MEWL: Few-shot multimodal word learning with referential uncertainty

Guangyuan Jiang [1 2 3] **Manjie Xu** [3 4] **Shiji Xin** [5] **Wei Liang** [4 6] **Yujia Peng** [1 3 7] **Chi Zhang** [3] **Yixin Zhu** [1]

Abstract

Without explicit feedback, humans can rapidly learn the meaning of words. Children can acquire a new word after just a few passive exposures, a process known as fast mapping. This word learning capability is believed to be the most fundamental building block of multimodal understanding and reasoning. Despite recent advancements in multimodal learning, a systematic and rigorous evaluation is still missing for human-like word learning in machines. To fill in this gap, we introduce the MachinE Word Learning (MEWL) benchmark to assess how machines learn word meaning in grounded visual scenes. MEWL covers human's core cognitive toolkits in word learning: cross-situational reasoning, bootstrapping, and pragmatic learning. Specifically, MEWL is a few-shot benchmark suite consisting of nine tasks for probing various word learning capabilities. These tasks are carefully designed to be aligned with the children's core abilities in word learning and echo the theories in the developmental literature. By evaluating multimodal and unimodal agents' performance with a comparative analysis of human performance, we notice a sharp divergence in human and machine word learning. We further discuss these differences between humans and machines and call for human-like few-shot word learning in machines.

Code and data: https://github.com/jianggy/MEWL. [1]Institute for AI, Peking University [2]Yuanpei College, Peking University [3]National Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence [4]School of Computer Science & Technology, Beijing Institute of Technology [5]School of EECS, Peking University [6]Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing [7]School of Psychological and Cognitive Sciences, Beijing Key Laboratory of Behavior and Mental Health, Peking University. Correspondence to: Guangyuan Jiang <jgy@stu.pku.edu.cn>, Chi Zhang <zhangchi@bigai.ai>, Yixin Zhu <yixin.zhu@pku.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

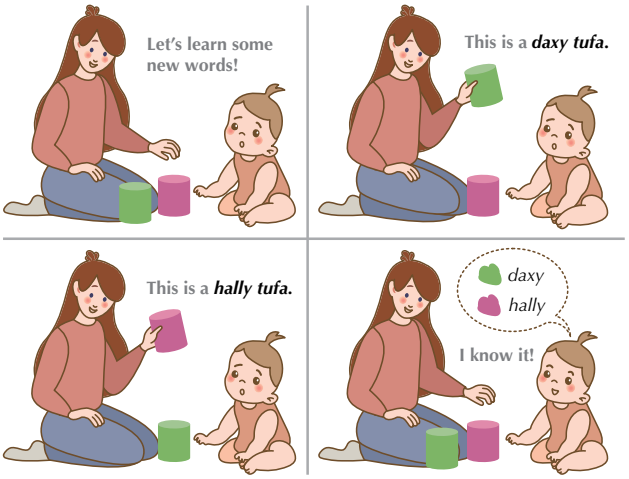


Figure 1: **Illustration of few-shot word learning.** Children can acquire a novel word after only few exposures using cross-situational information, even with referential uncertainty. In this example, a child induces that *daxy* refers to the color green and *hally* magenta, all from the experience of a *daxy tufa* (green cylinder) and a *hally tufa* (magenta cylinder) without explicit guidance.

1. Introduction

Learning words and a language is one of the most fundamental stages of human cognitive development, serving as the foundation for other crucial capabilities that come later, such as learning new object categories, forming abstractions of conceptual structures, making generalizations, and developing the ability to communicate (Lake & Murphy, 2021; Murphy, 2004; Smith & Gasser, 2005; Tenenbaum et al., 2011). Remarkably, we acquire the meaning of words rapidly and effortlessly, even without explicit feedback (Bloom, 2001). One striking observation is that young children can understand a novel word's meaning merely from a few examples, also known as fast mapping (Carey & Bartlett, 1978; Heibeck & Markman, 1987); a child can learn about 12 words per day by the age of eight (Bloom, 2002). These quickly learned words constitute our understanding of the world and the basis of symbol representation for concepts.

Human learning is inherently few-shot and open-ended, even without explicit guidance (Landau et al., 1988; Lake et al., 2015). Children experience substantial referential ambiguity while learning new words, yet they are nevertheless

able to comprehend the word-referent mappings; see Figure 1 for an illustration. How can we learn so many words from so little? Prior developmental studies indicate these capabilities come in many ways.

- We learn words from co-occurrence in multiple contexts (Scott & Fisher, 2012). Children are young statisticians (Gopnik et al., 1999; Abdulla, 2001); they employ **cross-situational** statistics (Smith & Yu, 2008) and perform Bayesian-like inference (Tenenbaum, 1998) to understand the meaning of words from multiple scenes (Xu & Tenenbaum, 2007).
- We leverage semantic and syntactic cues to **bootstrap** novel word learning (Quine, 1960; Pinker, 2009). For instance, we can use familiar relational words to infer an unknown word's meaning: hearing *beef and dax*, we may infer that *dax* is a noun and edible; it may represent a similar kind of food to *beef*.
- We comprehend word meanings with **pragmatics**, a social account of word learning with the help of other speakers. The fundamental premise is to leverage informative descriptions of the referent (Frank & Goodman, 2014; Horowitz & Frank, 2016; Stacy et al., 2022). For example, if we have a *blue cube*, a *blue ball*, and a *green cube* in a line, a speaker would use the word "*ball*" to refer to the object in the middle, which is the most informative word to tell them apart (Frank & Goodman, 2012).

Human-like word learning is quintessential towards building machines that learn and reason like people (Lake et al., 2017; Zhu et al., 2020; Fan et al., 2022). Despite recent development in language-only and vision-language pre-training, it is still unknown if these models acquire word meaning in a manner similar to that of humans (Lake & Murphy, 2021; Bender & Koller, 2020; Mitchell & Krakauer, 2023). Concerns have been raised regarding the pre-training paradigm's inability to capture the core components of human language and conceptual structure, such as compositionality (Thrush et al., 2022), concept association (Yamada et al., 2022), relational understanding (Conwell & Ullman, 2022), and conceptual meaning (Piantasodi & Hill, 2022). These concerns can be linked to the differences in how humans and machines acquire the primitives of words (Fodor et al., 1988; Tenenbaum et al., 2011). To the best of our knowledge, a systematic and rigorous evaluation for human-like word learning in machines is still missing.

To fill in this gap, we devise the MachinE Word Learning (MEWL) benchmark to assess machine word learning in grounded visual scenes, covering human's core cognitive toolkits in word learning. MEWL serves as a testbed for few-shot vision-language reasoning with referential uncertainty. It includes nine tasks covering four types of scenarios: basic attribute naming, relational word learning, number word learning, and pragmatic word learning.

We build MEWL in the CLEVR universe (Johnson et al., 2017). Each MEWL problem consists of six *context images* and corresponding descriptive novel words or phrases (*i.e.*, *utterances*). Agents (either humans or learning algorithms) are tasked to rapidly understand the meaning of novel words from context and choose the option that best matches the target *query image*. These settings closely mimic children's fast cross-situational word learning (Goodman et al., 2007; Smith et al., 2011; Carey & Bartlett, 1978).

In experiments, we deploy MEWL to analyze machines' and humans' ability to perform few-shot word learning under the nine scenarios. We first benchmark machines on MEWL by analyzing multimodal (*i.e.*, pre-trained vision-language) and unimodal models (*i.e.*, Large Language Models (LLMs)). Our experimental results indicate that pre-trained vision-language models struggle to learn word meaning with only a few examples, lagging far behind what humans can do. For LLMs, we turn the word learning problem into a concept binding problem, formulated as in-context learning with images captioned into texts and utterances as labels. LLMs perform well on attribute and object naming tasks but far worse on all others. Next, we benchmark human performance on MEWL for comparison. A comparative analysis reveals misalignment between humans and machines. Finally, we analyze and compare unimodal and multimodal learning algorithms using the rubrics of human-like word learning.

This paper makes three primary contributions:

1. We highlight the significance of human-like word learning in machines. To support this claim, we devise MEWL for probing and comparing few-shot word learning capabilities in machines and humans.
2. We craft MEWL to ensure its similarity to the human counterpart in learning new words. MEWL consists of nine tasks, all directly inspired by the established findings in human word learning.
3. We present a comprehensive benchmark of multimodal and unimodal models on MEWL. A comparative analysis of the experimental results shows that large models are generally not human-like in few-shot word learning, calling for future research on building human-like machine models on word and language understanding.

2. Related work

Word learning in machines Despite extensive studies in human word learning, how machines acquire word meaning is almost untouched. Recent attempts use infants' egocentric videos in SAYCam (Sullivan et al., 2022) and deep learning methods to mimic children's word learning experience (Orhan et al., 2020; Vong et al., 2021; Rane et al., 2022; Berger et al., 2022; Vong & Lake, 2020; Frank et al., 2017;

Table 1: **Comparison between MEWL and prior arts.** We compare MEWL and related benchmarks in six dimensions: multimodality, few shot, referential uncertainty, relational reasoning, pragmatic reasoning, and human baseline.

	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)	✓	✓	✓	✓	✓	✓

Wang et al., 2022; Zhuang et al., 2021). While most of these works focus on reverse-engineering the human word learning process, few supporting benchmarks or tasks probe machines' few-shot word learning capabilities.

Notably, Horst & Hout (2016) introduces the Novel Object and Unusual Name (NOUN) dataset for experimental research. This dataset is relatively small in size (64 images); nonetheless, it supports building word learning algorithms in machines (Krishnamohan et al., 2020; Vong & Lake, 2022). In vision-language learning, Tsimpoukelli et al. (2021) introduces Fast VQA, which presents fast concept binding as a new evaluation task for few-shot vision-language models.

In comparison, the proposed MEWL benchmark has three distinctions.

1. MEWL focuses not only on basic object categories but also on attribute, relational, numerical, and pragmatic word learning, offering a significantly more comprehensive benchmark suite in human-like word learning.
2. MEWL is akin to human word learning with referential uncertainty, where cross-situational learning is required. This particular setting is almost untouched but is at the core of human-like few-shot word learning.
3. Similar to other visual reasoning tasks (Johnson et al., 2017; Barrett et al., 2018; Depeweg et al., 2018; Edmonds et al., 2018; 2019; 2020; Zhang et al., 2019a;b; 2021a;b; 2022a; Nie et al., 2020; Zhang et al., 2020; Xie et al., 2021; Vedantam et al., 2021; Xu et al., 2022; Li et al., 2022a; 2023), MEWL is light in visual perception but richer in context. It has 37,800 questions, a significantly larger benchmark suite compared to 2,500 in Fast VQA and 64 in NOUN.

In a nutshell, we regard MEWL as the first systematic and rigorous benchmark suite for machine word learning.

Human-like few-shot learning Children are few-short learners, learning the meaning of new words after merely a single or few exposures (Carey & Bartlett, 1978; Heibeck & Markman, 1987). Modern benchmarks include various few-shot reasoning problems, including the Omniglot challenge (Lake et al., 2015; 2019b), intelligence measurements (Barrett et al., 2018; Zhang et al., 2019a; Chollet, 2019; Zhang et al., 2020), Bongard problems (Depeweg et al., 2018; Nie et al., 2020), causal reasoning (Edmonds et al., 2018; Zhang et al., 2021a; Xu et al., 2022), and generalization tasks (Lake & Baroni, 2018; Lake et al., 2019a; Vedantam et al., 2021; Xie et al., 2021; Hsu et al., 2022; Li et al., 2022b; 2023).

However, these few-shot reasoning problems do not directly tackle the human-like multimodal crux in word learning. Conversely, modern multimodal abstract reasoning benchmarks (Kuhnle & Copestake, 2017; Suhr et al., 2017; Ji et al., 2022) are not few-shot by design. MEWL perfectly fills in this gap as a testbed of few-shot multimodal word learning with referential uncertainty. It requires cross-situational grounding of novel words to the learned visual concepts via bootstrapping and pragmatic reasoning. Table 1 provides a comprehensive comparison of MEWL with prior arts.

3. Creating MEWL

When creating MEWL, we draw inspiration from and correspondingly highlight these methods in human word learning: cross-situational learning, bootstrapping, and pragmatic word learning. We design nine unique tasks in MEWL to comprehensively evaluate alignment between humans and machines: `shape`, `color`, `material`, `object`, `composite`, `relation`, `bootstrap`, `number`, and `pragmatic`. These tasks cover various aspects:

- 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`).

These tasks are crafted to be aligned with the core building blocks in human word learning and echo the theories in the developmental literature (Carey & Bartlett, 1978; Pinker, 2009; Bloom, 2002; Scott & Fisher, 2012; Smith et al.,

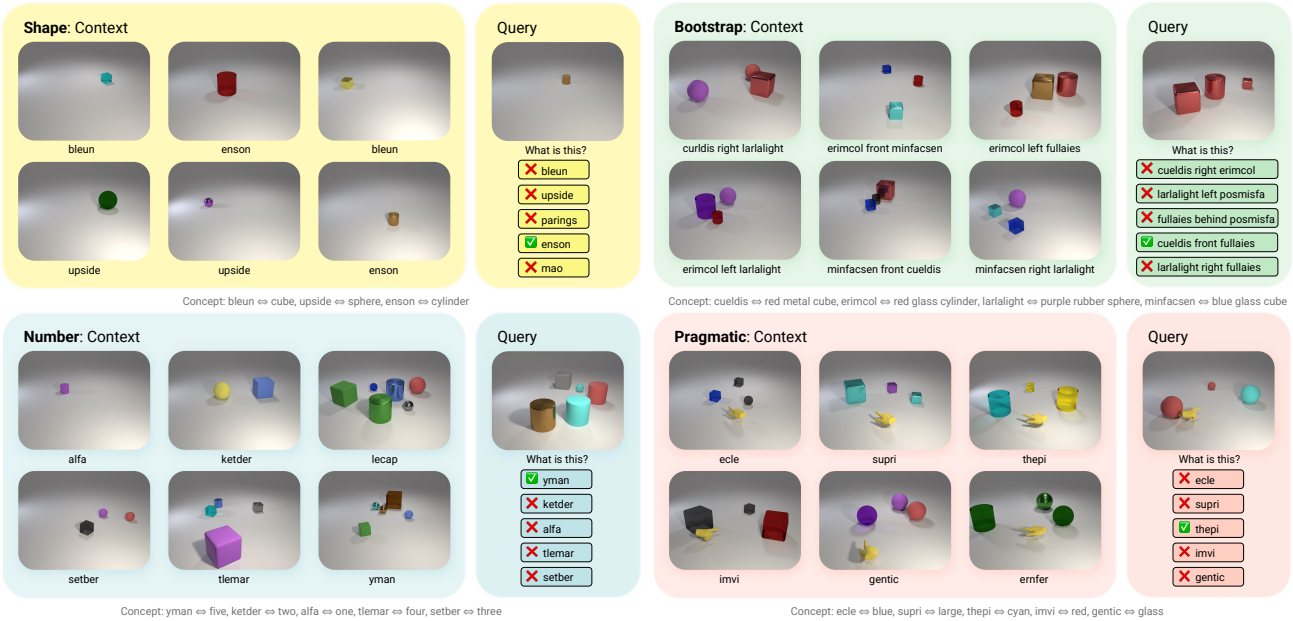


Figure 2: **Overview of the four categories of tasks in MEWL**: (i) basic naming (*e.g.*, `shape`), (ii) bootstrap relational word learning (*e.g.*, `bootstrap`), (iii) learning number words (*e.g.*, `number`), and (iv) pragmatic word learning (*i.e.*, `pragmatic`). Each episode consists of six context images and corresponding utterances. Agents need to choose the correct utterance matching the query image out of the given five options based on cross-situational reasoning from the six context panels. Ground-truth word-to-concept mappings are listed below examples (utterance ↔ concept). Please refer also to Appendix E for additional examples of the nine tasks.

2011; Horowitz & Frank, 2016; Frank & Goodman, 2014); we detail the setting of each task in Section 3.1. As such, MEWL constitutes a comprehensive suite for probing how machines learn words' meaning across various few-shot scenarios with referential uncertainty. In MEWL, all nine tasks involve referential uncertainty at varying extents and must be resolved from cross-situational disambiguation. We use the same referential uncertainty concept defined in previous word learning literature: "For any heard name, there are many candidate referents with variable perceptual properties" (Yu et al., 2021).

MEWL includes 27,000 problems for training, 5,400 problems for validation, and 5,400 problems for testing.[2] These problems are evenly divided among the nine tasks. As shown in Figure 2, each few-shot problem is an episode consisting of seven images, each containing a few randomly positioned objects. Among them are six context images; each 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, and therefore formulated as a multiple-choice problem. Following CLEVR (Johnson et al., 2017), images are rendered at a resolution of 320 × 240 with the Blender Cycles engine (Community, 2016). Apart from all CLEVR universe objects, we incorporate a glass material for more diversified textures, expanding the space for the `material` task. We also include a synthetic yellow rubber hand as the pragmatic pointer for the `pragmatic` task. We refer the readers to Appendices A and E for additional details and task examples of MEWL.

To assess the word learning capability in the context of few-shot instead of plain memorization, we create novel words unlikely to be genuine words in the English corpus across episodes. Specifically, we use the 175 most common syllables in the English language to generate more than 5 million pseudo words and associate them with the concepts in the images. Trisyllabic words are generated for `object`, `composite`, `relation`, and `bootstrap` tasks, whereas bisyllabic words are generated for `shape`, `color`, `material`, `number`, and `pragmatic`. As the words and concepts vary across episodes, the same word can be bound to different concepts in different problems; we assume mutual exclusivity: different novel words have different meanings in each episode (Merriman et al., 1989).

3.1. MEWL tasks

`shape` This task tests agents' understanding of novel words for shape concepts from the context panels. MEWL has three shapes: *cube*, *sphere*, and *cylinder*. To create word-shape mapping, we randomly assign three unique novel

[2] In theory, our environment for building MEWL allows for the creation of infinite problems. As a result, one can take advantage of this environment and train a foundation model for word learning. However, we contend that this is not how MEWL is meant to be used, as the primary objective is to examine the few-shot capability in machine word learning.

words to the shapes in each episode. When generating context, every context image has one object and a corresponding word as the utterance. Moreover, we control the object's shape to ensure it matches the utterance and leaves other properties (color, size, material) uniformly sampled. For the query panel, we choose one shape for testing. Agents are required to choose the correct word for the shape from the five options that include two distractors.

`color` Similar to the classic fast mapping experiment on children (Carey & Bartlett, 1978; Sandhofer & Smith, 1999), this task studies learning novel words representing colors. MEWL has eight distinctive colors: *gray*, *red*, *blue*, *green*, *brown*, *purple*, *cyan*, and *yellow*. We randomly sample three colors out of eight to appear in each episode. Other settings remain the same as in `shape`.

`material` We keep most of the settings unchanged from `shape` when designing this task. Instead of naming colors, we name three distinct materials: *rubber*, *metal*, and *glass*.

`object` Learning with referential uncertainty is challenging. Inspired by previous studies on infants (Smith & Yu, 2008; Smith et al., 2011; Vong & Lake, 2022), this subset of MEWL probes the agents' ability of cross-situational word learning for objects. In this task, novel words bind to the objects *per se* (*i.e.*, the quadruple of size, color, material, and shape). Unlike the `shape` task, each image in `object` has three objects and is paired with an utterance consisting of three words. Moreover, an episode has six unique word-object mappings, just within the working memory limit for humans (Miller, 1956). Because there is no one-to-one mapping from the words to objects in an image, agents must perform cross-situational reasoning to determine the correspondence between words and objects.

`composite` This task focuses on learning compositional multi-word phrases and uses syntax to bootstrap the word learning process. In detail, novel words represent basic attributes (*i.e.*, shape, color, and material), and phrases share the same syntax in an episode. The syntax can be any binary combination of the three types of attributes (*e.g.*, the first word represents color, and the second shape). We use the simplified syntax here because syntactic bootstrapping can accelerate the learning of new words (Abend et al., 2017; Gleitman, 1990). In accordance with the syntax, we selectively name two types of attributes out of three (*e.g.*, color, shape), followed by randomly choosing three instances (*e.g.*, *cyan*, *blue*, *yellow*) for each type of attribute, resulting in a lexicon size $3 + 3 = 6$. To succeed in this task, agents also need to possess systematic generalization because the answer may contain attribute combinations not shown in the context.

`relation` In this task, we probe agents' capability of learning relational words (*i.e.*, *left*, *right*, *front*, and *behind*). In humans, the understanding of spatial and temporal words is acquired later than object-centric words (Friedman & Seely, 1976). As temporal words are challenging to evaluate in most models, we only investigate spatial relation words in MEWL. To construct spatial relations, we place three objects (with one distractor) in an image and use two familiar English phrases to refer to objects and a novel spatial relation word in between to represent objects' relation (*e.g.*, "*cyan cube dax red sphere*"). We also replicate the ambiguity when children acquire spatial words. For example, "*dax*" can refer to *left* or *front* when inferring from a single image; agents must employ cross-situational reasoning to determine the exact meaning of the spatial words. We design each novel word to appear twice to ensure the problem is solvable. For example, from both *left behind* and *left front*, we understand "*dax*" means *left*. Hence, an episode only uses three spatial words, leaving one spatial relation untouched.

`bootstrap` Recall that we use syntactic cues to bootstrap the learning of attribute words in `composite`. In this `bootstrap` task, we flip the direction by inferring objects' names using familiar relational words. We include all four spatial relation words (*i.e.*, *left*, *right*, *front*, and *behind*) and use novel words to represent objects (similar to the setting in `object`). Each image includes three objects (with one distractor), and the utterance includes relations as cues (*e.g.*, "*tufa behind dax*"). Agents are tasked to learn the meaning of the six novel words with the help of relational description and choose the correct answer from the five options.

`number` Acquiring numerical words is a giant leap in children's word learning. Instead of acquiring the cardinal principle (the ability to count to infinity, usually acquired at an older stage), we only consider basic learning of counting words (Wynn, 1990; Fuson, 2012; Piantadosi et al., 2012). This task focuses on how to learn the numerical words, from *one* to *six*. As such, we design the six context images to contain different numbers of objects, ranging from one to six. Each utterance is a unique novel word corresponding to the number of objects in the scene. The query panel includes a random number of objects. To solve the problem, agents need to count how many objects are in the scene and determine the word-number mapping.

`pragmatic` A critical account in children's word learning is a social-pragmatic theory (Tomasello, 2000). Children learn words not only from the cross-situational or linguistic constraints demonstrated in previous tasks but also from inferring communicative intents. In this `pragmatic` task, we inspect this pragmatic word learning capability in machines. Inspired by previous studies on human (Frank & Goodman, 2012; 2014; Fay et al., 2010; 2014; 2018; Horowitz & Frank, 2016; Jiang et al., 2021; 2022; Chen et al., 2021; Qiu et al., 2022), we design a pragmatic word

learning scenario using rendered hands to represent pragmatic pointing. Specifically, every image has a set of three objects and a finger pointing to a referred one, such that the referred object has a unique attribute that can be identified from the context. For example, the targeted object is a *cube* while the other two are *cylinders*. In this case, an informative speaker should use the term "*cube*" instead of "*large cyan metal cube*" to refer to this object. In `pragmatic`, we select six attributes from all available attributes (two sizes, eight colors, three materials, and three shapes) and associate them with unique names. Each of the six context images is paired with a novel attribute word, and we randomly select one attribute to test in the query image.

For all tasks, we provide the following ground-truth scene information: task name, answer choice, word-to-concept mapping, object types, and coordinates (bounding boxes). `pragmatic` is additionally labeled with the pointed object.

4. Word learning with MEWL

To probe human-like word learning in artificial agents, we examine contemporary models on MEWL. Formulating MEWL as a few-shot vision-language learning problem, we choose models that fall into two categories: multimodal (vision-language) and unimodal (language-only) models. Please refer to Appendix C.1 for additional details on model implementation.

4.1. Multimodal models

As MEWL can be viewed as a vision-language task, we test representative multimodal models: pre-trained vision-language models and models with object-centric embedding.

CLIP Contrastive language-image pre-training (CLIP) on large-scale image-caption pairs produces embeddings in a joint image-text embedding space (Radford et al., 2021), showing superb performance on tasks such as zero-shot classification. We take CLIP's pre-trained vision and text encoder (*i.e.*, CLIP (w/ TE)) to extract features from input images and texts. These features are passed to a Transformer model for classification. We also train a model without using CLIP's pre-trained text encoder (*i.e.*, CLIP (w/o TE)).

Flamingo-1.1B Flamingo (Alayrac et al., 2022) is designed to tackle few-shot vision-language tasks. It aligns pre-trained vision and language models by training on large-scale multimodal data. Due to its limited availability, we fine-tune an open-sourced 1.1B version, built on the OPT-350M (Zhang et al., 2022b) and CLIP (ViT-L/14) (Radford et al., 2021), pre-trained on the Conceptual Captions (3M) dataset (Sharma et al., 2018).

Aloe As all the word learning tasks in MEWL are object-based, we additionally test Aloe (Ding et al., 2020), which uses the Transformer architecture to make predictions based on the learned object embeddings and has demonstrated outstanding performance on previous synthetic visual reasoning tasks (Yi et al., 2019; Girdhar & Ramanan, 2019; Zhang et al., 2021a). Therefore, we adopt Aloe in MEWL with object embeddings learned from MONet (Burgess et al., 2019).

4.2. Unimodal models

LLMs have been proven to be strong reasoners with few-shot learning abilities. Hence, we test models based on a caption-then-classify paradigm. First, we use a task-specific oracle captioner to parse the input visual scene to generate a scene description. Next, we use language models (*i.e.*, GPT-3.5 (Brown et al., 2020) and BERT (Devlin et al., 2018)) to classify the result as a multiple choice problem. Of note, such captions are injected with inductive biases that are precisely needed to solve those tasks, having less uncertainty and ambiguity than images used in the multimodal model. This design drastically simplifies the task difficulty, as it is easier for the unimodal model to map syntactic patterns in the captions to the answer. Specifically, inspired by Yang et al. (2021), we prompt GPT-3.5 with a zero-shot multiple-choice template based on full captions generated for the context scenes and the query. We also fine-tune a BERT model on ground-truth captions, resulting in a learned mapping from captions to the answer. Please refer to Appendix B for additional details.

5. Experiments

We examine few-shot word learning ability by experimenting with machine models and human participants on the proposed MEWL benchmark.

5.1. Experimental setup

The dataset consists of nine tasks for comprehensively evaluating agents' word learning capabilities. All models except GPT-3.5 are trained on the training sets of all tasks. We report the model performance on the test sets. All experiments run on eight NVIDIA A100 80GB GPUs. GPT-3.5 model is accessed via the OpenAI API (`text-davinci-003`) with temperature $t = 0$.

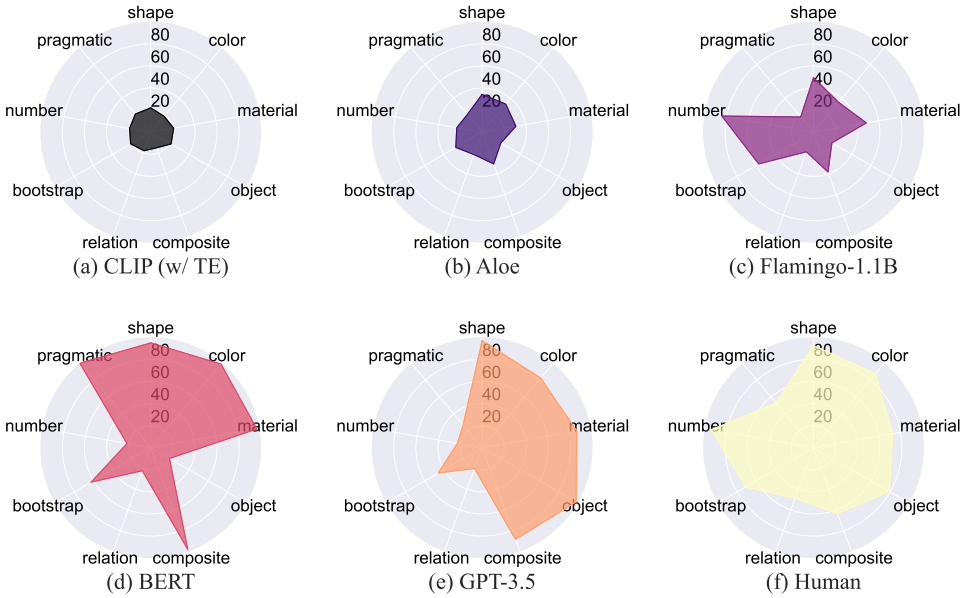
5.2. Human study

To establish a strong baseline to compare with the machines, we looked at how humans performed on the MEWL benchmark; this study was approved by the Institutional Review Board (IRB) at Peking University. We designed questionnaires for the human study based on Qualtrics.

Nine questionnaires were constructed, each of which corresponds to a task. To familiarize participants with our study, the Qualtrics workflow first walked them through a step-by-step tutorial. Next, participants were administered

Table 2: **Performance of baseline models and humans on MEWL.**

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



(a) CLIP (w/ TE) (b) Aloe (c) Flamingo-1.1B

(d) BERT (e) GPT-3.5 (f) Human

Figure 3: **Visualization of agents' performance on MEWL.**

tests and attention checks to ensure they comprehended the background and task settings. Specifically, ten questions and two attention check questions are randomly selected from the MEWL test set and shuffled in each questionnaire. Attention check questions were designed with a query image identical to one of six context images. Participants who failed these questions or checks were removed.

A total of 271 participants (169 female, mean age 42.8) from the US and UK were recruited from Prolific, an online participant pool, to complete the aforementioned nine tasks and were paid an hourly wage of £6, with a bonus of £0.25-£2. Of the 271 responses collected, 1 failed in familiarization, 52 were removed due to attention check failures, 1 outlier was not counted, and 217 were valid and included in the analysis below. For each of the nine tasks, every participant was presented with a randomly drawn ten-question subset from the task's test set. Please see also Appendix D for additional details on data processing and significance tests.

5.3. Results and analysis

Table 2 summarizes the performance of both machines and humans, with result visualization in Figure 3.

Multimodal models Overall, the best vision-language model is Flamingo-1.1B (41.0%), only about half as competent as humans (73.2%). Meanwhile, vanilla transformer models with CLIP features fail catastrophically, achieving only random-level performance on all tasks (less than 20%). Aloe's object-centric representation helps improve performance to 26.8% but may fare worse due to limited model capacity and lack of pre-training.

Peeking into task-specific results, we observe that vision-language pre-trained models perform relatively well on basic attribute naming tasks (*i.e.*, `shape`, `color`, `material`) but fail to generalize to object relations and reason with pragmatical cues. One interesting observation is that the Flamingo model can solve a small proportion of `bootstrap` tasks and some `number` tasks. This result may be attributed to the Flamingo model being language-model-based, capturing syntactic cues and understanding familiar words to bootstrap word learning.

Unimodal models As for unimodal language models, fine-tuned BERT has the best overall performance, with an average performance of 68.3%. Both BERT and GPT-3.5 achieve outstanding performance on object-level tasks

(*i.e.*, shape,color, material,object, composite, bootstrap), yet fail on tasks that require an understanding of more complex relations beyond one-to-one mapping (*i.e.*, relation, number). Fine-tuned on the training set, the BERT model also performs well on the pragmatic task, whereas GPT-3.5 (without fine-tuning) fails, indicating that certain capabilities can indeed be learned through task-specific fine-tuning. However, we also want to point out that detailed captions, with strong human bias injected, have been used: We give object-centric captioning to basic attribute naming tasks, relative spatial relations to relational tasks, and the ground-truth pointing to pragmatic tasks. In this sense, the problem is simplified into a translation-like problem, and the challenge of concept abstraction in human word learning is circumvented.

Human performance Based on 217 valid responses, our human study suggests that MEWL is well-designed and reflects core cognitive skills humans use for word learning. For example, we observe that humans have decent performance on basic naming tasks, with performance ranked shape $\approx$ color $>$ material $>$ composite, which echos prior psychological findings of shape bias (Landau et al., 1988) and fast mapping (Heibeck & Markman, 1987). Humans also perform counting effortlessly. Relational and pragmatic word learning tasks are more challenging than others; relational words often do not have referents to objects, and it is also known to be acquired at the later stage of development (McCune-Nicolich, 1981; Gentner, 2005). Our human study provides a critical reference for what human-level word learning should demonstrate on MEWL.

6. Discussion

6.1. Multimodal *vs.* unimodal

Comparing multimodal models (*i.e.*, CLIP, Flamingo, and Aloe) and unimodal models (*i.e.*, GPT-3.5 and fine-tuned BERT), we observe that text-based models with ground-truth captioning generally outperform pixel-based ones. This observation in machines seems counter-intuitive as it contrasts with the empirical observations and computational studies on human multimodal learning, which argue that multi-modality boosts the acquisition of words and concepts (Clark, 2006; Smith & Gasser, 2005). Why and how do contemporary unimodal agents outperform multimodal ones in few-shot word learning? We present some preliminary discussions on this phenomenon in the following.

First, we believe that part of the conceptual role, not all of it, in unimodal language models may be acquired in a way different from humans. Recently, some studies have shown that large language models can encode human-like conceptual structures, even perceptual ones, from unimodal training data (Piantasodi & Hill, 2022; Abdou et al., 2021), which are confirmed by experiments on human neural systems (Bi, 2021). In our experiments, GPT-3.5 successfully achieves comparable performance on some basic attribute naming tasks (*i.e.*, color, material, shape, object, and composite) and yet fails to learn complex relational words (*i.e.*, number, relation), indicating it already has some conceptual knowledge of shapes, colors, and materials from unimodal training. Nevertheless, GPT-3.5 fails to learn with pragmatic cues, supporting the claim that text-based models cannot infer the communicative word meaning without perceptual grounding (Lake & Murphy, 2021). This leads to the quest for perceptually grounded word learning in machines, to which our MEWL contributes.

Second, the unimodal version of MEWL is similar to the "Quine's *Gavagai* Problem" (Quine, 1960). Since we use ground-truth captioners specifically designed for each task, the unimodal language models do not need to undertake the original word learning as humans do with concept induction. Instead, they acquire the meaning of the novel words via few-shot translation from familiar English words, dramatically reducing the difficulty and ambiguity of multimodal word learning. In other words, the unimodal setting is not comparable with the multimodal one. From the experiment of fine-tuning the BERT model, some tasks that do not require complex cross-situational reasoning can be solved with satisfactory performance. By simplifying the problem as unimodal translation, fine-tuning a unimodal model transforms it into a pattern recognition problem, finding hidden statistical patterns from the training data without acquiring actual human-like few-shot word learning capabilities. Hence, we suggest that future work shall not perform specific fine-tuning on the unimodal captioned version of MEWL for improving performance but instead use it to compare unimodal and multimodal models.

6.2. Humans *vs.* machines

Efficacy of MEWL The results of human studies echo previous established developmental studies on human word learning, *e.g.*, the shape bias (Landau et al., 1988) and the shape $\approx$ color $>$ texture (material) relative preference in fast mapping (Heibeck & Markman, 1987), indicating that our design of MEWL indeed captures the essence of human-like word learning.

Failure of learning models Experiments with contemporary machine learning models show that they fail to demonstrate human-like word learning capabilities in various tasks. Most multimodal models fail catastrophically, reaching chance-level performance. Some cross-attention vision-language or object-centric models show relatively better performance on specific subtasks. Nonetheless, they still do not match overall human word learning capabilities. Although unimodal large language models achieve outstand-

ing performance on basic naming tasks but fall short in capturing relational and pragmatic word learning. In basic attribute naming tasks, large language models do not show human-like learning (e.g., shape versus texture bias (Geirhos et al., 2018; Tartaglini et al., 2022). Crucially, the unimodal paradigm fundamentally differs from human-like multimodal word learning.

Why should machines have human-like word learning capabilities? Few-shot word learning is one of the most basic human multimodal reasoning capabilities; it serves as the first step for language acquisition and facilitates learning concepts (Clark, 2006). Although recent large-scale vision-language contrastive pre-training (Radford et al., 2021) can be viewed as an approximate form of learning from referential uncertainty, it still diverges much from human-like learning: *e.g.*, the failure in social-pragmatics word learning (`pragmatic` task in MEWL and Lake & Murphy (2021)), difficulty in acquiring numerical and relational words (`number` and `relation` tasks in MEWL, Radford et al. (2021), and Conwell & Ullman (2022)), inability to understand compositionality (Thrush et al., 2022), and concept association bias (Yamada et al., 2022). These problems put it on display that current learning paradigms cannot capture the word meaning in a way similar to humans, leading to an alignment and efficiency problem. Whether human-like word learning should be a path to multimodal AI remains a debate, but it is a fundamental ability for human-AI alignment (Yuan et al., 2022).

Word learning represents a general form of human learning. We learn with referential uncertainty, whereas machines currently do not. We use cross-situational information to support few-shot learning of words and concepts, whereas models currently struggle. We learn with teaching and social-pragmatic cues, whereas artificial agents currently fail to understand. Before bridging the gap, how can we assess the capability of machines to learn words under the same conditions as humans? We take the first step by designing these word learning tasks in machines; MEWL is simple and intuitive to support these basic elements in word learning and, in a broader range, human-like learning.

7. Conclusion

We propose MachinE Word Learning (MEWL), a benchmark for human-like few-shot multimodal word learning with referential uncertainty. Inspired by prior developmental studies on how children learn the meaning of words, MEWL includes nine carefully designed tasks covering humans' core cognitive toolkits in word learning: cross-situational reasoning, bootstrapping, and pragmatic learning. These tasks make MEWL the first comprehensive suite for probing machines' word learning capabilities and echoing human word learning scenarios.

We further examine our tasks on contemporary multimodal and unimodal pre-trained models. By recruiting human participants for comparison on MEWL, we found unimodal large language models demonstrate few-shot word learning capabilities on certain subtasks but are still far from human-like. Multimodal vision-language models fail on most tasks, with only the largest language-model-based Flamingo performing better. Together, the results suggest a misalignment of machines' and humans' few-shot word learning capabilities.

We hope MEWL serves as the beginning of our journey to building multimodal agents with human-like few-shot learning. Many open problems and opportunities are left for the community to discuss further. For instance, how to build machines that can learn from uncertainty like children do? What role does social-pragmatic learning play in machine learning? Can unimodal LLMs acquire word meaning and conceptual roles in a way similar to humans without perceptual grounding? Will human-like word learning lead to human-like word meaning? As word learning is among the most fundamental cognitive skills for human multimodal understanding, concept learning, and language acquisition, it is undeniably an essential building block for human-like intelligence. We hope our psychologically informed MEWL can introduce human-like word learning to machines and motivate future research into this problem.

Broader impact and limitation Our MEWL launches a new initiative for modern multimodal learning and reasoning: Instead of focusing their performance on pure memorization tasks, we probe their ability of few-shot learning in context, starting with the fundamentals of human multimodal word learning. We hope our work will stimulate future research on developmentally realistic multimodal models that are endowed with the core capabilities and knowledge of human learning. As a first start, we incorporate nine tasks representing four types of word learning into MEWL. However, the MEWL benchmark is essentially synthetic and devoid of open-vocabulary concepts. As a result, if models are tweaked substantially on the training set, models may find shortcuts, making MEWL degenerate into a set of pattern recognition problems. Therefore, we suggest future research on MEWL to build core multimodal learning capabilities (inductive biases) in a small-data, developmentally plausible regime.

Acknowledgement We would like to thank Yuyang Li (THU) and Nan Jiang (PKU) for their helpful assistance, Liangru Xiang (THU) for constructive discussion, Miss Chen Zhen (BIGAI) for designing the figures, and NVIDIA for their generous support of GPUs and hardware. This work is supported in part by the National Key R&D Program of China (2022ZD0114900) and the Beijing Nova Program.

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., and Søgaard, A. Can language models encode perceptual structure without grounding? a case study in color. In *Computational Natural Language Learning*, 2021. 8

Abdulla, S. Statistics starts young. *Nature*, 2001. 2

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., and Steedman, M. Bootstrapping language acquisition. *Cognition*, 164:116–143, 2017. 5

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 6

Barrett, D., Hill, F., Santoro, A., Morcos, A., and Lillicrap, T. Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning (ICML)*, 2018. 3

Bender, E. M. and Koller, A. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 2

Berger, U., Stanovsky, G., Abend, O., and Frermann, L. A computational acquisition model for multimodal word categorization. *arXiv preprint arXiv:2205.05974*, 2022. 2

Bi, Y. Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, 25(10):883–895, 2021. 8

Bloom, P. Précis of how children learn the meanings of words. *Behavioral and Brain Sciences*, 24(6):1095–1103, 2001. 1

Bloom, P. *How children learn the meanings of words*. MIT press, 2002. 1, 3

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 6

Burgess, C. P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., and Lerchner, A. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019. 6

Carey, S. and Bartlett, E. Acquiring a single new word. *Papers and Reports on Child Language Development*, 15:17–29, 1978. 1, 2, 3, 5

Chen, Y., Li, Q., Kong, D., Kei, Y. L., Zhu, S.-C., Gao, T., Zhu, Y., and Huang, S. Yourefit: Embodied reference understanding with language and gesture. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

Chollet, F. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 3

Clark, A. Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences*, 10(8):370–374, 2006. 8, 9

Community, B. O. Blender–a 3d modelling and rendering package, 2016. 4

Conwell, C. and Ullman, T. Testing relational understanding in text-guided image generation. *arXiv preprint arXiv:2208.00005*, 2022. 2, 9

Depeweg, S., Rothkopf, C. A., and Jäkel, F. Solving bongard problems with a visual language and pragmatic reasoning. *arXiv preprint arXiv:1804.04452*, 2018. 3

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018. 6, 17

Ding, D., Hill, F., Santoro, A., Reynolds, M., and Botvinick, M. M. Attention over learned object embeddings enables complex visual reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 6

Edmonds, M., Kubricht, James, F., Summers, C., Zhu, Y., Rothrock, B., Zhu, S.-C., and Lu, H. Human causal transfer: Challenges for deep reinforcement learning. In *Annual Meeting of the Cognitive Science Society (CogSci)*, 2018. 3

Edmonds, M., Qi, S., Zhu, Y., Kubricht, J., Zhu, S.-C., and Lu, H. Decomposing human causal learning: Bottom-up associative learning and top-down schema reasoning. In *Annual Meeting of the Cognitive Science Society (CogSci)*, 2019. 3

Edmonds, M., Ma, X., Qi, S., Zhu, Y., Lu, H., and Zhu, S.-C. Theory-based causal transfer: Integrating instance-level induction and abstract-level structure learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 3

Engelcke, M., Kosiorek, A. R., Parker Jones, O., and Posner, I. GENESIS: Generative Scene Inference and Sampling with Object-Centric Latent Representations. In *International Conference on Learning Representations (ICLR)*, 2020. 16

Fan, L., Xu, M., Cao, Z., Zhu, Y., and Zhu, S.-C. Artificial social intelligence: A comparative and holistic view. *CAAI Artificial Intelligence Research*, 1(2):144–160, 2022. 2

Fay, N., Garrod, S., Roberts, L., and Swoboda, N. The interactive evolution of human communication systems. *Cognitive Science*, 34(3):351–386, 2010. 5

Fay, N., Ellison, M., and Garrod, S. Iconicity: From sign to system in human communication and language. *Pragmatics & Cognition*, 22(2):244–263, 2014. 5

Fay, N., Walker, B., Swoboda, N., and Garrod, S. How to create shared symbols. *Cognitive Science*, 42:241–269, 2018. 5

Fodor, J. A., Pylyshyn, Z. W., et al. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988. 2

Frank, M. C. and Goodman, N. D. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012. 2, 5

Frank, M. C. and Goodman, N. D. Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, 75:80–96, 2014. 2, 4, 5

Frank, M. C., Braginsky, M., Yurovsky, D., and Marchman, V. A. Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, 44(3):677–694, 2017. 2

Friedman, W. J. and Seely, P. B. The child's acquisition of spatial and temporal word meanings. *Child Development*, pp. 1103–1108, 1976. 5

Fuson, K. C. *Children's counting and concepts of number*. Springer Science & Business Media, 2012. 5

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 9

Gentner, D. The development of relational category knowledge. In *Building object categories in developmental time*, pp. 263–294. Psychology Press, 2005. 8

Girdhar, R. and Ramanan, D. Cater: A diagnostic dataset for compositional actions & temporal reasoning. In *International Conference on Learning Representations (ICLR)*, 2019. 6

Gleitman, L. The structural sources of verb meanings. *Language Acquisition*, 1(1):3–55, 1990. 5

Goodman, N., Tenenbaum, J., and Black, M. A bayesian framework for cross-situational word-learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007. 2

Gopnik, A., Meltzoff, A. N., and Kuhl, P. K. *The scientist in the crib: Minds, brains, and how children learn.* William Morrow & Co, 1999. 2

Heibeck, T. H. and Markman, E. M. Word learning in children: An examination of fast mapping. *Child Development*, pp. 1021–1034, 1987. 1, 3, 8

Horowitz, A. C. and Frank, M. C. Children's pragmatic inferences as a route for learning about the world. *Child Development*, 87 (3):807–819, 2016. 2, 4, 5

Horst, J. S. and Hout, M. C. The novel object and unusual name (noun) database: A collection of novel images for use in experimental research. *Behavior Research Methods*, 48(4):1393–1409, 2016. 3

Hsu, J., Wu, J., and Goodman, N. Geoclidean: Few-shot generalization in euclidean geometry. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 3

Ji, A., Kojima, N., Rush, N., Suhr, A., Vong, W. K., Hawkins, R. D., and Artzi, Y. Abstract visual reasoning with tangram shapes. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022. 3

Jiang, J. and Ahn, S. Generative neurosymbolic machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 15

Jiang, K., Stacy, S., Wei, C., Chan, A., Rossano, F., Zhu, Y., and Gao, T. Individual vs. joint perception: a pragmatic model of pointing as communicative smithian helping. *arXiv preprint arXiv:2106.02003*, 2021. 5

Jiang, K., Dahmani, A., Stacy, S., Jiang, B., Rossano, F., Zhu, Y., and Gao, T. What is the point? a theory of mind model of relevance. In *Annual Meeting of the Cognitive Science Society (CogSci)*, 2022. 5

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 4

Krishnamohan, V., Soman, A., Gupta, A., and Ganapathy, S. Audiovisual correspondence learning in humans and machines. In *INTERSPEECH*, 2020. 3

Kuhnle, A. and Copestake, A. Shapeworld-a new test methodology for multimodal language understanding. *arXiv preprint arXiv:1704.04517*, 2017. 3

Lake, B. and Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning (ICML)*, 2018. 3

Lake, B. M. and Murphy, G. L. Word meaning in minds and machines. *Psychological Review*, 2021. 1, 2, 8, 9

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 1, 3

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017. 2

Lake, B. M., Linzen, T., and Baroni, M. Human few-shot learning of compositional instructions. *arXiv preprint arXiv:1901.04587*, 2019a. 3

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. The omniglot challenge: a 3-year progress report. *Current Opinion in Behavioral Sciences*, 29:97–104, 2019b. 3

Landau, B., Smith, L. B., and Jones, S. S. The importance of shape in early lexical learning. *Cognitive Development*, 3(3):299–321, 1988. 1, 8

Li, Q., Zhu, Y., Liang, Y., Wu, Y. N., Zhu, S.-C., and Huang, S. Neural-symbolic recursive machine for systematic generalization. *arXiv preprint arXiv:2210.01603*, 2022a. 3

Li, Q., Huang, S., Hong, Y., Zhu, Y., Wu, Y. N., and Zhu, S.-C. A minimalist dataset for systematic generalization of perception, syntax, and semantics. In *International Conference on Learning Representations (ICLR)*, 2023. 3

Li, S., Wu, K., Zhang, C., and Zhu, Y. On the learning mechanisms in physical reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022b. 3

McCune-Nicolich, L. The cognitive bases of relational words in the single word period. *Journal of Child language*, 8(1):15–34, 1981. 8

Merriman, W. E., Bowman, L. L., and MacWhinney, B. The mutual exclusivity bias in children's word learning. In *Monographs of the society for research in child development*, pp. i–129. JSTOR, 1989. 4

Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81, 1956. 5

Mitchell, M. and Krakauer, D. C. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences (PNAS)*, 120(13):e2215907120, 2023. 2

Murphy, G. *The big book of concepts*. MIT press, 2004. 1

Nie, W., Yu, Z., Mao, L., Patel, A. B., Zhu, Y., and Anandkumar, A. Bongard-logo: A new benchmark for human-level concept learning and reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

Orhan, E., Gupta, V., and Lake, B. M. Self-supervised learning through the eyes of a child. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

Piantadosi, S. T., Tenenbaum, J. B., and Goodman, N. D. Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, 123(2):199–217, 2012. 5

Piantasodi, S. T. and Hill, F. Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*, 2022. 2, 8

Pinker, S. *Language learnability and language development: with new commentary by the author*, volume 7. Harvard University Press, 2009. 2, 3

Qiu, S., Xie, S., Fan, L., Gao, T., Joo, J., Zhu, S.-C., and Zhu, Y. Emergent graphical conventions in a visual communication game. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 5

Quine, W. V. O. *Word and object*. MIT press, 1960. 2, 8

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 6, 9

Rane, S., Nencheva, M. L., Wang, Z., Lew-Williams, C., Russakovsky, O., and Griffiths, T. L. Predicting word learning in children from the performance of computer vision systems. *arXiv preprint arXiv:2207.09847*, 2022. 2

Sandhofer, C. M. and Smith, L. B. Learning color words involves learning a system of mappings. *Developmental Psychology*, 35 (3):668, 1999. 5

Scott, R. M. and Fisher, C. 2.5-year-olds use cross-situational consistency to learn verbs under referential uncertainty. *Cognition*, 122(2):163–180, 2012. 2, 3

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. 6

Smith, K., Smith, A. D., and Blythe, R. A. Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, 35(3):480–498, 2011. 2, 3, 5

Smith, L. and Gasser, M. The development of embodied cognition: Six lessons from babies. *Artificial Life*, 11(1-2):13–29, 2005. 1, 8

Smith, L. and Yu, C. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008. 2, 5

Stacy, S., Parab, A., Kleiman-Weiner, M., and Gao, T. Overloaded communication as paternalistic helping. In *Annual Meeting of the Cognitive Science Society (CogSci)*, 2022. 2

Suhr, A., Lewis, M., Yeh, J., and Artzi, Y. A corpus of natural language for visual reasoning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017. 3

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., and Frank, M. C. Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind*, 5:20–29, 2022. 2

Tartaglini, A. R., Vong, W. K., and Lake, B. A developmentally-inspired examination of shape versus texture bias in machines. In *Annual Meeting of the Cognitive Science Society (CogSci)*, 2022. 9

Tenenbaum, J. Bayesian modeling of human concept learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1998. 2

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011. 1, 2

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 9

Tomasello, M. The social-pragmatic theory of word learning. *Pragmatics*, 10(4):401–413, 2000. 5

Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S., Vinyals, O., and Hill, F. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

Vedantam, R., Szlam, A., Nickel, M., Morcos, A., and Lake, B. M. Curi: a benchmark for productive concept learning under uncertainty. In *International Conference on Machine Learning (ICML)*, 2021. 3

Vong, W. K. and Lake, B. M. Learning word-referent mappings and concepts from raw inputs. In *Annual Meeting of the Cognitive Science Society (CogSci)*, 2020. 2

Vong, W. K. and Lake, B. M. Cross-situational word learning with multimodal neural networks. *Cognitive Science*, 46(4):e13122, 2022. 3, 5

Vong, W. K., Orhan, E., and Lake, B. Cross-situational word learning from naturalistic headcam data. In *CUNY Conference on Human Sentence Processing*, 2021. 2

Wang, W., Vong, W. K., Kim, N., and Lake, B. M. Finding structure in one child's linguistic experience, Dec 2022. URL psyarxiv.com/85k3y. 3

Wynn, K. Children's understanding of counting. *Cognition*, 36(2): 155–193, 1990. 5

Xie, S., Ma, X., Yu, P., Zhu, Y., Wu, Y. N., and Zhu, S.-C. Halma: Humanlike abstraction learning meets affordance in rapid problem solving. *arXiv preprint arXiv:2102.11344*, 2021. 3

Xu, F. and Tenenbaum, J. B. Word learning as bayesian inference. *Psychological Review*, 114(2):245, 2007. 2

Xu, M., Jiang, G., Zhang, C., Zhu, S.-C., and Zhu, Y. Est: Evaluating scientific thinking in artificial agents. *arXiv preprint arXiv:2206.09203*, 2022. 3

Yamada, Y., Tang, Y., and Yildirim, I. When are lemons purple? the concept association bias of clip. *arXiv preprint arXiv:2212.12043*, 2022. 2, 9

Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., and Wang, L. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 6

Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., and Tenenbaum, J. B. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2019. 6

Yu, C., Zhang, Y., Slone, L. K., and Smith, L. B. The infant's view redefines the problem of referential uncertainty in early word learning. *Proceedings of the National Academy of Sciences (PNAS)*, 118(52):e2107019118, 2021. 4

Yuan, L., Gao, X., Zheng, Z., Edmonds, M., Wu, Y. N., Rossano, F., Lu, H., Zhu, Y., and Zhu, S.-C. In situ bidirectional human-robot value alignment. *Science Robotics*, 7(68), 2022. 9

Zhang, C., Gao, F., Jia, B., Zhu, Y., and Zhu, S.-C. Raven: A dataset for relational and analogical visual reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019a. 3

Zhang, C., Jia, B., Gao, F., Zhu, Y., Lu, H., and Zhu, S.-C. Learning perceptual inference by contrasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019b. 3

Zhang, C., Jia, B., Edmonds, M., Zhu, S.-C., and Zhu, Y. Acre: Abstract causal reasoning beyond covariation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a. 3, 6

Zhang, C., Jia, B., Zhu, S.-C., and Zhu, Y. Abstract spatial-temporal reasoning via probabilistic abduction and execution. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b. 3

Zhang, C., Xie, S., Jia, B., Wu, Y. N., Zhu, S.-C., and Zhu, Y. Learning algebraic representation for systematic generalization in abstract reasoning. In *European Conference on Computer Vision (ECCV)*, 2022a. 3

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022b. 6

Zhang, W., Zhang, C., Zhu, Y., and Zhu, S.-C. Machine number sense: A dataset of visual arithmetic problems for abstract and relational reasoning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 3

Zhu, Y., Gao, T., Fan, L., Huang, S., Edmonds, M., Liu, H., Gao, F., Zhang, C., Qi, S., Wu, Y. N., Tenenbaum, J. B., and Zhu, S.-C. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020. 2

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., and Yamins, D. L. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences (PNAS)*, 118(3):e2014196118, 2021. 3

A. Details for MEWL dataset generation

This section provides additional details on the generation procedures of MEWL.

A.1. Additional task construction details

shape There are three shapes in MEWL: *cube*, *sphere*, and *cylinder*. We randomly assign three unique novel words to the shapes in each episode to create word-shape mappings. For example, *temmar* is *cylinder*, *subno* is *cube*, and *teis* is *sphere*. When generating the context, we first select a shape (*e.g.*, *cylinder*) and the corresponding word (*e.g.*, *temmar*) as utterance. We then create this context image containing one *cylinder* object and uniformly sample other properties (color, size, material). To make it solvable, we also ensure every panel appears more than once and avoid ambiguous scenarios where two concepts accidentally bind together across all contexts. For the query panel, we choose one shape for testing (with an object of the selected shape as a query image). The candidate options are three novel words (each corresponds to a shape concept) and two dummy words (randomly generated).

color There are eight colors in MEWL: *gray*, *red*, *blue*, *green*, *brown*, *purple*, *cyan*, and *yellow*. We randomly sample three colors out of eight to appear in each episode. Other settings remain the same as in `shape`.

material When designing this task, we keep most of the settings unchanged from `shape`. Instead of naming colors, we name three materials: *rubber*, *metal*, and *glass*.

object In this task, novel words bind to the objects (*i.e.*, the quadruple of size, color, material, and shape). There are 2 (sizes) $\times$ 8 (colors) $\times$ 3 (materials) $\times$ 3 (shapes) = 144 unique objects in MEWL. We first sample six unique objects and six novel words to construct images and utterances (*e.g.*, *daythetle* is *mall purple rubber cylinder*, *outsupac* is *small gray metal cube*, ..., and *peafcol* is *large cyan glass cylinder*). Unlike the `shape` task, each image in `object` has three objects. Each utterance has three words representing the three objects in the image (*e.g.*, *daythetle and outsupac and peafcol*). We randomly sample a subset of three objects from the six selected objects to construct a scene. Moreover, we ensure that there are no identical scenes between the six contexts and one query. The five options consist of one ground-truth utterance and four utterances corresponding to object subsets that have not appeared in the context or query.

composite This task focuses on learning compositional multi-word phrases and uses syntax to bootstrap the word learning process. In detail, novel words represent attributes (shape, color, and material), and utterance phrases share the same syntax in an episode. In each episode, we first sample two attribute types for naming (*e.g.*, color and shape). We also design an episode-specific syntax (*e.g.*, the first word represents color, and the second word is a shape). For each type of attribute, we randomly choose three instances (*e.g.*, for color, we choose *cyan*, *blue*, *yellow*), resulting in six named attributes (*e.g.*, three colors and three shapes). We map six novel words to these attributes. In each image and the corresponding utterance, we sample one object that satisfies the syntactic constraints (*e.g.*, the object's color must be in the three selected colors, and the object's shape also must be in the three selected shapes, but other attributes of the object are not restricted). We ensure that there are no duplicated attribute pairs (two objects have the same color and shape) among all objects in the images. The rest are the same as `object`.

relation In this task, we probe agents' capability of learning relational words (*i.e.*, *left*, *right*, *front*, and *behind*). To construct spatial relations, we place three objects (with one dummy object) in an image and use two familiar English words to refer to objects and a novel spatial relation word in between to represent objects' relation (*e.g.*, "*cyan cube dax brown sphere*"). We also replicate the ambiguity when children acquire spatial words. Because spatial locations are ambiguous: For example, *dax* can refer to *left* or *front* when inferring from a single image (the left is often confounded with at least one other orientation). Agents must use cross-situational reasoning to determine the exact meaning of the spatial words. We design each novel word to appear twice to ensure the problem is solvable (*e.g.*, from both *left behind* and *left front*, we can rule out the other confounded orientation and understand *dax* means *left*). Therefore, only three spatial words are used in an episode, leaving one orientation untouched. We use one of the three words (spatial relationships) in the query image.

bootstrap In `composite`, we have already used some syntactic cues to bootstrap the learning of attribute words. In this `bootstrap` task, we take a step further by inferring objects' names using familiar relational words. We include all four spatial relation words (*i.e.*, *left*, *right*, *front*, and *behind*) in familiar English and use novel words to represent objects (similar to the setting in `object`). Six word-concept mappings are required to figure out in every episode. Each image includes three objects (with one dummy object), and the utterance includes relations as cues (*e.g.*, "*tufa behind dax*"). Agents are tasked to learn the meaning of the six novel words with the help of relational description and choose the correct answer from the five options.

number This task focuses on how to learn the number words, from *one* to *six*. In this way, we design the six context images to contain different numbers of objects (ranging from one to six). Each utterance has a unique novel word corresponding to the number of objects in the scene (*e.g.*, *ure* is *one*, *manthe* is *two*, ..., and *sical* is *six*). The query panel includes a random number (within six) of objects. To solve the problem, agents need to count how many objects are in the scene and figure out the word-number mapping.

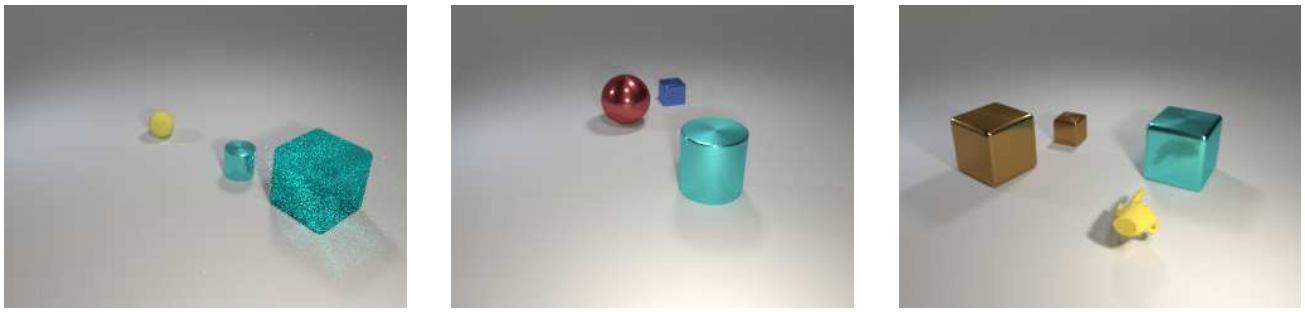
pragmatic We design a pragmatic word learning scenario using rendered hands to represent the pointing gesture. In detail, every image has a set of three objects and a finger pointing to a referred object. The referred object has a unique attribute that can be uniquely identified from the context. For example, suppose the targeted object is a *cube*, while the other two are *cylinders*. In that case, an informative speaker should use the term "*cube*" instead of "*large cyan metal cube*" to refer to this object. In practice, we select six attributes from all available attributes (two sizes, eight colors, three materials, and three shapes) and assign attributes with unique names (*e.g.*, *supcon* is *sphere*, *fuly* is *large*, ..., and *mainder* is *purple*). Each of the six context images represents a novel attribute word, while we randomly select one attribute for the query image. To generate three objects in a scene, we first sample one base object and modify this base object to become a referred object with a unique attribute. We then modify different attribute types to construct the third object. Agents must understand the correspondence of novel words to referred attributes. The rendering code for referring objects is based on Jiang & Ahn (2020).

B. Captioning and text input for unimodal models

For unimodal language models, we use ground truth scene caption for each figure as the input. In this section, we describe the caption generation process and provide some examples of the generated captions.

We use different captions for different types of questions.

The `object`, `shape`, `color`, `material`, `composite`, and `number` tasks require complete descriptions of all objects in the scene, including their attributes (*i.e.*, color, shape, material, size) to ensure all relevant information is provided. An example of the generated captions is shown in Figure A1a.



(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.

Figure A1: **Examples of generated captions.**

The `relation` and `bootstrap` tasks require knowledge of spatial relations between objects in the scene. To facilitate this, captions for these tasks include the relative position of objects using terms such as "*front*", "*behind*", "*left*", and "*right*" along with detailed descriptions of the objects. An example of the generated captions is shown in Figure A1b.

The `pragmatic` task necessitates not only knowledge of the objects and their descriptions but also the identification of the object being pointed at in the scene. An example of the generated captions is shown in Figure A1c.

To perform the task, we construct the final text input by combining three elements: i) a prompt that specifies the problem (*i.e.* "*Please name the target object according to the above context.*"). ii) captions for each figure in the context and their associated utterances. iii) captions for the query image along with the provided options.

A full example text input for a problem in `pragmatic` can is shown below:

Please name the target object according to the above context.
Context: A small cyan metal cylinder and a small cyan rubber cylinder and a small brown metal cylinder and a small yellow rubber arrow. And a finger is pointing to the small brown metal cylinder. Name: enre
Context: A large brown metal sphere and a large brown metal cylinder and a large brown rubber sphere and a small yellow rubber arrow. And a finger is pointing to the large brown metal cylinder. Name: taward
Context: 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. Name: facset
Context: A large brown rubber sphere and a large brown rubber cube and a large brown glass cube and a small yellow rubber arrow. And a finger is pointing to the large brown glass cube. Name: facov
Context: A small red metal cube and a small purple metal cube and a small red glass cube and a small yellow rubber arrow. And a finger is pointing to the small purple metal cube. Name: alim
Context: A small green glass sphere and a small green rubber sphere and a large green glass sphere and a small yellow rubber arrow. And a finger is pointing to the large green glass sphere. Name: tedfac
Context: A small yellow rubber cube and a large yellow rubber cube and a large purple rubber cube and a small yellow rubber arrow. And a finger is pointing to the large purple rubber cube. Name: `[Option]`

`[Option]` is a candidate utterance (*e.g.*, "*enre*", "*tedfac*", "*facset*", "*alim*", or "*facov*").

C. Experimental details

In this section, we describe the experimental details for the baseline models used in the paper.

For vision-language models, we use the image and the corresponding text as input. While for language-only models, we first use a captioner to parse the image into full scene descriptions, then use the scene descriptions and utterances as input.

C.1. Model details

CLIP The CLIP model utilizes a pre-trained image encoder (ViT-B/16) to extract features from images. Text features are calculated using either the text encoder of CLIP-ViT-B-16 (for CLIP (w/ TE)) or CLIP's token embedding (for CLIP (w/o TE)). The resulting features are concatenated in the format of `[image1, utterance1, image2, utterance2, ..., image6, utterance6, image query, option1, option2, ..., option5]`. These input features are then passed through a 6-layer Transformer model with an MLP head for classification. We freeze the CLIP when training. The model is trained on the training set for 600 epochs, dropout 0.1, batch size 64, learning rate 1×10^{-4}, and AdamW optimizer (weight_decay 0.01).

Aloe The Aloe model employs the MONet architecture as provided by Engelcke et al. (2020). The MONet model is pre-trained on the training set images resized into 128×128 for 600 epochs with Adam (lr $= 1 \times 10^{-5}$), 7 slots, and a latent dimension of 16. We take the mean as the object feature of the figure and use this feature as the visual input for the Transformer. As for the text input, we embed each word using a trainable embedding as follows: i) If the utterance is a typical English word (e.g. *yellow*, *metal*, *and*), the utterance is directly encoded by the embedder. ii) If the word is a novel one, it is embedded as a random placeholder. The visual inputs are concatenated with text embeddings of the context utterances and choices in a similar format as the CLIP model inputs. These inputs are then passed through a Transformer model with a head size of 8, a latent dimension of 512, and trained for 600 epochs using the Adam optimizer with a learning rate of 1×10^{-4} for classification. Hyperparameters for the Transformer model are as follows: training 200 epochs, learning rate 5×10^{-5}, Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-8}$, linear learning rate decay, and batch size 128.

Flamingo-1.1B Since the Flamingo models' pre-trained weights are not accessible, we use open-sourced implementations of the Flamingo model[1]. Specifically, we use a 1.1B Flamingo model built upon OPT-350M and CLIP pre-trained ViT-L/14 model. The model is pre-trained on the Conceptual Captions dataset.

For MEWL tasks, we formulated it as a multiple-choice problem (similar to how GPT performs multiple-choice tasks). Specifically, we add a binary classifier head to Flamingo's last layer outputs. Meanwhile, we concatenate episodic interleaved image and utterance pairs, the query image, and one option (candidate utterance) as input. We feed each episode five times to get the logits corresponding to the five options and pass through a softmax layer to get the final answer. We use cross-entropy loss for training. Hyperparameters are as follows: training steps 30000 ($\approx$ 106 epochs), learning rate 5×10^{-5}, Adam

[1] https://github.com/lucidrains/flamingo-pytorch and https://github.com/dhansmair/flamingo-mini

optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-8}$, linear learning rate decay, and batch size 96.

GPT-3.5 We use the `text-davinci-003` model provided by the OpenAI API. Formulated as a few-shot multiple choice question answering problem, we concatenate all the image captions and utterances as inputs. Formally following four steps: i). We caption six context images and a query image into `caption1, caption2, caption3, ..., caption6, query_caption`. ii). We concatenate them with the corresponding utterances (context) to get the context input $\mathcal{C}$ =`[caption1, utterance1, caption2, utterance2, ..., caption6, utterance6]`. iii). We then construct five inputs for GPT-3.5 by concatenating context input, query caption, and a possible option (*e.g.*, [$\mathcal{C}$, `query_caption, option1`], [$\mathcal{C}$, `query_caption, option2`], ..., [$\mathcal{C}$, `query_caption, option5`]). iv). Finally, we feed the prompt to GPT-3.5 and choose the one with the largest log probability as the answer.

BERT For the BERT model, we follow standard practice for utilizing BERT as a multiple-choice question answering; see Section 4.4 of Devlin et al. (2018) for details. We first generate ground truth captions for each figure using the captioner and construct the question input `[caption1, utterance1, caption2, utterance2, ..., caption6, utterance6, query_caption]`. Then, we construct five input sentences by concatenating the question input and a candidate option. We adopt a linear scoring head and a softmax layer on the last layer's `[CLS]` hidden state to calculate the class probability.

We fine-tune a BERT-base model on the training set for 200 epochs, with learning rate 5×10^{-5}, Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-8}$, linear learning rate decay, and batch size 64.

D. Human study

271 participants were recruited from Prolific (169 female; mean age 42.8) for the nine tasks. All of the participants are from UK or USA and have a Bachelor's degree or higher. The participants were paid an hourly wage of £6 (with a bonus of £0.25-£2). This study has been approved by an IRB. 270 of 271 responses are accepted (one failed in familiarization), 217 of which are valid (52 removed due to failures in attention checks, and one removed due to outlier).

D.1. Data processing

Responses from participants who failed attention check questions were removed when measuring human performance. Besides, Grubbs' tests were performed with the significance level $\alpha = 0.5$ in each group to remove outlier results. Only one outlier was detected and removed.

D.2. Tests of statistical significance

We used a *t*-test to determine if one task is significantly easier than the other. The *t*-test was performed between any two groups of human results with the significance level $p < .05$ (one-tailed). Our null hypothesis is that the group of results with a higher mean is not significantly better than the other group. the The average of human performance and *t*-test results are shown in Table A1 and Table A2, respectively.

Table A1: **Performance of humans on MEWL.**

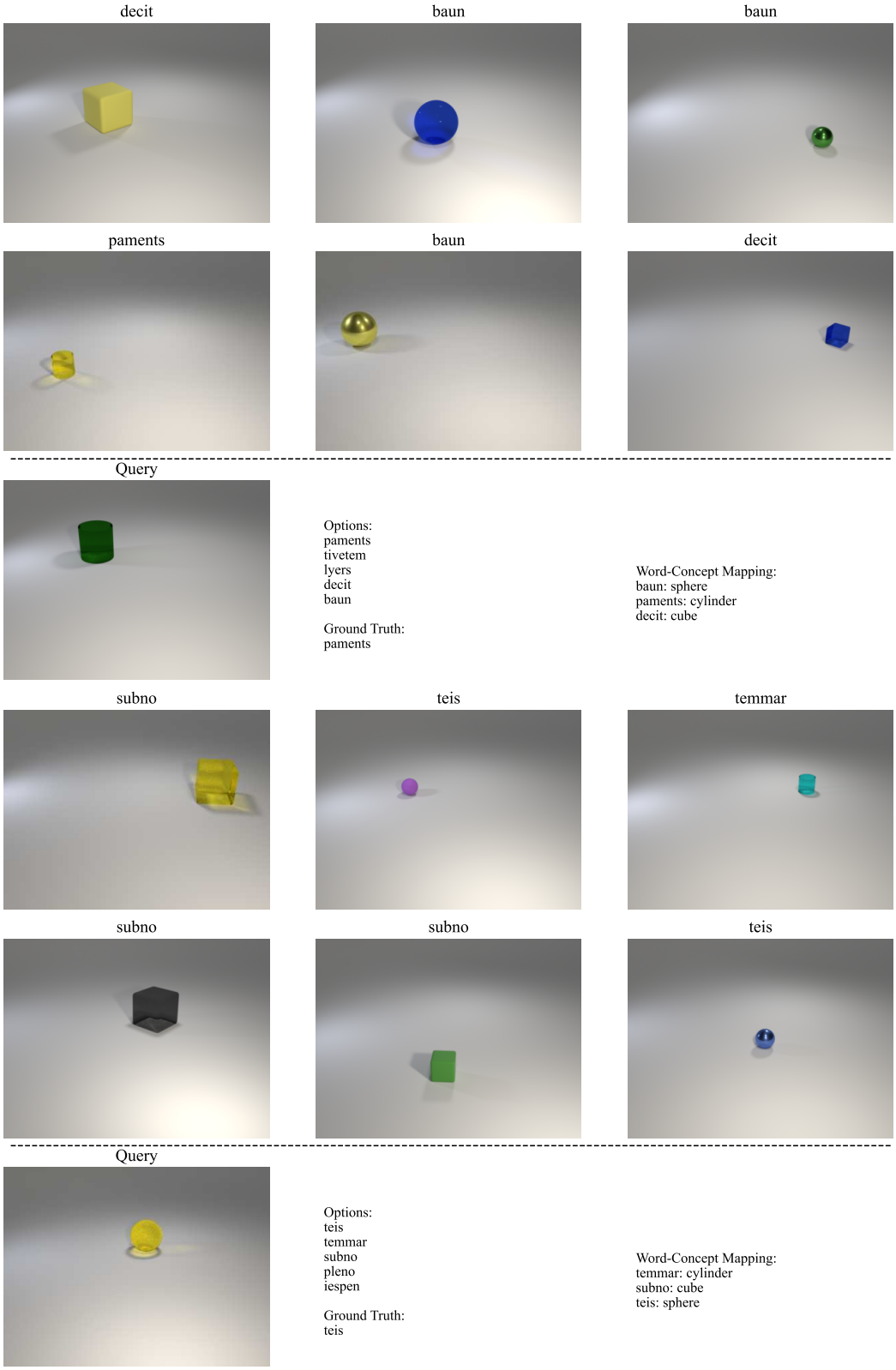
	shape	color	material	object	composite	relation	bootstrap	number	pragmatic	Avg.
Human	92.4	87.2	72.7	79.1	63.5	48.7	71.0	93.9	54.8	73.7

Table A2: **The *p*-values of human results on the MEWL**. Blue indicates that the task with a higher average is significantly easier for humans than the other, while Red indicates that there is no significant difference in difficulty between the two tasks.

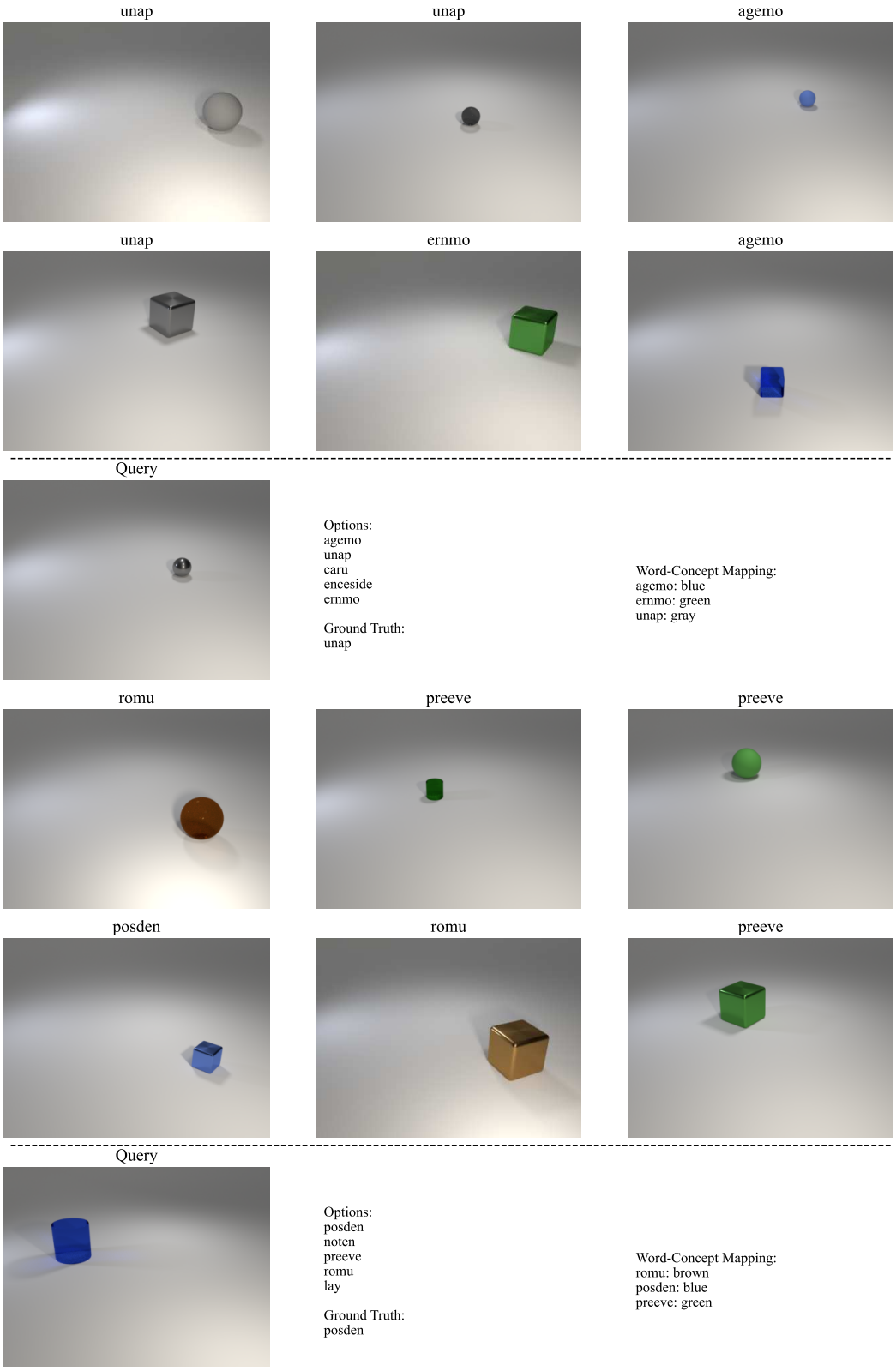
	color	material	object	composite	relation	bootstrap	number	pragmatic
shape	.098	$< .001$	.003	$< .001$	$< .001$	.001	.305	$< .001$
color	-	.006	.065	$< .001$	$< .001$	.002	.058	$< .001$
material	-	-	.133	.082	$< .001$	.385	$< .001$	.015
object	-	-	-	.009	$< .001$	.074	.002	.002
composite	-	-	-	-	.016	.120	$< .001$	.150
relation	-	-	-	-	-	$< .001$	$< .001$	.225
bootstrap	-	-	-	-	-	-	$< .001$	.021
number	-	-	-	-	-	-	-	$< .001$

E. More task examples

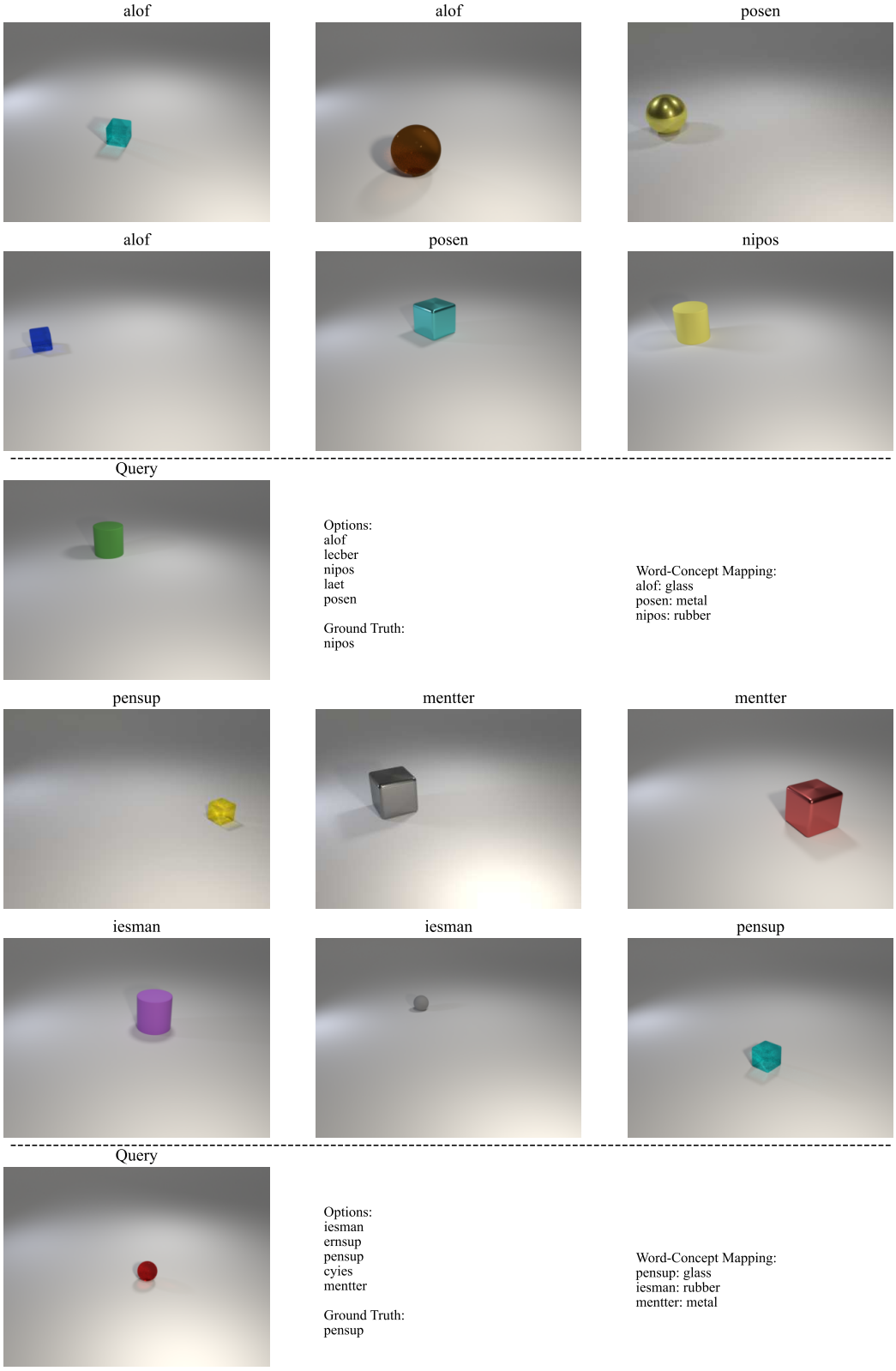
E.1. shape



E.2. `color`

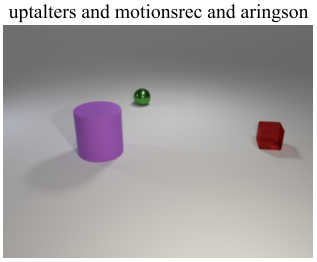

unap
unap
agemo
unap
ernmo
agemo
Query
Options:
agemo
unap
caru
enceside
ernmo
Ground Truth:
unap
Word-Concept Mapping:
agemo: blue
ernmo: green
unap: gray
romu
preeve
preeve
posden
romu
preeve
Query
Options:
posden
noten
preeve
romu
lay
Ground Truth:
posden
Word-Concept Mapping:
romu: brown
posden: blue
preeve: green

E.3. `material`

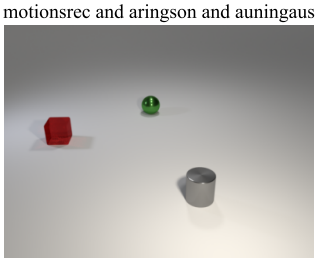

alof
alof
posen
alof
posen
nipos
Query
Options:
alof
lecber
nipos
laet
posen
Ground Truth:
nipos
Word-Concept Mapping:
alof: glass
posen: metal
nipos: rubber
pensup
mentter
mentter
iesman
iesman
pensup
Query
Options:
iesman
ernsup
pensup
cyies
mentter
Ground Truth:
pensup
Word-Concept Mapping:
pensup: glass
iesman: rubber
mentter: metal

E.4. object

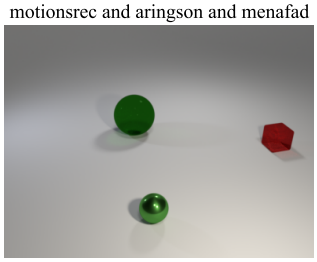
uptalters and motionsrec and aringson



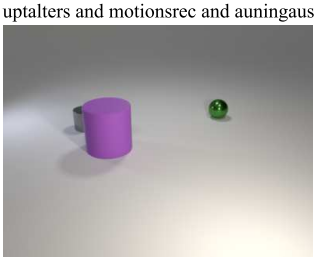
motionsrec and aringson and auningaus



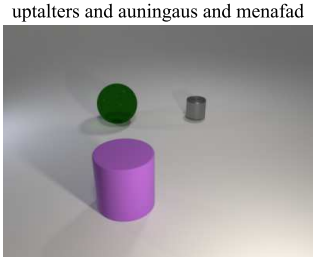
motionsrec and aringson and menafad



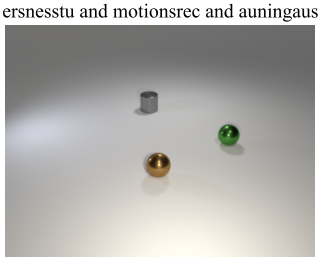
uptalters and motionsrec and auningaus



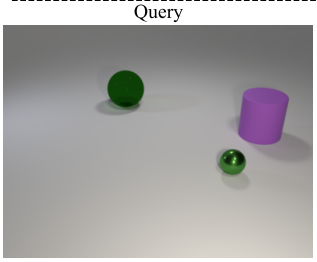
uptalters and auningaus and menafad



ersnesstu and motionsrec and auningaus



Query

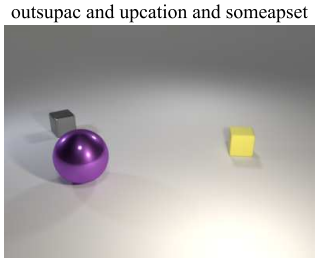


Options:
uptalters and ersnesstu and auningaus
uptalters and aringson and auningaus
ersnesstu and auningaus and menafad
uptalters and ersnesstu and motionsrec
uptalters and motionsrec and menafad

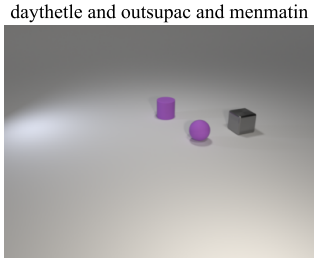
Ground Truth:
uptalters and motionsrec and menafad

Word-Concept Mapping:
uptalters: ['cylinder', 'purple', 'rubber', 'large']
ersnesstu: ['sphere', 'brown', 'metal', 'small']
motionsrec: ['sphere', 'green', 'metal', 'small']
aringson: ['cube', 'red', 'glass', 'small']
auningaus: ['cylinder', 'gray', 'metal', 'small']
menafad: ['sphere', 'green', 'glass', 'large']

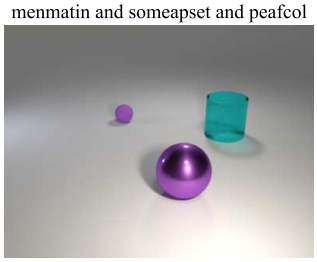
outsupac and upcation and someapset



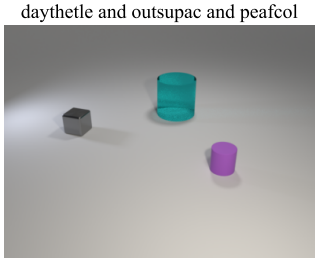
daythetle and outsupac and menmatin



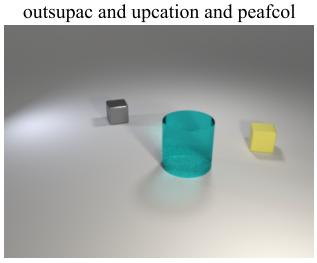
menmatin and someapset and peafcol



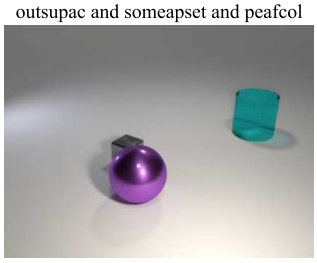
daythetle and outsupac and peafcol



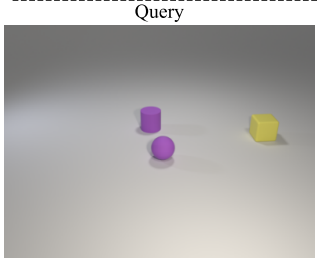
outsupac and upcation and peafcol



outsupac and someapset and peafcol



Query

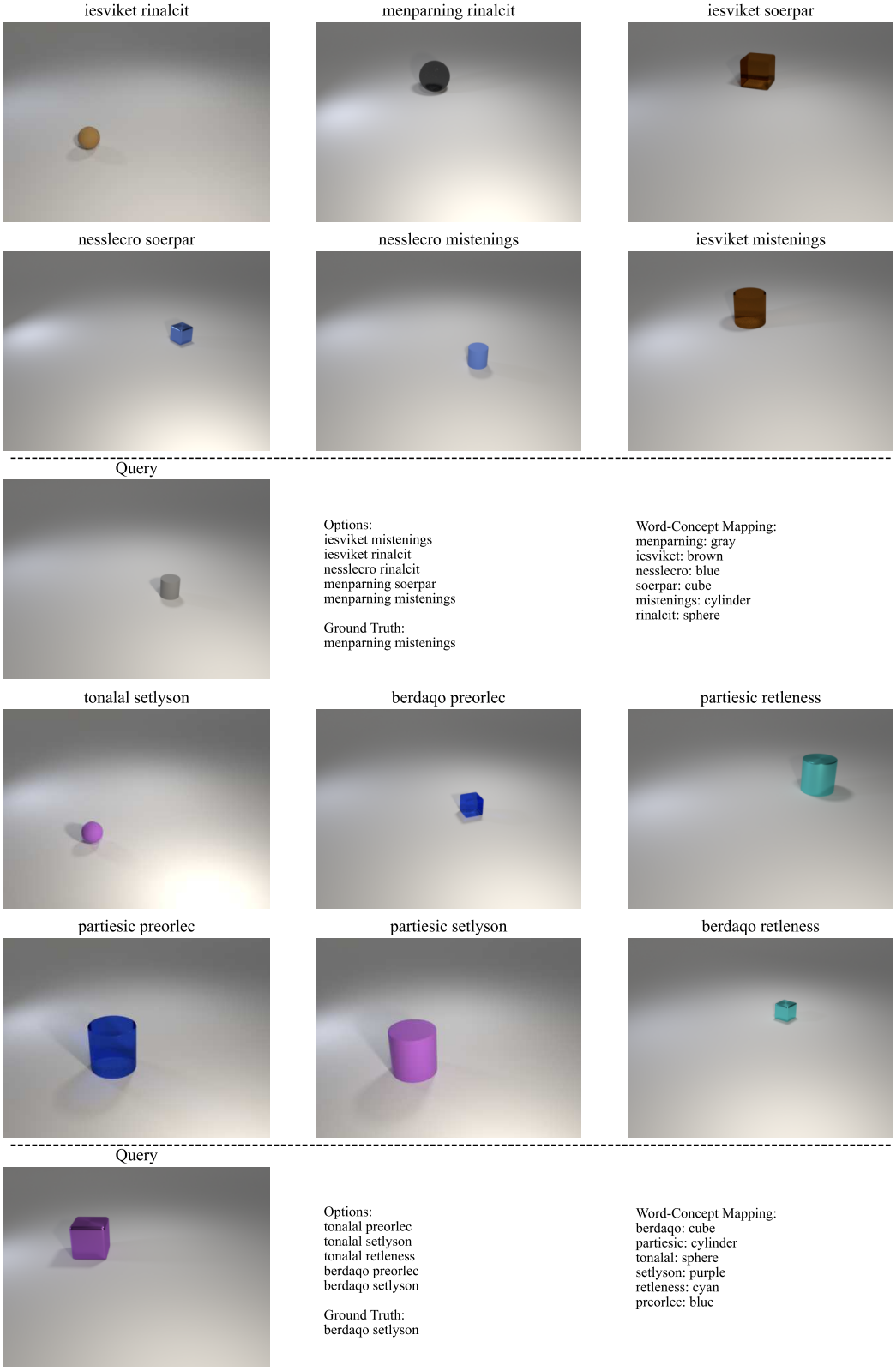


Options:
daythetle and menmatin and someapset
upcation and someapset and peafcol
daythetle and someapset and peafcol
daythetle and upcation and menmatin
upcation and menmatin and peafcol

Ground Truth:
daythetle and upcation and menmatin

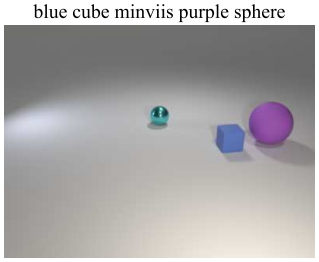
Word-Concept Mapping:
daythetle: ['cylinder', 'purple', 'rubber', 'small']
outsupac: ['cube', 'gray', 'metal', 'small']
upcation: ['cube', 'yellow', 'rubber', 'small']
menmatin: ['sphere', 'purple', 'rubber', 'small']
someapset: ['sphere', 'purple', 'metal', 'large']
peafcol: ['cylinder', 'cyan', 'glass', 'large']

E.5. composite

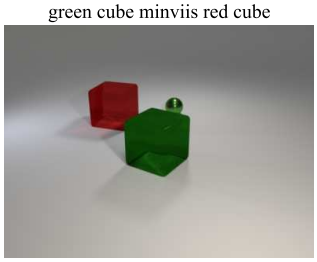


E.6. relation

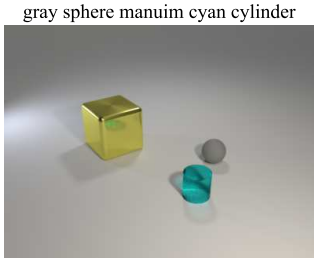
blue cube minviis purple sphere



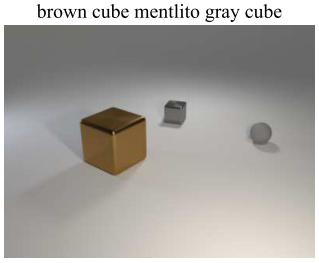
green cube minviis red cube



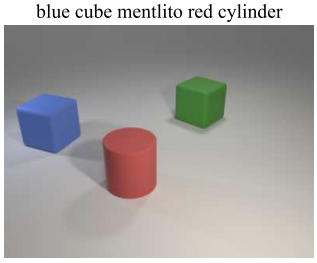
gray sphere manuim cyan cylinder



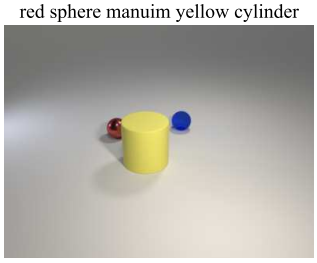
brown cube mentlito gray cube



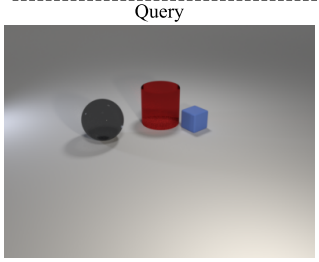
blue cube mentlito red cylinder



red sphere manuim yellow cylinder



Query

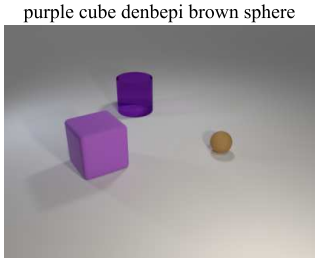


Options:
blue cube mentlito gray sphere
blue cube manuim gray sphere
red cylinder mentlito gray sphere
gray sphere manuim blue cube
blue cube minviis gray sphere

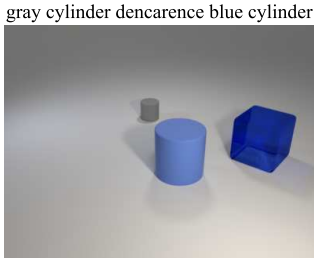
Ground Truth:
blue cube manuim gray sphere

Word-Concept Mapping:
manuim: behind
minviis: front
mentlito: left

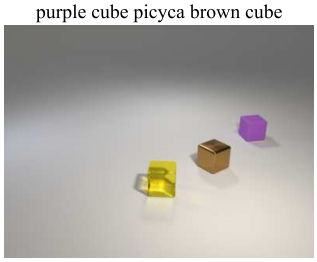
purple cube denbepi brown sphere



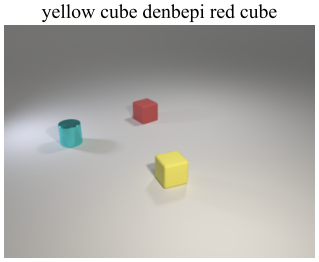
gray cylinder dencarence blue cylinder



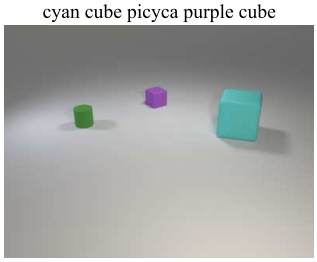
purple cube picyca brown cube



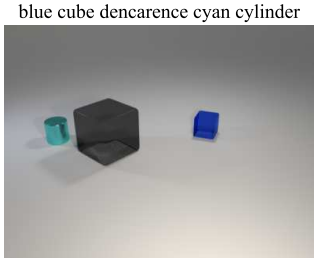
yellow cube denbepi red cube



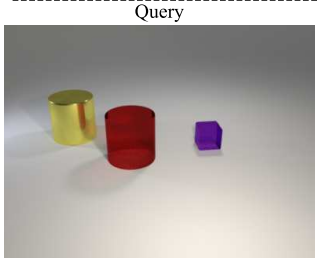
cyan cube picyca purple cube



blue cube dencarence cyan cylinder



Query



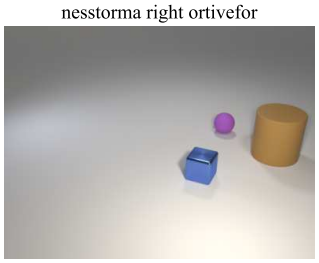
Options:
red cylinder dencarence yellow cylinder
yellow cylinder dencarence purple cube
purple cube denbepi red cylinder
yellow cylinder picyca purple cube
yellow cylinder denbepi purple cube

Ground Truth:
yellow cylinder dencarence purple cube

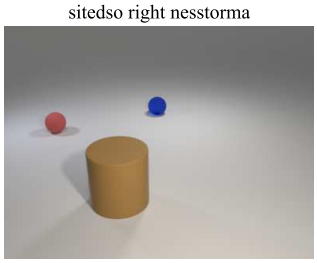
Word-Concept Mapping:
dencarence: behind
picyca: right
denbepi: front

E.7. bootstrap

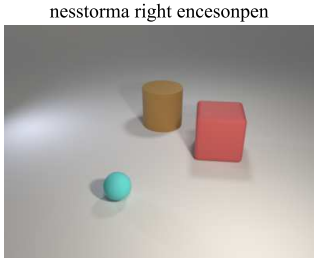
nesstorma right ortivefor



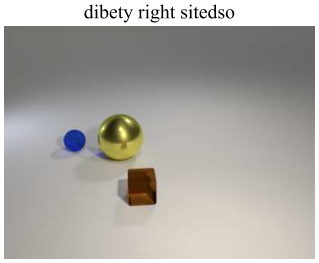
sitedso right nesstorma



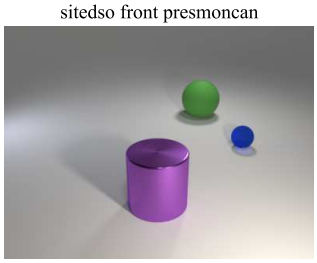
nesstorma right encesonpen



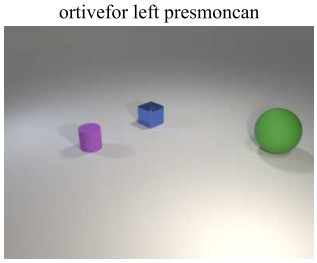
dibety right sitedso



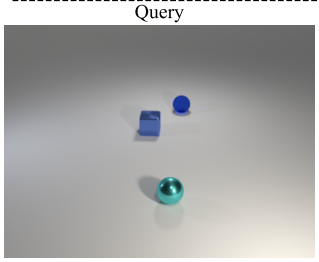
sitedso front presmoncan



ortivefor left presmoncan



Query

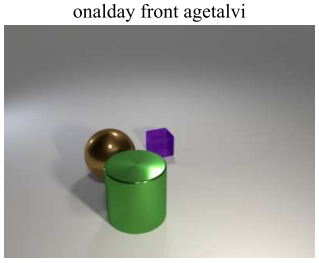


Options:
encesonpen front sitedso
sitedso behind ortivefor
ortivefor right dibety
presmoncan right dibety
encesonpen right dibety

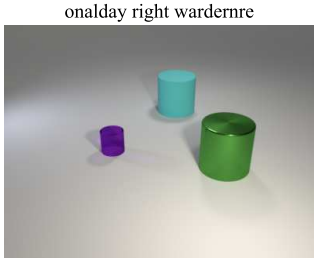
Ground Truth:
sitedso behind ortivefor

Word-Concept Mapping:
encesonpen: ['sphere', 'cyan', 'rubber', 'small']
presmoncan: ['sphere', 'green', 'rubber', 'large']
ortivefor: ['cube', 'blue', 'metal', 'small']
nesstorma: ['cylinder', 'brown', 'rubber', 'large']
sitedso: ['sphere', 'blue', 'glass', 'small']
dibety: ['cube', 'brown', 'glass', 'small']

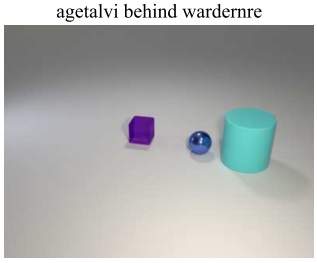
onalday front agetalvi



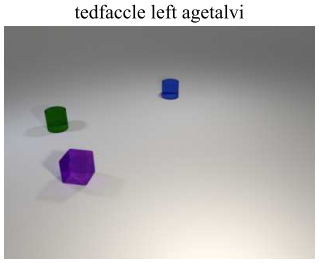
onalday right wardernre



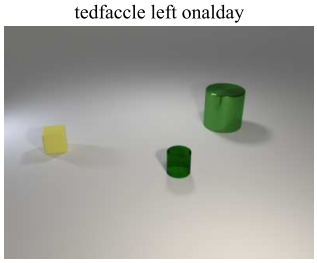
agetalvi behind wardernre



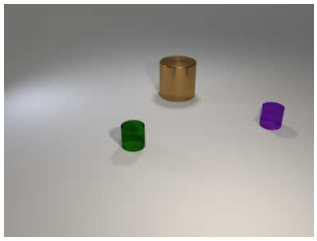
tedfaccle left agetalvi



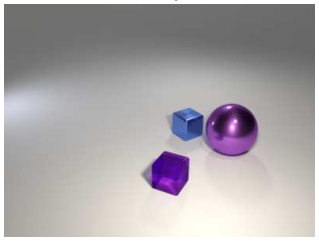
tedfaccle left onalday



tedfaccle front difculo



Query



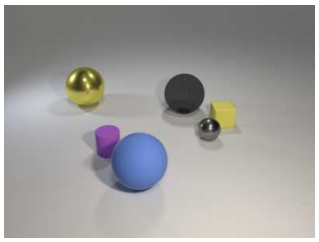
Options:
tionssionno right agetalvi
difculo left onalday
tionssionno left difculo
tionssionno right tedfaccle
wardernre front tedfaccle

Ground Truth:
tionssionno right agetalvi

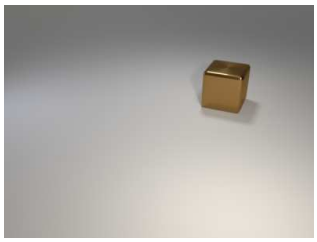
Word-Concept Mapping:
wardernre: ['cylinder', 'cyan', 'rubber', 'large']
agetalvi: ['cube', 'purple', 'glass', 'small']
tionssionno: ['sphere', 'purple', 'metal', 'large']
difculo: ['cylinder', 'brown', 'metal', 'large']
onalday: ['cylinder', 'green', 'metal', 'large']
tedfaccle: ['cylinder', 'green', 'glass', 'small']

E.8. number

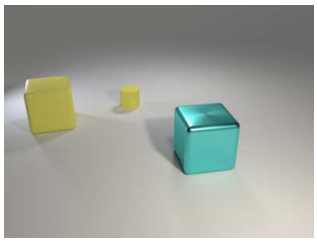
sical



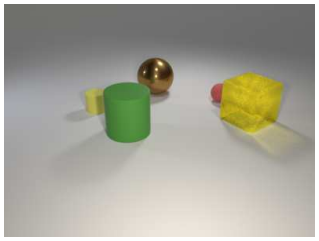
ure



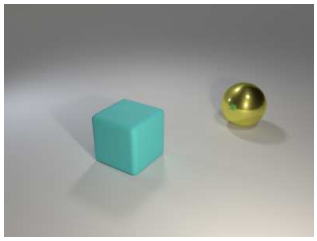
riside



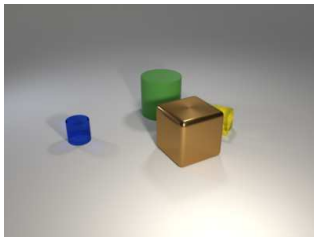
torden



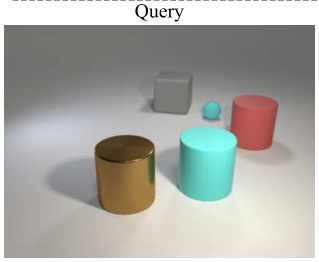
manthe



mentmo



Query

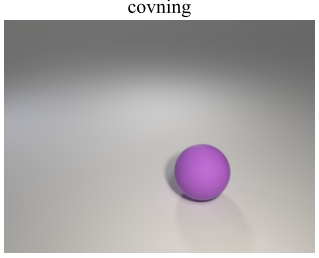


Options:
mentmo
ure
riside
torden
sical

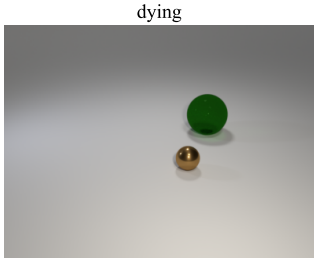
Ground Truth:
torden

Word-Concept Mapping:
ure: 1
manthe: 2
riside: 3
mentmo: 4
torden: 5
sical: 6

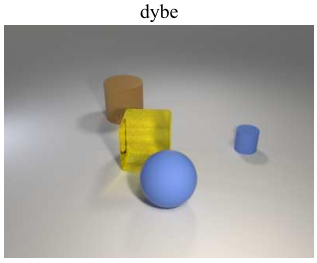
covning



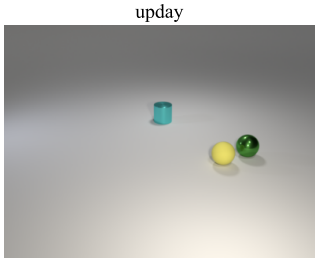
dying



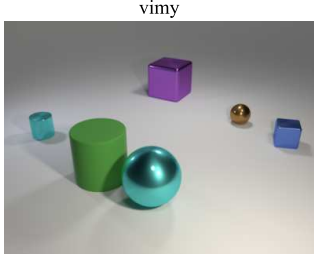
dybe



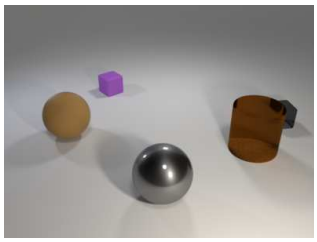
upday



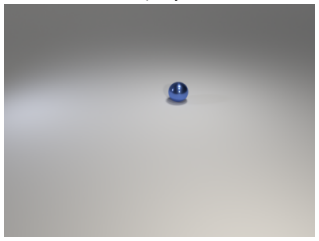
vimy



deneve



Query



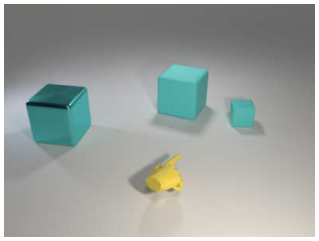
Options:
covning
dying
dybe
deneve
vimy

Ground Truth:
covning

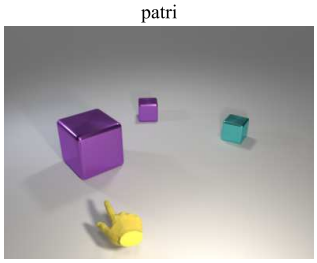
Word-Concept Mapping:
covning: 1
dying: 2
upday: 3
dybe: 4
deneve: 5
vimy: 6

E.9. pragmatic

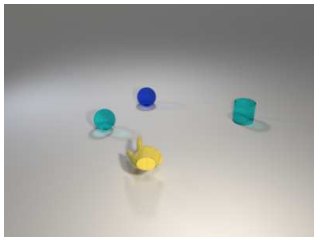
ite



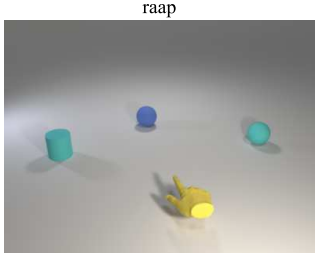
patri



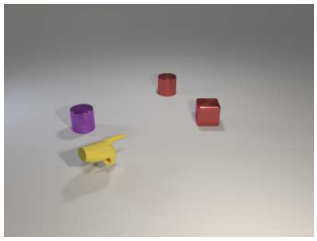
landward



raap

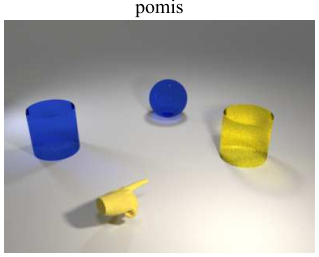


ane



sideex

pomis

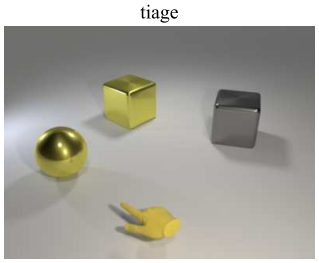


Options:
ingsan
pomis
unout
sideex
sidefa

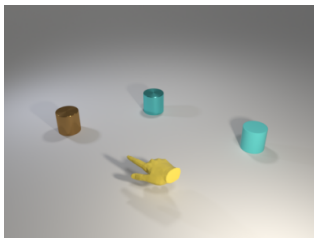
Ground Truth:
sidefa

Word-Concept Mapping:
unout: small
sideex: cylinder
sidefa: sphere
pomis: yellow
raap: blue
ingsan: cube

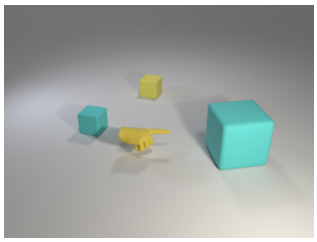
tiage



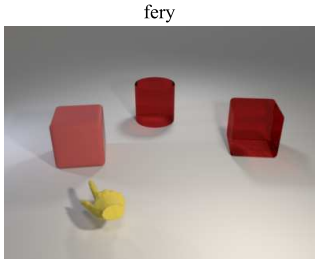
rations



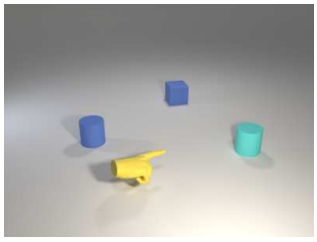
enments



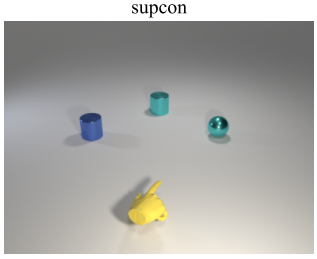
fery



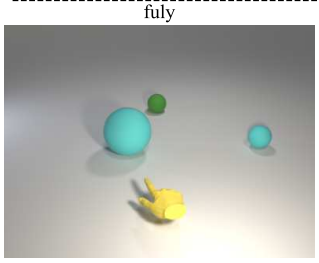
comaf



supcon



fuly



Options:
fuly
mainder
fery
supcon
nupen

Ground Truth:
fuly

Word-Concept Mapping:
supcon: sphere
fuly: large
rations: brown
nupen: gray
fery: rubber
mainder: purple