book]_Temporal.

Situated Grounding and Context

- Task-oriented dialogues are embodied interactions between agents, where language, gesture, gaze, and actions are situated within a common ground shared by all agents in the communication.
- Situated semantic grounding assumes shared perception of agents with co-attention over objects in a situated context, with co-intention towards a common goal.

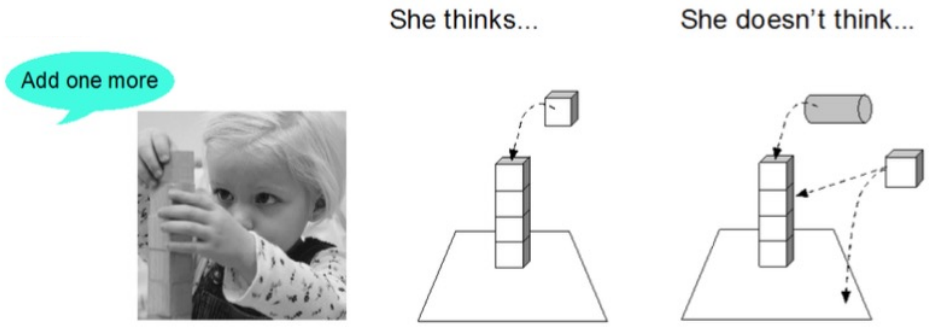
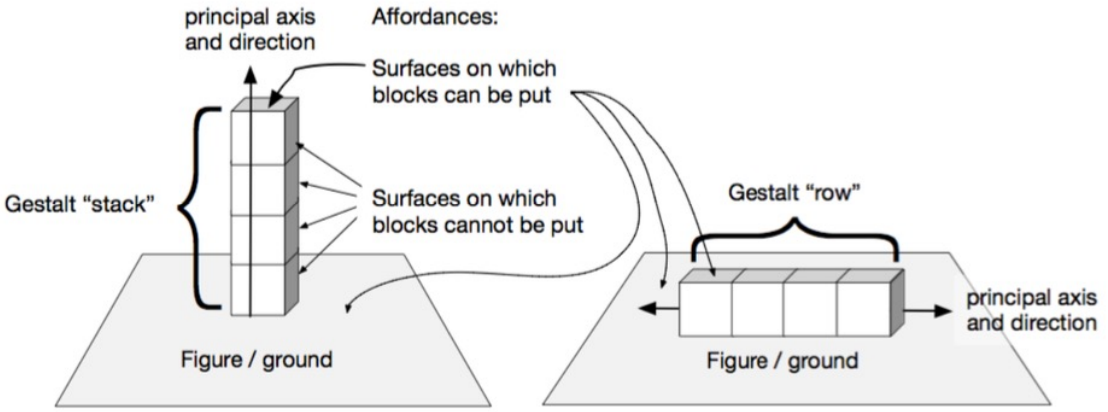


Figure 2. "Add one more" is ambiguous out of context, but given context it is remarkably precise.



DARPA Communicating with Computers BAA - Paul Cohen, PM (2015)

Motivating Example

Figure: "Put on your bee suit."

- What does "bee suit" mean in context?
  - An outfit used *for* beekeeping?
  - An outfit *resembling* a bee?

Motivating Example

Figure: "Put on your bee suit."

- What does "bee suit" mean in context?
  - An outfit used *for* beekeeping?
    - Generative Lexicon: $\textsc{telic} = \lambda z, e[\mathit{beekeeping}(e, z, x)]$
  - An outfit *resembling* a bee?
    - Generative Lexicon: $\textsc{formal} = \mathit{bee}(x)$

Motivating Example

Figure: "Put on your bee suit."

- Many such ways to make this distinction
  - An outfit used *for* beekeeping:
    - AMR: `ARG0 (w / wearer): (b / beekeeper)`
  - An outfit *resembling* a bee:
    - AMR: `ARG0 (w / wearer): (c / child)`

Motivating Example


Figure: "Put on your bee suit."

$\text{TELIC} = \lambda z, e[\textit{beekeeping}(e, z, x)]$	`ARG0 (w / wearer):  (b / beekeeper)`
$\text{FORMAL} = \textit{bee}(x)$	`ARG0 (w / wearer):  (c / child)`
Role-focused	Actor-focused

Motivating Example

- Different representations use different strategies.
- No matter the strategies, a *situationally complete* inference requires grounding representation to items in the discourse and in the environment.
- This requires merging deep semantic representation techniques and flexible neural estimation approaches.

Situated Grounding – Foundational Work

- Cassell, J., Nakano, Y., Reinstein, G., & Stocky, T. (2003). Towards a model of face-to-face grounding. ACL.
- Holroyd, A., Rich, C., Sidner, C. L., & Ponsler, B. (2011). Generating connection events for human-robot collaboration. *2011 RO-MAN,* IEEE.
- Traum, D., and Rickel, J. (2002). Embodied agents for multi-party dialogue in immersive virtual worlds. *Proc. Autonomous agents and multiagent systems*.
- Scheutz, M., Schermerhorn, P., Kramer, J., & Anderson, D. (2007). First steps toward natural human-like HRI. *Autonomous Robots*, *22*(4).
- Elliott, D., D. Kiela and A. Lazaridou (2016) Multimodal Learning and Reasoning, ACL Tutorial.
- Yatskar, M., Zettlemoyer, L., and Farhadi, A. (2016). Situation recognition: Visual semantic role labeling for image understanding. Proceedings of the IEEE CVPR.
- Hough, J., & Schlangen, D. (2017). A Model of Continuous Intention Grounding for HRI.
- Alikhani, M., and Stone, M. (2020). Achieving Common Ground in Multi-modal Dialogue. In Proceedings of the 58th Annual Meeting of ACL Tutorial
- Henlein, A., Gopinath, A., Krishnaswamy, N., Mehler, A., & Pustejovsky, J. (2023). Grounding human-object interaction to affordance behavior in multimodal datasets. Frontiers in artificial intelligence, 6, 1084740.
- Chen, Zhenyu, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X. Chang. "Unit3d: A unified transformer for 3d dense captioning and visual grounding." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 18109-18119. 2023.
- Wang, Haozhong, Hua Yu, and Qiang Zhang. "Detecting Zero-Shot Human-Object Interaction with Visual-Text Modeling." In 2023 9th International Conference on Virtual Reality (ICVR), pp. 155-162. IEEE, 2023.

Situated Grounding and Spatial Reasoning

- Frames of Reference
  - *Absolute (coordinate system)*
  - *Relative (from an agent view)*
  - *Intrinsic (inherent property of object)*

Levinson, S. C. (2003). Space in language and culture: Explorations in cognitive diversity. Cambridge: Cambridge University Press

☺ On the left side of the picture is a big tree.

☹ or ☺ A tree is in the center of the scene.

☹ The tree's shadow is in the lower left corner.

☹ A bench is in front of a tree.

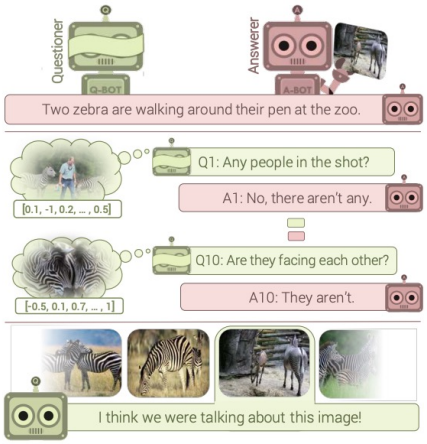
☹ On the left side of the picture is a big tree.

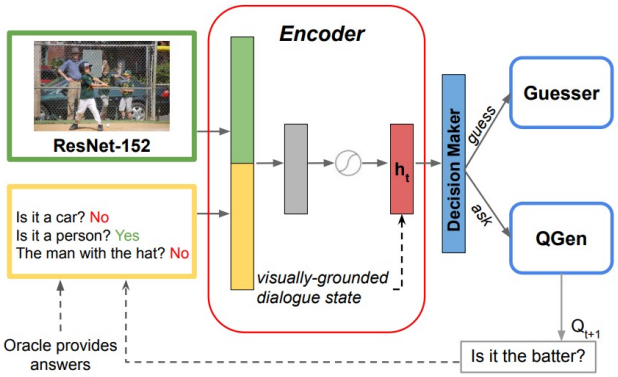
☺ A tree is in the center of the scene.

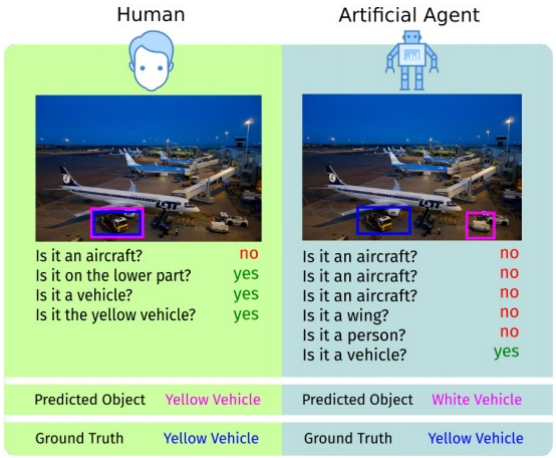
☺ The tree's shadow is in the lower left corner.

☺ A bench is in front of a tree.

Interactive Object Recognition in Dialogue







- Das, A., Kottur, S., Moura, J. M., Lee, S., & Batra, D. (2017). Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE international conference on computer vision* (pp. 2951-2960).
- Shekhar, R., Testoni, A., Fernández, R., & Bernardi, R. (2019). Jointly Learning to See, Ask, Decide when to Stop, and then GuessWhat. In *CLiC-it*.
- Shukla, P., Elmadjian, C., Sharan, R., Kulkarni, V., Turk, M., & Wang, W. Y. (2019). What Should I Ask? Using Conversationally Informative Rewards for Goal-Oriented Visual Dialog. *arXiv preprint arXiv:1907.12021*.
- Kim, Hyounghun, Hao Tan, and Mohit Bansal. "Modality-balanced models for visual dialogue." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05, pp. 8091-8098. 2020.

Spatial Reasoning in Collaborative Tasks

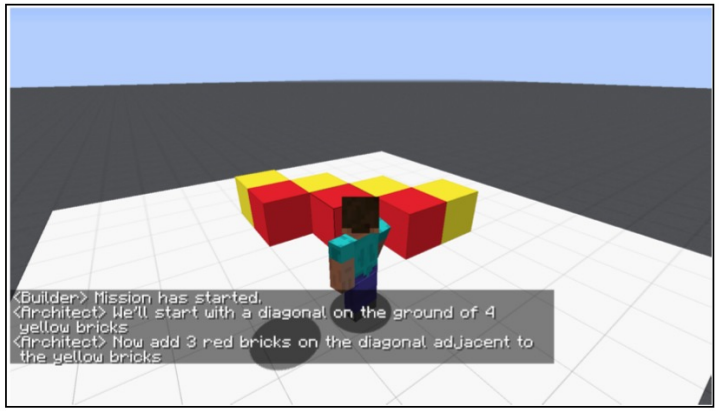


Figure 1: An instance of the collaborative building task. The last instruction was : Now add 3 red bricks on the diagonal adjacent to the yellow bricks.

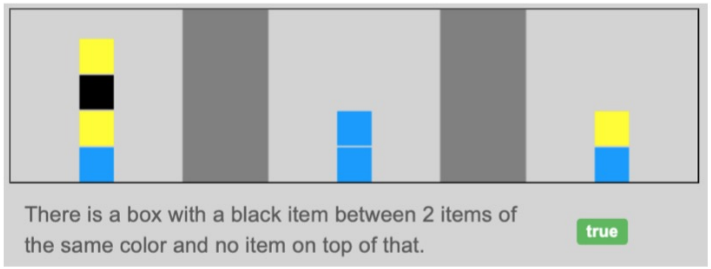


Figure 2: An example from the NLVR corpus that demonstrates *spatial focus shift* from the *black item* to the *yellow item.*

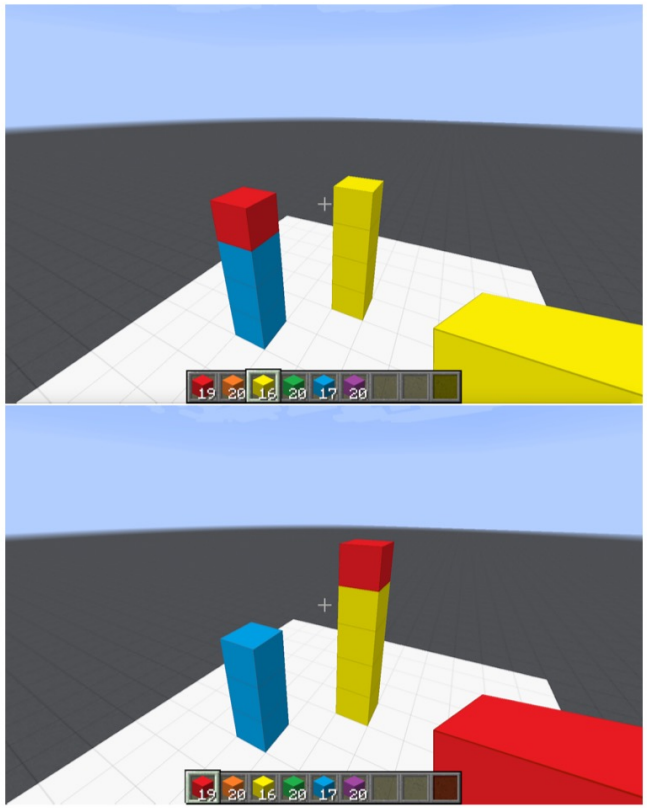
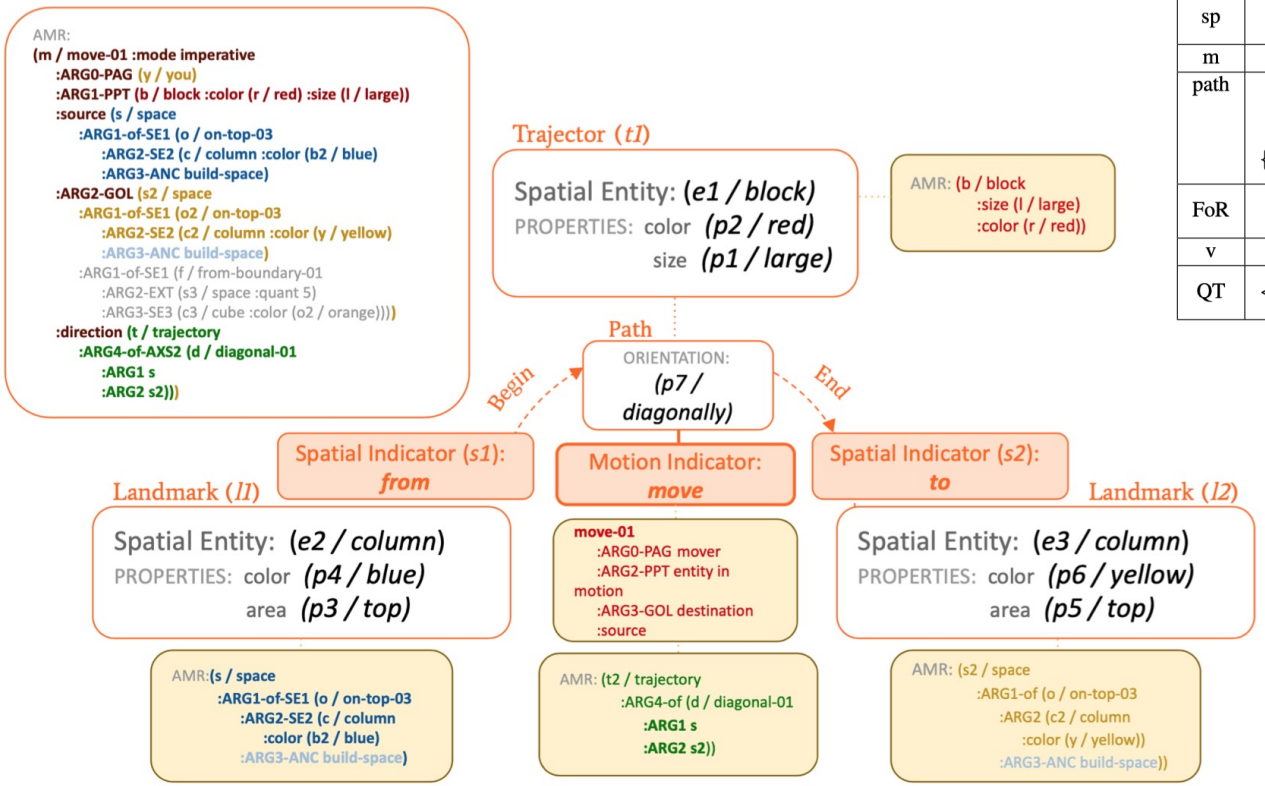


Figure 7: Move the large red block diagonally from the top of the blue column to the top of the yellow column ...

Dan, S., Kordjamshidi, P., Bonn, J., Bhatia, A., Cai, Z., Palmer, M., & Roth, D. (2020). From Spatial Relations to Spatial Configurations. *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 5855-5864).

Spatial Reasoning and AMRs



	Configuration 1	Configuration 2
tr	$< t1, e1 >$	$< t2, e3 >$
lm	$< l1, e2 >, < l2, e3 >$	$< l3, e4 >$
sp	$<s1,from >$ $<s2,to >$	$<s1,from,$ $\{metric = 5spaces\}>$
m	$<m1,move, >$	NULL
path	$< l1, s1, begin >$ $< l2, s2, end >$ $\{orientation = diagonally\}$	NULL
FoR	$< l1, relative >$ $< l2, relative >$	$< l3, relative >$
v	first-person	first-person
QT	$<directional, relative>$	$<distal, quantitative>$ $<topological, DC>$

Figure 3: Graphical Representation of Configuration 1 of Table 3 with aligned AMR : ***Move the large red block diagonally from the top of the blue column to the top of the yellow column****, which is 5 spaces from the orange cube.*

Dan, S., Kordjamshidi, P., Bonn, J., Bhatia, A., Cai, Z., Palmer, M., & Roth, D. (2020). From Spatial Relations to Spatial Configurations. *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 5855-5864).

Spatial Reasoning in Minecraft

Create models that generate spatial descriptions

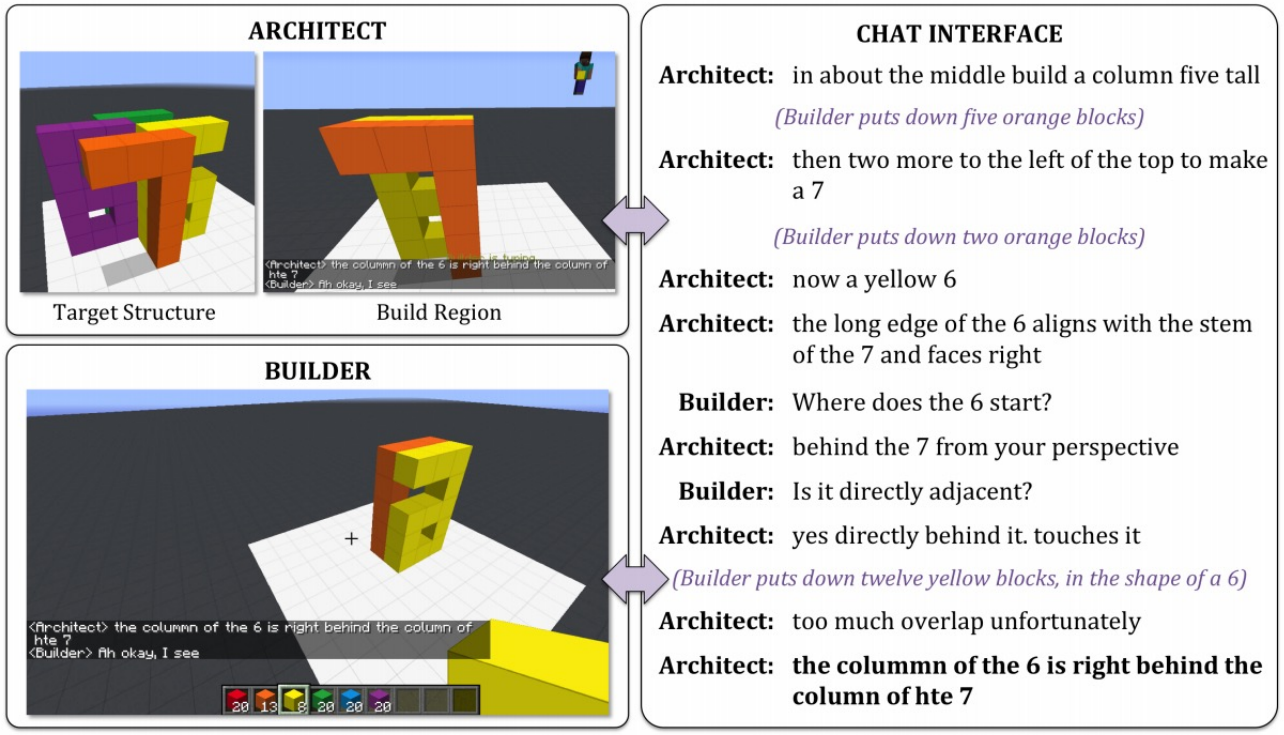


Figure 1: In the Minecraft Collaborative Building Task, the Architect (**A**) has to instruct a Builder (**B**) to build a target structure. **A** can observe **B**, but remains invisible to **B**. Both players communicate via a chat interface. (NB: We show **B**'s actions in the dialogue as a visual aid to the reader.)

Narayan-Chen, A., Jayannavar, P., & Hockenmaier, J. (2019). Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5405-5415).

Spatial Reasoning in Minecraft

Create models that execute spatial actions

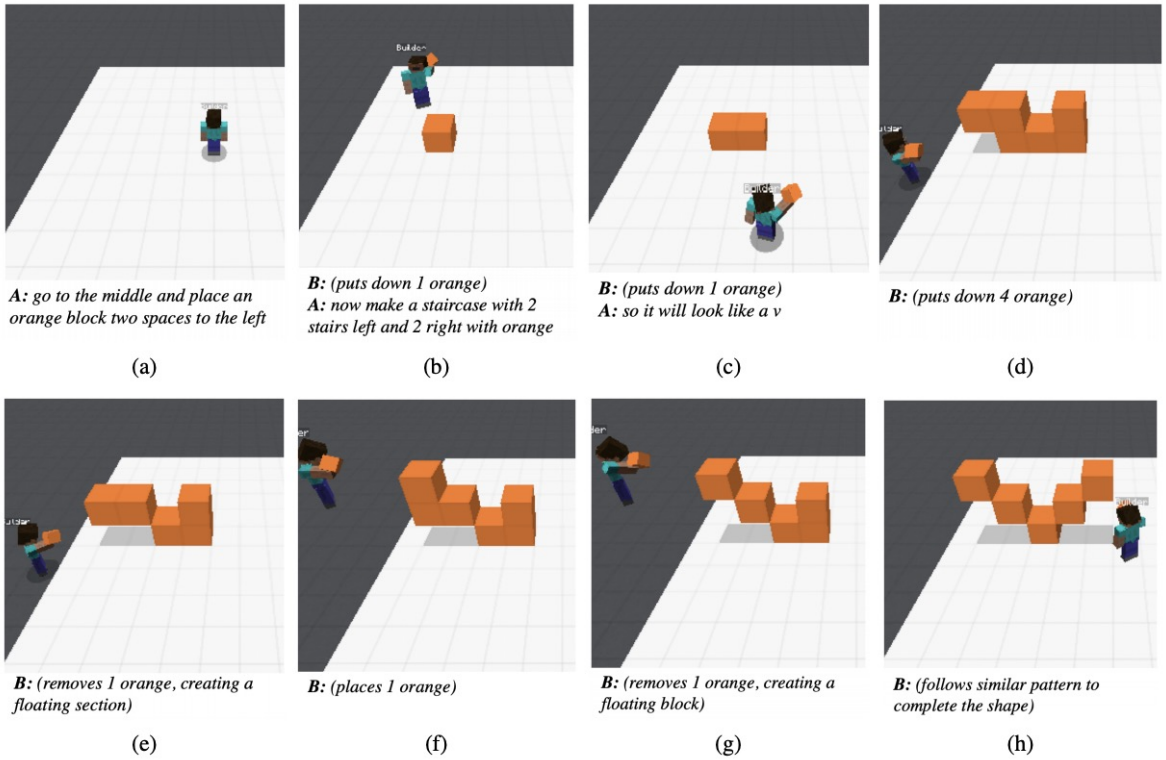
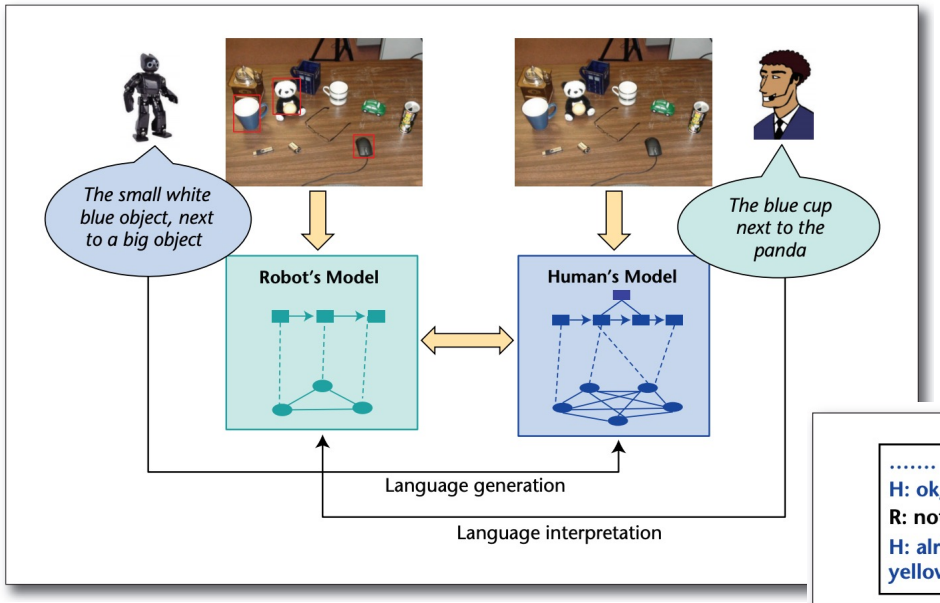


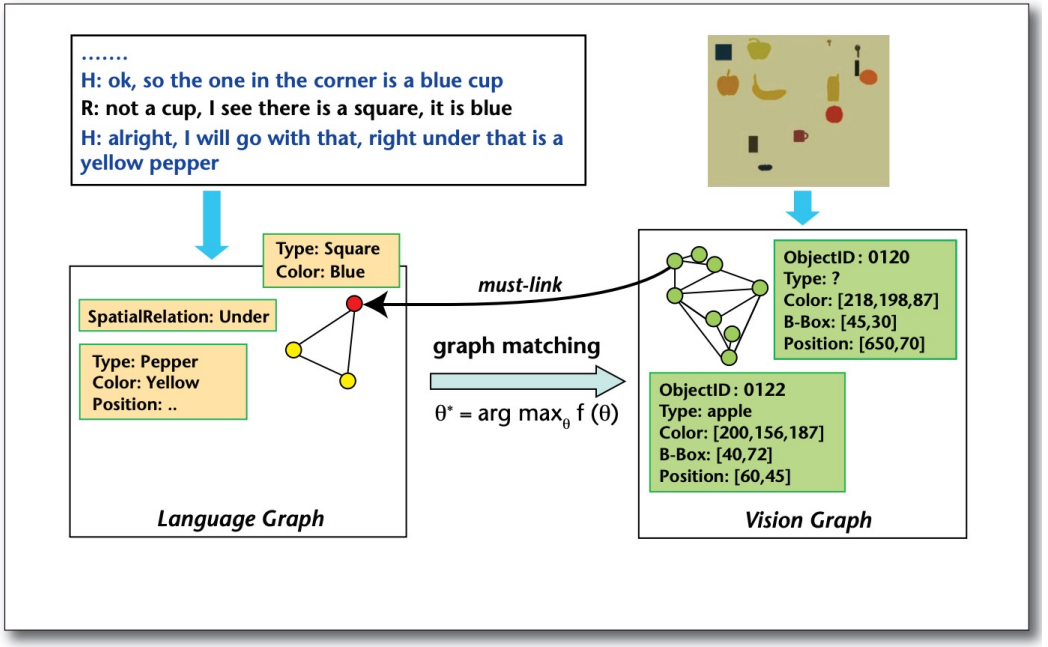
Figure 1: A sample sequence of human-human game states. The game starts with an empty grid and an initial **A** instruction (a), which **B** executes in the first action sequence (b) by placing a single block. In (c), **B** begins to execute the next **A** instruction given in (b). However, **A** interrupts **B** in (c), leading to two distinct **B** action sequences: (b)–(c) (single block placement), and (c)–(h) (multiple placements and removals).

Jayannavar, P., Narayan-Chen, A., & Hockenmaier, J. (2020). Learning to execute instructions in a Minecraft dialogue. In *Proceedings of the 58th Meeting of the Association for Computational Linguistics* (pp. 2589-2602).

Situated Grounding in Human Robot Dialogue



- Establish a Joint Perceptual Basis through language grounding



- Graph-Matching for Interpreting Referring Expressions

Chai, J. Y., Fang, R., Liu, C., & She, L. (2016). Collaborative language grounding toward situated human-robot dialogue. *AI Magazine*, *37*(4), 32-45.

Grounding - Multimodal Spatial Expressions

(1) Here$_{[deixis]}$ is the bus stop, a bit left of it$_{[deixis]}$ is a church and right in front of that$_{[deixis]}$ is the hotel.

Figure 1: Providing a multimodal description (*left*) of a scene (*right*).

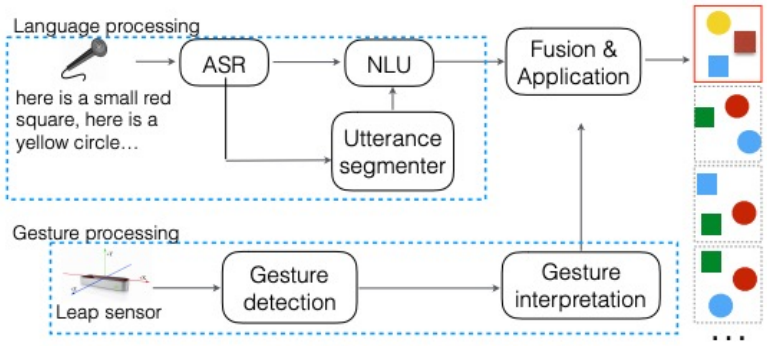


Figure 2: Multimodal system architecture.

- Interpreting multimodal spatial descriptions in route giving tasks.
- Gestures not only contribute information, but also help interpretations of speech incrementally, due to its parallel nature.

Han, T., Kennington, C., & Schlangen, D. (2018). Placing Objects in Gesture Space: Toward Real-Time Understanding of Spatial Descriptions. In *AAAI18.*

Situated Grounding and Pointing Actions

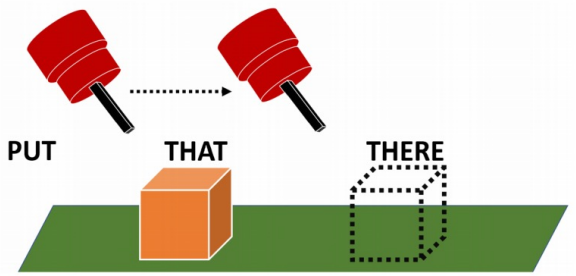


Figure 1: A pick-and-place task requires a *referential* pointing action to the object (orange cube) at the initial position, and a *locating* pointing action to a final placement position (dotted cube). Such an action by a robot (in red) can also be accompanied by verbal cues like *"Put that there."*

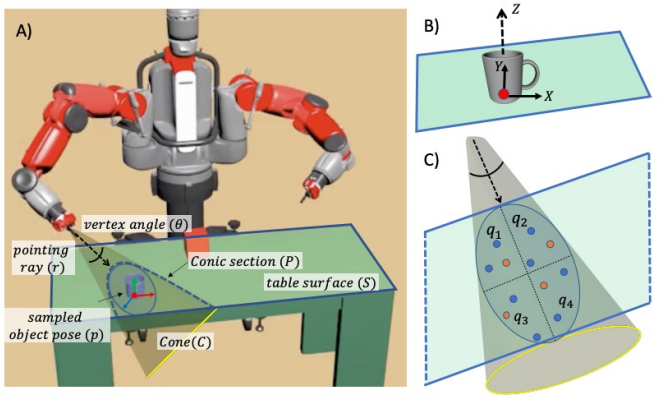


Figure 2: (A) Workspace setup showing the pointing cone and the corresponding conic section on the table. (B) The degrees-of-freedom considered for placement of the object on the table. (C) Sampling policy to sample object poses within the conic section.

- Pointing to something vs. somewhere
- Human subjects show greater flexibility in interpreting the intent of referential pointing compared to locating pointing, which needs to be more deliberate.

Alikhani, M., Khalid, B., Shome, R., Mitash, C., Bekris, K. E., & Stone, M. (2020). That and There: Judging the Intent of Pointing Actions with Robotic Arms. In AAAI (pp. 10343-10351).

Spatial Reasoning and Situated Meaning

Situated Meaning in a Joint Activity

- Son: *Put it there (gesturing with co-attention)?*
- Mother: *Yes, go down for about two inches.*
- Mother: *OK, stop there. (co-attentional gaze)*
- Son: *Okay. (stops action)*
- Mother: *Now, start this one (pointing to another cupcake).*

Krishnaswamy, N. and Pustejovsky, J. (2020). Neurosymbolic AI for Situated Language Understanding. In Annual Conference on Advances in Cognitive Systems (ACS). Cognitive Systems Foundation.

Situated Meaning and Common Ground

Agents	mother, son
Shared goals	baking, icing
Beliefs, desires, intentions	Mother knows how to ice, bake, etc. Mother is teaching son
Objects	Mother, son, cupcakes, plate, knives, pastry bag, icing, gloves
Shared perception	the objects on the table
Shared Space	kitchen

Stalnaker R., “Common ground”, *Linguistics and philosophy*, vol. 25, no 5-6, p. 701-721, 2002

Clark H. H., Brennan S. E., “Grounding in communication”, *Perspectives on socially shared cognition*, vol. 13, p. 127-149, 1991.

Embodiment and Situated Grounding

- Task-oriented dialogues are embodied interactions between agents, where language, gesture, gaze, and actions are situated within a common ground shared by all agents in the communication.
- Situated semantic grounding assumes shared perception of agents with co-attention over objects in a situated context, with co-intention towards a common goal.
- VoxWorld : a multimodal simulation framework for modeling Embodied Human-Computer Interactions and communication between agents engaged in a shared goal or task.
- Embodied HCI and robot control in action.

Pustejovsky, J., & Krishnaswamy, N. (2020). Embodied Human-Computer Interactions through Situated Grounding. In Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents.

Situated Grounding in Dialogue

A non-verbal interaction between a human and IVA using gesture, gaze, and action.

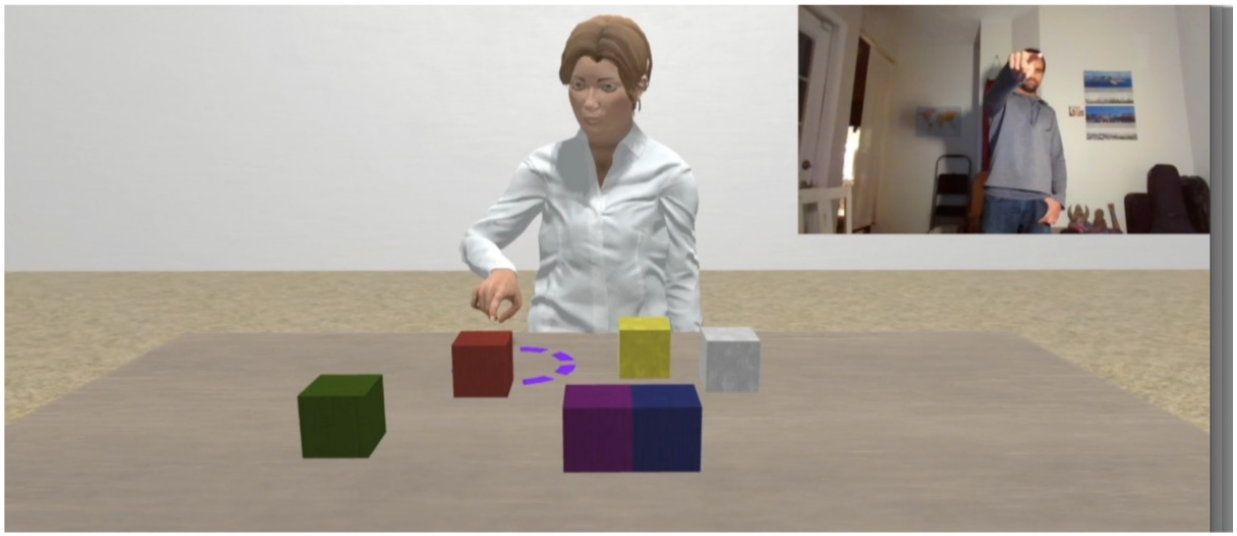


Figure: IVA Diana engaging in an embodied HCI with a human user.

Krishnaswamy, Nikhil, Pradyumna Narayana, Rahul Bangar, Kyeongmin Rim, Dhruva Patil, David McNeely-White, Jaime Ruiz, Bruce Draper, Ross Beveridge, and James Pustejovsky. "Diana's World: A Situated Multimodal Interactive Agent." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 09, pp. 13618-13619. 2020.

Embodiment and Situated Grounding

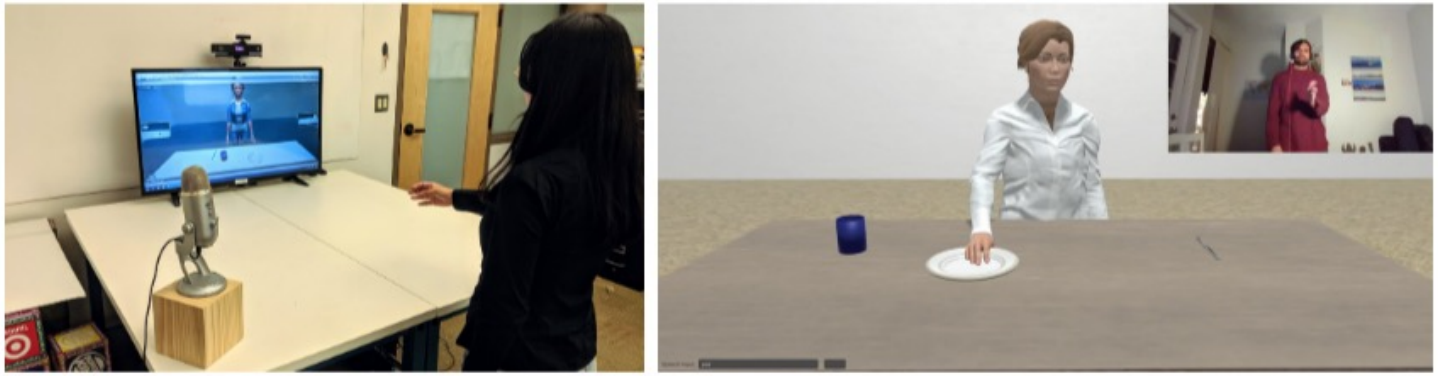


Figure 1: *Diana's interactive setup within the real world (L) and Diana's environment (human inset in upper right) (R)*

Pustejovsky, J., & Krishnaswamy, N. (2020). Embodied Human-Computer Interactions through Situated Grounding. In Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents.

HOI and Situated Grounding

- Identify requirements involved in developing a semantics for referential grounding in a situated context.
- This models a native human capability , so we study Human-human interactions (HHI) in multimodal communication.
- Modeling human-object interactions for communication
  - Object properties and behaviors
  - actions associated with objects

Spatial Semantics and Situated Grounding

- Identifying the *actions and consequences* associated with objects in the environment.
- Encoding a multimodal expression contextualized to the *dynamics of the discourse*
- *Situated grounding*: Capturing how multimodal expressions are anchored, contextualized, and situated in context

Spatial Properties of Objects

- Object size, shape, dimensionality, texture
- Orientation, frame of reference, facing (front/back)
- How we spatially interact with an object
- Space needed for Object function - affordance space
- Event space used for object function or purpose

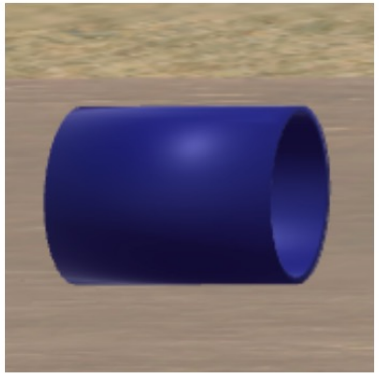
Pustejovsky, J., & Krishnaswamy, N. (2016). VoxML: A Visualization Modeling Language. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).

Krishnaswamy, N., & Pustejovsky, J. (2016). VoxSim: A visual platform for modeling motion language. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations.

Spatial Properties of Objects

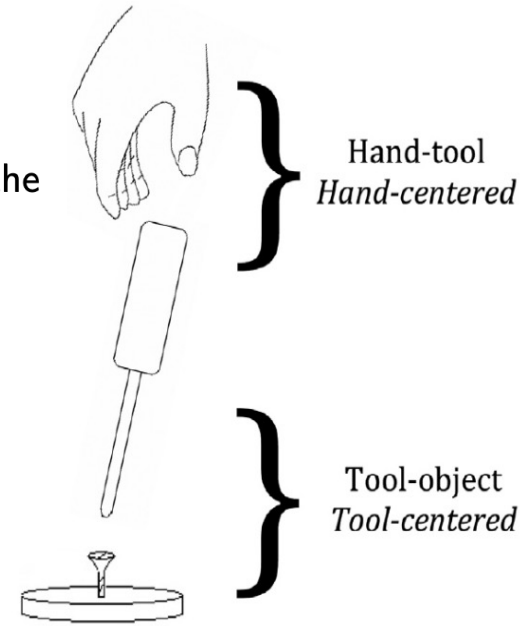
- Context of objects is described by their properties.
- Object properties cannot be decoupled from the events they facilitate.
  - *Affordances* (Gibson, 1979)
  - *Qualia* (Pustejovsky, 1995)

"He **slid** the cup across the table. Liquid spilled out."
"He **rolled** the cup across the table. Liquid spilled out."



Reference Frames and Affordances

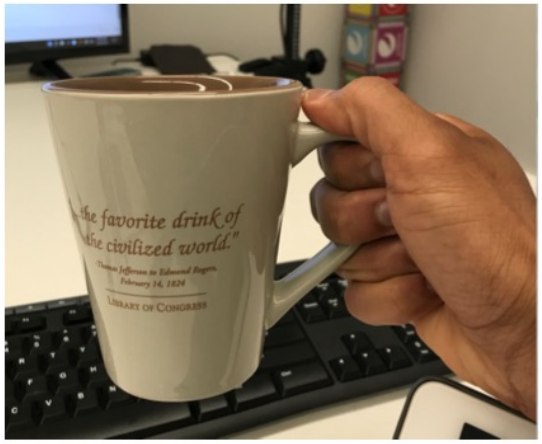
- Hand-tool Interface: based on the agent's biomechanical and morphological characteristics. For instance, a hammer is graspable by a human adult but not by a baby. Thus, the interface is centered on the agent.
- Tool-object Interface: independent of the agent's characteristics. The relationship is centered on objects external to the agent and the interaction is made possible because of the compatibility between the characteristics of the tool and the object.



Osiurak, F., Rossetti, Y., and Badets, A. (2017). What is an affordance? 40 years later. *Neuroscience & Biobehavioral Reviews*, 77, 403-417.

Affordance Space and Grasp Poses



Pustejovsky, J., Krishnaswamy, N., and Do, T. (2017). Object embodiment in a multimodal simulation. In *AAAI Spring Symposium: Interactive Multisensory Object Perception for Embodied Agents*.

Habitats and Affordances

Different Habitats for Object Use

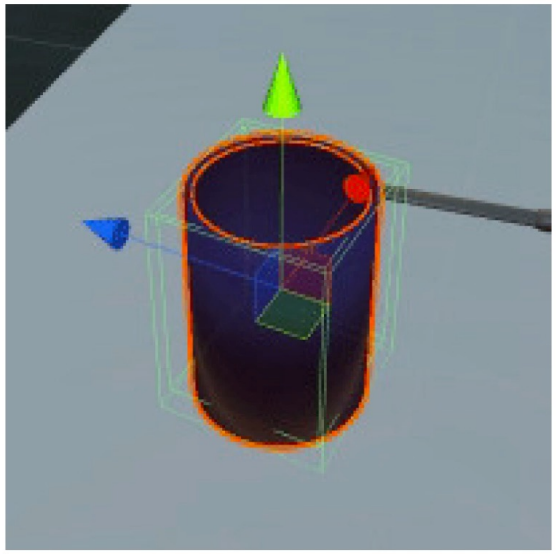
Top: *Spoon* allowing holding (left) and stirring (right).
Bottom: *Knife* allowing spreading (left) and cutting (right)

VoxML: Visual Object Concept Modeling Language

- Encodes afforded behaviors for each object
  - Gibsonian: afforded by object structure (Gibson,1977,1979)
    - grasp, move, lift, etc.
  - Telic: goal-directed, purpose-driven (Pustejovsky, 1995, 2013)
    - drink from, read, etc.
- Voxeme
  - Object Geometry: Formal object characteristics in R3 space
  - Habitat: Orientation, Situated context, Scaling
  - Affordance Structure:
    - What can one do to it
    - What can one do with it
    - What does it enable

VoxML - cup

$$
\begin{bmatrix}
\textbf{cup} \\
\text{LEX} = \begin{bmatrix} \text{PRED} = \textbf{cup} \\ \text{TYPE} = \textbf{physobj, artifact} \end{bmatrix} \\
\text{TYPE} = \begin{bmatrix} \text{HEAD} = \textbf{cylindroid[1]} \\ \text{COMPONENTS} = \textbf{surface, interior} \\ \text{CONCAVITY} = \textbf{concave} \\ \text{ROTATSYM} = \{Y\} \\ \text{REFLECTSYM} = \{XY, YZ\} \end{bmatrix} \\
\text{HABITAT} = \begin{bmatrix} \text{INTR} = {}_{[2]}\begin{bmatrix} \text{CONSTR} = \{Y > X, Y > Z\} \\ \text{UP} = align(Y, \mathcal{E}_Y) \\ \text{TOP} = top(+Y) \end{bmatrix} \\ \text{EXTR} = {}_{[3]}\begin{bmatrix} \text{UP} = align(Y, \mathcal{E}_{\perp Y}) \end{bmatrix} \end{bmatrix} \\
\text{AFFORD\_STR} = \begin{bmatrix} \text{A}_1 = H_{[2]} \rightarrow [put(x, on([1]))] support([1], x) \\ \text{A}_2 = H_{[2]} \rightarrow [put(x, in([1]))] contain([1], x) \\ \text{A}_3 = H_{[2]} \rightarrow [grasp(x, [1])] \\ \text{A}_4 = H_{[3]} \rightarrow [roll(x, [1])] \end{bmatrix} \\
\text{EMBODIMENT} = \begin{bmatrix} \text{SCALE} = \textbf{<agent} \\ \text{MOVABLE} = \textbf{true} \end{bmatrix}
\end{bmatrix}
$$



VoxML for actions and relations

$$
\left[\begin{array}{l}
\textbf{put} \\
\text{LEX} = \left[\begin{array}{l}\text{PRED} = \textbf{put} \\ \text{TYPE} = \textbf{transition\_event}\end{array}\right] \\
\text{TYPE} = \left[\begin{array}{l}\text{HEAD} = \textbf{transition} \\ \text{ARGS} = \left[\begin{array}{l}\text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:physobj} \\ \text{A}_3 = \textbf{z:location}\end{array}\right] \\ \text{BODY} = \left[\begin{array}{l}\text{E}_1 = grasp(x,y) \\ \text{E}_2 = [while(hold(x,y), move(x,y)] \\ \text{E}_3 = [at(y,z) \rightarrow ungrasp(x,y)]\end{array}\right]\end{array}\right]
\end{array}\right]
$$

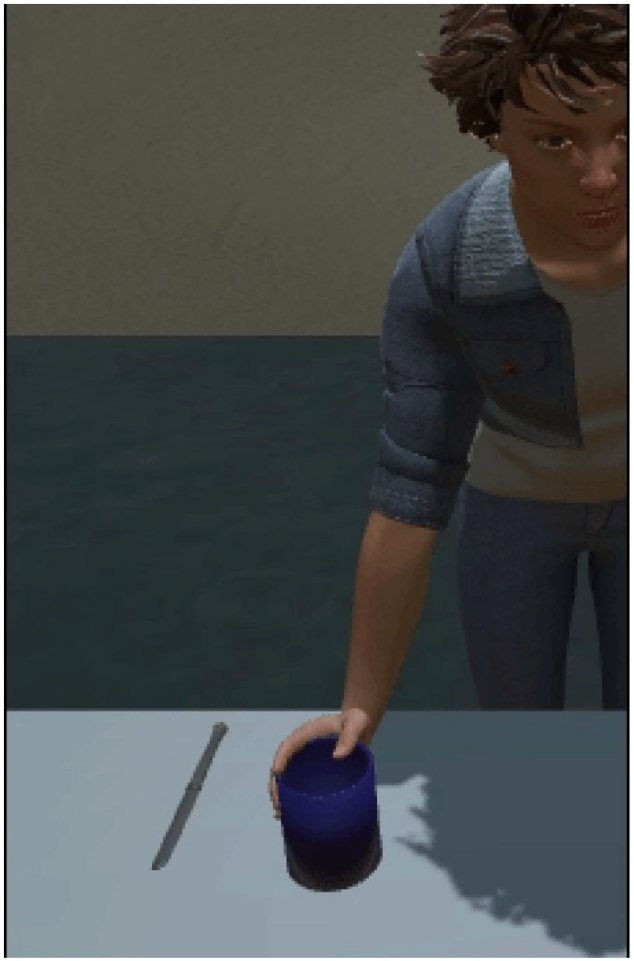
$$
\left[\begin{array}{l}
\textbf{on} \\
\text{LEX} = \left[\text{PRED} = \textbf{on}\right] \\
\text{TYPE} = \left[\begin{array}{l}\text{CLASS} = \textbf{config} \\ \text{VALUE} = \textbf{EC} \\ \text{ARGS} = \left[\begin{array}{l}\text{A}_1 = \textbf{x:3D} \\ \text{A}_2 = \textbf{y:3D}\end{array}\right] \\ \text{CONSTR} = \textbf{y}\rightarrow\text{HABITAT}\rightarrow\text{INTR}[align]\end{array}\right]
\end{array}\right]
$$

VoxML - grasp

$$
\left[\begin{array}{l}
\textbf{grasp} \\
\text{LEX} = \left[\begin{array}{l}\text{PRED} = \textbf{grasp} \\ \text{TYPE} = \textbf{transition\_event}\end{array}\right] \\
\text{TYPE} = \left[\begin{array}{l}\text{HEAD} = \textbf{transition} \\ \text{ARGS} = \left[\begin{array}{l}\text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:physobj}\end{array}\right] \\ \text{BODY} = \left[\text{E}_1 = grasp(x, y)\right]\end{array}\right]
\end{array}\right]
$$

VoxML – Composition [grasp + cup]

- Continuation-passing style semantics for composition
- Used within conventional sentence structures
- Used between sentences in discourse
- Used for gesture sequencing as well



Krishnaswamy, N., & Pustejovsky, J. (2019). Multimodal Continuation-style Architectures for Human-Robot Interaction. *arXiv preprint arXiv:1909.08161*.

QSRLib

Table 1: Description of supported qualitative spatial relation families

qualitative spatial relation families	type	num of relations / variations	kind of entities
Qualitative Distance Calculus	distance	user specified	2D points
Probabilistic Qualitative Distance Calculus	distance	user specified	2D points
Cardinal Directions	direction	9	2D rectangles
Moving or Stationary	motion	2	2D points
Qualitative Trajectory Calculus	motion	B11: 9, C21: 81	2D points
Rectangle/Block Algebra	topology & direction	169/2197	2D/3D rectangles
Region Connection Calculus	topology	2, 4, 5, 8	2D rectangles
Ternary Point Configuration Calculus	direction	25	2D points

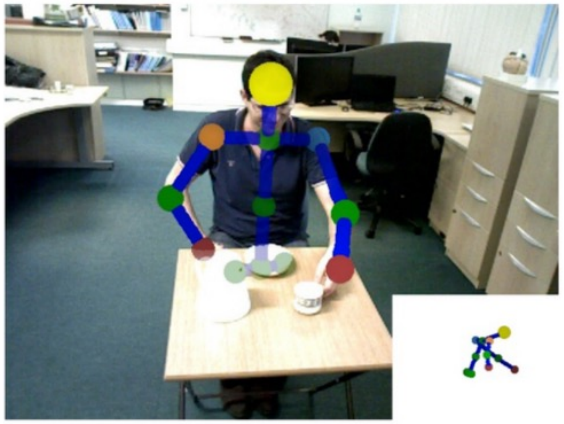


Figure 1: Activity recognition in a table top setting. Dyadic QSR relations between detected objects/skeleton points can be computed (bottom right inset).

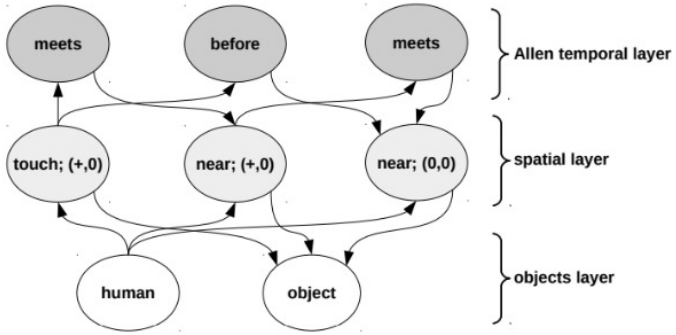
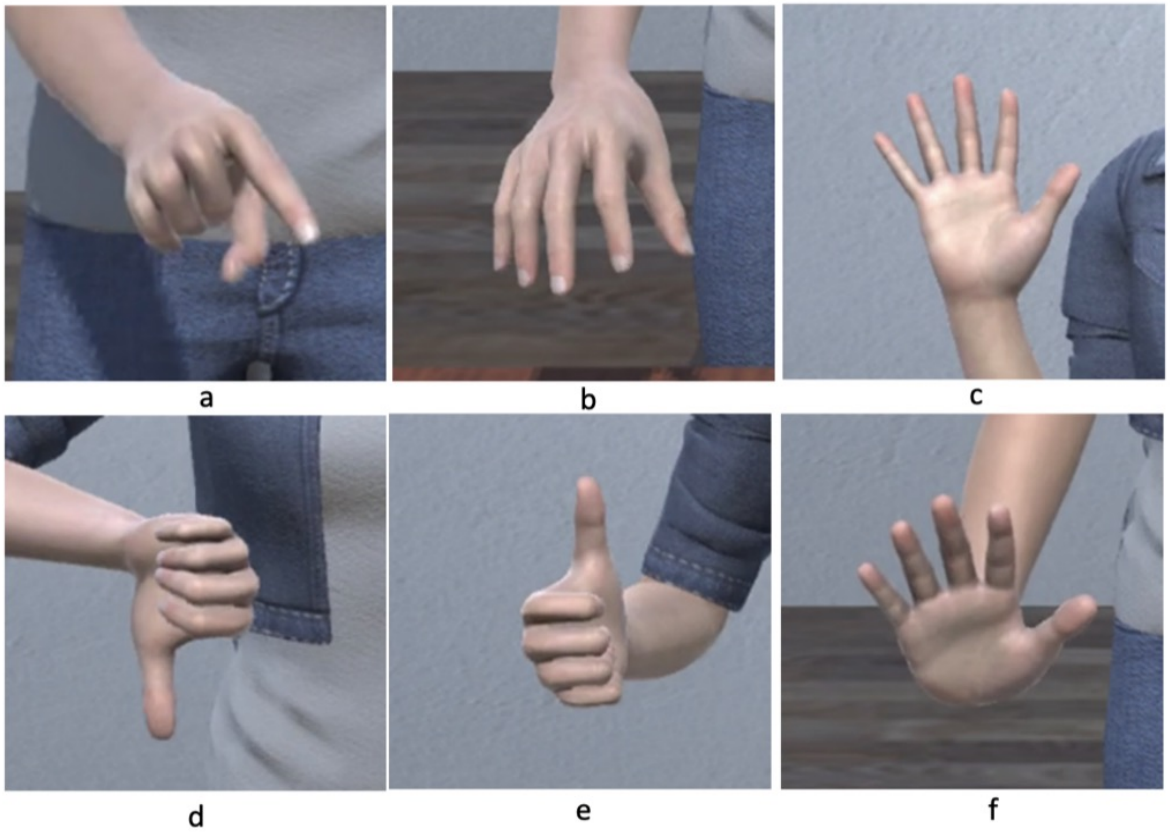


Figure 5: Example of a Qualitative Spatio-Temporal Activity Graph (QSTAG) between a human and an object; each spatial layer node encodes QSRs from two calculi: a QDC relation (touch/near) and a QTC_{B21} one ((+,0)/(0,0)).

- Gatsoulis, Yiannis, Muhannad Alomari, Chris Burbridge, Christian Dondrup, Paul Duckworth, Peter Lightbody, Marc Hanheide, Nick Hawes, D. C. Hogg, and A. G. Cohn. "Qsrlib: a software library for online acquisition of qualitative spatial relations from video." (2016).

Gestures Generated in VoxWorld

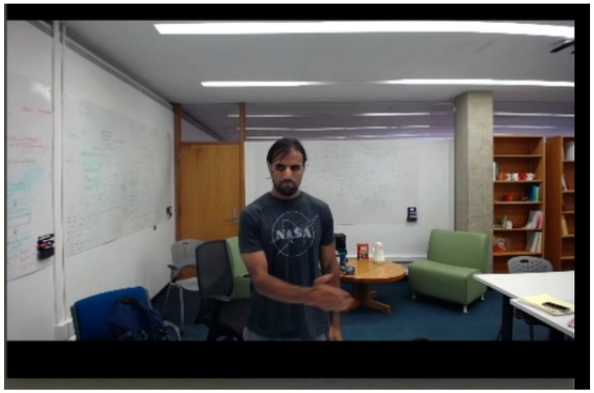
- VoxML encodes spatial configuration of gestures

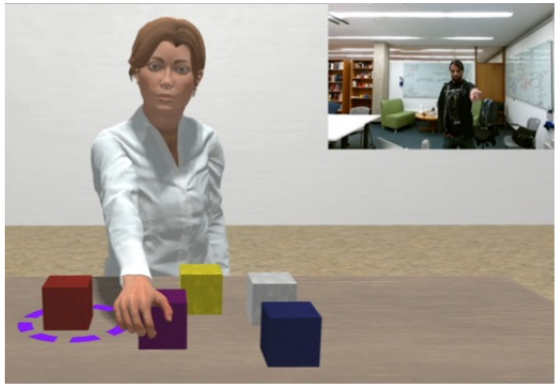


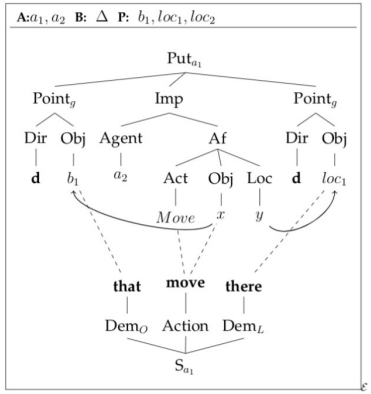
Pustejovsky, J., Krishnaswamy, N., Beveridge, R., Ortega, F. R., Patil, D., Wang, H., & McNeely-White, D. G. Interpreting and Generating Gestures with Embodied Human Computer Interactions, GENEA Workshop, IVA20, 2020.

Multimodal Dialogue

- Language and Gesture determine Situated Grounding
- “That block, move it there.”







$\lambda k'_s \otimes k'_g.\overline{(\langle \textbf{that}, Point_1\rangle\langle \textbf{move}, Move\rangle)}(\lambda r_s \otimes r_g.\overline{\langle \textbf{that}, Point_2\rangle}$
$(\lambda k_s \otimes k_g.k'_s \otimes k'_g(k_s \otimes k_g r_s \otimes r_g)))$

Multimodal Dialogue

- Gesture sequence command

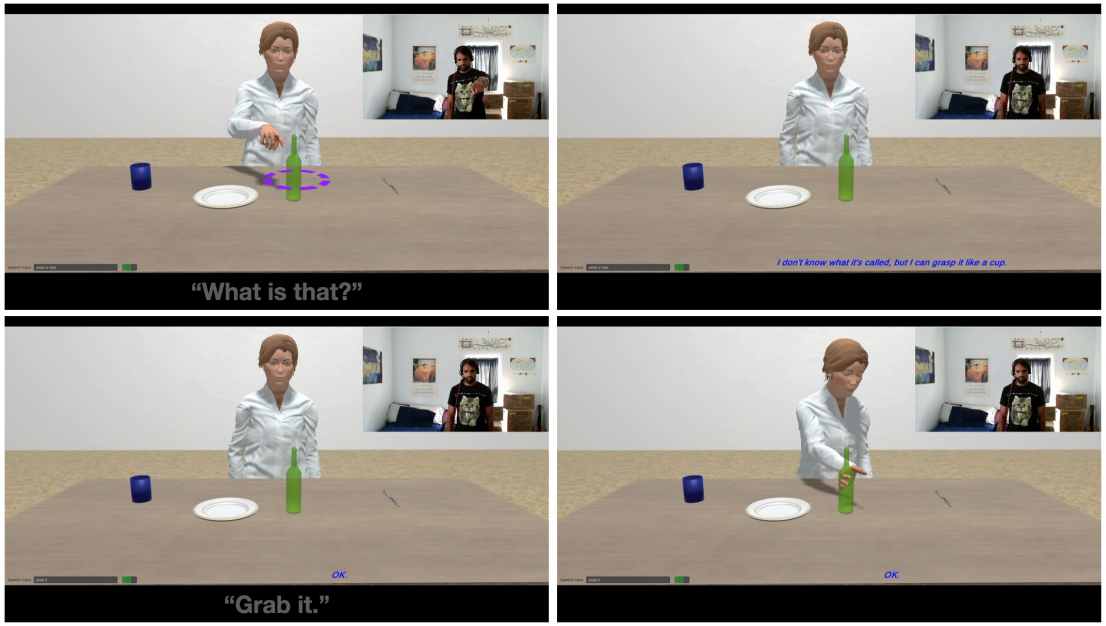
Single Modality (Gesture) Imperative

DIANA_1: $\mathcal{G} = [\textit{points to the purple block}]_{t1}$
DIANA_2: $\mathcal{G} = [\textit{makes move gesture}]_{t2}$
DIANA_3: $\mathcal{G} = [\textit{points to the blue block}]_{t3}$

Krishnaswamy, N., and Pustejovsky, J. (2018). Deictic Adaptation in a Virtual Environment. In *German Conference on Spatial Cognition* (pp. 180-196). Springer, Cham.

Spatial Reasoning and Affordance Learning

- Gibsonian/Telic affordances are associated with abstract properties:
  - spheres roll, sphere-like entities probably do too;
  - small cups are graspable, small cylindroid-shaped objects probably are too.
- Similar objects have similar habitats/affordances:
- This informs the way you can talk about items in context:
  - Q: "What am I pointing at?"
  - A: "I don't know, but it looks like {a ball/a container/etc.}



- Train over a sample of 17 different objects: blocks, KitchenWorld objects (apple, grape, banana, book, etc.)
- Trained 200 dimensional affordance and habitat embeddings using a Skip-Gram model, for 50,000 epochs with a window size of 3:
  - These embeddings serve as the inputs to the object prediction architectures
- Using the affordance embeddings in vector space, predict which object they belong to: using a 7-layer MLP; a 4-layer CNN with 1D convolutions

Affordance Embeddings

Krishnaswamy and Pustejovsky (2021)

- Situated grounding is particularly useful for transfer learning, because similar concepts often exist in similar situations (cf. analogical generalization, a la Forbus et al. (2017)).
  - e.g., "Build an X out of *these*," "Put *all those* in that X."
- Associate affordances with abstract properties—spheres roll, sphere-like entities probably do too.
- This informs the way you can talk about items (in real or virtual situations).
- Q: "What am I pointing at?" A: "I don't know, but it looks like [a container, something that rolls, etc.]"
- Similar objects have similar habitats/affordances.
- What happens when Diana encounters a new object?

Affordance Embeddings

- Exploit the correlations between habitats and affordances over known objects, and map those correspondences to novel objects
- Given: Object + A_1 + A_2 + _____ + A_4, predict A_3
- Goal: "Spheres roll. An apple is spherical. Apples probably roll."
- 17 distinct VoxML objects (~22 distinct affordance encodings):
  - e.g., $H_{[3]} = [\text{UP} = \textit{align}(\bar{Y}, \mathcal{E}_Y), \text{TOP} = \textit{top}(+Y)]$, $H_{[3]} \rightarrow [\textit{put}(x, \textit{in}(\textit{this}))]\textit{contain}(\textit{this}, x)$;
- Train 200-dimensional habitat or affordance embeddings using a Skip-Gram model;
- Represent objects as averaged habitat or affordance vectors.

Affordance Embeddings

- 2 architectures: 7-layer MLP and 4-layer CNN w/ 1D convolutions
- Evaluate against a ground truth of k-means clustered objects derived from human annotators
- Achieve ~80% accuracy with the predicted object clustering with the ground-truth object
  - ~40% of the time the predicted object *always* clusters with the ground truth in 5 randomized trials

Model	% predictions in correct cluster	% predictions always in correct cluster
MLP (Habitats)	78.82%	27.06%
MLP (Affordances)	**84.71%**	38.82%
CNN (Habitats)	78.82%	27.06%
CNN (Affordances)	81.18%	**40.00%**

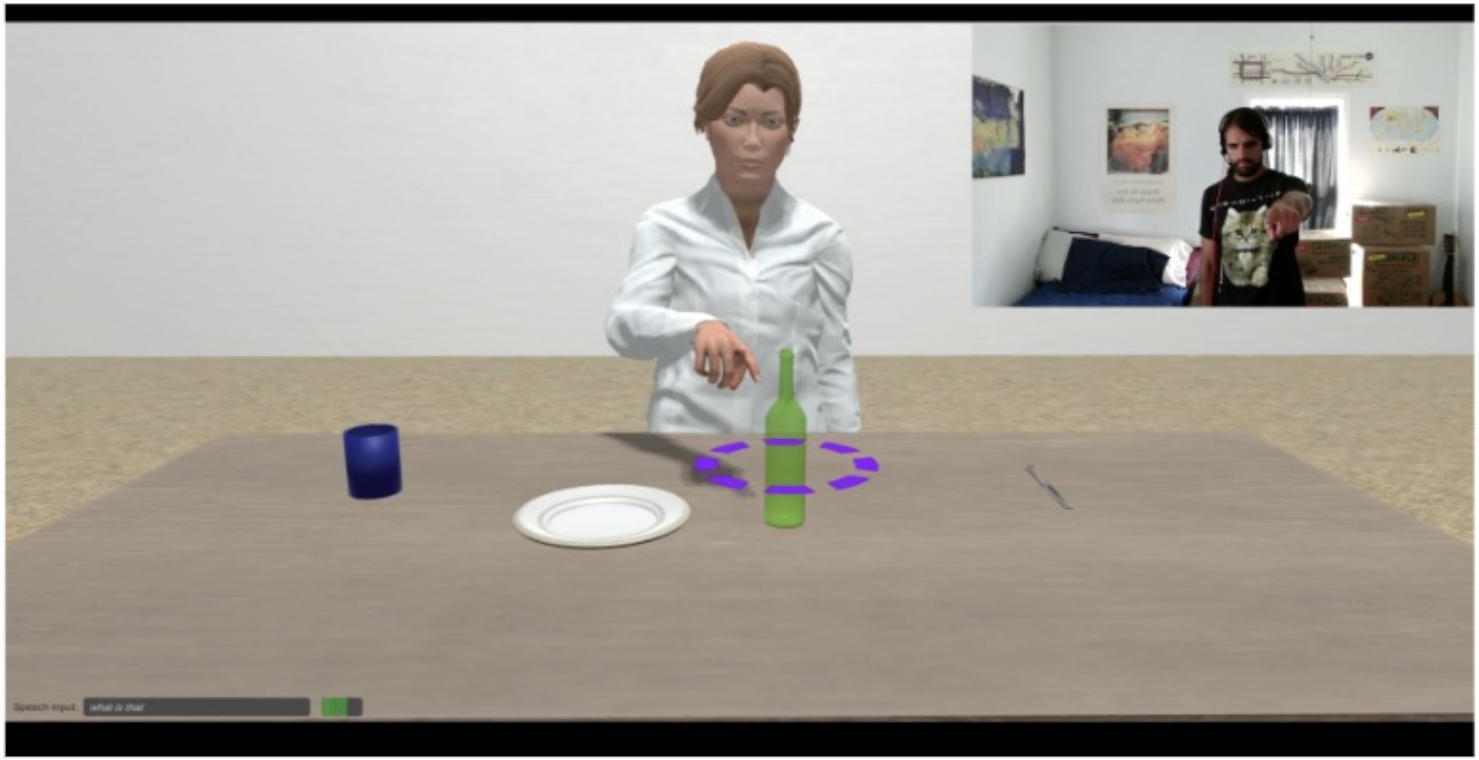
Affordance Embeddings

Tests on individual objects (plate):

Model	MLP-H	MLP-A	CNN-H	CNN-A
Predicted objects	book, cup, bowl, bottle	cup, bottle, apple	book	cup, bottle

- Habitat-based model typically better at capturing common behaviors (e.g., grasping), affordance-based model better at object-specific behaviors (e.g., rolling)

Example of Learning



- http://www.voxicon.net/wp-content/uploads/2020/07/DianaAffordanceTransferLearning.mp4

Data Augmentation

Captions Don't Describe Human-Object Interactions

Neither do conventional semantic representations

"Woman drinking coffee."

(1) a. $drink(w, c)$

b. $\exists x \exists y[\text{woman}(x) \wedge coffee(y) \wedge \text{drink}(x, y)]$

c. $\text{EVENT}(drink) \wedge \text{AGENT}(woman) \wedge \text{PATIENT}(coffee)$

Data Augmentation

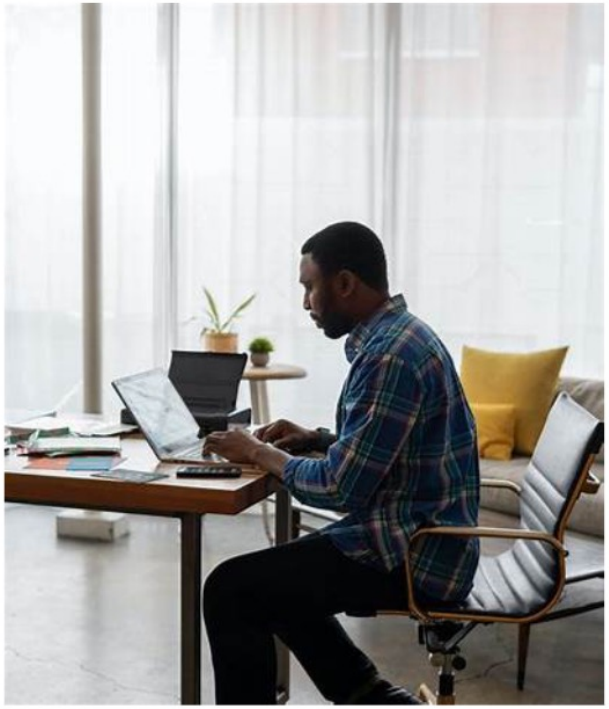
What the Caption Leaves Out

Dense Paraphrase

- *A woman drinking coffee.*
- A upright seated woman is holding in her hand, a cup filled with coffee while she drinks it.
- The cup is upright so the container portion (inside) is able to hold coffee.
- She is holding the cup by an attached handle.
- The cup is tilted towards her and touches her partially open mouth, in order to allow drinking.

Data Augmentation

Captions Don't Describe Human-Object Interactions



"A man working at a desk."

Data Augmentation

What the Caption Leaves Out

- *A man working at a desk.*
- A upright man is seated in a chair, typing with both hands on the keyboard of a laptop, which is on the top surface of a table.
- The chair he is seated in is close enough to the table for him to reach the keyboard.
- The laptop is open, with the keyboard exposed flat and the screen facing the man.
- The man is facing the computer screen and keyboard and the desk.

Situated Communications

- Multimodal Situated Grounding – co-perception and co-attention are necessary to understand deixis and relative spatial expressions

  Put the big one right here.

  I want the cookie on the left behind the donut.

  Show me a coffee shop around … here.

- Understanding Events and their Results – actions change the spatial nature of the environment

  Mary opened the door and left the room.

  Put the book in the bag. Take the bag to the car.

  Remove the seeds and cut into thin strips, then brown in oil.

- Appreciation of spatial properties of objects - intrinsic vs. relative Frame of Reference

  The tree behind the bench

  The bench in front of the tree

Paraphrase Grammars

Non-derivational Transformation Grammar

- Linguistic syntagmatic surface form variation is modeled in terms of transformations or sets of constructional variants (Harris, 1957, Hiz, 1964, Cullicover, 1968, Smaby, 1971)
- Formally, a paraphrase is a relation between two lexical, phrasal, or sentential expressions, E_i and E_j, where meaning is preserved (Smaby, 1971).
- Machine Translation adopted rule-based paraphrasing in the 1980s
- Statistical MT adopted it in 2013 (Bhagat and Hovy, 2013)
- Neural MT and QA exploit it (Weston et al, 2021)

Type-driven Dense Paraphrasing

Pustejovsky (1995)

(1) a. Mary likes to watch movies.
    b. Mary likes watching movies.
    c. Mary likes movies.
    d. Mary likes (for) John to watch movies with her.
    e. Mary likes that John watches movies with her.
    f. Mary likes it that John watches movies with her.

(2) a. Mary enjoys watching movies.
    b. Mary enjoys movies.

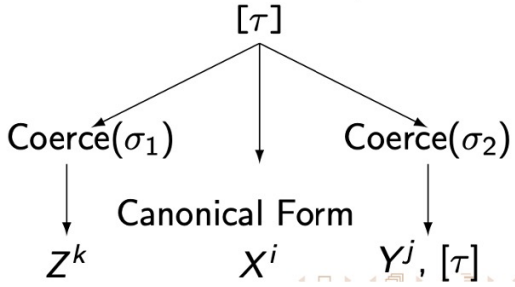
Canonical Syntactic Form – Structural Motifs

For any semantic type, τ, there is a unique *canonical syntactic form (csf)* that expresses this type as a syntactic object, X^i.

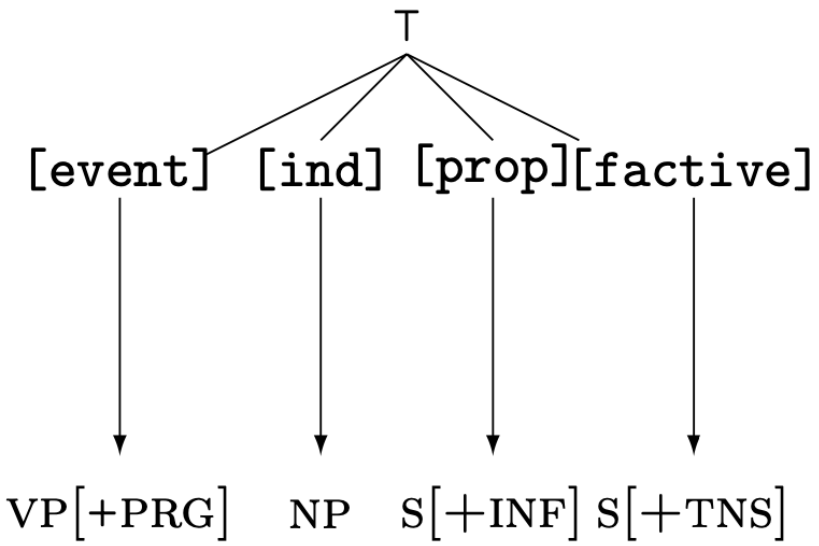
(3) For every type τ in the set of semantic types, there is a function, *canonical syntactic form (csf)*, such that $csf(\tau) = X^i$, except for
   a. when $\tau = \top$, or
   b. when $\tau = \bot$,
   in which case *csf* is undefined.

Y^j,where $csf(\sigma) = Y^j$, is substitutable for the *csf* of a type τ only if this type is fully recoverable from licensed semantic operations on σ.

(4)



CSF (Motifs) for these Semantic Types



Decontextualization

Choi et al (2021)

Decontextualization: Given a sentence-context pair (S, C), a sentence S' is a valid decontextualization of s if: (1) the sentence S' is interpretable in the empty context; and (2) the truth-conditional meaning of S' in the empty context is the same as the truth-conditional meaning of S in context C.

- Focuses on enriching text through anaphora resolution and knowledge base augmentation, such as Wikipedia.

Dense Paraphrasing

Pustejovsky et al (2021), Tu et al (2022)

Dense Paraphrasing: Given the pair, (S, P), where S is a source expression, and P is an expression, we say P is a valid *Dense Paraphrase* of S if: P is an expression (lexeme, phrase, sentence) that eliminates any contextual ambiguity that may be present in S, but that also makes explicit any underlying semantics that is not otherwise expressed in the economy of sentence structure, e.g., default or hidden arguments, dropped objects or adjuncts. P is both meaning preserving (consistent) and ampliative (informative) with respect to S.

Frame Saturation

Frame Saturation: recovering all logical hidden arguments to a predicate or function.

- **Drop argument**: A drop argument is an argument to a predicate that has been elided or left unexpressed in the syntax. Such elisions occur when the antecedent has been mentioned in a previous sentence and can be recovered from the context in the document.
- **Shadow argument**: A shadow argument is semantically incorporated in the meaning of the event predicate itself; e.g., an implicit tool or ingredient that is not mentioned but presupposed

Placing

Lexical Unit Index

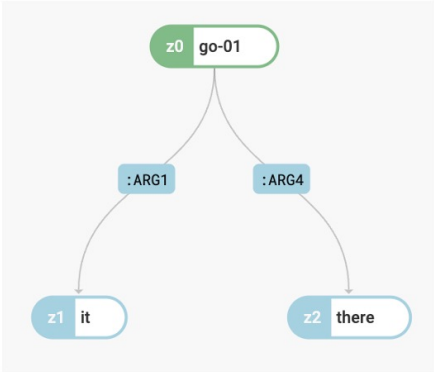
Definition:

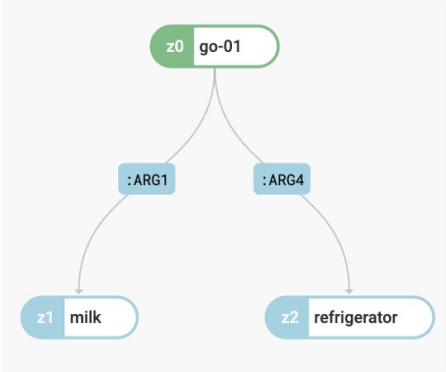
Generally without overall (translational) motion, an Agent places a Theme at a location, the Goal, which is profiled. In this frame, the Theme is under the control of the Agent/Cause at the time of its arrival at the Goal.
David PLACED his briefcase on the floor.
This frame differs from Filling in that it focuses on the Theme rather than the effect on the Goal entity. It differs from Removing in focusing on the Goal rather than the Source of motion for the Theme.

Frame Saturation – filling in missing roles

- Drop Arguments
  - *Combine sugar and water. Mix [. . .] until dissolved.*
  - $\Longrightarrow$ *Mix sugar and water until dissolved.*
  - *Chop the onion. Sauté [. . .] until browned.*
  - $\Longrightarrow$ *Sauté the onion until browned.*
- Shadow Arguments
  - *Stir [. . .] until firm.*
  - $\Longrightarrow$ *Stir with a spoon until firm.*
  - *Bake [. . .] at 350.*
  - $\Longrightarrow$ *Bake in an oven at 350.*

Dense Paraphrasing with situated grounding





The milk goes into the refrigerator.

Dense Paraphrasing

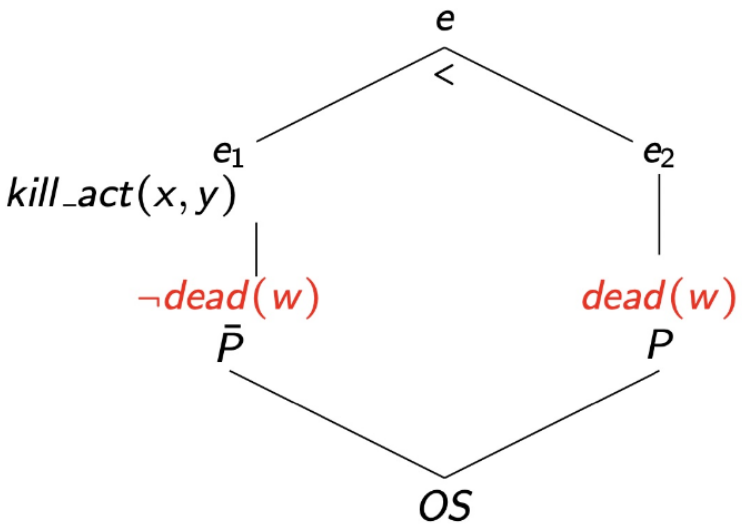
Subevent structure

- Expose the subevent structure of the lexical predicate
  - *arrive* is textually expressed as two subevents
  - *not_at_loc(x,y) and at_loc(x,y)*
- Dynamic tracking of event consequences
  - Recovering Entity Properties from Event Structure
  - *chop* applied to *onion* brings about *chopped onions*

GL Event Structure

Im and Pustejovsky (2010)

- Building Subevent Structures from Text



(10) **kill** in *John killed the plant*

se1: pre-state: not_dead(plant)

se2: process: killing(john,plant)

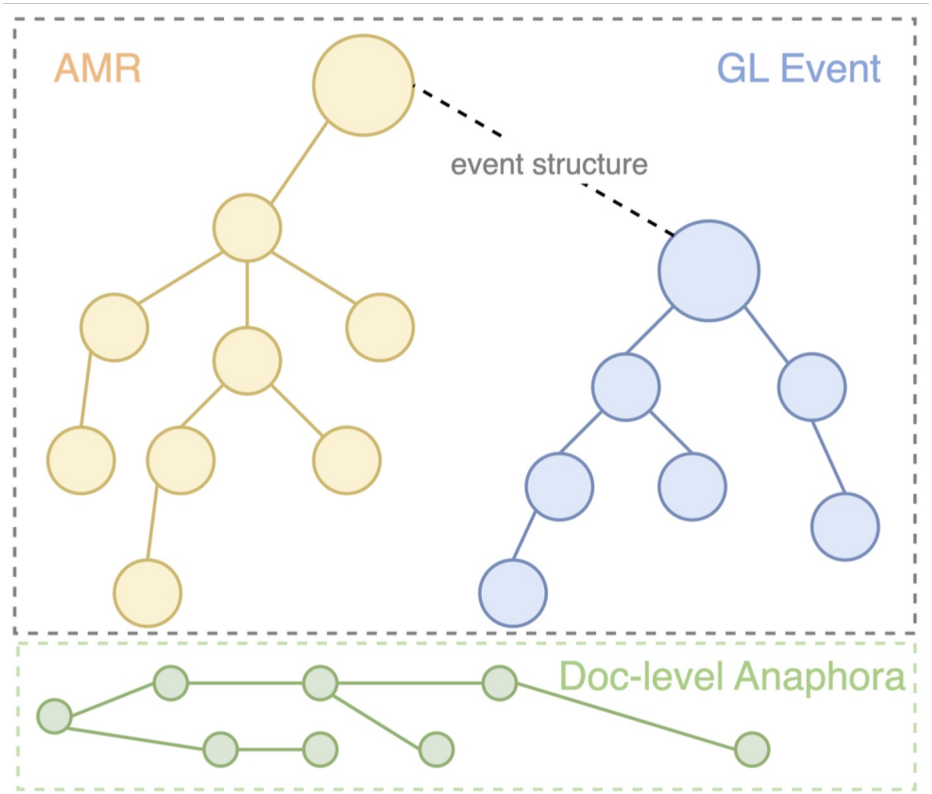
se3: post-state: dead(plant)

Generative Lexicon AMR (GLAMR)

- *Tu et al (2024) COLING-LREC*

- A new semantic representation extending AMR with Generative Lexicon event structure
- Propose a pipeline for automatic augmentation of AMR to GLAMR graphs
- Create a GLAMR dataset from procedural texts, e.g., cooking recipes
- Evaluate with baselines for converting text to GLAMR and GLAMR to text

Generic GLAMR graph



Background

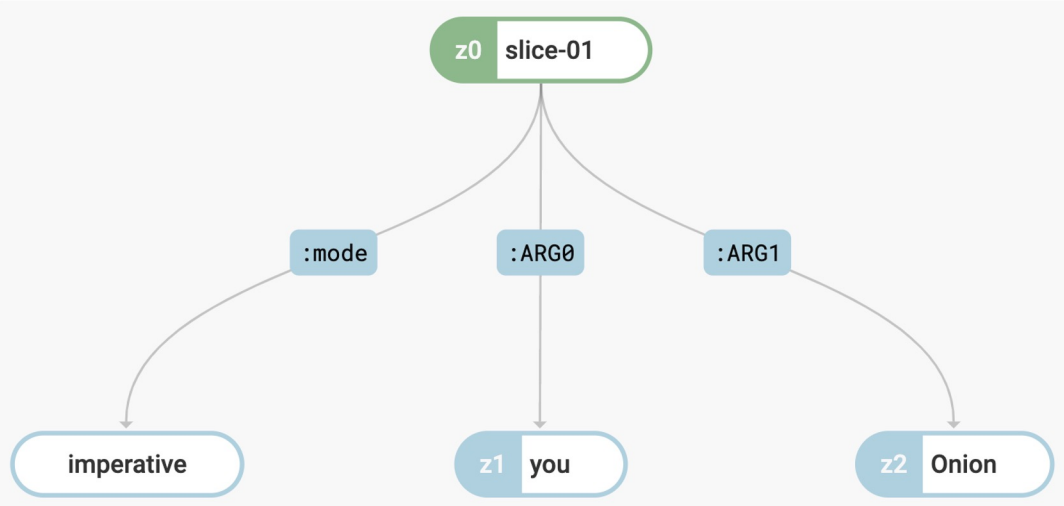
Abstract Meaning Representation (AMR)

- A semantic meaning representation that can encode the meaning of the texts in a structured way
- Able to go beyond sentences (e.g., DocAMR, UMR, etc)
- Flexible to be extended with other semantic information (e.g., dialogue, gesture, action, etc)

Background

Abstract Meaning Representation (AMR)

Sentence: *Slice the onion.*




```
(z0 / slice-01
    :mode imperative
    :ARG0 (z1 / you)
    :ARG1 (z2 / Onion))
```

Generated from https://nlp.uniroma1.it/spring/.

Background

Generative Lexicon - VerbNet (GL-VN)

- VerbNet provides semantic representations for a wide coverage of verb classes
- GL-VerbNet updates VN with representations for the GL event structure

Background

Generative Lexicon - VerbNet (GL-VN)

GL event structure of the VN class *pour-9.5*

- NP V NP PP.destination
- NP V NP ADVP
- NP V PP.destination
- NP V NP PP.initial_location PP.destination
- NP V PP.initial_location PP.destination

EXAMPLE:

Tamara poured water into the bowl.

SHOW DEPENDENCY PARSE TREE

SYNTAX:

Agent VERB Theme { PREP } Destination

SEMANTICS:

HAS_LOCATION(e1 , Theme , ?Initial_Location)
DO(e2 , Agent)
MOTION(ë3 , Theme , Trajectory)
¬ HAS_LOCATION(ë3 , Theme , ?Initial_Location)
CAUSE(e2 , ë3)
HAS_LOCATION(e4 , Theme , Destination)

FORCE DYNAMICS:

Volitional Apply FD representation

Background

Coreference under Transformation Labeling (CUTL) Dataset

- Contain the annotations of entities, their anaphoric and coreference relations, and the accompanying event semantics on the cooking recipes
- Annotate each event as an I/O process with the explicit and implicit arguments, as well as the anaphoric relations between the entities

Mapping from GL-VN to GLAMR

- `:event-structure` links the predicate to the root of subevents as the direct child of the predicate
- GL event structure is portable that can be added to or detached from original AMR graphs

***Pour** them into the bowl.*

```
(p / pour-01
  :ARG0 (y / you)
  :ARG1 (t / them)
  :ARG3 (b / bowl)
  :event-structure (s / subevents
    :E0 + (d / do
      :ACTION p)
    :E1 § (h / has_location
      :THEME t ‡
      :INITIAL_LOC ‡ N/A)
    :E3 (a / and
      :op1 ¶ (m / motion
        :THEME t
        :TRAJECTORY N/A)
      :op2 (h / has_location
        :polarity - ¶
        :THEME t
        :INITIAL_LOC N/A))
    :E4 (h1 / has_location
      :THEME t
      :DESTINATION b)
  :mode imperative )
```

Mapping from GL-VN to GLAMR

- New AMR roles `E1, E2, . . .` are added to represent the subevent indices
- The indices are aligned with the GL event structure of the predicate encoded in VN

***Pour** them into the bowl.*

```
(p / pour-01
  :ARG0 (y / you)
  :ARG1 (t / them)
  :ARG3 (b / bowl)
  :event-structure (s / subevents
    :E0 † (d / do
      :ACTION p)
    :E1 § (h / has_location
      :THEME t ‡
      :INITIAL_LOC ‡ N/A)
    :E3 (a / and
      :op1 ¶ (m / motion
        :THEME t
        :TRAJECTORY N/A)
      :op2 (h / has_location
        :polarity - ¶
        :THEME t
        :INITIAL_LOC N/A))
    :E4 (h1 / has_location
      :THEME t
      :DESTINATION b)
  :mode imperative )
```

Mapping from GL-VN to GLAMR

- The concepts and variables inside the subevent are synced with the outside through reentrance

***Pour** them into the bowl.*

```
(p / pour-01
  :ARG0 (y / you)
  :ARG1 (t / them)
  :ARG3 (b / bowl)
  :event-structure (s / subevents
    :E0 † (d / do
      :ACTION p)
    :E1 § (h / has_location
      :THEME t ‡
      :INITIAL_LOC ‡ N/A)
    :E3 (a / and
      :op1 ¶ (m / motion
        :THEME t
        :TRAJECTORY N/A)
      :op2 (h / has_location
        :polarity - ¶
        :THEME t
        :INITIAL_LOC N/A))
    :E4 (h1 / has_location
      :THEME t
      :DESTINATION b)
  :mode imperative )
```

Mapping from GL-VN to GLAMR

- `:ACTION` represent the action that has been performed on the objects during the event time
- The concept is the verb lemma of the predicate

***Pour** them into the bowl.*

```
(p / pour-01
  :ARG0 (y / you)
  :ARG1 (t / them)
  :ARG3 (b / bowl)
  :event-structure (s / subevents
    :E0 † (d / do
      :ACTION p)
    :E1 § (h / has_location
      :THEME t ‡
      :INITIAL_LOC ‡ N/A)
    :E3 (a / and
      :op1 ¶ (m / motion
        :THEME t
        :TRAJECTORY N/A)
      :op2 (h / has_location
        :polarity - ¶
        :THEME t
        :INITIAL_LOC N/A))
    :E4 (h1 / has_location
      :THEME t
      :DESTINATION b)
  :mode imperative )
```

Mapping from GL-VN to GLAMR

- Subevents with the same temporal index are stacked with the `:op` roles
- Negation is represented with the attribute `:polarity`

***Pour** them into the bowl.*

```
(p / pour-01
  :ARG0 (y / you)
  :ARG1 (t / them)
  :ARG3 (b / bowl)
  :event-structure (s / subevents
    :E0 † (d / do
      :ACTION p)
    :E1 § (h / has_location
      :THEME t ‡
      :INITIAL_LOC ‡ N/A)
    :E3 (a / and
      :op1 ¶ (m / motion
        :THEME t
        :TRAJECTORY N/A)
      :op2 (h / has_location
        :polarity - ¶
        :THEME t
        :INITIAL_LOC N/A))
    :E4 (h1 / has_location
      :THEME t
      :DESTINATION b)
  :mode imperative )
```

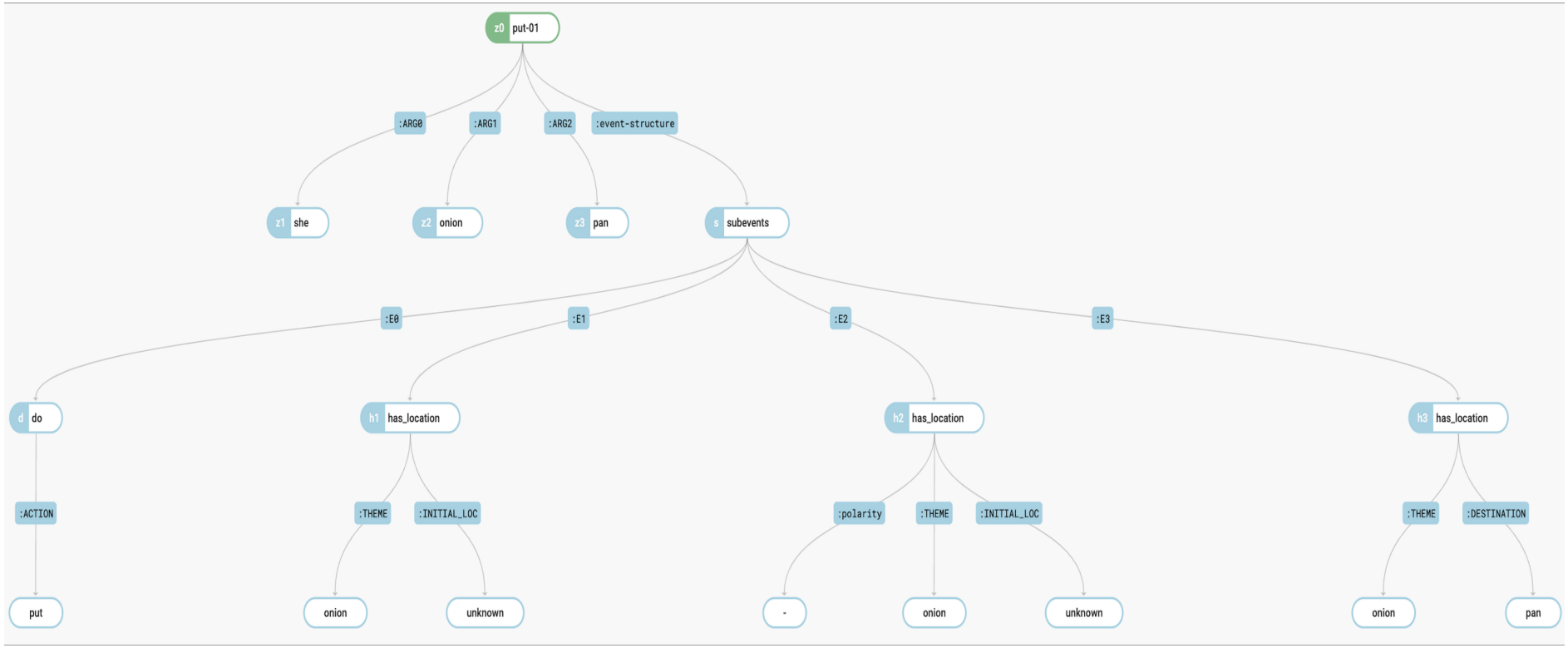
GLAMR Dataset

Subevent frequency

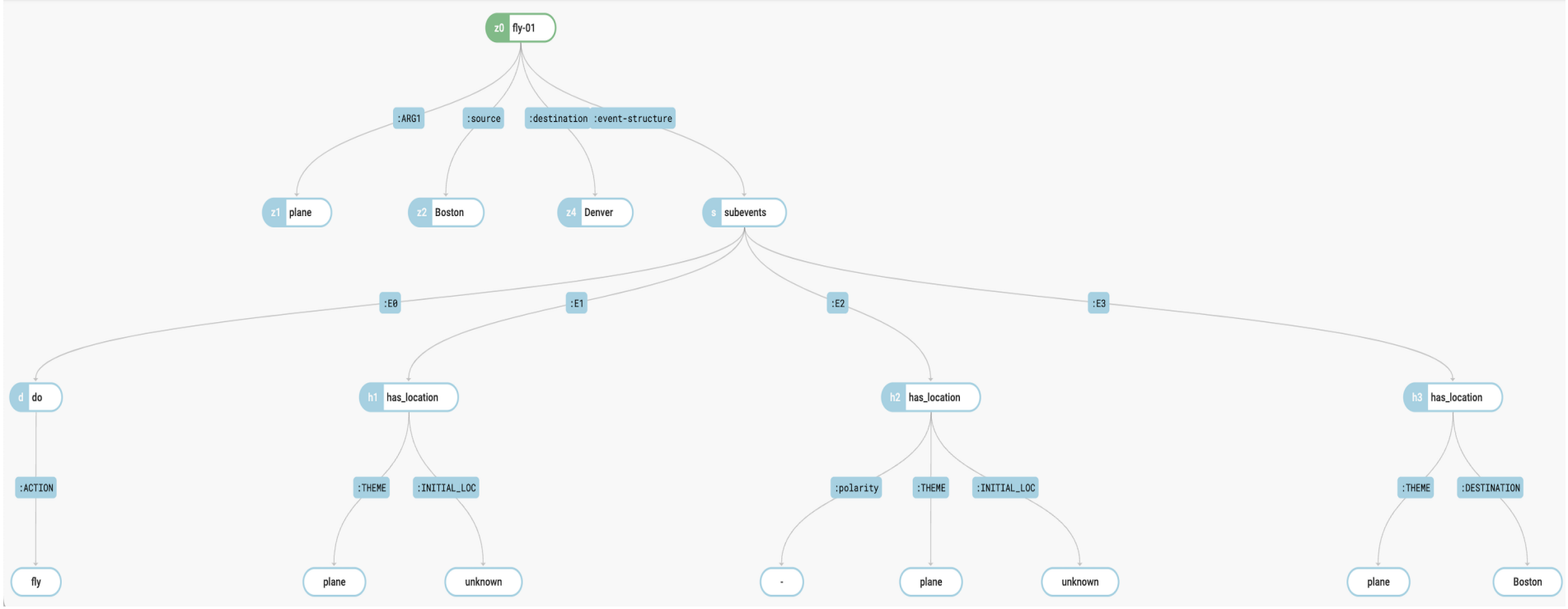
Sube. Names	Count	Sube. Roles	Count
[¬]has_loc.	636 (26%)	Patient	2038 (38%)
[¬]cooked	286 (11%)	Theme	1131 (21%)
[¬]MI_state	274 (11%)	V_Final_State	378 (7%)
[¬]together	212 (9%)	Initial_Loc.	371 (7%)
motion	211 (9%)	V_State	274 (5%)

GLAMR Event Enrichment

- Put the onions in the pan.



GLAMR Event Enrichment

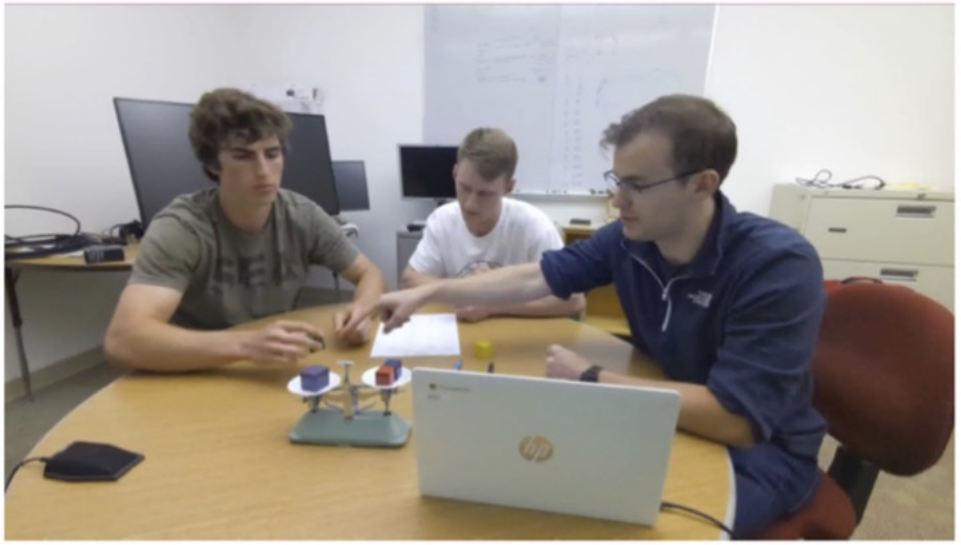


Multi-Modal Dense Paraphrasing

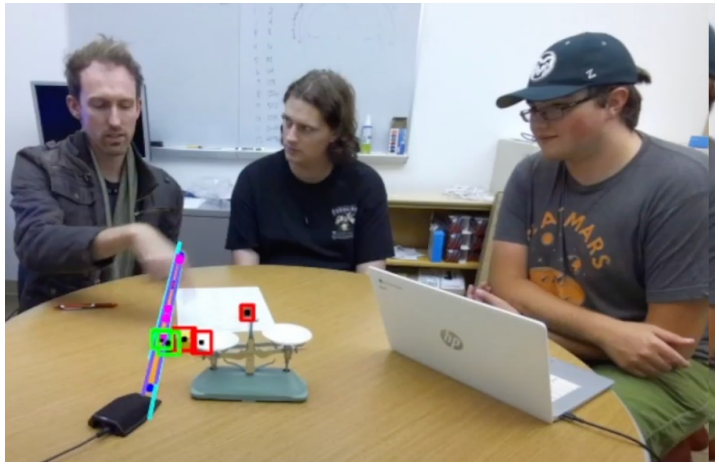
- Extend the DP to encode the **multimodal input** into **Machine Readable Paraphrases (MRP)**
- Apply LLMs to decode MRP into **Human Readable Paraphrases** for downstream tasks
- HRP encodes the potential **non-verbal information** and **situated grounding** from the interaction between participants and the objects
- Apply MMDP on the **Weights Task Dataset** for the common ground tracking problem

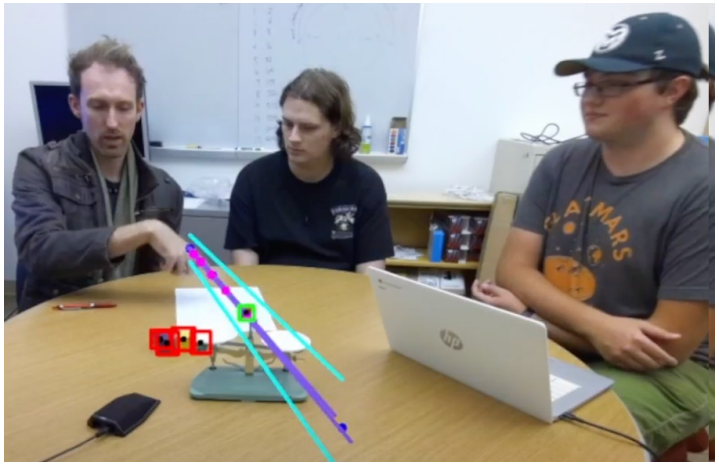
Weights Task Dataset

- Contains ten videos, in which groups of three were asked to determine the weights of five blocks using a balance scale
- Participants communicated with each other using multiple modalities, including language, gesture, gaze, and action
- Contains common ground annotation on the dialogue where participants reach the common grounds (agree on the statements on the weights of the tasks)
- Multimodal interaction
  - Speech
  - Gesture
  - Gaze
  - Action
  - Posture



Situated Multimodal Coreference





- Situated Grounding for Coreference
  - *It weighs 10 grams.* ⇒ *The red block* weighs 10 grams.
  - Point_Left$_{Gesture}$ ⇒ *The yellow block*
- Epistemic Framing: Express the epistemic attitude towards a sentence or action;
  - *Speaker A: It weighs 10 grams.* ⇒ *I think that* it weighs 10 grams.
  - Mary$_{AGENT}$ put the block on the scale. ⇒ *Mary believes that she put the block on the scale.*

Dense Paraphrasing through Multimodal Alignment

Original Utterance

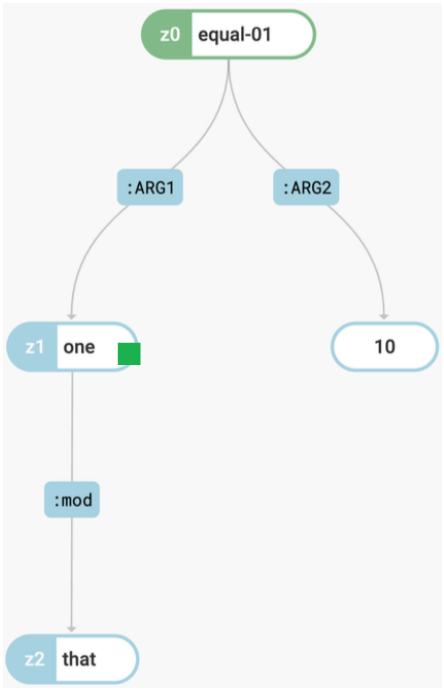


Figure: AMR of: That one is 10.

Dense Paraphrase

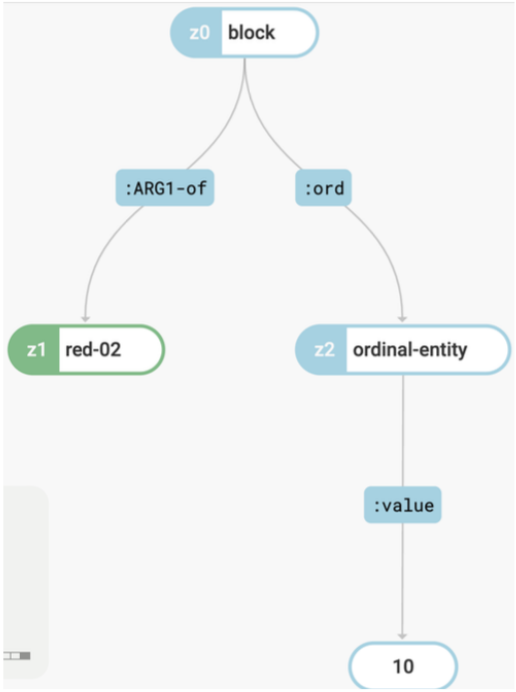
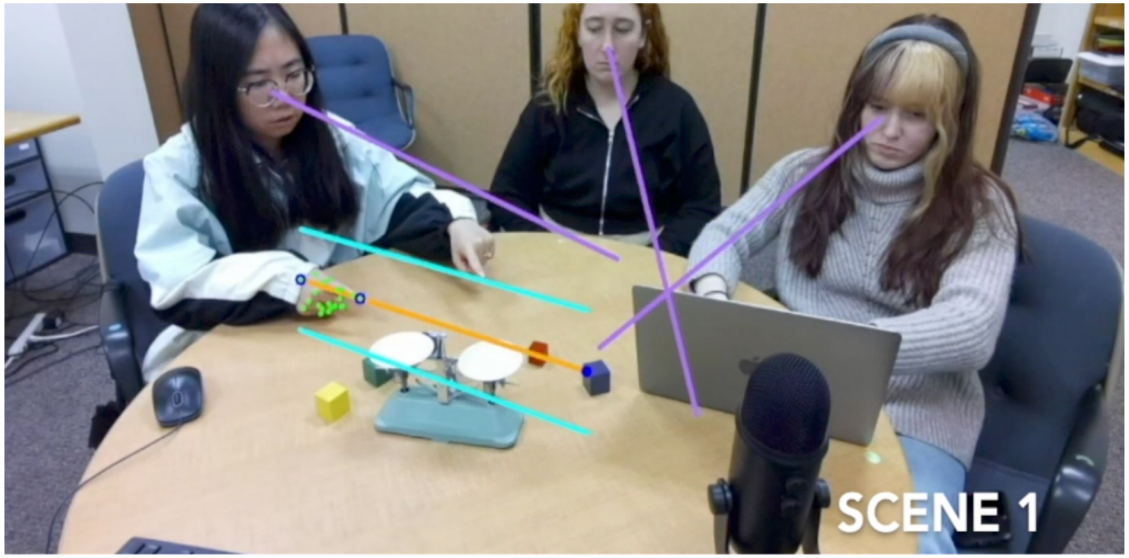


Figure: AMR of: Red block is 10.

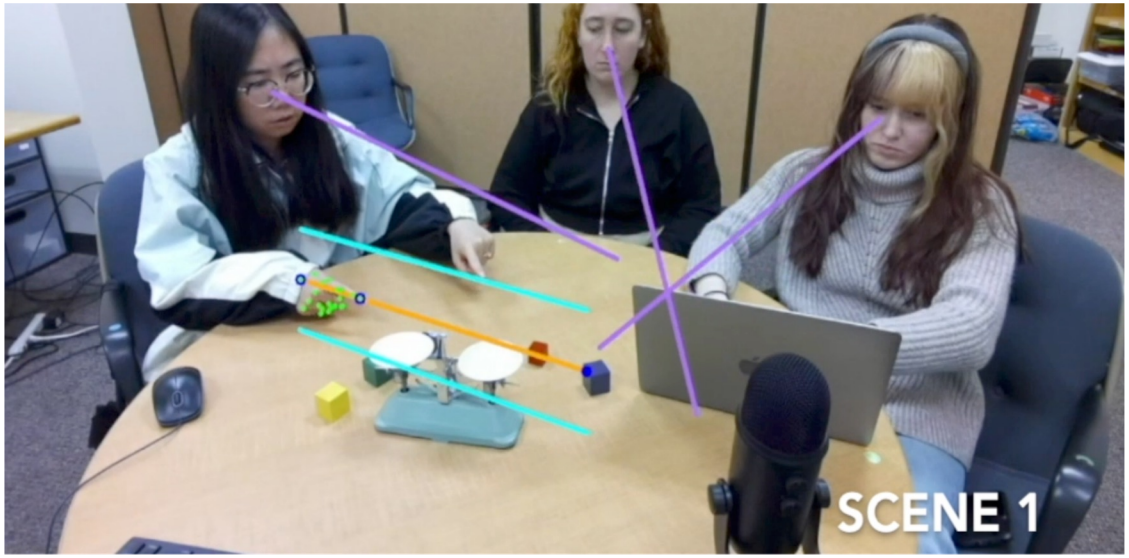
Encodings from the MMDP

- Utterance: *“Try this one”*
- Aligned Video frame with gesture, gaze, and speech



Encodings from the MMDP

- Utterance: *"Try this one"*
- Aligned Video frame with gesture, gaze, and speech



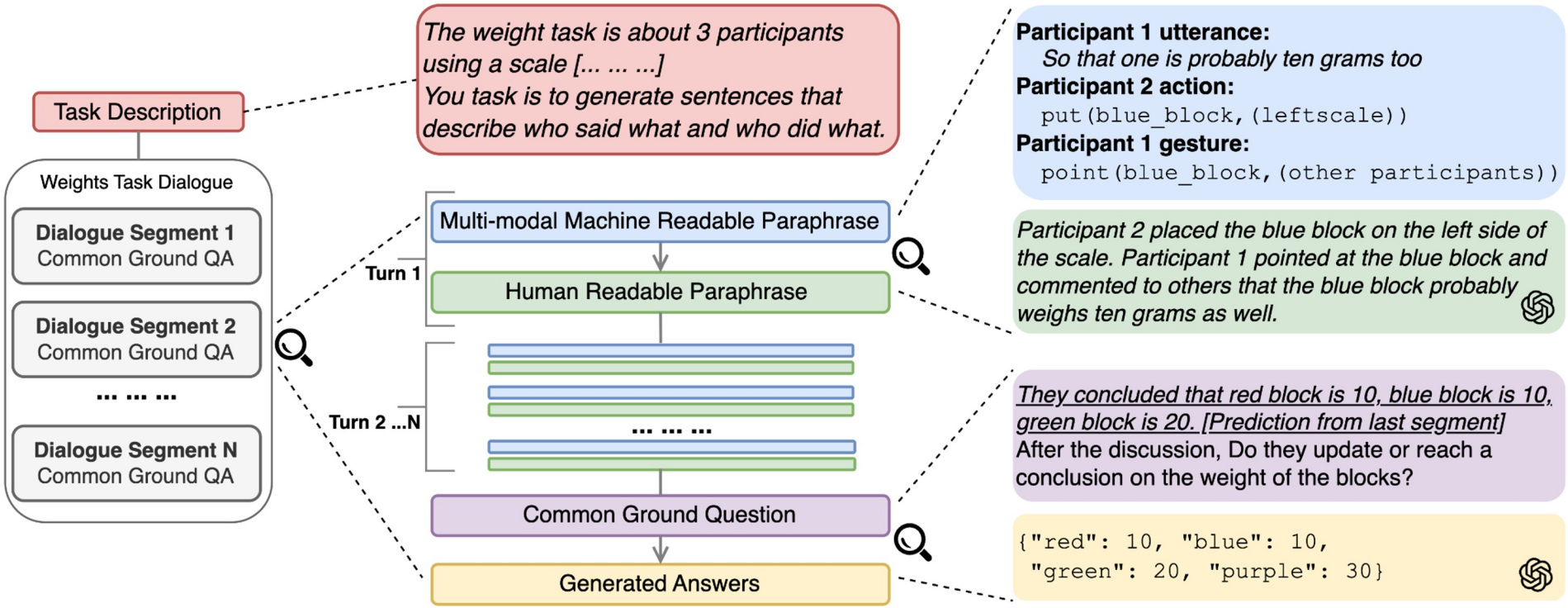
MRP

```
{P1-utterance: try this one,
 P1-gesture: point
(blue_block,others)}
```

HRP

Participant 1 pointed at the blue block and commented to others that they should try the blue block.

MMDP on Common Ground Tracking

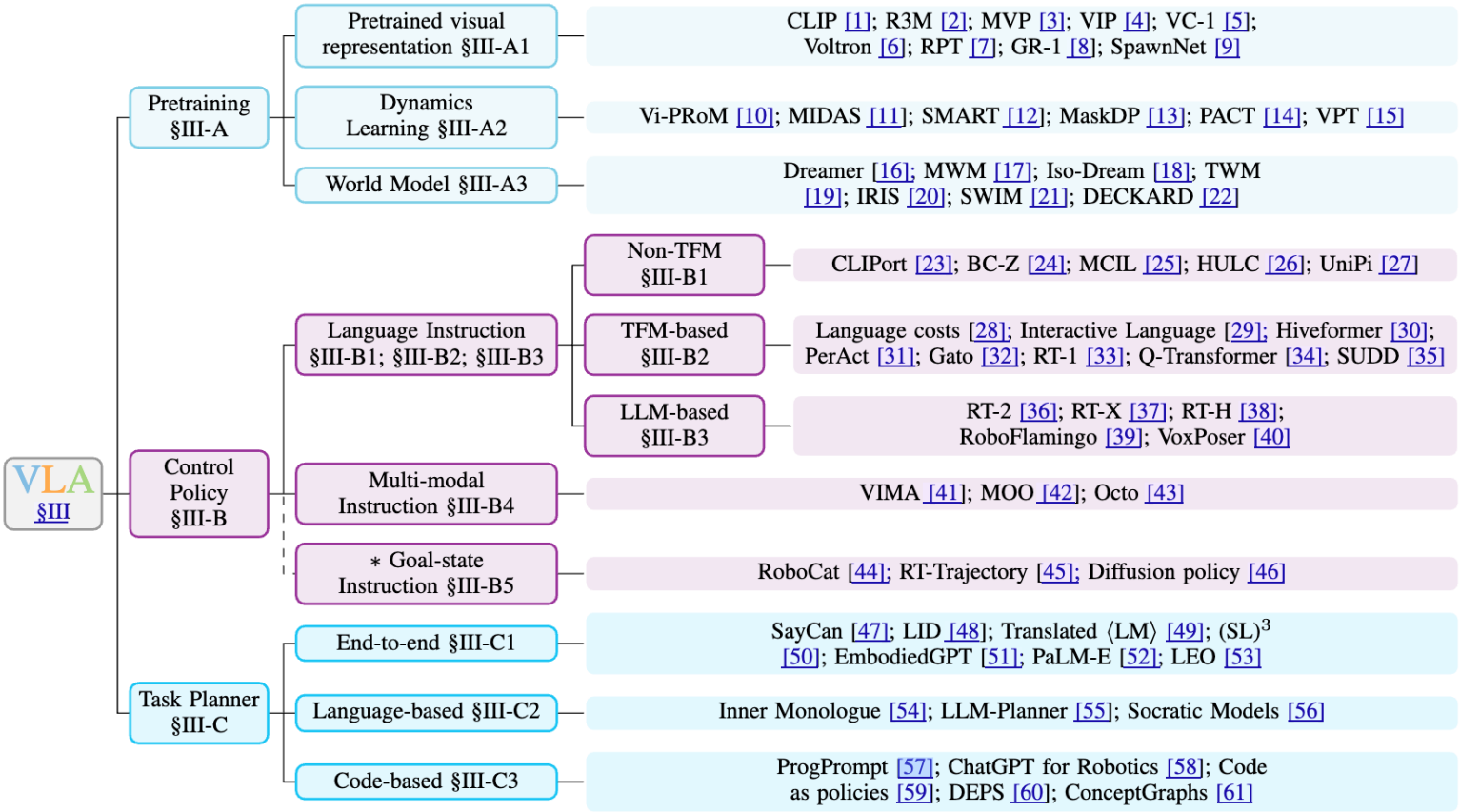


Multimodal Large Language Models (MLLMs)

- KOSMOS-2 :
  - Combines multiple modes of input and output, users can use their voice to give commands, gestures to navigate through menus, touch to interact with virtual objects, and gaze to control certain functions.
  - *Peng, Zhiliang, et al. "Kosmos-2: Grounding multimodal large language models to the world." arXiv preprint arXiv:2306.14824 (2023)*.
- GLaMM: Pixel Grounding Large Multimodal Model.
  - Key feature is pixel grounding, which involves associating specific pixels in an image with their corresponding textual concepts.
  - Global image encoder, a region encoder, a language-to-language model, a grounding image encoder, and a pixel decoder.
  - *Rasheed et al (2024). Glamm: Pixel grounding large multimodal model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

Vision Language Action Models (VLA)

- Coined by RT-2 (Brohan et al, 2023) – focuses on embodied AI for human-robot interaction
- Taxonomy of VLA Models



Vision Language Action Models (VLA)

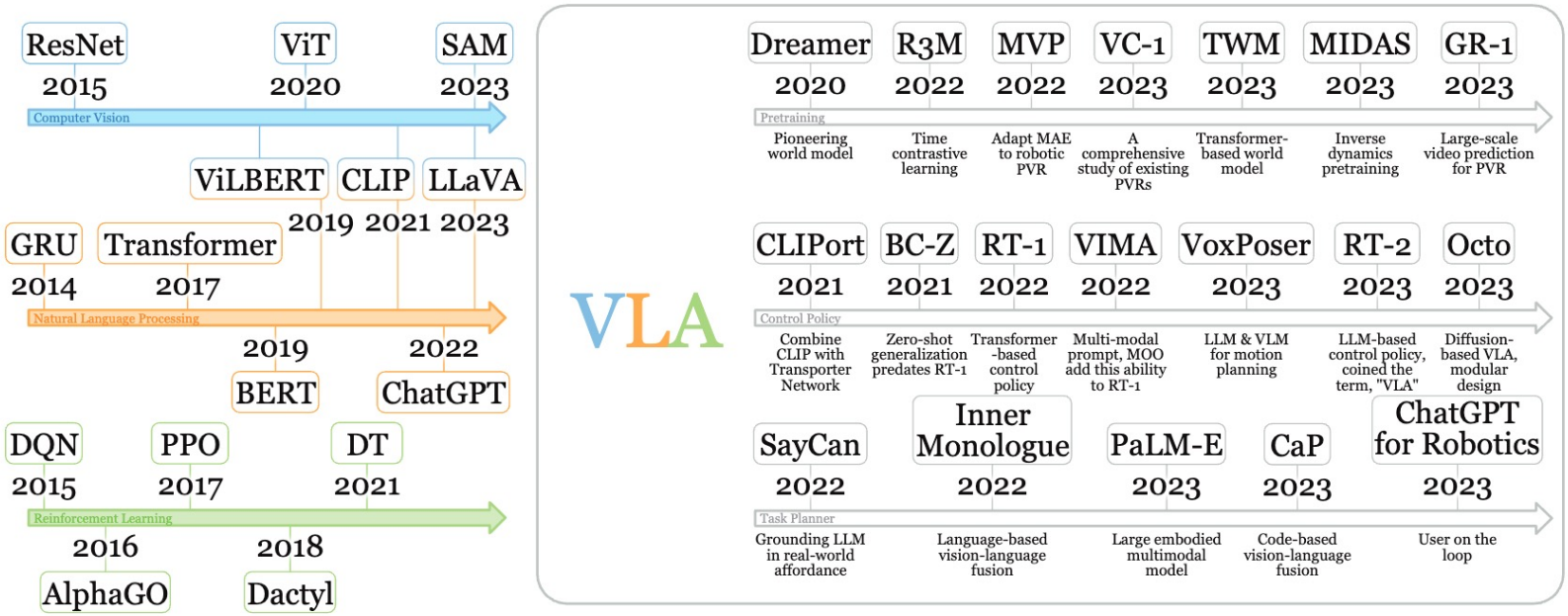


Figure 2: A brief timeline traces the evolution from unimodal models to multimodal models, laying the groundwork for the introduction of VLA models. Key advancements in computer vision (blue) include ResNet [85], ViT [86], and SAM [87]. Seminal works in natural language processing (orange) encompass GRU [88], Transformer [66], BERT [89], ChatGPT [62], etc. Reinforcement learning (green) has seen notable contributions from DQN [90], AlphaGo [91], PPO [92], Dactyl [93], and DT [94]. Vision-language models have emerged as a critical category of multimodal models, exemplified by ViLBERT [95], CLIP [1], and LLaVA [96]. The three main directions in VLA are: pretraining, control policy, and task planner.

Hierarchical Robot Policy

- Reinforcement learning has seen a shift towards employing Transformers to model the Markov Decision Process as autoregressive sequential data.

Figure 3: Illustration of a hierarchical robot policy comprising a high-level task planner and a low-level control policy. The high-level task planner generates a plan based on the user instruction, which is then executed step by step by the low-level control policy.

VLA Architectures

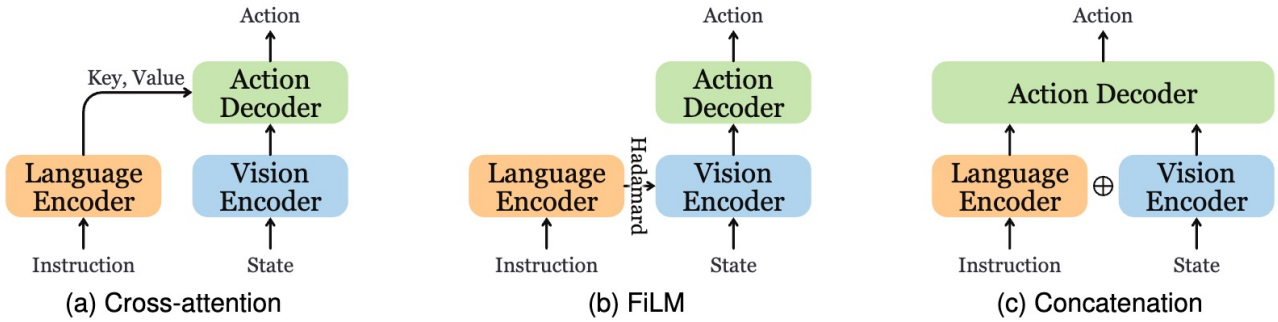


Figure 4: The three most common architectures of low-level control policies are characterized by their vision-language fusion methods. Some Transformer action decoders utilize cross-attention to condition on the instruction. FiLM layers are employed to fuse language and vision early in RT-1-based models. Concatenation is the prevailing method of vision-language fusion for Transformer action decoders.

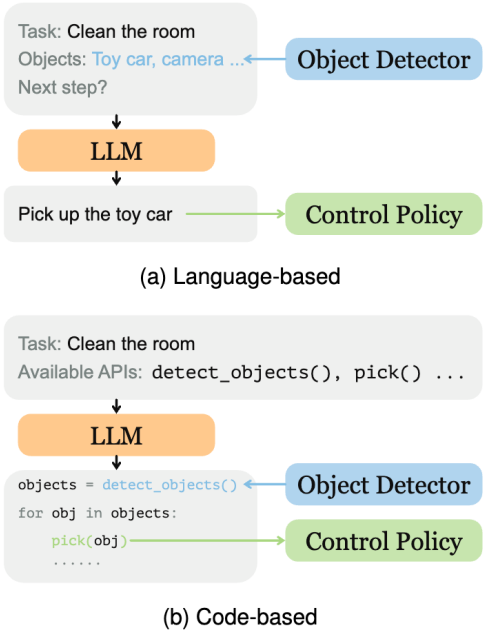
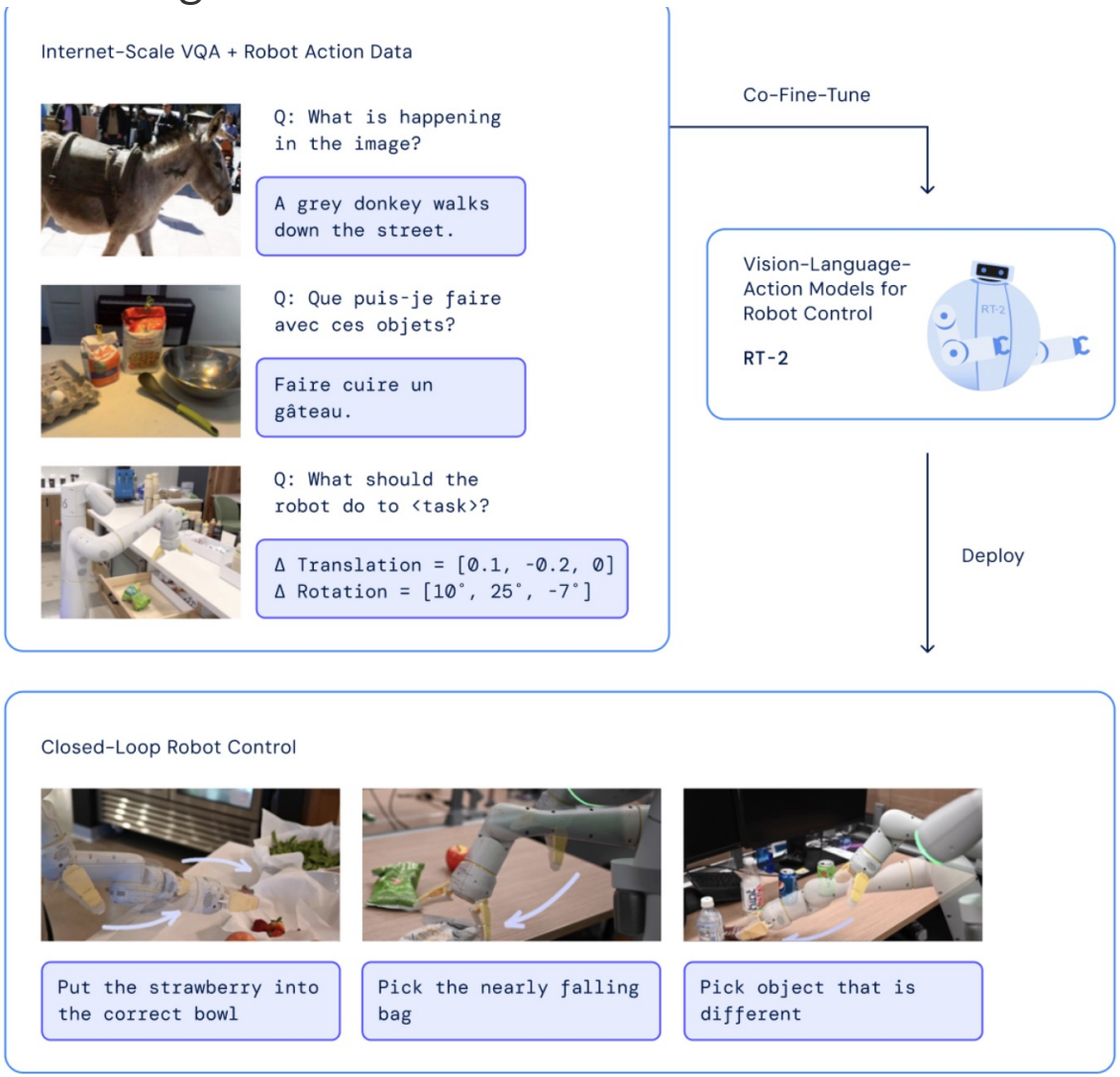


Figure 5: Different approaches to connect LLM to multi-modal modules in high-level task planner.

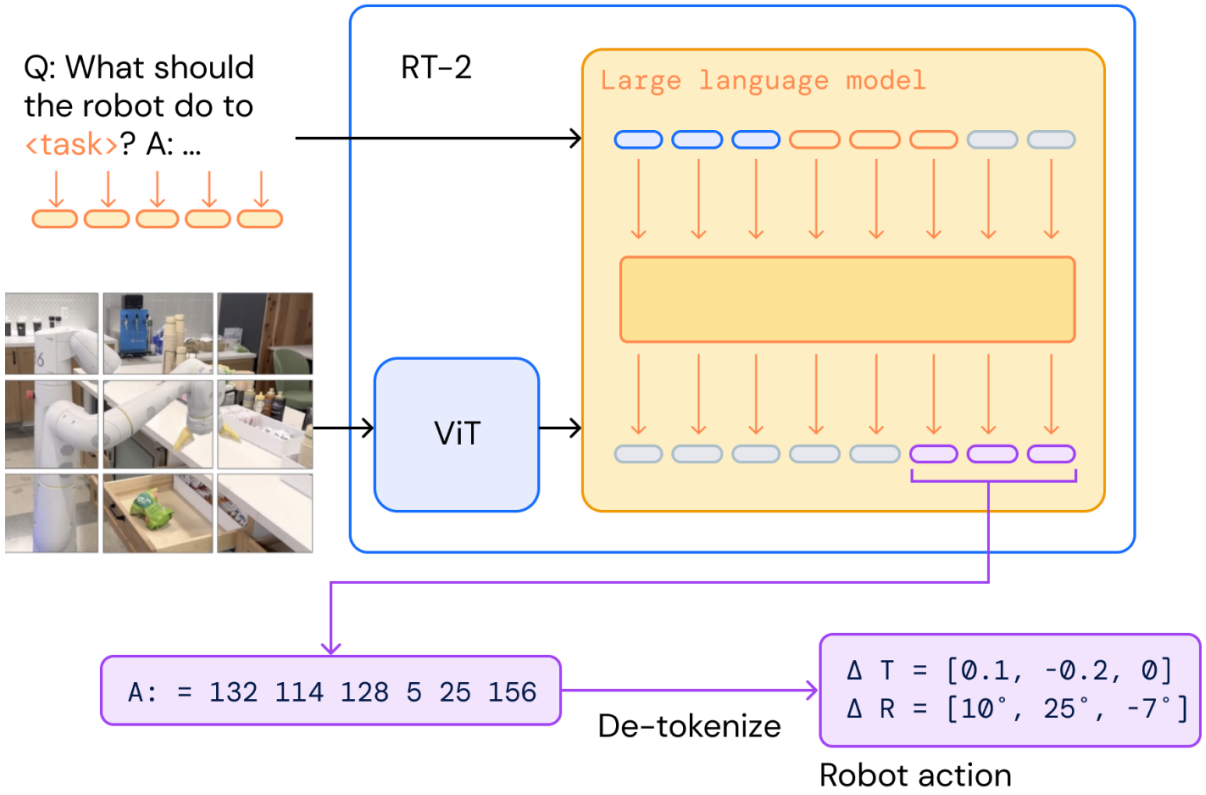
RT-2 – Google DeepMind

- Represent robot actions as another language, which can be cast into text tokens and trained together with Internet-scale vision-language datasets.



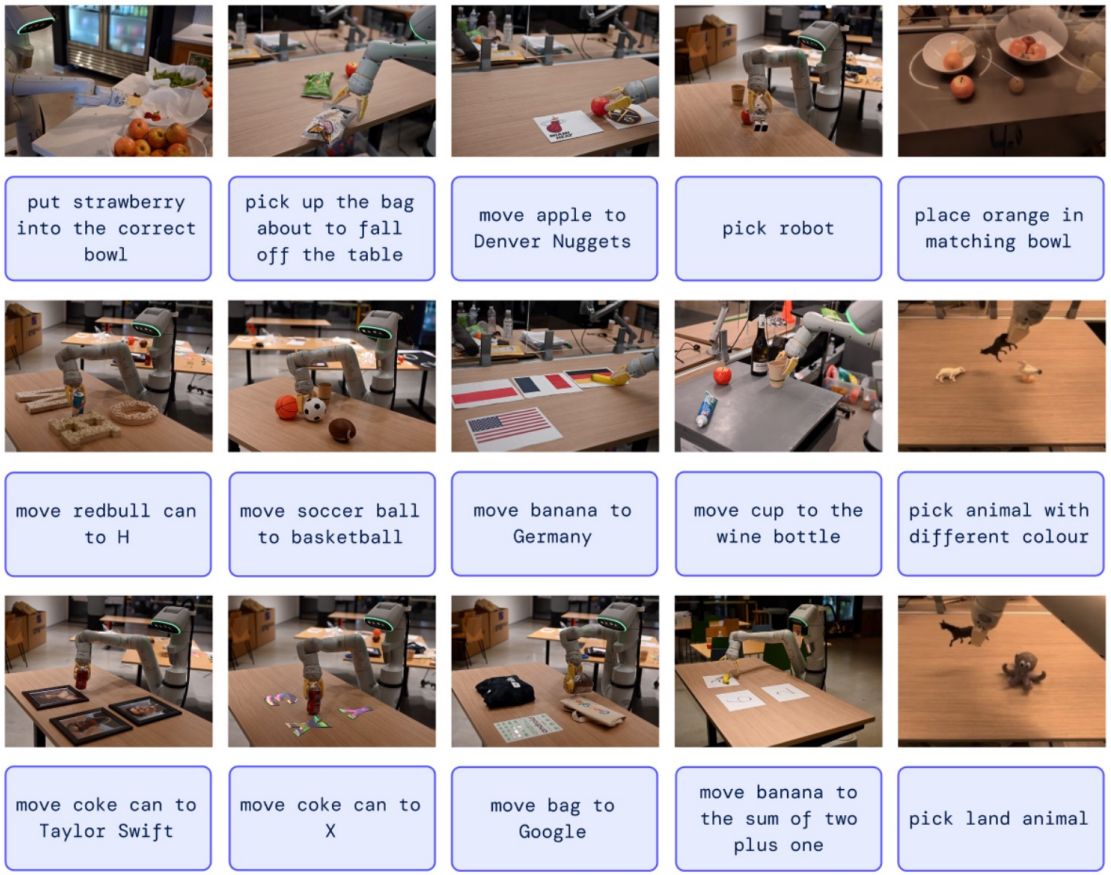
RT-2 – Google DeepMind

- During inference, the text tokens are de-tokenized into robot actions, enabling closed loop control. This allows us to leverage the backbone and pretraining of vision-language models in learning robotic policies, transferring some of their generalization, semantic understanding, and reasoning to robotic control.



RT-2 – Google DeepMind

- Each task required understanding visual-semantic concepts and the ability to perform robotic control to operate on these concepts. Commands such as “pick up the bag about to fall off the table” or “move banana to the sum of two plus one” – where the robot is asked to perform a manipulation task on objects or scenarios never seen in the robotic data – required knowledge translated from web-based data to operate.



RT-2 – Google DeepMind

- Affordance-like behavior is adaptive and transferable.

Situated Grounding -Future Research

- Integration of multimodal datasets, together with uniform encoding as textual form (linguistic dense paraphrasing) promises to provide additional training data for new modalities and context:
  - Situational dialogue variables
  - Environmental states
  - Epistemic states of agents
  - Other common ground knowledge