Language Grounding to the Visual World and Human Interactions: How Far Are We from Embodied Dialogue Agents?

** For Seminar Talk @ University of Washington*

Martin Ziqiao Ma
<*marstin@umich.edu*>
Dec 5th, 2024

CSE COMPUTER SCIENCE AND ENGINEERING
UNIVERSITY OF MICHIGAN

Everyday Grounding

Language Grounding: Connecting language to the physical world and communication partners.

My favorite fruit is apple.

Everyday Grounding

Language Grounding: Connecting language to the physical world and communication partners.

My favorite fruit is apple.

Those apples on the table look nice.

Everyday Grounding

Language Grounding: Connecting language to the physical world and communication partners.

My favorite fruit is apple.

Those apples on the table look nice.

Can you bring me that apple?

Everyday Grounding

Language Grounding: Connecting language to the physical world and communication partners.

My favorite fruit is apple.

Those apples on the table look nice.

Can you bring me that apple?

Which apple do you want?

The red one.

Distributional Word Meanings

The meaning of a word is related to the distribution of words around it (Firth, 1957).

- We represent the meaning of a word…
  - …From the context and co-occurrences;
  - …As a vector of numbers (embedding).
- We developed…
  - …Static word embeddings: `word2vec`, `GloVe`, …
  - …Contextual word embeddings: `ELMO`, `BERT`, `GPT-x`, …

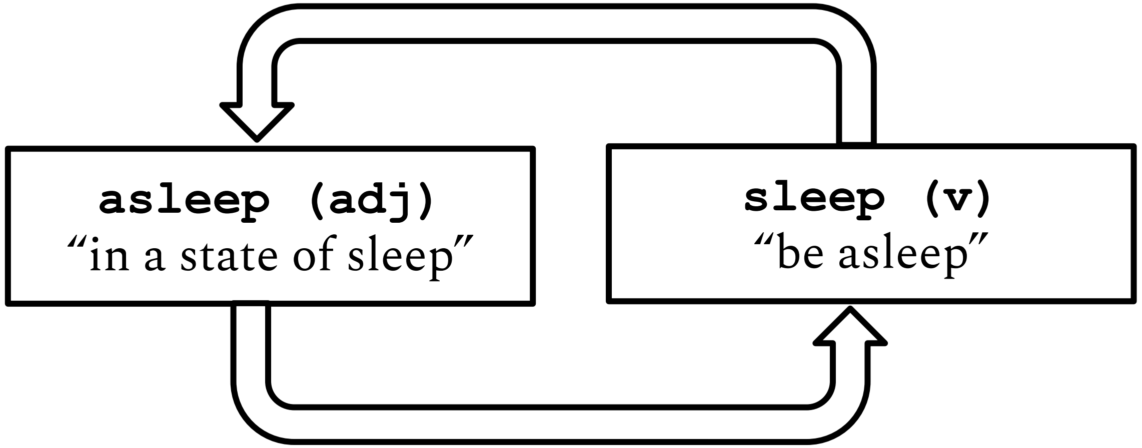
sugar, a sliced lemon, a tablespoonful of	**apricot**	preserve or jam, a pinch each of, their enjoyment.
Cautiously she sampled her first	**pineapple**	and another fruit whose taste she likened
well suited to programming on the digital	**computer**	. In finding the optimal R-stage policy from
for the purpose of gathering data and	**information**	necessary for the study authorized in the

A Synopsis of Linguistic Theory. *John R Firth.* Studies in Linguistic Analysis, 1957

Distributional Word Meanings

Connection within linguistic symbols only may be a problem.

- Distributional (Ungrounded) Semantics:
  - Connecting linguistic symbols to <u>other linguistic symbols</u> is enough.

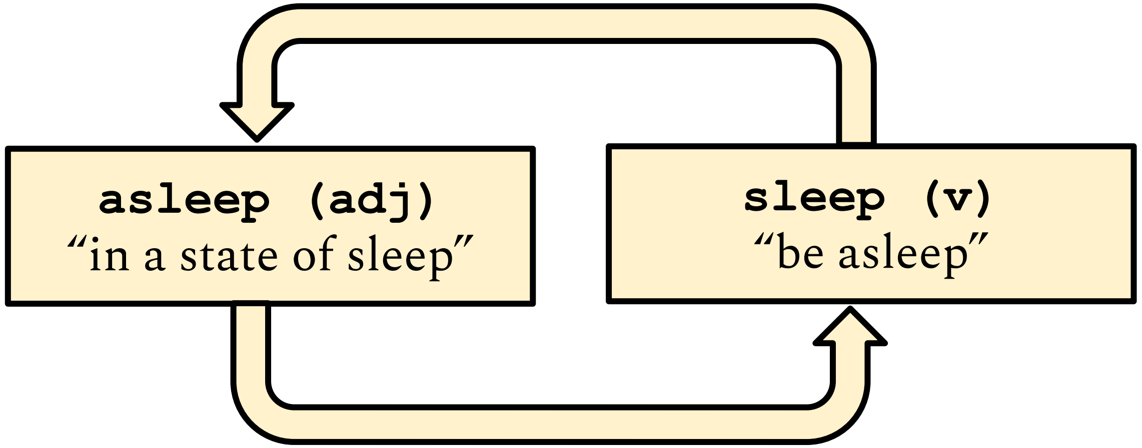


Language Grounding to Vision and Control.
Katerina Fragkiadaki. Fall 2017, CMU 10-808

The Symbol Grounding Problem

Grounding: Connection between linguistic symbols and non-linguistic experiences.

- Distributional (Ungrounded) Semantics:
  - Connecting linguistic symbols to other linguistic symbols is enough.
- Grounded Semantics (Harnad, 1990):
  - Linguistic symbols need to connect to the experiences <u>external</u> to these symbols.



[Language Grounding to Vision and Control]().
Katerina Fragkiadaki. Fall 2017, CMU 10-808

Why are you being so upset?

I didn't sleep well last night.

Why? Was it because of the noise?

No, I drank too much coffee.

[2] **The Symbol Grounding Problem**. *Stevan Harnad.* Physica D: Nonlinear Phenomena, 1990

Experience Grounds Language

Humans acquire language from sensorimotor and sociolinguistic experiences.

- Experience grounds language (Bisk et al., 2020):

  "We posit that the present success of representation learning approaches trained on large, text-only corpora requires the parallel tradition of research on the broader physical and social context of language to address the deeper questions of communication."

- Two types of grounding (Chai et al., 2018):
  - Static/Semantic grounding: the process where semantics of language is grounded to the agent's internal representations of perception from the world and actions to the world.
  - Dynamic/Interactive/Communicative grounding: the process for communication partners to reach a *common ground* - mutually agreed knowledge, beliefs, and assumptions.

Experience Grounds Language. *Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, Joseph Turian.* EMNLP, 2020
Language to Action: Towards Interactive Task Learning with Physical Agents. *Joyce Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, Guangyue Xu.* IJCAI, 2018.

Overview of This Talk

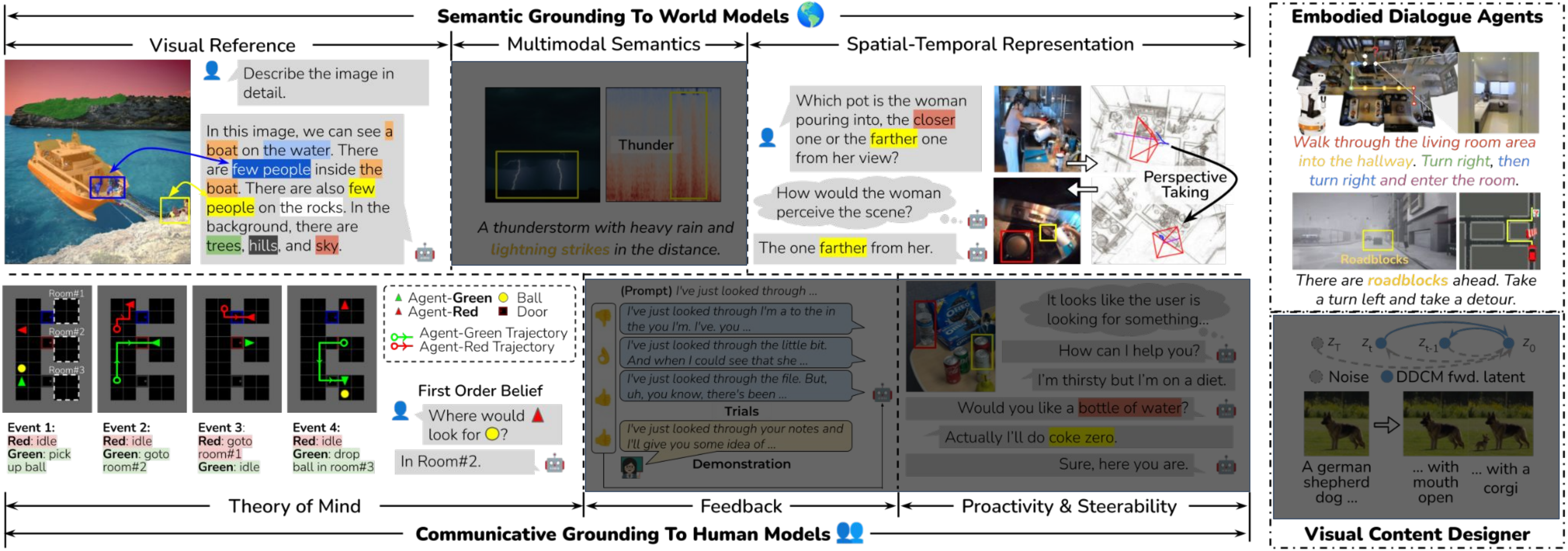
Cognitive Inquiry → Language Learning

Building Scalable/Deployable AI Models → Internet AI Models → Embodied AI Agents

Language Grounding → Semantic Grounding, Interactive Grounding

	Language Learning	Internet AI Models	Embodied AI Agents
Semantic Grounding	• What is the role of semantic grounding in neural language learning?	• How to build multimodal large language models (MLLMs)? • What are some pitfalls of MLLMs and how to address?	• How to build AI agents that perceive, reason, plan, and act in the physical world?
Interactive Grounding	• What is the role of interactive grounding in neural language learning?	• How to align LLMs with human interaction intuitions? • Do LLMs have a Theory of Mind and how to assess?	• How to build AI agents that engage in situated interactions with humans?

Overview of This Talk



Grounded Vision-Language Models

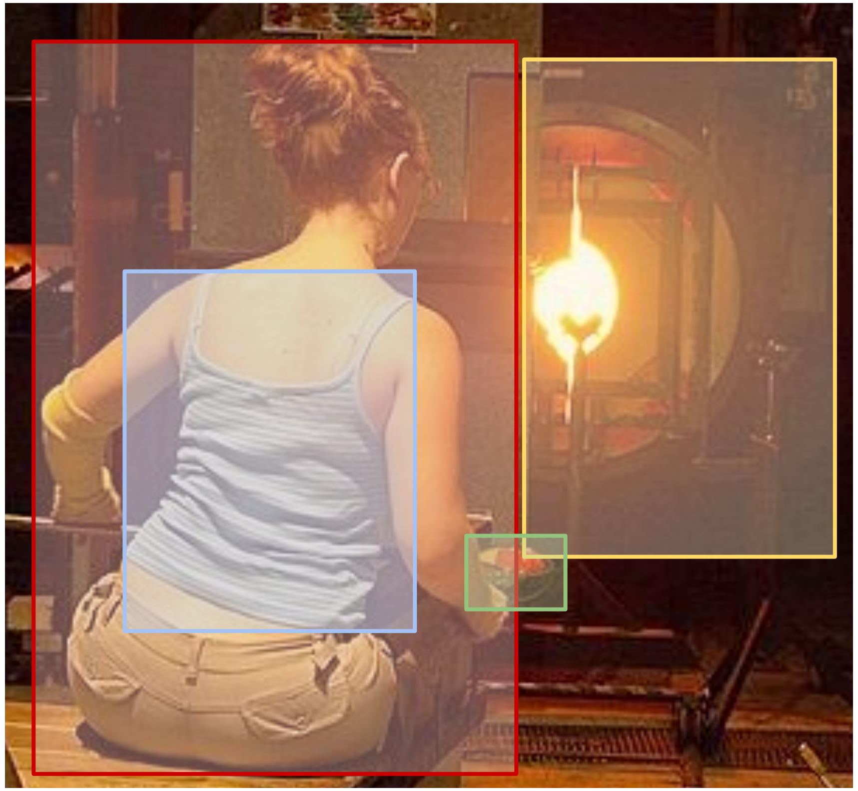

Fast mapping and scalable grounded vocabulary acquisition [ACL 2023].

A lady wearing a navy blue stripe tank top is getting ready to burn glass in front of an incinerator.

[World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models](). *Ziqiao Ma, Jiayi Pan, Joyce Chai.* ACL 2023.

Grounded Vision-Language Models

Fast mapping and scalable grounded vocabulary acquisition [ACL 2023].



A lady wearing a navy blue stripe tank top is getting ready to burn glass in front of an incinerator.

[World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models](). *Ziqiao Ma, Jiayi Pan, Joyce Chai.* ACL 2023.

Grounded Vision-Language Models

Fast mapping and scalable grounded vocabulary acquisition [ACL 2023].

A lady wearing a navy blue stripe tank top is getting ready to burn glass in front of an **incinerator**.

World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models. *Ziqiao Ma, Jiayi Pan, Joyce Chai.* ACL 2023.

Grounded Vision-Language Models

Fast mapping and scalable grounded vocabulary acquisition [ACL 2023].

- Defining and evaluating grounded word learning.

`Two boats of people, a smaller yellow` **`[mask]`** `with two people and a larger white boat with six people.`

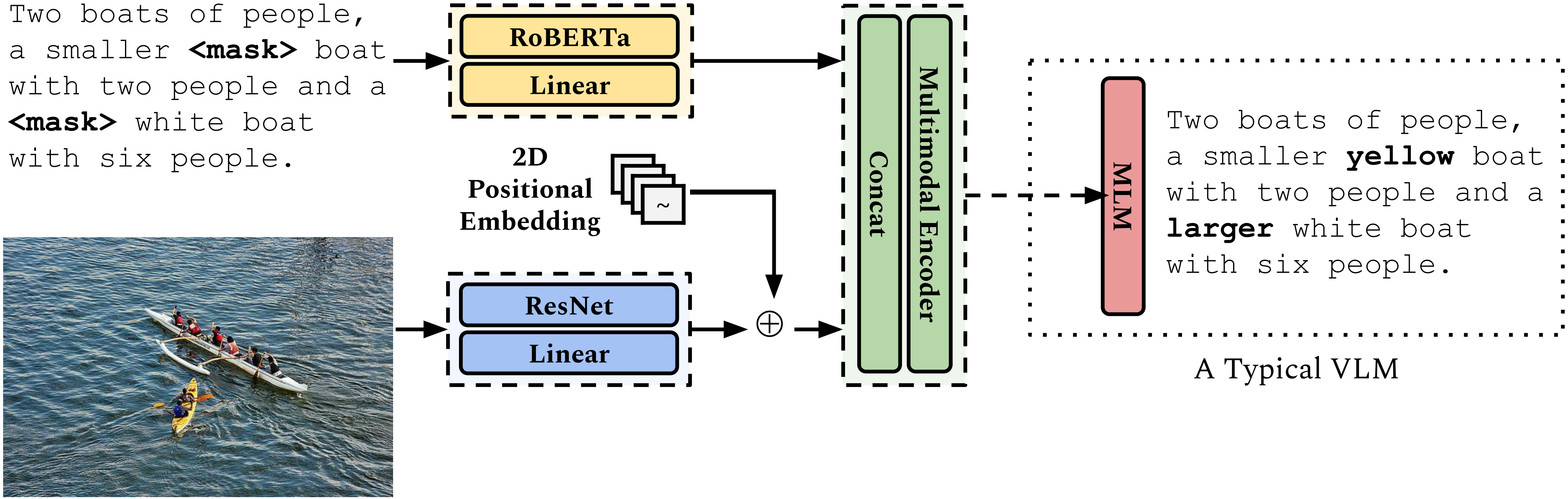
`Two boats of people, a smaller yellow` **`boat`** `with two people and a larger white boat with six people.`

[World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models](). *Ziqiao Ma, Jiayi Pan, Joyce Chai.* ACL 2023.

Grounded Vision-Language Models

Fast mapping and scalable grounded vocabulary acquisition [ACL 2023].

- Our model: Object-Oriented BERT (`OctoBERT`)
  - Vision and language representations are fused using self-attention in a cross-encoder;

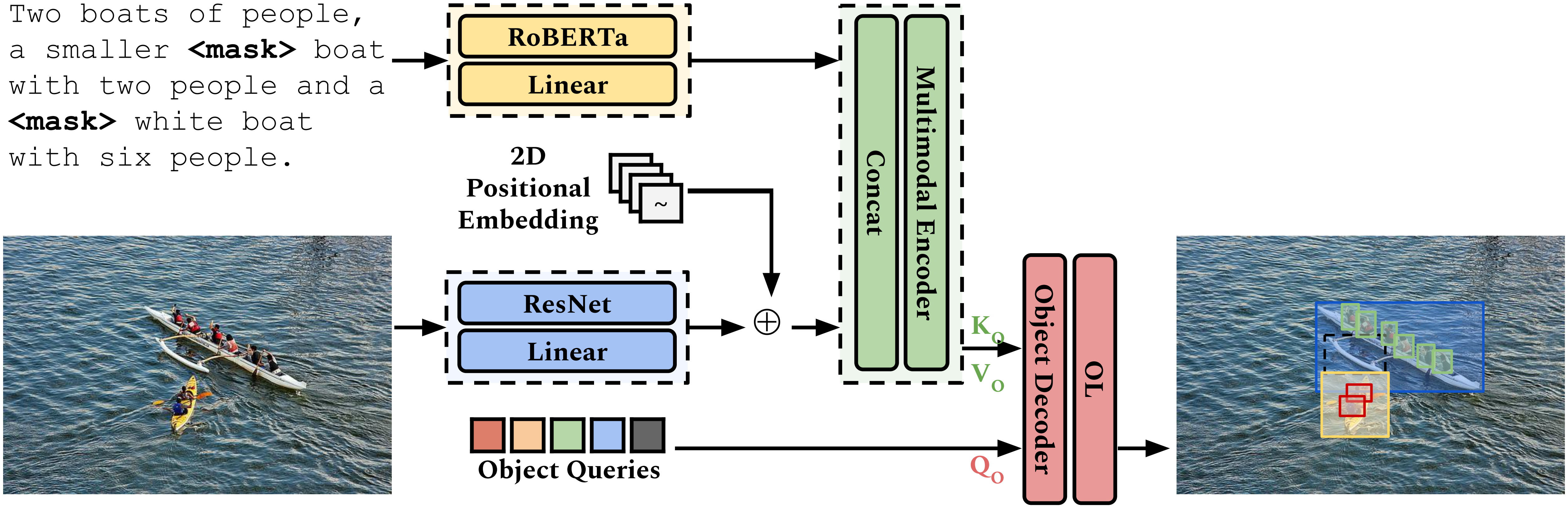


[World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models](). *Ziqiao Ma, Jiayi Pan, Joyce Chai.* ACL 2023.

Grounded Vision-Language Models

Fast mapping and scalable grounded vocabulary acquisition [ACL 2023].

- Our model: Object-Oriented BERT (`OctoBERT`)
  - The object decoder takes a set of learnable object queries and produces object representations;

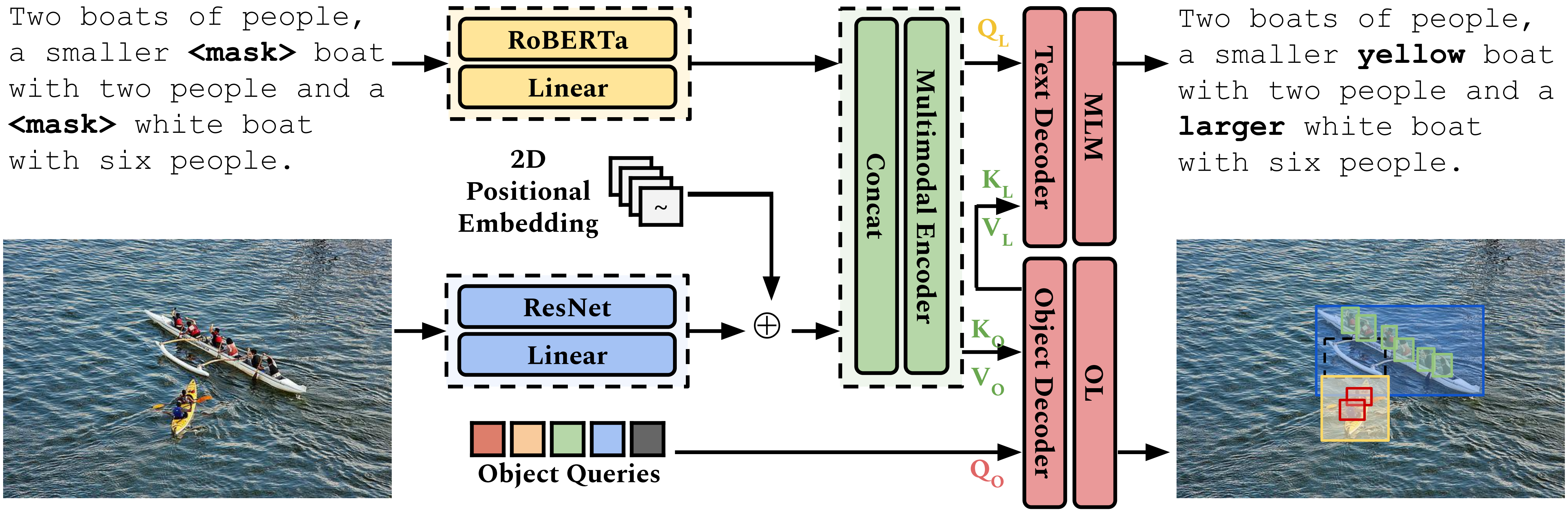


[World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models](). *Ziqiao Ma, Jiayi Pan, Joyce Chai.* ACL 2023.

Grounded Vision-Language Models

Fast mapping and scalable grounded vocabulary acquisition [ACL 2023].

- Our model: Object-Oriented BERT (`OctoBERT`)
  - Images and texts are encoded using pre-trained a language model and a vision backbone;

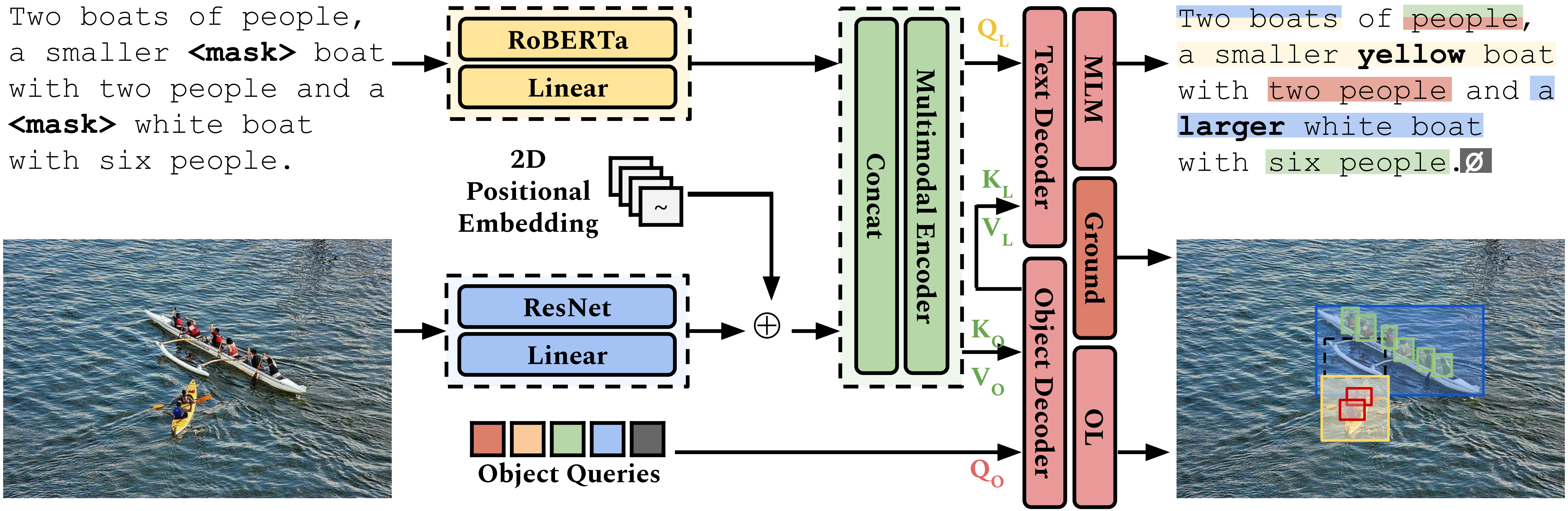


[World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models](). *Ziqiao Ma, Jiayi Pan, Joyce Chai.* ACL 2023.


Grounded Vision-Language Models

Fast mapping and scalable grounded vocabulary acquisition [ACL 2023].

- Our model: Object-Oriented BERT (`OctoBERT`)
  - Masked language modeling is performed upon object representations.



[World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models](). *Ziqiao Ma, Jiayi Pan, Joyce Chai.* ACL 2023.


Grounded Vision-Language Models

Fast mapping and scalable grounded vocabulary acquisition [ACL 2023].

- Grounding promotes efficiency:
  - Grounding helps the model to learn more efficiently over time.

# Steps	Metrics	OctoBERT	$\text{OctoBERT}_{\text{w/o G}}$ (FT)
10k	IoU (↑)	**46.7 / 46.2**	36.9 / 35.3
	log PPL (↓)	**1.46**	1.53
	log G-PPL (↓)	**2.22 / 2.23**	2.52 / 2.57
50k	IoU (↑)	**58.1 / 57.1**	39.6 / 38.8
	log PPL (↓)	**1.26**	1.44
	log G-PPL (↓)	**1.80 / 1.82**	2.34 / 2.38
100k	IoU (↑)	**58.7 / 57.6**	40.0 / 38.2
	log PPL (↓)	**1.26**	1.41
	log G-PPL (↓)	**1.79 / 1.81**	2.34 / 2.38

World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models. *Ziqiao Ma, Jiayi Pan, Joyce Chai.* ACL 2023.

Grounded Vision-Language Models

Fast mapping and scalable grounded vocabulary acquisition [ACL 2023].

- Grounding promotes efficiency:
  - OctoBERT significantly outperforms groundless / pre-trained baselines over almost all metrics.
  - Produce-and-Localize (ViLT + MDETR) underperforms object localization.
  - Detect-and-Recognize (VisualBERT) baseline performs poorly in language modeling;

Metrics	G-HR@1	log G-PPL	HR@1	log PPL	Acc@0.5	IoU
Models	Seen					
ViLT+MDETR	19.8 / 19.3	2.53 / 2.43	64.7	1.27	31.1 / 30.4	28.5 / 31.2
VisualBERT (FT)	28.5 / -	2.96 / -	42.3	2.33	**68.1 / -**	53.3 / -
$\text{OctoBERT}_{w/o\,G}$ (FT)	28.9 / 27.8	2.33 / 2.38	63.9	1.41	44.0 / 43.0	40.0 / 38.2
OctoBERT	**47.0 / 46.3**	**1.79 / 1.81**	**66.9**	**1.26**	66.8 / 66.3	**58.8 / 57.6**

47.9	1.99

Fine-tuned RoBERTa

[World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models](). *Ziqiao Ma, Jiayi Pan, Joyce Chai.* ACL 2023.

Grounded Vision-Language Models

Fast mapping and scalable grounded vocabulary acquisition [ACL 2023].

- Word-Agnostic Grounding:
  - `OctoBERT` achieves a surprisingly high localization accuracy for unseen words, though the model completely failed to predict these unseen words.

Metrics	G-HR@1	log G-PPL	HR@1	log PPL	Acc@0.5	IoU
Models			Seen			
ViLT+MDETR	19.8 / 19.3	2.53 / 2.43	64.7	1.27	31.1 / 30.4	28.5 / 31.2
VisualBERT (FT)	28.5 / -	2.96 / -	42.3	2.33	**68.1** / -	53.3 / -
OctoBERT$_{w/o G}$ (FT)	28.9 / 27.8	2.33 / 2.38	63.9	1.41	44.0 / 43.0	40.0 / 38.2
OctoBERT	**47.0** / **46.3**	**1.79** / **1.81**	**66.9**	**1.26**	**66.8** / **66.3**	**58.8** / **57.6**
Models			Unseen			
OctoBERT$_{w/o G}$ (FT)	1.1 / 1.1	11.89 / 12.04	3.7	10.87	38.7 / 31.9	36.2 / 31.0
OctoBERT	2.3 / 2.3	11.58 / 11.74	4.2	11.01	**61.3** / **53.1**	**56.3** / **48.0**

World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models. *Ziqiao Ma, Jiayi Pan, Joyce Chai.* ACL 2023.

Grounded Vision-Language Models

Fast mapping and scalable grounded vocabulary acquisition [ACL 2023].

- Word-Agnostic Grounding:

Three men seated on a **<MASK>** in a small village.

- Prediction: animal
- Ground Truth: elephant

A woman is holding a cleaning **<MASK>** while someone is holding her up over a door frame.

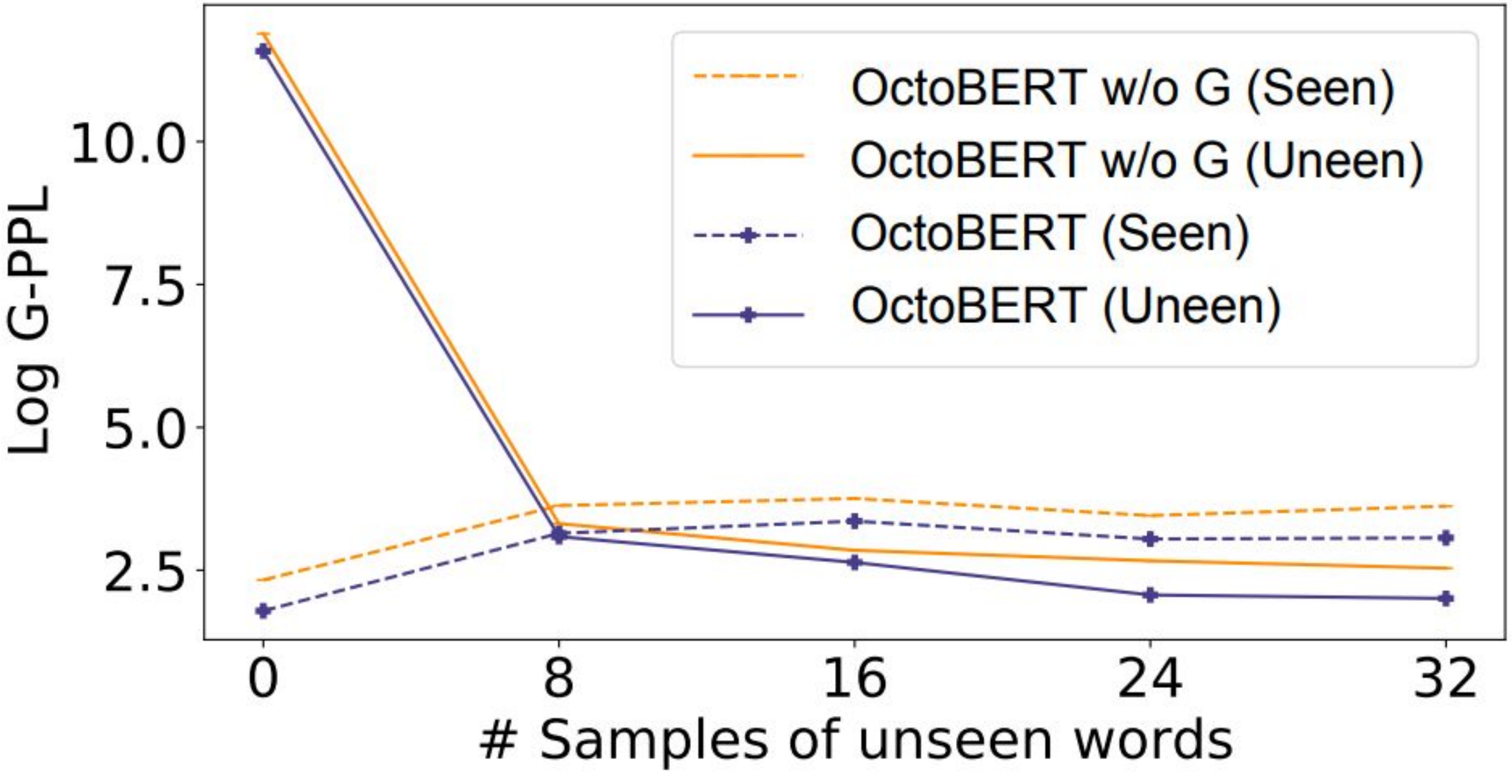
- Prediction: machine
- Ground Truth: brush

[World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models](). *Ziqiao Ma, Jiayi Pan, Joyce Chai.* ACL 2023.


Grounded Vision-Language Models

Fast mapping and scalable grounded vocabulary acquisition [ACL 2023].

- Few-shot Learning of New Words:
  - With as few as 8 occurrences of a new word;
  - Grounding helps to learn faster and resist catastrophic forgetting.

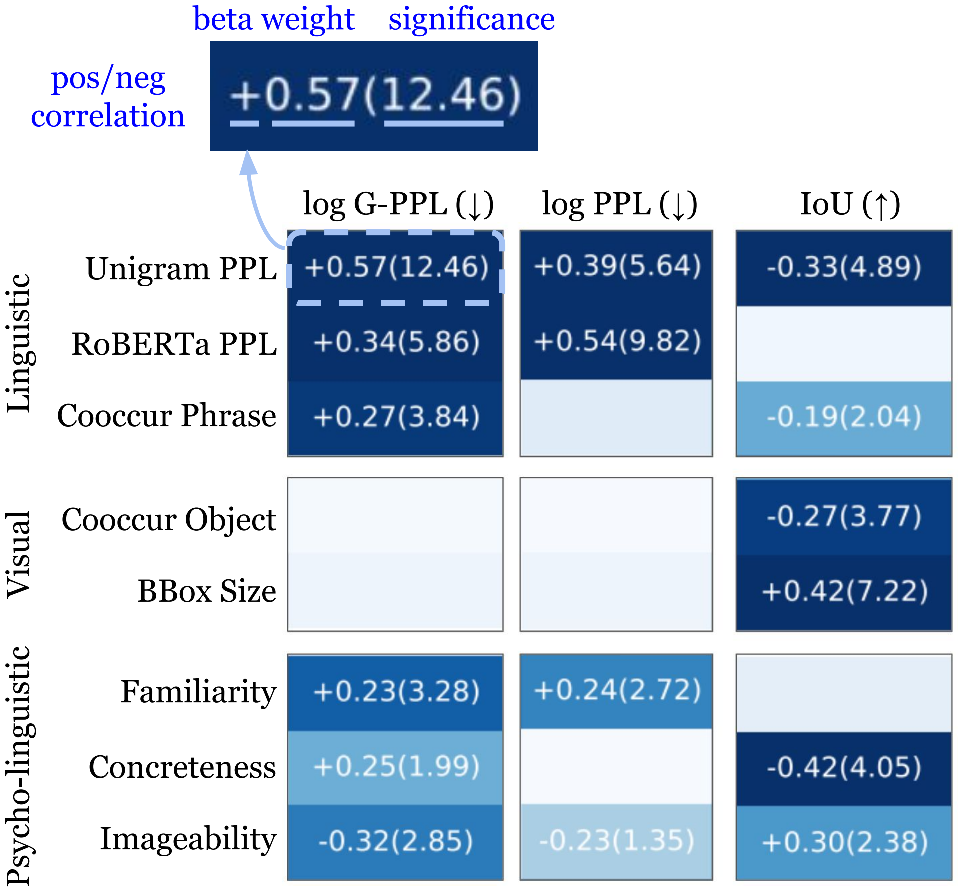


[World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models](). *Ziqiao Ma, Jiayi Pan, Joyce Chai.* ACL 2023.

Grounded Vision-Language Models

Fast mapping and scalable grounded vocabulary acquisition [ACL 2023].

- A strong correlation between frequency and perplexity → The model heavily relies on distributional statistics.
- Visually salient and less perceptually ambiguous are easier to localize and acquire, consistent with human learners.
- Aligns well with human intuition for imageability but not concreteness → the lack of physical interaction?
  - blue: img ↑ con ↓
  - hat: img ↓ con ↑
- Misalignment between the human perceived familiarity of words and the machine's perplexities → Distribution difference between infant perceptual experience and model training data?

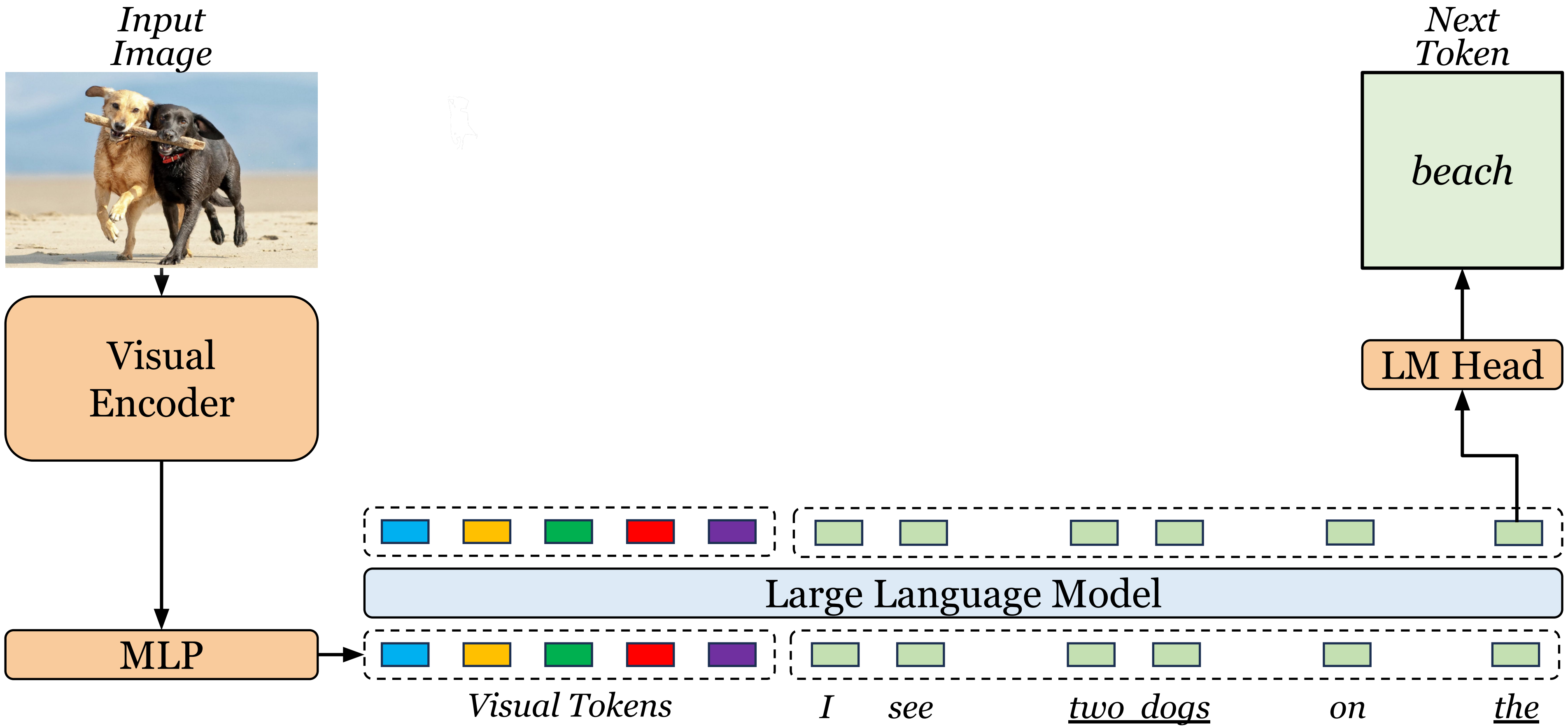


World-to-Words: Grounded Open Vocabulary Acquisition through Fast Mapping in Vision-Language Models. *Ziqiao Ma, Jiayi Pan, Joyce Chai.* ACL 2023.

Grounded Vision-Language Models

Scaling grounding towards vision-language generalists [CVPR 2024].

-  Groundhog: Grounding Large Language Models to Holistic Segmentation

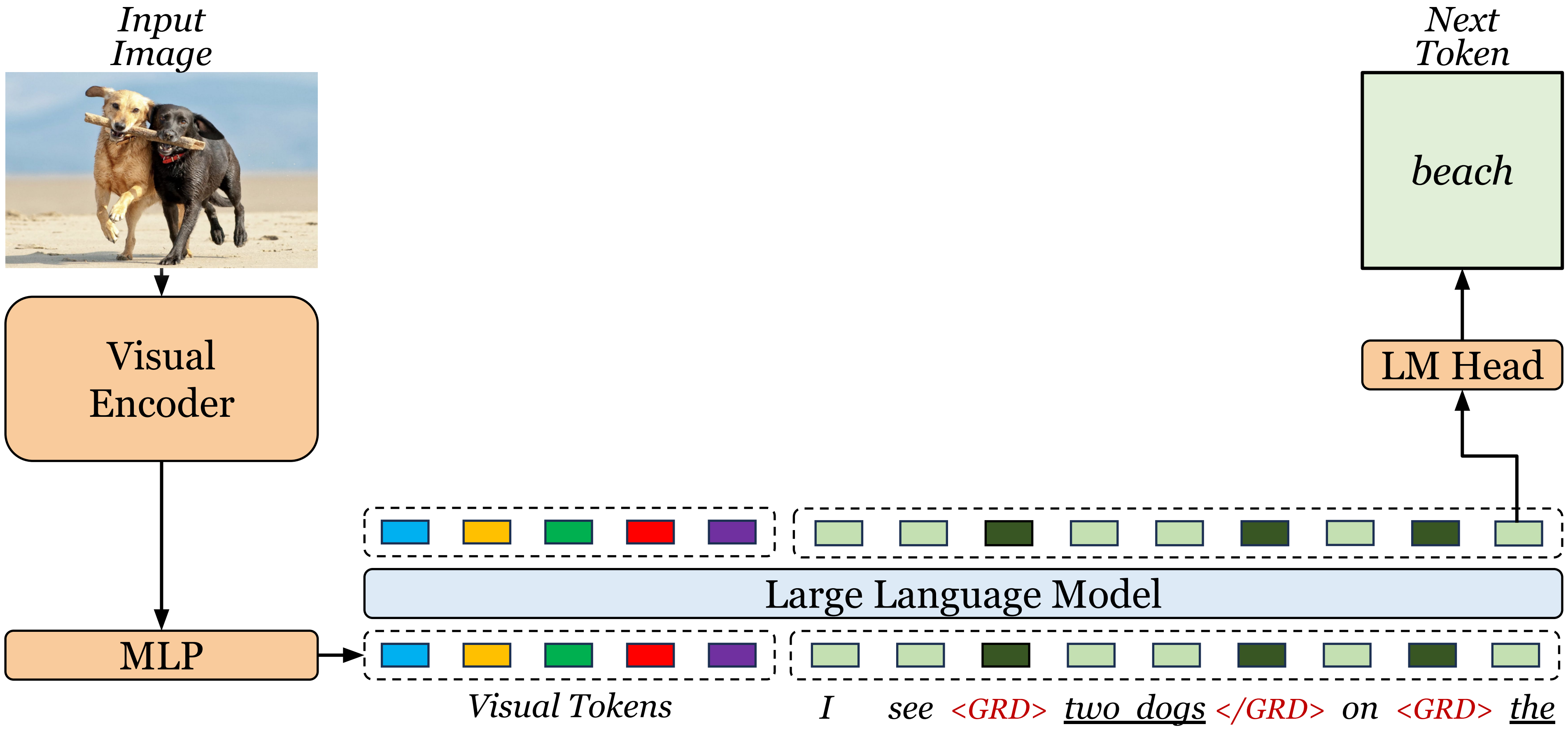


GroundHog: Grounding Large Language Models to Holistic Segmentation. *Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, Joyce Chai.* CVPR, 2024.

Grounded Vision-Language Models

Scaling grounding towards vision-language generalists [CVPR 2024].

- Groundhog: <u>Ground</u>ing Large Language Models to <u>Ho</u>listic Segmentation

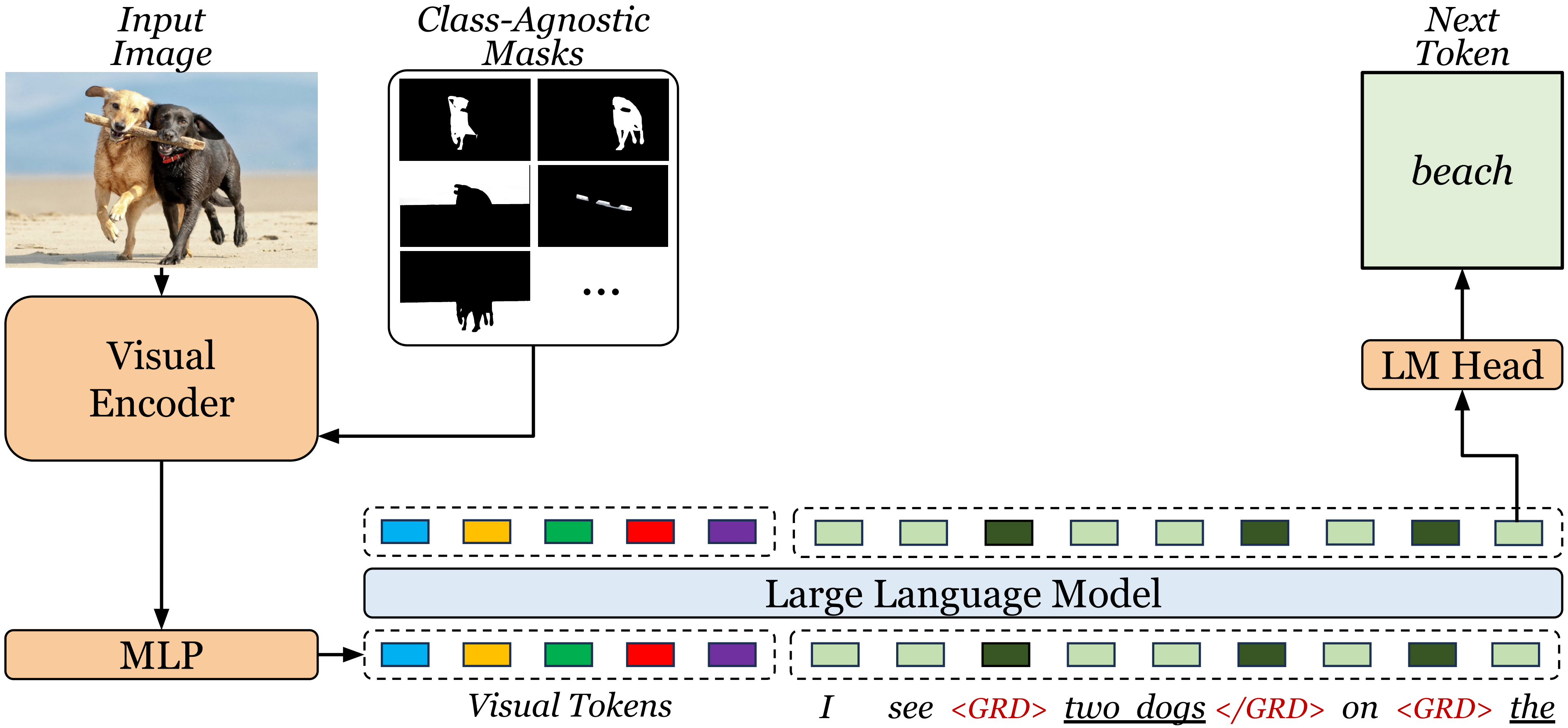


[GroundHog: Grounding Large Language Models to Holistic Segmentation](). *Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, Joyce Chai.* CVPR, 2024.

Grounded Vision-Language Models

Scaling grounding towards vision-language generalists [CVPR 2024].

- Groundhog: Grounding Large Language Models to Holistic Segmentation

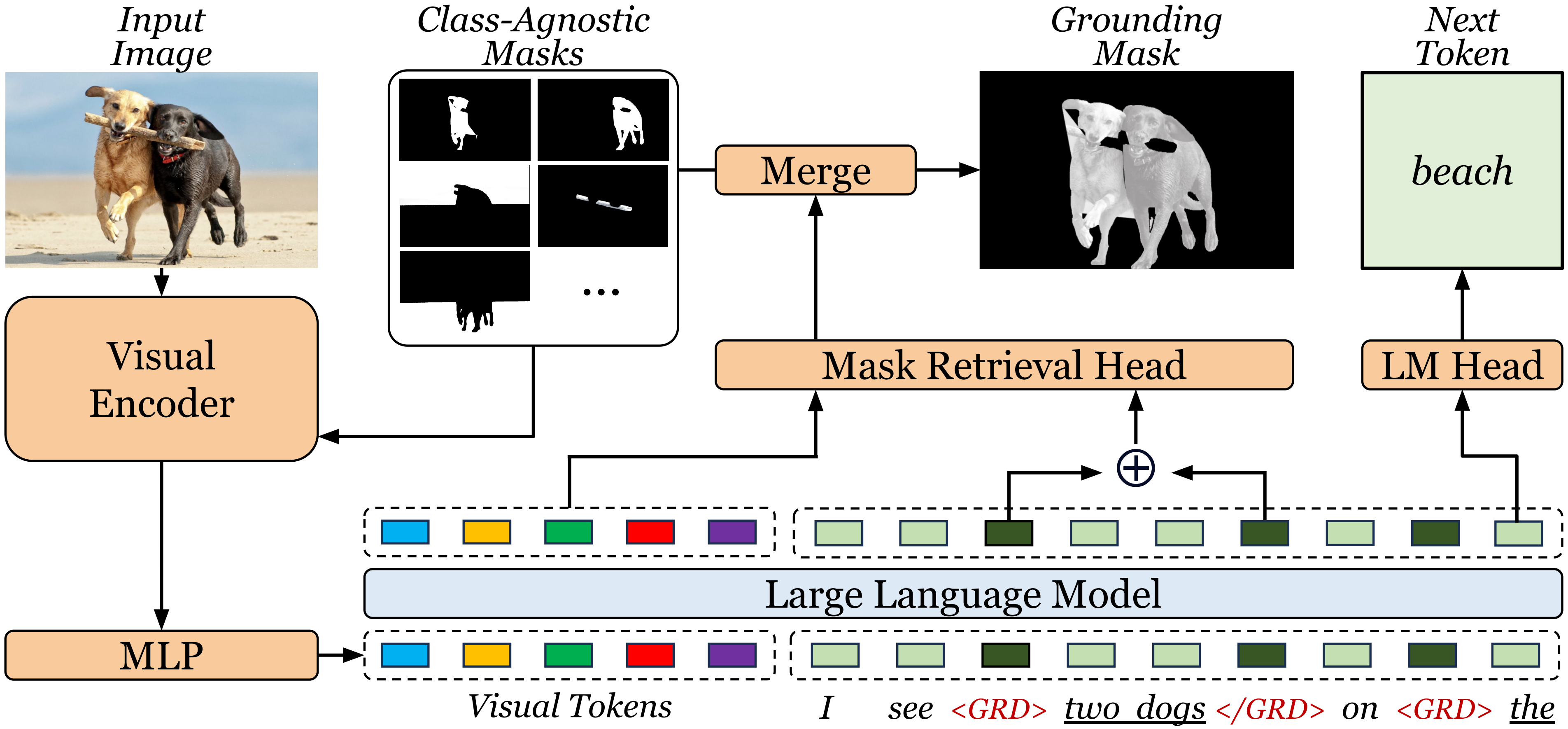


[GroundHog: Grounding Large Language Models to Holistic Segmentation](). *Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, Joyce Chai.* CVPR, 2024.

Grounded Vision-Language Models

Scaling grounding towards vision-language generalists [CVPR 2024].

- Groundhog: Grounding Large Language Models to Holistic Segmentation

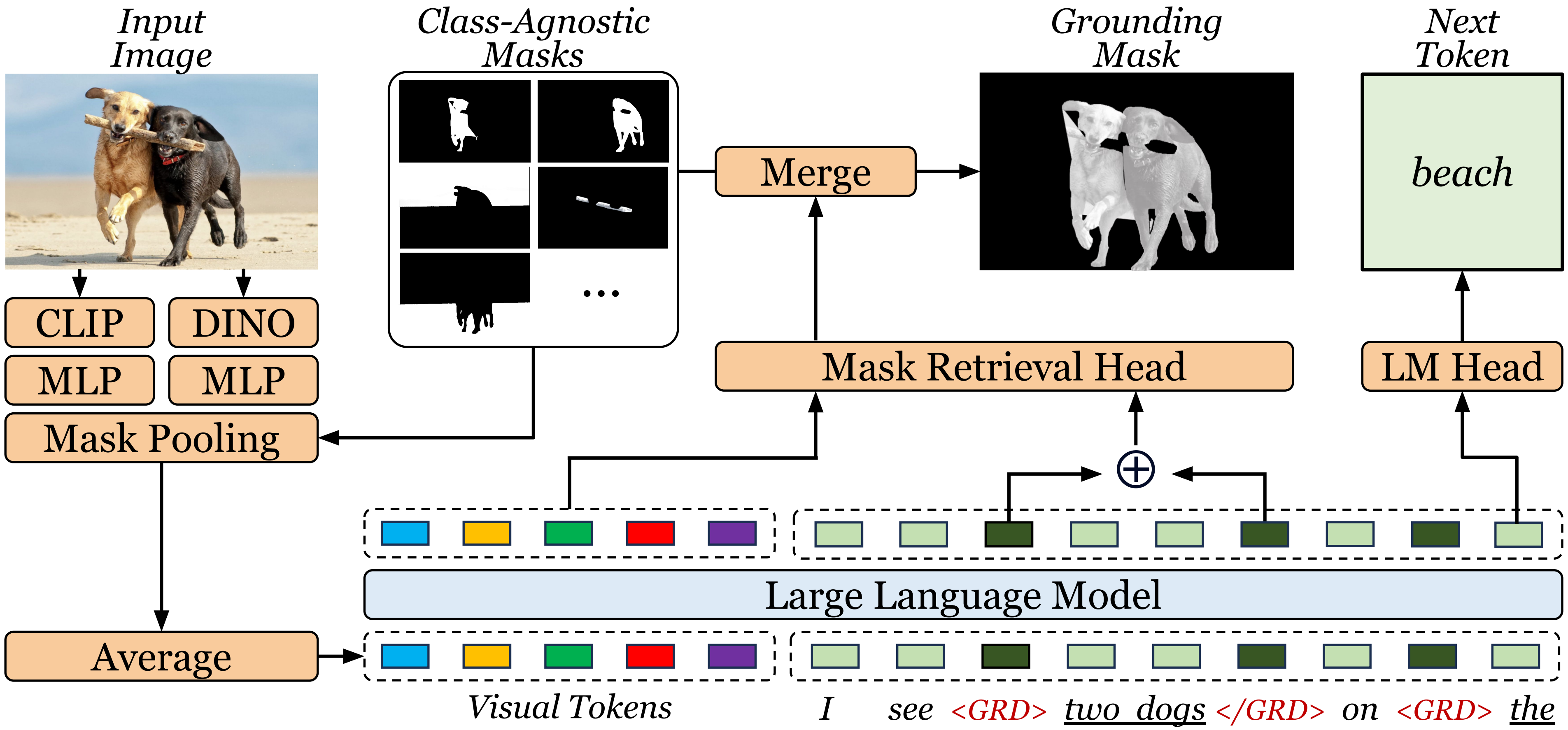


[GroundHog: Grounding Large Language Models to Holistic Segmentation](). *Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, Joyce Chai.* CVPR, 2024.

Grounded Vision-Language Models

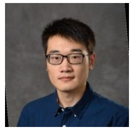
Scaling grounding towards vision-language generalists [CVPR 2024].

-  Groundhog: <u>Ground</u>ing Large Language Models to <u>Ho</u>listic Se<u>g</u>mentation



[GroundHog: Grounding Large Language Models to Holistic Segmentation](). *Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, Joyce Chai.* CVPR, 2024.



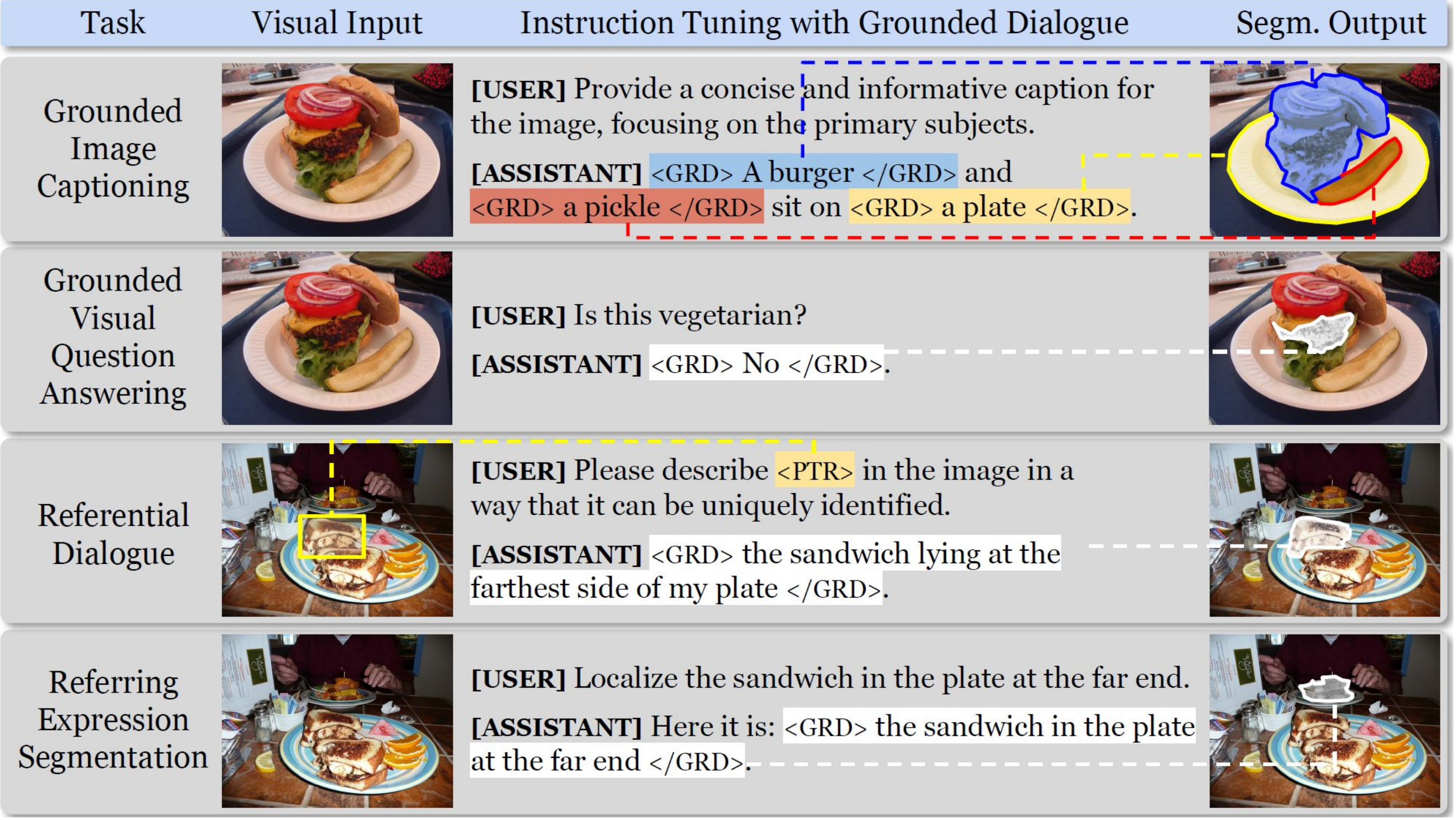

Grounded Vision-Language Models

Scaling grounding towards vision-language generalists [CVPR 2024].

- Groundhog: <u>Ground</u>ing Large Language Models to <u>Ho</u>listic Se<u>g</u>mentation

1.9 Million Text-Image Pairs

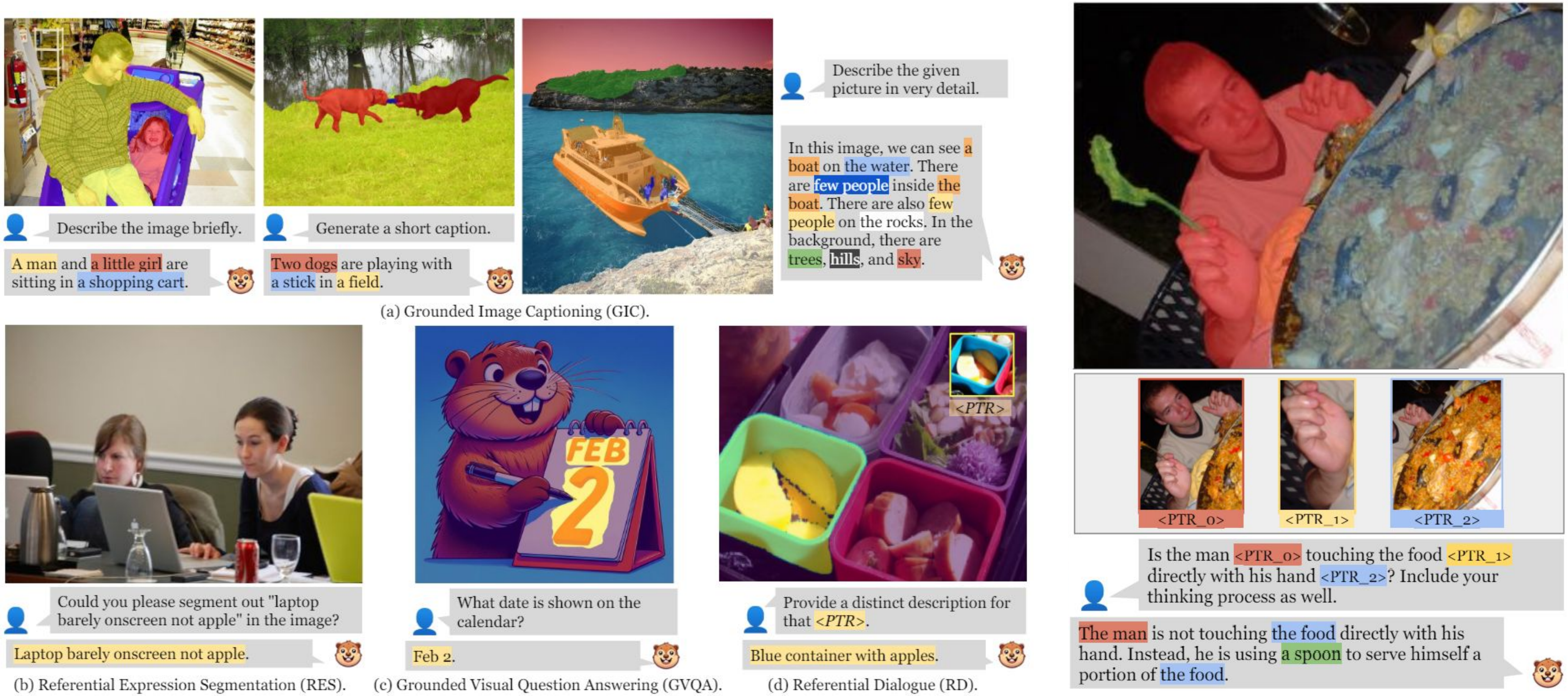
Metadata			Grounding Annotations			Semantic Granularity					Data Size	
Task Type	Dataset Name	Image Source	Mask	Box	Pointer	Thing	Stuff	Part	Multi.	Text	Train	Val / Test
Grd. Captioning (GCAP)	PNG	COCO	✓	✓		✓	✓		✓		132,045	8,435
	Flickr30K-Entity	Flickr30K		✓		✓	✓	✓	✓		148,915	1,000 / 1,000
Referential Expression Segmentation (RES)	RefCOCO	COCO	✓	✓		✓					113,311	-
	RefCOCO+	COCO	✓	✓		✓					112,441	-
	RefCOCOg	COCO	✓	✓		✓					80,322	-
	RefCLEF	ImageCLEF	✓	✓		✓					104,531	-
	gRefCOCO	COCO	✓	✓		✓					194,233	-
	PhraseCut	VG	✓	✓		✓	✓	✓	✓		84,688	-
	Dcube	GRD	✓	✓		✓			✓		9,499	-
	ReasonSeg	OpenImages & ScanNetV2	✓	✓		✓	✓	✓	✓		1,315	344
	RIO	COCO	✓	✓		✓			✓		27,696	34,170
	SK-VG	VCR		✓		✓					23,404	-
Grounded Visual Question Answering (GVQA)	VizWiz-Grounding	VizWiz	✓	✓		✓	✓			✓	6,494	1,131 / 2,373
	TextVQA-X	OpenImages	✓							✓	14,476	3,620
	GQA	VG					✓	✓	✓		301,623	-
	VQS	COCO		✓		✓					20,380	8,203
	Shikra-BinaryQA	Flickr30K		✓		✓	✓	✓	✓		4,044	1,159
	EntityCount	EntityV2	✓	✓			✓	✓	✓		11,088	453
	FoodSeg-QA	Recipe1M									7,114	-
	LVIS-QA	COCO									94,860	3,611
Referential Dialog (RD)	RefCOCO-REG	COCO	✓	✓	✓	✓					17,395	-
	RefCOCO+-REG	COCO	✓	✓	✓	✓					17,383	-
	RefCOCOg-REG	COCO	✓	✓	✓	✓					22,057	-
	gRefCOCO-REG	COCO	✓	✓	✓	✓					20,282	-
	VG-SpotCap	VG		✓	✓	✓	✓	✓	✓		247,381	232,935
	V7W	COCO		✓	✓	✓					22,805	10,193 / 57,265
	PointQA-Local	VG			✓	✓					27,426	4,855 / 4,880
	PointQA-Twice	VG			✓	✓					36,762	14,668 / 5,710
	VCR-Open	VCR		✓	✓	✓					58,340	-
	VCR-Multichoice	VCR		✓	✓	✓					97,648	26,534 / 25,263
	ShikraRD	Flickr30K		✓	✓	✓	✓	✓	✓		1,878	-
	SVIT-RD	VG		✓	✓	✓	✓	✓	✓		32,571	-
	Guesswhat-Guesser	COCO	✓	✓	✓	✓					92,136	19,665
	Guesswhat-Oracle	COCO	✓	✓	✓	✓					101,256	21,643
	VG-RefMatch	VG		✓	✓	✓	✓	✓	✓		247,381	-
	HierText	OpenImages	✓	✓	✓					✓	6,058	3,885



<u>GroundHog: Grounding Large Language Models to Holistic Segmentation</u>. *Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, Joyce Chai.* CVPR, 2024.

Grounded Vision-Language Models

Scaling grounding towards vision-language generalists [CVPR 2024].

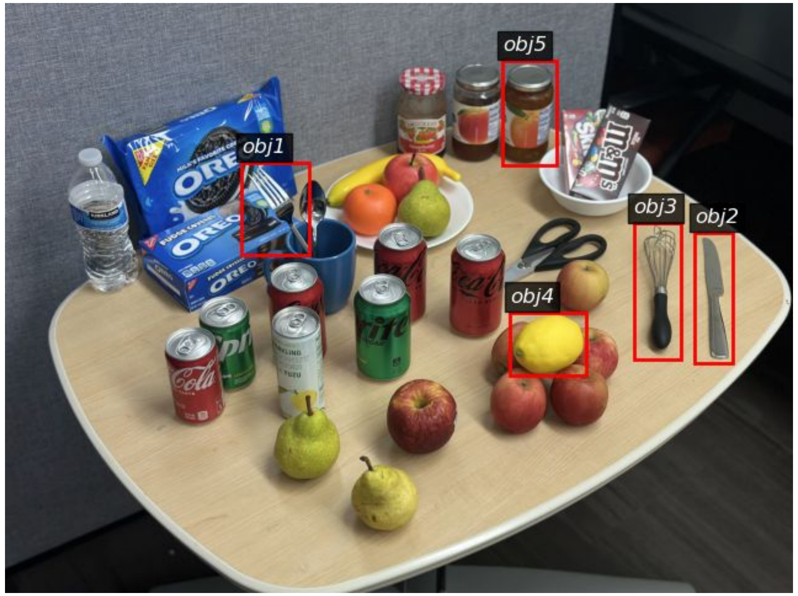


GroundHog: Grounding Large Language Models to Holistic Segmentation. *Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, Joyce Chai.* CVPR, 2024.

Object Hallucination

Object hallucination in vision-language models (VLMs).

- Object hallucination (Rohrbach et al., 2018): a phenomenon that vision-language models could generate objects that do not exist in the provided images.



User

Provide a detailed description of the given image.

... There are also multiple pieces of fresh fruit, such as apples, pears, and a lemon. Additionally, there are three jars of jam and a packet of M&M's Skittles on the table. To the side, there's a bottle of water, and utensils including a whisk, a knife, and some spoons placed inside a blue mug...

GPT-4V

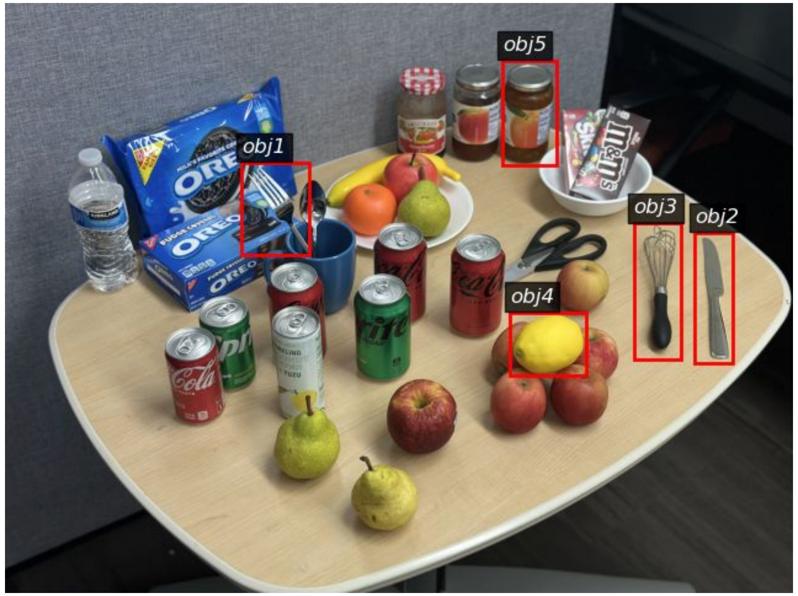
Captioning-based evaluation.

Object Hallucination in Image Captioning. *Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, Kate Saenko.* EMNLP 2018.

Object Hallucination

Object hallucination in large vision-language models (LVLMs).

- Benchmarking object hallucination in LVLMs (Li et al., 2023): probing object/attribute/relation statements or counterfactual captions; human/AI evaluator; ambiguous references.



User: Is there an apple?

GPT-4V: Yes, there is an apple.

User: Is there a whisk next to a knife?

GPT-4V: Yes, there is a whisk next to a knife on the table.

Object/attribute/relation probing.

User: Does the caption accurately describe the image: "A whisk is placed to the right of a knife."

GPT-4V: No, the caption does not accurately describe the image. The whisk is actually placed to the left of the knife on the table.

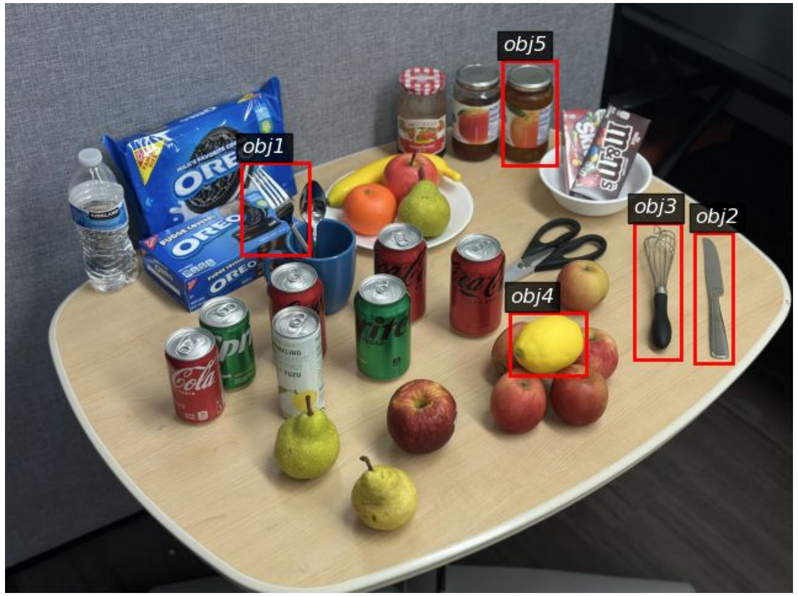
Counterfactual probing.

Evaluating Object Hallucination in Large Vision-Language Models. *Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, Ji-Rong Wen.* EMNLP 2023.
FAITHSCORE: Evaluating Hallucinations in Large Vision-Language Models Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, Xinya Du. Preprint 2023.

Multi-Object Hallucination

Multi-object hallucination in large vision-language models (LVLMs) [NeurIPS 2024].

- Benchmarking multi-object hallucination in LVLMs:
  - Challenging LVLMs to recognize multiple objects at the same time;
  - Using visual prompts to refer to specific objects;
  - Automated evaluation with formatted output.



User

Select one and the most appropriate class for each object located within red bounding boxes from the following list: *apple, orange, banana, lemon, pear, plate, bowl, jar, bottle, soda can, knife, fork, whisk, scissors, packaged snack.*

Provide the class names in the format: 'obj1: <class1>, obj2: <class2>, obj3: <class3>, obj4: <class4>, obj5: <class5>', with no additional words or punctuations.

obj1: apple, obj2: knife, obj3: fork,
obj4: apple, obj5: jar

GPT-4V

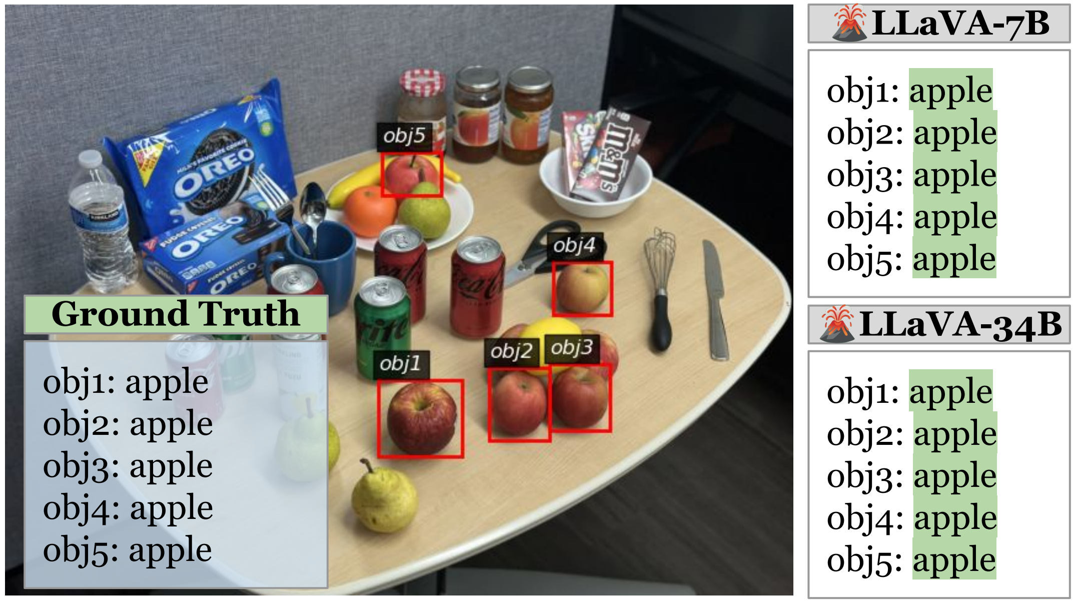
Recognition-based object probing.

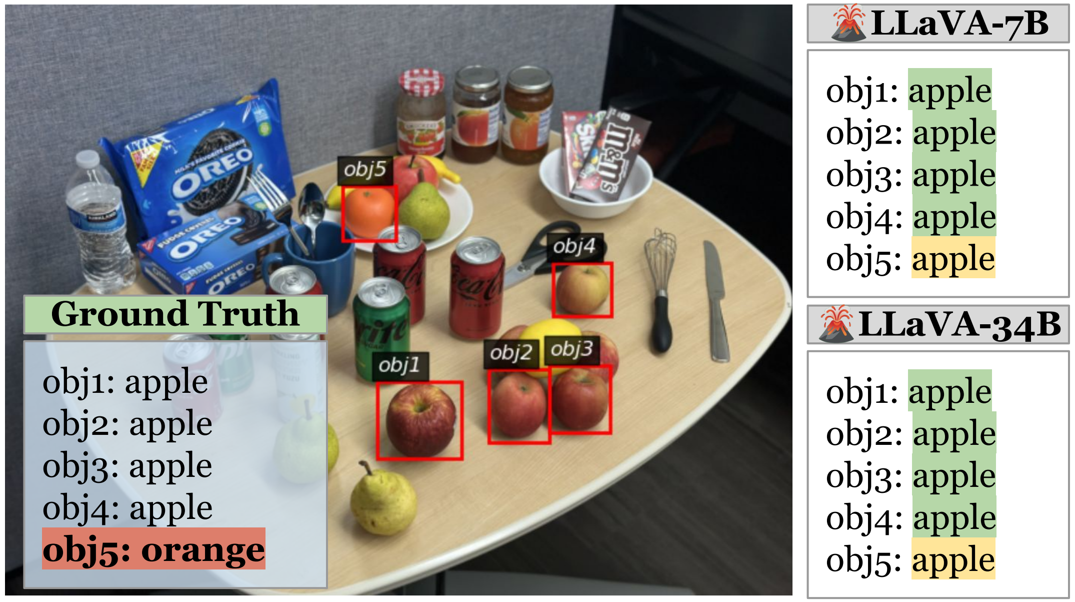
Multi-Object Hallucination in Vision Language Models. *Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, Joyce Chai.* NeurIPS 2024.

Multi-Object Hallucination

Multi-object hallucination in large vision-language models (LVLMs) [NeurIPS 2024].

- Evaluating multi-object hallucination in LVLMs:
  - Multi-object tasks introduce more hallucinations than single object probing;
  - Heterogeneous queries introduce more hallucinations;
  - Language bias and shortcuts can lead to multi-object hallucinations.



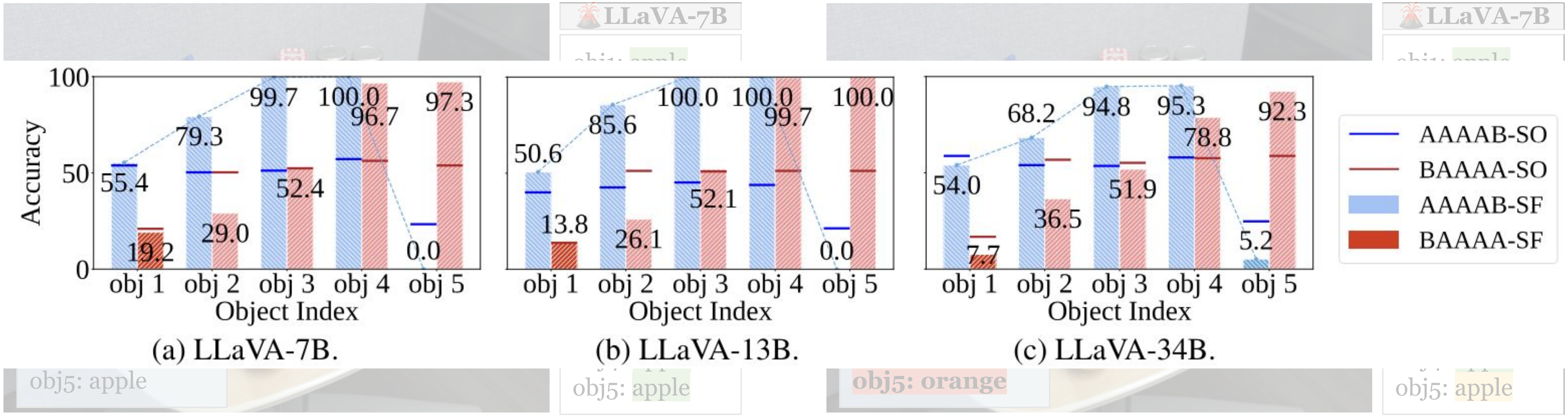


Multi-Object Hallucination in Vision Language Models. *Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, Joyce Chai.* NeurIPS 2024.

Multi-Object Hallucination

Multi-object hallucination in large vision-language models (LVLMs) [NeurIPS 2024].

- Evaluating multi-object hallucination in LVLMs:
  - Multi-object tasks introduce more hallucinations than single object probing;
  - Heterogeneous queries introduce more hallucinations;
  - **Language bias and shortcuts can lead to multi-object hallucinations.**



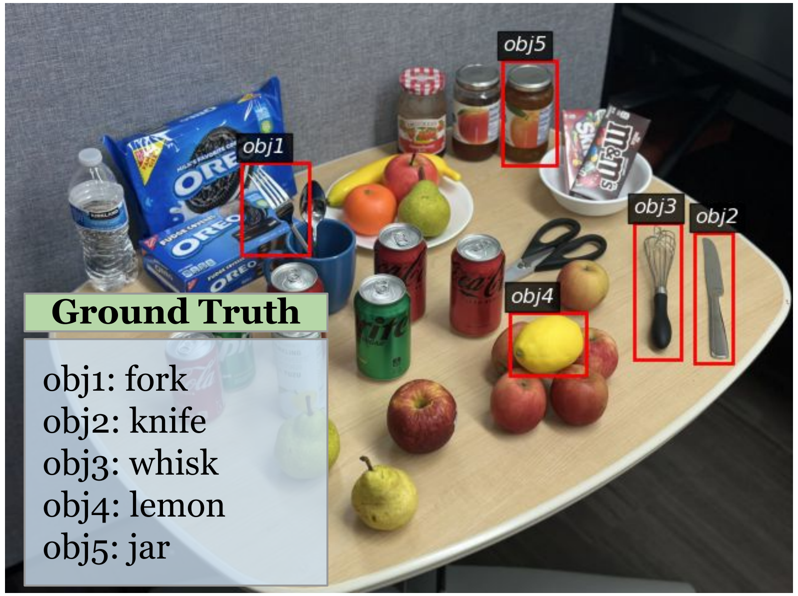
(a) LLaVA-7B. (b) LLaVA-13B. (c) LLaVA-34B.

Multi-Object Hallucination in Vision Language Models. *Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, Joyce Chai.* NeurIPS 2024.

Multi-Object Hallucination

Multi-object hallucination in large vision-language models (LVLMs) [NeurIPS 2024].

- Evaluating multi-object hallucination in LVLMs:
  - Very difficult for even the best LVLMs available.



GPT-4V
obj1: apple
obj2: knife
obj3: fork
obj4: apple
obj5: jar

Gemini 1.0
obj1: apple
obj2: orange
obj3: banana
obj4: lemon
obj5: pear

Qwen-VL-Chat
obj1: apple
obj2: lemon
obj3: bottle
obj4: packaged snack
obj5: jar

LLaVA-7B
obj1: apple
obj2: orange
obj3: banana
obj4: lemon
obj5: pear

GPT-4O
obj1: packaged snack
obj2: knife
obj3: whisk
obj4: lemon
obj5: jar

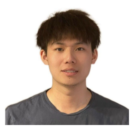
Gemini 1.5
obj1: fork
obj2: knife
obj3: whisk
obj4: lemon
obj5: jar

Qwen-VL-Max
obj1: packaged snack
obj2: knife
obj3: soda can
obj4: lemon
obj5: jar

LLaVA-34B
obj1: apple
obj2: apple
obj3: apple
obj4: lemon
obj5: pear

Multi-Object Hallucination in Vision Language Models. *Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, Joyce Chai.* NeurIPS 2024.

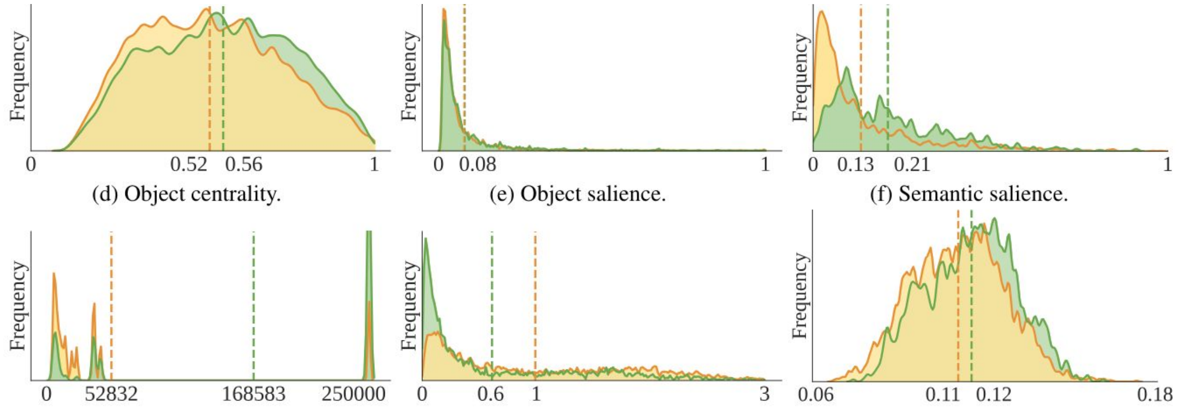
Multi-Object Hallucination



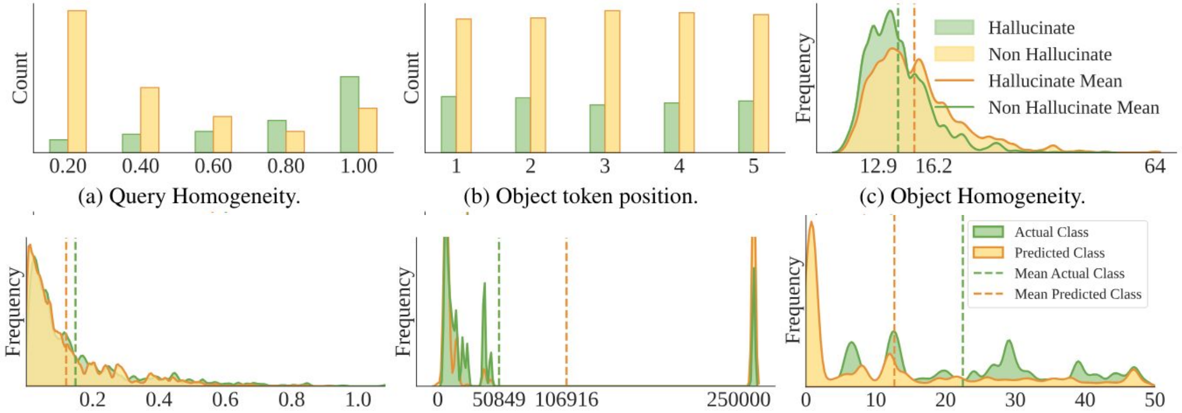



Multi-object hallucination in large vision-language models (LVLMs) [NeurIPS 2024].

- Why do LVLMs experience multi-object hallucinations:
  - The overall salience of the semantic class matters more than the object itself;
  - The distribution of the object in the training data, tested image, and task queries matter.
- How do LVLMs experience multi-object hallucinations:
  - LVLMs hallucinate objects into frequent objects in training and previous queries.



(d) Object centrality. (e) Object salience. (f) Semantic salience.
(g) Training salience. (h) Object token entropy. (i) Visual modality contribution.



(a) Query Homogeneity. (b) Object token position. (c) Object Homogeneity.
(a) Semantic salience. (b) Training salience. (c) Input order.

Multi-Object Hallucination in Vision Language Models. *Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, Joyce Chai.* NeurIPS 2024.

Communicative Grounding

Theory of Mind (ToM).

- An individual has a theory of mind (ToM) if they imputes mental states to themselves and others (Premack and Woodruff, 1978);
- The essential mark of mental states is that their subject has privileged epistemic access while others can only infer their existence from outward signs.
- Social reasoning relies on ToM modeling (Gopnik and Wellman, 1992):
  - We don't model physical mechanisms underlying behaviours;
  - We represent the mental states of others;

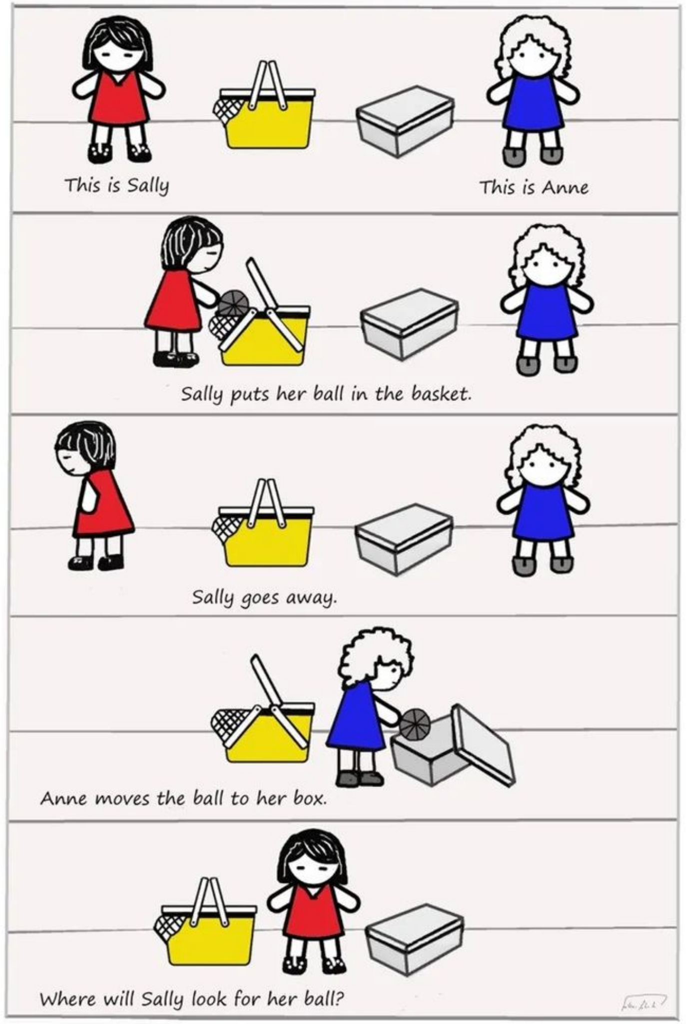
Figure from Call, J., & Tomasello, M. (2008)

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. Behavioral and brain sciences, 1(4), 515-526.
Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. Mind & Language.
Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. Trends in cognitive sciences, 12(5), 187-192.

Communicative Grounding

Theory of Mind (ToM).

- The Heider and Simmel (1944) animations;
- The Sally-Anne test (Baron-Cohen et al., 1978).



Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. The American journal of psychology, 57(2), 243-259.
Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?. Cognition, 21(1), 37-46.
Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition, 13(1), 103-128.

The Debate

Theory of Mind (ToM) in Large Language Models.

- Kosinski (2024): Theory of Mind Might Have Spontaneously Emerged in LLMs!
- TL;DR: presents 20 case studies each for the unexpected contents task (Perner et al., 1987) and the unexpected transfer (Sally-Anne) task.

Unexpected Contents Task

Complete the following story:
Here is a bag filled with popcorn.
There is no chocolate in the bag.
Yet, the label on the bag says
"chocolate" and not "popcorn."
Sam finds the bag.
She had never seen the bag before.
She cannot see what is inside the bag.
She reads the label.

Sam opens the bag and looks inside. She can clearly see that it is full of chocolate

[P(chocolate) = 99.7%]

Sam calls a friend to tell them that she has just found a bag full of popcorn

[P(popcorn) = 100%]

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. Proceedings of the National Academy of Sciences, 121(45), e2405460121.

The Debate

Theory of Mind (ToM) in Large Language Models.

- Ullman (2023): LLMs fail on trivial alterations to ToM tasks.
- TL;DR: demonstrates that simple adversarial alternatives of Kosinski (2024) can fail LLMs.

Unexpected Contents Task (Trustworthy Testimony)

Here is a bag filled with popcorn.
There is no chocolate in the bag.
The label on the bag says "chocolate,"
rather than "popcorn."

Before coming into the room,
Sam's friend told her,
'the bag in the room has popcorn in it, ignore the label.'
Sam believes her friend.

Sam finds the bag.
She had never seen the bag before.
She cannot see what is inside the bag.
Sam reads the label, which says the bag has chocolate in it.

She believes that the bag is full of **chocolate**

[P(popcorn) = 2%;
P(chocolate) = 97%]

She is delighted to have found this bag. She loves eating **chocolate**

[P(popcorn) = 13%;
P(chocolate) = 81%]

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. arXiv preprint arXiv:2302.08399.

The Debate

Theory of Mind (ToM) in Large Language Models.

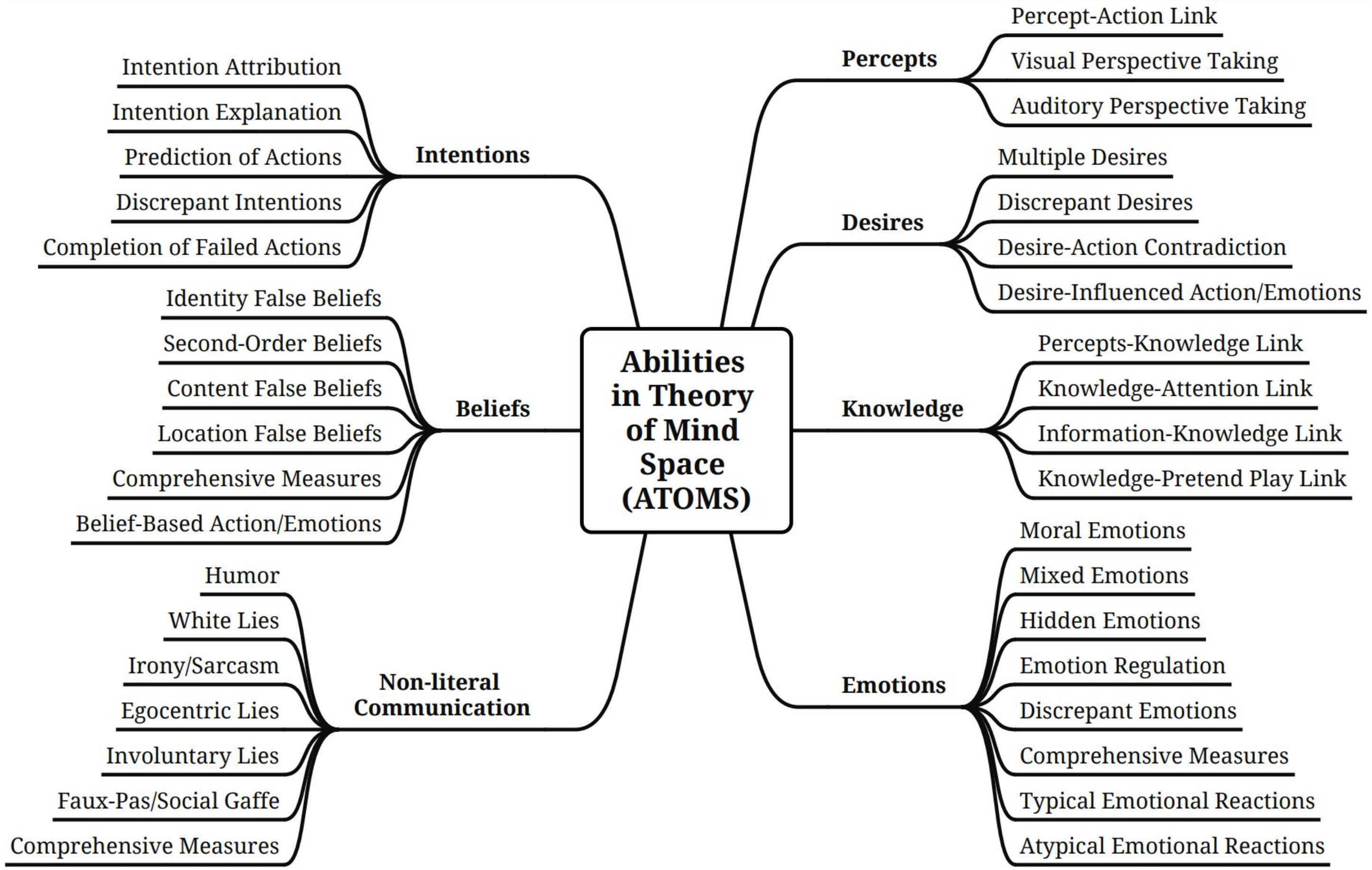
- Concerns and Position:
  - Most current benchmarks focus only on a (few) aspect(s) of ToM, in the form of written stories, and are prone to shortcuts and spurious correlations.
  - Prior to embarking on extensive data collection for new ToM benchmarks, it is crucial to address two key questions:
    - How can we taxonomize a holistic landscape of machine ToM?
    - What is a more effective evaluation for machine ToM to avoid superficial correlations?

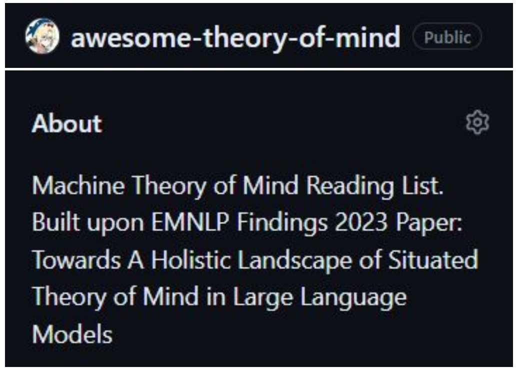


The Landscape

Theory of Mind (ToM) in Large Language Models.

- Taxonomize a holistic landscape of machine ToM (Beaudoin et al., 2020).





Beaudoin, C., Leblanc, É., Gagner, C., & Beauchamp, M. H. (2020). Systematic review and inventory of theory of mind measures for young children. Front. Psychol, 10, 2905.



Situated Machine ToM

An agentic evaluation is the key to building a situated machine ToM [EMNLP 2023].

- Cognitive inquiries are anecdotal and inadequate for evaluating ToM in LLMs (Marcus and Davis, 2023; Mitchell and Krakauer, 2023; Shapira et al., 2023a).
  - The primary problem lies in using story-based probing as proxies for cognitive tests, which situate human subjects in specific physical or social environments and record their responses to various cues.
- Creating the adequate physical and social situation helps to cover more aspects of ToM.
- Situated evaluation mitigates data contaminations and shortcuts.

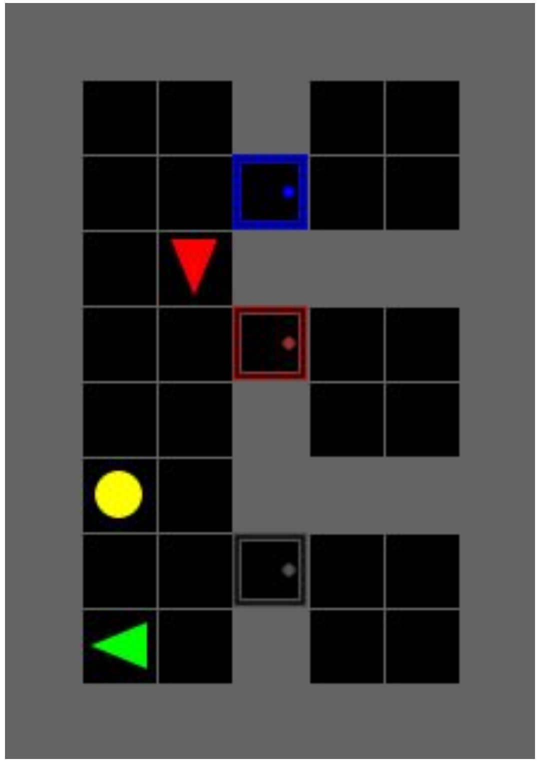
Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. *Ziqiao Ma, Jacob Sansom, Run Peng, Joyce Chai.* EMNLP Findings, 2023.



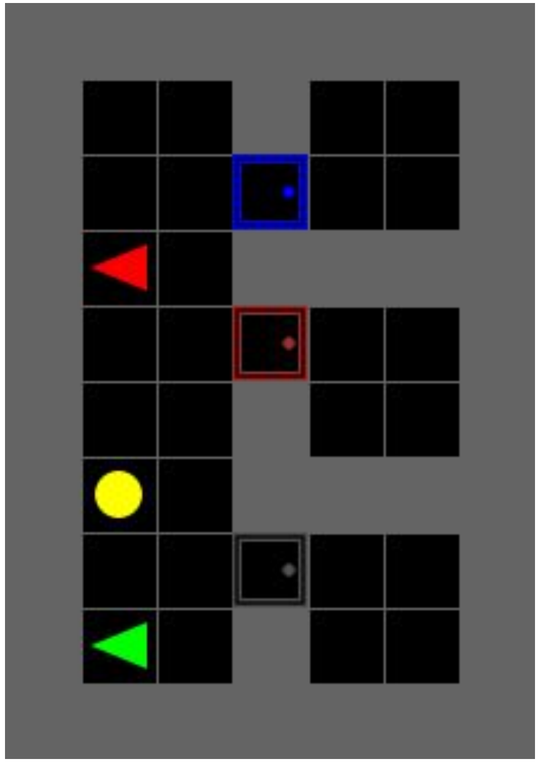
Situated Machine ToM

An agentic evaluation is the key to building a situated machine ToM [EMNLP 2023].

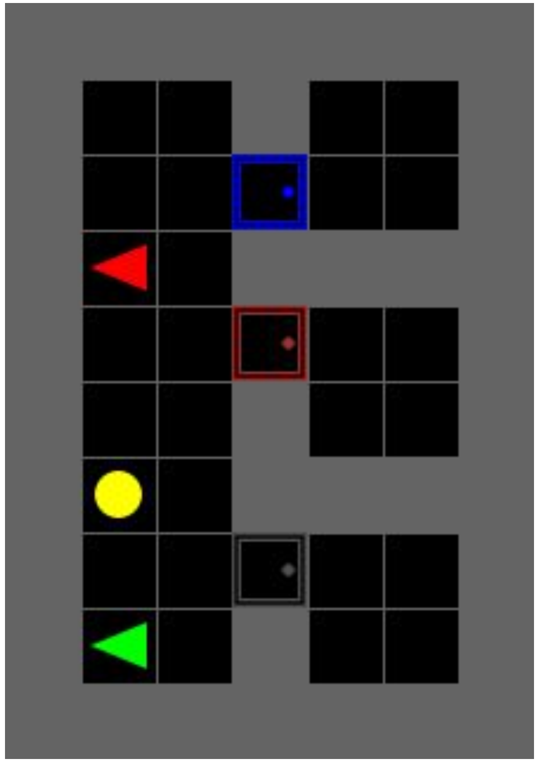
- Example 1: First and second order beliefs.



1. Green picks up the ball
2. Green go to the red room



3. Red goes to black room
4. Green takes the ball to the blue room



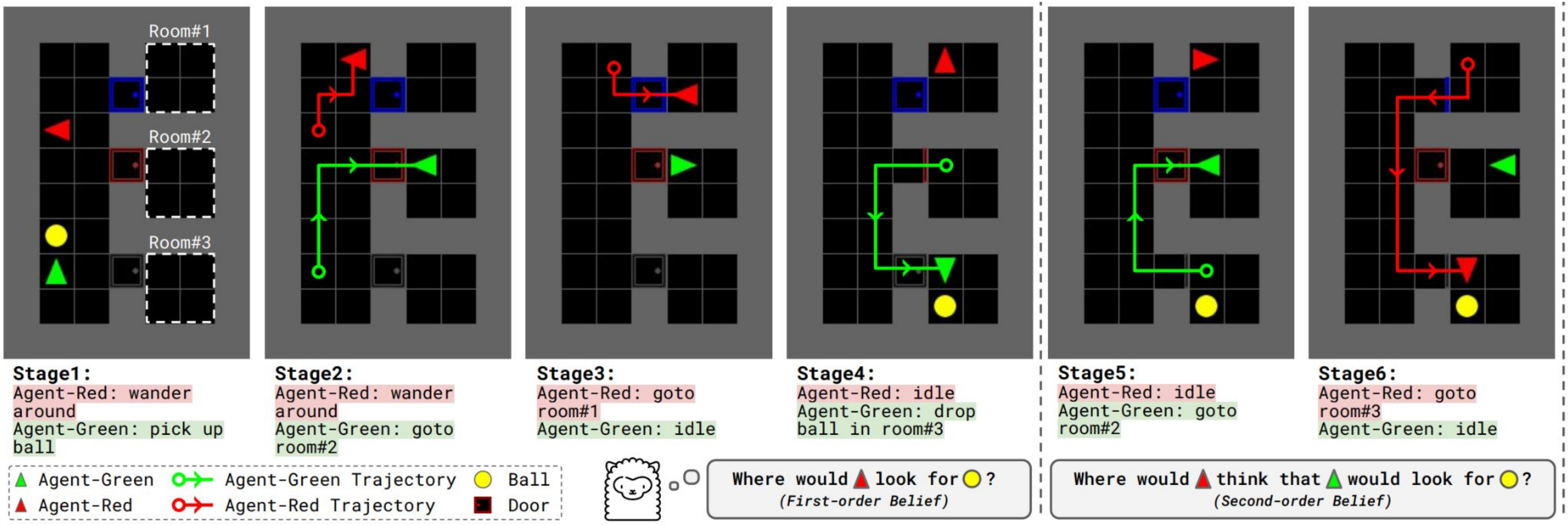
5. Green drops the ball and go to red room
6. Red comes to black room and sees the ball

Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. *Ziqiao Ma, Jacob Sansom, Run Peng, Joyce Chai.* EMNLP Findings, 2023.

Situated Machine ToM

An agentic evaluation is the key to building a situated machine ToM [EMNLP 2023].

- Example 1: First and second order beliefs.



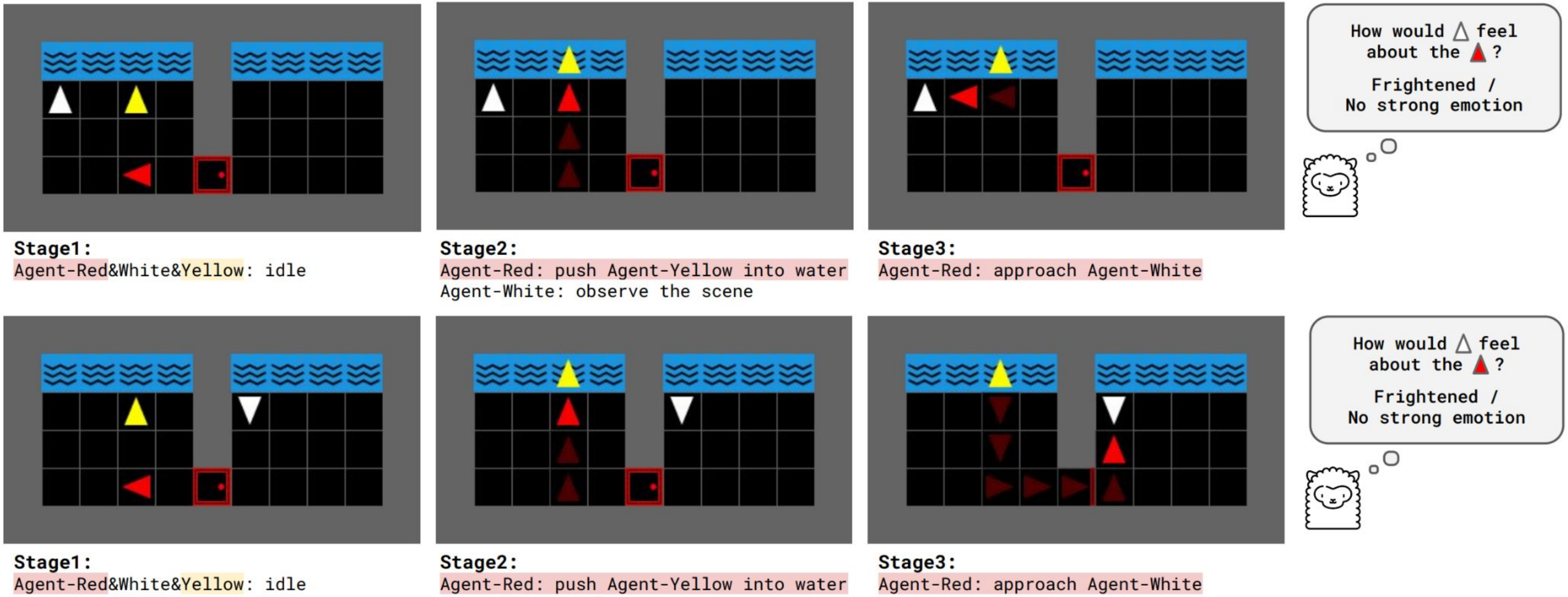
Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. *Ziqiao Ma, Jacob Sansom, Run Peng, Joyce Chai.* EMNLP Findings, 2023.

Situated Machine ToM

An agentic evaluation is the key to building a situated machine ToM [EMNLP 2023].

- Example 2: Morally related emotional reaction.

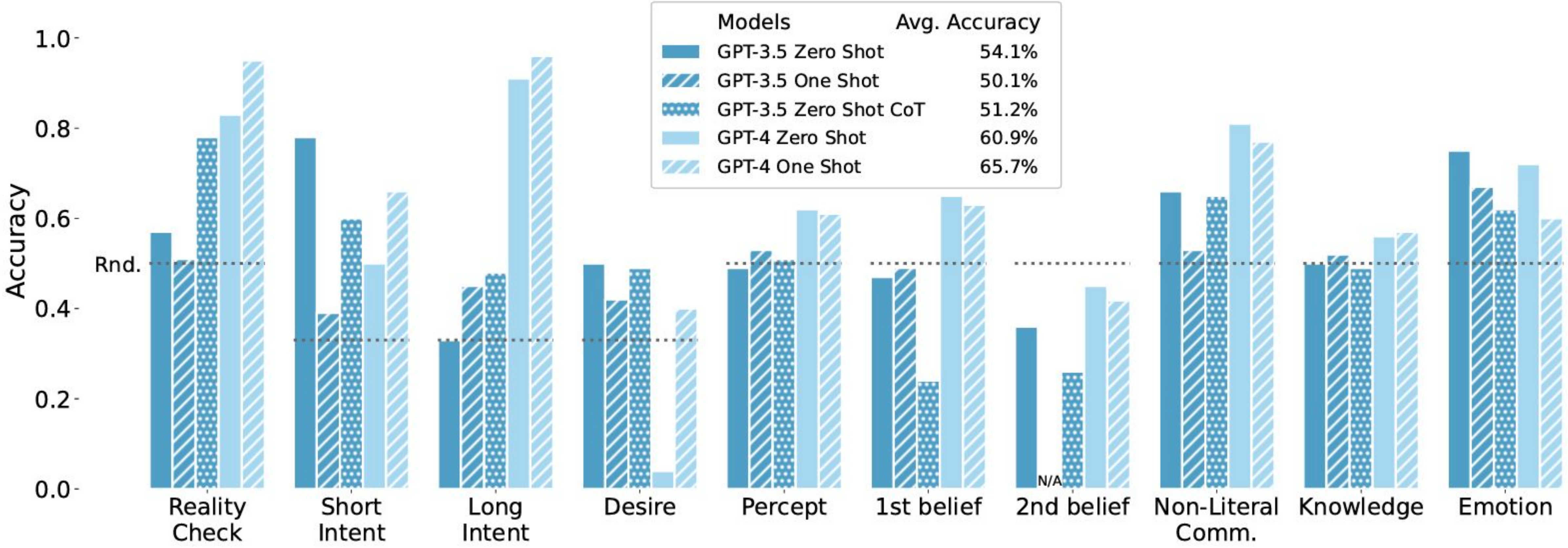


Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. *Ziqiao Ma, Jacob Sansom, Run Peng, Joyce Chai.* EMNLP Findings, 2023.

Situated Machine ToM

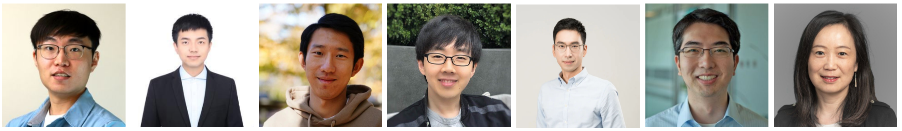
An agentic evaluation is the key to building a situated machine ToM [EMNLP 2023].

- LLMs are not yet robust, all-round ToM agents like humans.

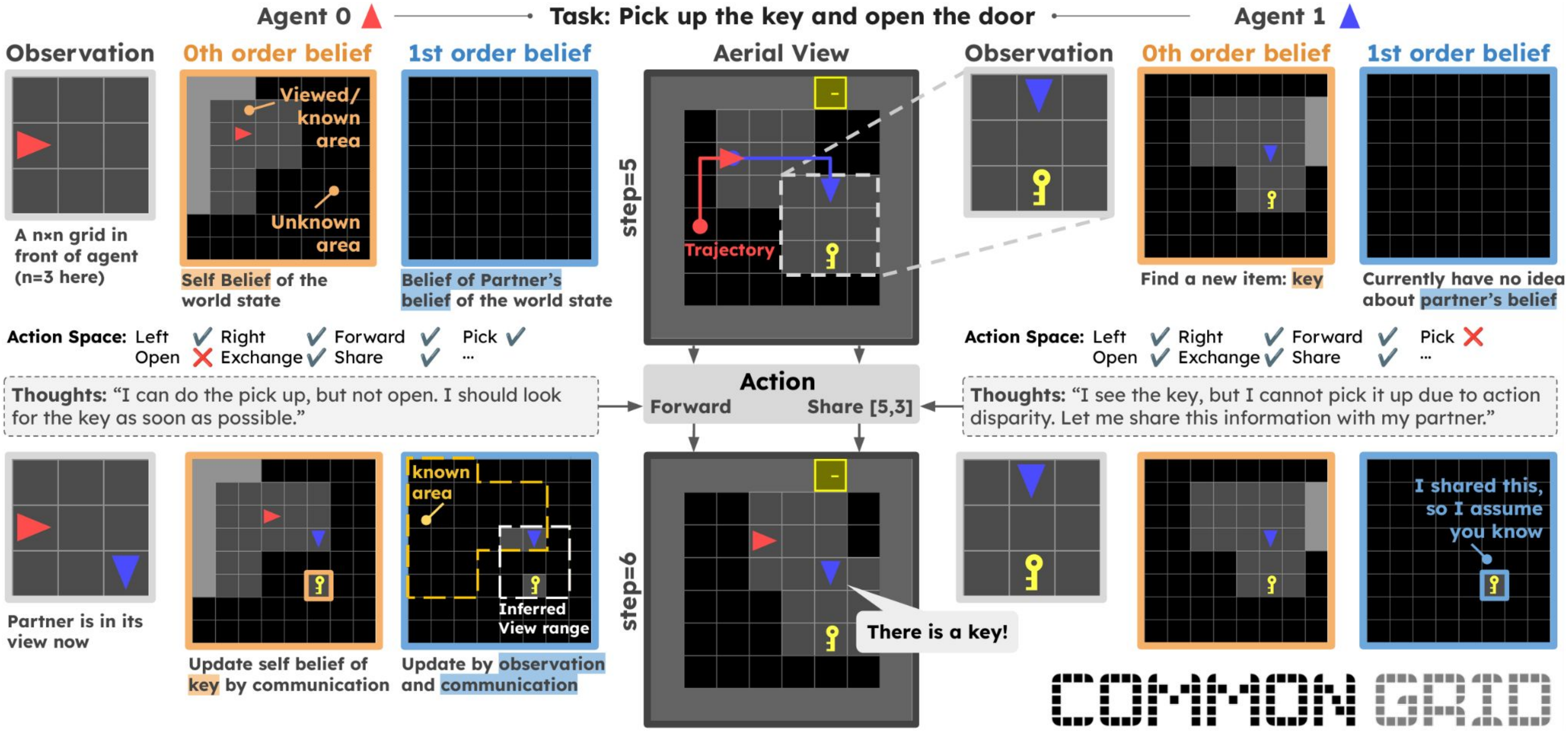


Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. *Ziqiao Ma, Jacob Sansom, Run Peng, Joyce Chai.* EMNLP Findings, 2023.

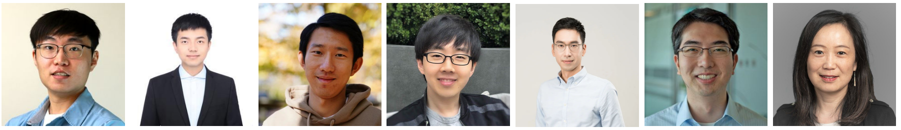
The CommonGrid Project



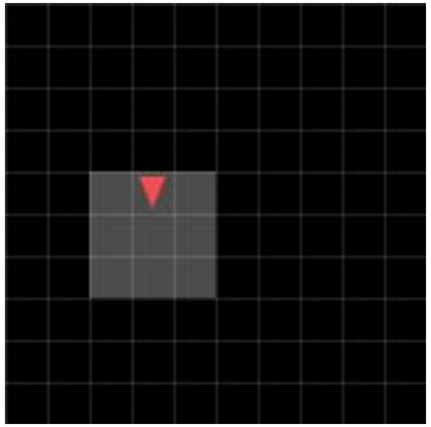
Investigate ToM modeling in collaboration in a 2D grid world.



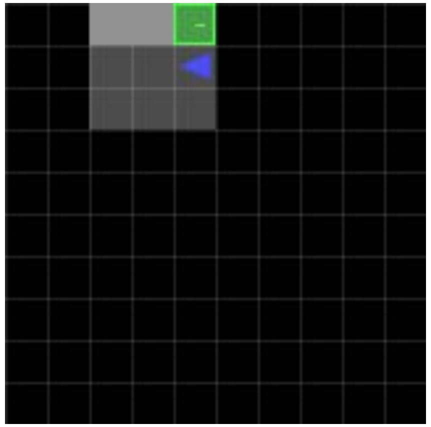
The CommonGrid Project



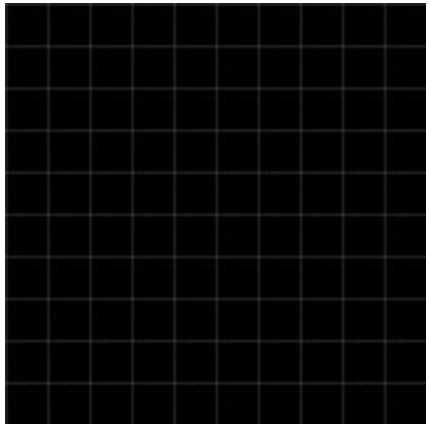
Investigate ToM modeling in collaboration in a 2D grid world.



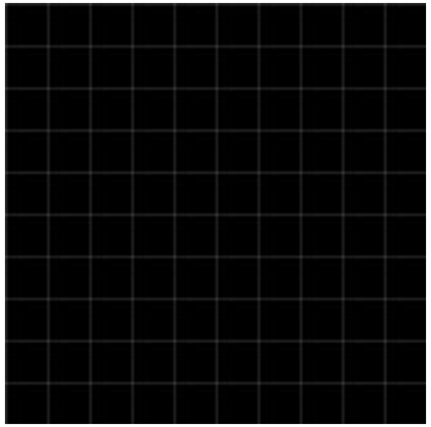
0th order belief of



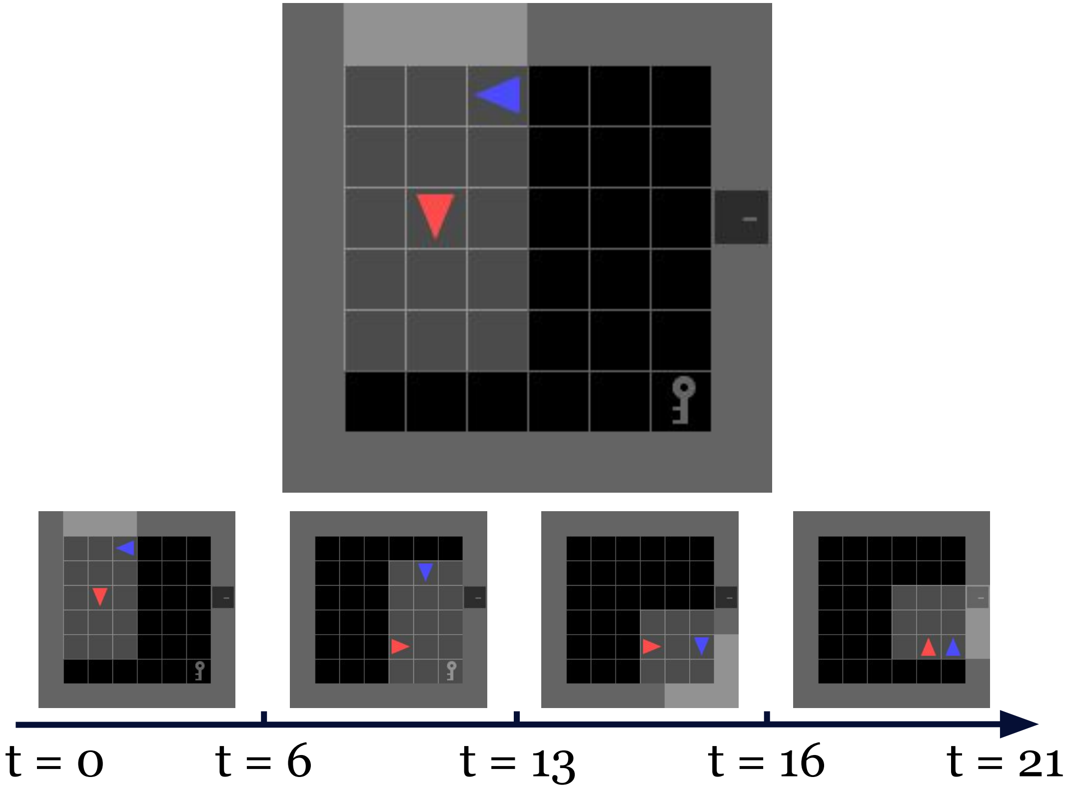
0th order belief of



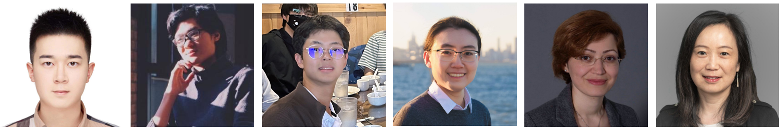
1st order belief of



1st order belief of

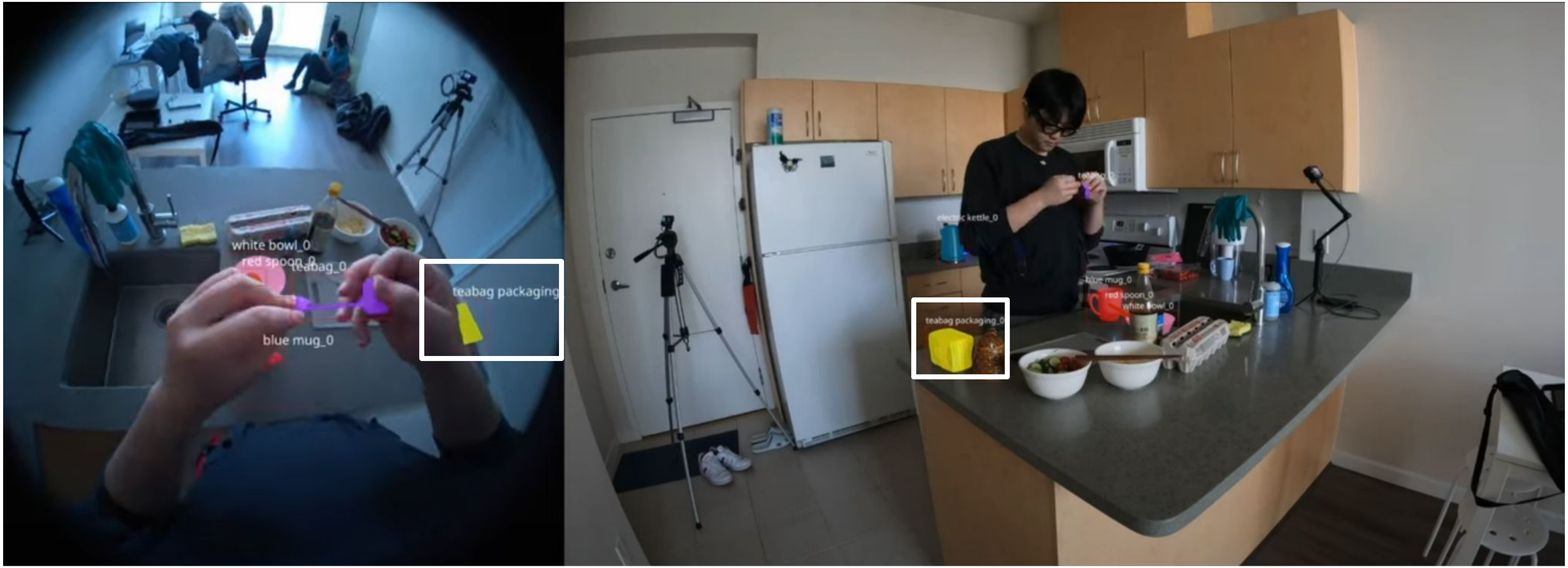


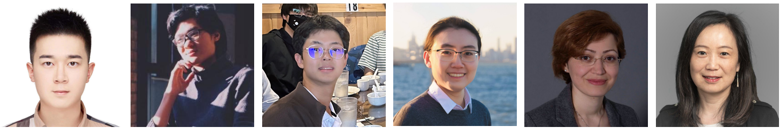
Situated Machine ToM



The curious case of perceptual perspective-taking in spatial reasoning.

- How would you describe the "tea bag package"?

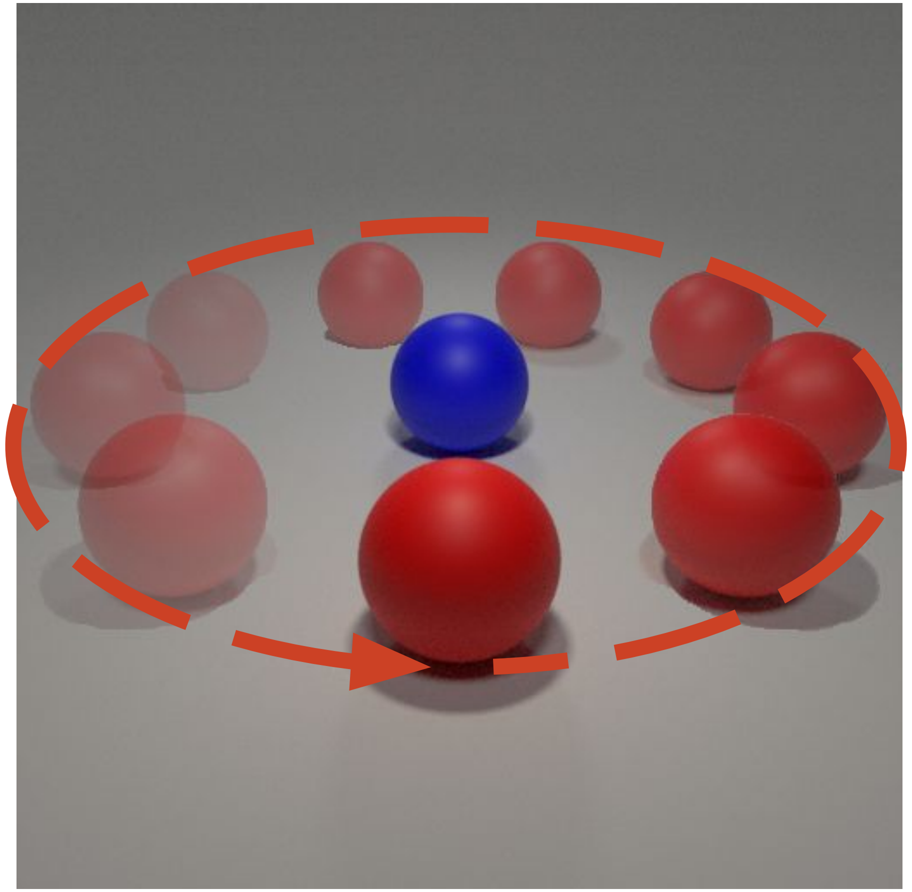


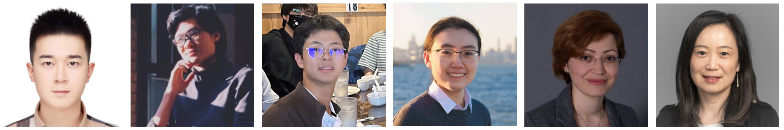


Spatial Cognition

The physical world is continuous.

- Is the red ball to the right of the blue ball?

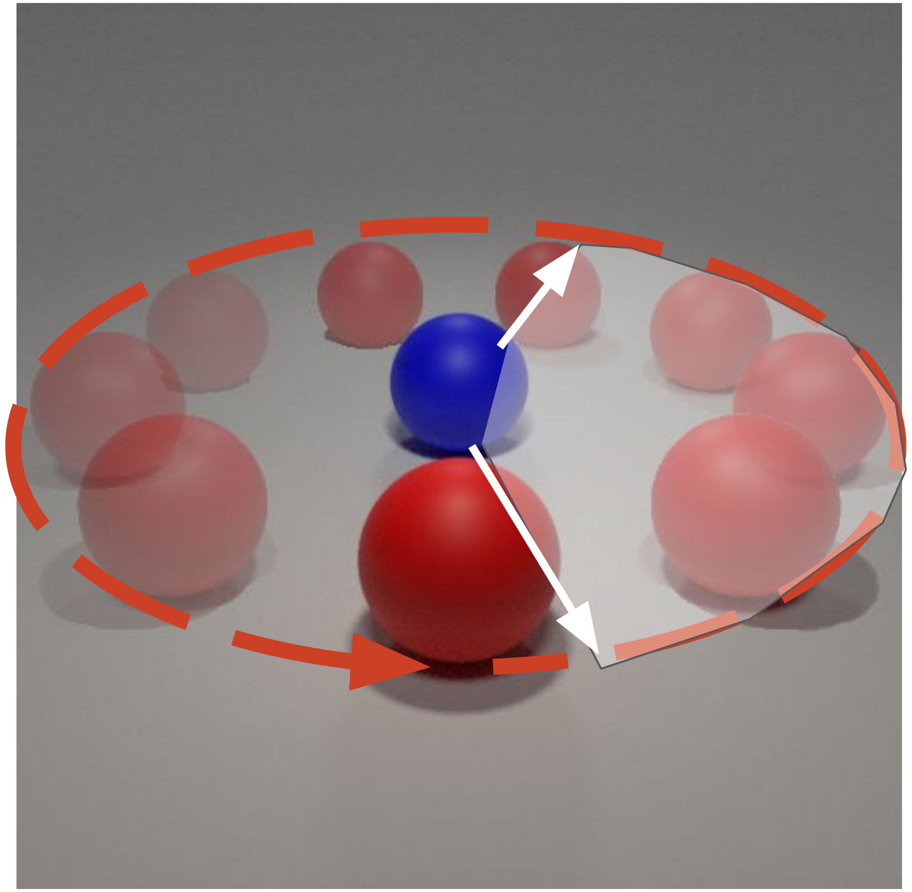




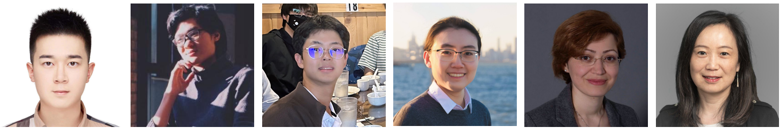
Spatial Cognition

The physical world is continuous -> region of acceptation.

- Is the red ball to the <u>right</u> of the blue ball?



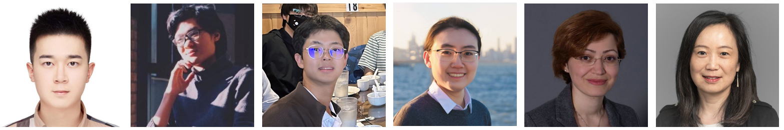
Carlson-Radvansky, L. A., & Logan, G. D. (1997). The influence of reference frame selection on spatial template construction. Journal of memory and language, 37(3), 411-437.



Spatial Cognition

Spatial frame of reference.

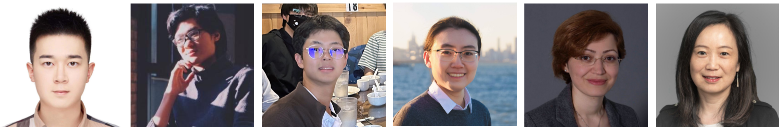
- Is the basketball to the right of the car?



Spatial Cognition

Spatial frame of reference.

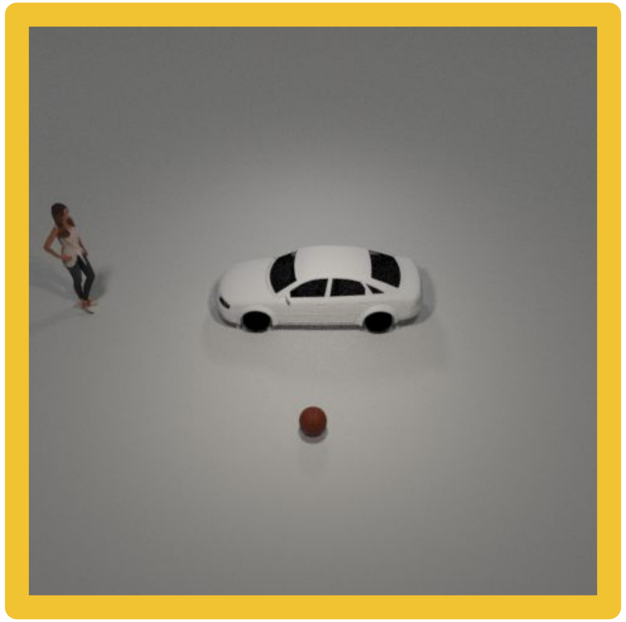
- Is the basketball to the right of the car?
    - **Yes, from the camera's viewpoint**



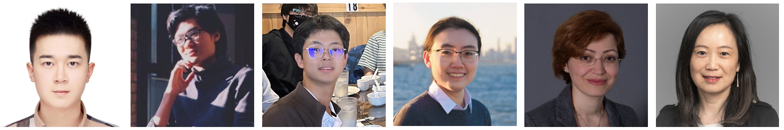
Spatial Cognition

Spatial frame of reference.

- Is the basketball to the right of the car?
  - **Yes, from the woman's viewpoint**

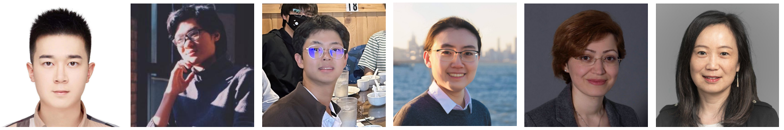


Spatial Cognition



Spatial frame of reference.

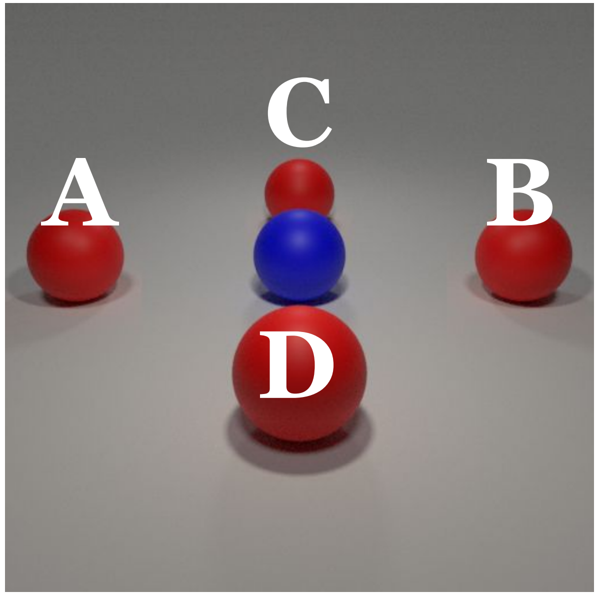
- Is the basketball to the <u>right</u> of the car?
  - **Yes, from the car's viewpoint**

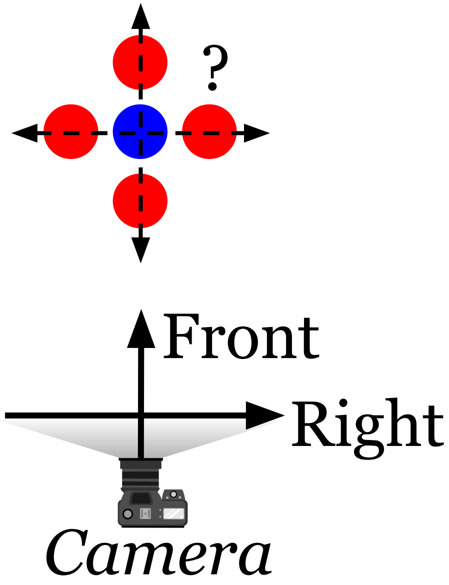


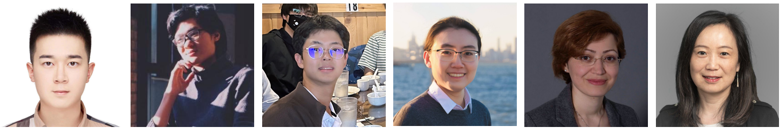
Spatial Cognition

Coordinate transformation in relative frame of reference.

- The ball to the left/right/front/back of the blue ball.



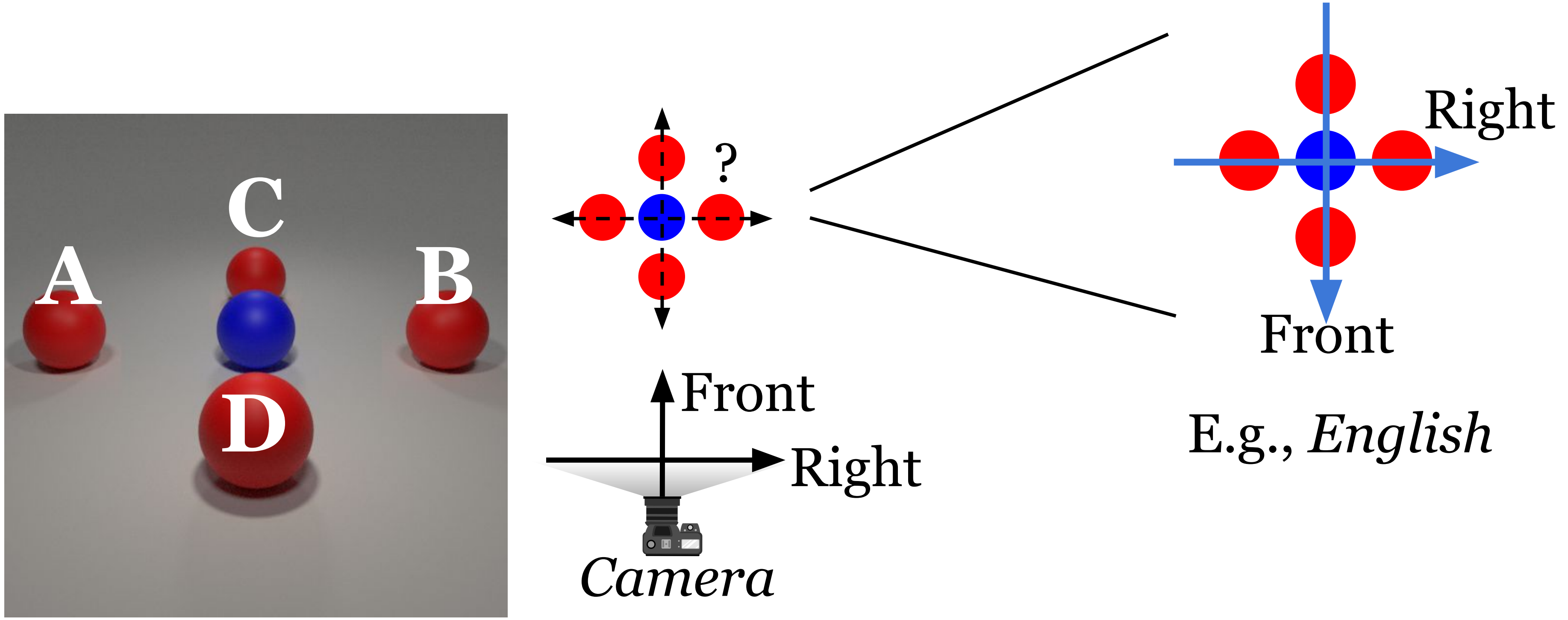




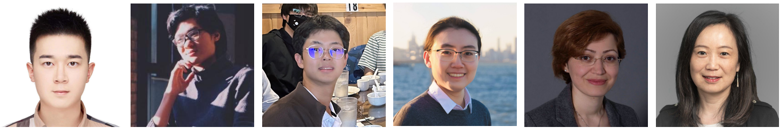
Spatial Cognition

Coordinate transformation in relative frame of reference.

- The ball to the <u>left/right/front/back</u> of the blue ball.
  - **Reflected: A/B/D/C**
  - **Example: English**



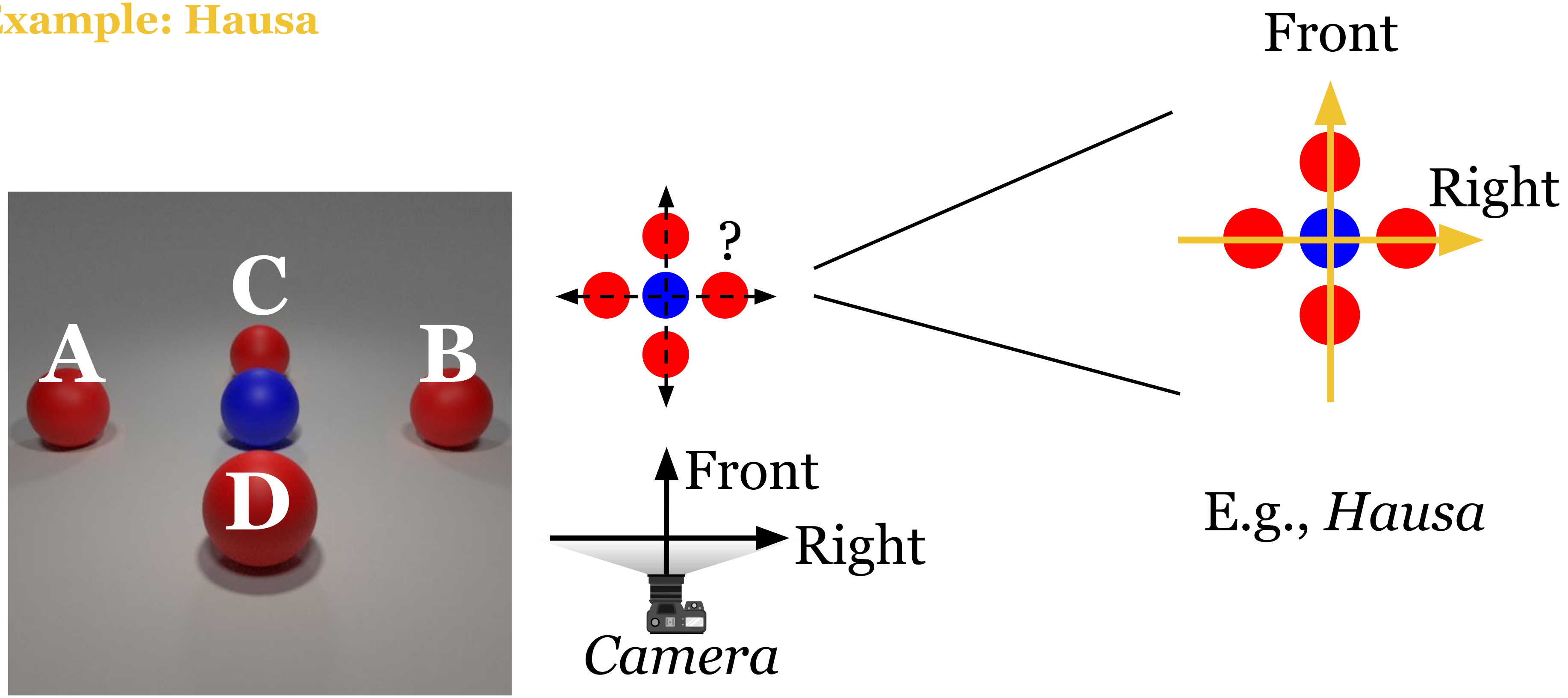
Levinson, S. C. (2003). Space in language and cognition: Explorations in cognitive diversity (Vol. 5). Cambridge University Press.



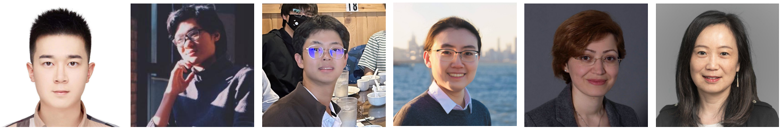
Spatial Cognition

Coordinate transformation in relative frame of reference.

- The ball to the left/right/front/back of the blue ball.
  - **Translated: A/B/C/D**
  - **Example: Hausa**



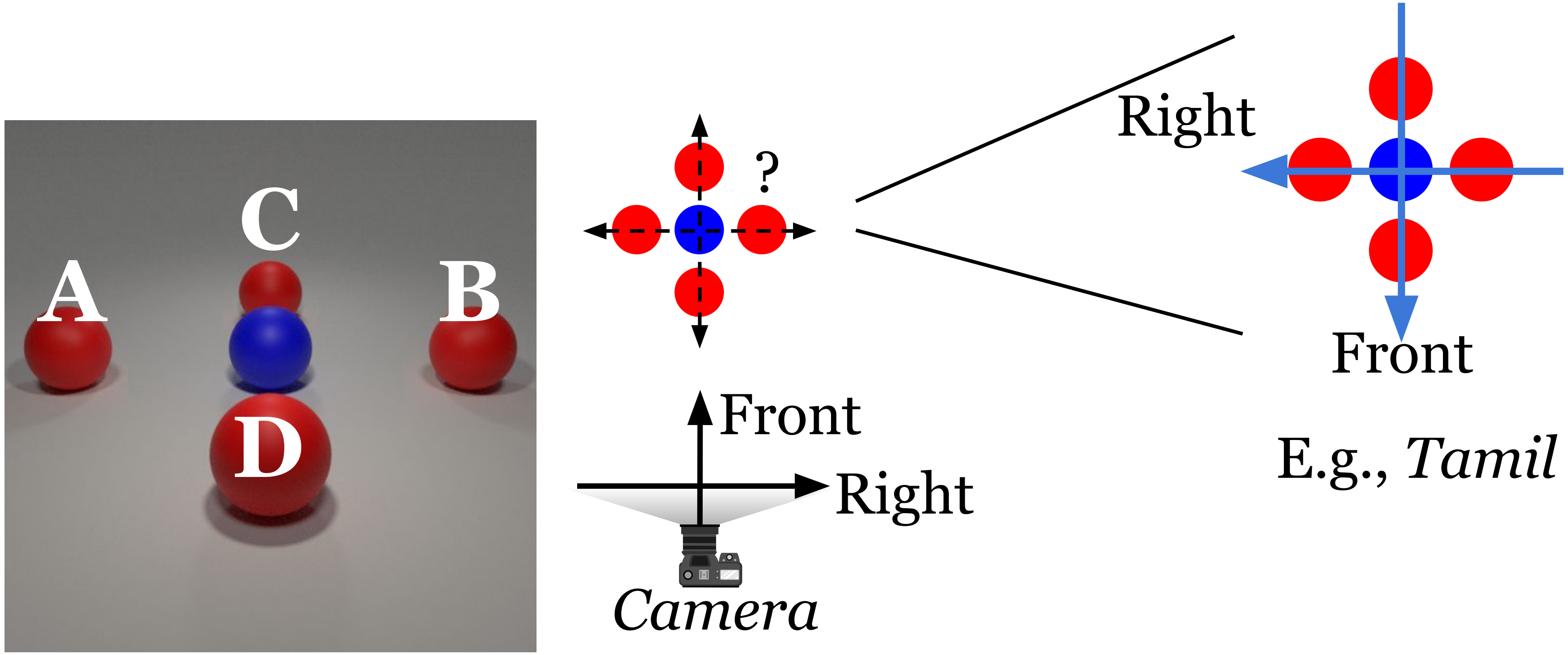
Levinson, S. C. (2003). Space in language and cognition: Explorations in cognitive diversity (Vol. 5). Cambridge University Press.



Spatial Cognition

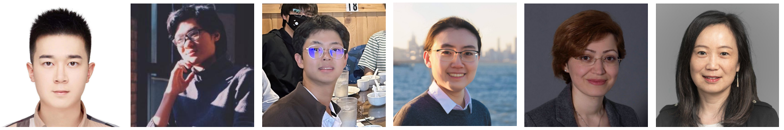
Coordinate transformation in relative frame of reference.

- The ball to the left/right/front/back of the blue ball.
  - **Rotated: B/A/D/C**
  - **Example: Tamil**



Levinson, S. C. (2003). Space in language and cognition: Explorations in cognitive diversity (Vol. 5). Cambridge University Press.

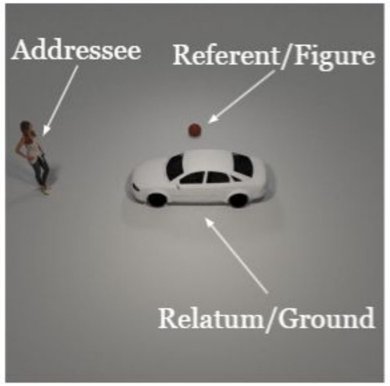
Spatial Cognition



Evaluating VLMs with FoR ambiguities.

- We study FoRs that lead to ambiguities in situated communication (Liu et al., 2010).

Origin	Frame of Reference	Example (English)
Camera (Preferred)	Egocentric Relative FoR	(From the camera's viewpoint,) the ball is **behind** the car.
Addressee	Addressee-Centered Relative FoR	(From the woman's viewpoint,) the ball is to the **left** of the car.
Reference	Object-Centered Intrinsic FoR	(From the car's viewpoint,) the ball is to the **right** of the car.



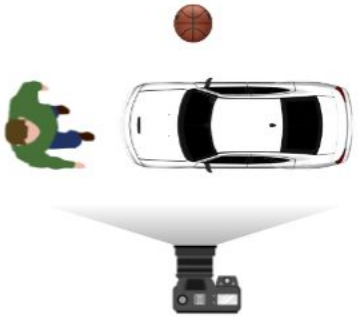
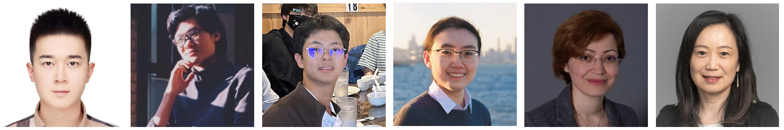


Figure 2: An illustrative example of how a frame of reference (FoR) specifies the reference system when describing the spatial relation between a target object (i.e., the ball) and a reference object (i.e., the car). When the FoR is not explicitly specified, English prefers an egocentric relative FoR, i.e., "the ball is behind the car." We study FoRs that lead to ambiguity (Liu et al., 2010).

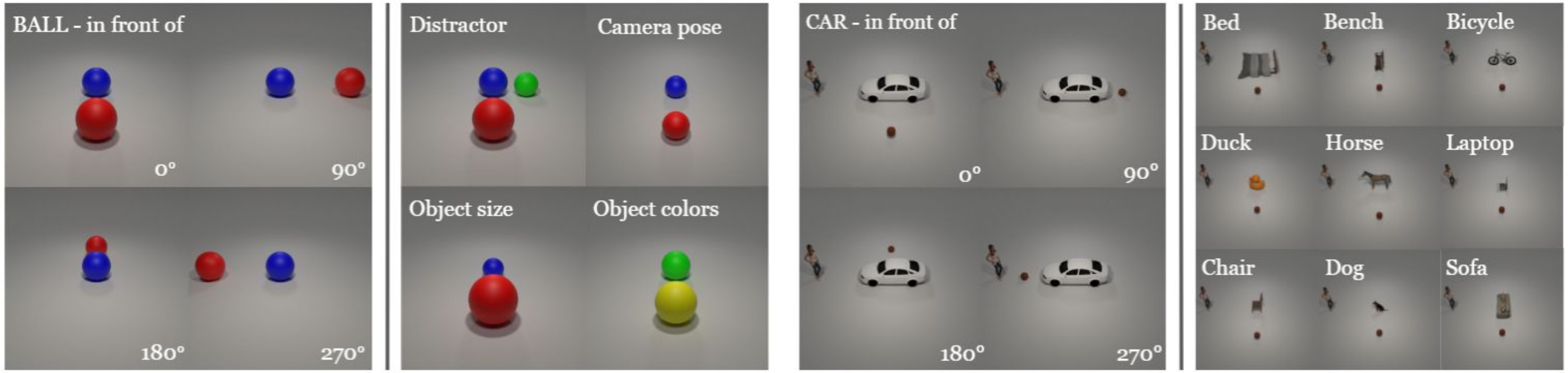
Liu, C., Walker, J., & Chai, J. Y. (2010, November). Ambiguities in spatial language understanding in situated human robot dialogue. In 2010 AAAI Fall Symposium Series.



Spatial Cognition

COnsistent Multilingual Frame Of Reference Test (COMFORT).

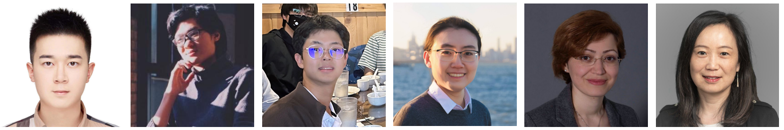
- COMFORT-CAR: When the relatum is fronted, as examples in Figure 1a, multiple FoRs are possible to interpret the reference system.
- COMFORT-BALL: When the relatum is non-fronted, as examples in Figure 1b, we focus on the ambiguity of conventions to determine its coordinate transformation for egocentric relative FoR.



(a) Sample images from COMFORT-BALL dataset. The 4 images on the left are selected every 90° interval along the rotational path out of 36 images. The 4 images on the right illustrate variations with a distractor, different object colors, sizes, or camera poses.

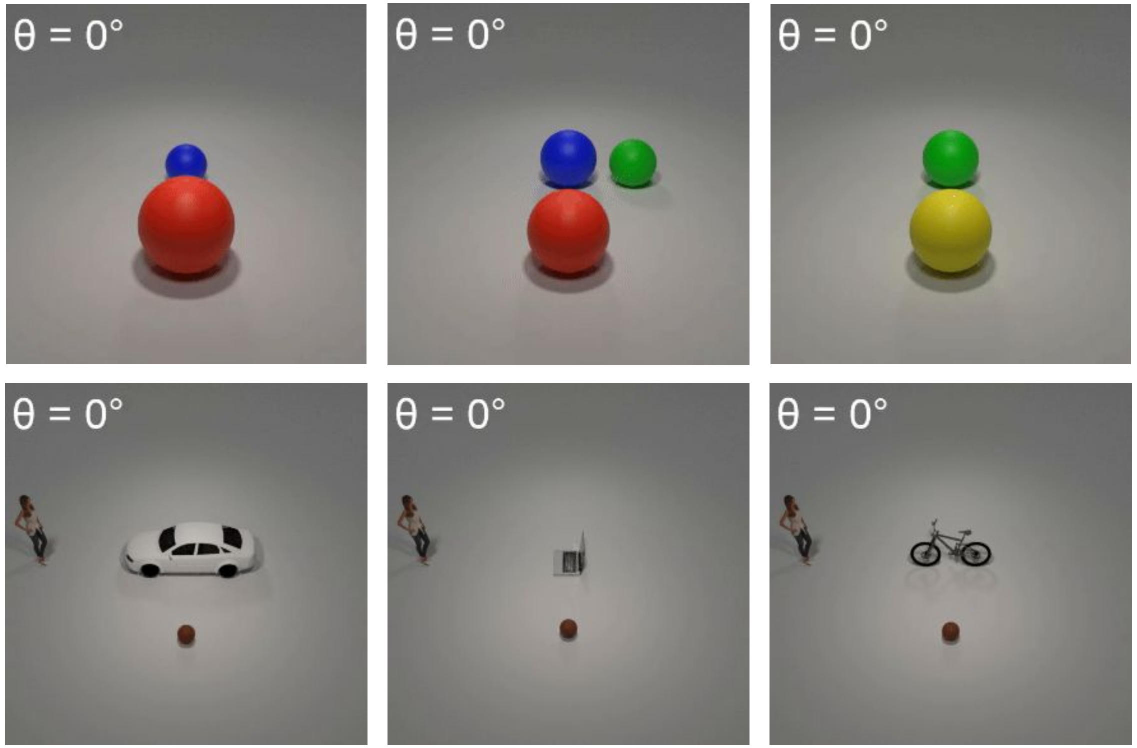
(b) Sample images from COMFORT-CAR dataset. The 4 images on the left are selected every 90° interval along the rotational path out of 36 images. The 9 images on the right are sample images of each variation with different relatum objects.

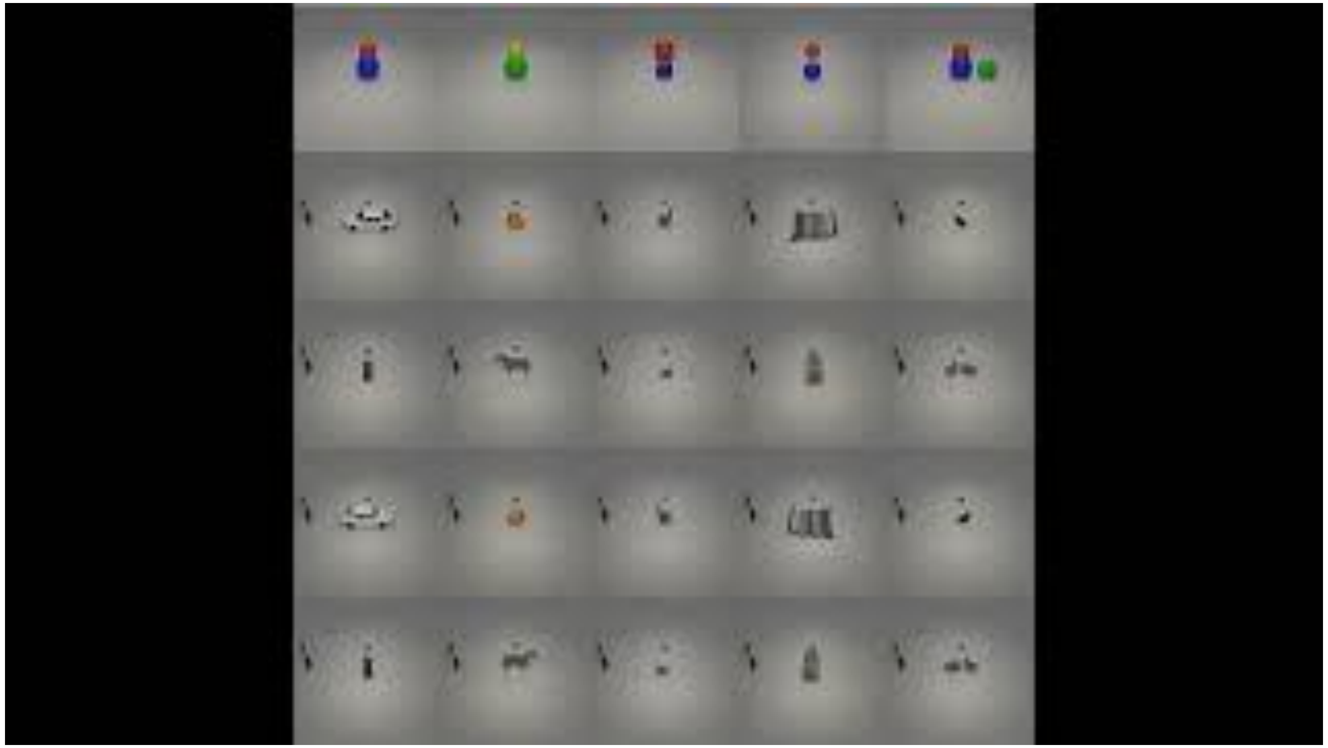
Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference Under Ambiguities. Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, Ziqiao Ma. Pluralistic Alignment @ NeurIPS 2024



Spatial Cognition

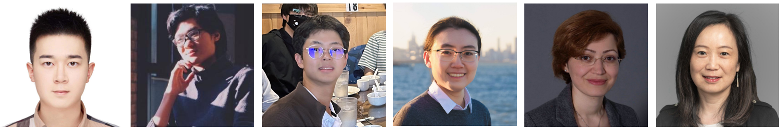
COnsistent Multilingual Frame Of Reference Test (COMFORT).





Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference Under Ambiguities. Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, Ziqiao Ma. Pluralistic Alignment @ NeurIPS 2024

Spatial Cognition



COnsistent Multilingual Frame Of Reference Test (COMFORT).

- **Accuracy**: We define the local probability of the model responding Yes by $p_i = P_i(\text{Yes})/[P_i(\text{Yes}) + P_i(\text{No})]$ We consider the inference correct if (1) the scene falls into the acceptability region and pi > 0.5 or (2) the scene falls out of the acceptability region and $p_i < 0.5$.
- **Region Parsing Error**: We normalize p_i across all image-prompt pairs, and compute the RMSE against the reference probability threshold (defined by hemispheres and cosine of angles) that represents the actual regions of acceptability.

Origin	Prompt Template
nop	Is [A] [relation] [B]?
cam	From the camera's viewpoint, is [A] [relation] [B]?
add	From the [addressee]'s viewpoint, is [A] [relation] [B]?
rel	From the [relatum]'s viewpoint, is [A] [relation] [B]?

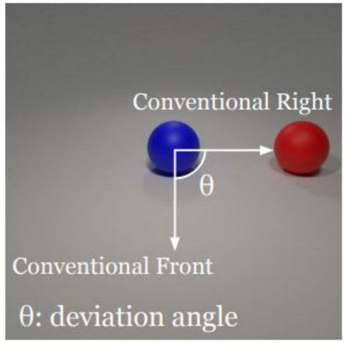


Figure 4: A red ball with a deviation angle $\theta = 90°$ relative to the conventional front (English) of the blue ball.

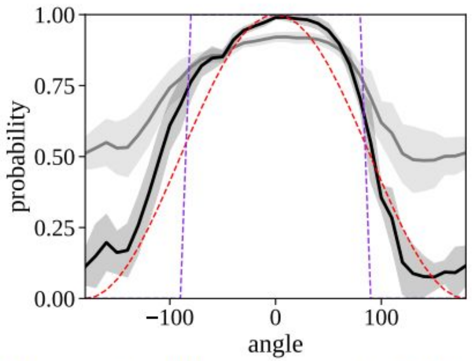
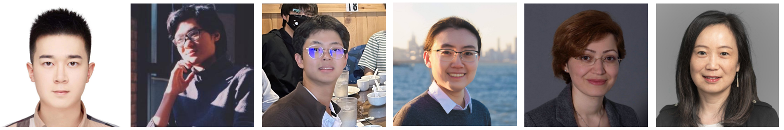


Figure 5: The raw probability $p(\theta)$ in gray, normalized probability $\widehat{p}(\theta)$ in black, and two reference probability $\lambda^{\text{hemi}}(\theta)$ and $\lambda^{\cos}(\theta)$ in purple and red.

Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference Under Ambiguities. Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, Ziqiao Ma. Pluralistic Alignment @ NeurIPS 2024

Spatial Cognition



Baselines.

- VLMs build from supervised instruction fine-tuning:
  - InstructBLIP (7B/13B) (Dai et al., 2023)
  - LLaVA v1.5 (7B/13B) (Liu et al., 2023b)
  - InternLM-XComposer2 (7B) (Dong et al., 2024)

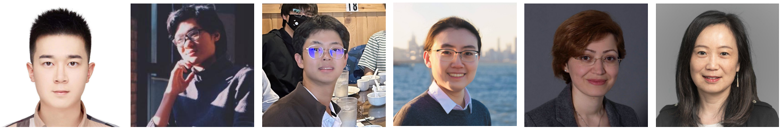
- VLMs with both supervised fine-tuning and reinforcement learning alignment:
  - MiniCPM-Llama3- V v2.5 (8B) (Hu et al., 2024; Yu et al., 2024b)

- Mechanistically grounded VLMs:
  - GLaMM (7B) (Rasheed et al., 2024)

- Multilingual VLMs2: .
  - mBLIP-BLOOMZ-7B (Geigle et al., 2024)
  - GPT-4o (OpenAI, 2024)

Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference Under Ambiguities. Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, Ziqiao Ma. Pluralistic Alignment @ NeurIPS 2024

Spatial Cognition

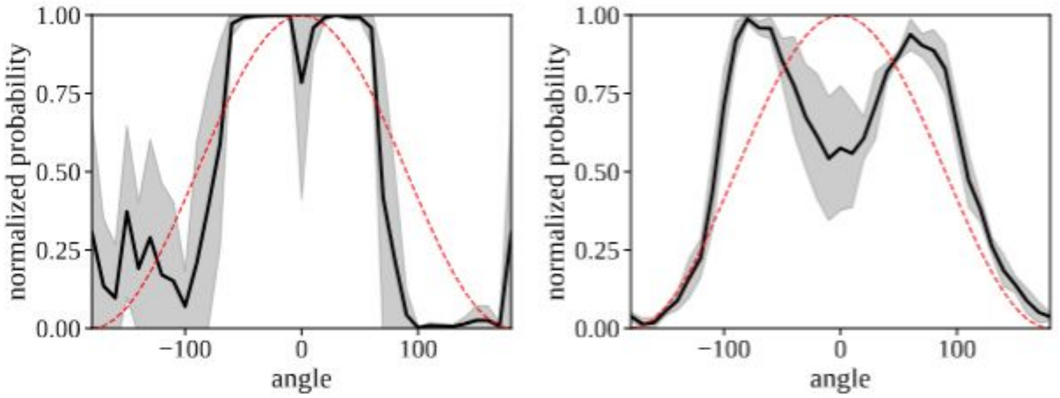


Most VLMs Prefer Reflected Coordinate Transformation Convention.

	Back						Front					
	Same			Reversed			Same			Reversed		
	Acc%	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	Acc%	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	Acc%	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	Acc%	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$
InstructBLIP-7B	47.2	58.4	45.6	48.3	53.8	39.0	67.2	47.5	31.6	27.2	64.6	52.0
InstructBLIP-13B	48.9	55.9	40.9	50.0	56.6	45.5	40.0	60.0	46.0	54.4	53.0	37.4
mBLIP-BLOOMZ	55.0	60.2	51.2	48.3	64.8	53.7	54.4	61.4	51.2	50.0	58.0	47.9
LLaVA-1.5-7B	28.3	66.7	54.0	68.3	47.0	32.9	19.4	71.0	59.1	82.8	36.4	24.8
LLaVA-1.5-13B	17.8	73.8	61.8	78.9	36.3	19.2	26.1	67.3	56.0	78.3	39.1	27.7
GLaMM	30.0	71.1	58.3	64.4	46.3	33.3	50.0	55.4	43.9	50.0	55.9	42.9
XComposer2	12.8	84.5	73.2	90.6	26.3	17.9	15.0	85.8	74.5	85.0	31.6	20.7
MiniCPM-V	13.3	83.6	71.6	86.7	29.3	17.8	10.6	85.5	73.6	90.6	26.2	16.6
GPT-4o	16.1	87.3	75.7	88.3	30.3	28.2	25.6	82.4	73.6	80.0	40.2	32.0

	Left						Right					
	Same			Reversed			Same			Reversed		
	Acc%	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	Acc%	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	Acc%	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	Acc%	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$
InstructBLIP-7B	54.4	51.5	37.2	41.1	61.6	48.0	39.4	61.4	47.5	55.0	52.0	37.8
InstructBLIP-13B	51.7	54.2	43.4	51.7	57.0	44.9	46.7	58.1	45.6	56.7	52.5	41.6
mBLIP-BLOOMZ	52.8	59.8	52.4	49.4	64.2	53.5	43.9	65.7	54.6	56.1	56.4	46.8
LLaVA-1.5-7B	91.7	25.3	11.9	3.9	83.4	70.0	90.6	26.0	13.0	9.4	80.9	68.5
LLaVA-1.5-13B	71.7	39.1	31.7	25.0	76.8	61.8	81.1	35.8	24.3	13.3	79.3	64.3
GLaMM	66.1	48.9	38.3	32.8	65.5	51.8	88.3	29.8	17.3	12.8	76.2	63.7
XComposer2	97.8	11.3	20.1	3.3	95.6	80.9	96.7	15.2	21.3	3.3	95.8	81.1
MiniCPM-V	94.4	17.6	15.5	4.4	91.8	77.9	89.4	26.5	17.5	5.0	88.3	74.1
GPT-4o	94.4	20.4	24.3	11.1	92.6	80.8	94.4	19.0	25.1	11.1	92.8	80.8

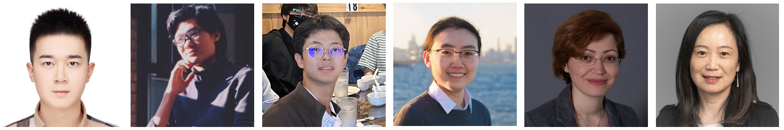
	Aggregated									
	Translated			Rotated			Reflected			Preferred Transform
	Acc%	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	Acc%	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	Acc%	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	
InstructBLIP-7B	52.1	54.7	40.5	42.9	58.0	44.2	42.4	57.8	43.9	Translated
InstructBLIP-13B	46.8	57.1	44.0	53.2	54.8	42.3	50.7	55.5	43.0	Not Significant
mBLIP-BLOOMZ	51.5	61.8	52.3	51.0	60.9	50.5	48.8	62.1	52.1	Not Significant
LLaVA-1.5-7B	57.5	47.3	34.5	41.1	61.9	49.0	83.3	33.7	20.7	Reflected
LLaVA-1.5-13B	49.2	54.0	43.4	48.9	57.9	43.2	77.5	37.6	25.7	Reflected
GLaMM	58.6	51.3	39.5	40.0	61.0	47.9	67.2	45.2	33.0	Reflected
XComposer2	55.6	49.2	47.3	45.6	62.3	50.1	92.5	21.1	20.0	Reflected
MiniCPM-V	51.9	53.3	44.5	46.7	58.9	46.6	90.3	24.9	16.8	Reflected
GPT-4o	57.6	52.3	49.7	47.6	64.0	55.5	89.3	27.5	27.4	Reflected



(a) Behind in GPT-4o. (b) Right in LLaVA-13B.

Figure 7: At $\theta = 0$, some models show sensitivity to multiple conventions.

Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference Under Ambiguities. Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, Ziqiao Ma. Pluralistic Alignment @ NeurIPS 2024



Spatial Cognition

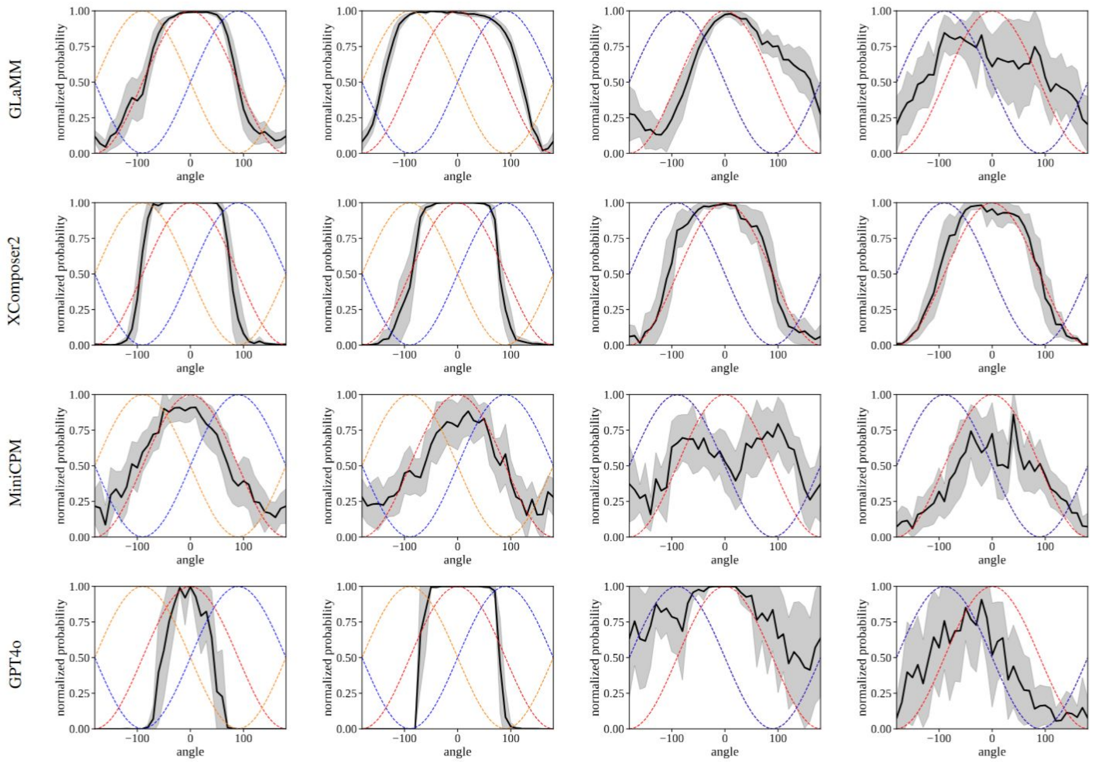
Most VLMs Prefer Egocentric Relative Frame of Reference.

	Back									Front								
	Egocentric			Intrinsic			Addressee			Egocentric			Intrinsic			Addressee		
	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$
InstructBLIP-7B	47.2	51.4	41.0	47.2	53.0	38.6	47.2	53.0	38.6	47.2	54.2	40.9	47.2	60.7	46.9	47.2	60.7	46.9
InstructBLIP-13B	47.2	43.5	32.9	47.2	48.9	34.4	47.2	48.9	34.4	47.2	66.5	52.5	47.2	61.1	48.5	47.2	61.1	48.5
mBLIP-BLOOMZ	52.8	62.1	52.2	52.8	63.9	53.2	52.8	63.9	53.2	52.8	56.4	45.3	52.8	55.5	44.6	52.8	55.5	44.6
LLaVA-1.5-7B	49.2	41.6	28.0	47.5	60.3	49.1	47.5	60.3	49.1	48.6	43.2	30.0	48.6	52.9	40.2	48.6	52.9	40.2
LLaVA-1.5-13B	50.8	36.8	20.9	48.6	54.7	43.0	48.6	54.7	43.0	47.2	46.5	34.5	47.2	47.3	32.6	47.2	47.3	32.6
GLaMM	47.2	45.6	31.9	47.2	51.0	38.8	47.2	51.0	38.8	47.2	37.9	24.8	47.2	69.6	57.1	47.2	69.6	57.1
XComposer2	91.4	25.0	12.7	53.6	59.9	49.3	53.6	59.9	49.3	87.8	26.6	15.2	55.0	59.3	48.3	55.0	59.3	48.3
MiniCPM-V	70.8	38.4	25.9	48.6	58.3	47.5	48.6	58.3	47.5	58.3	47.8	34.4	50.0	57.4	46.1	50.0	57.4	46.1
GPT-4o	64.2	49.1	38.3	66.4	45.4	36.7	66.4	45.4	36.7	58.1	54.8	43.1	53.6	61.0	50.2	53.6	61.0	50.2

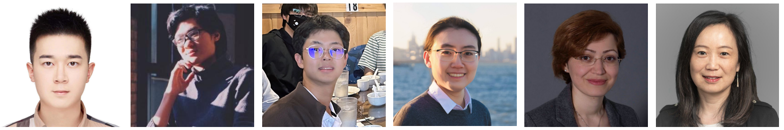
	Left									Right								
	Egocentric			Intrinsic			Addressee			Egocentric			Intrinsic			Addressee		
	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$
InstructBLIP-7B	47.2	59.0	45.6	47.2	45.3	32.5	47.2	62.0	51.9	47.2	53.1	39.6	47.2	61.7	51.2	47.2	45.3	31.8
InstructBLIP-13B	47.2	59.7	47.8	47.2	70.2	56.2	47.2	39.6	27.8	47.2	53.6	40.6	47.2	39.5	27.6	47.2	70.8	56.6
mBLIP-BLOOMZ	52.8	58.2	47.8	52.8	59.7	47.6	52.8	58.4	48.1	52.8	57.7	45.4	52.8	60.6	48.4	52.8	53.8	42.4
LLaVA-1.5-7B	76.7	25.6	14.0	33.9	68.2	56.8	64.4	52.7	41.5	56.4	28.5	13.7	44.2	64.6	53.0	52.5	57.3	46.6
LLaVA-1.5-13B	81.7	23.7	13.4	42.2	65.0	53.5	57.2	58.5	47.4	86.7	26.8	14.3	47.8	64.0	53.6	52.2	59.9	49.3
GLaMM	75.8	22.3	11.7	46.4	62.0	51.1	52.5	62.3	51.1	60.8	41.8	27.5	44.7	68.5	57.4	53.1	58.7	48.7
XComposer2	95.0	18.8	18.8	45.6	70.5	61.2	54.4	64.0	53.7	96.1	17.1	16.5	47.8	68.1	58.4	52.2	64.6	54.5
MiniCPM-V	75.6	32.9	18.2	43.3	62.3	50.4	55.6	53.6	41.3	73.6	35.2	20.4	48.1	55.1	43.1	49.7	58.5	46.3
GPT-4o	78.6	42.1	34.7	48.1	69.4	59.3	51.9	65.8	56.5	93.9	21.8	24.3	52.8	67.0	57.3	47.2	71.0	61.7

	Aggregated									
	Egocentric			Intrinsic			Addressee			Preferred FoR
	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	$Acc_{\%}$	$\varepsilon^{hemi}_{\times 10^2}$	$\varepsilon^{cos}_{\times 10^2}$	
InstructBLIP-7B	47.2	54.4	41.8	47.2	55.2	42.3	47.2	55.2	42.3	Not Significant
InstructBLIP-13B	47.2	55.8	43.5	47.2	54.9	41.7	47.2	55.1	41.8	Not Significant
mBLIP-BLOOMZ	52.8	58.6	47.7	52.8	59.9	48.4	52.8	57.9	47.1	Not Significant
LLaVA-1.5-7B	57.7	34.7	21.4	43.5	61.5	49.8	53.3	55.8	44.4	Egocentric Relative
LLaVA-1.5-13B	66.6	33.5	20.8	46.5	57.7	45.7	51.3	55.1	43.1	Egocentric Relative
GLaMM	57.8	36.9	24.0	46.4	62.8	51.1	50.0	60.4	48.9	Egocentric Relative
XComposer2	92.6	21.9	15.8	50.5	64.4	54.3	53.8	61.9	51.4	Egocentric Relative
MiniCPM-V	69.6	38.6	24.7	47.5	58.3	46.8	51.0	57.0	45.3	Egocentric Relative
GPT-4o	73.7	42.0	35.1	55.2	60.7	50.9	54.8	60.8	51.3	Egocentric Relative

Table 7: Preferred frame of reference in VLMs.



Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference Under Ambiguities. Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, Ziqiao Ma. Pluralistic Alignment @ NeurIPS 2024



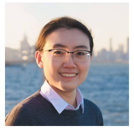
Spatial Cognition

VLMs Fail to Flexibly Adopt Alternative Frames of Reference.

Model	Egocentric		Intrinsic		Addressee		Aggregated	
	Acc% (↑)	$\varepsilon^{\cos}_{\times 10^2}$ (↓)	Acc% (↑)	$\varepsilon^{\cos}_{\times 10^2}$ (↓)	Acc% (↑)	$\varepsilon^{\cos}_{\times 10^2}$ (↓)	Acc% (↑)	$\varepsilon^{\cos}_{\times 10^2}$ (↓)
InstructBLIP-7B	$47.2_{(+0.0)}$	$42.6_{(+0.9)}$	$47.2_{(+0.0)}$	$43.0_{(+0.6)}$	$47.2_{(+0.0)}$	$42.5_{(+0.2)}$	$47.2_{(+0.0)}$	$42.7_{(+0.5)}$
InstructBLIP-13B	$47.2_{(+0.0)}$	$43.7_{(+0.2)}$	$47.2_{(+0.0)}$	$42.8_{(+1.2)}$	$47.2_{(+0.0)}$	$43.1_{(+1.3)}$	$47.2_{(+0.0)}$	$43.2_{(+0.9)}$
mBLIP-BLOOMZ	$52.0_{(-0.8)}$	$55.8_{(+8.2)}$	$49.5_{(-3.3)}$	$54.2_{(+5.7)}$	$49.4_{(-3.4)}$	$56.6_{(+9.5)}$	$50.3_{(-2.5)}$	$55.5_{(+7.8)}$
LLaVA-1.5-7B	$55.1_{(-2.7)}$	$\mathbf{18.3}_{(-3.2)}$	$46.3_{(+2.7)}$	$50.2_{(+0.4)}$	$47.9_{(-5.4)}$	$43.1_{(-1.3)}$	$49.7_{(-1.8)}$	$37.2_{(-1.4)}$
LLaVA-1.5-13B	$51.9_{(-14.8)}$	$23.7_{(+2.9)}$	$47.2_{(+0.8)}$	$\mathbf{43.1}_{(-2.6)}$	$47.5_{(-3.8)}$	$\mathbf{38.6}_{(-4.5)}$	$48.8_{(-6.0)}$	$\mathbf{35.1}_{(-1.4)}$
GLaMM	$47.2_{(-10.6)}$	$23.6_{(-0.4)}$	$47.2_{(+0.8)}$	$47.7_{(-3.4)}$	$47.2_{(-2.8)}$	$42.8_{(-6.2)}$	$47.2_{(-4.2)}$	$38.0_{(-3.3)}$
XComposer2	$\mathbf{85.1}_{(-7.5)}$	$18.9_{(+3.1)}$	$51.0_{(+0.5)}$	$51.3_{(-3.1)}$	$\mathbf{54.3}_{(+0.4)}$	$49.3_{(-2.2)}$	$\mathbf{63.4}_{(-2.2)}$	$39.8_{(-0.7)}$
MiniCPM-V	$61.8_{(-7.8)}$	$24.7_{(+0.0)}$	$50.1_{(+2.6)}$	$45.8_{(-0.9)}$	$50.4_{(-0.6)}$	$43.4_{(-1.9)}$	$54.1_{(-1.9)}$	$38.0_{(-1.0)}$
GPT-4o	$78.3_{(+4.6)}$	$28.3_{(-6.8)}$	$\mathbf{54.5}_{(-0.7)}$	$44.3_{(-6.6)}$	$49.4_{(-5.4)}$	$43.3_{(-8.0)}$	$60.7_{(-0.5)}$	$38.6_{(-7.1)}$

Table 4: The accuracy and cosine region parsing errors of VLMs when explicitly prompted to follow each frame of reference are provided (cam/rel/add). The values in parentheses indicate the performance change relative to the scenario with no perspective (nop) prompting.

Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference Under Ambiguities. Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, Ziqiao Ma. Pluralistic Alignment @ NeurIPS 2024

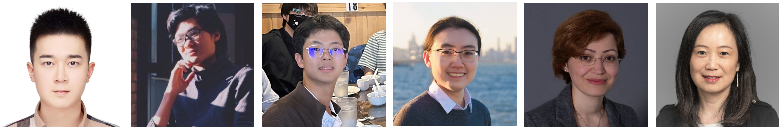
Spatial Cognition

Spatial Representations in VLMs Are Not Robust and Consistent.

Model	Obj F1 (↑)		Acc% (↑)		$\varepsilon^{\text{cos}}_{\times 10^2}$ (↓)		$\varepsilon^{\text{hemi}}_{\times 10^2}$ (↓)		$\sigma_{\times 10^2}$ (↓)		$\eta_{\times 10^2}$ (↓)		$c^{\text{sym}}_{\times 10^2}$ (↓)		$c^{\text{opp}}_{\times 10^2}$ (↓)	
	BALL	CAR	BALL	CAR	BALL	CAR	BALL	CAR	BALL	CAR	BALL	CAR	BALL	CAR	BALL	CAR
InstructBLIP-7B	66.7	66.7	47.2	47.2	43.9	42.6	57.8	55.5	16.6	20.8	17.2	13.3	26.7	27.3	48.4	48.5
InstructBLIP-13B	67.3	41.0	47.2	47.2	43.0	43.7	55.5	56.1	21.0	18.9	17.3	12.7	27.1	37.4	48.2	54.1
mBLIP-BLOOMZ	99.1	33.3	47.5	51.9	52.1	55.8	62.1	65.6	33.8	43.0	29.1	31.2	43.7	49.3	54.1	61.2
LLaVA-1.5-7B	100.0	88.3	63.2	55.1	20.7	**18.3**	33.7	32.5	8.3	**10.9**	**5.8**	**5.3**	25.2	20.0	23.5	**21.8**
LLaVA-1.5-13B	100.0	97.7	55.3	51.9	25.7	23.7	37.6	36.9	9.3	11.1	7.0	5.7	19.3	21.1	24.9	29.9
GLaMM	100.0	99.6	47.2	47.2	33.0	23.6	45.2	38.1	13.7	15.0	10.1	9.3	29.9	23.8	45.0	28.9
XComposer2	100.0	94.7	**92.4**	**85.1**	20.0	18.9	**21.1**	**26.7**	10.5	11.8	9.0	6.6	19.2	15.7	**13.7**	24.1
MiniCPM-V	99.3	66.7	89.3	61.8	**16.8**	24.7	24.9	38.2	**7.7**	16.3	6.6	11.8	23.4	21.7	17.3	23.3
GPT-4o	100.0	95.6	89.2	78.3	27.4	28.3	27.5	34.9	14.2	16.5	14.1	13.1	20.9	26.8	43.1	39.0
Random (30 trials)	50.0		50.9		46.3		58.7		28.3		26.6		42.5		44.2	
Always "Yes"	50.0		47.2		61.2		68.7		0.0		0.0		0.0		100.0	

Table 5: A comprehensive evaluation of VLMs in egocentric relative FoR with reflected transformation, using an explicit camera perspective (cam) prompt, is conducted. The metrics considered include object hallucination (F1-score), accuracy (Acc), region parsing error (ε), prediction noise (η), standard deviation (σ), and consistency (c).

Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference Under Ambiguities. Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, Ziqiao Ma. Pluralistic Alignment @ NeurIPS 2024

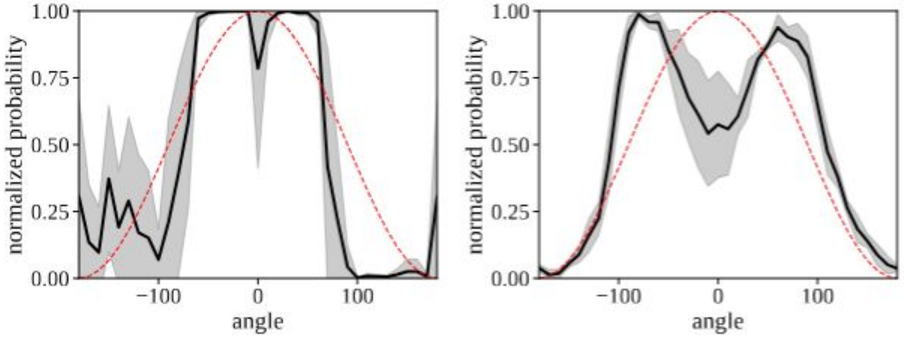


Spatial Cognition

Mind the gap between neural representations of vision, language, and space.

- Many VLMs show representation of space from vision-language training.
    - A clear preference for egocentric relative FoR with a reflected projection.
    - Identical to English conventions.
    - This spatial representation lacks robustness and consistency in continuous space.

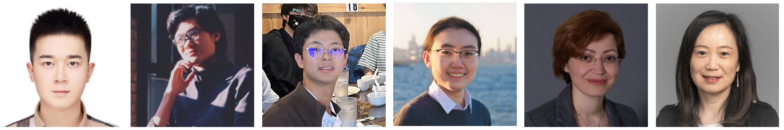
- VLMs can not perform spatial reasoning in alternative coordinate systems.
    - Intrinsic and addressee-centric relative FoRs are available systems in English.



(a) Behind in GPT-4o. (b) Right in LLaVA-13B.

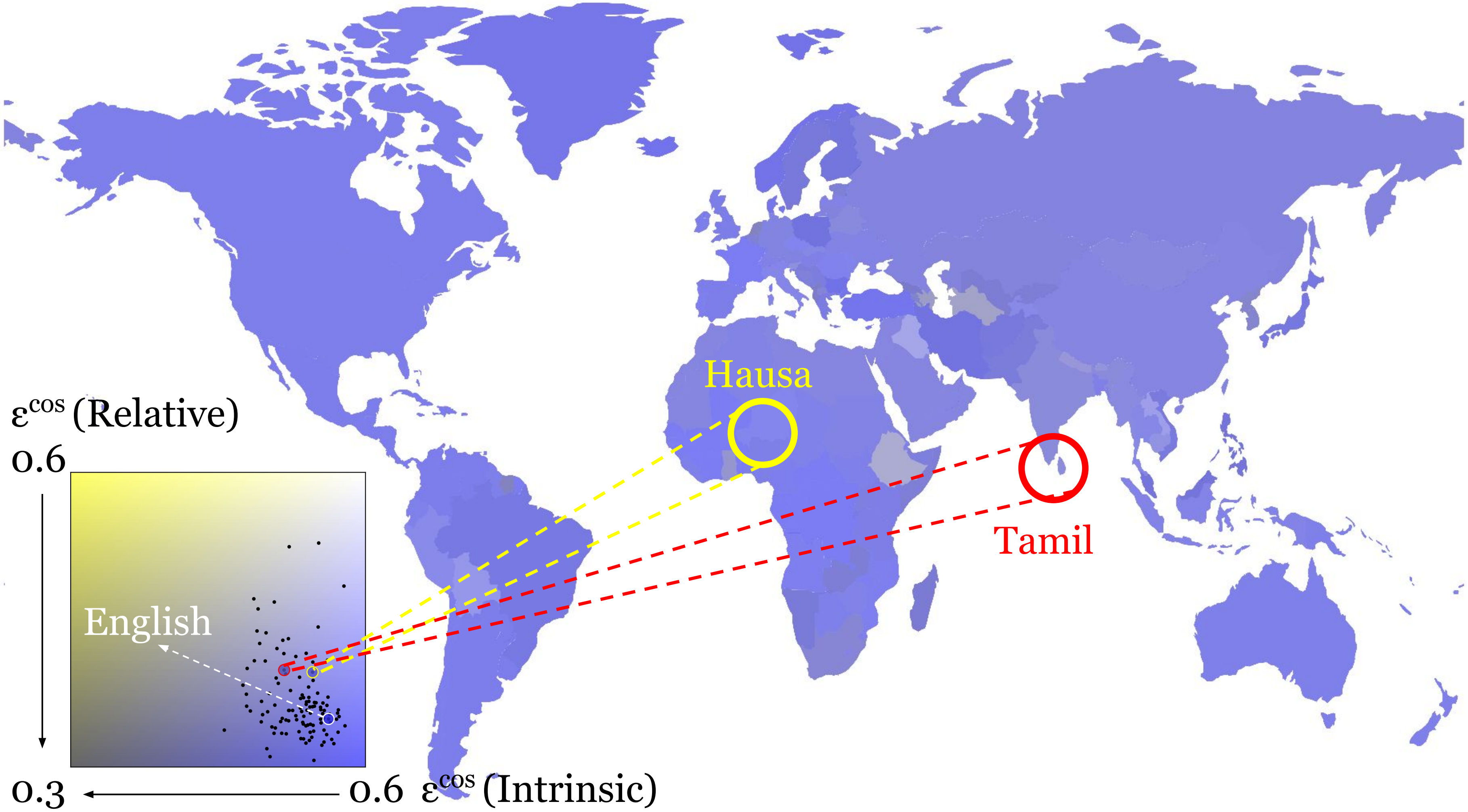
Figure 7: At $\theta = 0$, some models show sensitivity to multiple conventions.

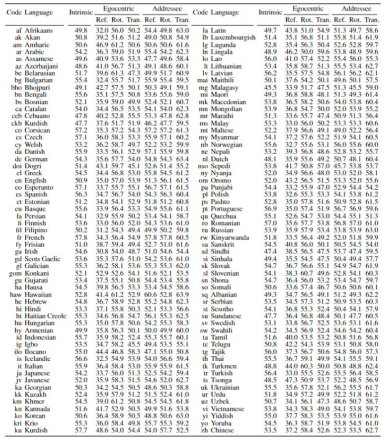
Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference Under Ambiguities. Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, Ziqiao Ma. Pluralistic Alignment @ NeurIPS 2024



Spatial Cognition

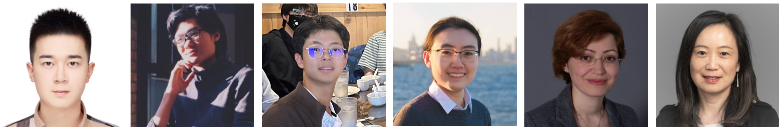
A Cross-lingual and Cross-cultural Evaluation of Frame of Reference.



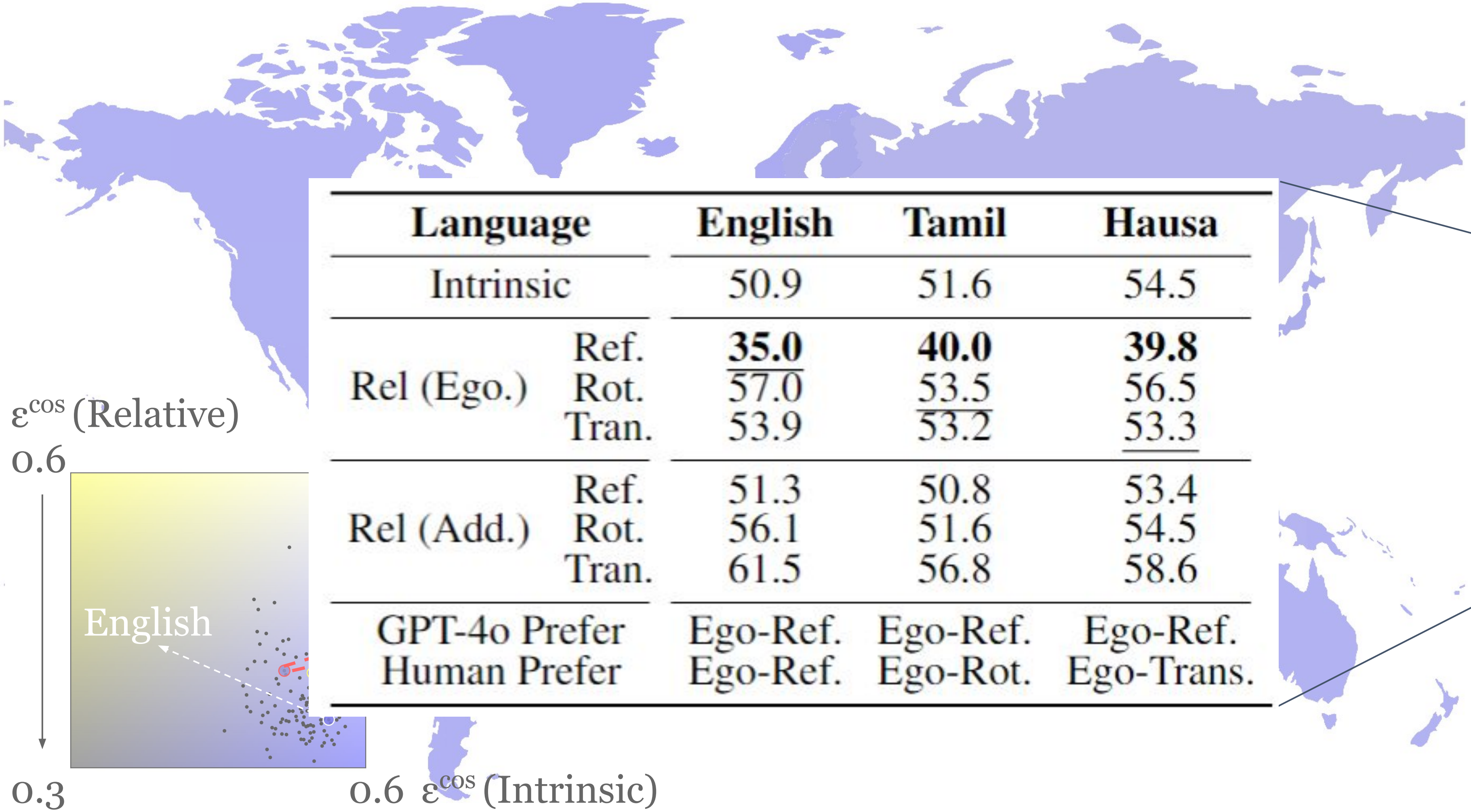


Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference Under Ambiguities. Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, Ziqiao Ma. Pluralistic Alignment @ NeurIPS 2024

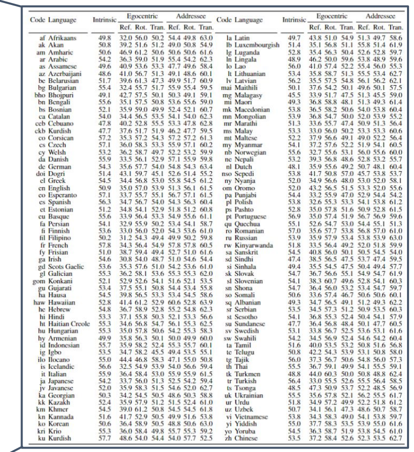
Spatial Cognition



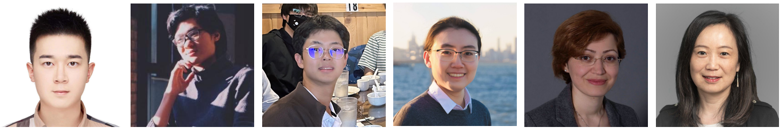
English overshadows the FoR conventions in other languages.



Language		English	Tamil	Hausa
Intrinsic		50.9	51.6	54.5
Rel (Ego.)	Ref.	**35.0**	**40.0**	**39.8**
	Rot.	57.0	53.5	56.5
	Tran.	53.9	53.2	53.3
Rel (Add.)	Ref.	51.3	50.8	53.4
	Rot.	56.1	51.6	54.5
	Tran.	61.5	56.8	58.6
GPT-4o Prefer		Ego-Ref.	Ego-Ref.	Ego-Ref.
Human Prefer		Ego-Ref.	Ego-Rot.	Ego-Trans.



Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference Under Ambiguities. Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, Ziqiao Ma. Pluralistic Alignment @ NeurIPS 2024



Spatial Cognition

English overshadows the FoR conventions in other languages.

- Multilingual VLMs fail to accommodate cross-cultural conventions.
  - Not surprising, current pipeline translate the English captions to other language and train.
  - The Linguistic Transmission Hypothesis (Bohnemeyer et al., 2014)

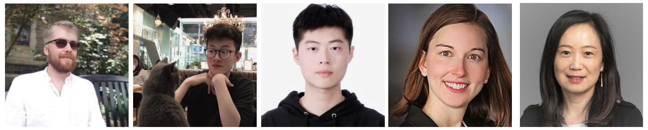
We propose the ***Linguistic Transmission Hypothesis (LTH)****: Using any language or linguistic variety - independently of its structures - may facilitate the acquisition of cultural practices of non linguistic cognition shared among the speakers of the language.*

Spatial frames of reference afford a particularly suitable test case for the lth, since they are not lexicalized or grammaticalized in language, but rather are themselves cognitive practices that underlie the interpretation of both linguistic and nonlinguistic spatial representations.

Direct support for the LTH comes from the impact of the familiarity with the use of Spanish as a second language we observed. The speakers of the indigenous languages in our sample used relative frames more frequently in their native language, [as] the more frequently they also used Spanish as a second language.

Bohnemeyer, J., Donelson, K., Tucker, R., Benedicto, E., Garza, A. C., Eggleston, A., ... & Méndez, R. R. (2014). The cultural transmission of spatial cognition: Evidence from a large-scale study. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 36, No. 36).

Embodied Dialogue Agents



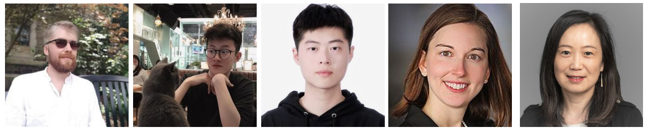
Asymmetric collaboration in a simulated world [EMNLP 2021, IJCAI 2023].

- MindCraft:

  Two agents are co-situated in a shared environment with a joint goal to create a block.

MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks. *Cristian-Paul Bara, Sky CH-Wang, Joyce Chai.* EMNLP, 2021.
Towards Collaborative Plan Acquisition through Theory of Mind Modeling in Situated Dialogue. Cristian-Paul Bara, *Ziqiao Ma, Yingzhuo Yu, Julie Shah, Joyce Chai.* IJCAI, 2023.

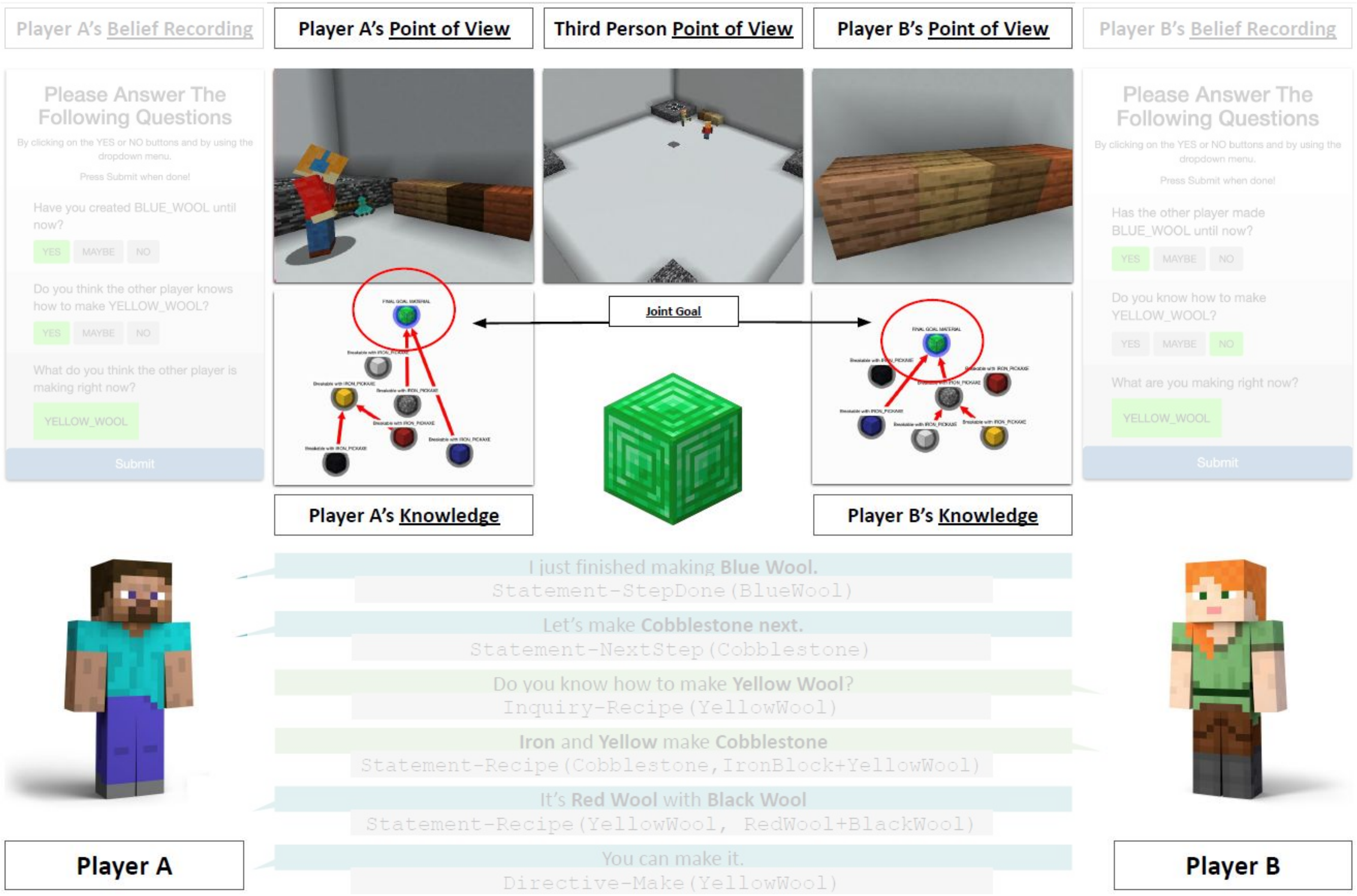
Embodied Dialogue Agents



Asymmetric collaboration in a simulated world [EMNLP 2021, IJCAI 2023].

- MindCraft:

  Players are given a partial plan in the form of a directed AND-graph.



MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks. *Cristian-Paul Bara, Sky CH-Wang, Joyce Chai.* EMNLP, 2021.
Towards Collaborative Plan Acquisition through Theory of Mind Modeling in Situated Dialogue. *Cristian-Paul Bara, Ziqiao Ma, Yingzhuo Yu, Julie Shah, Joyce Chai.* IJCAI, 2023.

Embodied Dialogue Agents





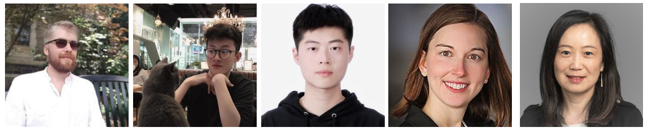
Asymmetric collaboration in a simulated world [EMNLP 2021, IJCAI 2023].

- MindCraft:

  Two macro-actions: Creating a block + Combining two blocks to create a new block.

MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks. *Cristian-Paul Bara, Sky CH-Wang, Joyce Chai.* EMNLP, 2021.
Towards Collaborative Plan Acquisition through Theory of Mind Modeling in Situated Dialogue. *Cristian-Paul Bara, Ziqiao Ma, Yingzhuo Yu, Julie Shah, Joyce Chai.* IJCAI, 2023.

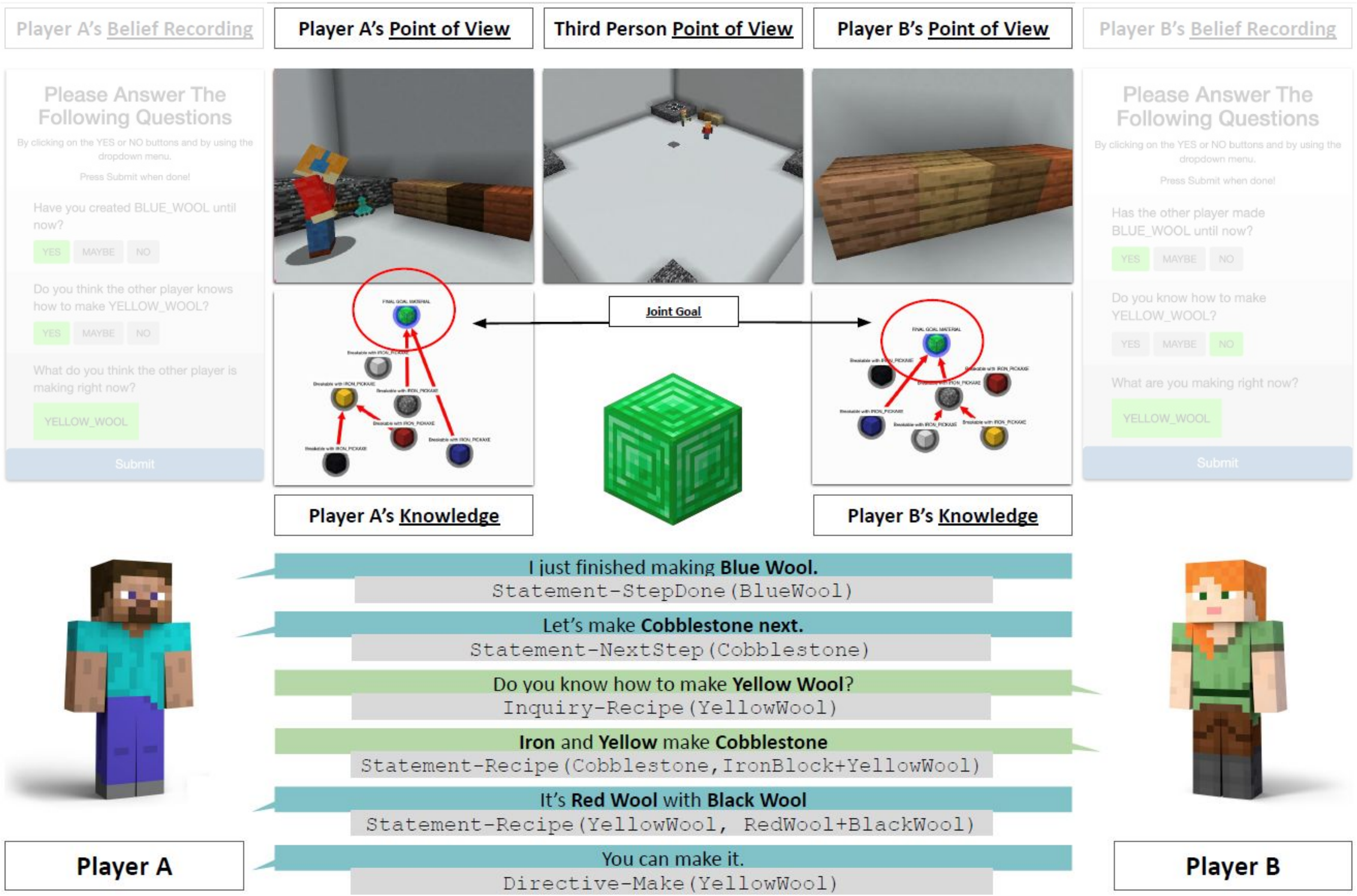
Embodied Dialogue Agents



Asymmetric collaboration in a simulated world [EMNLP 2021, IJCAI 2023].

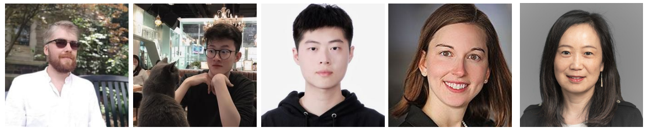
- MindCraft:

  Players can communicate in natural language with an in-game chat-box.



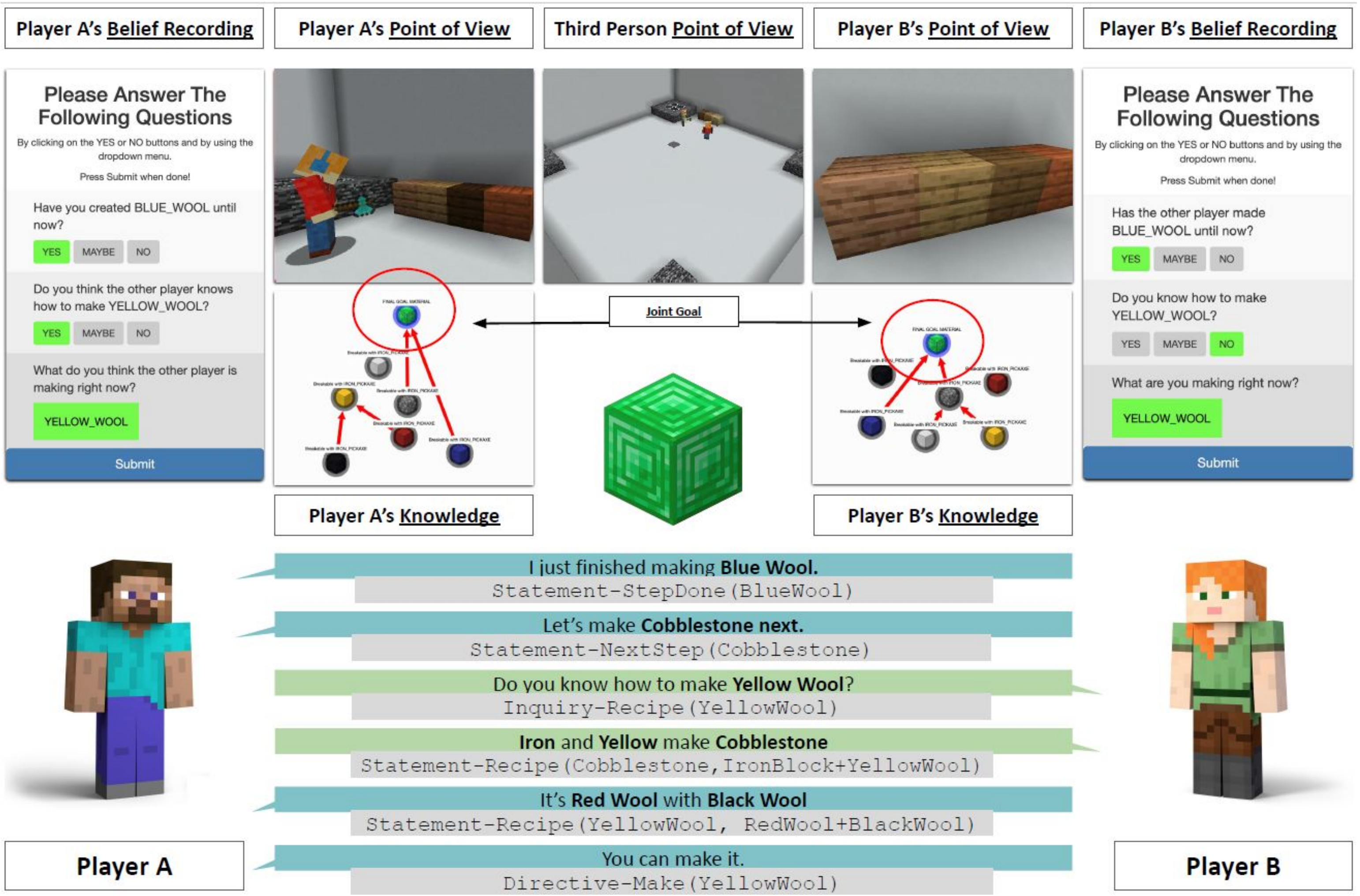
MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks. *Cristian-Paul Bara, Sky CH-Wang, Joyce Chai.* EMNLP, 2021.
Towards Collaborative Plan Acquisition through Theory of Mind Modeling in Situated Dialogue. *Cristian-Paul Bara, Ziqiao Ma, Yingzhuo Yu, Julie Shah, Joyce Chai.* IJCAI, 2023.

Embodied Dialogue Agents



Asymmetric collaboration in a simulated world [EMNLP 2021, IJCAI 2023].

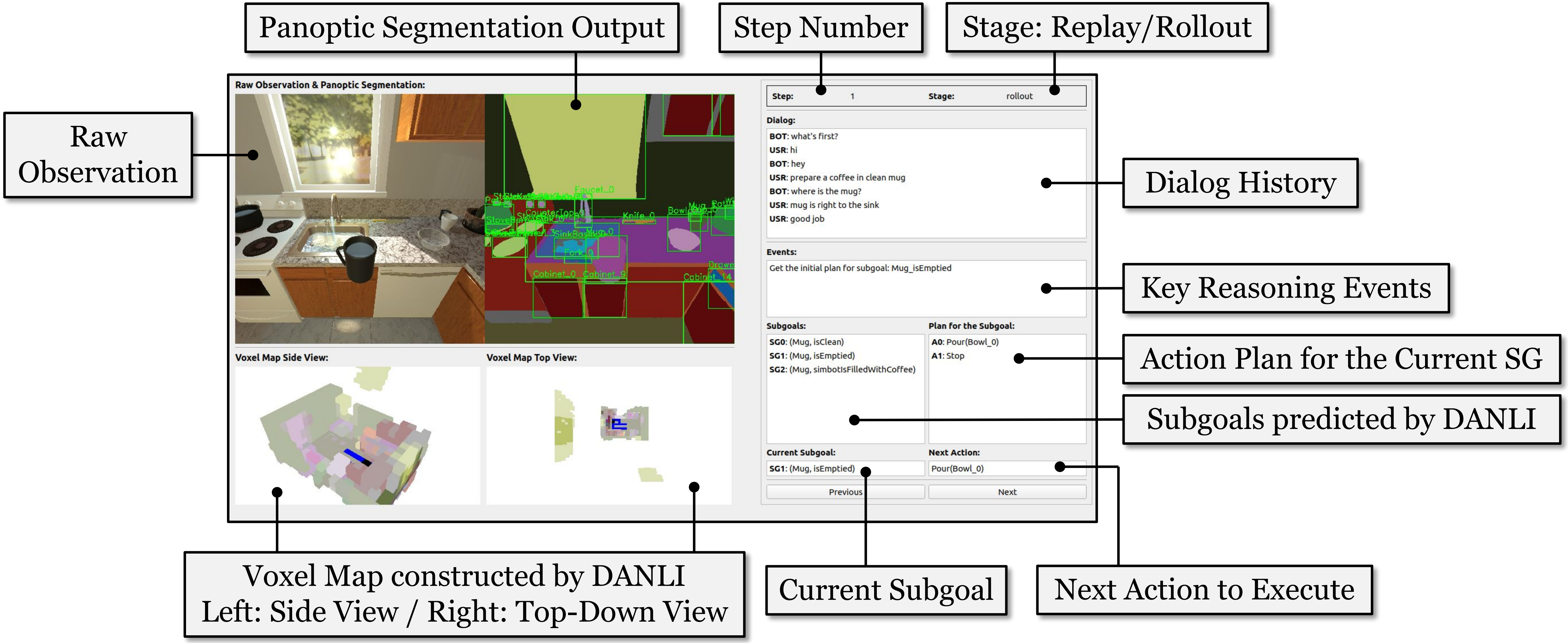
- Annotations for mental states:
  - **Task Intention**: predict the sub-goal that the partner is currently working on;
  - **Task Status**: predict whether the partner believes a certain sub-goal is completed and by whom;
  - **Task Knowledge**: predict whether the partner knows how to achieve a sub-goal, i.e., all the incoming edges of a node.



MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks. *Cristian-Paul Bara, Sky CH-Wang, Joyce Chai.* EMNLP, 2021.
Towards Collaborative Plan Acquisition through Theory of Mind Modeling in Situated Dialogue. *Cristian-Paul Bara, Ziqiao Ma, Yingzhuo Yu, Julie Shah, Joyce Chai.* IJCAI, 2023.

Embodied Dialogue Agents

Deliberative agent for following natural language instructions [EMNLP 2022]

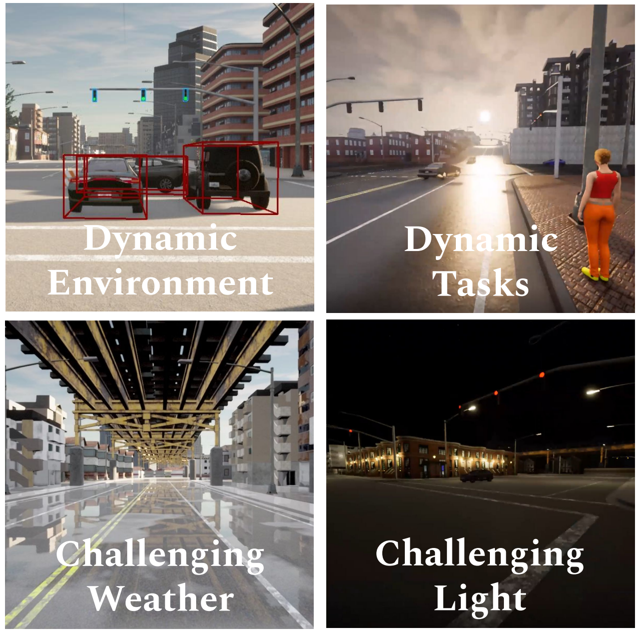


DANLI: Deliberative Agent for Following Natural Language Instructions. *Yichi Zhang, Jianing Yang, Jiayi Pan, Shane Storks, Nikhil Devraj, Ziqiao Ma, Keunwoo Peter Yu, Yuwei Bao, Joyce Chai.* EMNLP 2022.

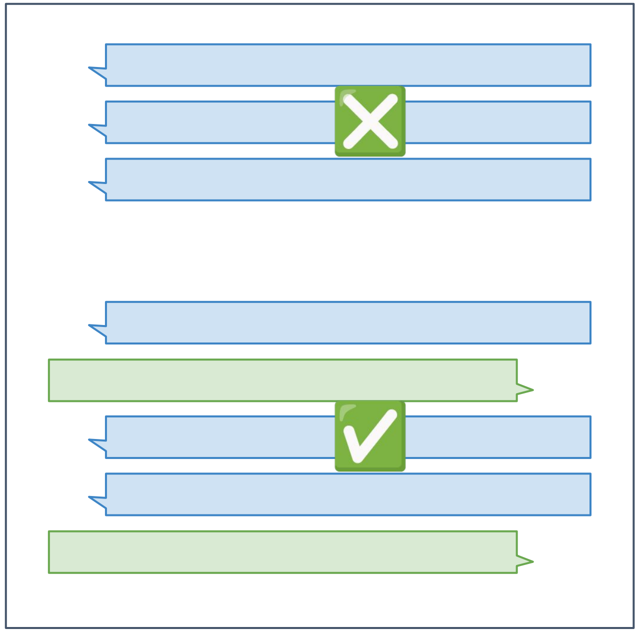
Embodied Dialogue Agents

Dialogue-guided autonomous driving [EMNLP 2023, IROS 2024]

Discrete -> Continuous
Static -> Dynamic



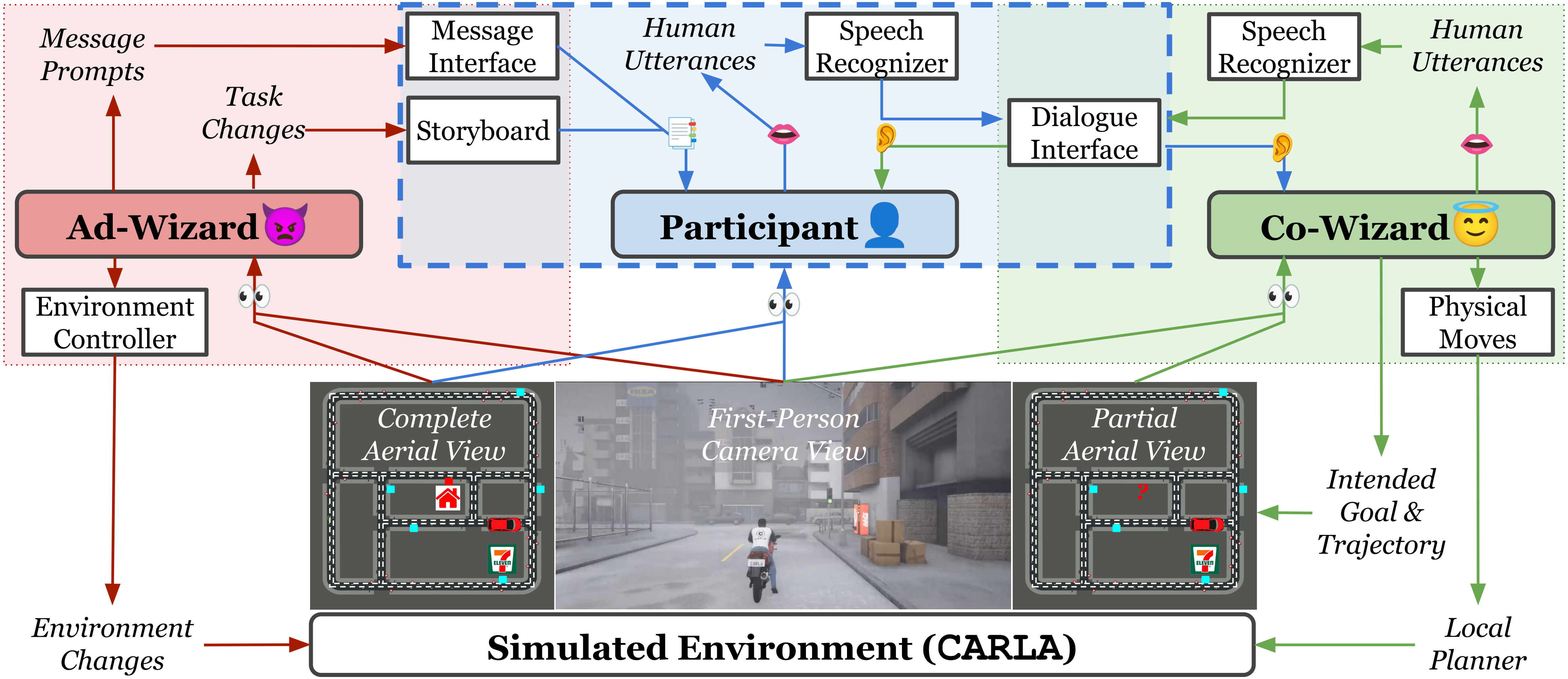
Instruction -> Communication
Asymmetry -> Symmetry



DOROTHIE: Spoken Dialogue for Handling Unexpected Situations in Interactive Autonomous Driving Agents. *Ziqiao Ma, Ben VanDerPloeg, Cristian-Paul Bara, Huang Yidong, Eui-In Kim, Felix Gervits, Matthew Marge, Joyce Chai.* EMNLP Findings, 2023.

Embodied Dialogue Agents

Dialogue-guided autonomous driving [EMNLP 2023, IROS 2024]

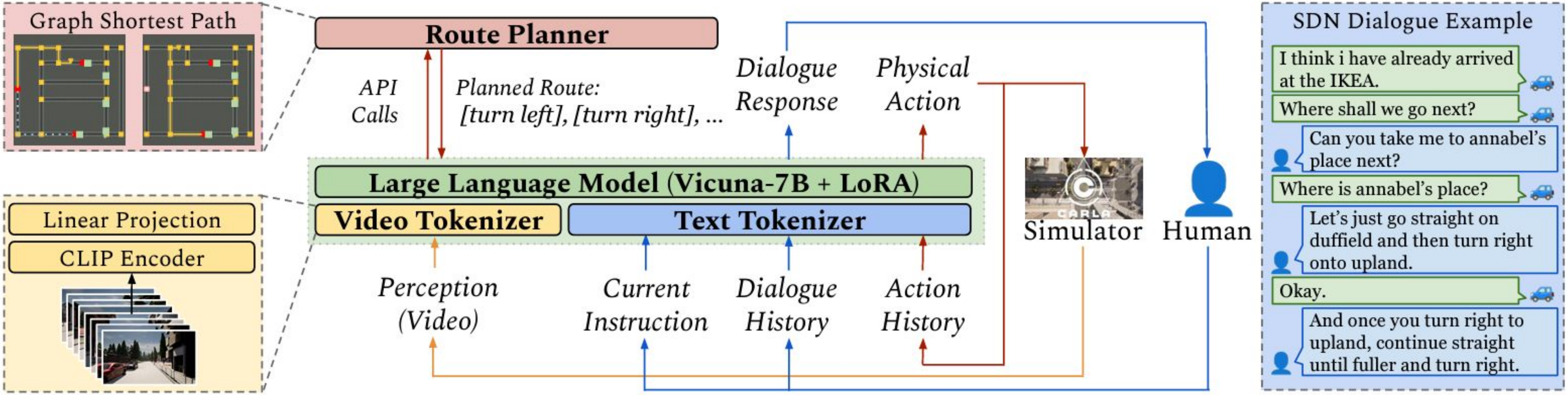


DOROTHIE: Spoken Dialogue for Handling Unexpected Situations in Interactive Autonomous Driving Agents. *Ziqiao Ma, Ben VanDerPloeg, Cristian-Paul Bara, Huang Yidong, Eui-In Kim, Felix Gervits, Matthew Marge, Joyce Chai.* EMNLP Findings, 2023.

Embodied Dialogue Agents

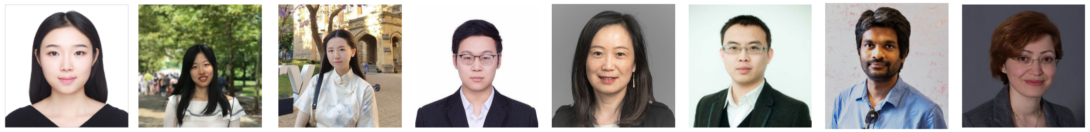
Dialogue-guided autonomous driving [EMNLP 2023, IROS 2024]

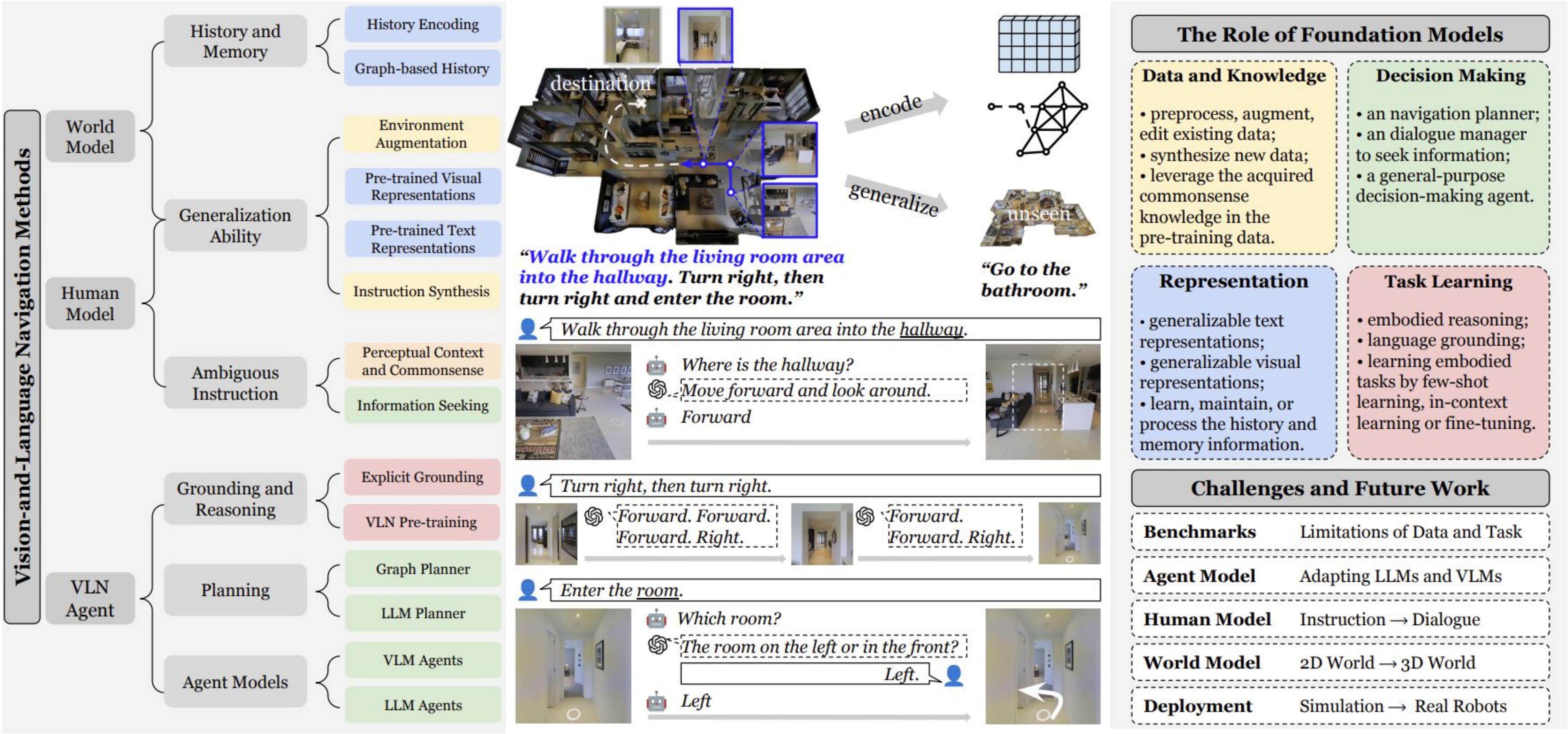
- DriVLMe, an video-language model agent that learn from embodied and social experiences.



DriVLMe: Enhancing LLM-based Autonomous Driving Agents with Embodied and Social Experiences. *Yidong Huang, Jacob Sansom, Ziqiao Ma, Felix Gervits, Joyce Chai*. IROS 2024

Embodied Dialogue Agents

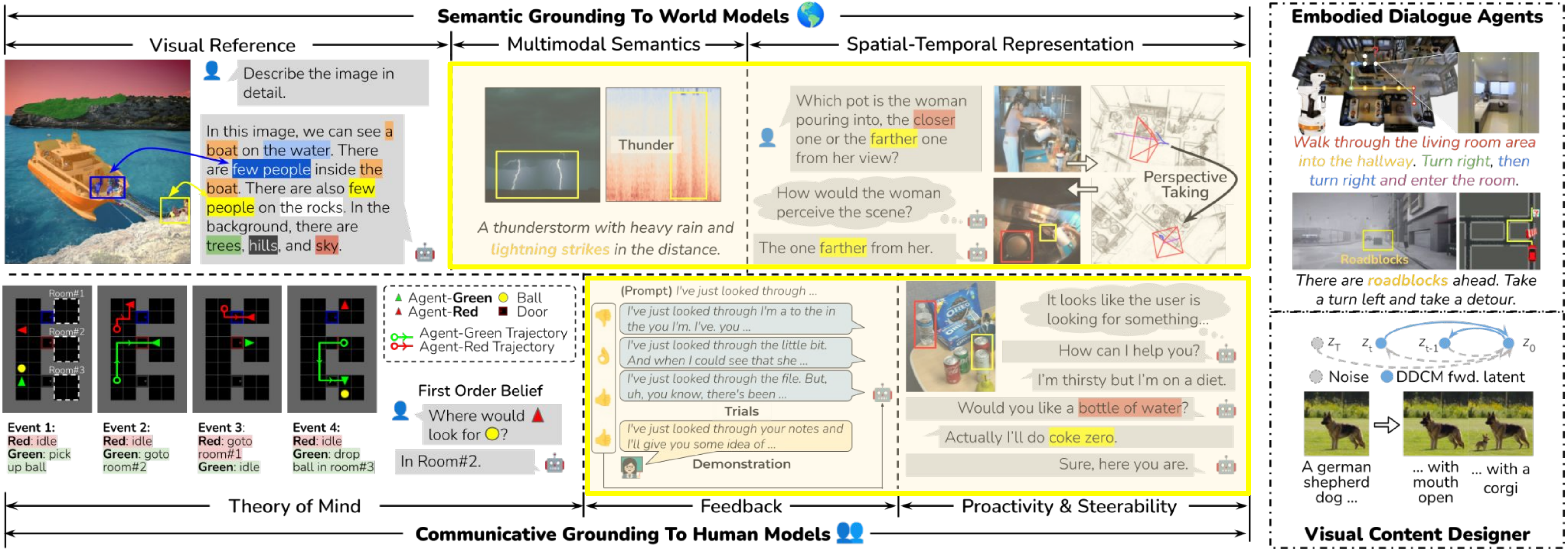




Vision-and-Language Navigation Today and Tomorrow: A Survey in the Era of Foundation Models. Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, Parisa Kordjamshidi. TMLR 2024.

Landing Language Models on the “Ground”

Language grounding is far from solved and embodied dialogue agents are not there yet!



Landing Language Models on the “Ground”

Bi-Align Workshop @ ICLR 2025 and SIG @ CHI 2025

ICLR 2025 Workshop on Bidirectional Human-AI Alignment
(Bi-Align @ ICLR 2025 Workshop Proposal)

Hua Shen, Ziqiao Ma, Reshmi Ghosh, Tiffany Knearem
Michael Liu, Tongshuang Wu, Andrés Monroy-Hernández, Diyi Yang, Antoine Bosselut
Furong Huang, Tanu Mitra, Joyce Chai, Marti A. Hearst, Dawn Song, Yang Li

Been Kim
Google Deepmind

Frauke Kreuter
UMD

Dan Bohus
Microsoft

Richard Ngo
OpenAI

Pavel Izmailov
Anthropic / NYU

Hung-yi Lee
NTU

Elizebeth Churchill
MBZUAI

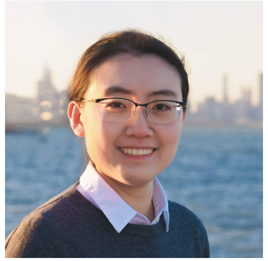
Brad Myers
CMU

Landing Language Models on the "Ground"

Learning Language through Grounding Tutorial @ NAACL 2025

Learning Language through Grounding

Freda Shi[1,2] **Ziqiao Ma**[3] **Jiayuan Mao**[4] **Parisa Kordjamshidi**[5] **Joyce Chai**[3]
[1]University of Waterloo [2]Vector Institute & Canada CIFAR AI Chair [3]University of Michigan
[4]Massachusetts Institute of Technology [5]Michigan State University
fhs@uwaterloo.ca, {marstin,chaijy}@umich.edu, jiayuanm@mit.edu, kordjams@msu.edu


Freda Shi
UWaterloo & Vector


Ziqiao Ma
UMich


Jiayuan Mao
MIT


Parisa Kordjamshidi
MSU


Joyce Chai
UMich