# MEWL: Few-shot multimodal word learning with referential uncertainty

Guangyuan Jiang[1,2], Manjie Xu[2,3], Shiji Xin[1], Wei Liang[3], Yujia Peng[1,2], Chi Zhang[2], Yixin Zhu[1]
[1]Peking University [2]Beijing Institute for General Artificial Intelligence [3]Beijing Institute of Technology

**Paper, code, and data are available:**
https://github.com/jianggy/MEWL

## Motivation: How children learn words


- Learn words from **cross-situational** information from multiple contexts.
- Leverage semantic and syntactic cues to **bootstrap** novel word learning.
- Comprehend word meanings with **pragmatics**, a social account of word learning with the help of other speakers.

## Motivation: Human-like word learning

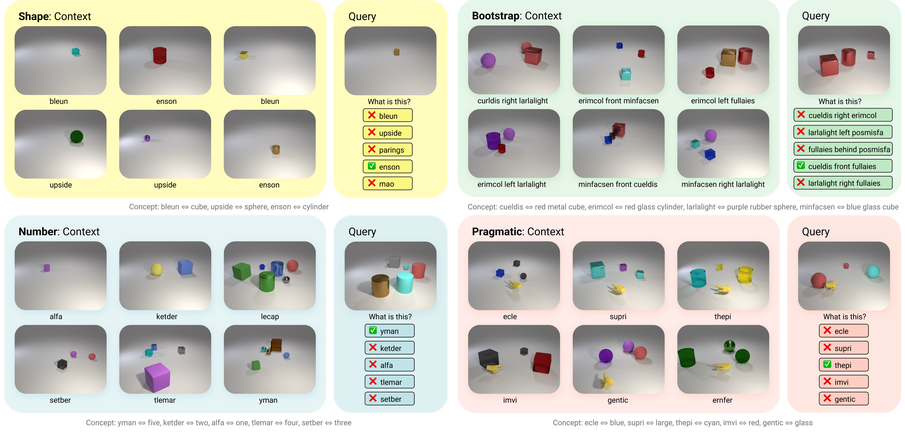
- **Word learning** facilitates critical **downstream tasks**:
  - learning new object categories
  - forming abstractions of conceptual structures
  - making generalizations
  - developing the ability to communicate
- It is the very **first step** in language learning:
  - born **multi-modal**: children need to ground every word to visual perceptions and relations.
  - closely related to the **learning of concepts**.
  - children can use many clues to facilitate word learning**: cross-situational statistics**, **bootstrapping**, and **pragmatic** word learning.

## Our contributions

| | multimodal | few-shot | uncertainty | relation | pragmatic | human |
|---|---|---|---|---|---|---|
| CLEVR (Johnson et al., 2017) | ✗ | ✗ | ✗ | ✓ | ✗ | ✓ |
| RAVEN (Zhang et al., 2019a) | ✗ | ✓ | ✗ | ✓ | ✗ | ✓ |
| NLVR (Suhr et al., 2017) | ✓ | ✗ | ✗ | ✓ | ✗ | ✓ |
| KiloGram (Ji et al., 2022) | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ |
| CURI (Vedantam et al., 2021) | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ |
| Fast VQA (Tsimpoukelli et al., 2021) | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ |
| MEWL (ours) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

- Highlight the significance of human-like word learning in machines.
- Devising and benchmarking MEWL's nine tasks, all directly inspired by the established findings in human word learning, for probing and comparing few-shot word learning capabilities in machines and humans.

## MEWL tasks




### Task design:

- Learn novel words or phrases that represent **basic object attributes** (i.e., `shape`, `color`, and `material`), the objects per se (i.e., `object`), and the composition of basic attributes (i.e., `composite`).
- Use familiar words to **bootstrap** learning novel (spatial) **relational** words (i.e., `relation`) or vice versa (i.e., `bootstrap`).
- Learn **counting** and **number** words from one to six (i.e., `number`).
- Use **pragmatic cues** to learn novel words by assuming the speaker is informative (i.e., `pragmatic`).

- Each few-shot problem is an episode consisting of seven images, each containing a few randomly positioned objects and has an utterance consisting of a novel word/phrase describing the image.
- After seeing context images, a query image is presented with five candidate utterances, with one answer that correctly describes the scene.

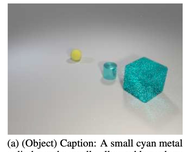
- 27,000 problems for training
- 5,400 problems for validation
- 5,400 problems for testing
- evenly divided among the nine tasks

## Results

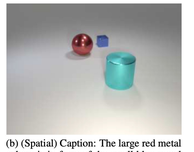
### Models:

- CLIP
- Aloe
- Flamingo-1.1B
- BERT
- GPT-3.5
- Human

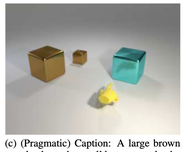
### Captioning for unimodal models:



(a) (Object) Caption: A small cyan metal cylinder and a small yellow rubber sphere and a large cyan glass cube.

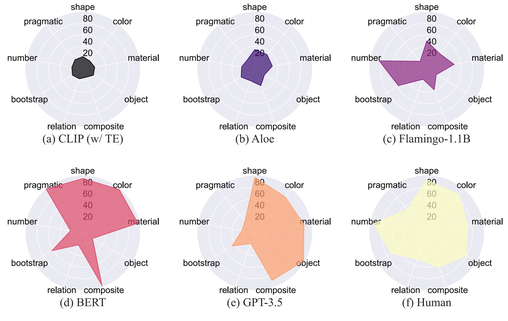


(b) (Spatial) Caption: The large red metal sphere is in front of the small blue metal cube and behind the large cyan metal cylinder. The small blue metal cube is on the left of the large cyan metal cylinder and on the right of the large red metal sphere.



(c) (Pragmatic) Caption: A large brown metal cube and a small brown metal cube and a large cyan metal cube and a small yellow rubber arrow. And a finger is pointing to the large cyan metal cube.

| Models | shape | color | material | object | composite | relation | bootstrap | number | pragmatic | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| CLIP (w/o TE) | 16.2 | 18.0 | 19.3 | 17.0 | 22.2 | 20.8 | 18.7 | 19.2 | 20.2 | 19.1 |
| CLIP (w/ TE) | 22.0 | 18.8 | 21.0 | 21.2 | 15.0 | 17.8 | 21.0 | 19.5 | 21.5 | 19.8 |
| Aloe | 34.2 | 33.2 | 31.0 | 19.5 | 30.5 | 21.5 | 27.5 | 23.3 | 20.8 | 26.8 |
| Flamingo-1.1B | 49.3 | 35.3 | 48.5 | 19.2 | 38.2 | 18.8 | 57.3 | 84.2 | 18.0 | 41.0 |
| BERT | 94.8 | 98.8 | 97.5 | 19.5 | 97.8 | 22.2 | 62.2 | 21.8 | 99.8 | 68.3 |
| GPT-3.5 | 96.8 | 82.3 | 87.0 | 98.2 | 88.3 | 20.0 | 45.8 | 22.7 | 26.7 | 63.1 |
| Human | 92.4 | 87.2 | 72.7 | 79.1 | 63.5 | 48.7 | 71.0 | 93.9 | 54.8 | 73.7 |




### Discussion:

- Multimodal *vs.* unimodal
- Humans *vs.* machines
  - Efficacy of MEWL
  - Failure of learning models
  - Why should machines have human-like word learning capabilities?