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

A long tradition in the cognitive sciences posits that language, perception, and thought are grounded in the same underlying conceptual representations (Fodor, 1975; Talmy, 1988; Pinker, 1989; Jackendoff, 1992). The structure of this underlying system has been argued to shape the organization of language, constrain cross-linguistic variation, and provide an entry point for children acquiring language. However, most existing evidence comes from language itself, particularly from research on the semantics of argument structure and its acquisition (Pinker, 1989; Hartshorne et al., 2016). By contrast, evidence that non-linguistic systems are structured in similar ways is far more limited, focusing primarily on whether perceptual processes exhibit analogs of thematic roles such as agent or patient (Gao et al., 2010; Strickland, 2017; Papeo et al., 2024). In the present study, we test whether visual reasoning draws on the same conceptual distinctions that Pinker (1989), Jackendoff (1992), and others have argued underlie verb argument structure. If these distinctions are shared across systems, then humans should find it easier to infer novel categories that align with them. This hypothesis intersects with ongoing debates about the degree to which argument structure is semantically structured and about which semantic distinctions are most relevant (Levin & Rappaport Hovav, 2005).

**Experiment 1** examined whether humans can spontaneously categorize novel action videos from a few positive examples. We tested four classes of verbs with different argument structures: (1) continuous force; (2) instantaneous force; (3) attachment via contact; and (4) destruction of an object. Participants first viewed six videos depicting three verbs from a single verb class, with no linguistic input. They were then shown new videos depicting unseen verbs from either the same class or other classes and judged whether each action belonged to the learned category. We used a between-subjects design with 80 adults from English-speaking countries evenly assigned to four conditions. Despite minimal training and no negative examples, participants achieved 77.1% accuracy and strong sensitivity ($d' = 1.28$). They exhibited a high hit rate (73.3%) and a relatively low false-alarm rate (25.6%), indicating robust discrimination among the action types. The results broadly support our hypothesis, though they also revealed an unexpected reduction in accuracy for the continuous-force class (64.1%).

**Experiment 2** followed the same paradigm as Experiment 1 but presented English sentences instead of videos, with each sentence explicitly using the target verb from the assigned class. This language-based version tested how strongly few-shot learning depends on visual representations. Participants successfully categorized actions from linguistic input, achieving 68.5% accuracy and solid sensitivity ($d' = 1.20$). However, performance declined for the continuous- and instantaneous-force classes, which showed markedly reduced sensitivity ($d' = .49$ and .46, respectively) compared to the visual task. These results suggest that perceptual input contributes uniquely to action verb categorization, and linguistic descriptions alone may be more indistinct.

To evaluate whether current models can perform this task without domain-specific knowledge, we tested the Experiment 1 paradigm on three vision models (VideoMAE, CLIP, and TimeSformer). All models performed poorly, indicating that they do not yet capture the underlying conceptual distinctions humans rely on. Taken together, our findings bridge linguistics with visual cognition, supporting the view that language is grounded in domain-general conceptual representations. We discuss implications for children's acquisition of sentence structure and for patterns of cross-linguistic variation.


**Figure 1 Overview of the verb classes and selected verbs for experiments, and examples of videos.**

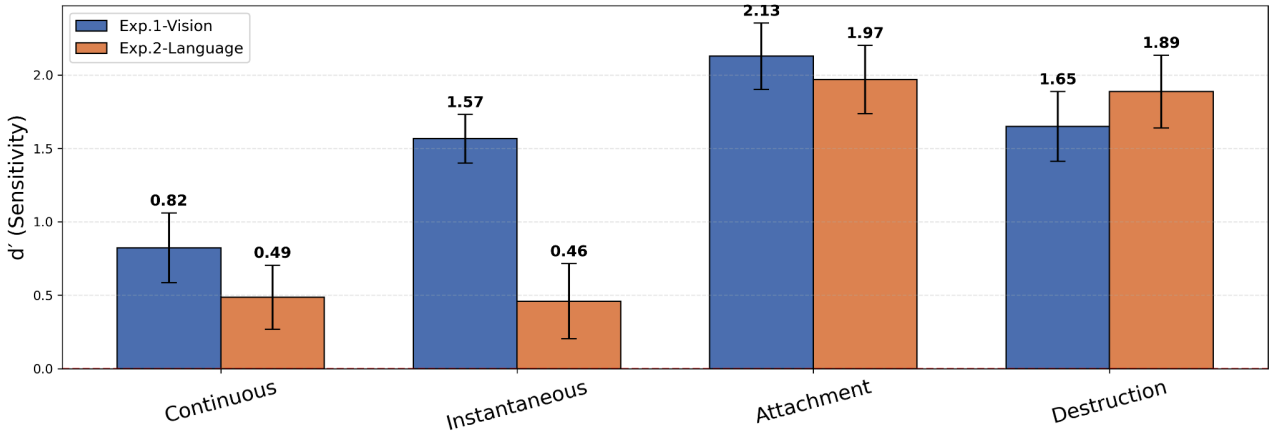


**Figure 2 Sensitivity (d') across the four verb classes in Experiment 1 (vision) and Experiment 2 (language)**

**Table 1 Comparison between humans and vision models**

| | d' | Hit rate | False alarm rate |
|---|---|---|---|
| **Humans** | **1.28** | **77.3%** | **23.4%** |
| VideoMAE | -0.69 | 7.8% | 10.9% |
| CLIP | -0.42 | 5.0% | 63.8% |
| TimeSformer | -0.08 | 0 | 0.6% |

**References:**

Fodor (1975) *The language of thought*; Gao et al. (2010) *Psychol Sci;* Hartshorne et al. (2016) *Cognition;* Jackendoff (1992) *Semantic Structures;* Levin & Rappaport Hovav (2005) *Argument Realization;* Papeo et al. (2024) *Curr Biol;* Pinker (1989) *Learnability and Cognition;* Pinker (2007) *The Stuff of Thought;* Strickland (2017) *Cogn Sci;* Talmy (1988) *Cogn Sci.*