Human Inspired Progressive Alignment and Comparative Learning for Grounded Word Acquisition

Yuwei Bao[†]   **Barrett Martin Lattimer**[§*]   **Joyce Chai**[†]

[†]Computer Science and Engineering, University of Michigan   [§]ASAPP

{yuweibao, lattimer, chaijy}@umich.edu

Abstract

Human language acquisition is an efficient, supervised, and continual process. In this work, we took inspiration from how human babies acquire their first language, and developed a computational process for word acquisition through comparative learning. Motivated by cognitive findings, we generated a small dataset that enables the computation models to compare the similarities and differences of various attributes, learn to filter out and extract the common information for each shared linguistic label. We frame the acquisition of words as not only the information filtration process, but also as representation-symbol mapping. This procedure does not involve a fixed vocabulary size, nor a discriminative objective, and allows the models to continually learn more concepts efficiently. Our results in controlled experiments have shown the potential of this approach for efficient continual learning of grounded words.

1 Introduction

Two of the important word acquisition problems are: 1) what must be learned to acquire a word, and 2) how to learn the word? To the first question, cognitive studies have shown that several critical steps in learning language naturally comes from joint attention establishment (Tomasello and Farrar, 1986), and symbol grounding (Harnad, 1990). Children's attention are usually redirected through a mother or teacher's guidance, and they learn to map these attended sensory inputs (e.g. color, sound, heat) with their corresponding words or sentences. Living in a rich and diverse world enabled by our multiple body sensors, we learned to filter out the noise and pay attention to specific aspects of an input which we assign linguistic labels to. This attention establishment and information filtration process is the first step of word acquisition. After filtering out the noise, we are left with a mental representation of

what a word entails. Just as the word "car" triggers certain impressions of a common transportation in head, we store these representations as they could come in handy later when we use them to reason, imagine, and express ourselves. To acquire a word, humans learn to filter out noise to focus on key information from sensory inputs that contributes to its meaning (Gentner, 1983; Tomasello and Farrar, 1986), and store that meaning representation for future use (Harnad, 1990; Kuehne et al., 2000).

As for the second question, one of the common but under-explored methods is implicit or explicit comparisons. Caretakers may lay out stuffed animals around a baby and name them one by one to differentiate them. In school, teachers may compare different components of the learning material, e.g. "Today we learn 'colors'. This is red/blue/yellow...". Comparison is the process of finding commonalities and highlighting differences (Gentner and Markman, 1994). It allows children to attend to matching relational structures of inputs (Gentner, 1983), filter out background noise, and learn to generalize as well as abstract. With comparisons, especially clean well-structured comparisons, children can learn a lot of condensed knowledge efficiently and cultivate their capabilities to tackle noisier challenges outside of the classroom (Ball and Thompson, 2018; Anderson et al., 2018; Shao and Gentner, 2019).

From these findings, we propose a new method of word acquisition for artificial intelligent (AI) agents. We mimic the classroom learning setting and constructed a small clean dataset named **SOLA** – **S**imulated **O**bjects for **L**anguage **A**cquisition. This dataset allows the model to draw efficient similarity and difference comparisons, learn to filter out noise, pay attention only to key information that contributes to a word meaning, and store these word-representation mappings continually as new words are introduced. While a larger scale evaluation is needed in the future, through controlled

---

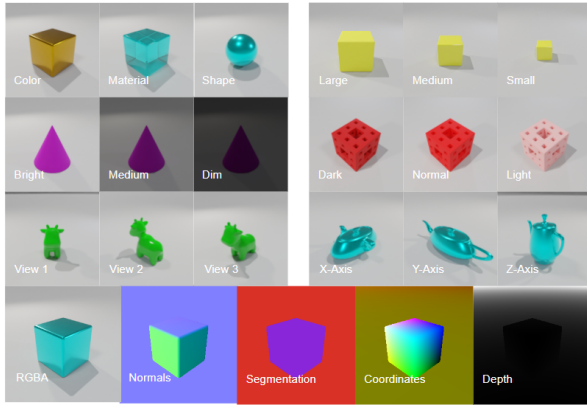* Work done during master study at the University of Michigan.

Figure 1: SOLA Dataset

experiments, our preliminary results have demonstrated the potential of this model in efficient continual learning of grounded words. The dataset and code are available at `https://github.com/sled-group/Comparative-Learning`.

The contributions of this work include:

1. Constructed a small, clean dataset SOLA for studying efficient comparisons.

2. Framed the acquisition of words as both an information filtration process, and as a representation-symbol mapping.

3. Proposed a new method of grounded word acquisition through comparative learning

4. Demonstrated the performance, usability, and generalizability of the acquired representations through multiple tasks.

2 Related Work

2.1 Human Language Acquisition

Language acquisition is the process of putting linguistic labels onto abstracted features, and structuring them into sentences following publicly recognized grammatical rules to express intention. The simple process of abstracting features, takes input attention filtering and relational generalization to pinpoint the learning concept, and associate them with linguistic labels (Harnad, 1990; Kuehne et al., 2000). Studies show that the amount of mother-child joint attention facilitation time is positively correlated to a child's early vocabulary size growth (Tomasello and Farrar, 1986), and that human infants are capable of comparison and abstraction through same/different relation comprehension (Anderson et al., 2018).

Comparison is a central component of human cognition which results in our own uniquely structured knowledge representations (Markman and Gentner, 1993; Gentner and Maravilla, 2017). The theory of Structure-Mapping predicts that similarity comparison allows subjects to attend to matching relational structures of inputs (Gentner, 1983), highlight the differences, and that human infants are able to learn such relations in very few examples (Hespos et al., 2020).

The difficulty of establishing a structural mapping, however, is influenced by the ease of the alignment process (Gentner, 1983). Progressive alignment (Hespos et al., 2020; Kotovsky and Gentner, 1996) suggests that constructing an alignment among highly similar comparisons can invite young children to reason about relational structures and serve as base knowledge for future hierarchical abstractions and complex characteristic learning.

Our work took inspiration from the above two theories by constructing grouped multimodal samples for similarity and difference comparisons. We also start progressive alignment with highly aligned pairings during early word acquisition.

2.2 Continual Learning

There are two major limitations that current neural network models face. Models either take the large-pretrained approach, throwing as much data as possible during training, and hope to learn everything and achieve AGI (Reed et al., 2022) all at once without the need for continual learning. Or models take the architectural (Lomonaco and Maltoni, 2017)/ rehearsal (Lopez-Paz and Ranzato, 2017)/ replay (Parisi et al., 2018)/ regularization (Kirkpatrick et al., 2017) approaches hoping to retain previously learned knowledge amid newly introduced data distribution shift (Ring, 1998; Nguyen et al., 2017; Schlimmer and Fisher, 1986).

Humans, however, are lifelong learners. We are constantly adapting to new environments, learning new concepts & tasks, and evolving together as a society. Human execution of this process is simple, natural, and cost effective, without catastrophically forgetting previously learned knowledge (Van de Ven and Tolias, 2019), nor having to retrain from scratch every time new knowledge is introduced (Schlimmer and Fisher, 1986).

Our method follows the human learning approach and gradually learns more concepts as they are introduced. We demonstrate the model's resistance against catastrophic forgetting and the data learning efficiency in our experiments.

2.3 Contrastive Learning

Contrastive learning is a paradigm that enables models to learn feature representations through contrasting examples without explicit labeling (Chen et al., 2020a; Wu et al., 2018; Khosla et al., 2020). Contrastive learning uses a single contrastive loss function that pushes similar classes together and dissimilar classes apart (Dosovitskiy et al., 2014). In this paper, we introduce **Comparative Learning** which adapts the general definition of contrastive learning by explicitly separating the similarity training from the difference training. On top of encoding each input as in contrastive learning, we took additional steps to further extract information about similarities and differences separately given the same amount of inputs. Supervised by associated words, we use the similarity batches to learn the process of noise filtration and a shared feature representation. We use the difference batches to refine and differentiate these feature representations.

2.4 Multimodal Grounding

A large number of previous works try to draw connections between language and different modalities, such as VILBERT (Lu et al., 2019), LXMERT (Tan and Bansal, 2019), UNITER (Chen et al., 2020b), OSCAR (Li et al., 2020a), Vokenization (Tan and Bansal, 2020), and more (Radford et al., 2021; Wang et al., 2021; Zhang et al., 2021; Cho et al., 2021; Bao et al., 2022; Hill et al., 2020; Tsimpoukelli et al., 2021). These models demonstrated their state of the art multimodal representations on a range of downstream tasks, including image/video captioning, image/text retrieval, visual question answering, and text-to-image generation (Kamath et al., 2021; Wu et al., 2022; Mao et al., 2019; Zheng et al., 2022; Ahmetoglu et al., 2022). A large portion of these works focusd on visual recognition and language production tasks such as captioning, retrieval, and some visual question answering. These works embed visual and textual inputs into the same latent space for similarity comparison and retrieval. These models can learn a great language-vision matching filter, but often do not preserve a grounded concept representation given the linguistic labels.

Another line of works focus on language comprehension and image/video generation. They take a pre-trained language embedding and use it to generate high resolution images, and have achieved extraordinary performance. Notably, (Liu et al., 2022; Du et al., 2021, 2020) achieved composi-

tional visual generation with energy based models. (Brooks et al., 2022) worked on image editing given instructions with paired training images. Also others demonstrated language grounding through compositional text to image generations (Feng et al., 2022; Pearl et al., 2022). These models rely on great grounded language representations to generate meaningful images.

Our work frames the language acquisition process as both input information filtration and representation learning. A few methods include both parts of this definition. CLIP (Radford et al., 2021) used contrastive learning on massive number of weakly linked image-text pairs to project each modality into the same embedding space, which allows the encoders to filter inputs, and store the representations through text embeddings. Several works including (Liu et al., 2022; Du et al., 2021, 2020) used a set of energy based models on recognition tasks for input filtration, and iteratively refined the representations through the Langevin dynamics procedure (Xie et al., 2016). Our work proposes a human inspired approach for word acquisition. We jointly train both the input filtration process and the representations, and map them to their corresponding words through comparative learning.

3 Dataset

Inspired by the classroom teaching setting and the Progressive Alignment theory (Kotovsky and Gentner, 1996), we created a new dataset **SOLA** (**S**imulated **O**bjects for **L**anguage **A**cquisition). SOLA has little noise and clearly defined attributes to isolate different concepts for efficient sample comparisons and grounded language-feature mapping. We generated SOLA using the open-source simulation software Kubric (Greff et al., 2022) designed for semi-realistic image/video synthesis.

SOLA (Figure 1) contains images of individual simulated objects with three associated learning attributes: color, material, and shape. Each object is a composition of one of 8 colors, 11 shapes, and 4 materials. We also diversify the images by capturing each object at 3 different light settings and 6 different camera angles. A total of 6336 Red Green Blue Alpha (RGBA) images were generated. To evaluate the generalizability and robustness of the models on nosier inputs, we also composed a Variation Test set (D_{test_v}) of 989 RGBA images by applying a stretch, shade change, or size transformation. An object in this test set is either stretched along one of the x, y, and z axis, colored with a

darker or lighter shade, or shrunk to a medium or small size. Although not used in this work, we rendered the Depth, Surface Normal, Segmentation Map, and Object Coordinates images for each corresponding RGBA image for future research.

Split	Total	D_{known}	$D_{unknown}$
Vocab Size	23	20	3
D_{train}	5094	3006	2088
D_{test_nc}	1242	744	468
D_{test_v}	989	580	409

Table 1: Splits on RGBA Images

To evaluate the novel composition capability of the methods, we reserved 9 learning attribute pairs exclusively in the Novel Composition Test set (D_{test_nc}). The rest were assembled into the Train set (D_{train}) for word acquisition training. To evaluate models' abilities to continual learning, we split the vocabulary into two sets: a Known vocabulary and an Unknown vocabulary set, which leads to two datasets D_{known} and $D_{unknown}$. The $D_{unknown}$ dataset includes images describable by at least one of the three attributes: [yellow, glass, torus_knot], and the rest of the images are in D_{known}. Each training and testing dataset is broken down into Known and Unknown versions accordingly. More details about SOLA can be found in the Appendix.

Several existing datasets offer dense compositional attribute annotations that can be helpful for language grounding, such as MIT-States (Isola et al., 2015), UT-Zappos (Yu and Grauman, 2014), CUB (Wah et al., 2011), ShapeNet (Chang et al., 2015), Visual Genome (Krishna et al., 2017), and PACO (Ramanathan et al., 2023). These datasets are great resources for scaling attribute concept learning, especially from noisy real world images, but are not designed to form clean structural alignment for comparative language acquisition. Our work took the baby step of progressive alignment (Hespos et al., 2020; Kotovsky and Gentner, 1996; Childers) by offering the model structured and denoised sets of inputs for easier structural comparison and efficient feature extraction. Following these works, we believe that equipping the model with a set of clean base knowledge can help it extend to messier inputs in the future.

Other abstract datasets such as CLEVR (Johnson et al., 2017) focus on diagnosing and probing the reasoning or interpretability of models through visual question and answering, and are not designed for language acquisition. Additionally, our dataset includes 2 more materials and 8 more shapes than CLEVR, providing a lot more variance and opportunities for vocabulary learning. We also diversify lighting, camera angles, and further object transformations in the variation test set for generalization and composition analysis. We introduce SOLA as it offers clean, grouped images for structured comparative learning. More detailed dataset comparisons can be found in Table 3.

4 Method

4.1 Comparative Learning

Comparative Learning is the process of finding the similarities and differences from a set of inputs. It is a general learning strategy that can be applied to different input modalities, sizes, and duration. The general formulation can be found below.

For each label/word/symbol l_i in an unconstrained set $L = \{l_1, l_2, \cdots\}$, we first assemble a batch of samples $\mathcal{B}_s = \{a_1^{l_1}, a_2^{l_1}, \cdots, a_n^{l_1}\}$, that share the label l_i for similarity learning, and a batch of samples $\mathcal{B}_d = \{b^{l_1}, \cdots, b^{l_j}, \cdots\}_{j \neq i}$ that cannot be described by l_i for difference learning. The process of SIM_{l_i} (Eq.1) finds similarities across examples in $\mathcal{B}_s$, and extracts out its representation Rep_{l_i}. The process of DIFF_{l_i} (Eq.2) highlights the differences between l_i and other non-compatible labels, and refines the representation Rep_{l_i}. Non-compatible labels are the ones that cannot be assigned to the same entity at the same time, e.g.(up, down). Comparable to the positive and negative batches in contrastive learning, these labels naturally occur through difference comparisons, and are organized by the supervisor. Both the computations and the representation are stored to map the label: $\{l_i: [\text{SIM}_{l_i}, \text{DIFF}_{l_i}, \text{Rep}_{l_i}]\}$.

$$\text{Rep}_{l_i} = \text{SIM}_{l_i}(\{a^{l_i} \in \mathcal{B}_s\}) \qquad (1)$$

$$\text{Rep}_{l_i} = \text{DIFF}_{l_i}(a^{l_i}, \{b^l \in \mathcal{B}_d\}) \qquad (2)$$

In this work, we contextualize the method of comparative learning in word acquisition through a set of still visual inputs (Figure 2). For each concept, e.g. "red", we assemble a batch of images that share the word "red" for similarity training. We also assemble a batch of images that are of any other color (non-compatible) but "red" for difference refinement. We keep the rest of the attributes the same for better structural alignment.

As illustrated in Algorithm 1 and Figure 2, given a batch of training samples (sim and diff) for word l_i: $\mathcal{B} = \{\mathcal{B}_s, \mathcal{B}_d\}$, we first take a shortcut by having each image a_u go through a frozen pre-trained
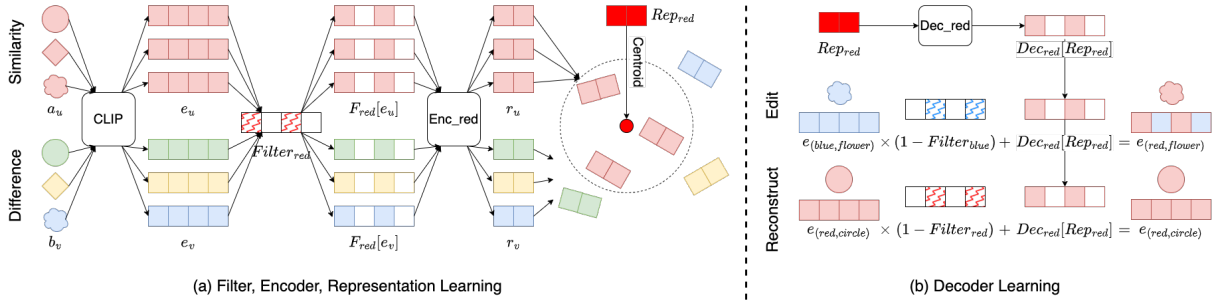
Figure 2: **Model Architecture and Learning Process:** (a) The encoder model learns to extract the shared features from the similarity batch, computes the centroid, and differentiates from the difference batch examples. (b) The decoder model learns to decompress the representation through image reconstruction and editing.

CLIP (Radford et al., 2021) image embedding as the starting point. This shortcut bypasses a few structural alignment steps, and encodes the raw images into the same 512 dimensions e_u available for direct comparisons. The information denoising and attention establishment process is composed of two parts for each word l_i: the filter $\texttt{F}_{l_i}$ and the encoder $\texttt{Enc}_{l_i}$. The filter maintains a vector the same size as the embedding e_u, and computes the element-wise product of the two. It is a learning vector that masks the input embedding by selecting only the relevant dimensions that contributes to the word l_i. This masked embedding goes through two fully connected layers of $\texttt{Enc}_{l_i}$ to output a condensed representation r_u.

On top of learning the attention filtration process ($\texttt{F}_{l_i}, \texttt{Enc}_{l_i}$), we then calculate the centroid of all the sample representations r_u from the similarity batch $\mathcal{B}_s$ as the condensed representation Rep_{l_i} for l_i. For difference learning, we have all the $\mathcal{B}_d$ samples to go through the same filtration and encoding process for word l_i. Since none of them can be described by the word l_i, the output should be nothing like Rep_{l_i}. Therefore, the loss function pushes the distance between each sim batch sample and the centroid close, and pushes the diff batch sample representations apart from the centroid.

This process filters out and abstracts the shared representations of l_i, and differentiates it from other non-compatible words. It jointly trains the filter $\texttt{F}_{l_i}$, the encoder $\texttt{Enc}_{l_i}$, and the representation Rep_{l_i}. We store the mapping $\{l_i: [\texttt{F}_{l_i}, \texttt{Enc}_{l_i}, \text{Rep}_{l_i}]\}$ for each word in memory for later use.

4.2 Generative Decoder Learning

Due to input filtration, the dimensions of the condensed word representations come from selective, word-specific subsets of the original 512 dimensions of e. They are, therefore, not aligned in the same space across different words and cannot be

---

Algorithm 1: Comparative Learning-Word l_i

1 **for** *Sim and Diff data batches:* $\{\mathcal{B}_s, \mathcal{B}_d\}$ **do**
2   // Similarity Learning
3   **for** $a_u \in \mathcal{B}_s$ **do**
4      $e_u = \texttt{CLIP\_emb}[a_u]$
5      $r_u = \texttt{Enc}_{l_i}[\texttt{F}_{l_i}(e_u)]$
6    $\text{Rep}_{l_i} = \texttt{Centroid}[\{r_u\}_{u \in \{1, \cdots, n\}}]$
7   // Difference Learning
8   **for** $b_v \in \mathcal{B}_d$ **do**
9      $e_v = \texttt{CLIP\_emb}[b_v]$
10      $r_v = \texttt{Enc}_{l_i}[\texttt{F}_{l_i}(e_v)]$
11   // Loss
12    $\text{loss}_s = \sum_u \texttt{Dist}[r_u, \text{Rep}_{l_i}]$
13    $\text{loss}_d = \sum_v \texttt{Dist}[r_v, \text{Rep}_{l_i}]$
14    $\text{loss} = (\text{loss}_s)^2 + (1 - \text{loss}_d)^2$
15   Backpropagate and Optimize

Output: $\{l_i: [\texttt{F}_{l_i}, \texttt{Enc}_{l_i}, \text{Rep}_{l_i}]\}$

---

used for direct interactions. To allow compositional reasoning with all the words and their grounded representations, we trained a decoder (Figure 2) to revert the condensed representations back to the same space as the CLIP embedding e.

To train the decoder $\texttt{Dec}_{l_p}$ for word p, we adopted two strategies in parallel: Editing and Reconstruction (Figure 2). About editing, given an image of a (blue, flower), for example, if we filter out blue add red, we should get an image of a (red, flower). Following this logic as in Eq. 3, we mask out feature q from input embedding e_q by multiplying the opposite of filter q: $(1 - \texttt{F}_{l_q})$. We then add back the decoded ($\texttt{Dec}_{l_p}$) representation of Rep_{l_p} for word p. Both the filter $\texttt{F}_{l_q}$ and the representation Rep_{l_p} were trained in the previous step and frozen. The output (out_q2p) aims to resemble the embedding of e_p. Similarly, for reconstruction as in Eq. 4, if we filter out feature p

from input embedding e_p and add back the decoded representation of Rep_{l_p}, we should get the original embedding of e_p. Both passes are trained jointly to learn the decoder of p (Eq. 5).

Each decoder is stored together in the mapping $\{l: [\text{F}_l, \text{Enc}_l, \text{Dec}_l, \text{Rep}_l]\}$. The decoded representations open the door for zero-shot compositional comprehension, generation, and reasoning. For illustration purpose, we also trained a small image generator that upsamples the CLIP embedding back to an RGB image. Details about the models and training can be found in the Appendix.

$$\text{out\_q2p} = e_q(1 - \text{F}_{l_q}) + \text{Dec}_{l_p}[\text{Rep}_{l_p}] \quad (3)$$

$$\text{out\_p2p} = e_p(1 - \text{F}_{l_p}) + \text{Dec}_{l_p}[\text{Rep}_{l_p}] \quad (4)$$

$$\text{loss} = \text{Dist}[e_p, \text{out\_q2p}] + \text{Dist}[e_p, \text{out\_p2p}] \quad (5)$$

5 Experiments

With the training described above, each word will have a mapping $\{l: [\text{F}_l, \text{Enc}_l, \text{Dec}_l, \text{Rep}_l]\}$ stored in the memory. These acquired word representations can be used during inference time for downstream tasks. In this section, we evaluate these representations on several tasks that test models' robustness, generalizability, flexibility, and ability to continual learning.

5.1 Multi-Attribute Recognition

In this task, the models are challenged with zero-shot recognition of all the attributes (color, shape, material) of a given test image a under two evaluation settings: (1) *Novel composition setting* where the image with a combination of attributes which is not seen during training (i.e., $a \in D_{test\_nc}$); and (2) *Noisy setting* where the images were injected with noise in the variation test set ($a \in D_{test\_v}$). The models were trained on the training data (D_{train}). For each test image (Figure 3), we go through the memory, apply the corresponding filter and encoder of each word to the input embedding, and picked the top 3 words with the shortest mean squared error (MSE) between the learned word representation and image embedding.

We took the essence of several zero-shot compositional learning methods such as (Li et al., 2020b; Anwaar et al., 2021; Mancini et al., 2021), and implemented them as variations of the CLIP model for a better experiment control and fairer comparison. More specifically, we compare our method with the following baselines:

CLIP Zero Shot computes the highest matching words for each test image. We experimented with
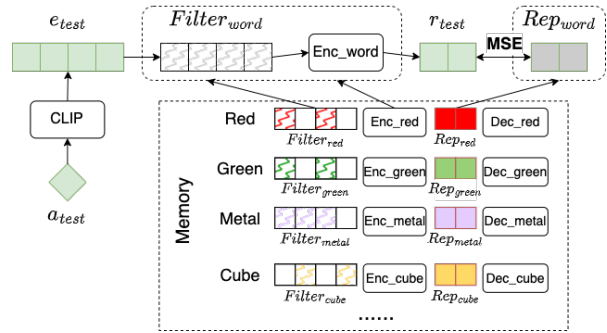

Figure 3: Multi-Attribute Recognition Inference
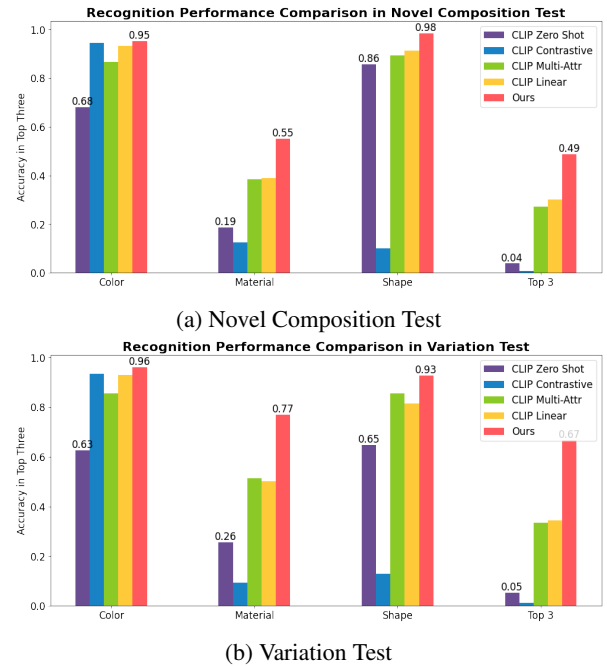

(a) Novel Composition Test

(b) Variation Test

Figure 4: **Multi-Attribute Recognition Performance Comparison**: The percentage of each ground truth attribute (color, material, shape, or all 3) being among the top 3 model predictions.

different prompts, and reported the highest performances using the prompt "a photo of a x".

CLIP Contrastive adds two fully connected layers to the image encoder, and fine tune the model on the same training dataset with a contrastive loss.

CLIP Linear also adds two fully connected layers to the image encoder, but with an output dimension of the vocabulary size. It predicts 1s for all corresponding word dimensions, and 0s otherwise. This method is subject to a fixed vocabulary size, and can be hard to expand to new concepts.

CLIP Multi-Attr finetunes two fully connected layers out of the image encoder for each word, and predicts 1s and 0s based on its confidence measured by the word-image matchability (i.e., similarity).

The performance of all the methods over two test datasets can be found in Figure 4. For each image, we evaluate whether its corresponding color,
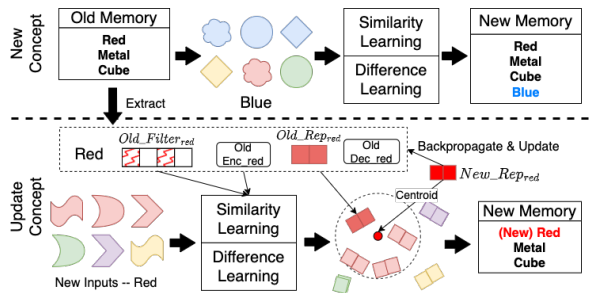
Figure 5: **Continual Learning**: 1) New concepts can be continually added to memory using the same method in Figure 2; 2) Existing concepts can be updated and refined as more samples are introduced.

material, shape, or all three of them are among the top 3 selected words.

It is observed that our method consistently outperforms all baselines across two test datasets and four categories. CLIP Zero Shot showed decent baseline performance on the multi-attribute recognition task, as this model was pre-trained on massive weakly linked language-image pairs. However, our model and the finetuned models are able to surpass this baseline with a significant margin. CLIP Contrastive overfits to the color features, mainly guessing colors in its top three resulting in high color performance but lagging behind in all other attributes. CLIP Linear and CLIP Multi-Attr showed an improved performance compared to CLIP Zero Shot, but couldn't catch up with our method.

Among the 3 attributes, the material attribute was the hardest to learn for all the methods. Humans generally learn textures through touching, tapping an object for sound, and other sensors so a visual input alone may not be sufficient to grasp the meaning of materials, especially under a dim light. However, our method was still able to lead in performance on materials, which consequently also increased the accuracy for all top 3. This is likely because our model is able to pay attention to specific aspects (e.g. light reflection, transparency) of the images better through explicit comparisons.

5.2 Continual Word Acquisition

We investigated models' capability to continually acquire new words on the same multi-attribute recognition task in comparison with the models mentioned in Section 5.1. As mentioned in Section 3, we split all the training and testing datasets into two parts based on the vocabulary (D_{known}, $D_{unknown}$). The D_{known} datasets include 20 words, and the $D_{unknown}$ datasets include an additional 3 new words for continual word ac-
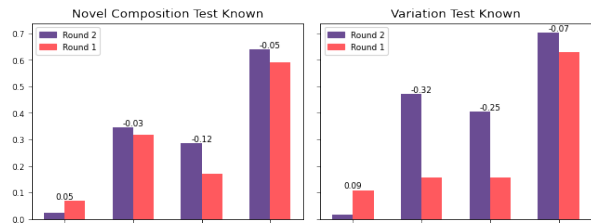
(a) Catastrophic Forgetting on Test Datasets D_{known}
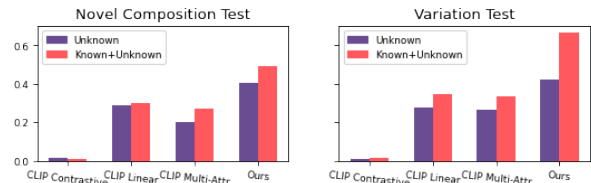
(b) Data Efficiency in Round 2: Trained on $D_{unknown}$ versus $D_{known}+D_{unknown}$

Figure 6: **Continual Word Acquisition Evaluation**: Accuracy of model's top 3 predictions being all ground truth attributes.

quisition and evaluation. Any image that shares at least one of the 3 new words is part of $D_{unknown}$.

Our model conducts continual learning in two ways (Figure 5) it can learn new concepts using the exact same way as described in Figure 2, and add the word-representation mapping to the memory; 2) It can also update and refine the learned concepts, whenever new samples are available. More specifically, we extract the relevant $\{l: [\text{F}_l, \text{Enc}_l, \text{Dec}_l, \text{Rep}_l]\}$ from the memory for word l. The new samples go through similarity and difference learning with the old F_l and Enc_l to get a batch of condensed representation $\{r\}$'s. Together with the old Rep_l, we can calculate a new centroid with these embeddings, and a loss. Through backpropogation and training, the new centroid will be the refined Rep_l, and both the encoder and filter are updated in the memory for word l.

We first evaluate the severity of catastrophic forgetting of the methods (Figure 6a). In Round 1, the models were trained on D_{known} datasets, and the $D_{unknown}$ sets in Round 2. We evaluate the accuracy of the models on two D_{known} test sets by computing the percentage of models' top 3 predictions all being the ground truth attributes.

For CLIP Contrastive, we increased the vocab size for Round 2 training. For CLIP Multi-Attr and our method, we introduced additional models for each new concept. The CLIP Linear model was the hardest to grow as the output dimension was fixed to the previous vocab size. We initialized the first linear weights with the learned weights in Round 1, and had to retrain the model in Round 2.

15481

In Figure 6a, except for the CLIP Contrastive model, most models suffered from catastrophic forgetting between Round 1 and Round 2. Our method had a mild performance decrease as more vocab was introduced. This is likely due to the top 3 label selection competitions among increasing vocab size. CLIP Linear and CLIP Multi-Attr suffered severe catastrophic forgetting on the Variation Test D_{known} set, likely due to lack of generalizability.

We also evaluated the continual training data efficiency for different models. During Round 2, we compare how much replay data would the models need in order to achieve a decent performance by training them on either the $D_{unknown}$ datasets only (new lessons) or both the $D_{known}+D_{unknown}$ datasets (new and old lessons). Round 2 trained on only $D_{unknown}$ receives significantly less data, and does not require reviewing old lessons.

In Figure 6b, when trained only with the $D_{unknown}$ set, our method had already outperformed all other methods even compared to their performances when trained with both $D_{known}+D_{unknown}$ datasets. When more data was available, our method was able to improve performance even further on identifying all attributes.

These results showed early signs of efficient continual learning capabilities and resistance against catastrophic forgetting. Unlike discriminative methods such as CLIP Linear, which has a fixed output dimension based on the vocab size, our method is a lot more flexible to grow for new concepts, and achieved higher performance without the need to review old concepts. Further investigations are needed for larger scale evaluation.

5.3 Compositional Imagination and Reasoning

Another way of evaluating acquired words is through compositional imagination and reasoning given words. With the stored meaning representations, we will be able to flexibly compose different meanings together for reasoning, imagination, simulation, and language understanding. We evaluate this capability in two use cases: composition reasoning and generation.

Most traditional multimodal methods, such as the ones in Section 5.1 only focus on learning a feature extractor given an image. They do not store a grounded representation for each word for reasoning or generation. We therefore, have to turn to the text embedding part of CLIP for comparison as they were trained to be in the same space as the image
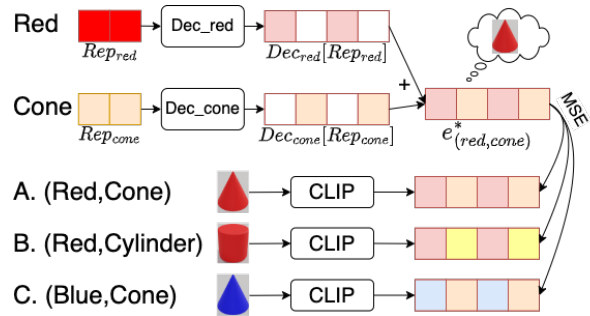

Figure 7: Novel Composition Imagination Multiple Choice Evaluation

embeddings. Text embeddings have been shown to carry grounded semantic meanings through high resolution image generations, but also have been found to struggle at grounding certain attributes (Saharia et al., 2022; Ramesh et al., 2022).

In this section, we compare our method to **CLIP Zero Shot** and **CLIP Finetune** on the following tasks. We use the text embedding of both methods to do image editing and compositional generation. For CLIP Finetune, we added two fully connected layers on the text embedding and fintuned with our D_train dataset.

Composition Generation

Without any given images, humans are able to build mental representations of objects given linguistic descriptions. These representations are built upon abstract word associated features, and can be flexibly composed and manipulated as more features are added. Unlike previous works that emphasize on high resolution image generation, we focus on building compositional mental representations that can be used for downstream reasoning.

Quantitatively, we evaluate the fidelity of a pair of concept composition through a multiple choice task. Given any two random compatible words, e.g. (red, cone), and the CLIP embedding of two hard distractors (each sharing at least one attribute as the original pair, e.g. A. (Red, Cone), B. (Red, Cylinder), C. (Blue, Cone)), we challenge the models to generate a mental representation such that it is closest to the correct image embedding. Each choice is a randomly selected image with the two specified features. As shown in Figure 7, for example, we decode representations of both "red" and "cone" and then add the two resulting vectors to create our "mental image" embedding of a red cone. The multiple choice embedding with the smallest MSE is chosen to be the desired attribute composition imagination.

	C+M	**C+S**	**M+S**	**All**
Zero Shot	53.8±3.2	74.5±4.7	51.0±6.1	59.8±3.3
Finetune	53.9±4.4	75.3±4.2	57.9±6.2	62.4±3.1
Ours	**58.1±5.1**	**82.3±4.4**	**59.7±3.3**	**66.7±2.6**

Table 2: Composition Generation Multiple Choice Task Performance. **C**=Color, **M**=Material, **S**=Shape.
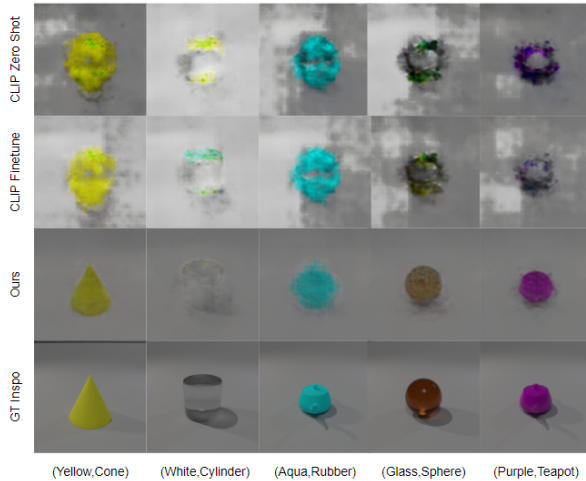
Figure 8: **Novel Composition Generation**: Our method is able to accurately reflect the word, and zero-shot compose mental images with novel feature combinations.

The performance can be found in Table 2. Over 15 runs of 100 randomly selected questions each, our model is able to outperform the text embedding composition of both CLIP Zero Shot and CLIP Text Finetune. Among those, the Color+Shape combo is the easiest one to assemble, likely due to the challenges of learning the material features for the other two combinations. Our method is better at extracting different aspects of the inputs, and learn to pin down the exact word meanings through efficient similarity and difference comparisons.

Qualitatively, we also generated the representations of the novel combinations in our testing data (Figure 8), and see how visually close they are to the GT pictures. The visualization shows that CLIP Zero Shot and CLIP Finetune both struggle at representing the precise definition of some shape concepts, such as "Cone", "Sphere", and "Teapot", but are good at embedding color concepts. The last row serves as an inspiration of possible ground truth images given a pair of concepts.

Composition Reasoning

Another simple compositional reasoning task on different object features is to do 'arithmetic' with them. For example, a (red, cone) - red + blue = (blue, cone). With the Decoder training in Figure 2, we can flexibly edit a given image to a desired feature. In this section, given an input image, and a
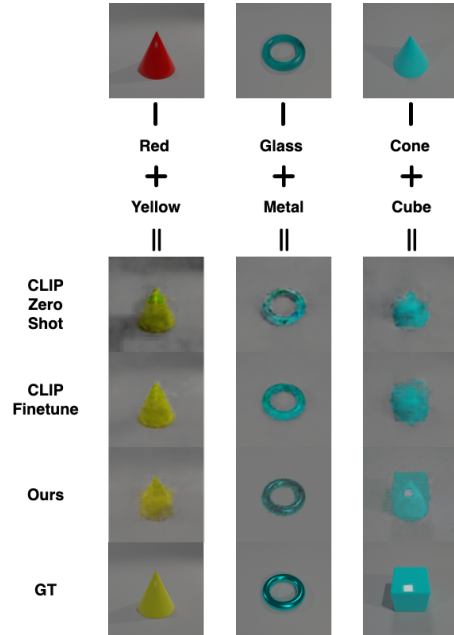

Figure 9: **Composition Reasoning**: Our method can perform image feature editing given grounded word representations, and performs better than CLIP Zero Shot and CLIP Finetune on material and shape editing.

random pair of attribute switching, we qualitatively evaluate if the edited image resembles the desired attribute while keeping all other features the same.

Figure 9 shows three qualitative examples on feature switching over color, material, and shape, compared to CLIP Zero Shot and CLIP Finetune. It is observed that CLIP trained text embeddings excel at extracting color related concepts, but struggle at material and shape. This could be due to its unfamiliarity with the specific material and shape words that we use in our dataset, whereas color concepts are more universal and are easier to learn from pixels. Finetuning helps improve the performance, but still lagged behind our method. More qualitative examples can be found in the Appendix.

6 Conclusion

In this work, we took a human inspired approach to acquire multi-attribute concepts. We define the acquisition of word as learning both an information filtration process, and a representation-symbol mapping. We mimic the classroom setting, constructed a small clean dataset SOLA for efficient comparative and continual learning. We evaluated the learned representations in multi-attribute recognition, compositional simulation and reasoning tasks. Our experiment results outperformed CLIP variations in controlled settings, and showed early signs of a promising new method for continual grounded word acquisition through comparative learning.

Limitations

As exciting as this work is, it does have several limitations and a lot of opportunities for future improvement.

How to scale? We demonstrated our method in a highly constrained environment with very limited concepts, whereas humans are able to pick up new concepts in the noisy world with few shots. How could these representations learned in a clean environment be useful in real world? Would comparative learning still be useful outside of the classroom? We followed the baby step of progressive alignment and hoping that establishing a set of clean base knowledge, can ease the acquisition of future more complex concepts through comparisons with existing knowledge, analogy and hierarchical abstraction. This hypothesis remains to be investigated in the future.

What about other words? Some concepts can be learned through just visual inputs, like color, whereas other concepts require grounding through different sensory types or modalities, like "hot", "loud" and "fast". Even more concepts are built upon existing words through abstraction and generalization, e.g. "philosophy", "momentum". Comparisons can still be used to ground these words, but input to these comparisons could vary from data modalities to computation methods, to abstract representations. We leave these for future work.

How to put words into sentences? This work only focused on the grounding of individual words into visual representations, whereas sentence syntax, grammar, and article structure are yet to be learned. For future work, we could treat language as its own modality, and learn the structure through comparisons as well. Just like in an elementary linguistic class, a teacher would list out several examples "I shower"/"You shower"/"He shower". Humans can learn grammar through what's changing and what's constant. This could be an interesting next step to look into.

Who can offer the supervision? As mentioned at the beginning, human language acquisition is a highly supervised learning process. Babies are rarely inventing new words but learning how adults label objects through generations of conventions. A classroom setting with highly structured curriculum and clean dataset takes a lot of curriculum design and heavy annotation. This is the cost that humans are willing to spend in order to educate human children from kindergarten to college. Maybe it is a fair price that we have to pay in order for artificial intelligence to learn what we want them to learn.

About the current work itself, there are several constraints that we are limited to.

First of all, due to limited computation resources and data size, we had to take a shortcut by using a pre-trained CLIP embedding as a starting point for our models. In theory, we could and would love to train our models from scratch, just like how a new born would learn their first language. A dataset like Toys-200 (Stojanov et al., 2019) could mimic the process of babies interacting with the objects, get a 360 view and help build 3D mental representations.

Second of all, like many other continual learning methods, an unbounded memory space is an unrealistic assumption. As more concepts are learned, the memory space would grow fast, so as the search time. An interesting next step could be to reorganize the memory according to the association distances and hierarchical structures.

Lastly, our work aims at proposing a novel language acquisition definition and the comparative continual learning method. We used somewhat simple model architecture and image generation models for proof-of-concept demonstration on the method. More sophisticated model architecture and training can be switched for different input modalities and applications.

Listed above are several major limitations and future directions based on current work. We are more than happy to take constructive suggestions and criticism to help improve this and future works.

Ethics Statement

This work took the human inspired approach to learn word acquisition for artificial intelligent agents. We generated a small clean dataset using the open-source simulation software Kubric, which was designed for semi-realistic image/video synthesis. All of our training was done on a single machine with 8GB GPU and an Interl i9 processor, with very limited environmental cost. This work does not involve human subject, nor can be can be used to directly interact with humans.

Acknowledgements

References

Alper Ahmetoglu, Erhan Oztop, and Emre Ugur. 2022. Learning multi-object symbols for manipulation with attentive deep effect predictors.

Erin M Anderson, Yin-Juei Chang, Susan Hespos, and Dedre Gentner. 2018. Comparison within pairs promotes analogical abstraction in three-month-olds. *Cognition*, 176:74–86.

Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. 2021. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1140–1149.

Linden J. Ball and Valerie A. Thompson, editors. 2018. *The Routledge international handbook of thinking and reasoning*. Routledge, New York.

Yuwei Bao, Sayan Ghosh, and Joyce Chai. 2022. Learning to mediate disparities towards pragmatic communication.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2022. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Jane B Childers. Language and concept acquisition from infancy through childhood.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. 2014. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27.

Yilun Du, Shuang Li, and Igor Mordatch. 2020. Compositional visual generation with energy based models. *Advances in Neural Information Processing Systems*, 33:6637–6647.

Yilun Du, Shuang Li, Yash Sharma, Josh Tenenbaum, and Igor Mordatch. 2021. Unsupervised learning of compositional energy concepts. *Advances in Neural Information Processing Systems*, 34:15608–15620.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2022. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.

Dedre Gentner and Francisco Maravilla. 2017. Analogical reasoning. In *The Routledge International Handbook of Thinking and Reasoning*, pages 186–203. Routledge.

Dedre Gentner and Arthur B. Markman. 1994. Structural alignment in comparison: No difference without similarity. *Psychological Science*, 5(3):152–158.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. 2022. Kubric: a scalable dataset generator.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42:335–346.

Susan J Hespos, Erin Anderson, and Dedre Gentner. 2020. Structure-mapping processes enable infants' learning across domains including language. In *Language and concept acquisition from infancy through childhood*, pages 79–104. Springer.

Felix Hill, Olivier Tieleman, Tamara Von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark. 2020. Grounded language learning fast and slow. *arXiv preprint arXiv:2009.01719*.

Phillip Isola, Joseph J Lim, and Edward H Adelson. 2015. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. 2021. MDETR - modulated detection for end-to-end multimodal understanding. *CoRR*, abs/2104.12763.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Laura Kotovsky and Dedre Gentner. 1996. Comparison and categorization in the development of relational similarity. *Child Development*, 67(6):2797–2822.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Sven E Kuehne, Dedre Gentner, and Kenneth D Forbus. 2000. Modeling infant learning via symbolic structural alignment. In *Proceedings of the twenty-second annual conference of the cognitive science society*, pages 286–291.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020a. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Y. Li, Y. Xu, X. Mao, and C. Lu. 2020b. Symmetry and group in attribute-object compositions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11313–11322, Los Alamitos, CA, USA. IEEE Computer Society.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. *arXiv preprint arXiv:2206.01714*.

Vincenzo Lomonaco and Davide Maltoni. 2017. Core50: a new dataset and benchmark for continuous object recognition. *CoRR*, abs/1705.03550.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continuum learning. *CoRR*, abs/1706.08840.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

M. Mancini, M. Naeem, Y. Xian, and Z. Akata. 2021. Learning graph embeddings for open world compositional zero-shot learning. *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, (01):1–1.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *CoRR*, abs/1904.12584.

Arthur B Markman and Dedre Gentner. 1993. Structural alignment during similarity comparisons. *Cognitive psychology*, 25(4):431–467.

Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. 2017. Variational continual learning. *arXiv preprint arXiv:1710.10628*.

German Ignacio Parisi, Jun Tani, Cornelius Weber, and Stefan Wermter. 2018. Lifelong learning of spatiotemporal representations with dual-memory recurrent self-organization. *CoRR*, abs/1805.10966.

Ofek Pearl, Itai Lang, Yuhua Hu, Raymond A Yeh, and Rana Hanocka. 2022. Geocode: Interpretable shape programs. *arXiv preprint arXiv:2212.11715*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. 2023. Paco: Parts and attributes of common objects. *arXiv preprint arXiv:2301.01795*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. 2022. A generalist agent. *Transactions on Machine Learning Research*. Featured Certification.

Mark B Ring. 1998. Child: A first step towards continual learning. In *Learning to learn*, pages 261–292. Springer.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding.

Jeffrey C Schlimmer and Douglas Fisher. 1986. A case study of incremental concept induction. In *AAAI*, volume 86, pages 496–501.

Ruxue Shao and Dedre Gentner. 2019. Symmetry: Low-level visual feature or abstract relation? In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, Proceedings of the 41st Annual Meeting of the Cognitive Science Society: Creativity + Cognition + Computation, CogSci 2019, pages 2790–2796. The Cognitive Science Society.

Stefan Stojanov, Samarth Mishra, Ngoc Anh Thai, Nikhil Dhanda, Ahmad Humayun, Chen Yu, Linda B. Smith, and James M. Rehg. 2019. Incremental object learning from contiguous views. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8769–8778.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *arXiv preprint arXiv:2010.06775*.

Michael Tomasello and Michael Jeffrey Farrar. 1986. Joint attention and early language. *Child Development*, 57(6):1454–1463.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Gido M Van de Ven and Andreas S Tolias. 2019. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Tailin Wu, Megan Tjandrasuwita, Zhengxuan Wu, Xuelin Yang, Kevin Liu, Rok Sosič, and Jure Leskovec. 2022. Zeroc: A neuro-symbolic model for zero-shot concept recognition and acquisition at inference time.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. 2016. A theory of generative convnet. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2635–2644, New York, New York, USA. PMLR.

Aron Yu and Kristen Grauman. 2014. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 192–199.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

Qi Zheng, Chaoyue Wang, Dadong Wang, and Dacheng Tao. 2022. Visual superordinate abstraction for robust concept learning.

A Dataset SOLA

Here is a detailed description of the Simulated Objects for Language Acquisition (SOLA) dataset:
Learning Attributes (Figure 10):

1. Color: 8

2. Material: 4

3. Shape: 11

Changing Attributes (Figure 11):

1. Lighting: 3

2. Camera Angle: 6

Variation Attributes (Figure 12):

1. Shade: 3

2. Size: 3

3. Stretch: 4

Image Types:

1. RGBA

2. Depth

3. Surface Normal

4. Segmentation

5. Object Coordinates Coordinates

This amounts to 7325 RGBA images in total, with 6336 originals and 989 with variations. A training and testing split can be found in Table 1. The original image set was first broke down into Novel Composition Training and Novel Composition Testing. 9 pairs of attributes are:
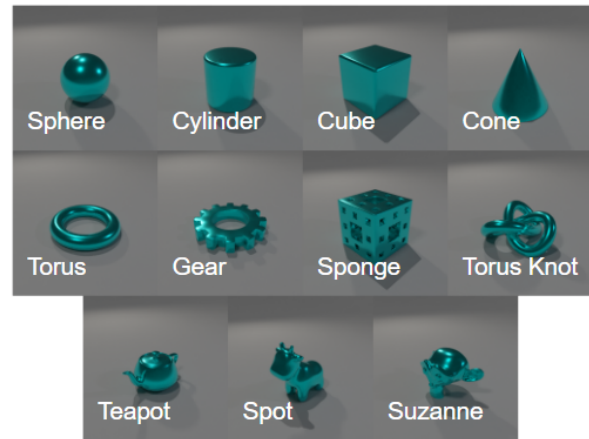
1. (yellow, cone)

2. (green, metal)

3. (plastic, cube)

4. (purple, teapot)

5. (red metal)

6. (glass, torus_knot)

7. (white, cylinder)

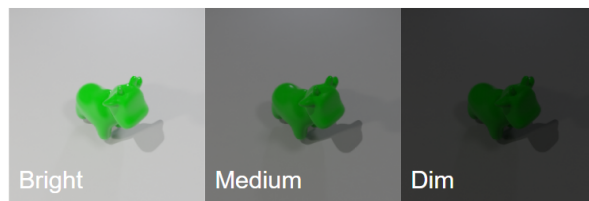8. (aqua, rubber)

9. (glass, sphere)
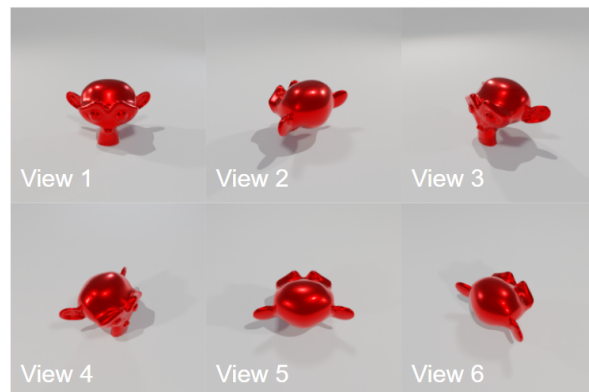

(a) 8 Colors
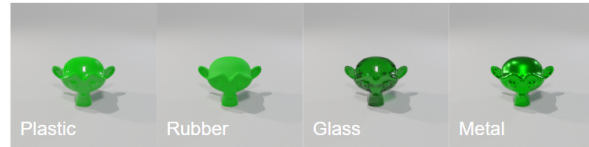
(b) 4 Materials

(c) 11 Shapes

Figure 10: SOLA Learning Attributes

(a) 3 Lighting Conditions

(b) 6 Views

Figure 11: SOLA Environment Augmentation

(a) 3 Shades per Color

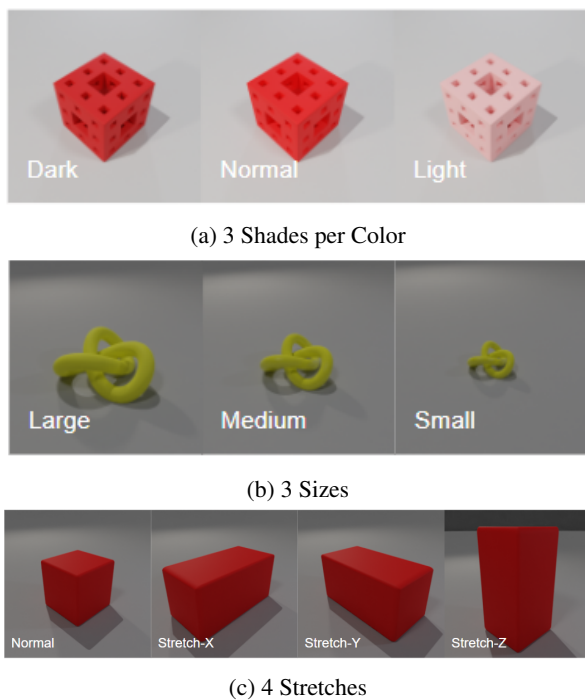(b) 3 Sizes

(c) 4 Stretches

Figure 12: SOLA Variations

For continual learning evaluation, we split the vocabulary into the following two sets. Any images associated at least one of the concepts in $D_{unknown}$ are assembled into $D_{unknown}$ train/test datasets, and the rest in D_{known}. The number of samples in each split can be found in Table 1.

Known = [brown, green, blue, aqua, purple, red, white, rubber, material, plastic, cube, cylinder, sphere, cone, torus, gear, sponge, spot, teapot, suzzane]

Unknown = [yellow, glass, torus_knot]

B Model Architecture and Training

For the encoder training, we used the pretrained CLIP image encoder (frozen) to embed the input images, going through a filter of 512 dimensions, and two fully connected layers with a hidden dimension of 128 and latent dimension of 16. Each round is trained on a similarity batch and a difference batch of size 128 each. The training moves on to the next concept when the loss went down below 0.008 or hit 200 rounds. The whole vocabulary was trained for 5 epochs with a learning rate of $1e^{-3}$.

For the decoder training, we froze the weights of the filter and the pre-trained representations from the previous step, and trained four fully connected layers with a dimension upsampling $16 \rightarrow 64 \rightarrow 64 \rightarrow 96 \rightarrow 512$ with a dropout rate of 0.2. Each concept was trained for 100 round with a batch size of 128. The whole vocabulary was trained for 5

epochs with a learning rate of $1e^{-3}$.

For comparisons, CLIP Contrastive embedded both image inputs, and text inputs. The image embeddings went through two fully connected layers with a hidden dimension of 128 and output dimension of the vocabulary size. CLIP Linear trained two fully connected layers on top of the image embeddings with a hidden dimension of 128 and output dimension of the vocabulary size. CLIP Multi-Attr did the same thing for each word, and the output dimension was 1 over softmax predictions. CLIP Text Finetune trained two fully connected layers on top of the text embeddings, with an input & output dimensions of 512, and hidden dimension of 66. We tried to keep all the model architecture relative the same or having similar number of parameters for a fair comparison. The models were each trained for 50 epochs with a learning rate of $1e^{-3}$.

The small image generator contains 5 upsampling convolution layers with dimensions going from (512,1) to (3,224,224). The number of channels are [128, 64, 32, 16, 3]. We trained 100 epochs on our original dataset with a learning rate or $1e^{-3}$.

All experiments done on a single NVIDIA(R) GeForce(R) RTX 2070 SUPER(TM) 8GB GDDR6 and 10th Gen Intel(R) Core(TM) i9-10900K processor.

C SOLA and Other Dataset Comparisons

| Dataset	Size	Language	Image Type(s)	Purpose	Structural Alignment
CUB	11.8k	Sentences	rgb, bbox	fine-grain classification	No
UT-Zappos	50k	Words	rgb	attribute comparison	No
ShapeNet	51k	Words	3D	3D shapes	No
MIT-States	53k	Words	rgb	state transformation	No
PACO	81.5k	Words	rgb, seg	part segmentation	No
CLEVR	100k	Sentences	rgb	reasoning diagnosis	No
Visual Genome	108k	Sentences	rgb, bbox	question answering	No
SOLA	36.6k	Words	rgba, depth, seg, surf form, obj cords	comparative acquisition	Yes

Table 3: Multimodal Dataset Comparisons
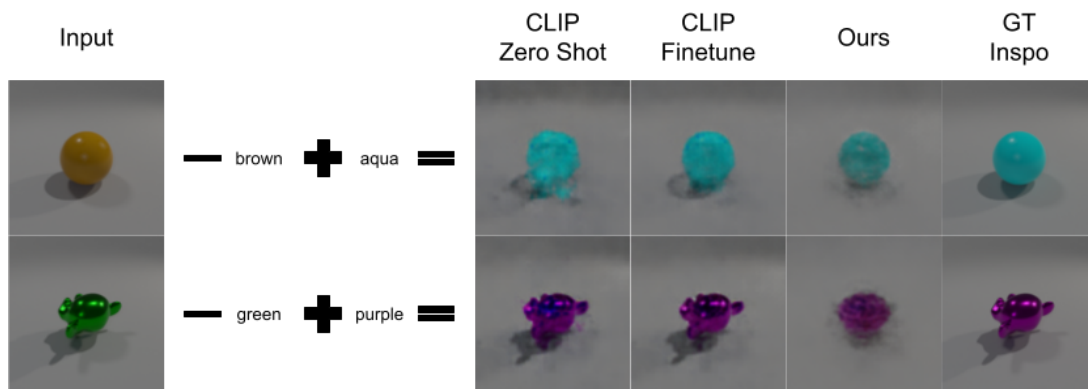
(a) Color Editing
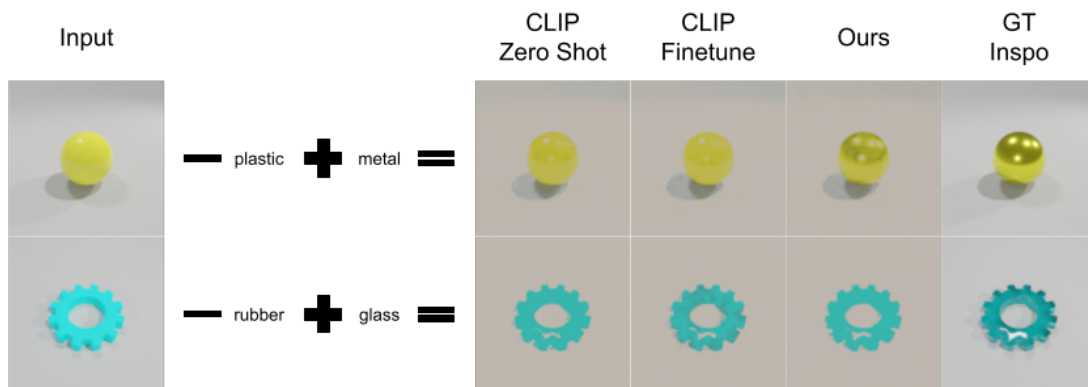
(b) Material Editing
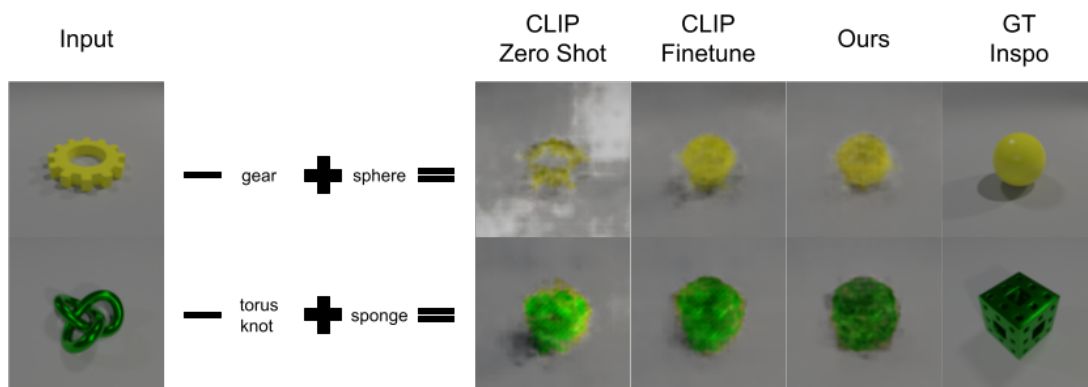
(c) Shape Editing

Figure 13: More Qualitative Examples for Composition Reasoning

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes. Section Limitation.

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes. Section Abstraction and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
☒

B ☑ Did you use or create scientific artifacts?

Yes. Section 3.

☑ B1. Did you cite the creators of artifacts you used?
Yes. Section 3.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Yes. Section 3.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Yes. Section 3.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes. Section 3 and Appendix A.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Yes. Section 3 and Appendix A.

C ☑ Did you run computational experiments?

Yes. Section 5.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Yes. Section 5 and Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. Section 5 and Appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Yes. Section 5.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Yes. Section 5 and Appendix B

D   ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
No.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.