Grounding Verbs in Intuitive Physics:
Few-Shot Categorization of Action Events

Jingyi Wu[1,2], Enjie Xu[1,2], Ella Bitterman[3], Joshua Tenenbaum[4], Joshua Hartshorne[3], Tao Gao[1]

[1]UCLA [2]Zhejiang University [3]MGH Institute of Health Professions [4]MIT

UCLA Visual Intelligence Lab

HSP 2026 Language and Thought — Massachusetts Institute of Technology, Conference on Human Sentence Processing, Cambridge, MA

A Longstanding Debate in Language Study

Relation Between Language and Thought

Language determinism ⟷ (linguistic relativity / instrumental view) ⟷ Thought determinism

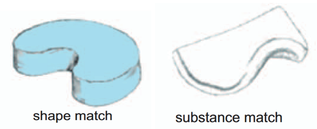
Methods to Explore the Relation

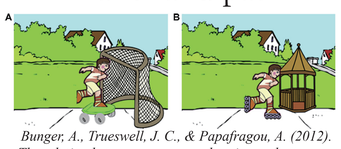
- **Cross-linguistic variation** — Comparisons of different language speakers' understaning of the world
- **Pre-linguistic cognition** — Basic concept understanding in infants and animals
- **Vision-Language Consistency** — Examining shared representations in vision and language

a key research topic in both cognitive science and current AI research

Parallel Structure in Language and Vision

count and mass — maner and path





Barner, D., Li, P., & Snedeker, J. (2010). Words as windows to thought: The case of object representation. Current Directions in Psychological

Bunger, A., Trueswell, J. C., & Papafragou, A. (2012). The relation between event apprehension and utterance formulation in children: Evidence from linguistic omissions. Cognition, 122(2), 135-149.

Grounding Verbs in Vision

Verb Argument Structure Is Shaped by Subtle, Nonlinguistic Physical Properties

=> Certain verbs resist argument structures shared by similar verbs.

Example 1: Perceptual Constraint

1.Someone poured the water into the glass.
2.Someone filled the glass with water.
3.Someone poured the glass with water. (unsuitable)
4.Someone filled the water into the glass. (unsuitable)

"Pour" → directional motion
"fill" → result state

Example 2: Dative Alternation

1.Danielle brought the cat to her mother. ("causing-to-go")
2.Danielle brought her mother the cat. ("causing-to-have")
3.Danielle lifted the crate to him. ("causing-to-go")
4.Danielle lifted him the crate. (unsuitable)

"lift"→ continuous force do not encode transfer and caused possession.

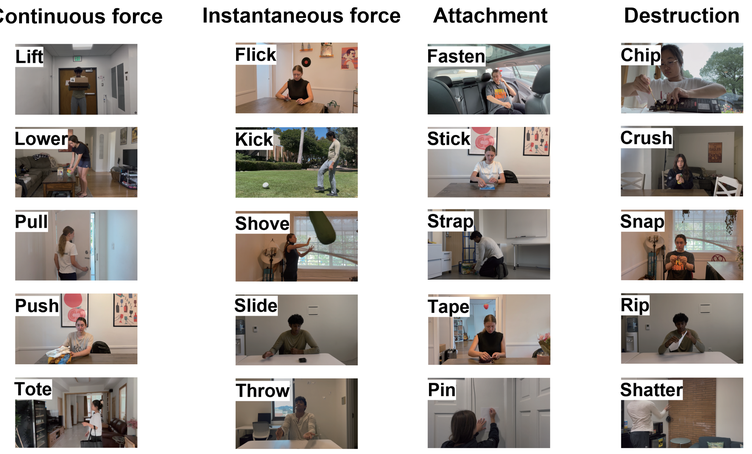
=> Humans tend to categorize these verbs based on the nonlinguistic physical causal properties.

Our Prediction:

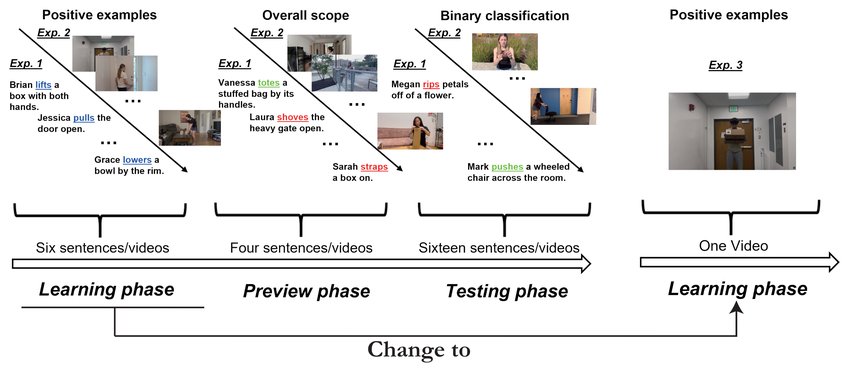
Do physical causal properties that shape verb argument structure also guide few-shot learning of event categories in non-linguistic visual tasks?

Expt. 1-3: Few-Shot (Vision & Language) and One-Shot (Vision) Learning

Four Classes of Verbs



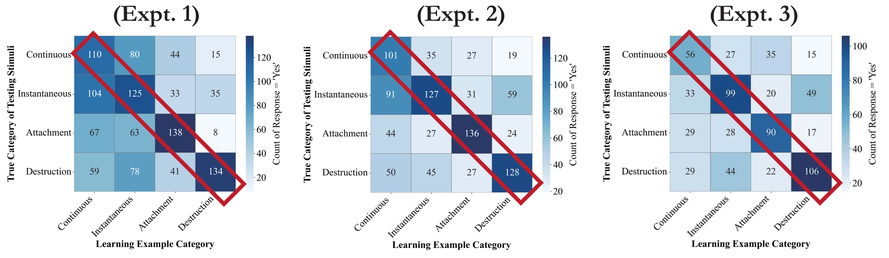
Few-shot Learning



One-shot Learning

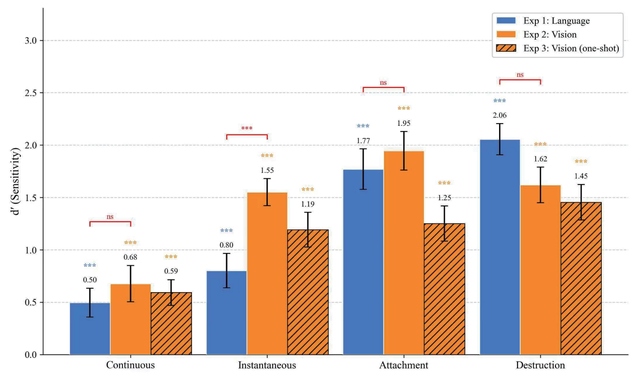
Confusion Matrix of ("Yes") Responses

Few-shot learning in language (Expt. 1) — Few-shot learning in vision (Expt. 2) — One-shot learning in vision (Expt. 3)



=> highest accuracy along the diagonal

Overal Sensitivity (d')



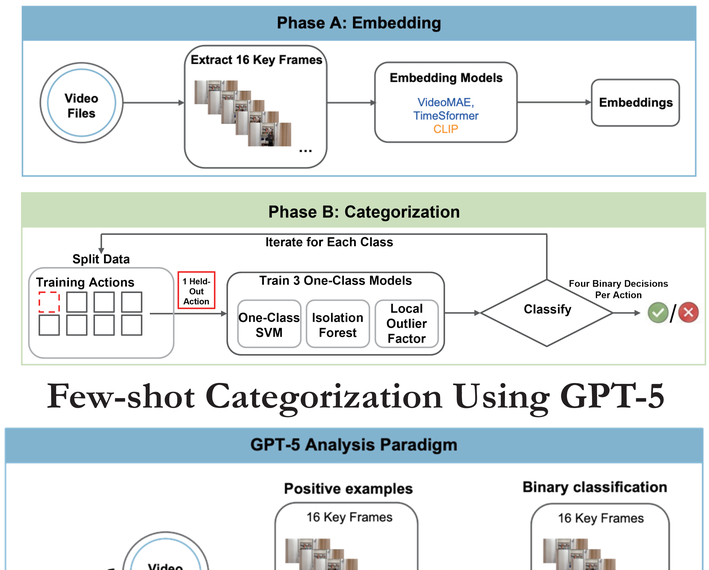
=> all sensitivity values above chance in four verb classes

=> language and vision show similar patterns

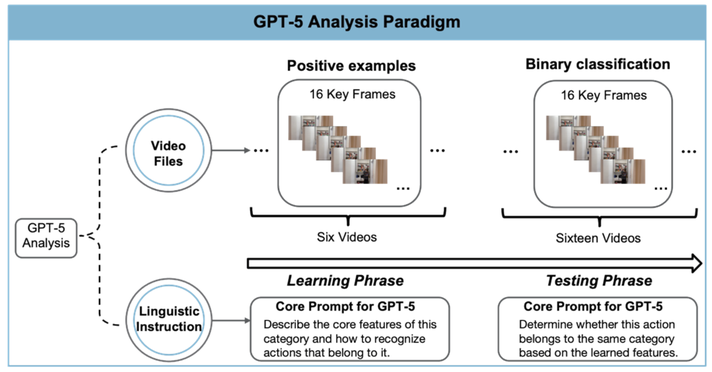
Humans reliably perform few-shot learning and generalize event categories across both language and vision; vision even supports one-shot learning.

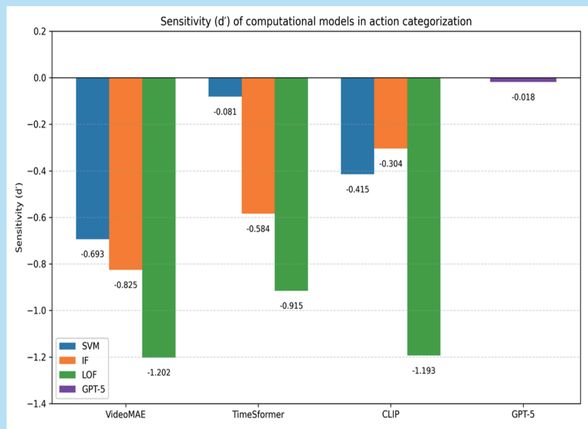
Expt. 4: Computational Vision Models Baseline

Categorization Based on Models' Embeddings



Few-shot Categorization Using GPT-5





All models show lower performance compared to chance

Conlusion

1. Humans successfully learned and generalized event categories with few-shot learning (even with one-shot learning in vision).

2. Humans showed parallel patterns of evevnt categorization in both language and vision.

3. Current computational vision models lack the human-like event representation.

=> Suggesting that human have shared, structured event representations across language and vision, which current AI systems do not capture.