Grounding Meaning Representation for Situated Reasoning

Nikhil Krishnaswamy and James Pustejovsky

AACL-IJCNLP 2022 Tutorial

Taipei, Taiwan

November 20, 2022

The Big Picture Goal

- NLP advances have driven the growth and mainstreaming of AI:
  - Once brittle, now everywhere.
- Text-based language models demonstrate impressive performance in **tasks** like inference, QA, knowledge extraction, etc.
- Transformers have been applied to vision tasks.
- NLP/AI is still *task specific*.

The Big Picture Goal

- Multimodal tasks have pushed the boundaries of relating language to visuals.
- Inherent in this is an understanding of how entities relate to *situations* and how humans relate to the entities.
- A wealth of modalities is implicated (language, gesture, gaze, posture, facial expressions).

The Big Picture Goal

- As AI becomes more multimodal, we need to understand the role of affordances and human-object interactions in reasoning.
- We examine how our knowledge of object interactions is rarely reflected in linguistic descriptions of actions (or images).
- We demonstrate how object and situational conditions on actions need to be identified and encoded, just as importantly as the actions themselves.
- Identifying and encoding situated conditions requires grounding deep semantic representations to the environment.

Tutorial Learning Goals

- Identify requirements involved in developing a deep 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
- Developing the notion of embodiment: an agent's actions are *situated* by the context and constrained by its *embodiment* in the context.
  - embodiment in a simulation
  - an embodied environment allows us to bridge and integrate formal symbolic and statistically oriented reasoning approaches, to create grounding and situated reasoning.

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.

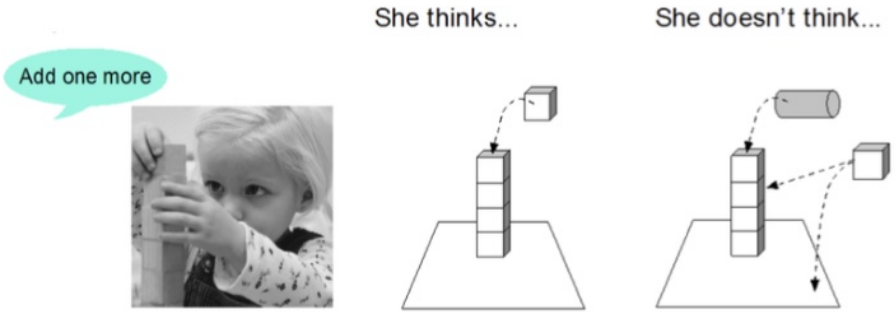
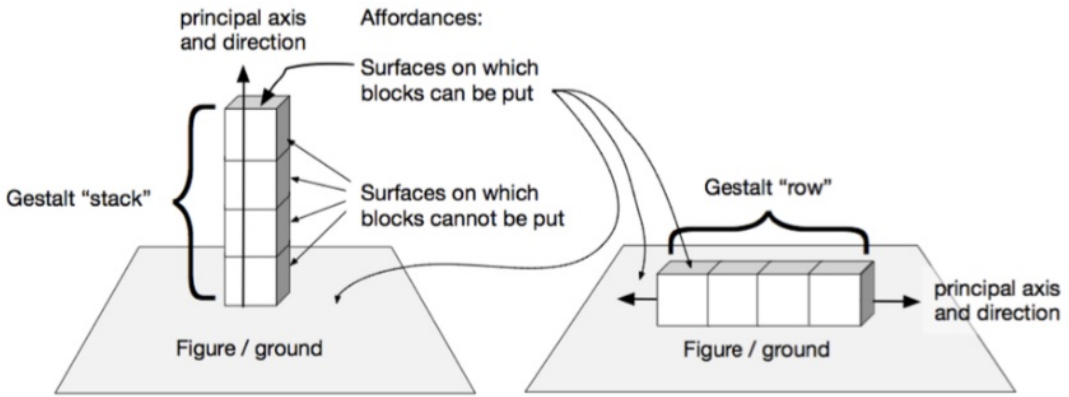


Figure 2. "Add one more" is ambiguous out of context, but given context it is remarkably precise.



Motivating Example


Figure: "Put on your bee suit."

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: $\text{TELIC} = \lambda z, e[\mathit{beekeeping}(e, z, x)]$
  - An outfit *resembling* a bee?
    - Generative Lexicon: $\text{FORMAL} = \mathit{bee}(x)$

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.

Tutorial Outline

- Approaches to Multimodal Grounding
- Meaning Representation Frameworks
- Communicating with Multimodal Common Ground
- Reasoning with and about Affordances

Challenges to Situated Grounding

Implicit spatial semantics

Put the milk in the coffee vs. Put the milk in the refrigerator

Fly a kite vs. Carry a kite

Approaches to Multimodal Grounding

- VQA: Visual Question Answering (Antol et al., 2015)
- Combines common-sense knowledge with visual interpretation

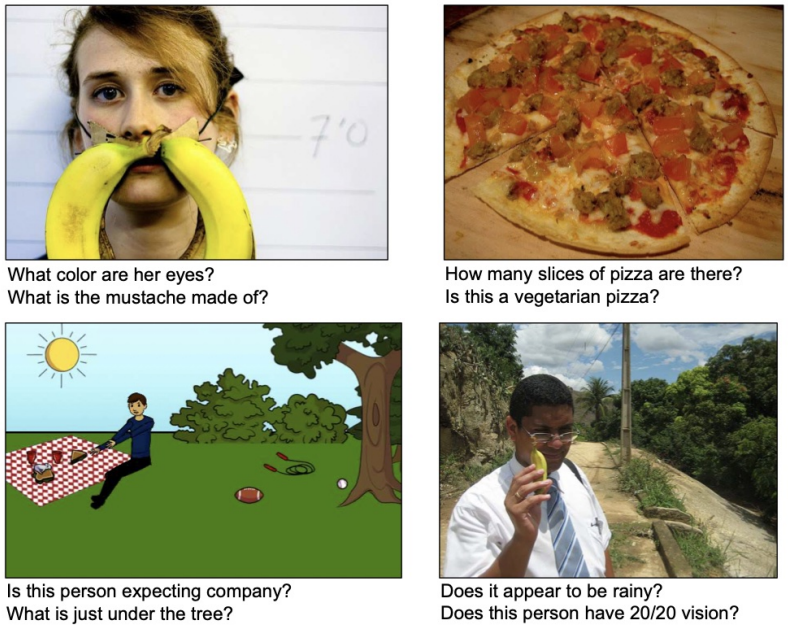


Figure: Example questions from VQA (Antol et al., 2015)

Approaches to Multimodal Grounding

VQA: Visual Question Answering (Antol et al., 2015)

- Natural images from MS-COCO and synthetic images

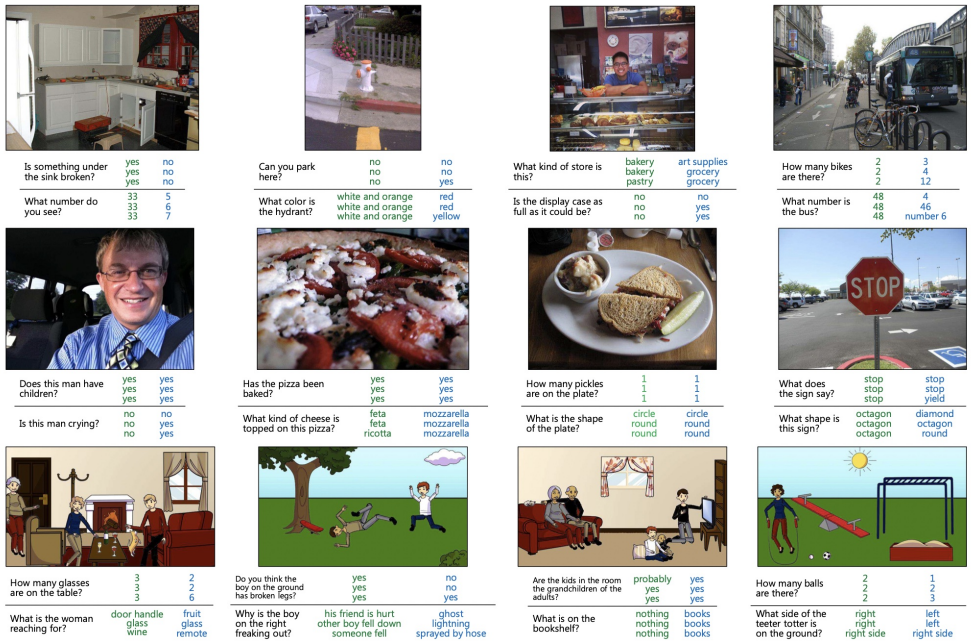


Figure: Example questions and answers (Antol et al., 2015)

Approaches to Multimodal Grounding

VQA: Visual Question Answering (Antol et al., 2015)

- Use BOW representation of question and caption, VGGNet image embedding as input
- Use MLP and LSTM model, supervised approach

	Open-Answer				Multiple-Choice			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
Question	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
Image	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
Q+I	52.64	75.55	33.67	37.37	58.97	75.59	34.35	50.33
LSTM Q	48.76	78.20	35.68	26.59	54.75	78.22	36.82	38.78
LSTM Q+I	53.74	78.94	35.24	36.42	57.17	78.95	35.80	43.41
Caption	26.70	65.50	02.03	03.86	28.29	69.79	02.06	03.82
Q+C	54.70	75.82	40.12	42.56	59.85	75.89	41.16	52.53

Figure: VQA accuracy (Antol et al., 2015)

Approaches to Multimodal Grounding

- Situation Recognition (Yatskar et al., 2016)
- Visual semantic role labeling linking FrameNet frames and ImageNet images


CLIPPING			
ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	PLACE	ROOM


JUMPING			
ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR


SPRAYING			
ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	FIREMAN
SOURCE	SPRAY CAN	SOURCE	HOSE
SUBSTANCE	PAINT	SUBSTANCE	WATER
DESTINATION	WALL	DESTINATION	FIRE
PLACE	ALLEYWAY	PLACE	OUTSIDE

Figure: Image/role/activity links (Yatskar et al., 2016)

Approaches to Multimodal Grounding

Situation Recognition (Yatskar et al., 2016)

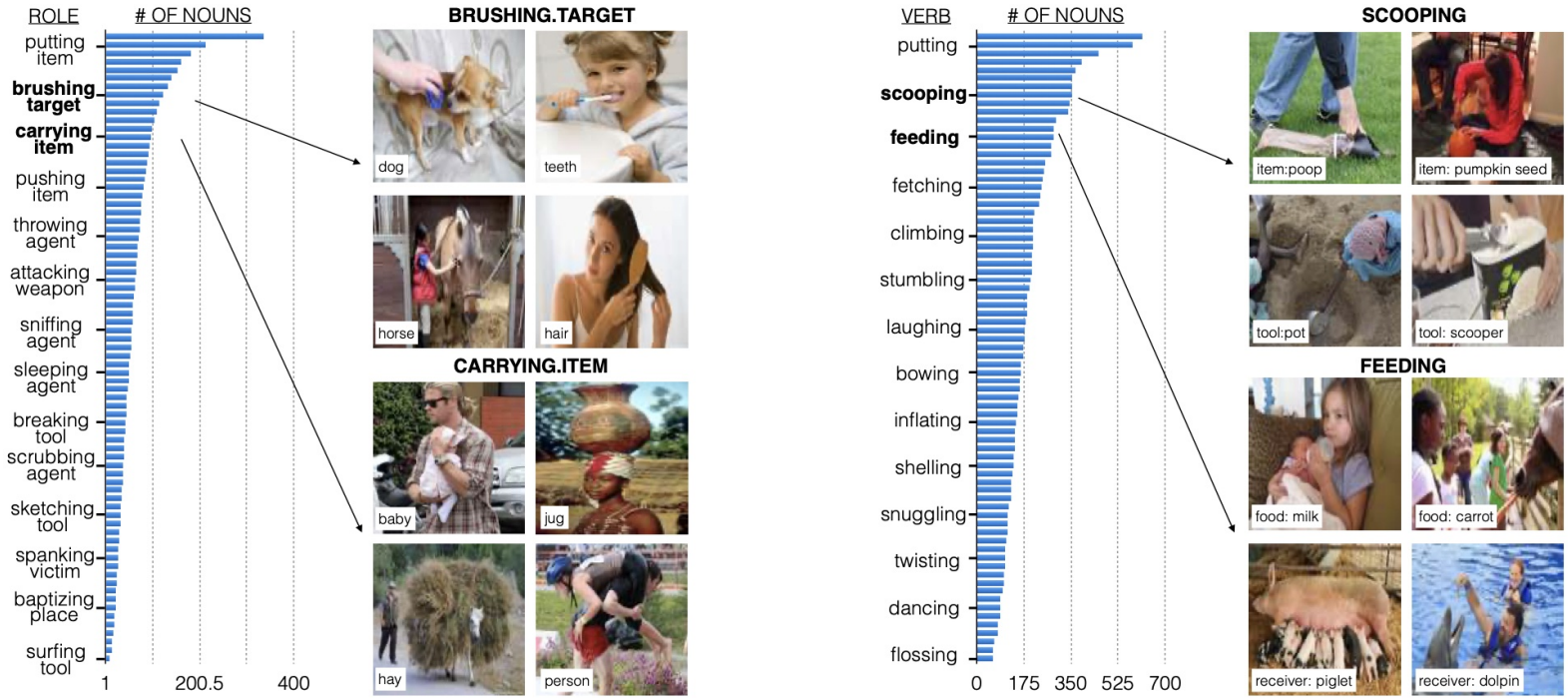


Figure: Nouns that fill multiple roles [L] and nouns that participate in multiple activities [R]

Approaches to Multimodal Grounding

Situation Recognition (Yatskar et al., 2016)

- Uses CRF to predict situation given image i, decomposing i over verb v and role-value pairs

		top-1 predicted verb				top-5 predicted verbs				ground truth verbs		
		verb	value	value-any	value-full	verb	value	value-all	value-full	value	value-all	value-full
dev	Discrete Classifier	26.4	4.0	0.4	0.2	51.1	7.8	0.6	0.4	14.4	0.9	0.6
	CRF	32.2	24.6	14.3	11.2	58.6	42.7	22.7	17.5	65.9	29.5	22.3
test	Discrete Classifier	26.8	4.1	0.3	0.2	51.2	7.8	0.5	0.4	14.4	0.8	0.6
	CRF	32.3	24.6	14.2	11.2	58.9	42.8	22.5	17.5	65.7	29.0	22.0

Figure: CRF performance (Yatskar et al., 2016)

- "[I]n activity-centric images, situation-driven prediction of objects and activities outperforms independent object and activity recognition."

Approaches to Multimodal Grounding

- Grounded Semantic Role Labeling (Yang et al., 2016)
- Integrates language and vision to ground arguments of verbs to action participants in the physical world.

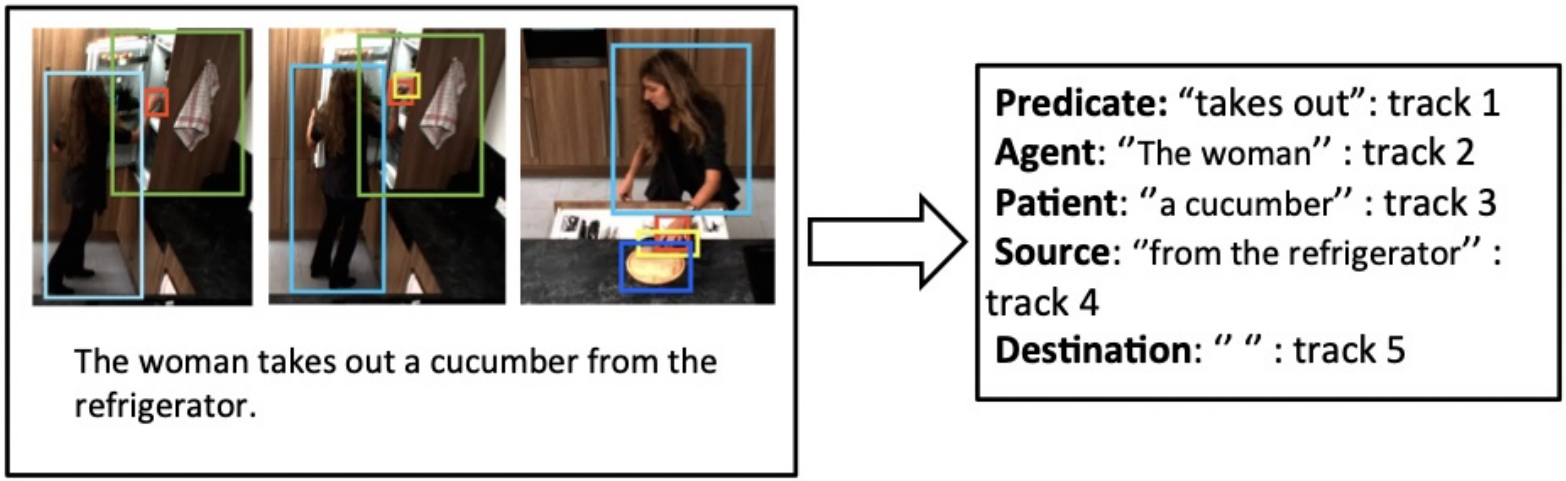


Figure: Example grounded semantic role labeling (Yang et al., 2016)

Approaches to Multimodal Grounding

Grounded Semantic Role Labeling (Yang et al., 2016)

- Given sentence S and corresponding video clip V, ground explicit/implicit roles associated with a verb in S to video tracks in V.
- Focus on {`predicate`, `patient`, `location`, `source`, `destination`, `tool`} roles.
- Also uses CRF approach, which decomposes visual features across text, semantic role, and track-level features.

Approaches to Multimodal Grounding

Grounded Semantic Role Labeling (Yang et al., 2016)

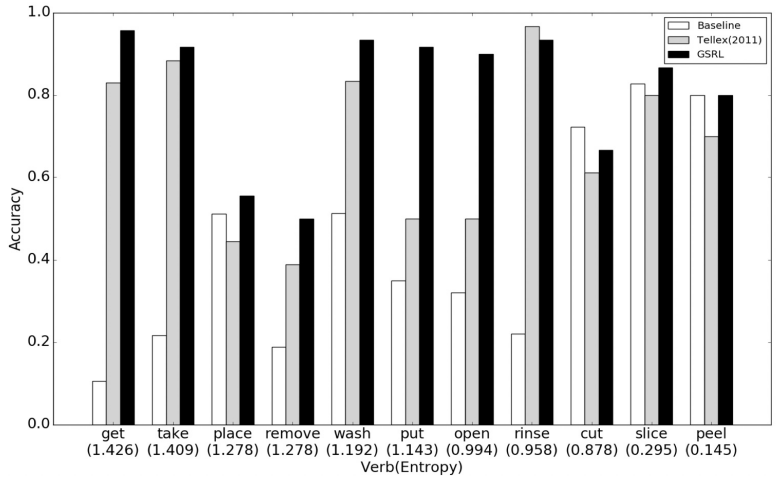


Figure: Relation between accuracy and verb entropy (Yang et al., 2016)

- When using of gold object labeling, incorporating visual features improves performance.
- Effect not observed using automated recognition.

Approaches to Multimodal Grounding

- Where to Look (Shih et al., 2016)
- Answers visual questions by selecting image regions relevant to the text-based query.

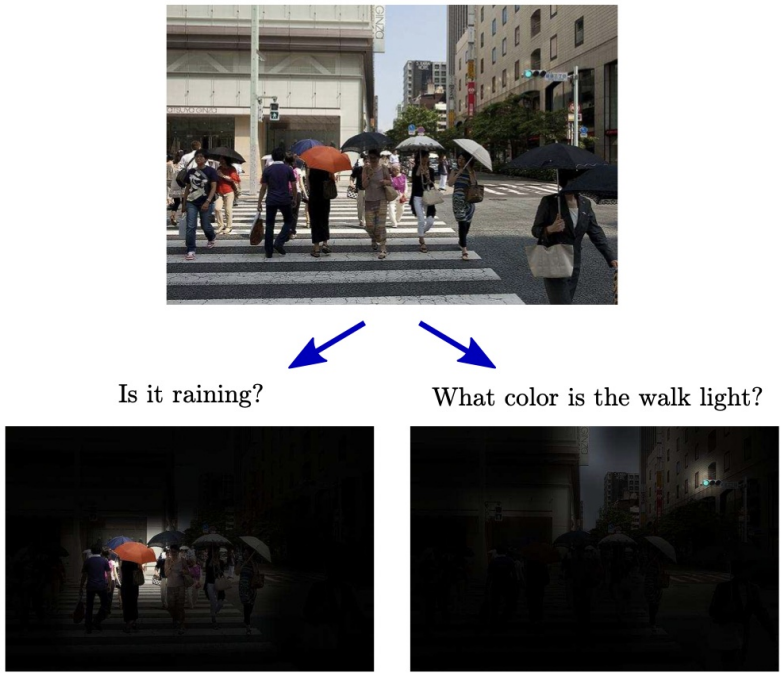


Figure: Identifying salient regions of an image (Shih et al., 2016)

Approaches to Multimodal Grounding

Where to Look (Shih et al., 2016)

- Maps textual queries and visual features into shared space.
- Compared for relevance using inner product.

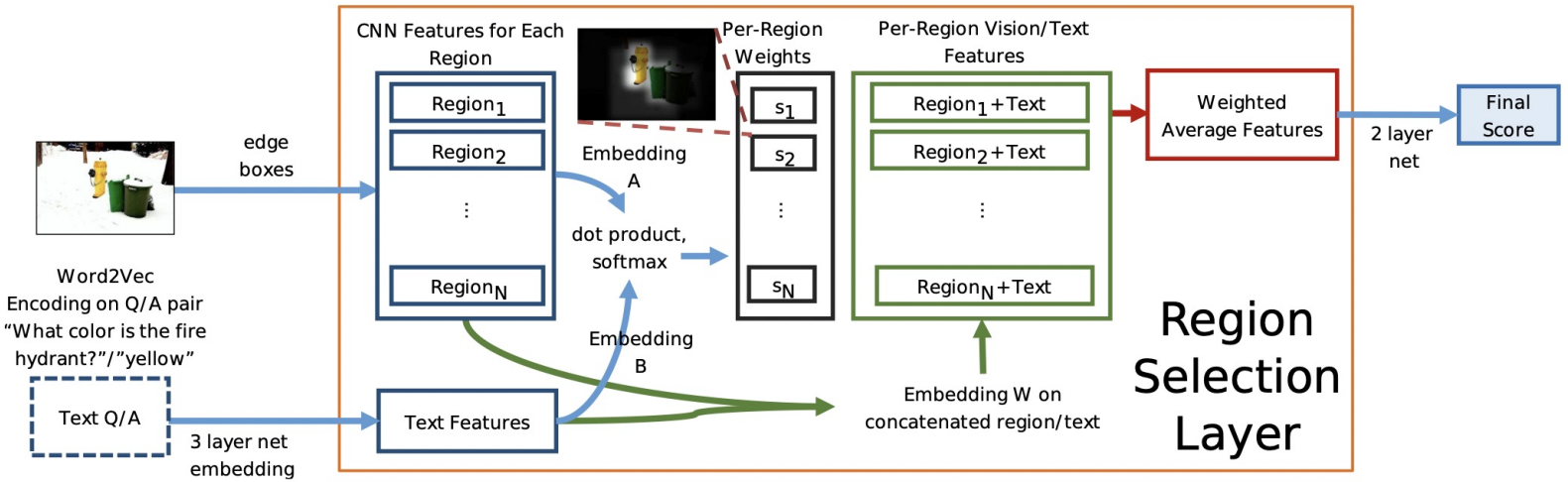


Figure: Where to Look architecture (Shih et al., 2016)

Approaches to Multimodal Grounding

Where to Look (Shih et al., 2016)

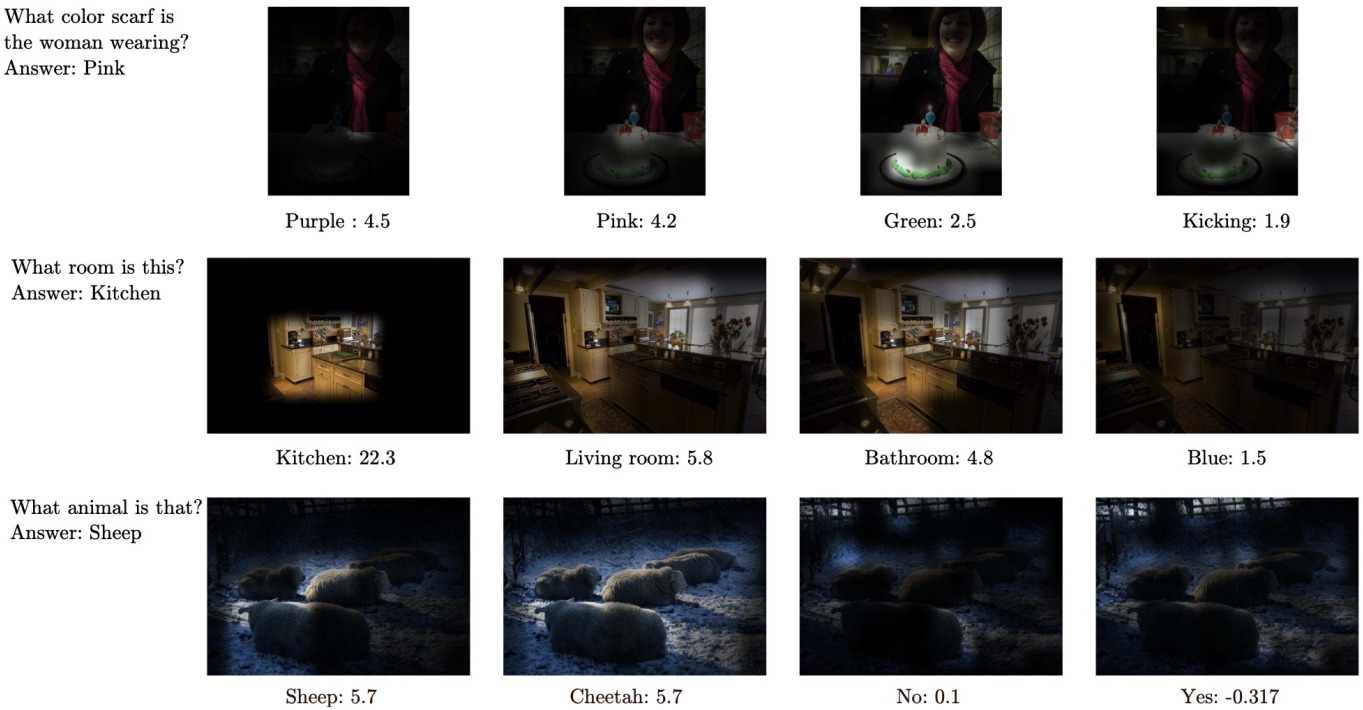


Figure: Where to Look sample results (Shih et al., 2016)

Approaches to Multimodal Grounding

Where to Look (Shih et al., 2016)

- Using dot product attention, Where to Look can identify salient regions even if it can't answer the question.

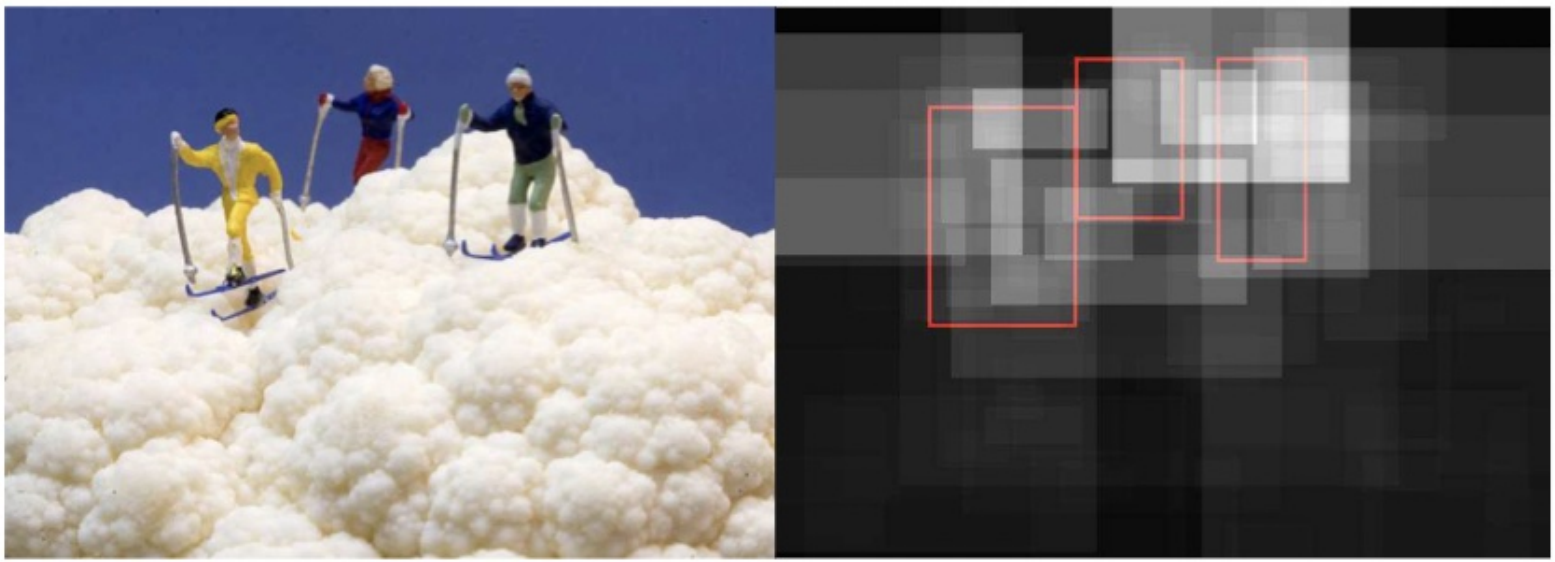


Figure: Identifying salient regions for "Are the people real?" (Shih et al., 2016)

Approaches to Multimodal Grounding

- Multimodal Summarization with Multimodal Output (MSMO) (Zhu et al., 2018)
- Multimodal attention to generate text and select relevant images.

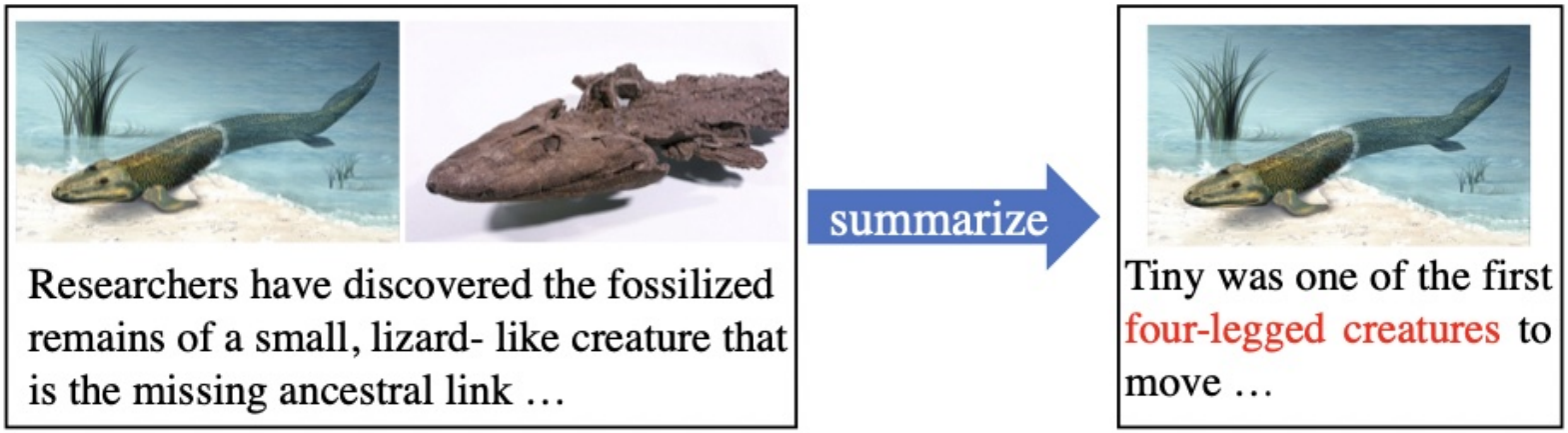


Figure: Summary of MSMO approach (Zhu et al., 2018)

Approaches to Multimodal Grounding

Multimodal Summarization with Multimodal Output (Zhu et al., 2018)

- Uses a "pointer-generator" network for text encoding and summary decoding:

$$e_i^t = v^T \tanh(\mathbf{W}_h h_i + \mathbf{W}_s s_t + \mathbf{W}_c cov^t) \quad (1)$$
$$\alpha^t = \text{softmax}(e^t) \quad (2)$$
$$c_t = \sum_i \alpha_i^t h_i \quad (3)$$

Figure: Pointer-generator network (See et al., 2017) as used in Zhu et al., 2018. cov^t denotes sum of attentions over previous decoder timestep.

Approaches to Multimodal Grounding

Multimodal Summarization with Multimodal Output (Zhu et al., 2018)

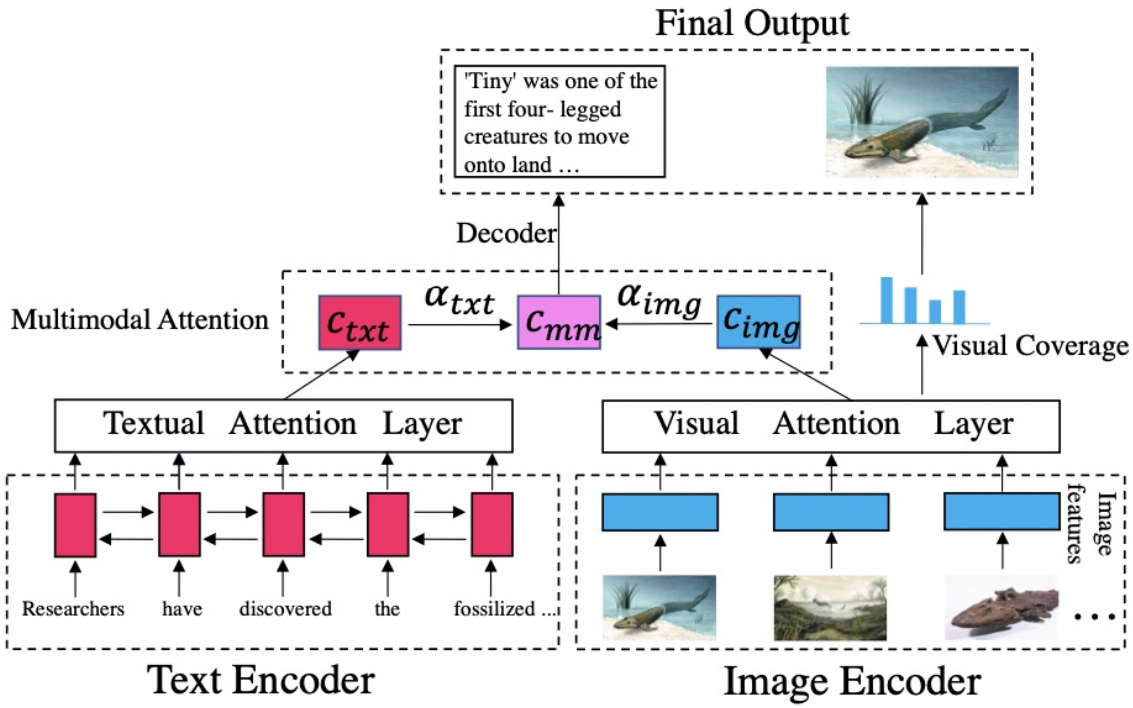


Figure: MSMO framework.

Approaches to Multimodal Grounding

- Improving natural language processing tasks with human gaze-guided neural attention (Sood et al., 2020)
- Hybrid *text saliency model* that combines a cognitive model of reading with human gaze supervision.

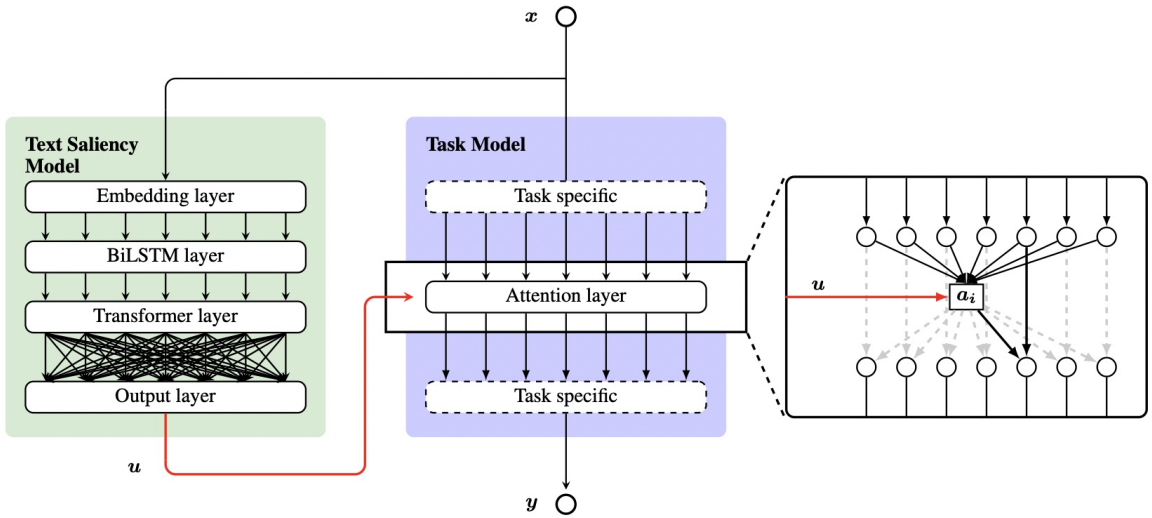


Figure: Sood et al. (2020) architecture.

Approaches to Multimodal Grounding

Improving natural language processing tasks with human gaze-guided neural attention (Sood et al., 2020)

- Pretrain on synthetic gaze data and fine tune on human gaze data.
- Paraphrase generation, sentence compression.

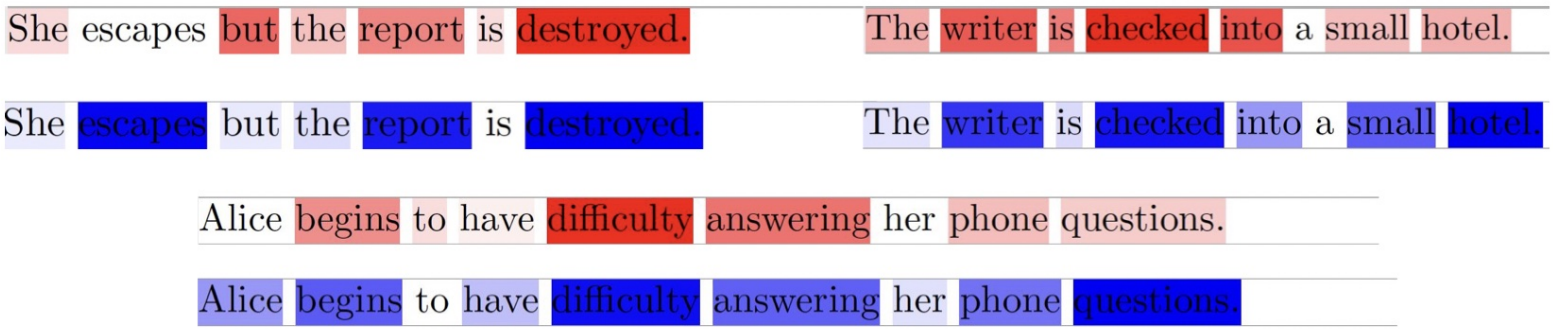


Figure: Human fixation durations (red) and model fixation predictions (blue) (Sood et al., 2020).

Approaches to Multimodal Grounding

- Visual semantic role labeling for video understanding (Sadhu et al., 2021)
- VidSitu: annotates event sequences with roles and dependencies.

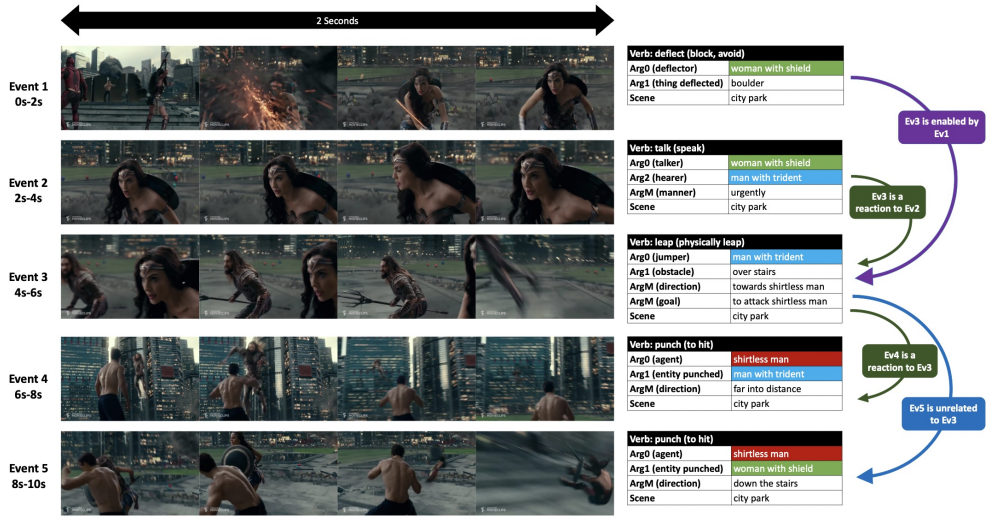
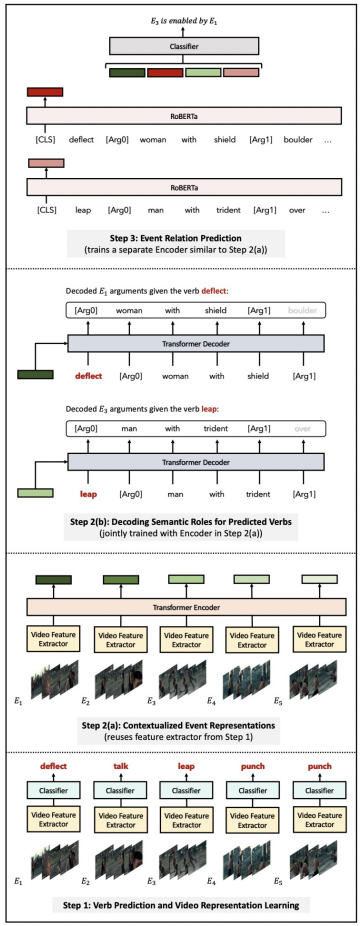


Figure: Sample VidSitu video and annotation (Sadhu et al., 2021).

Approaches to Multimodal Grounding

Visual semantic role labeling for video understanding (Sadhu et al., 2021)



- VidSRL task: Given video V, predict set of related salient events constituting a situation.
- Events must be localized in time and space.
- Events must be appropriately described.
- Coreferring entities must be resolved across events.
- Relations between events must be correctly inferred.

Approaches to Multimodal Grounding

Visual semantic role labeling for video understanding (Sadhu et al., 2021)

- No surprise, including video information improves performance.

Model	Vis	Enc	Val						Test					
			C	R-L	C-Vb	C-Arg	Lea	Lea-S	C	R-L	C-Vb	C-Arg	Lea	Lea-S
GPT2	✗	✗	34.67	40.08	42.97	34.45	48.08	28.1	36.48	41.33	44.27	36.51	49.38	30.24
TxDec	✗	✗	35.68	41.19	47.5	32.15	**51.76**	28.6	35.34	41.45	44.44	32.06	**52.46**	29.18
Vid TxDec	SlowFast	✗	44.78	40.61	49.97	41.24	37.88	28.69	44.95	41.12	49.46	41.98	38.91	30.21
Vid TxEncDec	SlowFast	✓	45.52	**42.66**	**55.47**	**42.82**	**50.48**	**31.99**	47.25	**43.46**	**52.92**	**45.48**	**50.88**	**33.5**
Vid TxDec	I3D	✗	**47.14**	40.67	51.61	41.29	37.89	30.38	**47.9**	41.5	51.29	43.62	38.77	31.73
Vid TxEncDec	I3D	✓	**47.06**	**42.41**	**51.67**	**42.76**	48.92	**33.58**	**48.51**	**42.96**	**53.88**	**44.53**	49.61	**35.46**
Human*			84.85	39.77	91.7	80.15	72.1	70.33	83.68	40.04	87.78	79.29	71.77	70.6

Figure: Semantic role prediction and co-referencing metrics.

Datasets

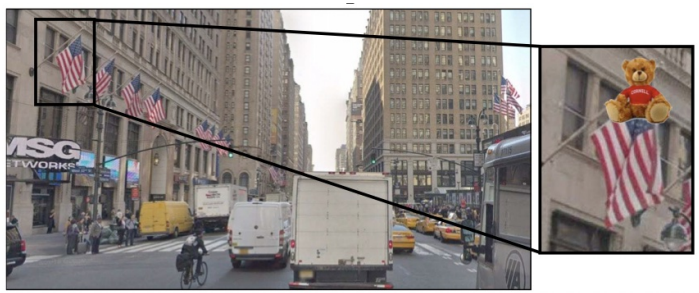
A Multimodal Corpus for Mutual Gaze and Joint Attention in Multiparty Situated Interaction (Kontogiorgos et al., 2018)

- Participants collaborated on moving virtual objects on a touch screen.
- Recordings of speech, eye gaze and gesture using wearable eye trackers, motion capture and A/V.
- 30 participants, 15 interactions.
- 15 hours of triadic interactions (2 participants, 1 moderator).
- Annotated for referring expression targets, eye gaze patterns.
- Alignment: on average, gaze targeted object 0.796s before RE utterance.

Datasets

TOUCHDOWN: *Natural Language Navigation and Spatial Reasoning in Visual Street Environments* (Chen et al., 2019)

- Interactive visual navigation environment based on Google Street View.
- Follow instructions to reach a goal, then resolve spatial description to find Touchdown (hidden teddy bear) in the goal location.
- 29,641 panorama and 61,319 edge images from NYC.



Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic light, As you come to the third traffic light you will see a white building on your left with many American flags on it. Touchdown is sitting in the stars of the first flag.

Datasets

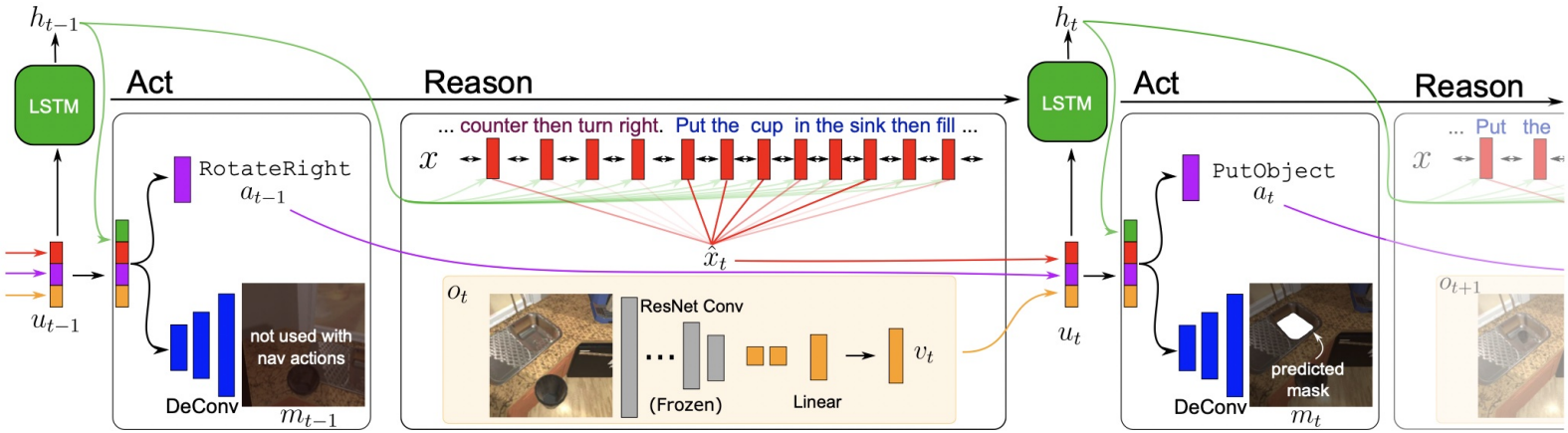
ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks (Shridhar et al., 2020)

- 25K language directives corresponding to expert demonstrations of household tasks.

Datasets

ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks (Shridhar et al., 2020)

- Shridhar et al. (2020)'s model:
  - At each step, reweight instruction based on history;
  - Combine current observation features + previous action.
  - Input to LSTM cell → current hidden state.
  - New hidden state combined with previous features to predict next action + pixelwise interaction mask.



CLIP

- Learning Transferable Visual Models From Natural Language Supervision (Radford et al., 2021)

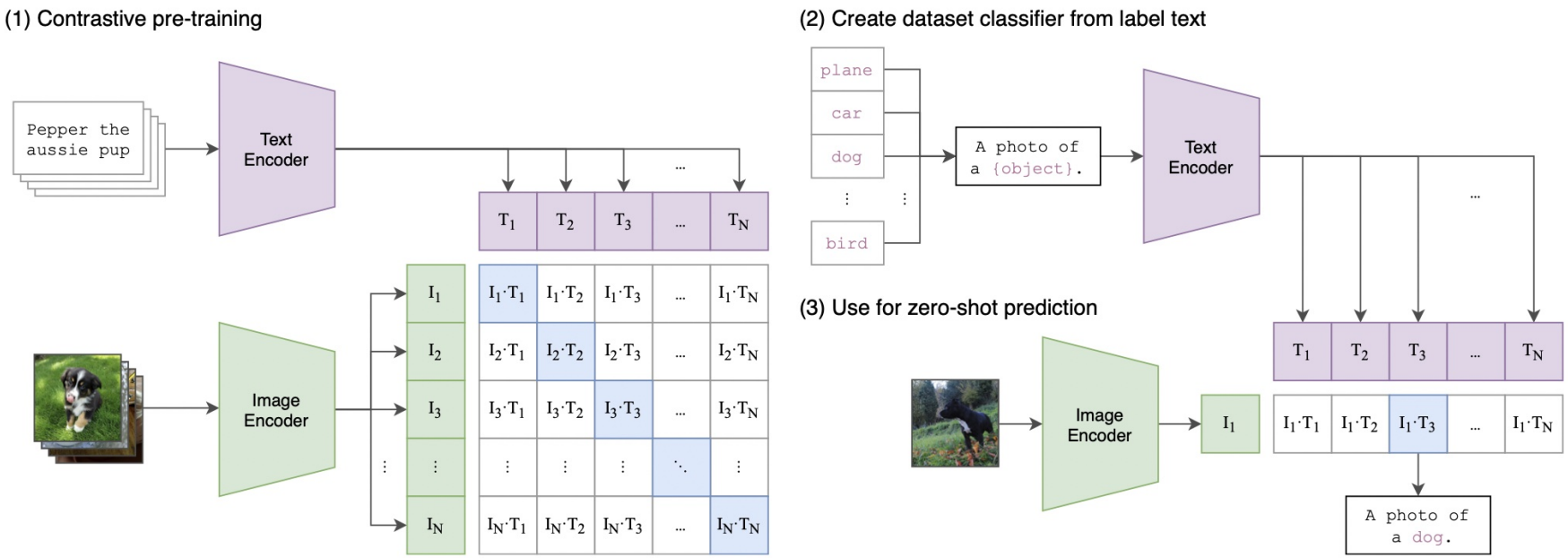
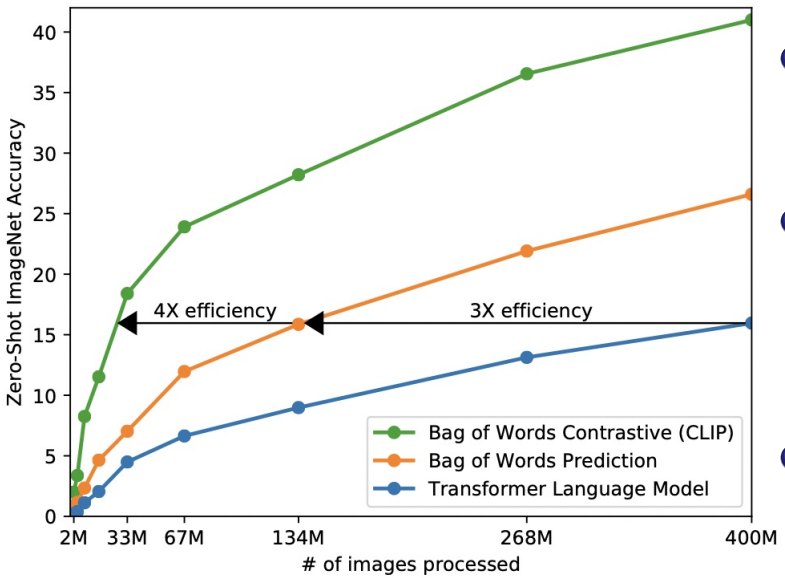


Figure: CLIP overview.

CLIP

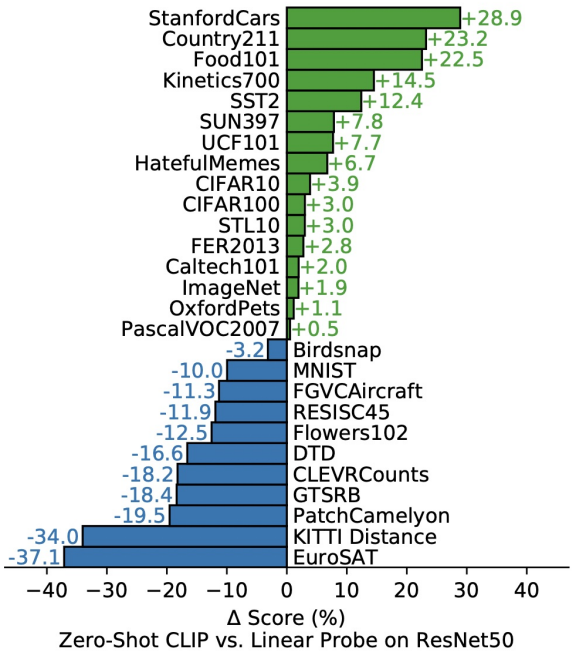
Learning Transferable Visual Models From Natural Language Supervision (Radford et al., 2021)



- Pre-trained on text-image pairs.
- Predicts image encoding correlation to text encoding using contrastive training.
- Four times more efficient at zero-shot ImageNet accuracy than baseline Transformer.

CLIP

Learning Transferable Visual Models From Natural Language Supervision (Radford et al., 2021)



- Zero-shot CLIP still underperforms on:
  - Highly complex, specialized, or abstract tasks;
  - Many of these tasks are easy for non-expert humans.
  - Many of these tasks also have specific methods for implicit meaning representation and interpretation.

Spatial Question Answering

Do we have relevant corpora to evaluate spatial meaning representations and their impact on downstream tasks?

- SQuAD, Hotpot QA, WiQA
- bAbi (task 17 on spatial reasoning), BoolQA

Checking samples of these datasets, we realized:

- No complex spatial descriptions included
- Spatial reasoning is not a key issue for solving these tasks

Spatial Question Answering

A new Benchmark: SpartQA

Formal Representations:

- Topological relations. (contains, part-of, overlap,...)
- Relative directions. (Left, Right, under, above)
- Qualitative distance. (near to, close to, far from)

The girl is on the left of the bookcase. She holds a box with a cat in it.
What is to the right of the cat? The girl or the bookcase?

Rules of Reasoning:

- Symmetry : near to (girl, cat) -> near to (cat, girl)
- Transitivity : left (girl, bookcase) & left (cat, girl) -> left (cat, bookcase)
- Reverse : left (girl, bookcase) -> right (bookcase, girl)

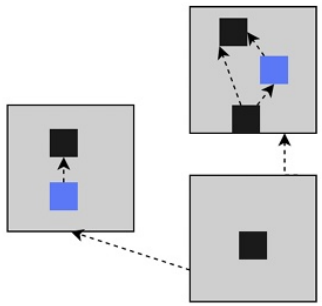
Roshanak Mirzaee, et. al,. 2021. SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning. *2021 Conference of the North American Chapter of the Association for Computational Linguistics(NAACL): Human Language Technologies*, pages 4582–4598

Spatial Reasoning QA dataset

Generate Dataset (SPARTQA) use Visual info and Rules of reasoning as a distant source of supervision

NLVR1 image

Random Sampling




```
[[{"y_loc":80,"size":20,"type":"square","x_loc":40,"color":"Black"},
  {"y_loc":59,"size":20,"type":"square","x_loc":40,"color":"#0099ff"},
  {"y_loc":38,"size":20,"type":"square","x_loc":40,"color":"#0099ff"},
  {"y_loc":17,"size":20,"type":"square","x_loc":40,"color":"Black"}],
```


NLVR1 scene graph (image data)

Story

We have three blocks, A, B and C. Block B is to the right of block C and it is below block A. Block A has two black medium squares. Medium black square number one is below medium black square number two and a medium blue square. It is touching the bottom edge of this block. The medium blue square is below medium black square number two. Block B contains one medium black square. Block C contains one medium blue square and one medium black square. The medium blue square is below the medium black square.

Questions

YN: Is there a square that is below medium square number two above all medium black squares that are touching the bottom edge of a block? Yes

CO: Which object is above a medium black square? the medium black square which is in block C or medium black square number two? medium black square number two

FR: What is the relation between the medium black square which is in block C and the medium square that is below a medium black square that is touching the bottom edge of a block? Left

FB: Which block(s) has a medium thing that is below a black square? A, B, C

FB: Which block(s) doesn't have any blue square that is to the left of a medium square? A, B

Improve Language Models for Spatial Grounding

Evaluating BERT on spatial Understanding and Reasoning.

Fine-tune BERT on MLM task (using auto-SpartQA stories).

Fine-tune BERT on auto-SpartQA's training set.

#	Model	FB	FR	CO	YN	Avg
1	Majority	28.84	24.52	40.18	53.60	36.64
2	BERT	16.34	20	26.16	45.36	30.17
3	BERT (Stories only; MLM)	21.15	16.19	27.1	**51.54**	32.90
4	BERT (SpartQA-Auto; MLM)	19.23	29.54	**32.71**	47.42	34.88
5	BERT (SpartQA-Auto)	**62.5**	**46.66**	**32.71**	47.42	**47.25**
6	Human	91.66	95.23	91.66	90.69	92.31

Try more LMs and various test sets

#	Models	FB			FR			CO			YN		
		Seen	Unseen	Human*	Seen	Unseen	Human*	Seen	Unseen	Human*	Seen	Unseen	Human*
1	Majority	48.70	48.70	28.84	40.81	40.81	24.52	20.59	20.38	40.18	49.94	49.91	**53.60**
2	BERT	87.13	69.38	62.5	85.68	73.71	46.66	71.44	61.09	32.71	78.29	76.81	47.42
3	ALBERT	97.66	83.53	56.73	91.61	83.70	44.76	95.20	84.55	49.53	79.38	75.05	41.75
4	XLNet	**98.00**	**84.85**	**73.07**	**94.60**	**91.63**	**57.14**	**97.11**	**90.88**	**50.46**	**79.91**	**78.54**	39.69
5	Human	85		91.66	90		95.23	94.44		91.66	90		90.69

find relation (FR), find blocks (FB), choose object (CO), and yes/no/DK (YN)

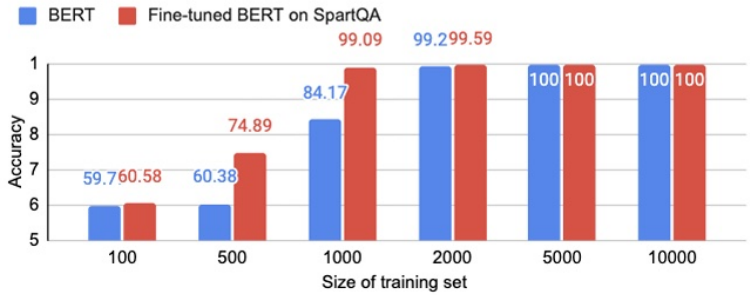
Fine-tuned LM with SpartQA

bAbI dataset (task 17)

The **pink rectangle** is to the *left* of the **red square.**

The **blue square** is to the *right* of the **red square.**

Is the **blue square** to the *left* of the **pink rectangle**? No
Is the **red square** to the *left* of the **blue square**? Yes



Model	Accuracy
Majority baseline	62.2
Recurrent model (ReM)	62.2
ReM fine-tuned on SQuAD	69.8
BERT (our setup)	71.89
ReM fine-tuned on QNLI	71.4
ReM fine-tuned on NQ	72.8
BERT fine-tuned on auto-SPARTQA	**74.18**

boolQ dataset

Q: Has the UK been hit by a hurricane?
P: The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands ...
A: Yes. [An example event is given.]

A paper with similar idea in AAAI-2022 created a dataset called **StepGame**. There is no citation of SpaRTQA of NAACL-2021. So there is a disconnect between the results and comparisons.

Limitations of SpRTQA for Transferability

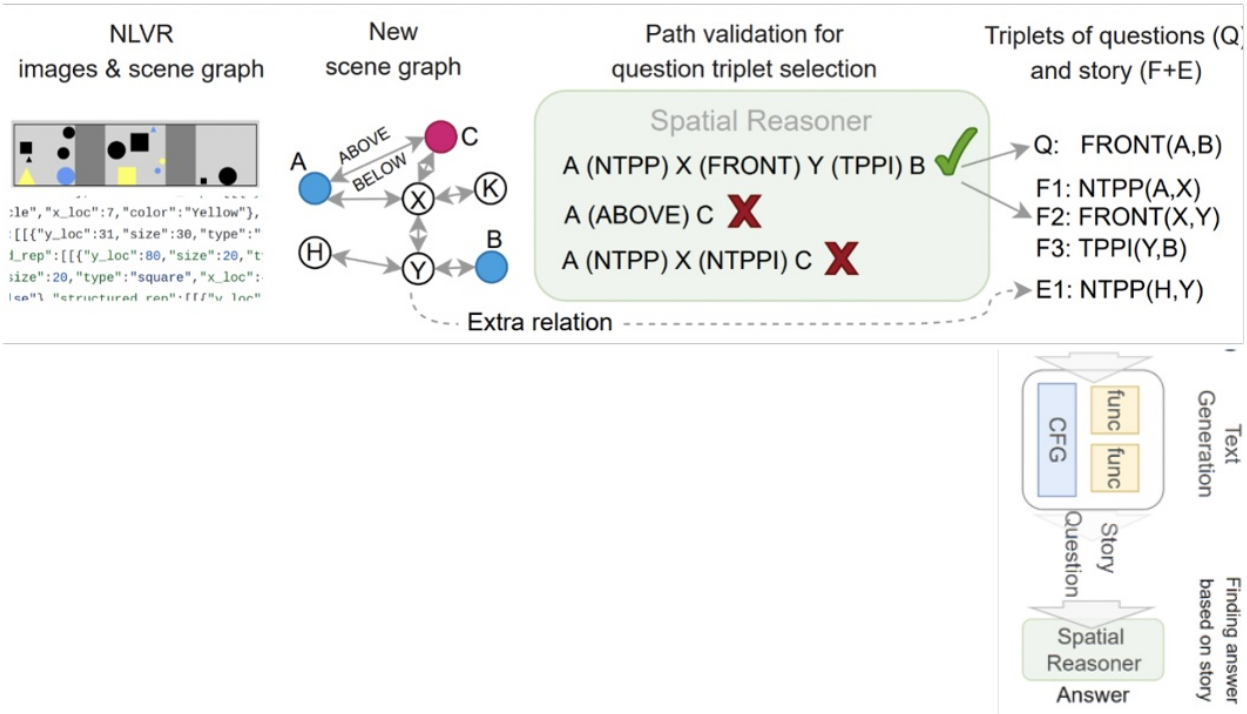
Problems

- Few spatial types
  - Compared to real world
- Low generalizability
  - Limited vocabulary
- Complex question text with nesting relations
  - Hard to process the question

Solution

- Gather more relation types
- Extend the vocabulary
- Simplified the questions and kept the reasoning difficulty/multi-hop reasoning on the text side

SpaRTUN extension for Transferability



SpaRTUN ReSQ

- On two tasks (annotation):
  - Spatial Question Answering
  - Spatial Role Labeling
    - Spatial Entity/indicator extraction
    - Spatial Relation Extraction

Dataset	Train	Dev	Test
SPARTUN(YN)	20334	3152	3193
SPARTUN(FR)	18400	2818	2830

R. Mirzaee, P. Kordjamshidi, 2022. Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning. The 2022 Conference on Empirical Methods in Natural Language Processing

Three boxes called one, two and three exist in an image. Box one contains a big yellow melon and a small orange watermelon. **Box two has a small yellow apple**. A small orange apple is inside and touching this box. Box one is in box three. **Box two** is to the **south** of, **far from** and to the **west** of **box three**. A **small yellow watermelon** is **inside box three**.

Q: Is **the small yellow apple** to the **west** of the **small yellow watermelon**? Yes

Q: Where is **box two** relative to the **small orange watermelon**? Left, Below, Far

(a) SPARTUN - A synthetic large dataset provided as source of supervision

A **grey car** is parking in **front** of a **grey house with brown window frames** and **plants on** the **balcony**.

Q: Are the **plants in front of** the **car**? No

Q: Are the **plants in** the **house**? Yes

(b) RESQ - A human-generated dataset for probing the models on realistic spatial problems

Experimental Results on QA

Model	DS	17^{900}	19^{500}
MB	-	51.9	10.6
BERT	-	87.39	34.53
BERT	SPARTQA-A	90.42	**100**
BERT	StepGame	87.39	99.89
BERT	SPARTUN-S	**92.43**	98.99
BERT	SPARTUN	90.02	99.89

Impact of using synthetic supervision on the bAbI tasks. All the models are further fine-tuned on the training set of task 17 (size = 1k) and 19 (size = 500), and test on bAbI test sets.

Model	DS	YN	FR
MB	-	53.60	24.52
BERT	-	**49.65**	18.18
BERT	SPARTQA-A	39.86	48.05
BERT	StepGame	44.05	11.68
BERT	SPARTUN-S	44.75	37.66
BERT	SPARTUN	48.25	**50.64**
Human	-	90.69	95.23

Impact of transfer learning on SPARTQA-HUMAN. SPARTQA-A stands for SPARTQA-AUTO.

		k steps of reasoning									
Model	DS	1	2	3	4	5	6	7	8	9	10
TP-MANN	-	85.77	60.31	50.18	37.45	31.25	28.53	26.45	23.67	22.52	21.46
BERT	-	98.44	94.77	91.78	71.7	57.56	50.34	45.17	39.69	35.41	33.62
BERT	SPARTQA-A	98.63	94.95	91.94	77.74	68.37	61.67	57.95	50.82	46.86	44.03
BERT	SPARTUN-S	**98.70**	**95.21**	**92.46**	77.93	69.53	62.14	57.37	48.79	44.67	42.72
BERT	SPARTUN	98.55	95.02	92.04	**79.1**	**70.34**	**63.39**	**58.74**	**52.09**	**48.36**	**45.68**

Result of models with and without extra supervision on StepGame.

Experimental Results SpRL

- Spatial Roles and Relations

Model	DS	mSpRL	SpartQA-H
R-Inf	-	80.92	-
SAE	-	**88.59**	55.8
SAE	SpartQA-A	88.41	57.28
SAE	SpaRTUN	88.03	**72.43**

Evaluate spatial argument extraction(SAE) on two mSpRL and SpartQA-Human(SpartQA-H) datasets with and without synthetic supervision.

Model	DS	mSpRL	SpartQA-H
R-Inf	-	68.78	-
SRE	-	69.12	S: 48.58 Q: 49.46
SRE	SpartQA-A	68.38	S: 58.32 Q: 55.17
SRE	SpaRTUN	**74.74**	S: **61:53** Q: **63.22**

Spatial relation extraction on mSpRL and SpartQA-Human(SpartQA-H) with and without synthetic supervision. Since the questions(Q) and stories(S) in SpartQA-Human have different annotations (questions have empty spatial_indicators), we separately train and test this model on each.

ReSQ (Real World Spatial Questions)

A **grey car** is parking in **front** of a **grey house** **with brown window frames** and **plants on** the **balcony**.

Q: Are the **plants in front of** the **car**? No

Q: Are the **plants in** the **house**? Yes

Model	DS	Accu
MB	-	50.21
BERT	-	57.37
BERT	SPARTQA-AUTO	55.08
BERT	StepGame	60.14
BERT	SPARTUN-S	58.03
BERT	SPARTUN	**63.60**
Human	-	90.38

Results of models with and without extra supervision on the ReSQ. The Human accuracy is the performance of human on answering a part of test set.

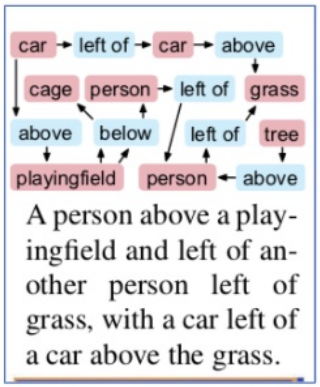
Text-to-image Synthesis

Text-to-image or text-to-scene synthesis techniques aimed at naturally composing and visualizing **text** instances:

- Many practical applications in education, gaming, creating virtual realities steered and manipulated through language

Text-to-image synthesis: integration of a scene graph

- Text is first translated into a scene graph (= symbolic representation expressing the objects and their semantic/spatial relationships)
- The spatial layout is generated from the scene graph



- Use of a graph convolution network composed of several graph convolution layers to represent objects and their relationships
- Followed by steps of layout prediction and pixel prediction

Text-to-image synthesis: integration of a scene graph

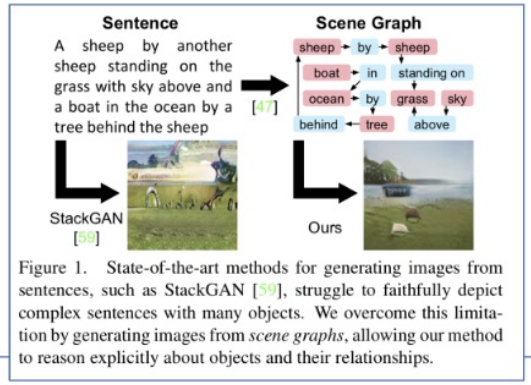


Figure 1. State-of-the-art methods for generating images from sentences, such as StackGAN [59], struggle to faithfully depict complex sentences with many objects. We overcome this limitation by generating images from *scene graphs*, allowing our method to reason explicitly about objects and their relationships.

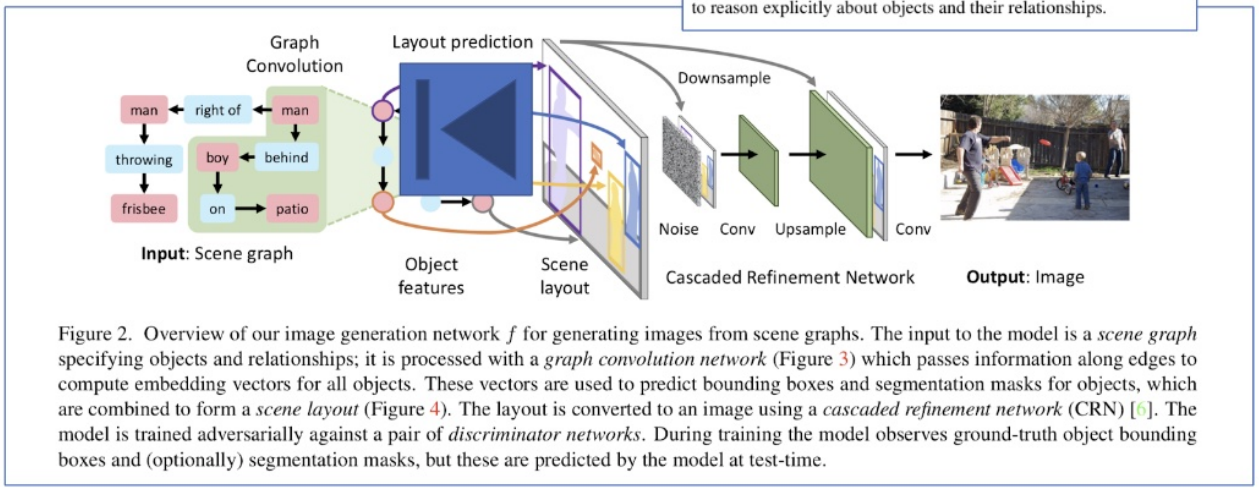


Figure 2. Overview of our image generation network f for generating images from scene graphs. The input to the model is a *scene graph* specifying objects and relationships; it is processed with a *graph convolution network* (Figure 3) which passes information along edges to compute embedding vectors for all objects. These vectors are used to predict bounding boxes and segmentation masks for objects, which are combined to form a *scene layout* (Figure 4). The layout is converted to an image using a *cascaded refinement network* (CRN) [6]. The model is trained adversarially against a pair of *discriminator networks*. During training the model observes ground-truth object bounding boxes and (optionally) segmentation masks, but these are predicted by the model at test-time.

Justin Johnson, Agrim Gupta & Li Fei-Fei (2018). Image generation from scene graphs. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (CVPR).

Text-to-image synthesis

The LSTM encoder provides a representation (embedding) of each object mentioned in the input text

From this representation a bounding box of the object is predicted in the 2D space:
The output of the box generator is a set of bounding boxes **B** $= \{B_1, \ldots, B_n\}$ where each bounding box B_t defines the location, size and category label of the t-th object

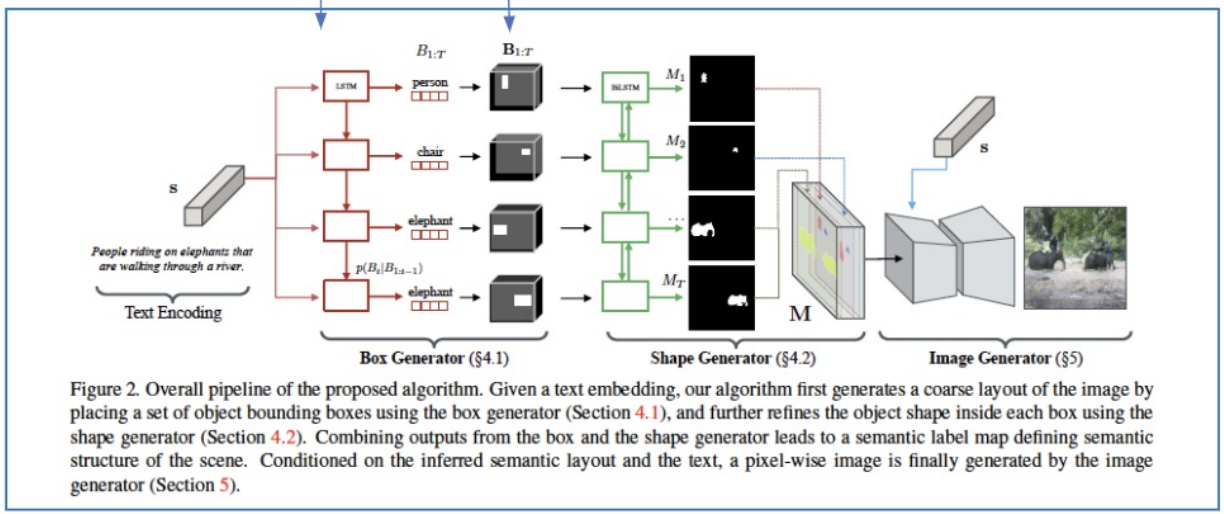


Figure 2. Overall pipeline of the proposed algorithm. Given a text embedding, our algorithm first generates a coarse layout of the image by placing a set of object bounding boxes using the box generator (Section 4.1), and further refines the object shape inside each box using the shape generator (Section 4.2). Combining outputs from the box and the shape generator leads to a semantic label map defining semantic structure of the scene. Conditioned on the inferred semantic layout and the text, a pixel-wise image is finally generated by the image generator (Section 5).

Seunghoon Hong, Dingdong Yang & Jongwook Choi (2018). Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (CVPR).

Text-to-image synthesis

- Qualitative evaluation of the full image generation process

Seunghoon Hong, Dingdong Yang & Jongwook Choi (2018). Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (CVPR).

Visual Language Modeling

Depending on the context, spatial language might have different meaning in terms of targeted geometry

a man riding a motorcycle in front of an orange bus

The distance between the man and the motorcycle is usually much smaller in a city environment compared to a highway environment

Spatial Reasoning BERT

Quantitative and qualitative evaluation

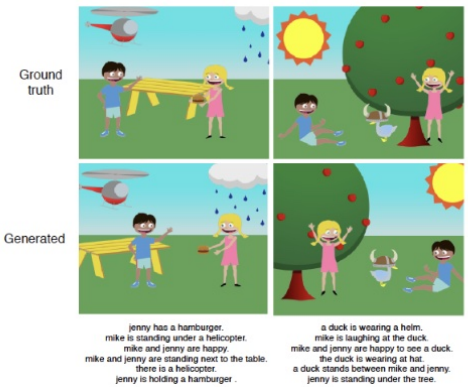


Figure 1: Ground truth and generated spatial arrangements expressed as 2D scenes of clip-arts. For each scene, there is a set of spatial relationships expressed in both implicit and explicit language.

Method	Prec	Rec	Pose	Expr	Abs. sim.
(Zitnick et al., 2013)	72.2	65.5	40.7	30.0	0.449
(Tan et al., 2018)	76.0	69.8	**41.8**	37.5	0.409
ClipPredict + SR-BERT	**82.7**	**72.5**	40.4	**38.0**	**0.512**

Table 4: Per-object precision and recall, pose and expression classification accuracies, and abs. sim. using the test split provided by Tan et al. (2018).

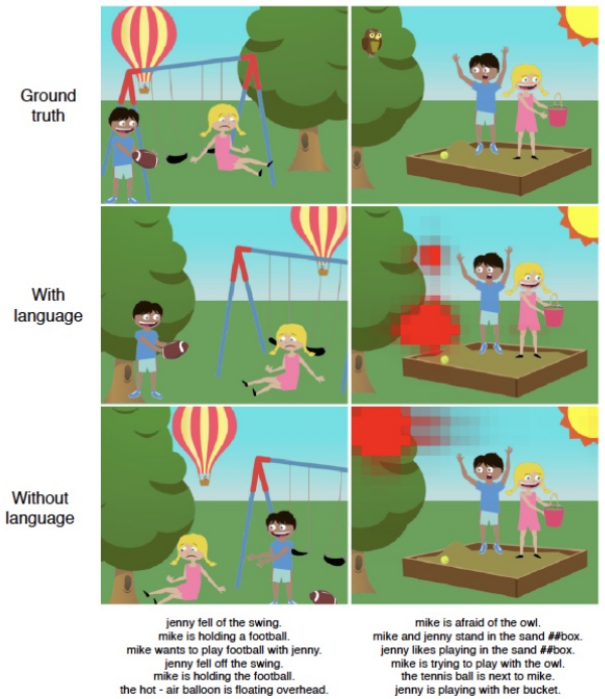


Figure 4: Ground truth (top), generated scenes (middle-left, bottom-left) and heat-maps (middle-right, bottom-right) with and without conditioning on the language.

Gorjan Radevski, Guillem Collell, Marie-Francine Moens & Tinne Tuytelaars (2020). Decoding language spatial relations to

Break

BREAK
We will resume in 15 minutes

Abstract Meaning Representations (AMR)

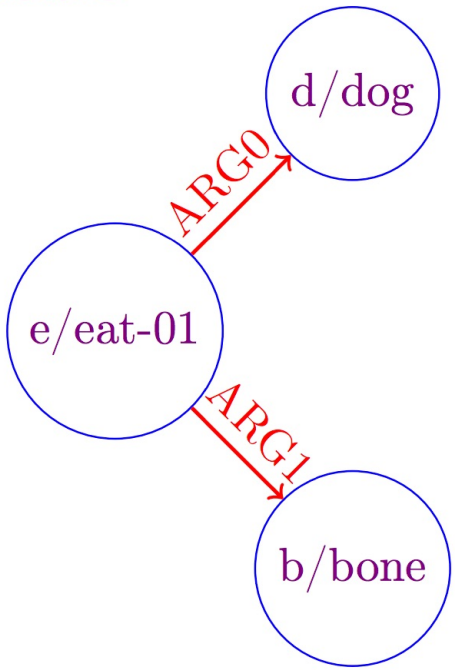
Banarescu et al (2013)

- Rooted, acyclic, directed graph used to represent the meaning of English sentences.
- Builds on valency lexicon developed for Propbank, where senses of each predicate, verbal or nominal, and their semantic roles are specified.
- Syntactically it uses the Penman notation and the semantic structure of each sentence closely resembles that of a dependency tree ...
- except when there is re-entrancy, where the same concept is an argument for multiple predicates.
- Each argument of a predicate is labeled with a semantic role; interpretation does not depend on the order it has in syntax, so it abstracts away from its morpho-syntactic variations.

Abstract Meaning Representations (AMR)

NAACL-HLT Tutorial 2015
Nathan Schneider, Jeffrey Flanigan, and Tim O'Gorman

- The edges (ARG0 and ARG1) are **relations**
- Each node in the graph has a **variable**
- They are labeled with **concepts**
- **d** / **dog** means "**d** is an instance of **dog**"

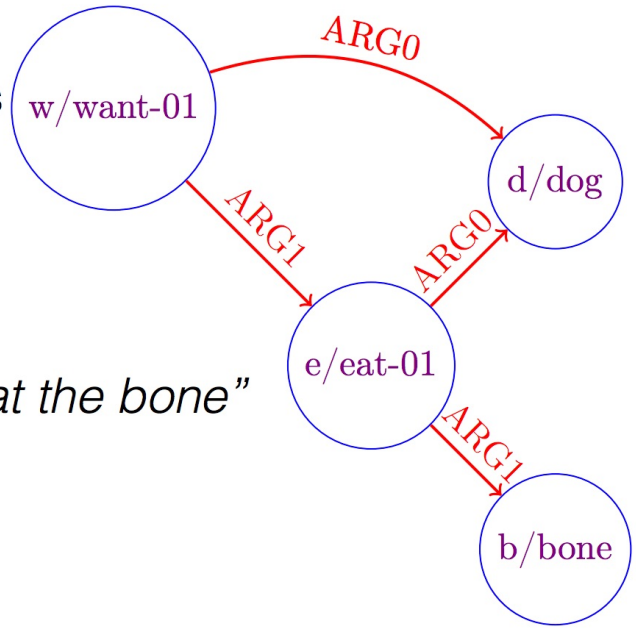


"The dog is eating a bone"
(**e** / **eat-01**
:**ARG0** (**d** / **dog**)
:**ARG1** (**b** / **bone**))

Abstract Meaning Representations (AMR)

NAACL-HLT Tutorial 2015
Nathan Schneider, Jeffrey Flanigan, and Tim O'Gorman

- What if something is referenced multiple times?
- Notice how **dog** has two incoming roles now.
- To do this in PENMAN format, repeat the variable. We call this a **reentrancy**.



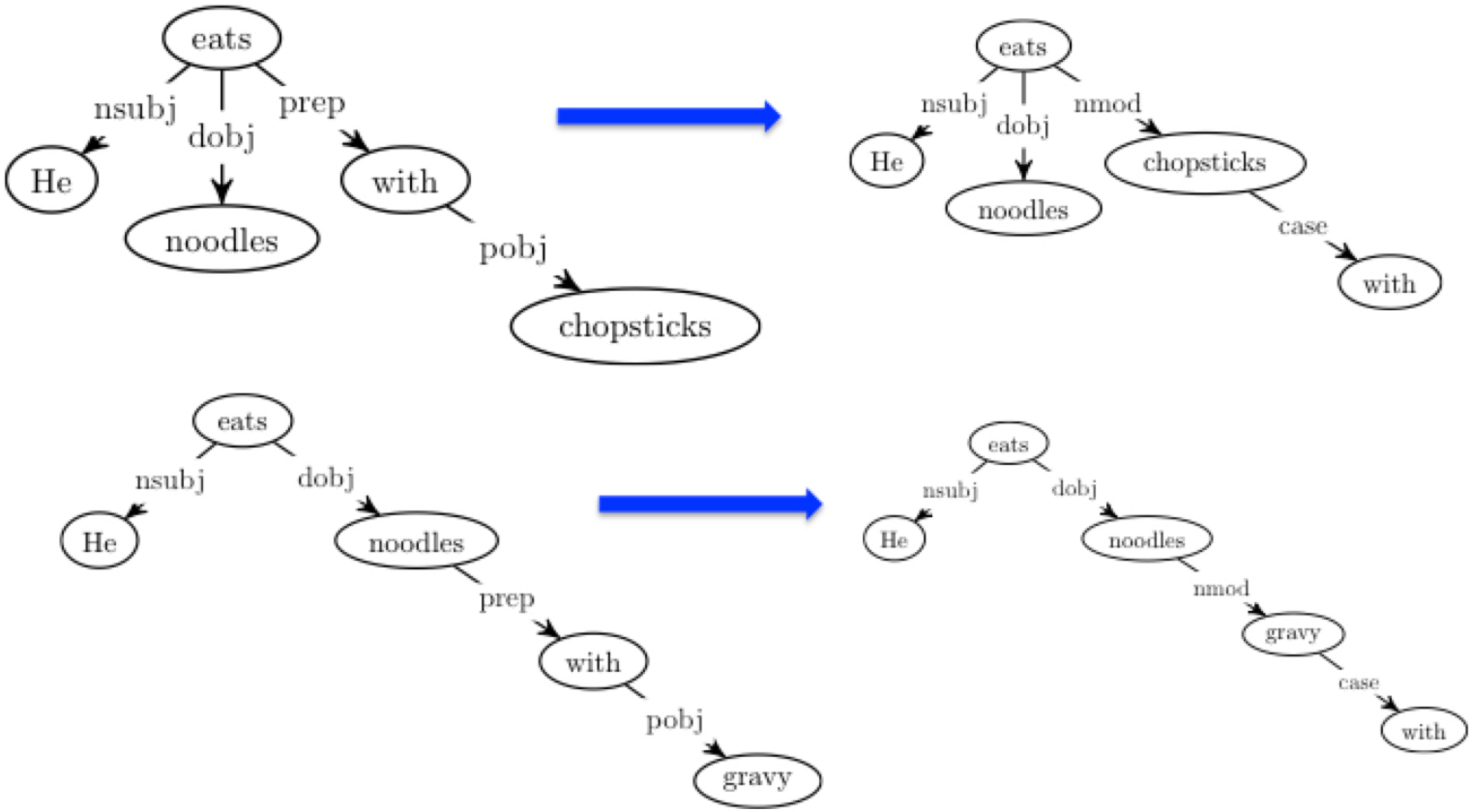
*"The dog **wants to** eat the bone"*

```
(want-01
   :ARG0 (d / dog)
   :ARG1 (e / eat-01
      :ARG0 d
      :ARG1 (b / bone)))
```

Universal Dependencies

- The primacy of content words: Dependency relations hold primarily between content words, rather than being indirect relations mediated by function words.
- Function words normally do not have dependents of their own. Multiple function words related to the same content words are typically siblings
- The dependency relations are described by a mixture of functional and structural notions: advmod vs nmod
- There is some machinery to account for word order variations: nsbj vs nsbjpass
- In coordination structures, the first conjunct is the head, and all other conjuncts depend on it. So are the coordinating conjunctions.

Primacy of Content Words



Comparison between UD, SDP and AMR

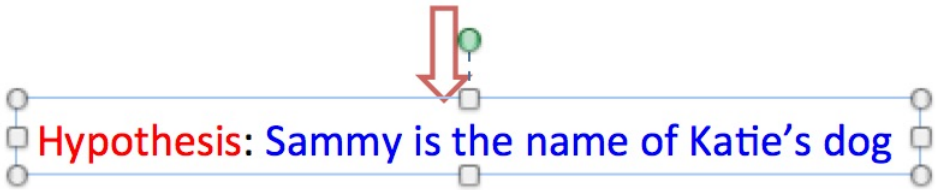
Attributes	SRL	UD	SDP	AMR
Normalizes syntactic variations	yes	no	yes	yes
Predicate sense	yes	no	no	yes
Primacy of content words	n/a	yes	yes	yes
Leaves function words unattached	no	no	yes	yes
A complete and connected structure for a sentence	no	yes	yes	yes
Allows re-entrancy (graph)	n/a	no	yes	yes
Named entities	no	no	no	yes
Relations between named entities	no	no	no	yes
Polarity and modality	no	no	no	yes

Why is abstraction good?

- "Normalizes" different realizations of the same meaning, supports similarity based inference that helps applications such as text comprehension.

Text: ... Katie also has a dog, but he does not like Bows.
... His name is Sammy. ...

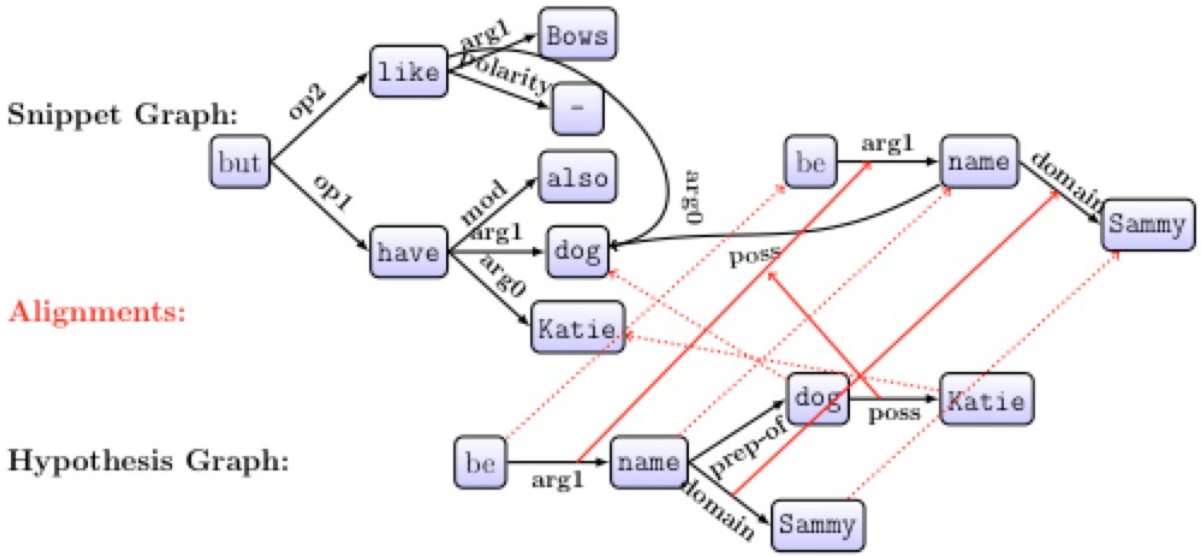
Question: What is the name of Sammy's dog? Answer: Sammy.



From *(Sachan and Xing 2016)*

Why is abstraction good?

- "Normalizes" different realizations of the same meaning, supports similarity based inference that helps applications such as text comprehension.

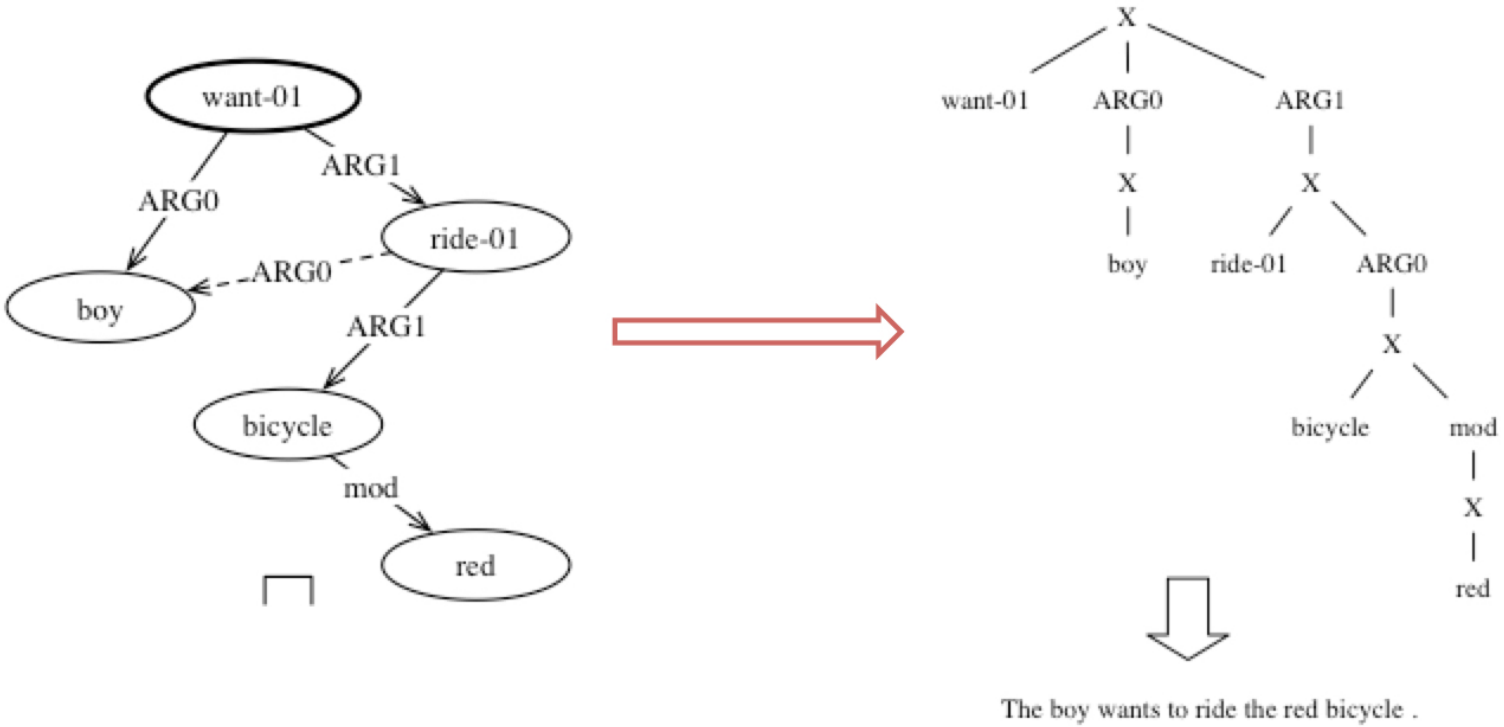


Hypothesis: Sammy is the name of Katie's dog.
Question: What is the name of Katie's dog. Answer: Sammy

From *(Sachan and Xing 2016)*

Why is abstraction good?

- Supports natural language generation

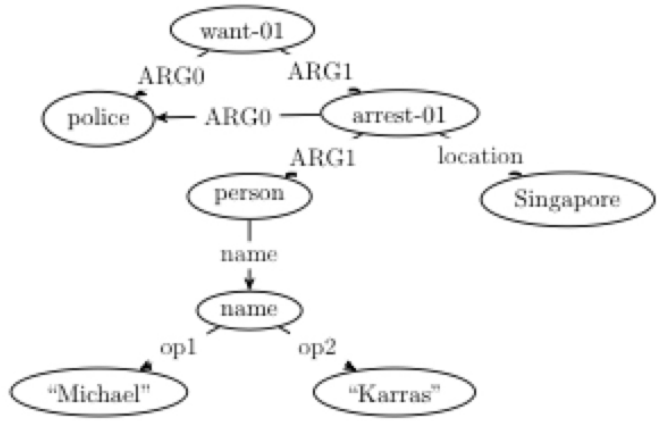


AMR parsing approaches

- Approaches that focus on addressing the "abstract" nature of AMR
  - Dependency Tree to AMR graph transduction (Wang et al 2015a, Wang et al 2015b, Wang et al 2016)
  - String to tree translation (Pust et al 2015)
- Approaches that focus on the "graph" aspect of the AMR
  - AMR parsing based on Synchronous Hyper-edge Replacement Grammar (SHRG) (Peng et al 2015)
  - MSCG: Maximum Spanning Connected Graph (Flanigan et al 2014)

Tree-to-AMR graph transduction

(a) Dependency tree

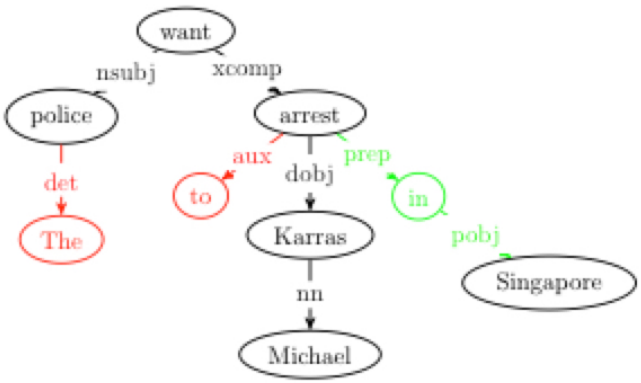


(b) AMR graph

The police wants to arrest Micheal Karras in Singapore.

Linguistically, there are many similarities between an AMR and the dependency structure of a sentence.

Tree-to-AMR graph transduction



(a) Dependency tree

(b) AMR graph

The police wants to arrest Micheal Karras in Singapore.

Wang et al, 2015a, 2015b

Situated Reasoning in Collaborative Tasks

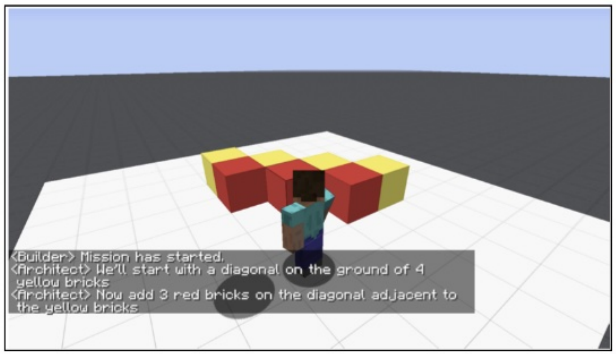


Figure 1: An instance of the collaborative building task. The last instruction was : Now add 3 red bricks on the diagonal adjacent to the yellow bricks.

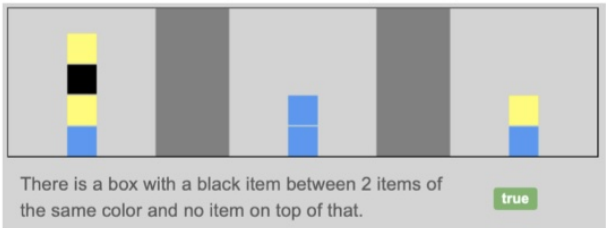


Figure 2: An example from the NLVR corpus that demonstrates *spatial focus shift* from the *black item* to the *yellow item.*

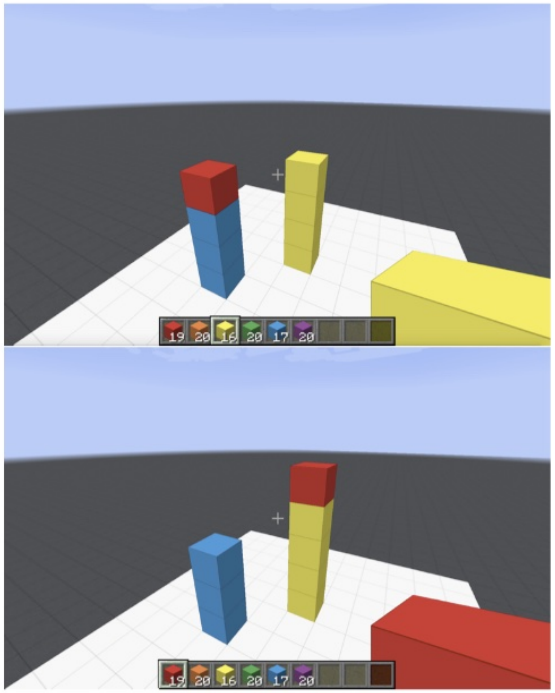


Figure 7: Move the large red block diagonally from the top of the blue column to the top of the yellow column ...

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).

Situated 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>

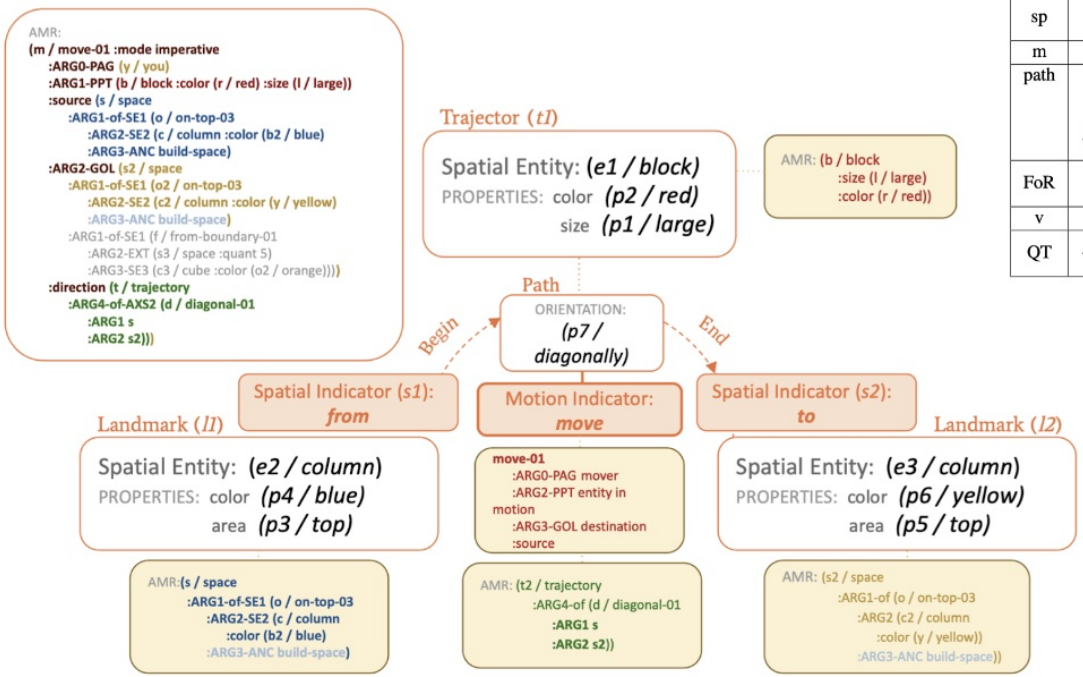


Figure 3: Graphical Representation of Configuration 1 of Table 3 with aligned AMR : ***Move the large red block diagonally from the top of the blue column to the top of the yellow column***, *which is 5 spaces from the orange cube.*

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

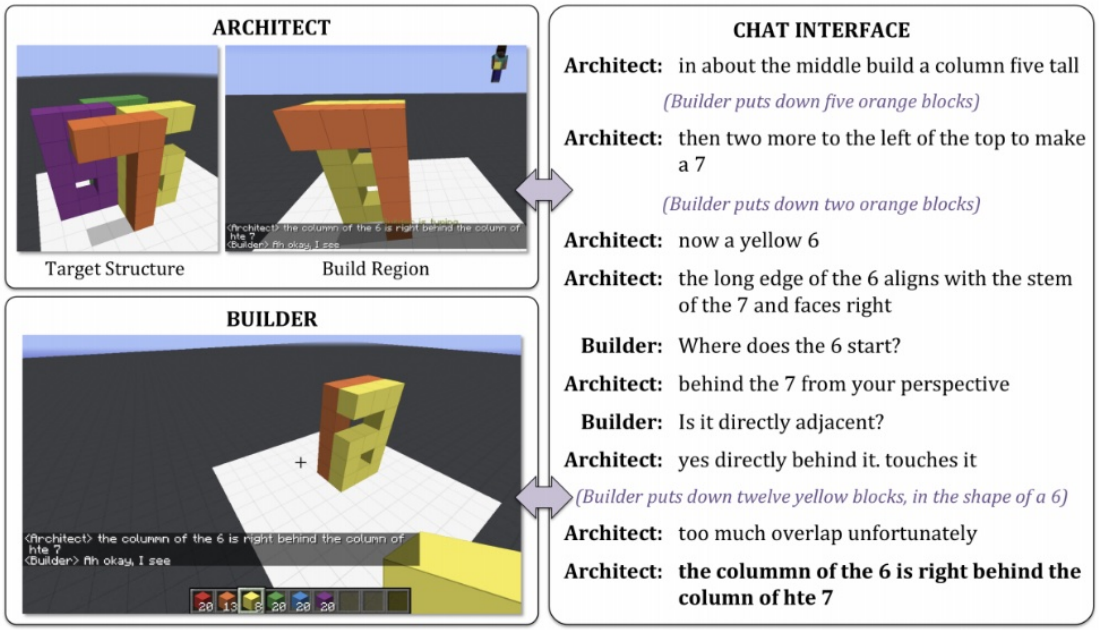
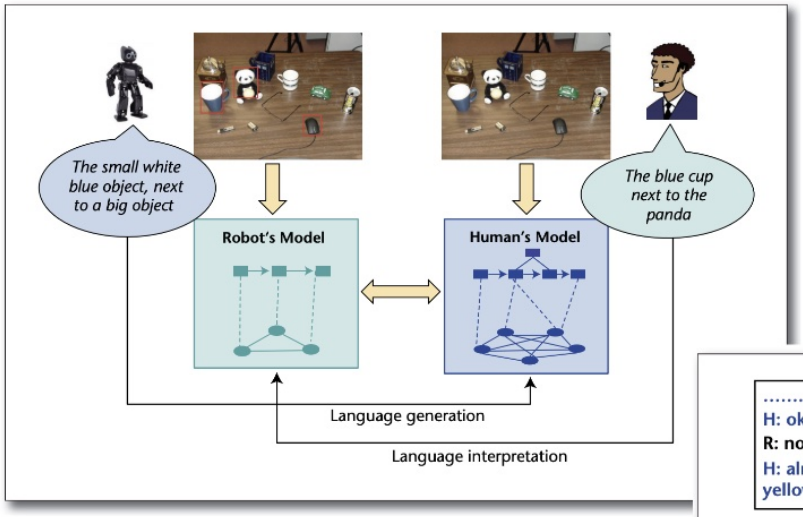


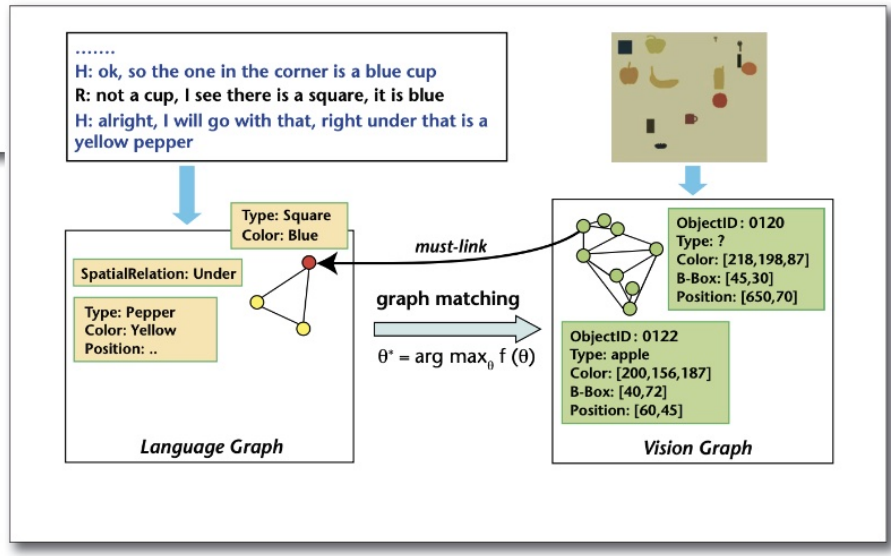
Figure 1: In the Minecraft Collaborative Building Task, the Architect (**A**) has to instruct a Builder (**B**) to build a target structure. **A** can observe **B**, but remains invisible to **B**. Both players communicate via a chat interface. (NB: We show **B**'s actions in the dialogue as a visual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

Situated Grounding in Human Robot Dialogue



- Establish a Joint Perceptual Basis through language grounding



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

- Graph-Matching for Interpreting Referring Expressions

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*).

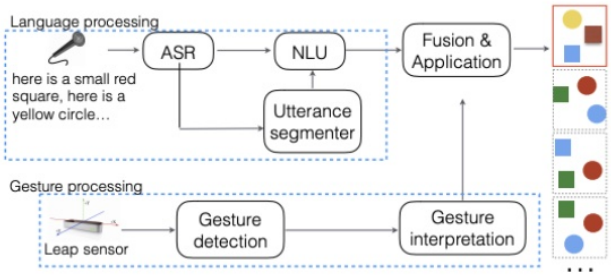


Figure 2: Multimodal system architecture.

- 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

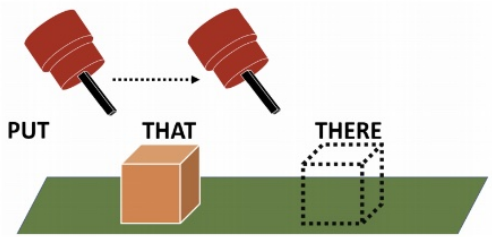


Figure 1: A pick-and-place task requires a *referential* pointing action to the object (orange cube) at the initial position, and a *locating* pointing action to a final placement position (dotted cube). Such an action by a robot (in red) can also be accompanied by verbal cues like *"Put that there."*

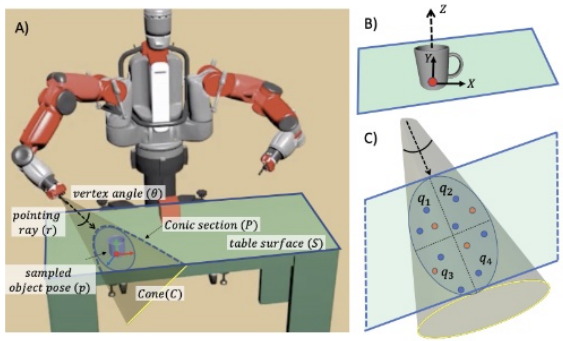


Figure 2: (A) Workspace setup showing the pointing cone and the corresponding conic section on the table. (B) The degrees-of-freedom considered for placement of the object on the table. (C) Sampling policy to sample object poses within the conic section.

- 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).

The Components of Multimodal Communication

- Achieving and maintaining common ground
  - shared conceptual space
- Context-aware interpretation of communicative acts
  - language, gesture, gaze
- Recognizing Object-specific knowledge and behavior
- Objects and actions are situated in the interaction
- Agents are embodied in the interaction:
  - all actions (communicative or physical) are interpreted through embodiment.
- Generative Lexicon object semantics

Situated Semantic Grounding and Embodiment

- 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.

Situated Meaning

Mother and son interacting in a shared task of icing cupcakes

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).*

Situated Meaning

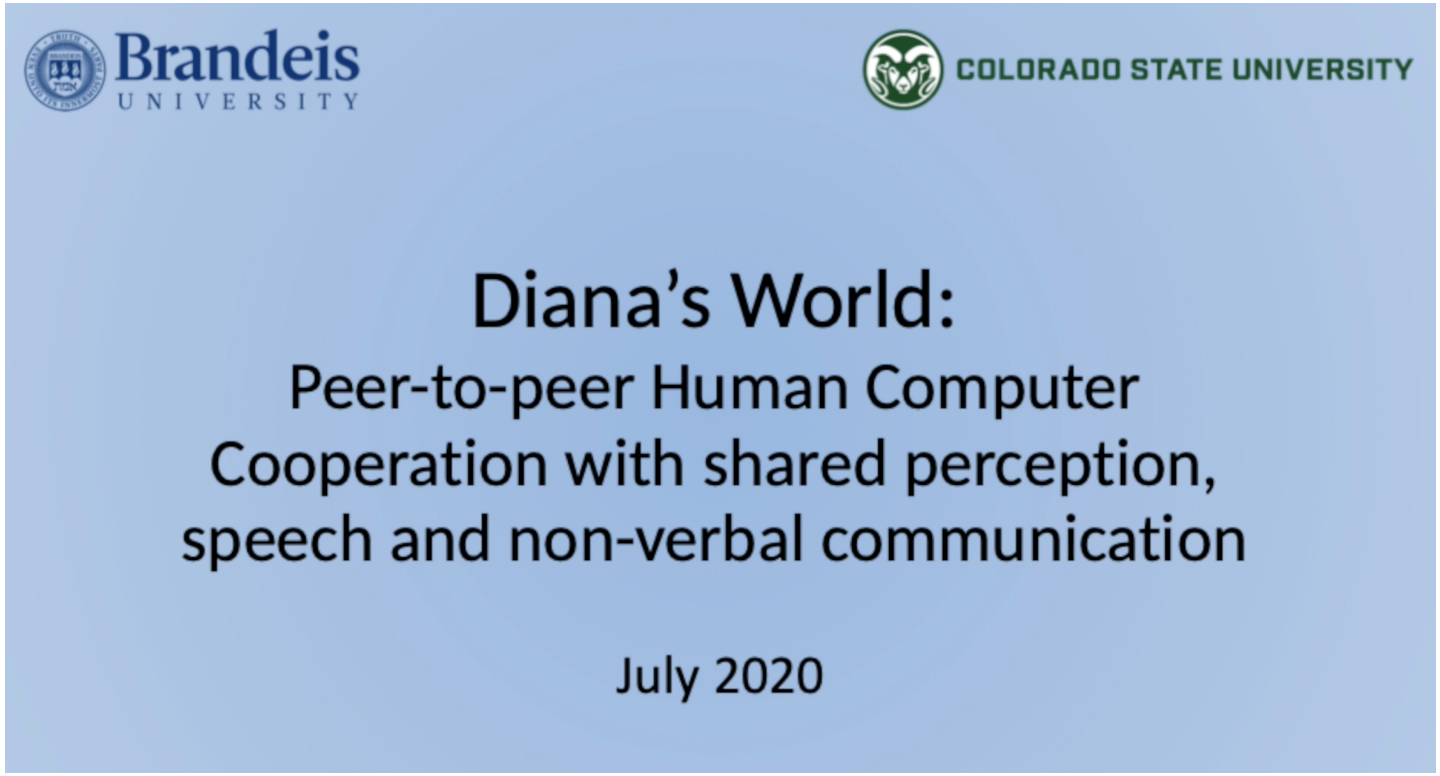
Elements from the Common Ground

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

The Challenge of Situated Grounding

1. Human-computer/robot interactions require at least the following capabilities:
   - Robust recognition and generation within multiple modalities
     - language, gesture, vision, action;
   - understanding of contextual grounding and co-situatedness;
   - appreciation of the consequences of behavior and actions.
2. Multimodal simulations provide an approach to modeling human-computer communication by situating and contextualizing the interaction, thereby visually demonstrating what the computer/robot sees and believes.

Diana's World



▸ Link

The Meaning of Embodiment in Communication

- Agent has situated meaning for the objects and actions in the environment;
- Recognition of the human's embodiment; agent has awareness of people's linguistic and gestural expressions, facial expressions, and actions.
- Self-embodiment of the agent: the agent has "spatial presence" within the domain of the interaction

Shared Conceptual Space

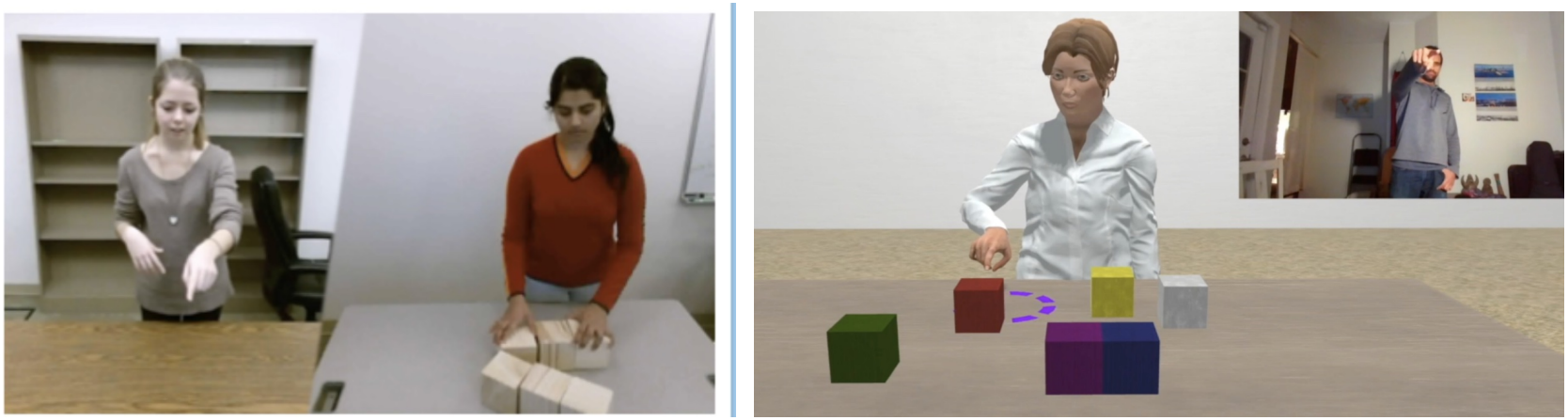


Figure: *Left:* Human-human collaborative interaction; *Right:* Human-avatar interaction.

Embodiment and Situated Meaning

- Elements of Situated Meaning
  - 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
- Modalities Deployed
  - gesture recognition and generation
  - language recognition and generation
  - affect, facial recognition, and gaze
  - action generation

Recognition of Human's Embodiment

Awareness of the partner's:

- linguistic and gestural expressions
- gestural expressions
- facial expressions
- gaze and eye tracking
- actions

The agent continuously constructs and maintains a representation of the embodiment of its human partner.

Agent Self-embodiment

"spatial presence" within the domain of the interaction

- facial "countenance"
- explicit effectors for action
- explicit sensors for audio and visual input.
- Constraints on its behavior are imposed by the physical extents and limitations of the embodiment (e.g., how far it can reach, degrees of freedom on the joints, etc.).

Intelligent Virtual Agents

Embodied Environments

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

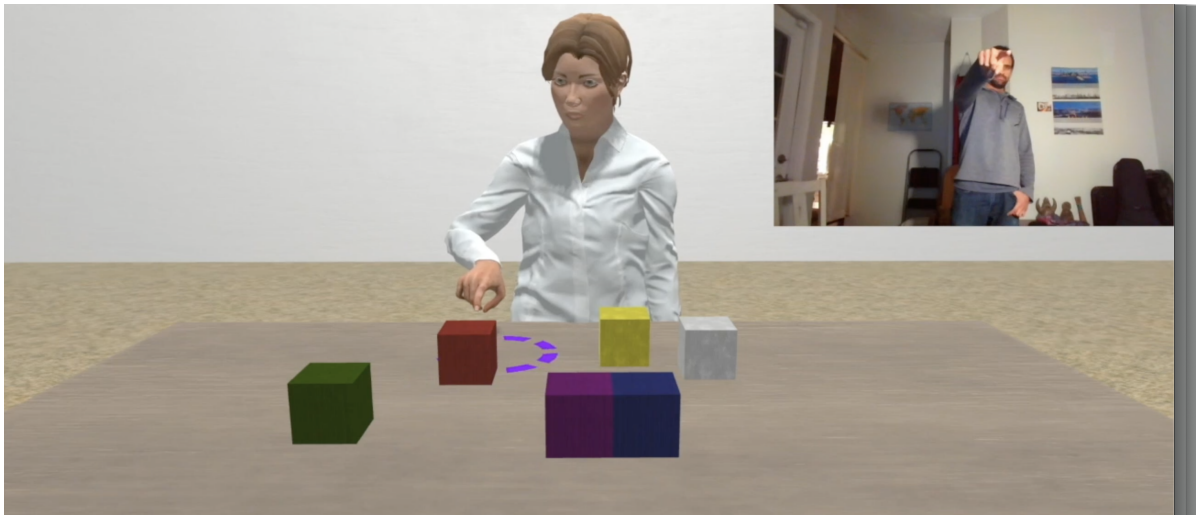


Figure: IVA Diana engaging in an embodied HCI with a human user.

▸ Link

Modeling Human Object Interactions (HOIs)

- The objects in a dialogue carry much more semantic information than conventionally assumed.
- This includes knowledge for how the objects can be manipulated by an agent in space and time, their *Gibsonian affordances*, and how they can be used, their *Telic affordances*.
- Such information also includes knowledge of how an object is situated in the environment relative to an agent for specific purposes and actions, that is, its *habitat*.
- We show how affordance encoding and recognition can improve object and action classification in HCI tasks.

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

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.

Captions Don't Describe Human-Object Interactions

"A man working at a desk."

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.

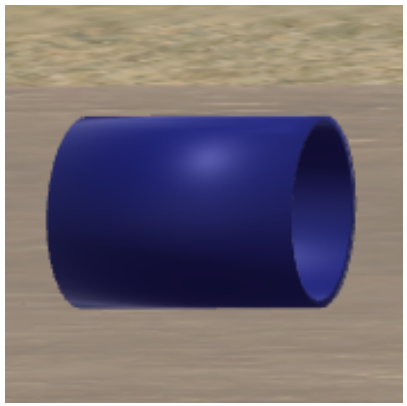
Affordance and Goal Recognition

1. Perceived purpose is an integral component of how we interpret situations and reason about utterances in communicative contexts.
   - Events are purposeful and directed;
   - Places are functional;
   - Objects are usable and manipulable.
2. Affordances are latent action structures of how an agent interacts with objects in the environment, in different modalities:
   - language, gesture, vision, action;
3. Qualia Structure provides a link to such latent actions structures associated with objects in utterances and the context.

Encoding Object Behavior

- 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."



Extending Qualia to Modeling Affordances

The affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill. It implies the complementarity of the animal and the environment. (J. J. Gibson, 1979/1986)

- Gibson (1979), Turvey (1992), Steedman (2002), Sahin et al (2007), Krippendorff (2010);
- Affordance: a correlation between an agent who acts on an object with a systematic or prototypical effect.

Habitats and Simulations

Pustejovsky (2013)

- Habitat: a representation of an object situated within a partial minimal model; Enhancements of the qualia structure.
- With multi-dimensional affordances that determine how habitats are deployed and how they modify or augment the context.
- Compositional combinations of procedural (simulation) and operational (selection, specification, refinement) knowledge.
- A habitat:
  - embeds;
  - orients;
  - positions.

Different Habitats for Object Use

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

Habitat-Affordance Pairs

If **Habitat** *then* **Action**

spoon

(2) a. If *spoon's concavity is vertical*, then it can *support containment of a substance*;
b. If *spoon's major axis is vertical*, then it can *support mixing*.

knife

(3) a. If *knife's zero convexity (sheet) is horizontal*, then it can *support spreading of a substance*;
b. If *knife's zero convexity (sheet) is vertical*, then it can *support cutting or separating*.

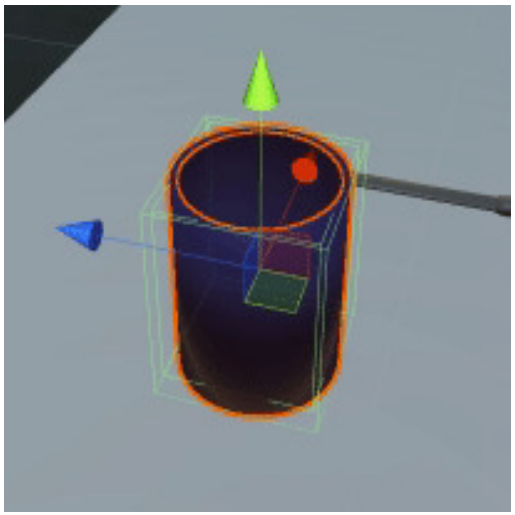
Visual Object Concept Modeling Language (VoxML)

Pustejovsky and Krishnaswamy (2016)

- 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: Conditioning environment affecting object **affordances** (behaviors attached due to object structure or purpose);
  - 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

VoxML for Actions and Relations

$$\left[\begin{array}{l}\textbf{slide} \\ \text{LEX} = \left[\begin{array}{l}\text{PRED} = \textbf{slide} \\ \text{TYPE} = \textbf{process}\end{array}\right] \\ \text{TYPE} = \left[\begin{array}{l}\text{HEAD} = \textbf{process} \\ \text{ARGS} = \left[\begin{array}{l}\text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:physobj} \\ \text{A}_3 = \textbf{z:physobj}\end{array}\right] \\ \text{BODY} = \left[\begin{array}{l}\text{E}_1 = grasp(x,y) \\ \text{E}_2 = [while(hold(x,y), \\ \qquad while(EC(y,z), move(x,y)))]\end{array}\right]\end{array}\right]\end{array}\right]$$

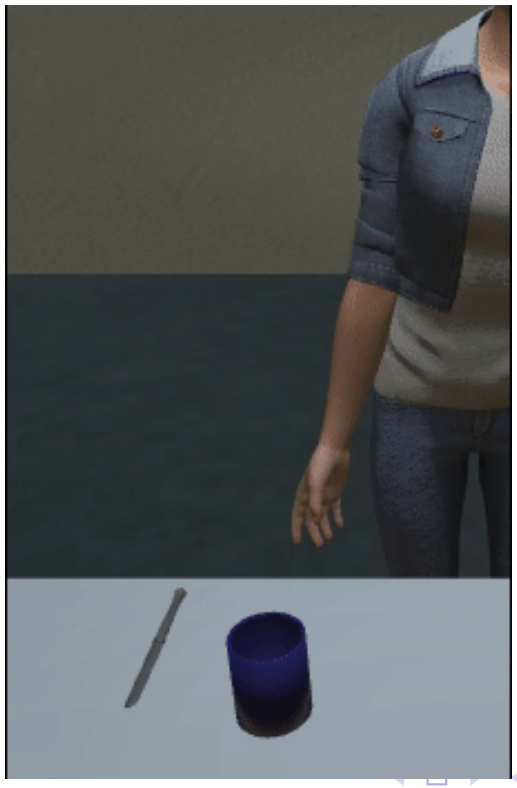
$$\left[\begin{array}{l}\textbf{put} \\ \text{LEX} = \left[\begin{array}{l}\text{PRED} = \textbf{put} \\ \text{TYPE} = \textbf{transition\_event}\end{array}\right] \\ \text{TYPE} = \left[\begin{array}{l}\text{HEAD} = \textbf{transition} \\ \text{ARGS} = \left[\begin{array}{l}\text{A}_1 = \textbf{x:agent}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

VoxML - grasp

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

VoxML - grasp cup

- Continuation-passing style semantics for composition
- Used within conventional sentence structures and between sentences in discourse in MSG



Multimodal Simulations

- Human understanding depends on a wealth of common-sense knowledge; humans perform much reasoning qualitatively.
- To simulate events, every parameter must have a value
  - "Roll the ball." How fast? In which direction?
  - "Roll the block." Can this be done?
  - "Roll the cup." Only possible in a certain orientation.
- VoxML: Formal semantic encoding of properties of objects, events, attributes, relations, functions.
- VoxSim: What can situated grounding do? (Krishnaswamy, 2017)
  - Exploit numerical information demanded by 3D visualization;
  - Perform qualitative reasoning about objects and events;
  - Capture semantic context often overlooked by unimodal language processing.

VoxWorld Architecture

Pustejovsky and Krishnaswamy (2016), Krishnaswamy (2017), Pustejovsky et al (2017), Narayana et al (2018)

- Dynamic interpretation of actions and communicative acts:
  - Dynamic Interval Temporal Logic (DITL)
  - Dialogue Manager
- VoxML: Visual Object Concept Modeling Language
- EpiSim: Visualizes agent's epistemic state and perceptual state in context;
  - Public Announcement Logic
  - Public Perception Logic
- VoxSim: 3D visualizer of actions, communicative acts, and context.
  - Built on Unity Game Engine

Dynamic Discourse Interpretation

- Common Ground Structure
  - Co-belief
  - Co-perception
  - Co-situatedness
- Multimodal communication act:
  - language
  - gesture
  - action
- Dynamic tracking and updating of dialogue with:
  - Discourse Sequence Grammar
  - Gesture Grammar
  - Action Grammar

Multimodal Communicative Acts

- A communicative act, performed by an agent, a, is a tuple of expressions from the modalities available to a, involved in conveying information to another agent.
- We restrict this to the modalities of speech, S, gesture, G, facial expression F, gaze Z, an explicit action A.
  1. $C_a = \langle S, G, F, Z, A \rangle$
- These modal channels can be aligned or unaligned in the input.

Co-belief and Co-perception in the Common Ground

- *Public announcement logic* (PAL)
  - $[\alpha]\varphi$ denotes that an agent "α *knows* φ".
  - Public Announcement: $[!\varphi_1]\varphi_2$
  - Any proposition, φ, in the common knowledge held by two agents, α and β, is computed as: $[(\alpha \cup \beta)^*]\varphi$.
- *Public perception logic* (PPL)
  - $[\alpha]_\sigma\varphi$ denotes that agent "α perceives that φ".
  - $[\alpha]_\sigma\hat{x}$ denotes that agent "α perceives that there is an x."
  - Public Display: $[!\varphi_1]_\sigma\varphi_2$
  - The co-perception by two agents, α and β includes φ : $[(\alpha \cup \beta)^*]_\sigma\varphi$

Multimodal Semantics for Common Ground

Common Ground Structure (CGS)

The situated common ground consists of the following state information:

(4) a. **A**: The agents engaged in communication;
b. **B**: The shared belief space;
c. **P**: The objects and relations that are jointly perceived in the environment;
d. $\mathcal{E}$: The embedding space that both agents occupy in the communication.

(5)

A:a_1, a_2 **B**:Δ **P**:b
$\mathcal{S}_{a_1}$ = "You$_{a_2}$ see it$_b$"

$\mathcal{E}$

Multimodal Semantics for Common Ground

Modeling the Current Context

- State Monad: $M\alpha = State \rightarrow (\alpha \times State)$
- Context is a stack of items and the type of left contexts is a list of entities, $[e]$.
- Right contexts will be interpreted as continuations: a discourse that requires a left context to yield a truth value., of type $[e] \rightarrow t$.
- Hence, context transitions are of type $[e] \rightarrow [e] \rightarrow t$;
- Given the current discourse, T, and a new expression, C, C updates T as follows:
- $[\![\overline{(\mathbf{T}.\mathbf{C})}]\!]^{M,cg} = \lambda k.[\![\overline{\mathbf{T}}]\!](\lambda n.[\![\overline{\mathbf{C}}]\!](\lambda m.k(m\ n)))$

Adding Gesture to Common Ground

Multimodal Contextualized Reference

- Representing how gestures denote
- Encoding co-perception of situated objects under reference
- Situated alignment of expressions from distinct modalities

Gesture Types in Multimodal Interactions

1. Deixis (pointing) gestures, generated to request information regarding an object, a location, or a direction when performing a specific action;
2. Iconic action gestures, generated to request clarification on how (what manner of action) to perform a specific task;
3. Affordance-denoting gestures, generated to describe how the agent can interact with an object, even when it does not know what it is or what it might be used for;
4. Direct situated actions, where the agent responds to a command or request by acting in the environment directly.

Gestures used in VoxWorld

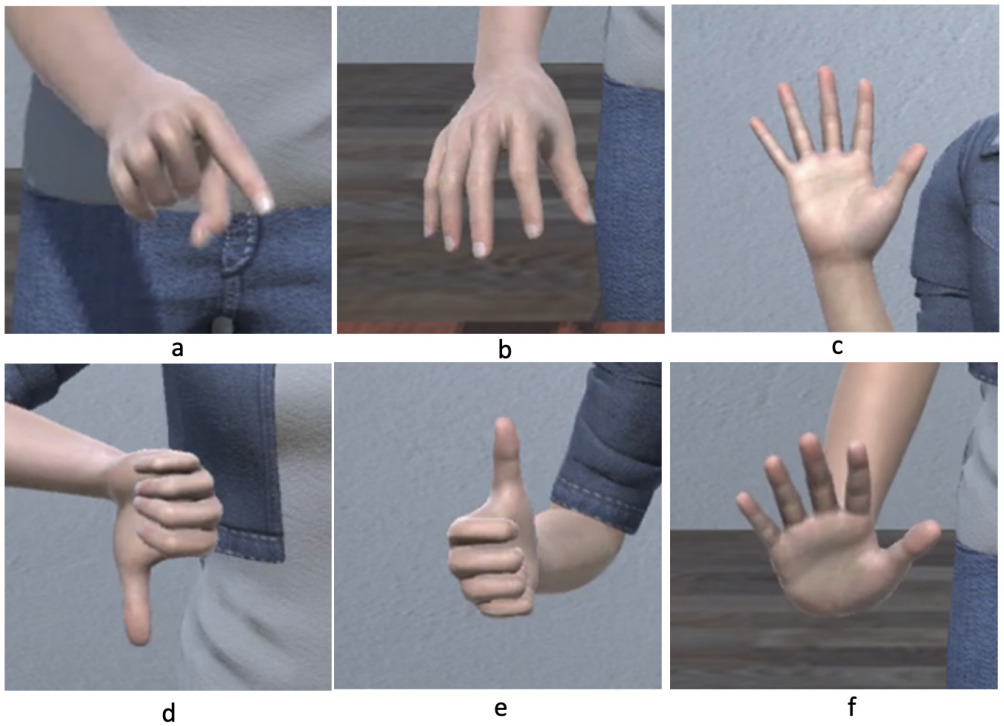
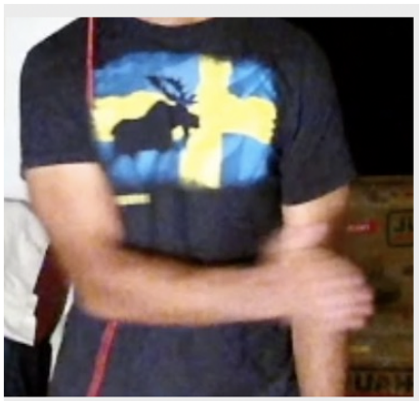


Figure: Some of the gestures generated by VoxWorld: pointing, grab, five, no, yes, push back.

Bidirectional Gesture Recognition and Generation

- On the left, a human is action gesturing to move an object to the left:
- On the right, the IVA is performing the identical gesture.




Actions as Described by Gesture

Kendon (2004), Lascarides and Stone (2009)

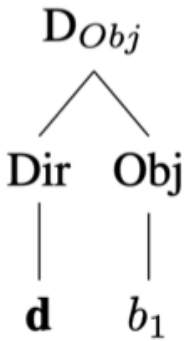
- $G \rightarrow (Prep)\ (Pre\_stroke\ Hold)$ Stroke $Retract$

The stroke is the content-bearing phase, **d**, and in a pointing gesture, will convey the deictic orientational information.

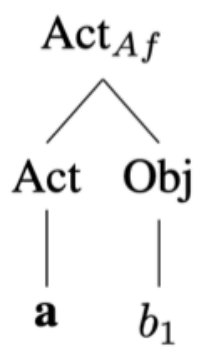
- $[\![point]\!] = [\![End(cone(\mathrm{d}))]\!]$

Interpreted Gesture Tree:

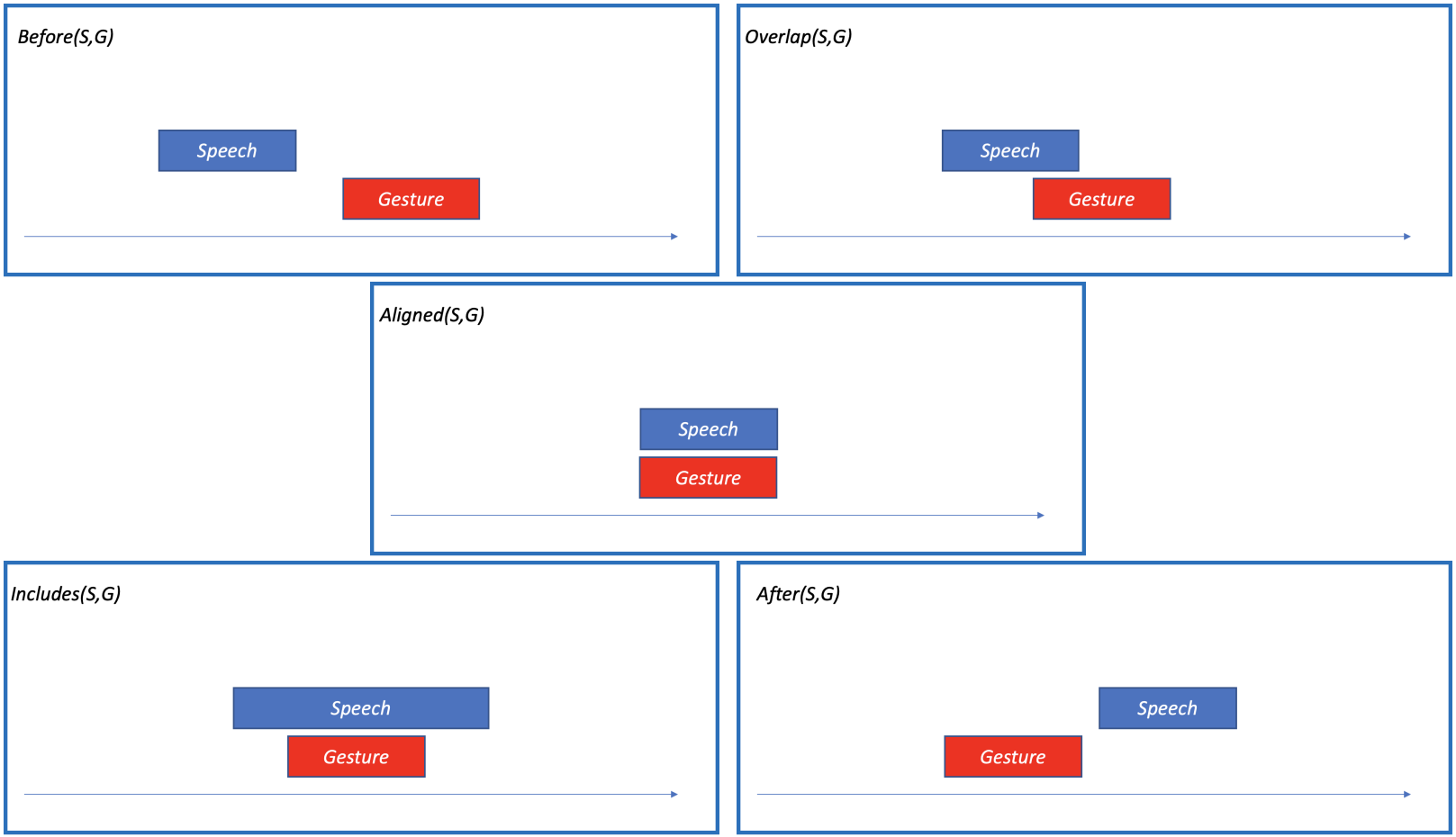
a. **Deixis**: $D_{obj} \rightarrow Dir\ \ Obj$



b. **Action**: $G_{Af} \rightarrow Act\ \ Obj$



Aligning Speech and Gesture in Dialogue



Break

BREAK
We will resume in 15 minutes

Reasoning in Multimodal Simulations

- Human understanding depends on a wealth of common-sense knowledge; humans perform much reasoning qualitatively.
- To simulate events, every parameter must have a value
  - "Roll the ball." How fast? In which direction?
  - "Roll the block." Can this be done?
  - "Roll the cup." Only possible in a certain orientation.
- VoxML: Formal semantic encoding of properties of objects, events, attributes, relations, functions.
- VoxSim: What can situated grounding do? (Krishnaswamy, 2017)
  - Exploit numerical information demanded by 3D visualization;
  - Perform qualitative reasoning about objects and events;
  - Capture semantic context often overlooked by unimodal language processing.

Transfer Learning of Object Affordances

- 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.}

Transfer Learning of Object Affordances

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., $H_{[3]} = [\text{UP} = \mathit{align}(\bar{Y}, \mathcal{E}_Y), \text{TOP} = \mathit{top}(+Y)]$, $H_{[3]} \rightarrow [\mathit{put}(x, \mathit{in}(\mathit{this}))]\mathit{contain}(\mathit{this}, x)$;
- Train 200-dimensional habitat or affordance embeddings using a Skip-Gram model;
- Represent objects as averaged habitat or affordance vectors.

Transfer Learning of Object Affordances

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

Transfer Learning of Object Affordances

Affordance Embeddings

- 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%**

Learning with Affordances

Affordance Embeddings

Tests on individual objects (plate):



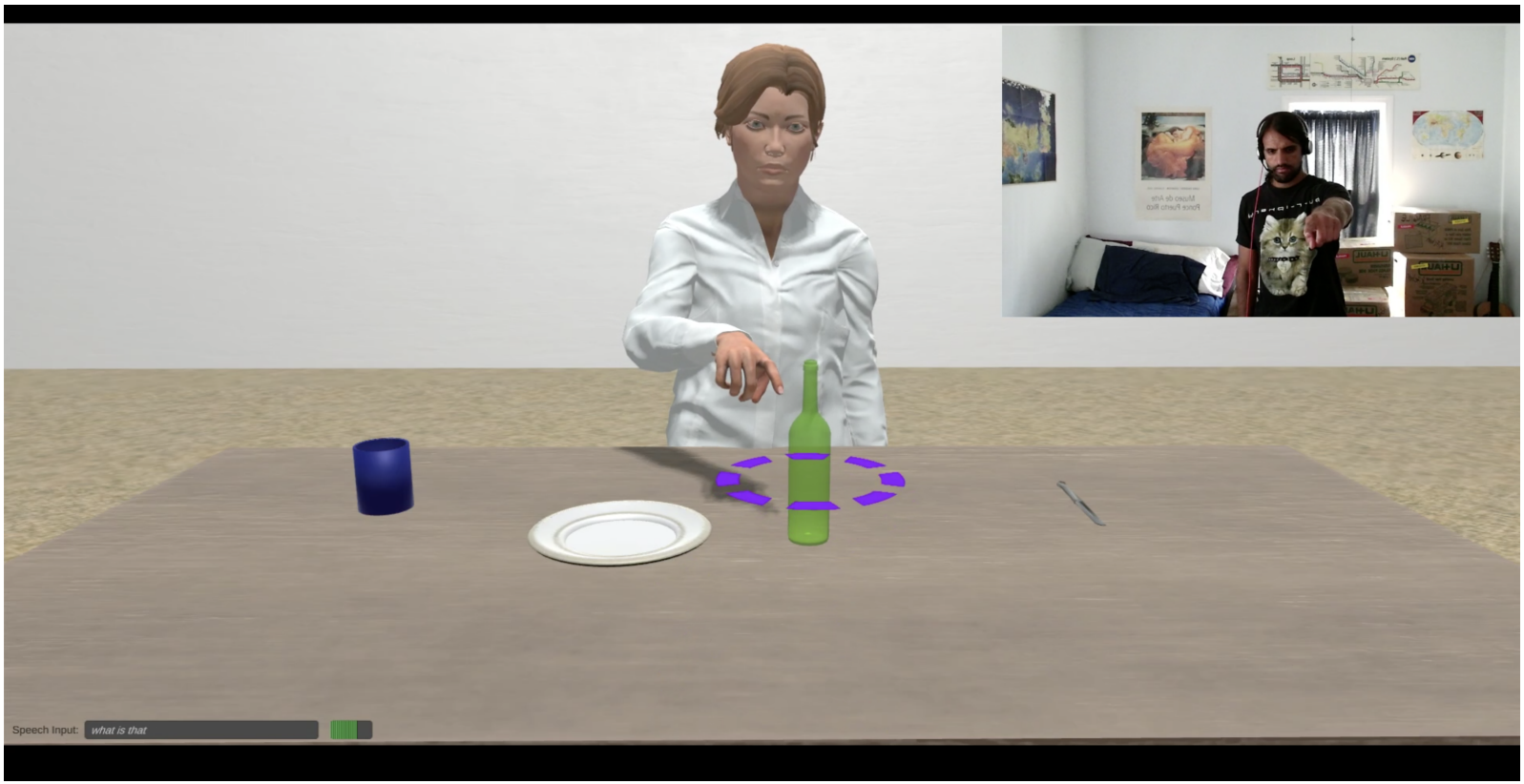


Model	MLP-H	MLP-A	CNN-H	CNN-A
Predicted objects	book, cup, bowl, bottle	cup, bottle, apple	book	cup, bottle

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

Transfer Learning of Object Affordances

Affordance Embeddings



▶ Play!

Reasoning with Affordances

Learning how to stack a cube

- An agent can interact with various objects and see how they behave differently under the same circumstances.
- An agent can learn to distinguish objects based on behaviors.

Reasoning with Affordances

Learning how to stack a cube

- Goal: select best numerical action to keep stacked block stable
- Optimal action: places theme block centered atop destination block
  - Can be moved (perturbed) anywhere within the search space
- Episode terminates on success, or 10 failures
- Reward shaping:
  - 1000 for stacking successfully first time
    - -100 for each additional attempt (e.g., 900 for success on second try)
  - 9 for touching destination block but falling off
  - -1 for missing destination block entirely e.g., a 3-attempt episode where agent 1) misses destination block; 2) touches destination block but doesn't stack; 3) stacks successfully $= -1 + 9 + 800 = 808$ total return

Reasoning with Affordances

Learning how to stack a cube

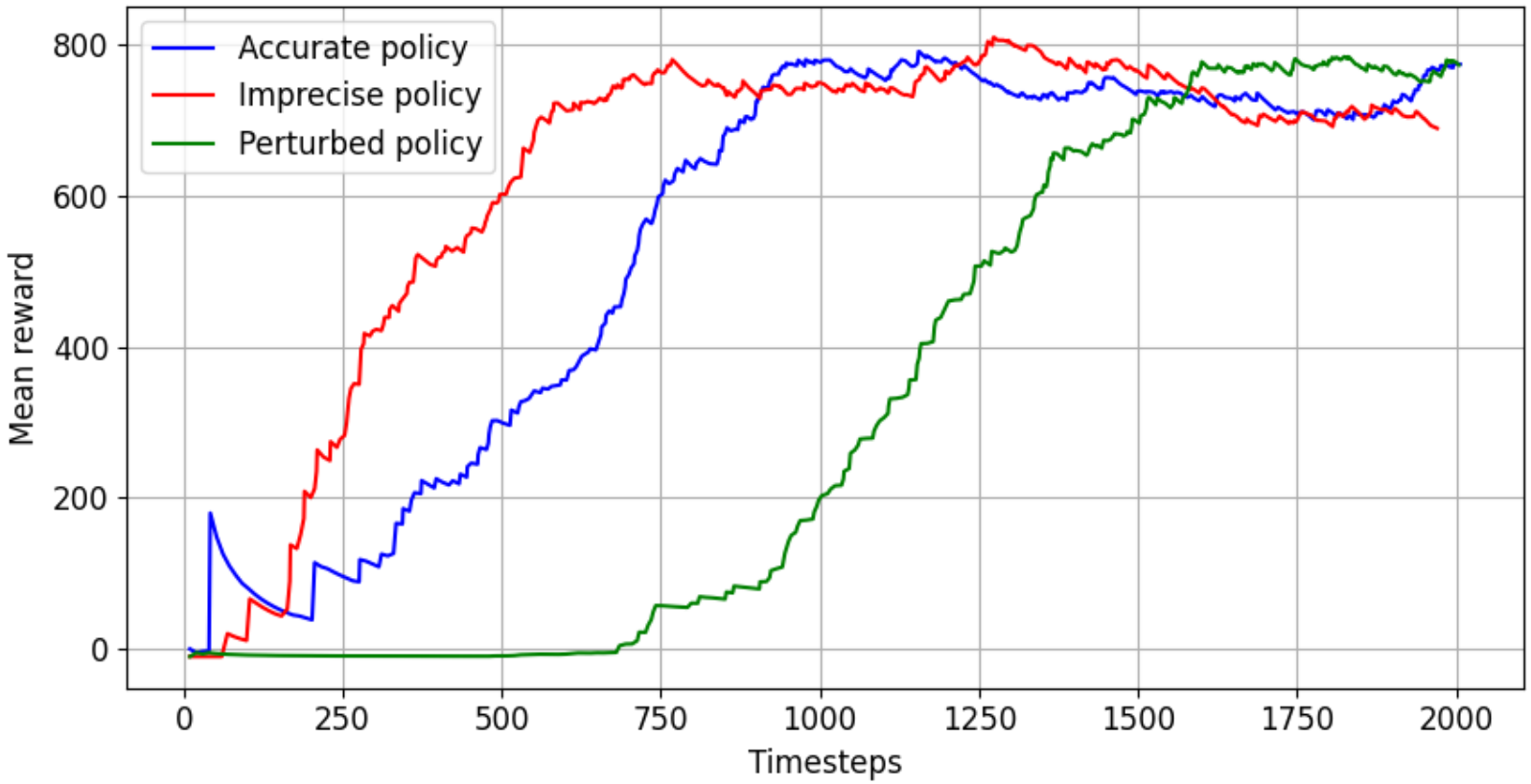
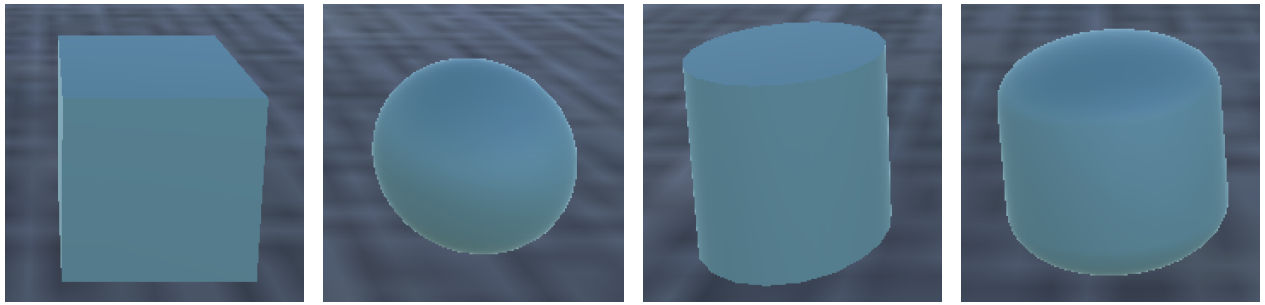


Figure: TD3 training reward plots (2,000 training episodes)

Reasoning with Affordances

Learning how to stack a cube

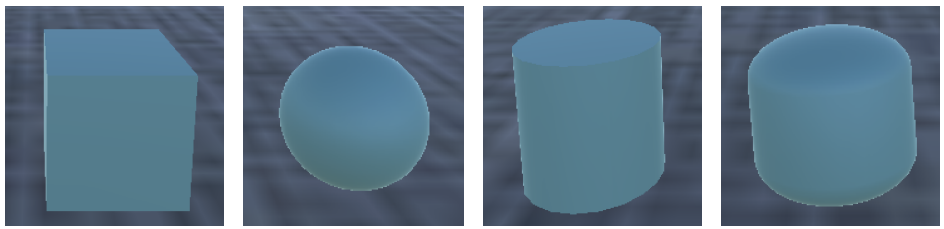
- An RL policy that can successfully stack cubes can... successfully stack cubes
  - and not much else
- We then use the successful cube-stacking policy to make the agent attempt to stack other spheres, cylinders, and capsules on a block:
  - forcing it to stack the other objects *as if they were cubes*.



Reasoning with Affordances

Learning the different affordances of objects

- This control structure allowed us to identify differences in the behaviors of the different objects in the stacking task.
- These behaviors can be described in terms like cubes stack successfully, spheres roll off, cylinders stack if oriented properly, etc.
  - Differences in behavior can be characterized in terms of the object's *affordances*.



Reasoning with Affordances

Learning the different affordances of objects

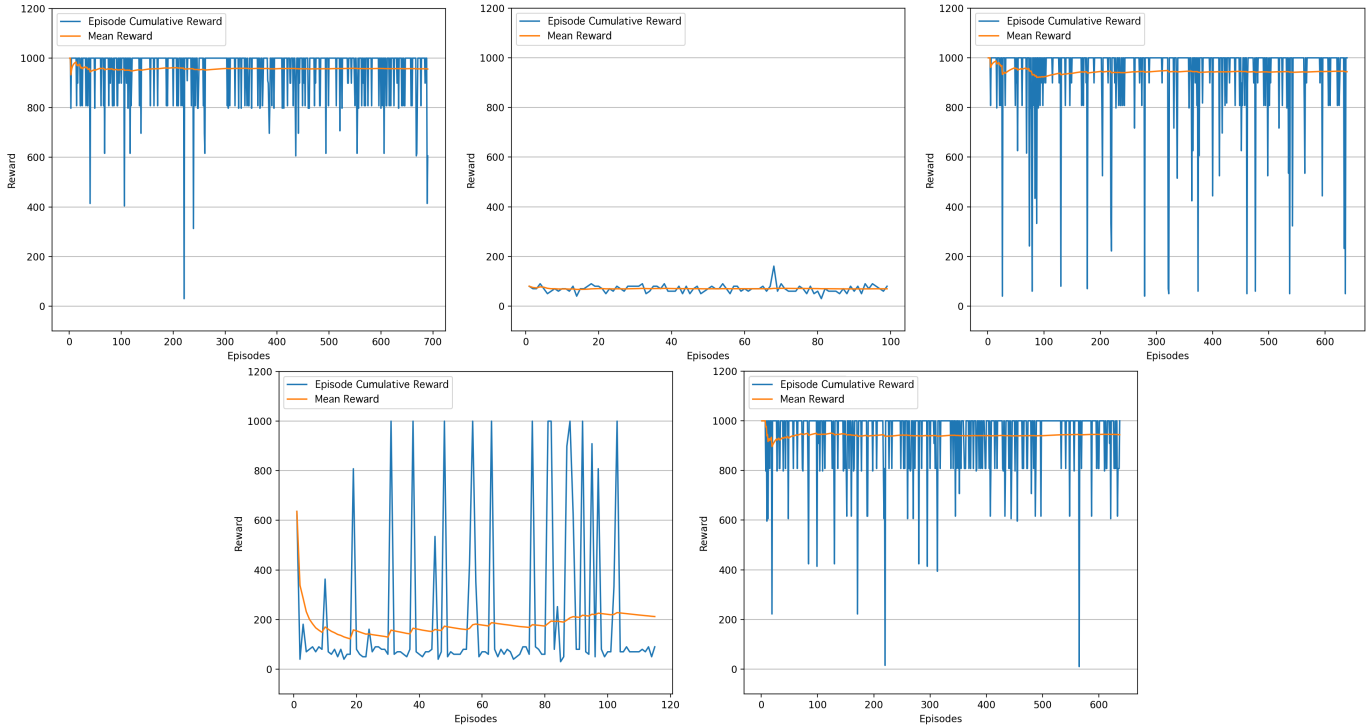


Figure: Evaluation reward plots for stacking (in order) a **cube**, **sphere**, **cylinder**, **capsule**, and **small cube** on a cube

Executing the Trained Policy over Objects

Collect object information for each stacking attempt in VoxSim

- During evalution, we store: Object type, rotation at start, offset (radians) between world and object upright axes at start, action executed, rotation and offset from world after action, state after action, reward, cumulative total and mean rewards over episode.
- At the end of the action, a small "jitter" is applied, to simulate the small force exerted on an object when it is released from a grasp, in an otherwise hyper-precise virtual environment.
- The post-action jitter is a small force perpendicular to the object's axis of symmetry, therefore implicitly encoding information about the object's habitat, *as encoded in VoxML*.
- The stackability (or lack thereof), encoded in the distribution of state observations, implicitly encodes an affordance.

Executing the Trained Policy over Objects

Collect object information for each stacking attempt in VoxSim

Object	Jitter			θ after Action	Stack Height
cube	-1.472×10^{-4}	0	2.021×10^{-4}	0.02238165	2
sphere	8.165×10^{-5}	0	-2.363×10^{-5}	2.134116	1
cylinder	0	0	2.5×10^{-4}	0.01457105	2
cylinder	0	0	2.5×10^{-4}	1.570793	1

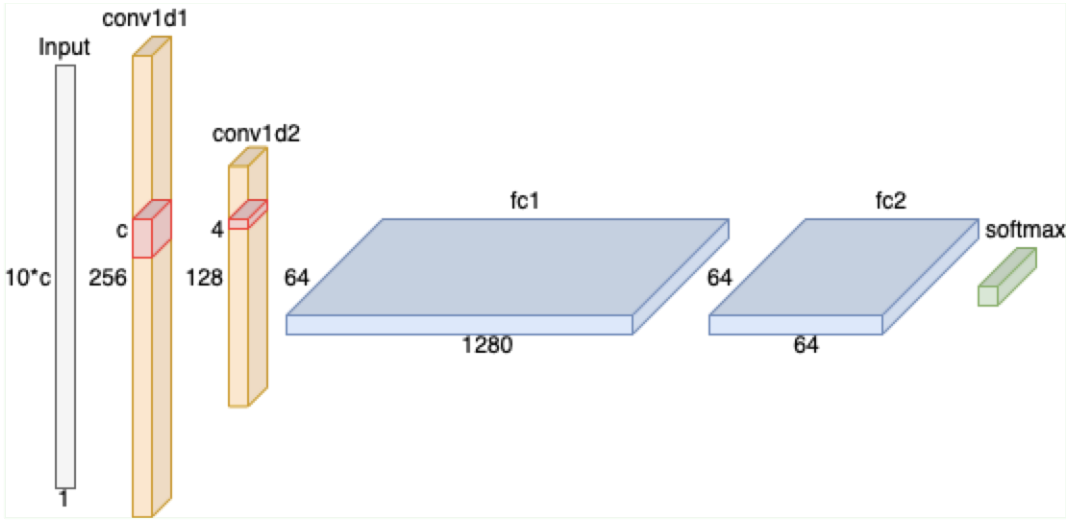
Table 1: Observations gathered during stacking task with multiple objects.

- A cube, which is flat on all sides, can rest stably (multiples of $\frac{\pi}{2}$)
- The sphere rolls off, does not stack (height 1), comes to rest at an arbitrary angle.
- The cylinder shares properties with cubes (flat ends) and others with spheres (round sides), in the last two rows.
  - the cylinder stacks successfully (height 2), and is resting upright ($\theta \approx 0$)
  - it rolls off with the cylinder on its side ($\theta \approx \frac{\pi}{2}$)

Using Habitat and Affordance Embeddings

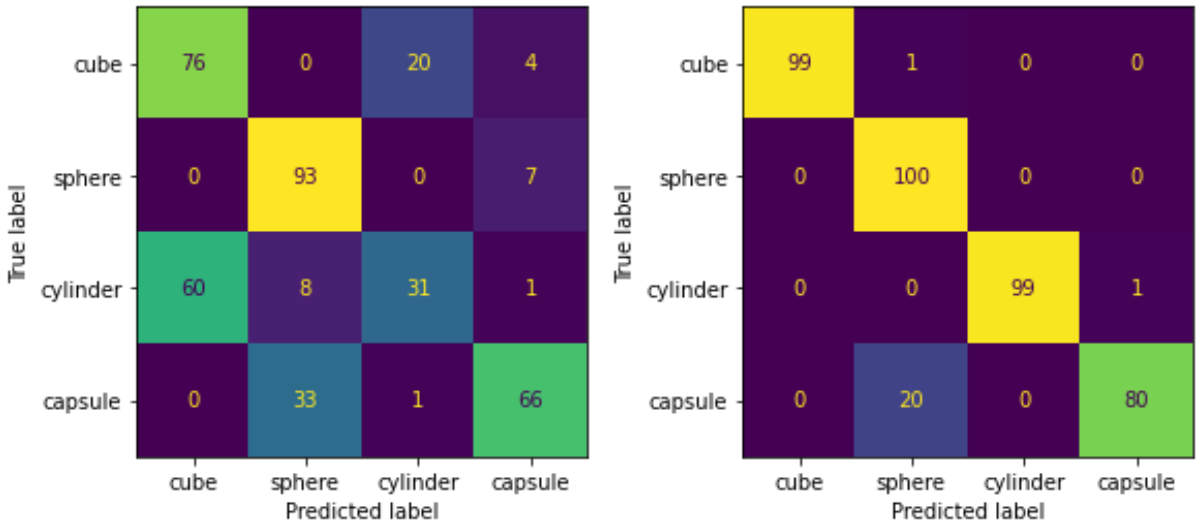
Training a model to predict object type from its behavior in stacking task

- 1D convolutional neural net (2 convolutional layers—256 and 128 hidden units, 2 64-unit fully-connected layers, and a softmax layer)
- We train for 500 epochs using Adam optimizer, batch size of 100 (= 10 episodes), learning rate of 0.001.



Using Habitat and Affordance Embeddings

Training a model to predict object type from its behavior in stacking task



Left chart shows results without input of implicit habitat and affordance information encoded in post-action jitter (66.5%). Right chart shows results with those input features (94.5%).

Using Habitat and Affordance Embeddings

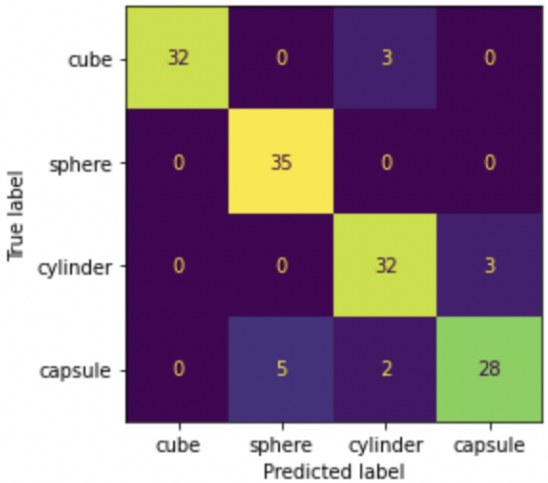
Training a model to predict object type from its behavior in stacking task

Can't you just tell if objects are different by looking at them?

Comparing Behavior with Visual Clues

Humans can tell that objects are different by visual clues

- We compared the performance of the behavior-based classifier to a 2D CNN CIFAR-10 style object detector.
- This classifier achieves a validation accuracy of 97.5%, but when evaluated against an unseen test set of 140 novel images of the four object classes (35 images each), accuracy falls below the behavior-based classifier, to **90.7%**.



Mistakes Made by the Visual Classifier

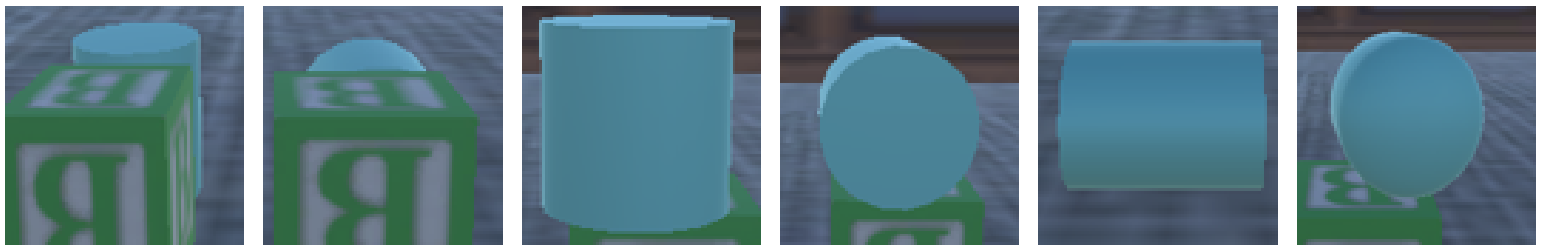
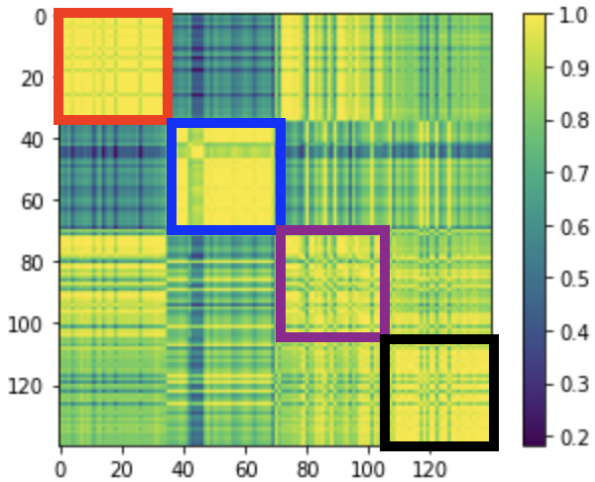


Figure: Sample of misrecognized objects. From left: cylinder misrecognized as cube (1), and capsule (2), cylinder misrecognized as cube (3), sphere (4), and capsule (5), and capsule misrecognized as sphere (6).

Evaluating the Embedding Vectors

- While some objects are visually distinct, like cube vs. sphere, other object classes are more difficult to distinguish visually.
- To confirm this, we draw out the 64-dimensional embedding vectors from the final fully-connected layer.
- These can be used to quantitatively assess the similarity of different input samples to each other



Cosine similarity matrix of visual embedding vectors from 2D CNN.

Multimodal Semantics

Sight and behavior

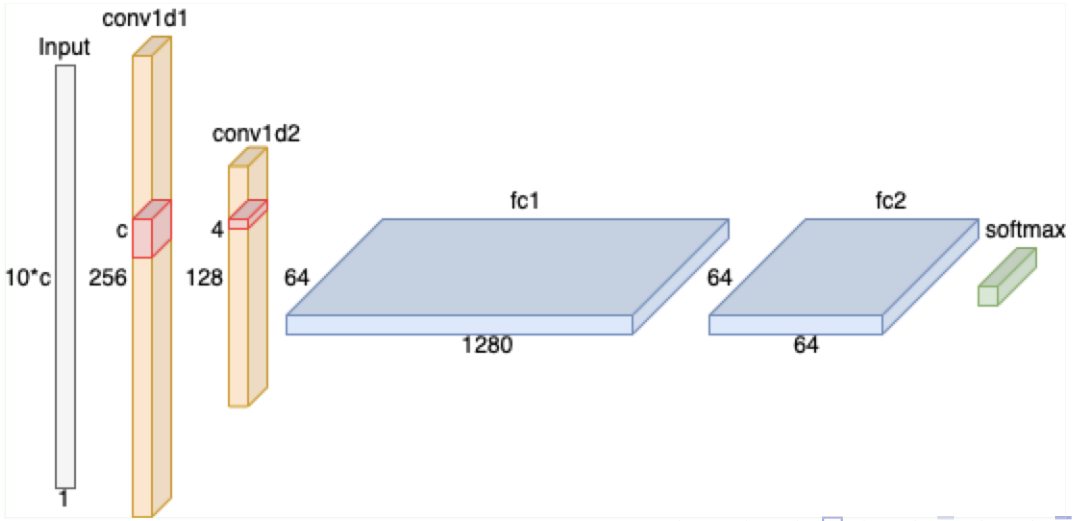
- Where vision may be ambiguous, behavior may be a stronger signal
- **Embodiment** drives **affordance** reasoning
- Embodiment (and grounding) is more than only language+vision
- Integrated multimodal semantics a la VoxML facilitates extraction of quantified *embodied* information from an environment
- Which leads to...

Novel Class Detection

- Weakness of a forced-choice classifier:
  - *Will* output one of its known classes for any input;
  - Softmax layer obscures probability, confidence, representation.
- Strength of a forced-choice classifier:
  - Embedding-level representations preserve similarity across dimensions.
- Novelty detection procedure:
  1. Identify which known class an object is most similar to;
  2. Determine if new object is different enough from most similar known class to be considered novel.

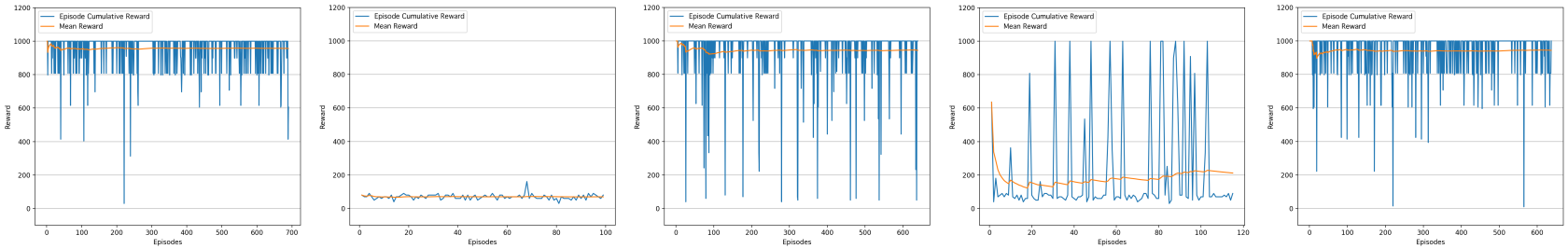
Novel Class Detection

- CCA shows that cube and sphere are the two mos dissimilar objects
- Correlates to prototypical "stackable" and "unstackable" objects
- Instantiate behavior CNN classifier with two classes: cube/sphere



Novel Class Detection

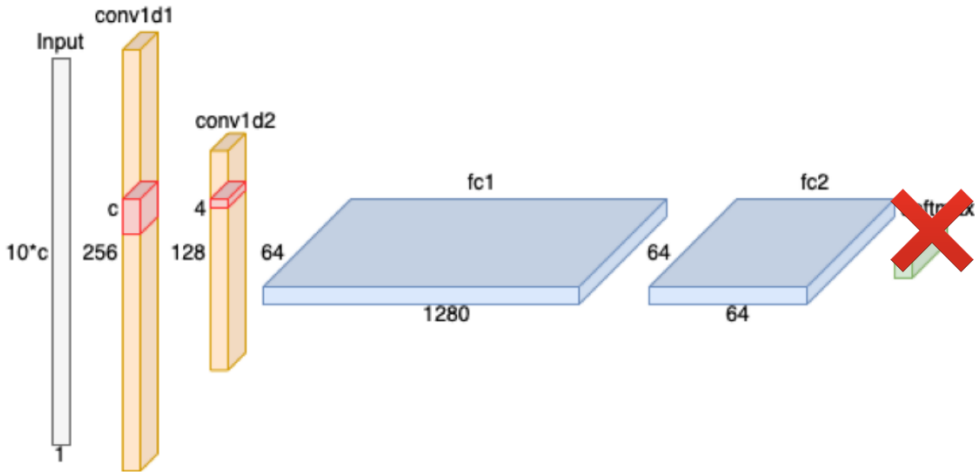
- Given known classes cube and sphere, cylinders usually classified as cubes, capsules usually classified as spheres



- Recapitulates observations from CCA and evaluation reward plots

Novel Class Detection

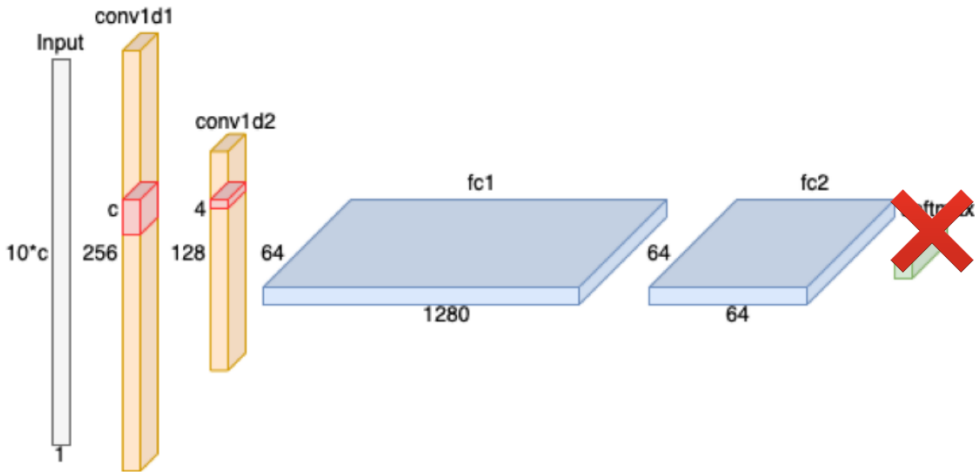
- If something is classified as *sphere* is it actually a sphere?
  - Or just sphere-*like* and I don't know any better?



- Get the embeddings for the input sample(s) and compare them to embeddings model "knows" belong to the predicted class (i.e., training embeddings)

Novel Class Detection

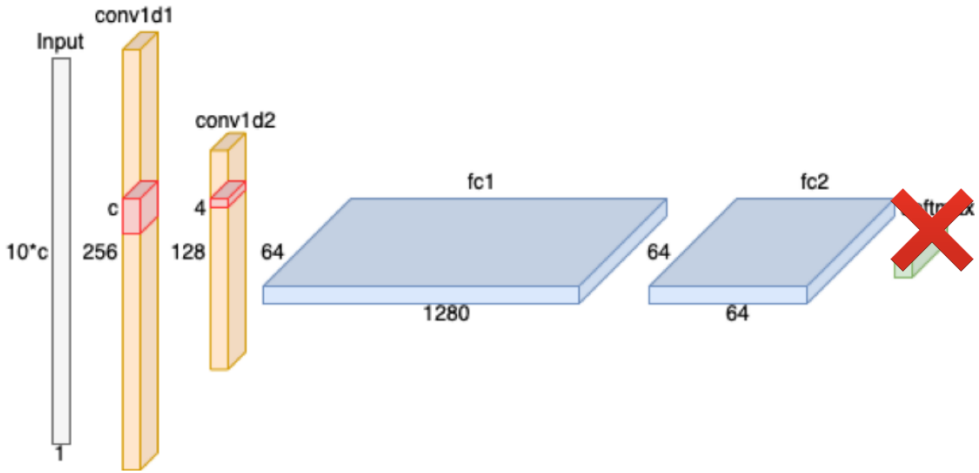
Intriguing properties of embedding spaces



- Not as simple as constant similarity threshold between embedding vectors
- Differences in weight initialization may mean embeddings are differently distributed each time

Novel Class Detection

Intriguing properties of embedding spaces



- Same network, same data, same training, different initial weights may mean
  - One trained model has dispersed or even near-isotropic vectors
  - One trained model clusters all vectors close together in high-D space; absolute distance between classes is small

Novel Class Detection

- **This is now an outlier detection problem**
- Let $\overrightarrow{\mu_S}$, $\overrightarrow{\sigma_S}$ be the mean, stdev of the known class
- Let $\overrightarrow{\mu_N}$ be the mean of the new batch
- Let $\overrightarrow{v}$ be a single sample

Novel Class Detection

- Let $\overrightarrow{\mu_S}$, $\overrightarrow{\sigma_S}$ be the mean, stdev of the known class
- Let $\overrightarrow{\mu_N}$ be the mean of the new batch
- Let $\overrightarrow{v}$ be a single sample
- Compute $\rho_{\overrightarrow{v}}$, the ratio of $\overrightarrow{v}$'s distance from $\overrightarrow{\mu_S}$ to the overall spread of embedding vectors in class S
  - $\rho_{\overrightarrow{v}} = \frac{\cos(\overrightarrow{\mu_S}, \overrightarrow{v})}{\cos(\overrightarrow{\mu_S}, \overrightarrow{\mu_S} + \overrightarrow{\sigma_S})}$
- If $\rho_{\overrightarrow{v}} > 1$, add $\overrightarrow{v}$ to the set of outliers $\overrightarrow{o} \in O$
  - (outliers are defined relative to a sample set)

Novel Class Detection

- Compute $\rho_{\vec{v}}$, the ratio of $\vec{v}$'s distance from $\vec{\mu_S}$ to the overall spread of embedding vectors in class S
  - $\rho_{\vec{v}} = \frac{\cos(\vec{\mu_S}, \vec{v})}{\cos(\vec{\mu_S}, \vec{\mu_S} + \vec{\sigma_S})}$
- If $\rho_{\vec{v}} > 1$, add $\vec{v}$ to the set of outliers $\vec{o} \in O$
  - (outliers are defined relative to a sample set)
- Outlying samples may still belong to the known class
  - (e.g., sometimes a cube fails to stack properly due to bad placement)
- If $\frac{\rho_{\vec{o}} - \mu_\rho}{\mu_\rho} > 3$, remove $\vec{o}$ from outlier set

Novel Class Detection

- Outlying samples may still belong to the known class
  - (e.g., sometimes a cube fails to stack properly due to bad placement)
- If $\frac{\rho_{\overrightarrow{o}} - \mu_\rho}{\mu_\rho} > 3$, remove $\overrightarrow{o}$ from outlier set
- **What this does is define a subspace in 64D space bounded by non-extreme outlier vectors of each sample**
- Compute ratio of outliers in known class to outliers in new batch:
  - Outlier ratio $OR = \frac{\sum_{\overrightarrow{o_N} \in O_N} \rho_{\overrightarrow{o_N}}}{\sum_{\overrightarrow{o_S} \in O_S} \rho_{\overrightarrow{o_S}}}$
- Scale OR by distance between known class and new batch means, then normalize by spread of embeddings in known class times denominator OR

Novel Class Detection

- What this does is define a "core" subspace in 64D space bounded by non-outlier vectors of each sample
- Compute ratio of outliers in known class to outliers in new batch:
  - Outlier ratio $OR = \frac{\sum_{\overrightarrow{o_N} \in O_N} \rho_{\overrightarrow{o_N}}}{\sum_{\overrightarrow{o_S} \in O_S} \rho_{\overrightarrow{o_S}}}$
- Scale OR by distance between known class and new batch means, then normalize by spread of embeddings in known class times denominator OR
- Define a "dissimilarity threshold" T
- **If** $\frac{OR \times \mathbf{cos}(\overrightarrow{\mu_S}, \overrightarrow{\mu_N})}{\mathbf{cos}(\overrightarrow{\mu_S}, \vec{\mu_S} + \overrightarrow{\sigma_S}) \times \sum_{\overrightarrow{o_S} \in O_S} \rho_{\overrightarrow{o_S}}} > T$**, the new input samples likely belong to new class!**

Novel Class Detection

Results

- Evaluation: for a model beginning with cube and sphere labels:
  - there are 2 novel objects that may occur: cylinder and capsule
  - and small cube, which is the same as cube
    - (not considering size as a distinguishing factor as parameters indicating size are not captured in the raw data)
- Correct result: detecting cylinder and capsule as novel, and cube, sphere, and small cube as not novel
  - Correct result for model including cylinder: detecting capsule as novel, and all other classes as not novel, etc.
- Evaluation is performed 5 times in each condition to allow for differences in weight initialization

Novel Class Detection

Results

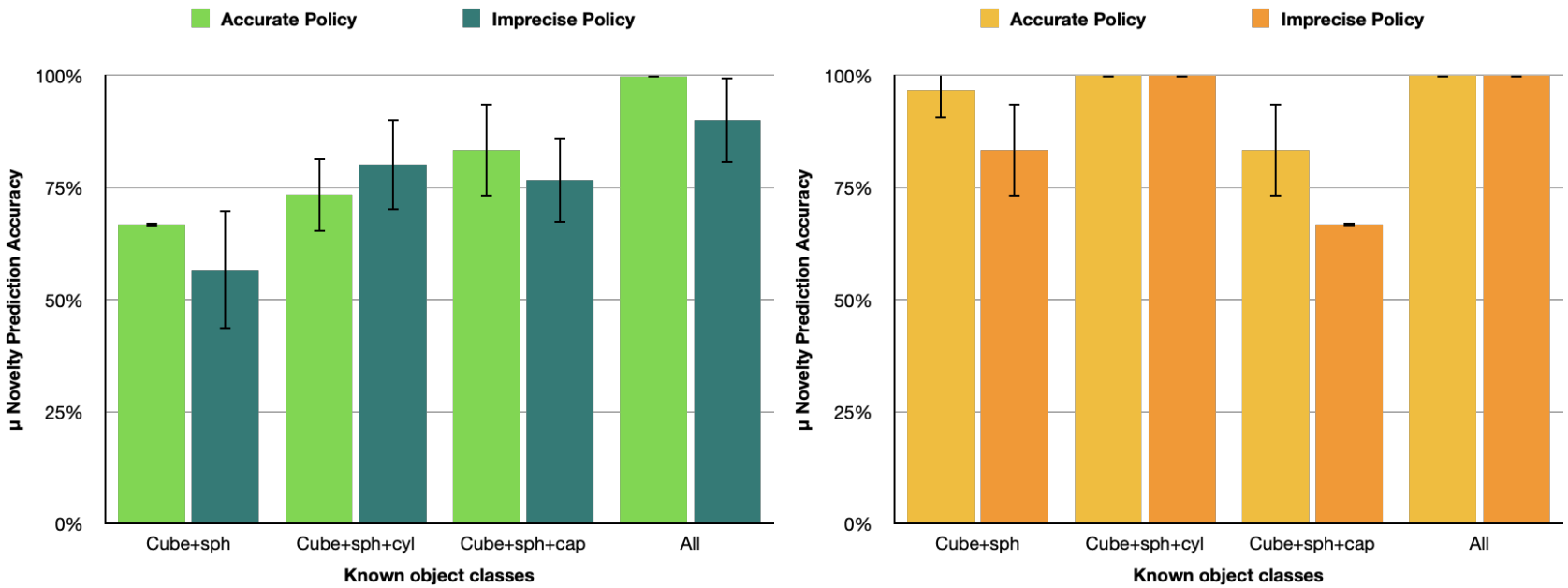
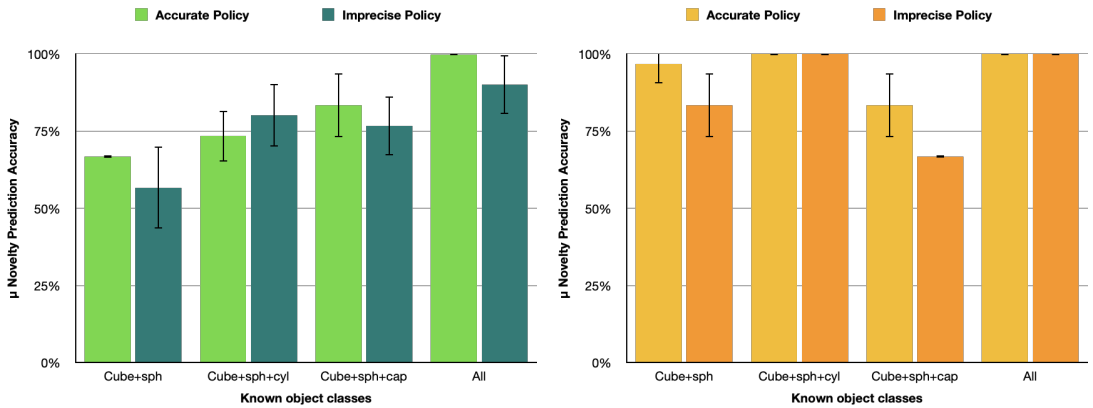


Figure: Novel class detection accuracy without implicit habitat enoding (left) and with (right)

Novel Class Detection

Discussion



- Can correctly identify the novelty of cylinders and capsules based on behavior alone
- Small cubes identified as same type as large cubes
- Imprecise policy data slightly more challenging
- Including implicit habitat/affordance information increases performance by 25%

Novel Class Detection

Discussion

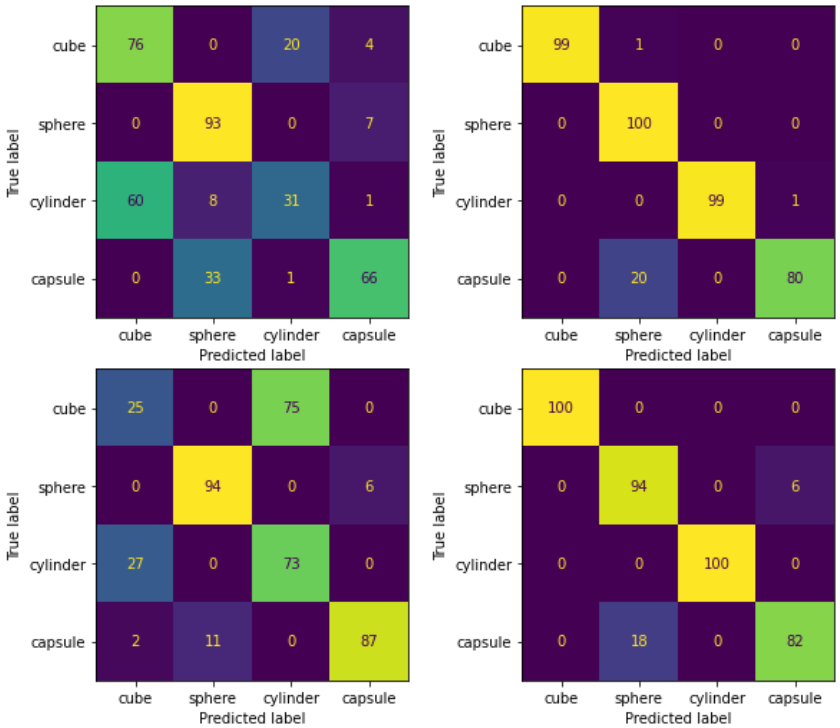
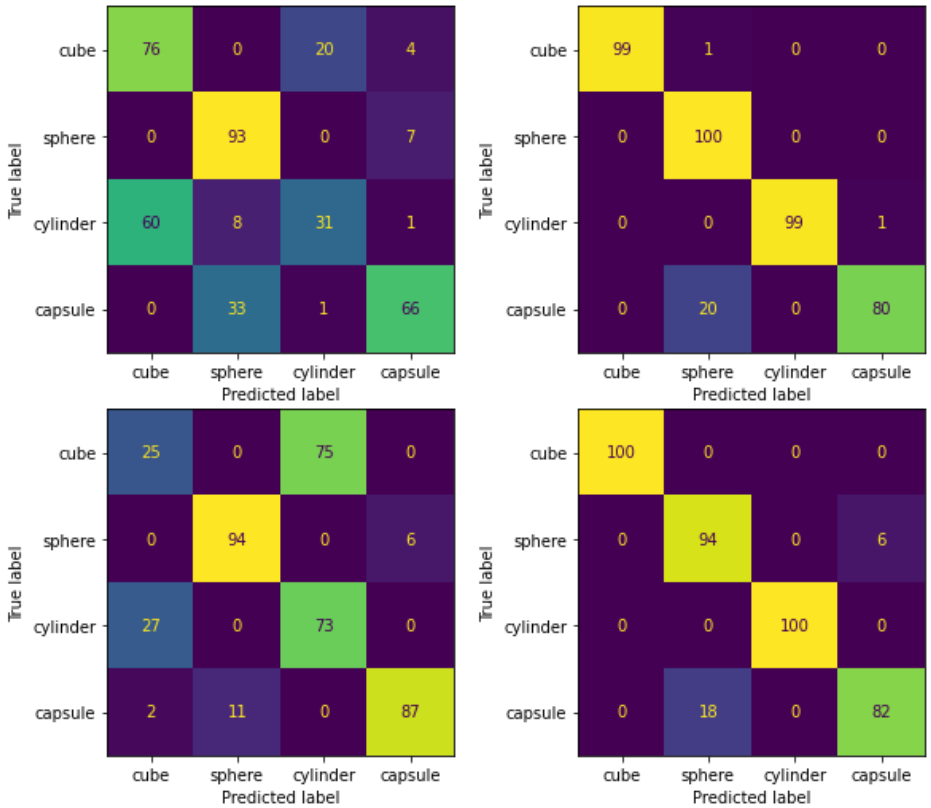


Figure: Aggregated 1D CNN classifier outputs over dev-test set. T: accurate policy evaluation; B: imprecise policy evaluation; L: without VoxML-derived inputs; R: with VoxML-derived inputs

Novel Class Detection

Discussion

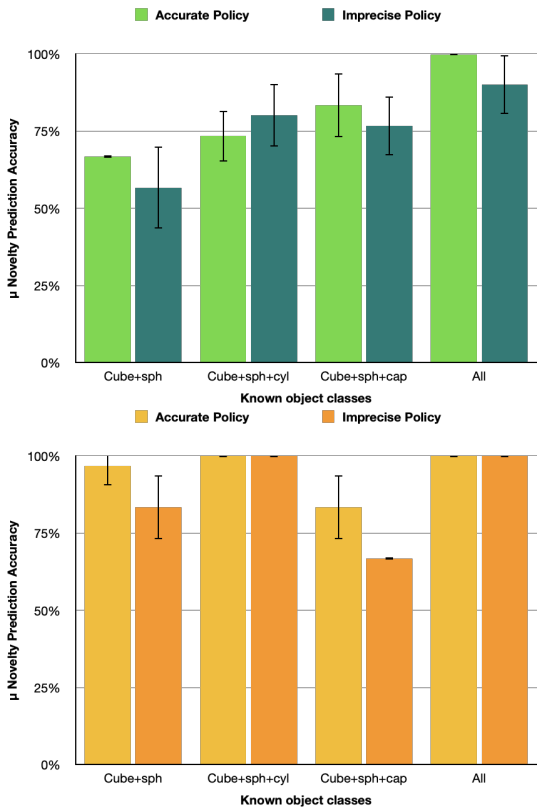
- Without habitat information, classifier confuses cylinders and cubes, capsules and spheres
  - Same patterns as original two-class classifier



Novel Class Detection

Discussion

- Order of concept acquisition matters
  - Detecting capsule concept before cylinder impedes cylinder detection
- Capsule is more distinct from sphere than cylinder is from cube in its behavior
- Capsule embedding vectors take up more "space"
  - Makes the subtle cube/cylinder distinction harder to detect



Multimodal Grounding

A new approach

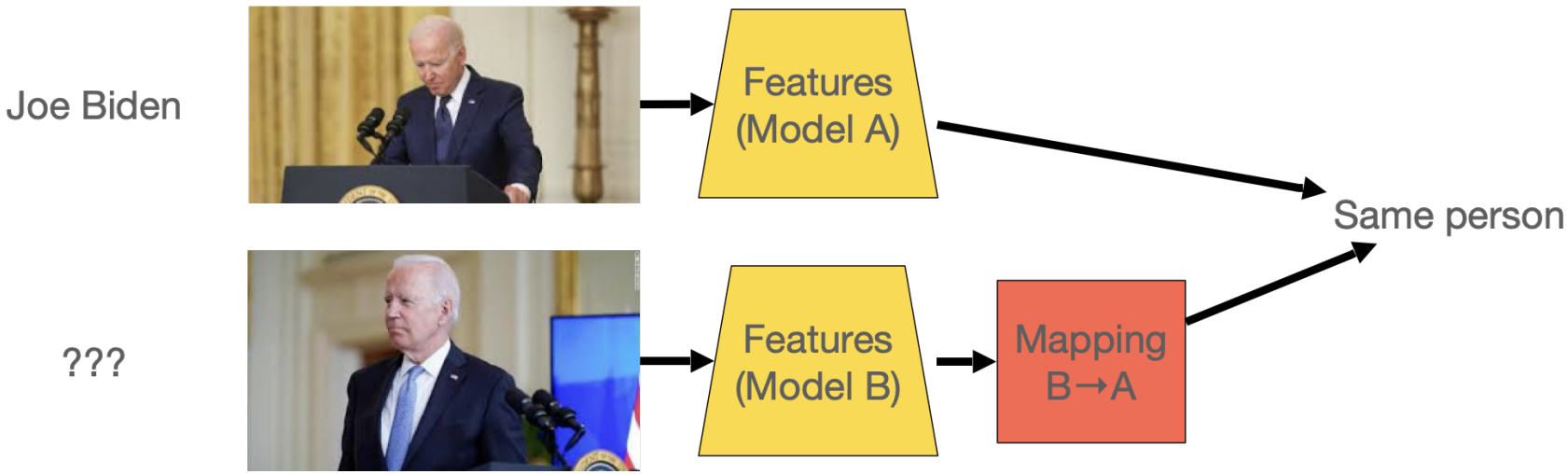


Figure: Transforming features from different embedding spaces into common space

- Findings from the vision community: CNN embedding spaces share interchangeability up to an affine transformation $M_{A \to B}$

Multimodal Grounding

A new approach

- Properties of transformation $M_{A \to B}$:
- Given two embedding spaces A, B, where embeddings in each are of dimensionality d_A, d_B
  - $M_{A \to B} \in \mathbb{R}^{d_A} \times \mathbb{R}^{d_B}$ that transforms the representation of $C \in \mathbb{R}^{d_A}$ to $\sim C \in \mathbb{R}^{d_B}$
  - Computed by minimizing distance between paired (equivalent) points in A and B
- Given two sets of objects (vectors) X and Y, are they the "same" under an affine transformation?
- Task is now to recover that transformation

Multimodal Grounding

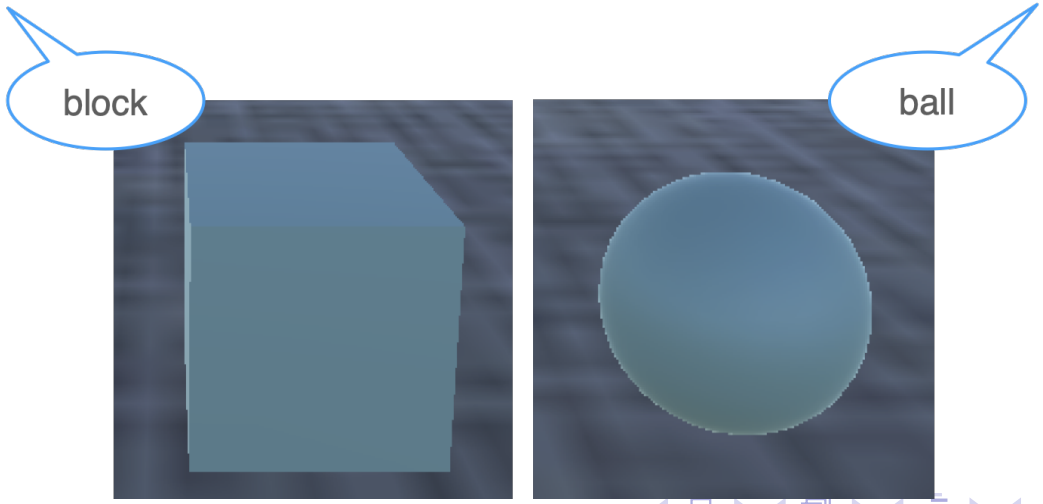
A new approach

- Different images of the same person, if both classified correctly, should have similar but not identical embedding vectors
- **Similarly, object embeddings from behavioral data are distinct**
  - They define a subspace
- Similar to contextualized word embeddings a la BERT
  - Not identical, but point in the same direction
- Use one set of vectors as inputs to a regressor and the other as outputs
- **If inputs and outputs preserve similar information and distinctions...**

Multimodal Grounding

A new approach

- $M_{A \to B}$ is the weights of a mapping function $f(x; W)$
- $f(x; W) : U \subset \mathbb{R}^{d_A} \to V \subset \mathbb{R}^{d_B}$, where $d_B \leq d_A$, s.t. $f(x \in U) \approx x \in V$
  - inputs: U; outputs: V
  - Application of f to any element of U should approximate the equivalent element of V



Multimodal Grounding

A new approach

- Imagine two agents, a "child" agent playing with objects, and a "parent" agent generating utterances containing object references.
  - Model child agent as 1D CNN behavior classifier, parent agent as Transformer language model

Child:

Parent:

"The **block** is flat."

Multimodal Grounding

A new approach

Child:

Parent:

"The **block** is flat."

- Take 64D embeddings from CNN model (grounded object instances) as outputs
- Take 768D embeddings from Transformer model (contextualized token embeddings) as inputs
- Compute $M_{A \to B}$ as 768×64 affine matrix
  - Do new mentions of the same contextualized tokens cluster with the correct objects?
  - Are different senses of the same word distinct from the object mentions?

Multimodal Grounding

A new approach

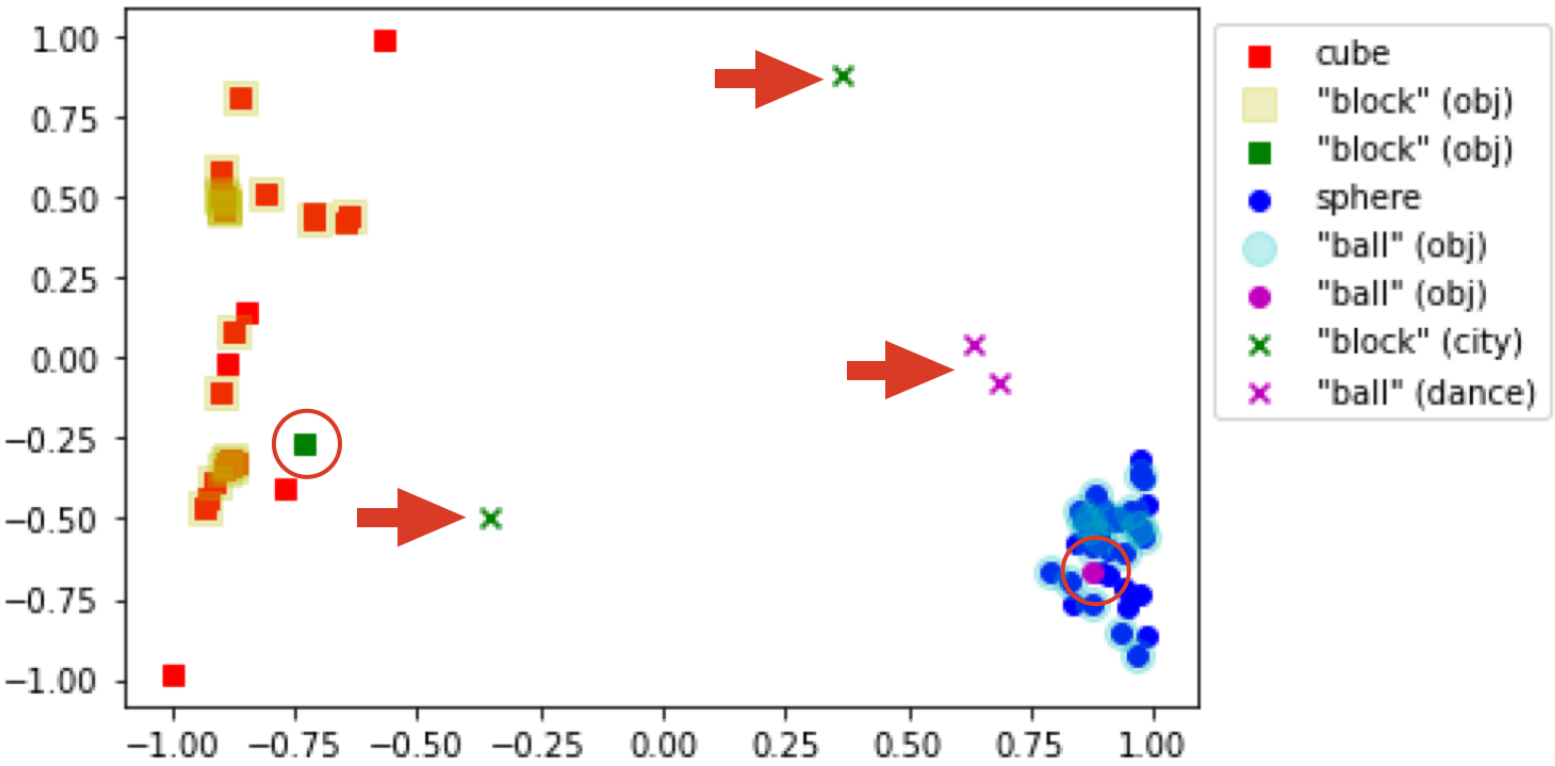
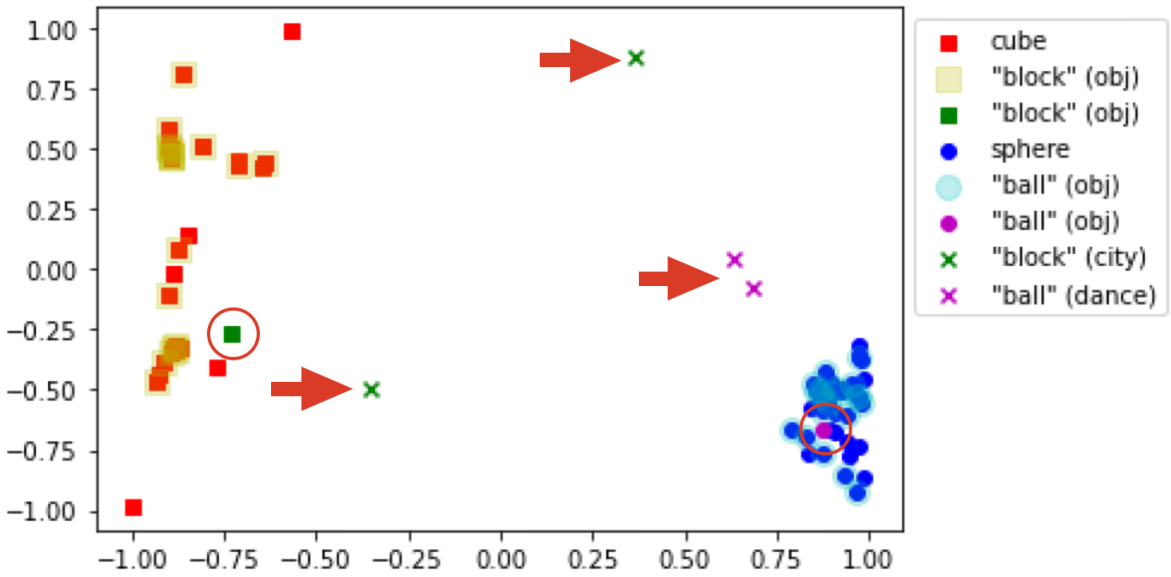


Figure: BERT word vectors mapped into object vector space (reduced to 2 dimensions)

Multimodal Grounding

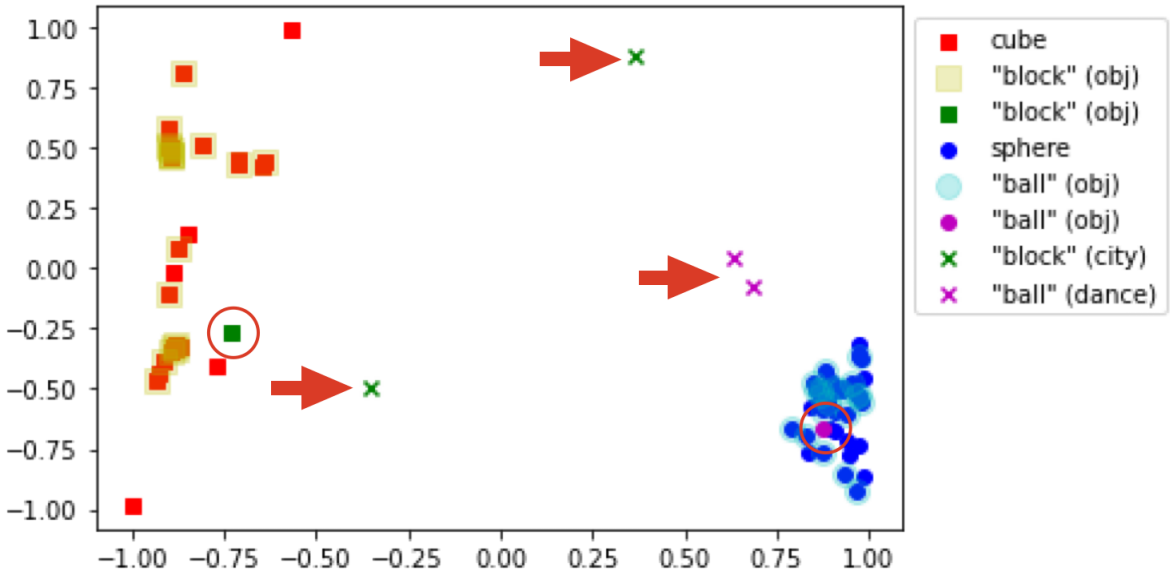
A new approach



- Novel instances of the known sense of the word “block” and “ball” map into the sphere/cube subspaces
- Different senses of the same words fall outside of those subspaces

Multimodal Grounding

A new approach



- Enables "decontextualized reference"
- Agent can discuss items not present while maintaning grounding to concept
- Facilitates transfer between otherwise dissimilar models

Gesture in Multimodal Communication

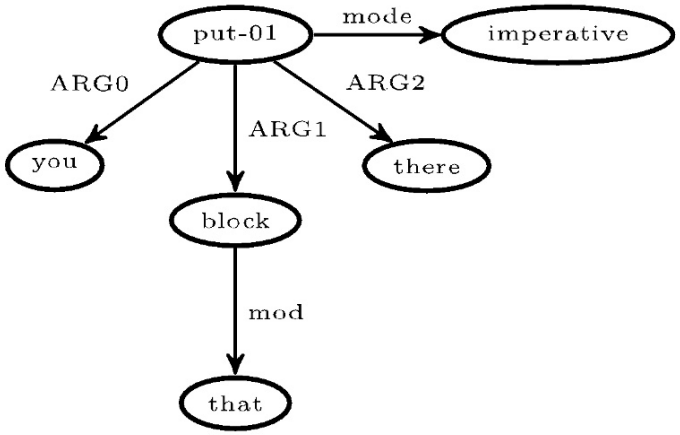
- Gesture can ground and situate linguistic expressions in dialogue
- Gestures can be formalized as interpreted grammars

Abstract Meaning Representation (AMR)

Predicate-argument Binding

- Put that block there

```
(p / put-01
      :mode imperative
      :ARG0 (y / you)
      :ARG1 (b / block
               :mod (t / that))
      :ARG2 (t2 / there))
```

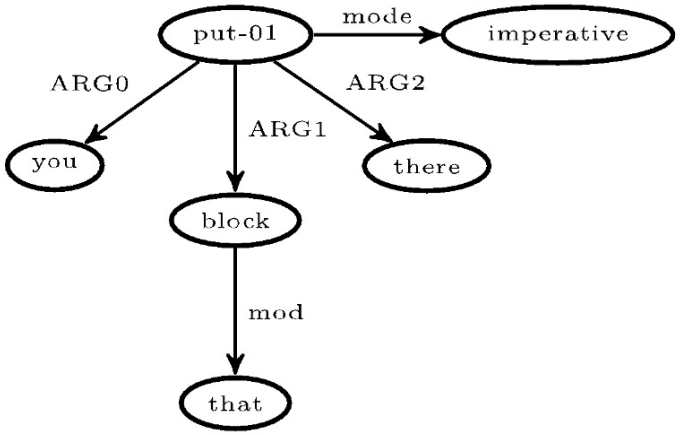


Abstract Meaning Representation (AMR)

Event Predicate

- Put that block there

```
(p / put-01
      :mode imperative
      :ARG0 (y / you)
      :ARG1 (b / block
               :mod (t / that))
      :ARG2 (t2 / there))
```

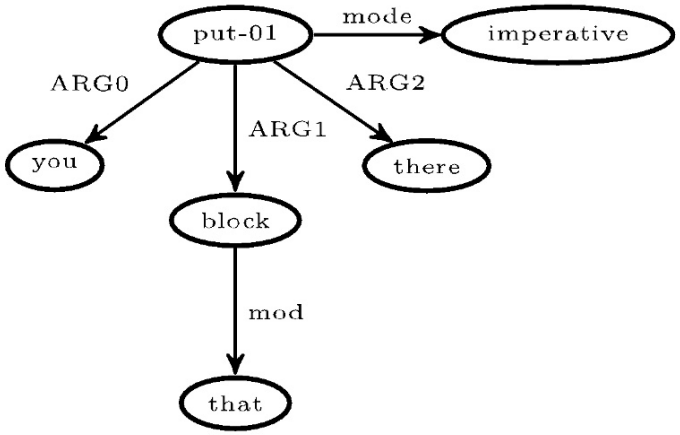


Abstract Meaning Representation (AMR)

Subject

- Put that block there

```
(p / put-01
      :mode imperative
      :ARG0 (y / you)
      :ARG1 (b / block
               :mod (t / that))
      :ARG2 (t2 / there))
```

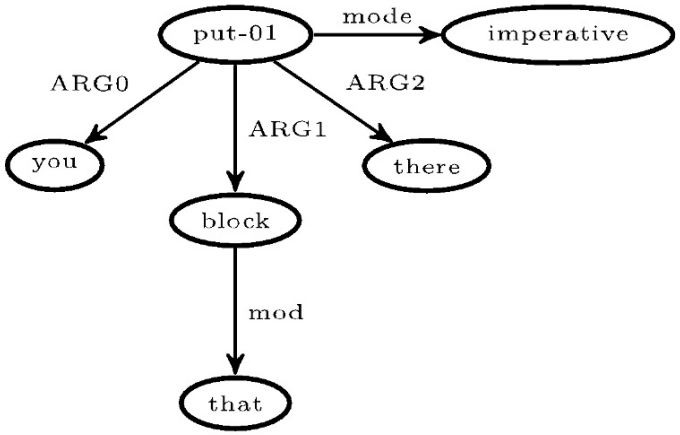


Abstract Meaning Representation (AMR)

Object

- Put that block there

```
(p / put-01
      :mode imperative
      :ARG0 (y / you)
      :ARG1 (b / block
               :mod (t / that))
      :ARG2 (t2 / there))
```

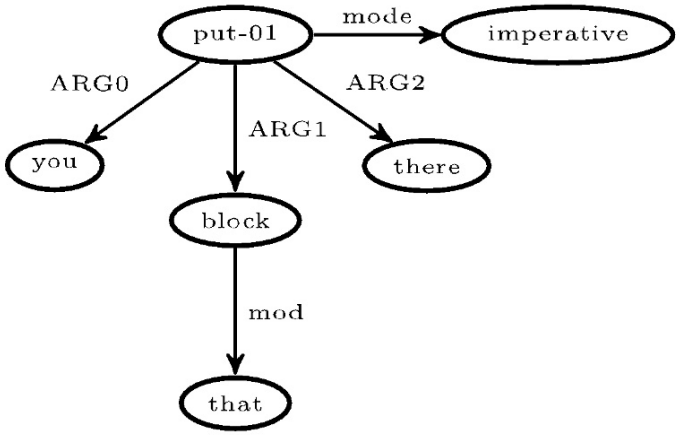


Abstract Meaning Representation (AMR)

Locative

- Put that block there

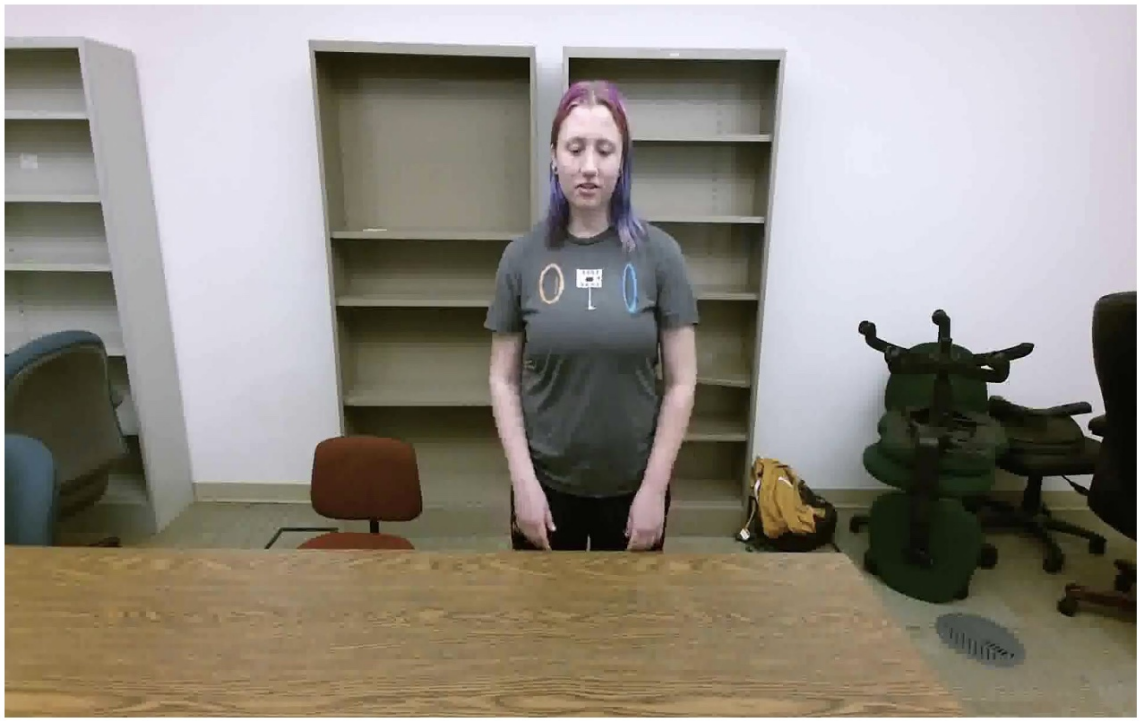
```
(p / put-01
      :mode imperative
      :ARG0 (y / you)
      :ARG1 (b / block
              :mod (t / that))
      :ARG2 (t2 / there))
```



Gesture AMR: EGGNOG corpus

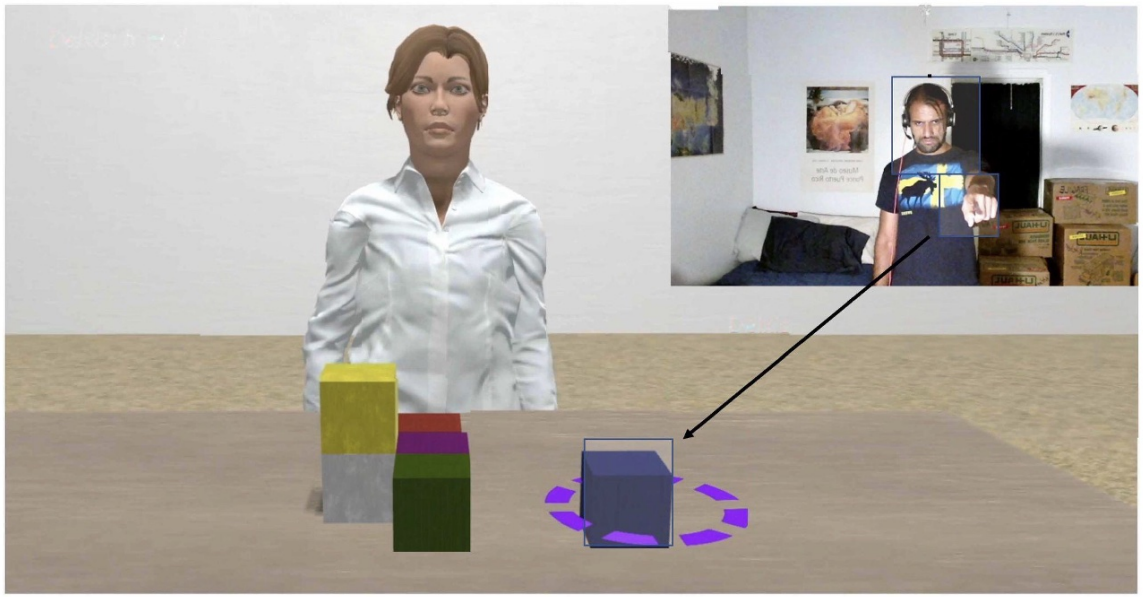
- 8 hours of video
- 40 participants
- Shared task of arranging block structures
- *Signaler* gives instructions
- *Actor* interprets and builds structures
- Wang et al, 2017

Gesture in Multimodal Communication



- Gesture: how speakers move hands when they communicate information
- Existing schemes typically focus on hand shape and movement
- We focus on gesture tied to speech: iconic, deictic, metaphoric, & emblematic

Gesture AMR: Approach

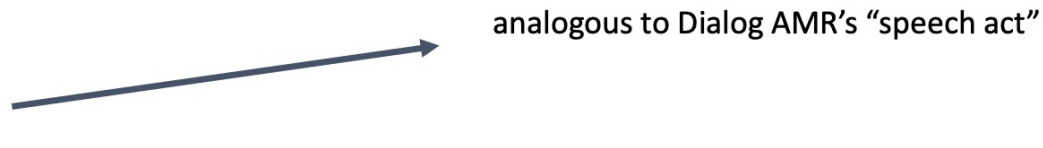


- Needs to represent:
  - situated meaning
  - common ground, objects, participants
  - modes of communication
- While abstracting away from physical descriptions

Gesture AMR: Schema

```
( g / [gesture]-GA
      :ARG0 [gesturer]
      :ARG1 [content]
      :ARG2 [addressee])
```

Gesture AMR: Schema




```
( g / [gesture]-GA
    :ARG0 [gesturer]
    :ARG1 [content]
    :ARG2 [addressee])
```

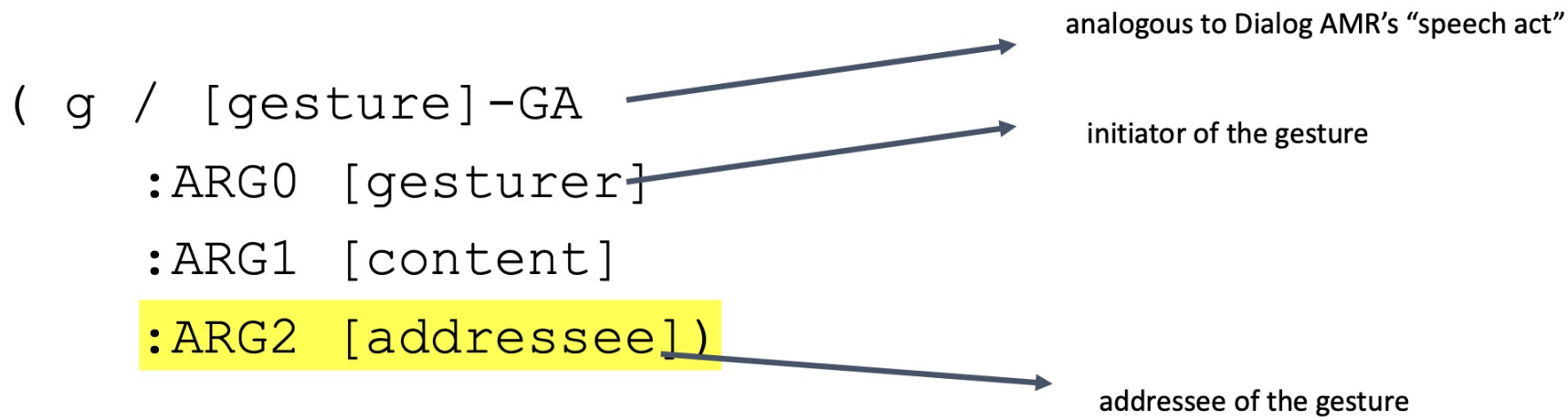
Gesture AMR: Schema

```
( g / [gesture]-GA
    :ARG0 [gesturer]
    :ARG1 [content]
    :ARG2 [addressee])
```

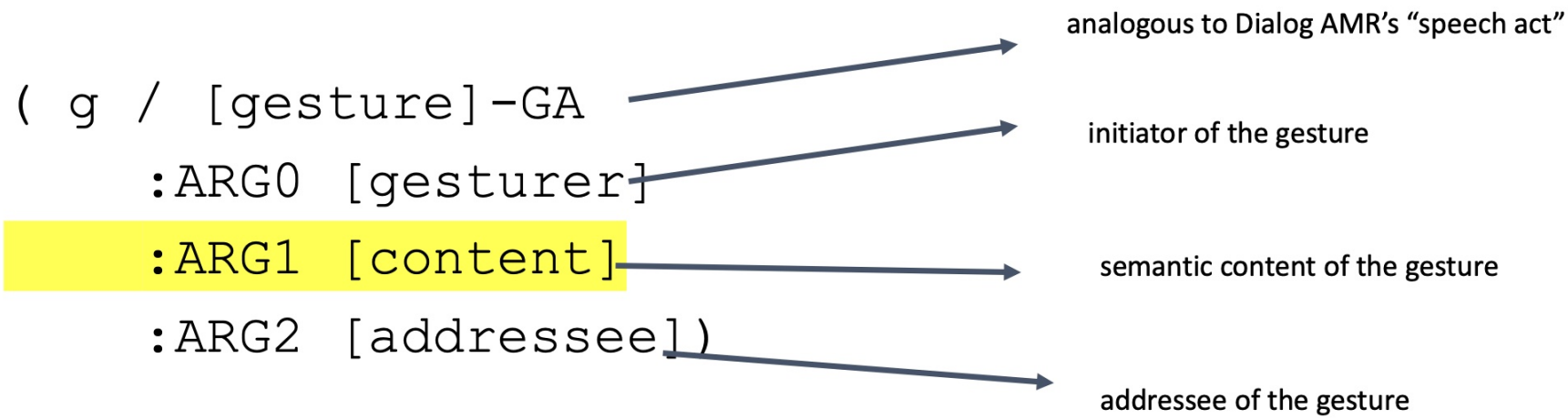
analogous to Dialog AMR's "speech act"

initiator of the gesture

Gesture AMR: Schema



Gesture AMR: Schema



Gesture AMR: Deixis

```
(d / deixis-GA
      :ARG0 (g / gesturer)
      :ARG1 (b / block)
      :ARG2 (a / addressee
```

Gesture AMR: Deixis

```
(d / deixis-GA
     :ARG0 (g / gesturer)
     :ARG1 (b / block)
     :ARG2 (a / addressee))
```

Gesture AMR: Deixis

```
(d / deixis-GA
      :ARG0 (g / gesturer)
      :ARG1 (b / block)
      :ARG2 (a / addressee
```

Gesture AMR: Deixis

```
(d / deixis-GA
      :ARG0 (g / gesturer)
      :ARG1 (b / block)
      :ARG2 (a / addressee
```

Gesture AMR: Icon

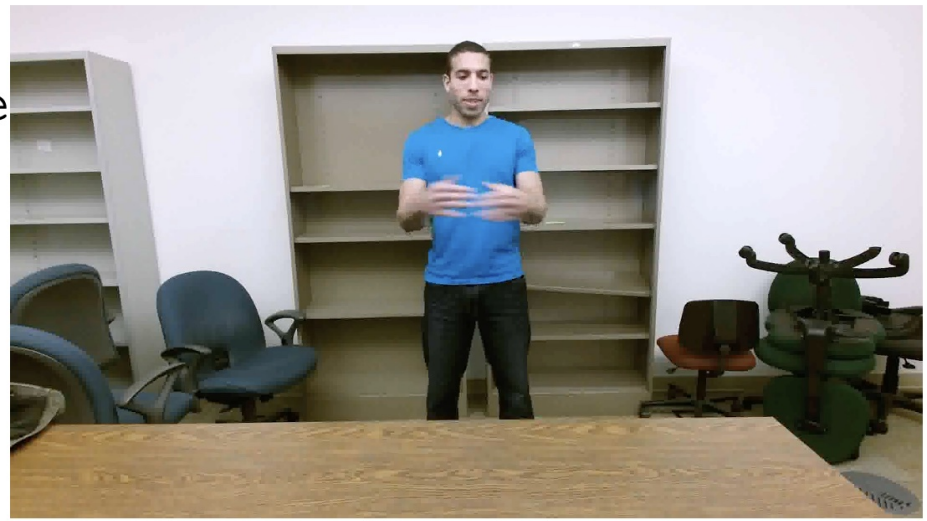
```
(i / icon-GA
      :ARG0 (g / gesturer)
      :ARG1 (s / slide-01)
            :direction (f / forward))
      :ARG2 (a / addressee))
```

Gesture AMR: Icon

```
(i / icon-GA
      :ARG0 (g / gesturer)
      :ARG1 (s / slide-01)
            :direction (f / forward))
      :ARG2 (a / addressee))
```

Gesture AMR: Metaphor

```
(m / metaphor-GA
      :ARG0 (g / gesturer)
      :ARG1 (s / sound wave
      :ARG2 (a / addressee)
```



Gesture AMR: Metaphor

```
(m / metaphor-GA
     :ARG0 (g / gesturer)
     :ARG1 (s / sound wave)
     :ARG2 (a / addressee))
```

Gesture AMR: Emblem

```
(e / emblem-GA
      :ARG0 (g / gesturer)
      :ARG1 (y / yes)
      :ARG2 (a / addressee))
```

Gesture AMR: Emblem

```
(e / emblem-GA
     :ARG0 (g / gesturer)
     :ARG1 (y / yes)
     :ARG2 (a / addressee))
```

Gesture AMR: Gesture Unit

```
(g / gesture-unit
     :op1 (i / icon-GA
          :ARG0 (g2 / gesturer)
          :ARG1 (b / block)
          :ARG2 (a / addressee))
     :op2  (d / deixis-GA
          :ARG0 g2
          :ARG1 (l/ location)
          :ARG2 a))
```

Gesture AMR: Gesture Unit

```
(g / gesture-unit
     :op1 (i / icon-GA
          :ARG0 (g2 / gesturer)
          :ARG1 (b / block)
          :ARG2 (a / addressee))
     :op2  (d / deixis-GA
          :ARG0 g2
          :ARG1 (l/ location)
          :ARG2 a))
```

Gesture AMR: Gesture Unit

```
(g / gesture-unit
     :op1 (i / icon-GA
          :ARG0 (g2 / gesturer)
          :ARG1 (b / block)
          :ARG2 (a / addressee))
     :op2  (d / deixis-GA
          :ARG0 g2
          :ARG1 (l/ location)
          :ARG2 a))
```

Gesture AMR: Temporal alignment

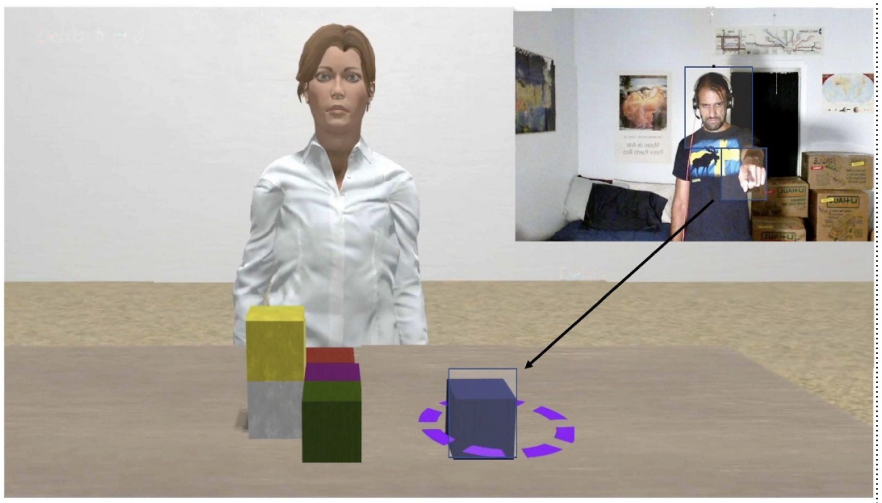
Click to add text

(2) **Co-gestural Speech Ensemble**:

$$\begin{bmatrix} \mathcal{G} & g_1 & \dots & g_i & \dots & g_n \\ \mathcal{S} & s_1 & \dots & s_i & \dots & s_n \end{bmatrix}$$

CO-GESTURAL SPEECH

HUMAN: s_1 = Put
$g_1 = \varnothing$
HUMAN: s_2 = [that block]
g_2 = [*points to the blue block*]
HUMAN: s_3 = there.
g_3 = [*points to the purple block*]



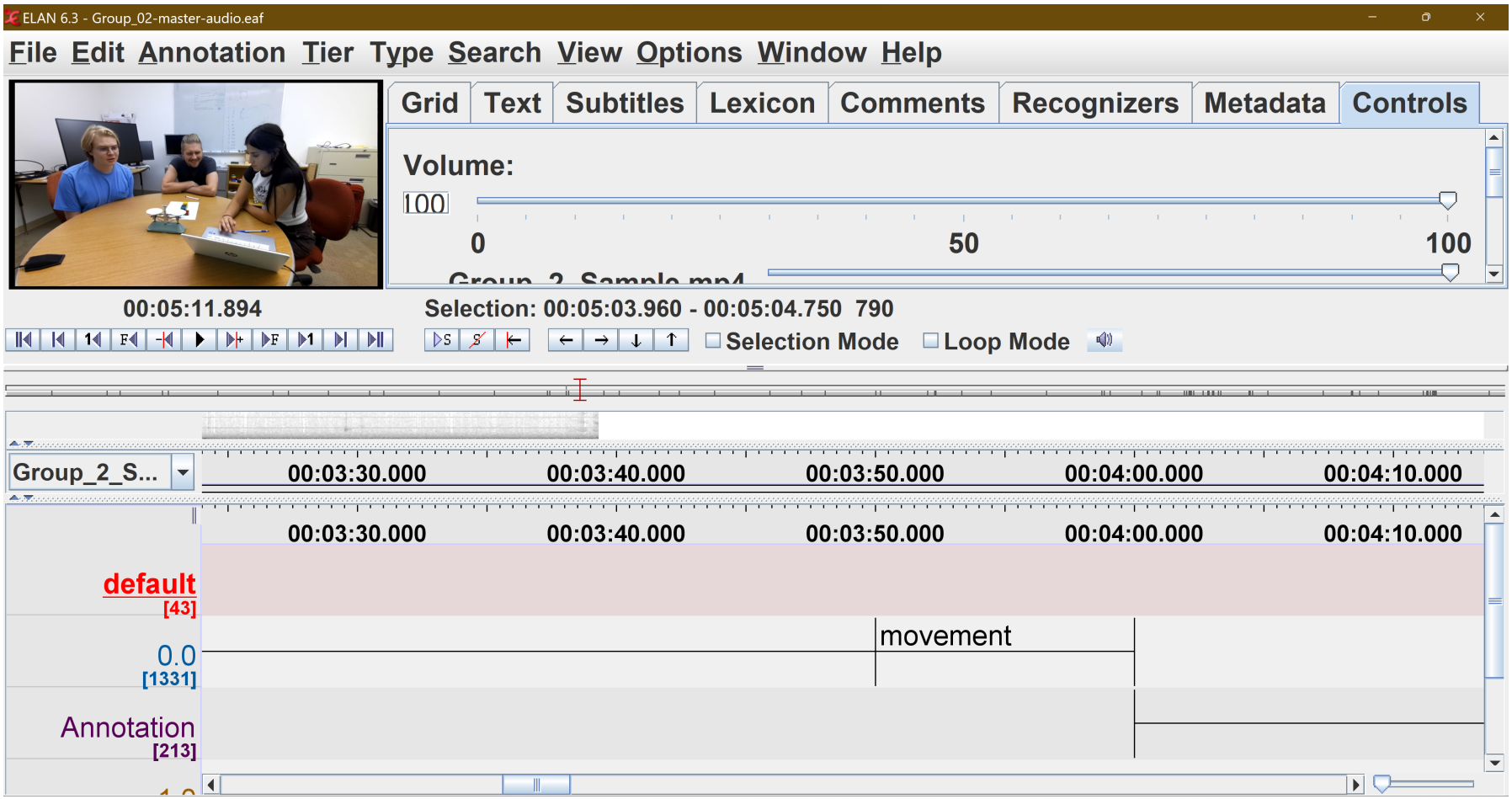
Gesture AMR: Temporal alignment

Before(S, G)

Overlap(S, G)

After(S, G)

Gesture AMR: Annotation



Concluding Remarks

- As AI becomes more multimodal, we need to understand the role of affordances and human-object interactions in reasoning.
- We examined how our knowledge of object interactions is rarely reflected in linguistic descriptions of actions (or images).
- We demonstrated how object and situational conditions on actions need to be identified and encoded, just as importantly as the actions themselves.
- Identifying and encoding situated conditions requires grounding deep semantic representations to the environment.