Grounding and Distinguishing Conceptual Vocabulary Through Similarity Learning in Embodied Simulations

Sadaf Ghaffari and Nikhil Krishnaswamy

IWCS 2023, Nancy, France

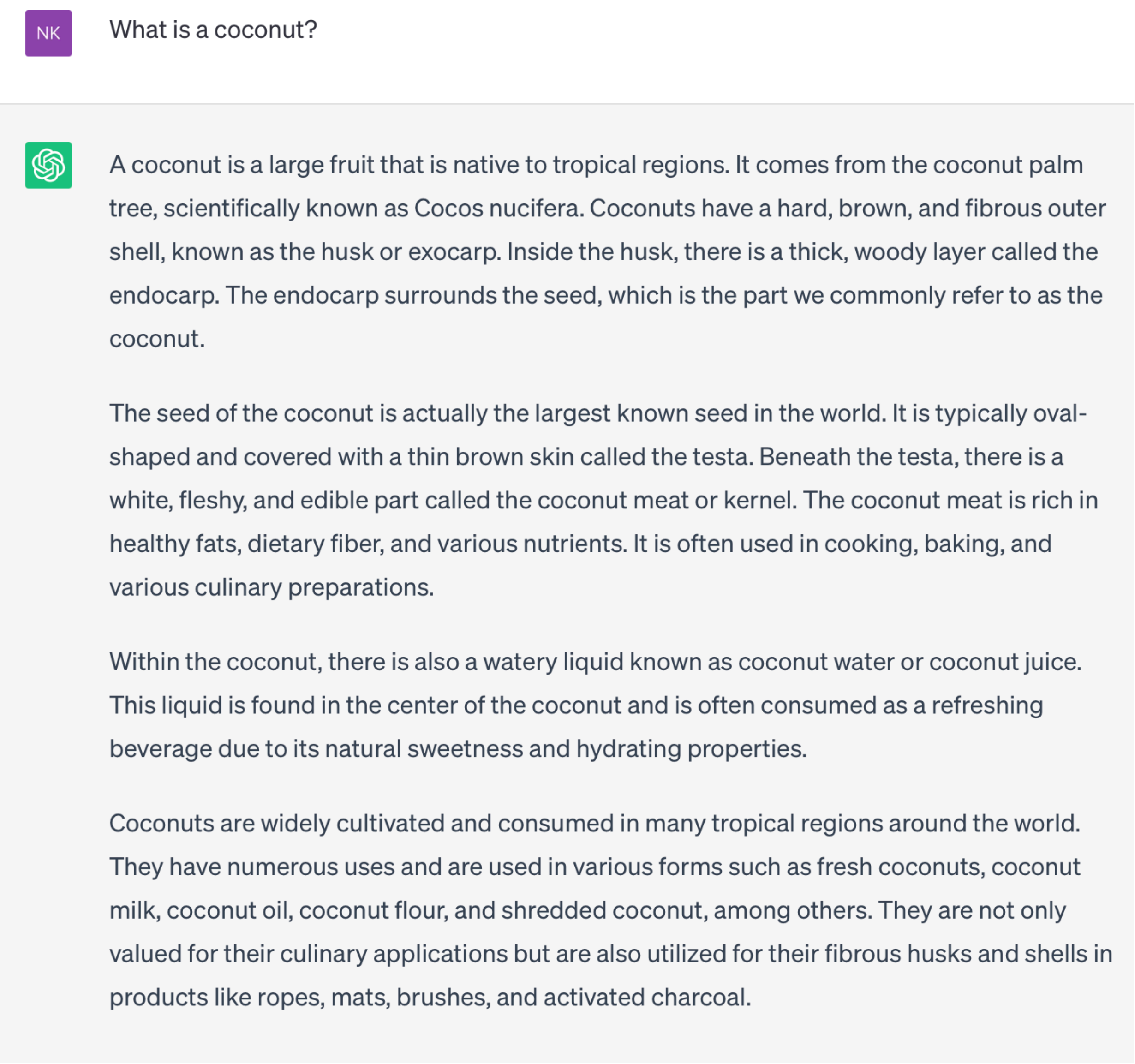
June 23, 2023

Colorado State University

Introduction

Introduction

- A common critique of language models is that they are "ungrounded"
- Lack ability to link an utterance to a communicative intent
- A sophisticated language model can generate text about an object without understanding what the object is 👉
- Grounding usually involves recourse to another modality
- Wealth of research in cross-modal semantic linking, language grounding to images



What is a coconut?

A coconut is a large fruit that is native to tropical regions. It comes from the coconut palm tree, scientifically known as Cocos nucifera. Coconuts have a hard, brown, and fibrous outer shell, known as the husk or exocarp. Inside the husk, there is a thick, woody layer called the endocarp. The endocarp surrounds the seed, which is the part we commonly refer to as the coconut.

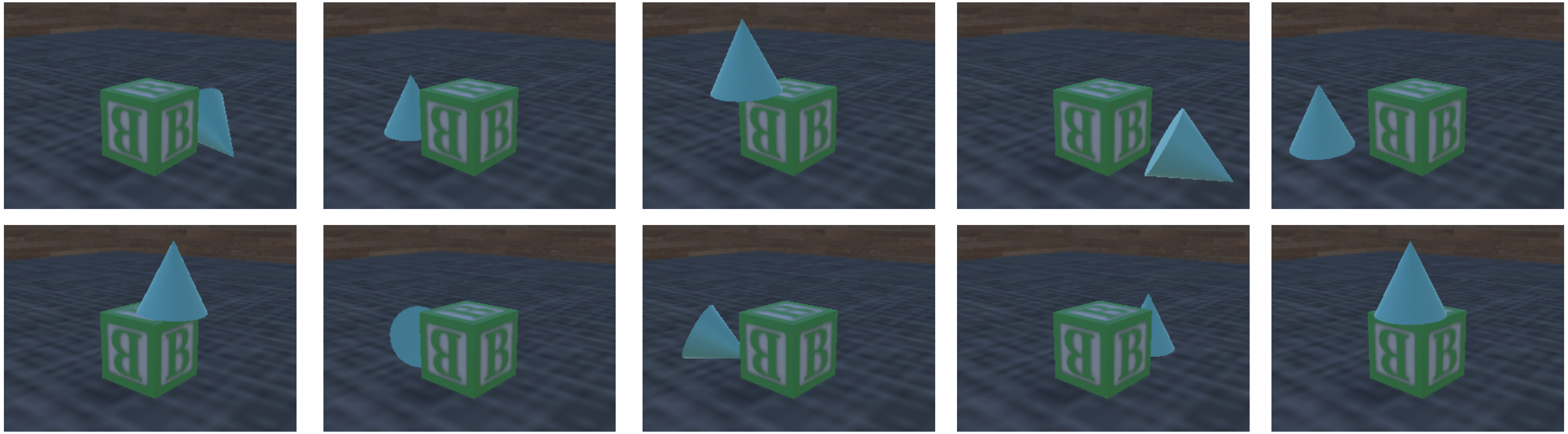
The seed of the coconut is actually the largest known seed in the world. It is typically oval-shaped and covered with a thin brown skin called the testa. Beneath the testa, there is a white, fleshy, and edible part called the coconut meat or kernel. The coconut meat is rich in healthy fats, dietary fiber, and various nutrients. It is often used in cooking, baking, and various culinary preparations.

Within the coconut, there is also a watery liquid known as coconut water or coconut juice. This liquid is found in the center of the coconut and is often consumed as a refreshing beverage due to its natural sweetness and hydrating properties.

Coconuts are widely cultivated and consumed in many tropical regions around the world. They have numerous uses and are used in various forms such as fresh coconuts, coconut milk, coconut oil, coconut flour, and shredded coconut, among others. They are not only valued for their culinary applications but are also utilized for their fibrous husks and shells in products like ropes, mats, brushes, and activated charcoal.

Introduction

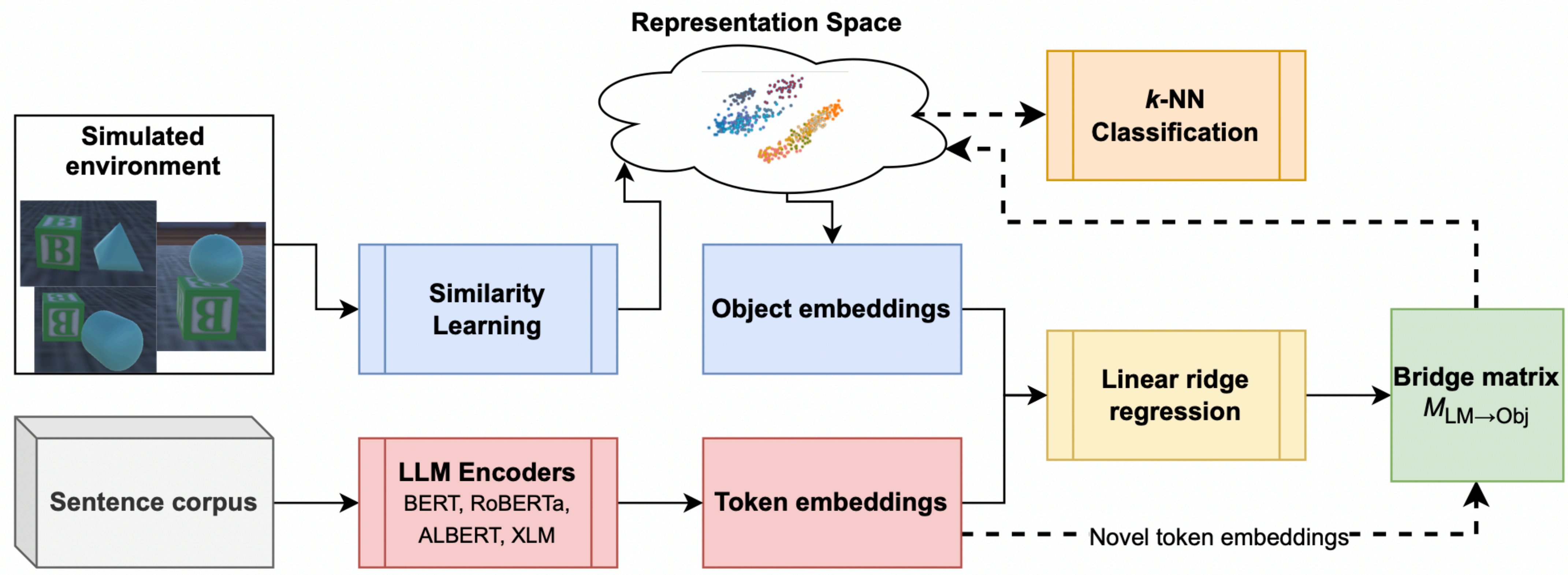
- “Multimodal” doesn’t just mean language+vision
- Humans don’t just use images as their only non-linguistic modality
- A wealth of environment and sensory experience is implicated in learning



Introduction

- We take an *embodied simulation* approach to grounding
- Use a virtual environment to create experiences for an agent interacting with various objects
- Object motions leave **trajectories in space** based on their geometric properties and **affordances**
- Similarity learning over agent's experience can make analogical comparisons between object types …
- … and appears to learn more abstract properties of the objects
- Words for concrete nouns (object types) are easy to ground to this representation space using **affine transformation**
- Grounding concrete terms **provides a scaffold** for learning and distinguishing the meaning of other terms in context
- Explore the properties of different language models for grounding concrete objects to the learned space vs. abstract terms (verbs, properties, etc.)

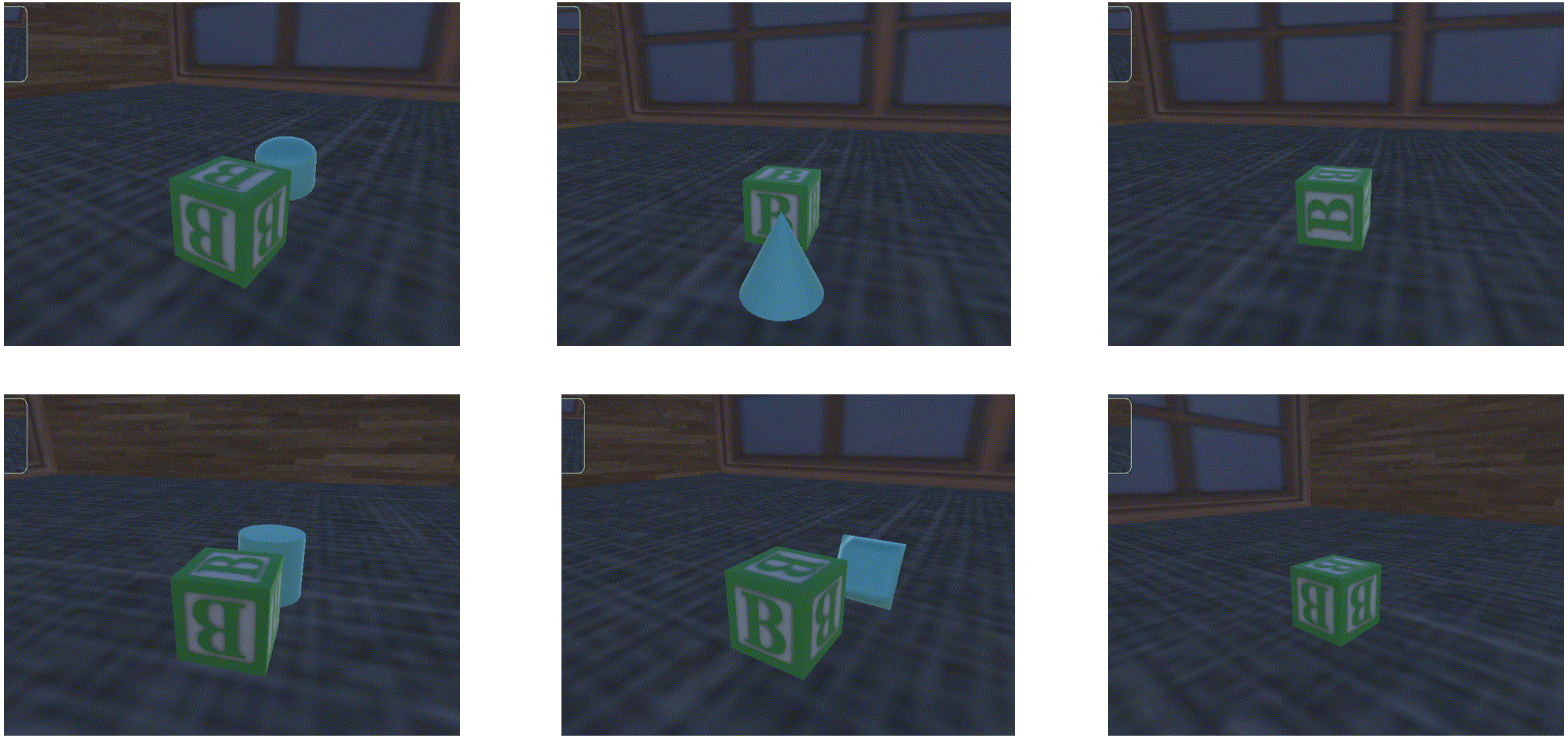
Schematic Overview



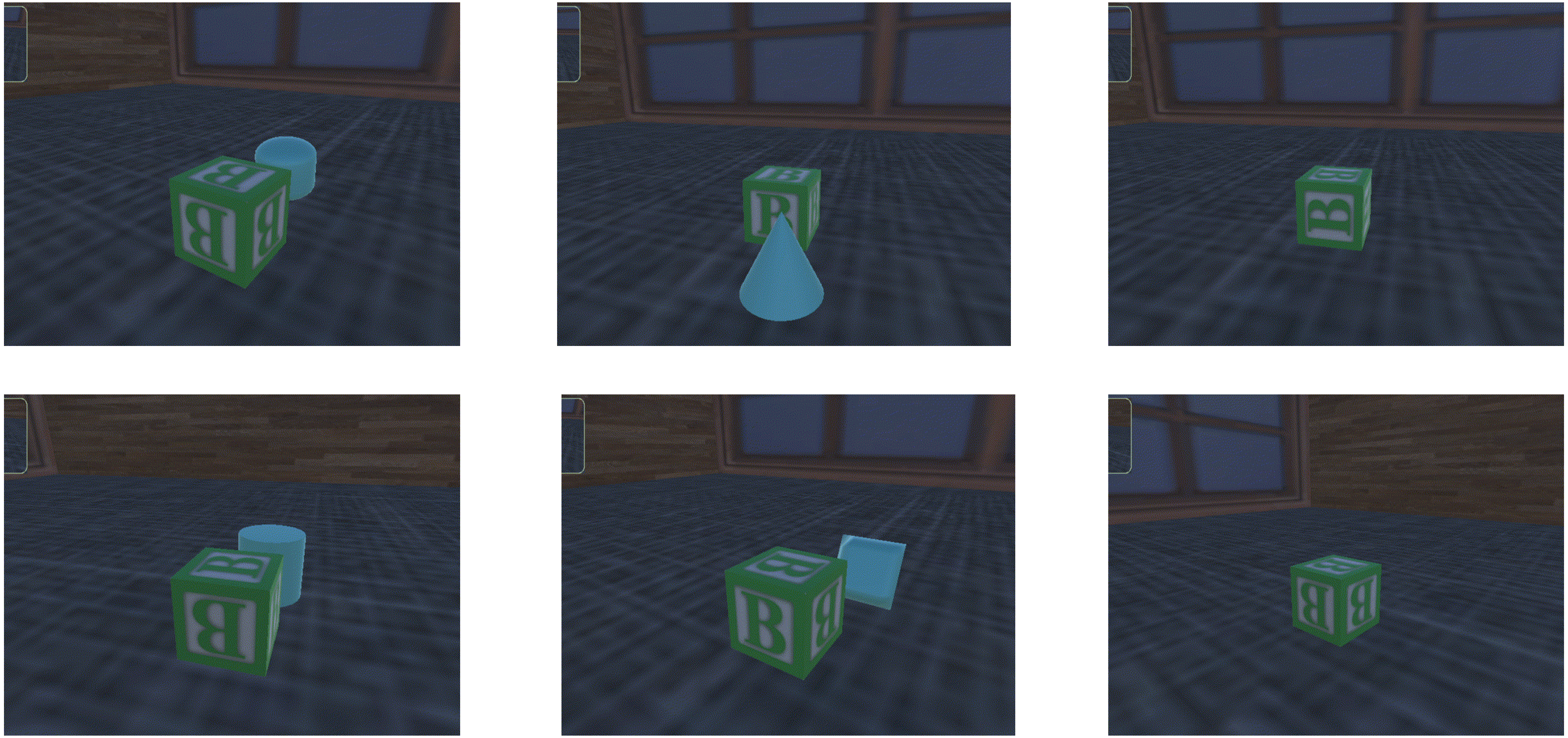
Why Do We Think This Will Work?

- As humans map object concept representations to nouns, they also learn to individuate them from the perceptual flow, based on experience and interaction, not just visual features but also (Spelke, 1985; Spelke et al., 1989; Spelke, 1990; Baillargeon, 1987)
- Gentner (2006) argues that variability in verbal semantics (Talmy, 1975) helped explain why nouns are typically learned before verbs
- Problem: neural language and “object” (based on visual or tabular data) representation spaces are not directly comparable
- Affine transformation technique between embedding spaces has been successfully used in vision and language use cases (McNeely-White et al., 2022; Nath et al., 2022)
    - Here we explore applicability in a cross-modal setting

Sampling Object Properties

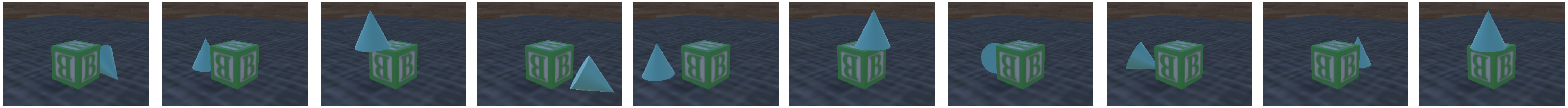


Sampling Object Properties



Similarity Learning of Object Properties

Similarity Learning of Object Properties



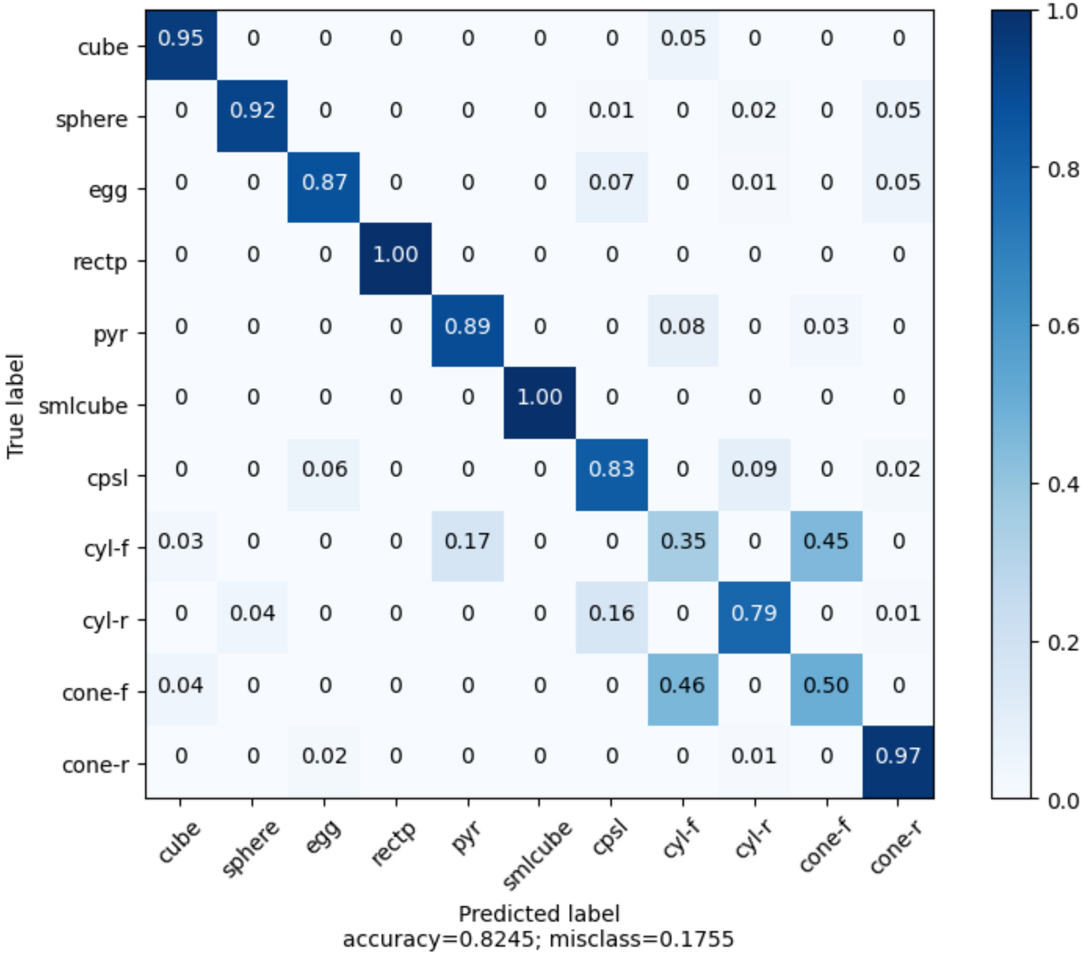
- Dataset: Samples gathered from an agent stacking objects in a virtual environment
- Behavior of different object types under interaction (stacking)
- Object behavior bootstrapped with ontological knowledge about symmetry
- E.g., Objects placed on their rounded edges are more likely to roll
- Object trajectories, geometric features captured through simulated environment
- Object types: cube, sphere, cylinder, capsule, egg, pyramid, cone, rectangular prism, small cube
- Captured 43 numerical values describing each object interaction
  - action taken, object position before and after action, orientation before and after action, spatial relations between objects, etc. (Ghaffari and Krishnaswamy, 2022)

Similarity Learning of Object Properties

- Use a CNN with multisimilarity loss $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\alpha}\log\left(1+\sum_{k\in P_i}1+e^{-\alpha(S_{ik}-\lambda)}\right)+\frac{1}{\beta}\log\left(1+\sum_{k\in N_i}1+e^{\beta(S_{ik}-\lambda)}\right)\right)$ to classify objects by type
- Construct a similarly matrix S where S_{ik} is the similarity of samples $\{x_i, x_k\}$ according to neural network f with weights θ
- Adam optimization, LR 5x10^{-6}, batch size 70, 20 epochs, embedding size 64
- Train only on a **subset of objects** (cube, sphere, egg, capsule, small cube, rectangular prism, pyramid)
  - Objects that are all flat-sided or all round-sided
- Two objects (cylinder, cone) have both flat and round sides
  - Split these samples based on their stacking behavior (stays stacked vs. falls off)
- Test set includes seven seen classes and four unseen classes

Similarity Learning of Object Properties

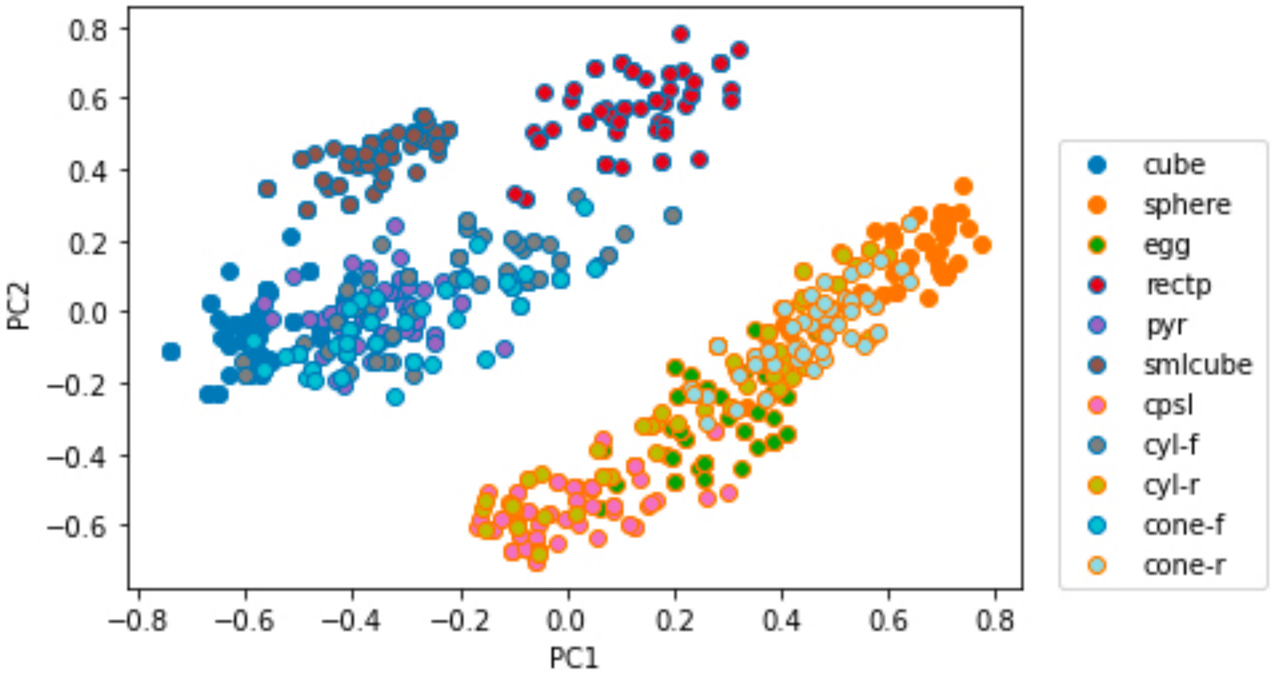
- Neural approaches successfully classify these geometric features into different object types
- Unseen classes are also classifiable using KNN over embeddings
- Where confusion arises, they are between different flat and different round objects
  - Never between these classes
- Model appears to learn contrast between “flatness” and “roundness”



True label \ Predicted label	cube	sphere	egg	rectp	pyr	smlcube	cpsl	cyl-f	cyl-r	cone-f	cone-r
cube	0.95	0	0	0	0	0	0	0.05	0	0	0
sphere	0	0.92	0	0	0	0	0.01	0	0.02	0	0.05
egg	0	0	0.87	0	0	0	0.07	0	0.01	0	0.05
rectp	0	0	0	1.00	0	0	0	0	0	0	0
pyr	0	0	0	0	0.89	0	0	0.08	0	0.03	0
smlcube	0	0	0	0	0	1.00	0	0	0	0	0
cpsl	0	0	0.06	0	0	0	0.83	0	0.09	0	0.02
cyl-f	0.03	0	0	0	0.17	0	0	0.35	0	0.45	0
cyl-r	0	0.04	0	0	0	0	0.16	0	0.79	0	0.01
cone-f	0.04	0	0	0	0	0	0	0.46	0	0.50	0
cone-r	0	0	0.02	0	0	0	0	0	0.01	0	0.97

Similarity Learning of Object Properties

- Neural approaches successfully classify these geometric features into different object types
- Unseen classes are also classifiable using KNN over embeddings
- Where confusion arises, they are between different flat and different round objects
    - Never between these classes
- Model appears to learn contrast between “flatness” and “roundness”
- **Two distinct “flat” and “round” clusters!**

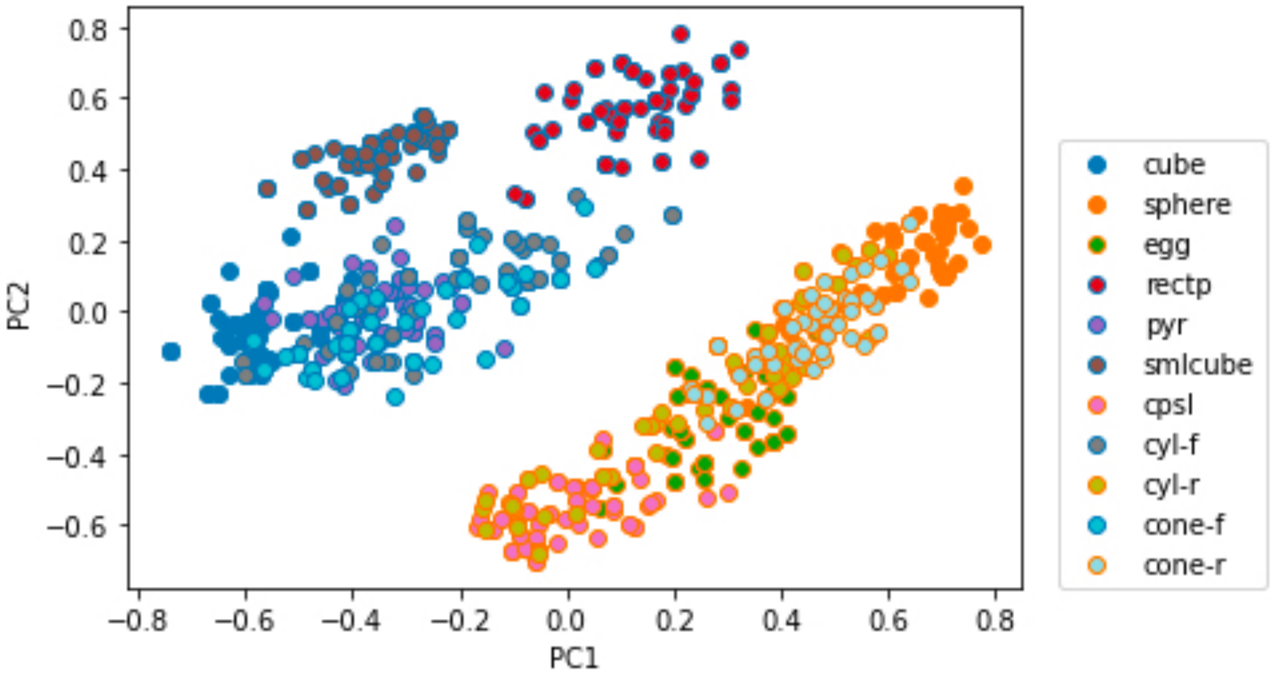


800 test samples

Language Grounding to Environment

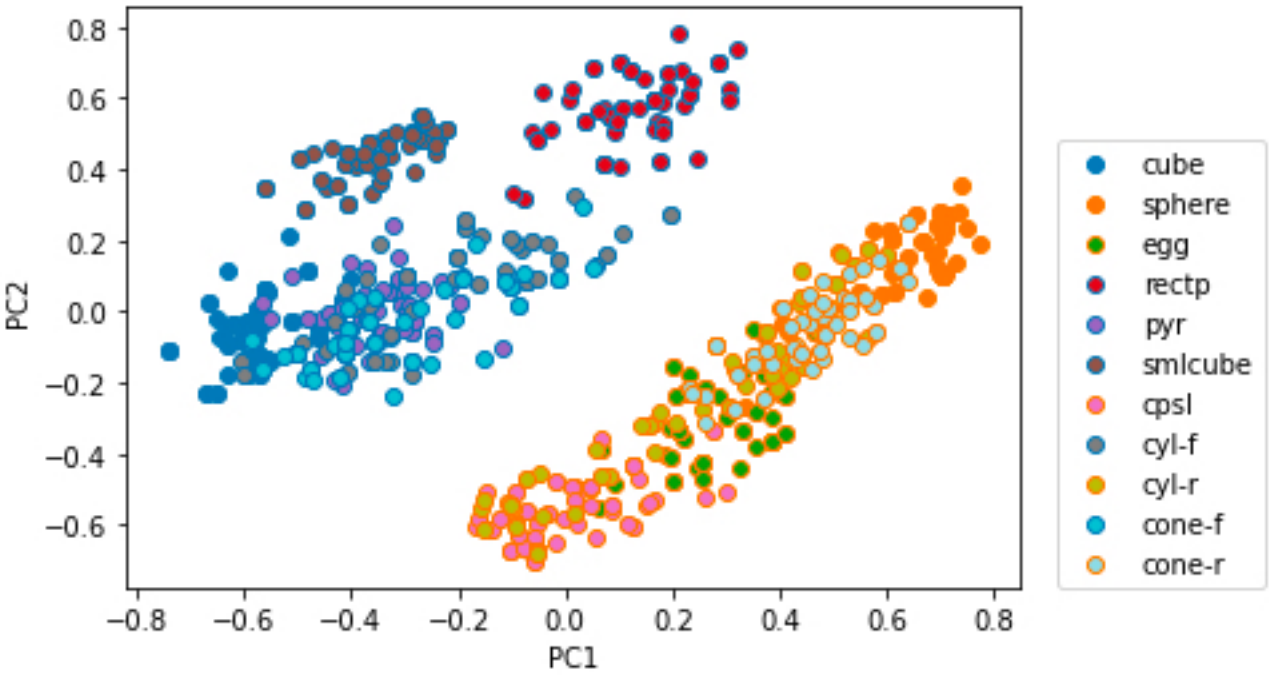
Language Grounding to Environment

- Individual embeddings of the same object type form a high-dimensional region defining the object representation
- Ethayarajh (2019) observed similar phenomena in the representations of contextualized token vectors from LLMs
- Suggests a structure-preserving mapping exists between equivalent regions in different embedding space
- **Affine transformation** preserves collinearity, parallelism



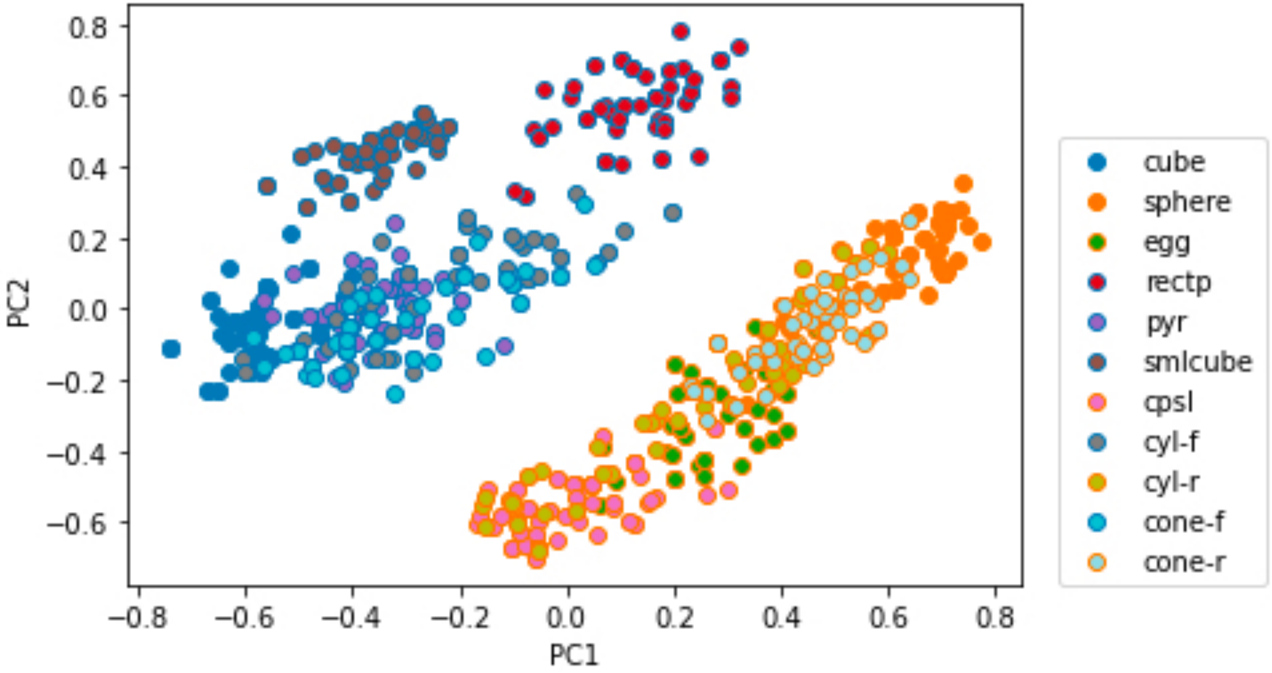
Language Grounding to Environment

- **RQ: If agent encounters novel objects, can it learn words for them by grounding tokens to the object representation space?**
  1. Prompt language model: Generate sentences containing target term to be grounded
  2. Extract token representations: For each instance of target token, extract contextualized numerical representation
  3. Linear regression between paired embedding vectors: Compute transformation $\mathscr{M} \in \mathbb{R}^{d_A} \times \mathbb{R}^{d_B}$ with ridge regression



Language Grounding to Environment

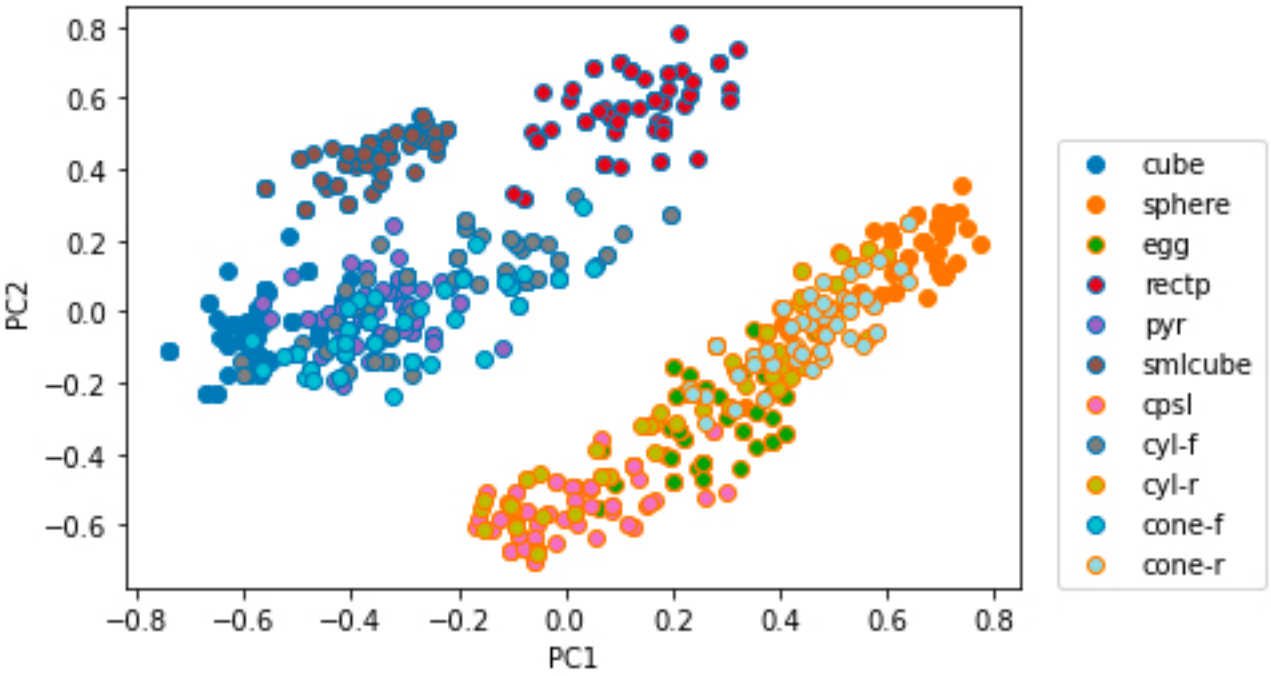
- Given two models: language model A and object representation model B, token representations in A and object representations in B form subspaces in the embedding spaces
- Vectors chosen from respective embedding spaces form at minimum the spanning set of the respective regions
- Vectors chosen to compute $\mathscr{M}$ represent similar, non-identical instances of token/object
- Optimal $\mathscr{M}$: **structure-preserving** transformation that transforms new object-denoting tokens into the object-representation region



Language Grounding to Environment

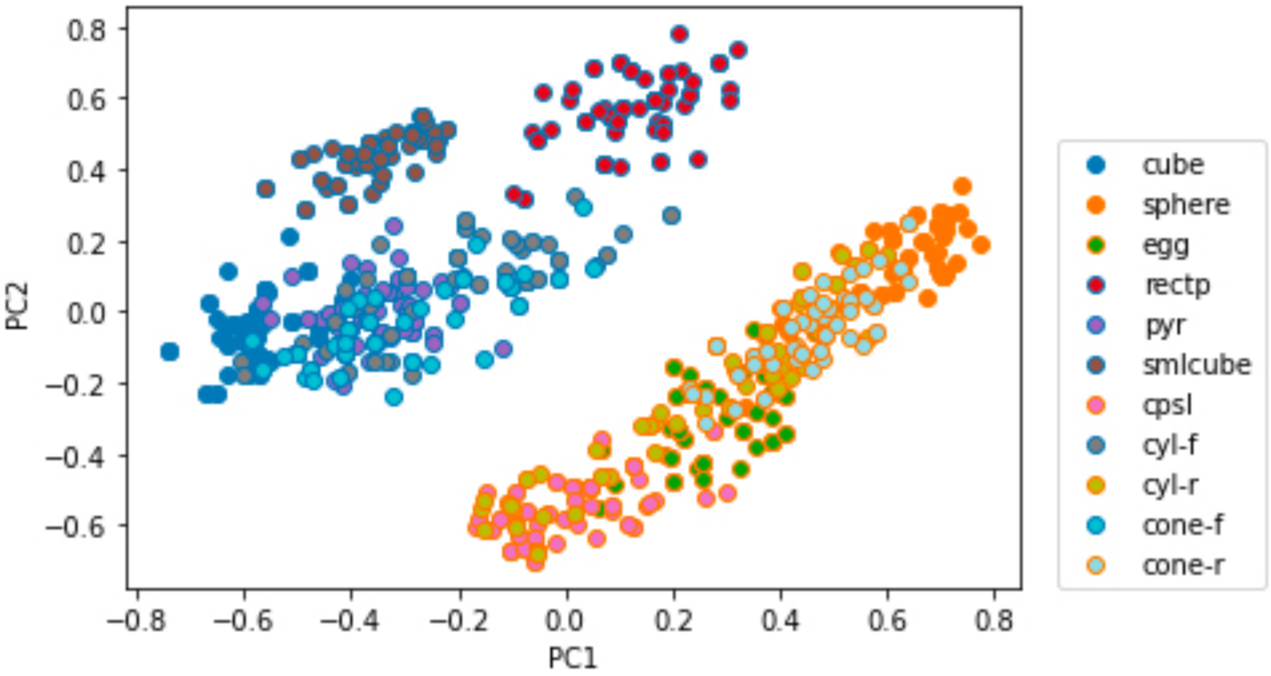
4. Dialogue with a language model: Generate *novel* sentences containing instances of the transformed ("grounded") target term
5. Extract token-level representations: cf. Step 2
6. Transform new tokens into object space: evaluation - do the expected senses of grounded terms cluster with the right objects?

- Sentence generation: ChatGPT
- Prompt ChatGPT with sentences about objects and their properties, e.g., "Write 40 short sentences about how blocks are flat on all sides and can be stacked" (total corpus of 440 sentences)
- Take the most frequently occurring domain-appropriate tokens as candidate target terms



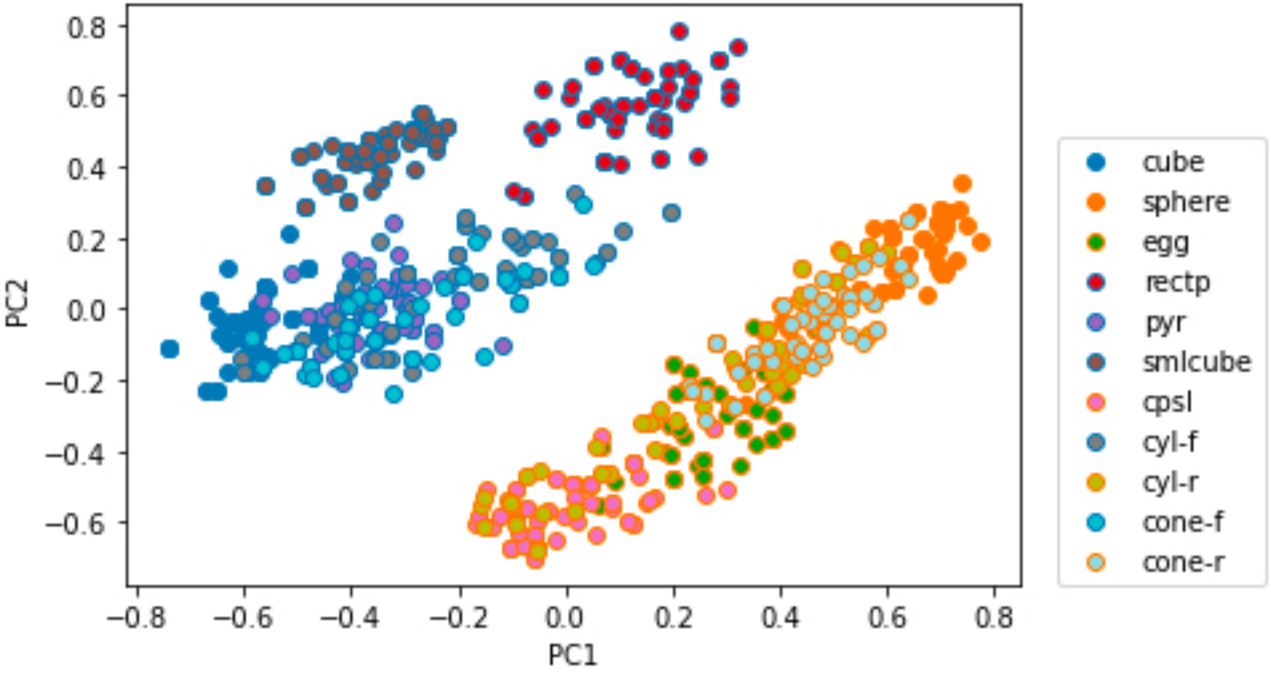
Language Grounding to Environment

- Take most frequently occurring domain appropriate tokens as candidate target terms
  - Not just object terms
  - Properties: “flat”/“round”, “stable”/“unstable"
  - Behaviors: “stack”/“roll”, “stand”/“fall”
- What happens to these terms as object terms are grounded?
- Pull token representations from 4 LMs:
  - BERT, RoBERTa, ALBERT, XLM
- Compute transformation using only object token vectors and object behavior vectors



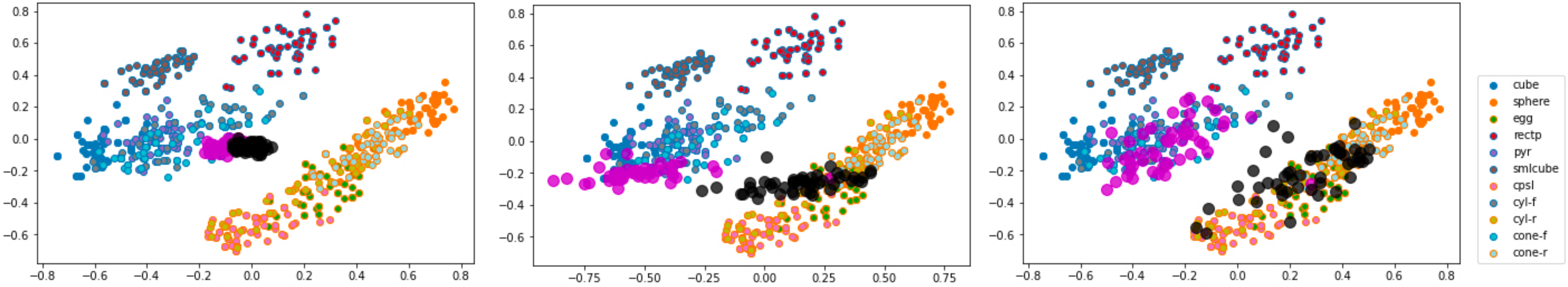
Language Grounding to Environment

- 5 contextualized embeddings of each target word paired with 5 randomly selected representations of the associated object
- Calculate “bridge matrix” using ridge regression
- Perform iterative experiments, incrementally adding new object concepts to improve the transformation
  - Order adapted from Ghaffari and Krishnaswamy (2022)
- Evaluate transformation by transforming word vectors for concepts not used in computing bridge matrix
  - Evaluate KNN classification of terms, and separation of cluster centers
- Final step: explicit “hint” including 5 embeddings of the novel concept to be grounded and a co-occurring object word



Language Grounding to Environment

- Transformed token clusters separate even when they're never explicitly seen in transformation

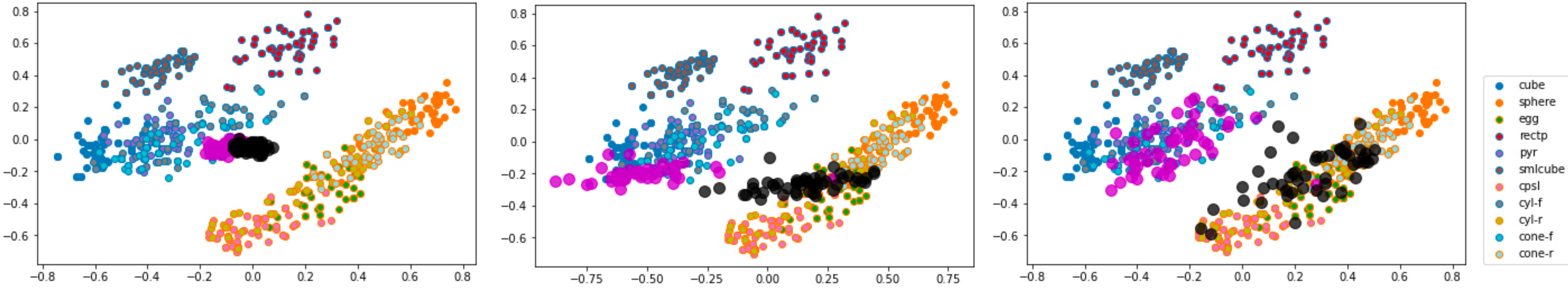


XLM: "flat" (pink) vs. "round" (black)

Language Grounding to Environment

- Transformed token clusters separate even when they're never explicitly seen in transformation

Transformation using only cube, sphere, and egg



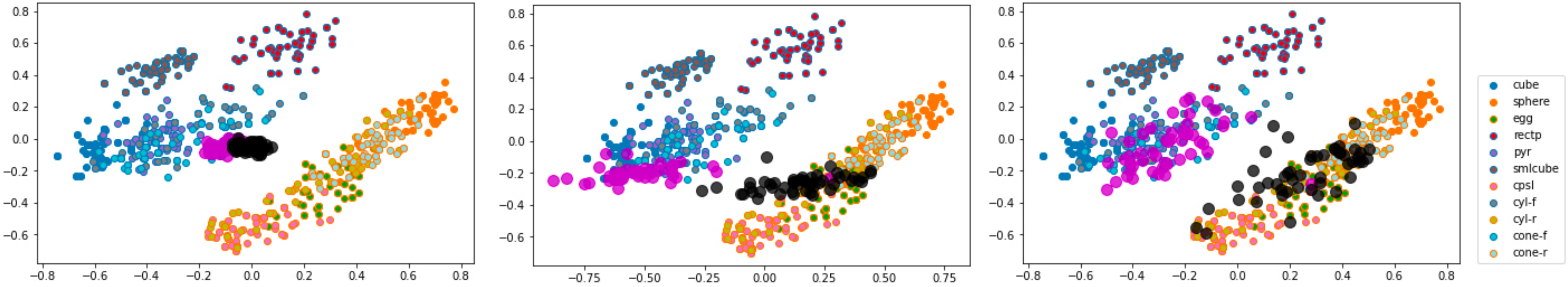
XLM: "flat" (pink) vs. "round" (black)

Language Grounding to Environment

- Transformed token clusters separate even when they’re never explicitly seen in transformation

Transformation using only cube, sphere, and egg

Using all objects



XLM: “flat” (pink) vs. “round” (black)

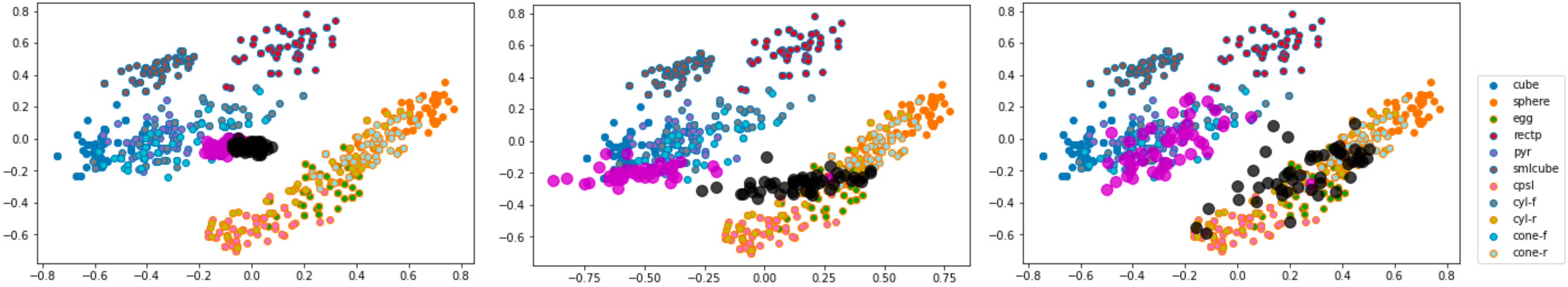
Language Grounding to Environment

- Transformed token clusters separate even when they're never explicitly seen in transformation

Transformation using only cube, sphere, and egg

Using all objects

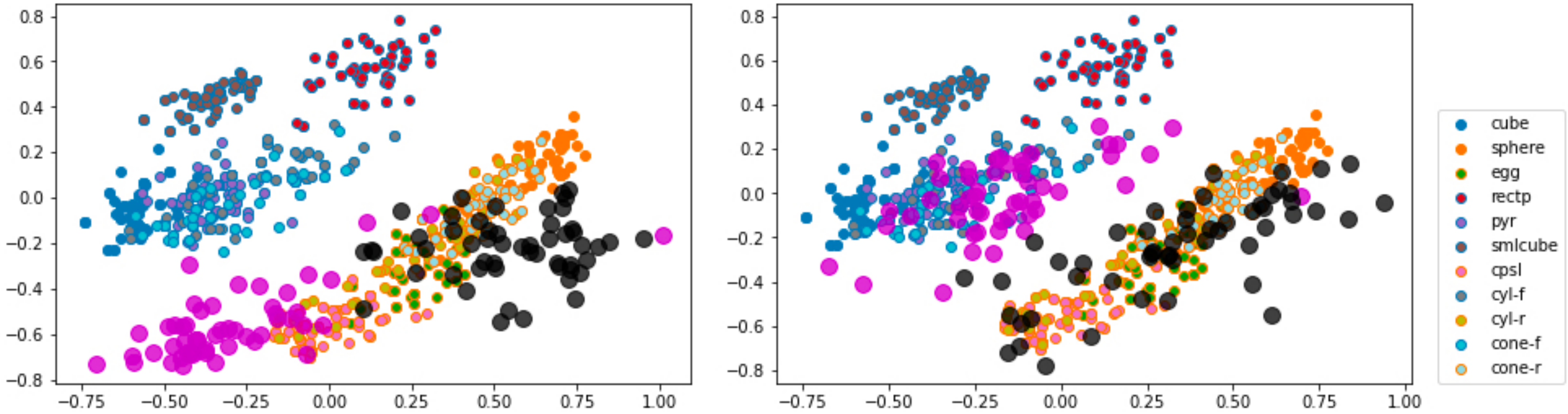
Using all objects + "hint"



XLM: "flat" (pink) vs. "round" (black)

Language Grounding to Environment

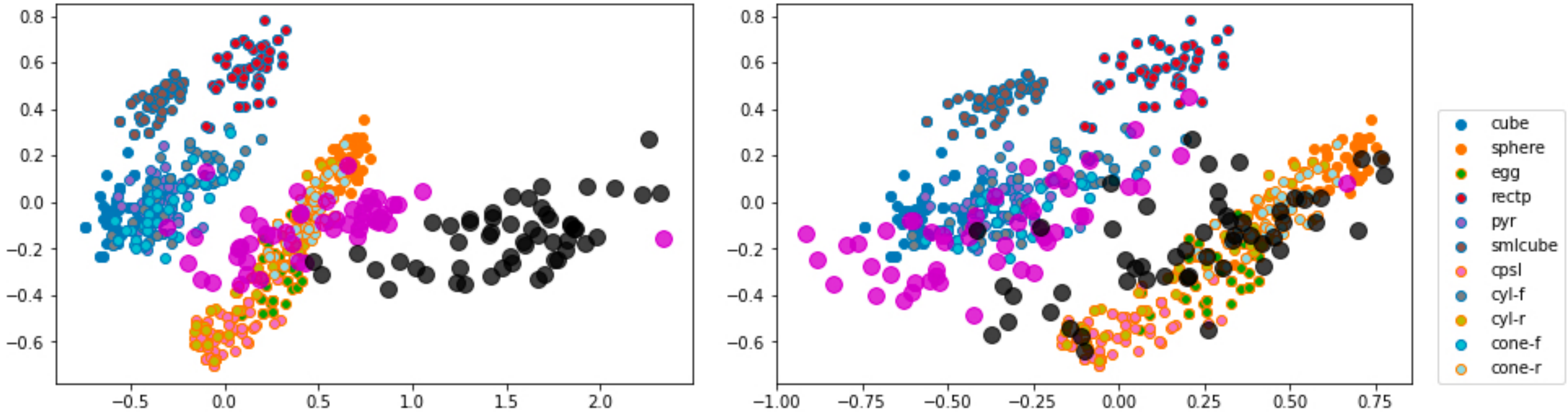
- Clusters of transformed abstract concept vocabulary separate when transformation includes object words
    - but different models representations behave differently



BERT: “flat” vs. “round” (black) without hinting [L] and with [R]

Language Grounding to Environment

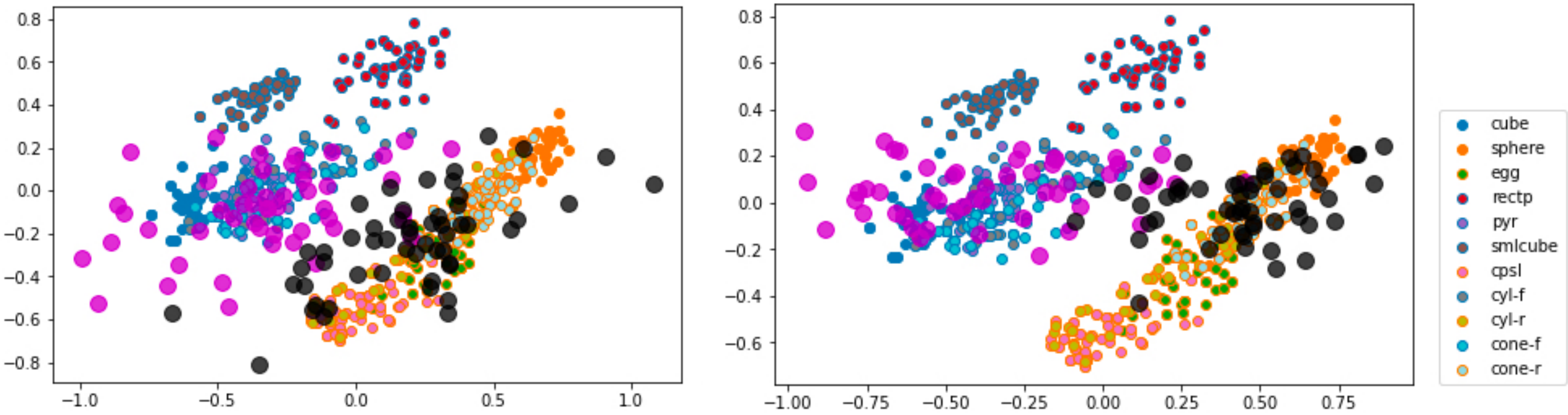
- Clusters of transformed abstract concept vocabulary separate when transformation includes object words
  - but different models representations behave differently



RoBERTa: "flat" vs. "round" (black) without hinting [L] and with [R]

Language Grounding to Environment

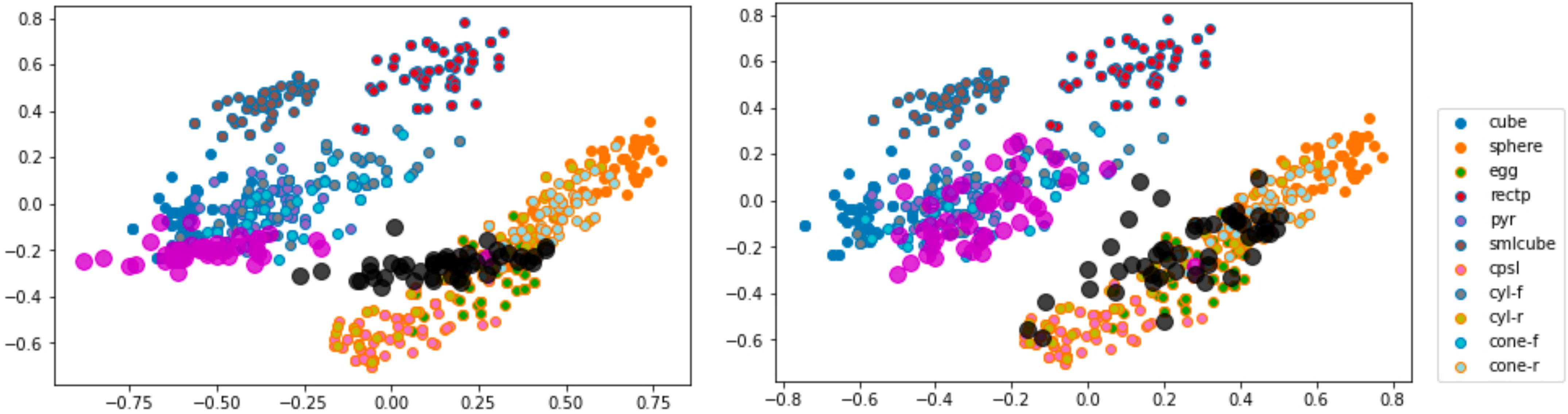
- Clusters of transformed abstract concept vocabulary separate when transformation includes object words
  - but different models representations behave differently



ALBERT: “flat” vs. “round” (black) without hinting [L] and with [R]

Language Grounding to Environment

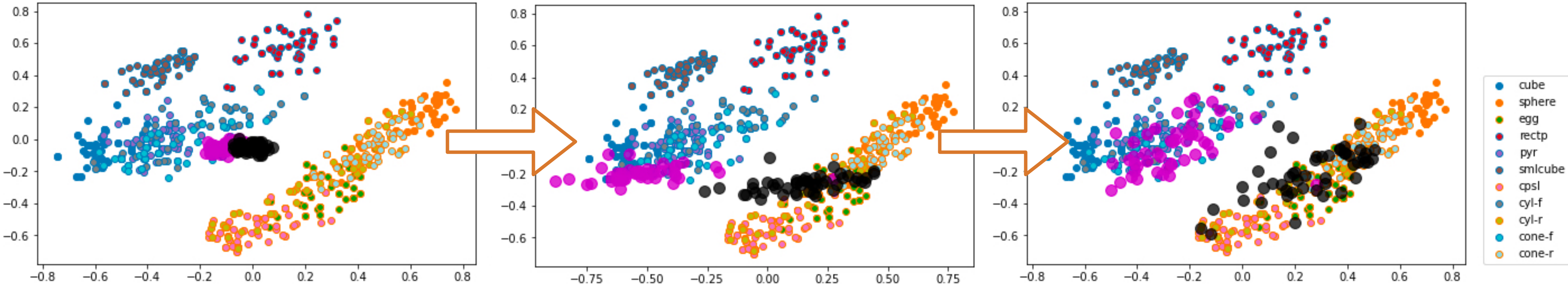
- Clusters of transformed abstract concept vocabulary separate when transformation includes object words
  - but different models representations behave differently



XLM: “flat” vs. “round” (black) without hinting [L] and with [R]

Language Grounding to Environment

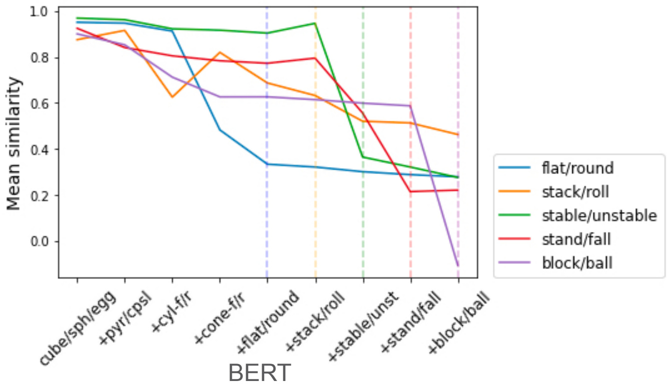
- How does addition of information to the transformation affect the similarity of concept clusters?



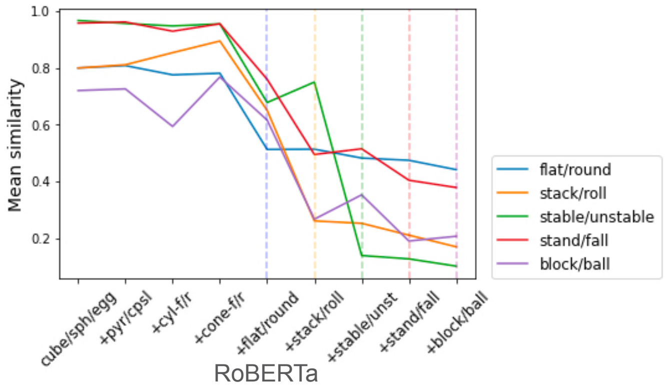
XLM: “flat” (pink) vs. “round” (black)

Language Grounding to Environment

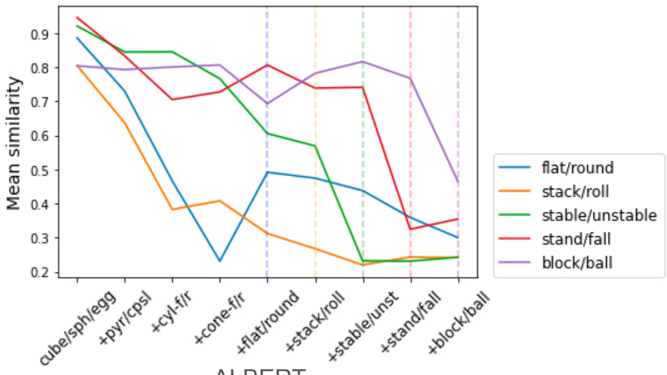
- Separation of cluster centers (object terms first)



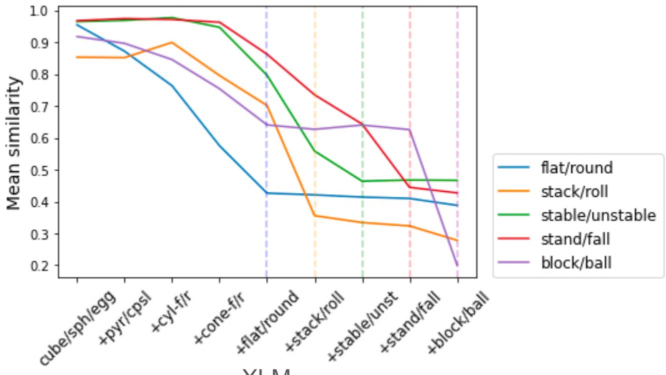
BERT



RoBERTa



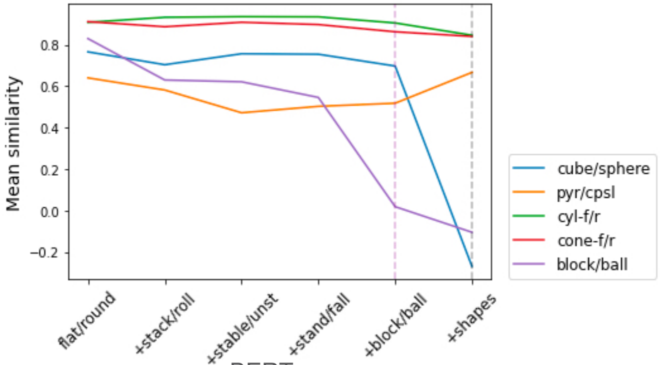
ALBERT



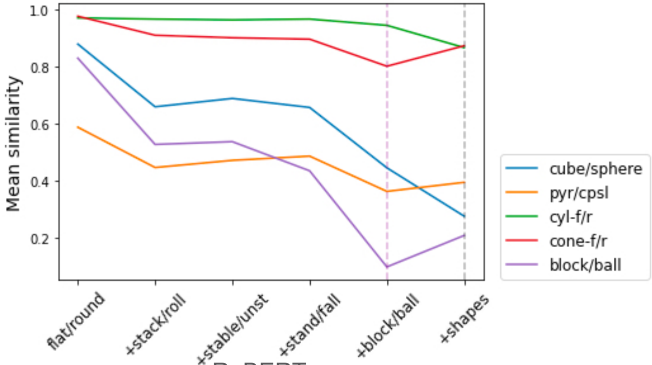
XLM

Language Grounding to Environment

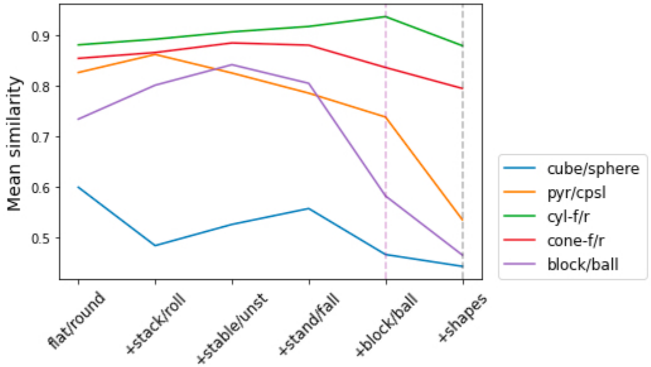
- Separation of cluster centers (abstract terms first)
- Grounding object terms helps distinguish related actions or properties (cf. Gentner (2006))
- But grounding abstract terms isn't as helpful in making distinctions between object terms!



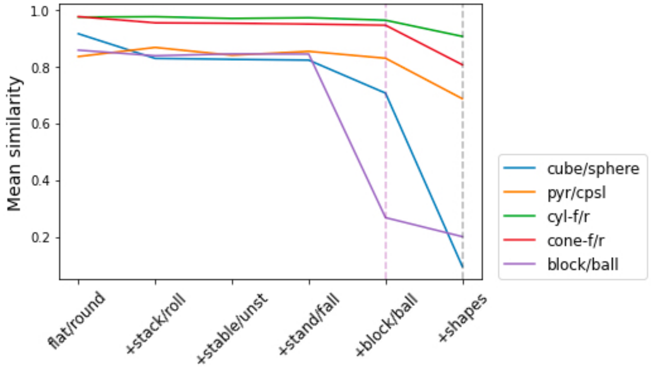
BERT



RoBERTa



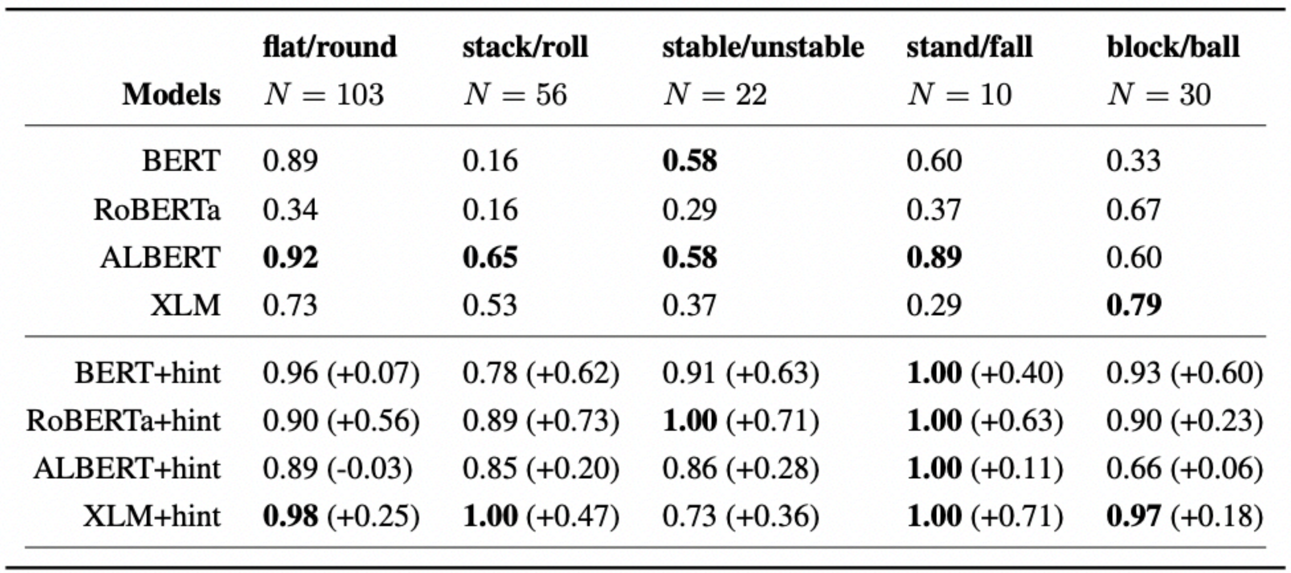
ALBERT



XLM

Language Grounding to Environment

- Formulation as classification task: KNN (k=5) after all object/abstract terms are grounded
- Do transformed word vectors for object-related concept terms cluster with the correct set of objects?
- XLM (largest model): best with hinting
- XLM’s larger training and embedding size may make it better able to represent multiple word senses
- ALBERT (smallest model): best without hinting

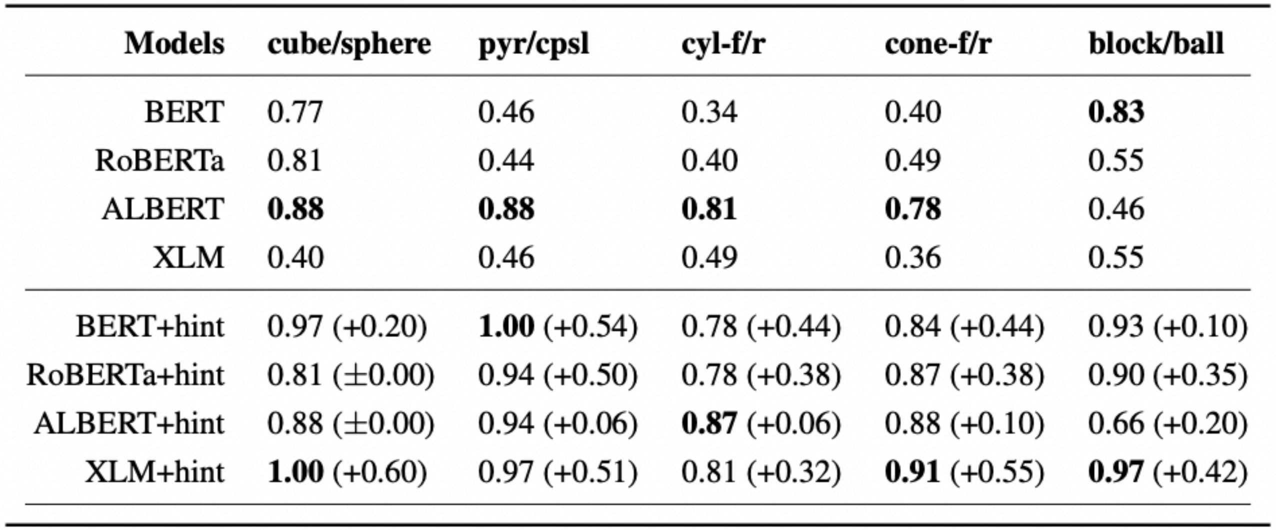


	flat/round	**stack/roll**	**stable/unstable**	**stand/fall**	**block/ball**
Models	$N = 103$	$N = 56$	$N = 22$	$N = 10$	$N = 30$
BERT	0.89	0.16	**0.58**	0.60	0.33
RoBERTa	0.34	0.16	0.29	0.37	0.67
ALBERT	**0.92**	**0.65**	**0.58**	**0.89**	0.60
XLM	0.73	0.53	0.37	0.29	**0.79**
BERT+hint	0.96 (+0.07)	0.78 (+0.62)	0.91 (+0.63)	**1.00** (+0.40)	0.93 (+0.60)
RoBERTa+hint	0.90 (+0.56)	0.89 (+0.73)	**1.00** (+0.71)	**1.00** (+0.63)	0.90 (+0.23)
ALBERT+hint	0.89 (-0.03)	0.85 (+0.20)	0.86 (+0.28)	**1.00** (+0.11)	0.66 (+0.06)
XLM+hint	**0.98** (+0.25)	**1.00** (+0.47)	0.73 (+0.36)	**1.00** (+0.71)	**0.97** (+0.18)

Table 1: Macroaveraged KNN F1 over transformed attribute/verb/synonym word embedding test sets (mapping computed using object embeddings). Numbers in parentheses show performance increase with “hinting.”

Language Grounding to Environment

- Hypothesis:
  - XLM representation is less “entangled,” more compositional
  - ALBERT representation is more “entangled,” possible bias toward noun-verb/adj. correlations
- Object words and related terms may already have some overlap in ALBERT space
- Grounding concepts to a physical environment without explicit nudging may be more challenging for larger models than smaller ones



Models	cube/sphere	pyr/cpsl	cyl-f/r	cone-f/r	block/ball
BERT	0.77	0.46	0.34	0.40	**0.83**
RoBERTa	0.81	0.44	0.40	0.49	0.55
ALBERT	**0.88**	**0.88**	**0.81**	**0.78**	0.46
XLM	0.40	0.46	0.49	0.36	0.55
BERT+hint	0.97 (+0.20)	**1.00** (+0.54)	0.78 (+0.44)	0.84 (+0.44)	0.93 (+0.10)
RoBERTa+hint	0.81 (±0.00)	0.94 (+0.50)	0.78 (+0.38)	0.87 (+0.38)	0.90 (+0.35)
ALBERT+hint	0.88 (±0.00)	0.94 (+0.06)	**0.87** (+0.06)	0.88 (+0.10)	0.66 (+0.20)
XLM+hint	**1.00** (+0.60)	0.97 (+0.51)	0.81 (+0.32)	**0.91** (+0.55)	**0.97** (+0.42)

Table 2: Macroaveraged KNN F1 over transformed object word embedding test sets (mapping computed using attribute/verb embeddings). Numbers in parentheses show performance increase with “hinting.” $N = 30$ for all.

Conclusion and Future Work

Conclusion

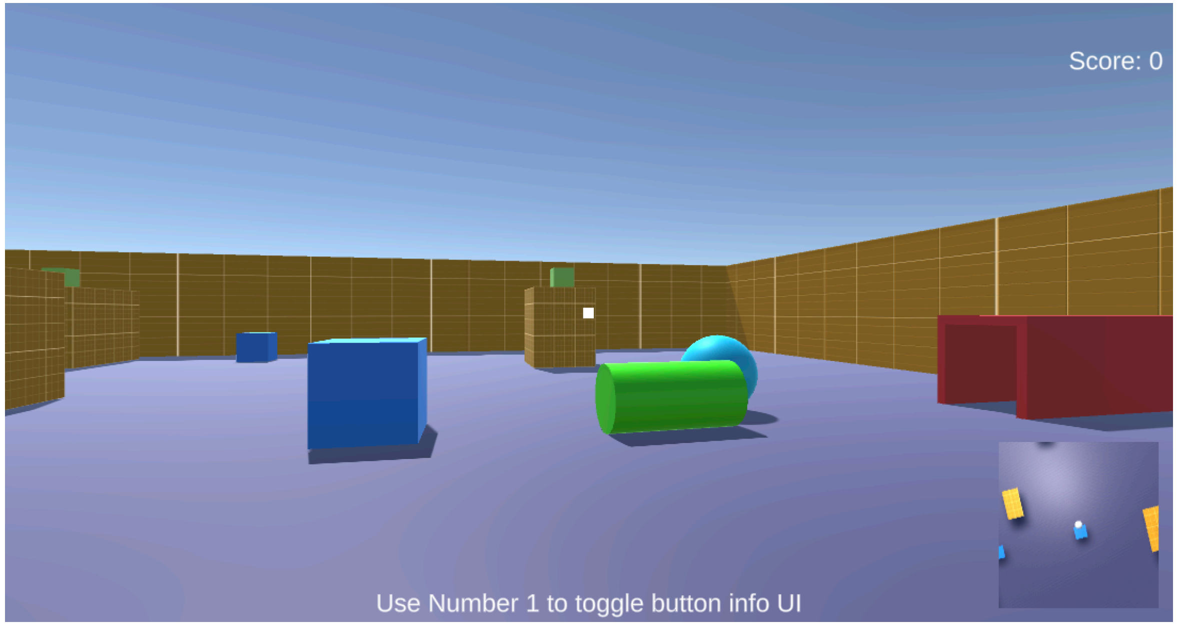
- **Similarity learning over rich data from an embodied simulation create a representation space that..**
  - **successfully classifies concrete objects**
  - **make analogical comparisons based on abstract properties that inhere across multiple object types**
- **Used the resulting representation space to investigate the properties of different LLMs regarding object and concept representations**
- **Simple ridge regression preserves interchangeability across modalities**
  - **Technique has previously been used in vision-only (McNeely-White et al., 2022) and language-only conditions (Nath et al., 2022)**
- **Can use these techniques to make AI models behave similarly to human learning, or to examine the properties of AI models themselves**
  - (Not saying humans use the same technique)

Conclusion

- **Our embodied approach allows us to build a model without visual artifacts like occlusion**
- **Embodiment is influenced by factors like events and habitats (Pustejovsky and Krishnaswamy, 2022)**
- **Purely linguistic representations of tokens may not capture these factors**
    - **ChatGPT-generated corpus is likely *not* representative of these aspects**
- **Our embodied approach enables correlating representations extracted from unembodied models to representations learned from embodied data**
    - **Provides evidence that the ability to ground real-world entities, properties, or actions to lexical items could enable LLMs to simulate the human ability to link utterances to specific communicative intents**

Future Work

- Investigating different orders in transformation
- Incorporating images
- Other embodied tasks to investigate other concepts
- Evaluating representations directly from a GPT-like decoder
- Having an agent learn such correlations in the environment in real time



Thanks to…

DEVCOM
ARMY RESEARCH
LABORATORY

SCIENCE TO SHAPE THE FUTURE OF THE ARMY
ARMY RESEARCH OFFICE
1951

NSF
National Science Foundation

Thank you!

{sadafgh,nkrishna}@colostate.edu
https://signallab.ai

Colorado State University

References

- Renee Baillargeon. 1987. Object permanence in 3½- and 4½-month-old infants. Developmental psychology, 23(5):655.
- Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 55–65, Hong Kong, China. Association for Computational Linguistics.
- Dedre Gentner. 2006. Why verbs are hard to learn. *Action meets word: How children learn verbs*, pages 544–564.
- Sadaf Ghaffari and Nikhil Krishnaswamy. 2022. Detecting and accommodating novel types and concepts in an embodied simulation environment. In *Proceedings of the Tenth Annual Conference on Advances in Cognitive Systems*.
- David McNeely-White, Ben Sattelberg, Nathaniel Blanchard, and Ross Beveridge. 2022. Canonical face embeddings. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(2):197–209.
- Abhijnan Nath, Rahul Ghosh, and Nikhil Krishnaswamy. 2022. Phonetic, semantic, and articulatory features in Assamese-Bengali cognate detection. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 41–53.
- James Pustejovsky and Nikhil Krishnaswamy. 2022. Multimodal semantics for affordances and actions. In Human-Computer Interaction. Theoretical Approaches and Design Methods: Thematic Area, HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part I, pages 137–160. Springer.

References

- Elizabeth S. Spelke. 1985. Perception of unity, persistence, and identity: Thoughts on infants’ conceptions of objects.
- Elizabeth S. Spelke. 1990. Principles of object perception. *Cognitive science*, 14(1):29–56.
- Elizabeth S. Spelke, Claes von Hofsten, and Roberta Kestenbaum. 1989. Object perception in infancy: Interaction of spatial and kinetic information for object boundaries. *Developmental Psychology*, 25(2):185.
- Elizabeth S. Spelke, Claes von Hofsten, and Roberta Kestenbaum. 1989. Object perception in infancy: In- teraction of spatial and kinetic information for object boundaries. *Developmental Psychology*, 25(2):185.