Learning to Learn Words from Visual Scenes

Dídac Surís[1⋆], Dave Epstein[1∗], Heng Ji[2], Shih-Fu Chang[1], and Carl Vondrick[1]

[1]Columbia University [2]UIUC
expert.cs.columbia.edu

Abstract. Language acquisition is the process of learning words from the surrounding scene. We introduce a meta-learning framework that *learns how to learn* word representations from unconstrained scenes. We leverage the natural compositional structure of language to create training episodes that cause a meta-learner to learn strong policies for language acquisition. Experiments on two datasets show that our approach is able to more rapidly acquire novel words as well as more robustly generalize to unseen compositions, significantly outperforming established baselines. A key advantage of our approach is that it is data efficient, allowing representations to be learned from scratch without language pretraining. Visualizations and analysis suggest visual information helps our approach learn a rich cross-modal representation from minimal examples.

1 Introduction

Fig. 1: **What is "ghee" and "roti"?** The answer is in the footnote.[1] Although the words "ghee" and "roti" may be unfamiliar, you are able to leverage the structure of the visual world and knowledge of other words to acquire their meaning. In this paper, we propose a model that learns how to learn words from visual context.

Language acquisition is the process of learning words from the surrounding environment. Although the sentence in Figure 1 contains new words, we are able to leverage the visual scene to accurately acquire their meaning. While this process comes naturally to children as young as six months old [54] and represents a major milestone in their development, creating a machine with the same malleability has remained challenging.

The standard approach in vision and language aims to learn a common embedding space [13,50,27], however this approach has a number of key limitations. Firstly, these models are inefficient because they often require millions of examples to learn. Secondly, they consistently generalize poorly to the natural compositional structure of language [16]. Thirdly, fixed embeddings are unable to adapt to novel words at inference time, such as in realistic scenes that

⋆ Equal contribution

[1] Answer: "ghee" is the butter on the knife, and "roti" is the bread in the pan

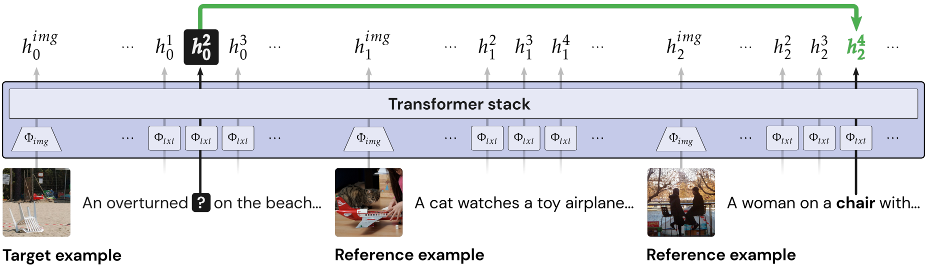


Fig. 2: **Learning to Learn Words from Scenes:** Rather than directly learning word embeddings, we instead learn the *process* to acquire word embeddings. The input to our model is an episode of image and language pairs, and our approach meta-learns a policy to acquire word representations from the episode. Experiments show this produces a representation that is able to acquire novel words at inference time as well as more robustly generalize to novel compositions.

are naturally open world. We believe these limitations stem fundamentally from the process that models use to acquire words.

While most approaches learn the word embeddings, we propose to instead learn the *process* for acquiring word embeddings. We believe the language acquisition process is too complex and subtle to handcraft. However, there are large amounts of data available to learn the process. In this paper, we introduce a framework that *learns how to learn* vision and language representations.

We present a model that receives an episode of examples consisting of vision and language pairs, where the model meta-learns word embeddings from the episode. The model is trained to complete a masked word task, however it must do so by copying and pasting words across examples within the episode. Although this is a roundabout way to fill in masked words, this requires the model to learn a robust process for word acquisition. By controlling the types of episodes from which the model learns, we are able to explicitly learn a process to acquire novel words and generalize to novel compositions. Figure 2 illustrates our approach.

Our experiments show that our framework meta-learns a strong policy for word acquisition. We evaluate our approach on two established datasets, Flickr30k [62] and EPIC-Kitchens [8], both of which have a large diversity of natural scenes and a long-tail word distribution. After learning the policy, the model can receive a stream of images and corresponding short phrases containing unfamiliar words. Our model is able to learn the novel words and point to them to describe other scenes. Visualizations of the model suggest strong cross-modal interaction from language to visual inputs and vice versa.

A key advantage of our approach is that it is able to acquire words with orders of magnitude less examples than previous approaches. Although we train our model from scratch without any language pre-training, it either outperforms or matches methods with massive corpora. In addition, the model is able to effectively generalize to compositions outside of the training set, *e.g.* to unseen compositions of nouns and verbs, outperforming the state-of-the-art in visual language models by over fifteen percent when the compositions are new.

Our primary contribution is a framework that meta-learns a policy for visually grounded language acquisition, which is able to robustly generalize to both new words and compositions. The remainder of the paper is organized around this contribution. In Section 2, we review related work. In Section 3, we present our approach to meta-learn words from visual episodes. In Section 4, we analyze the performance of our approach and ablate components with a set of qualitative and quantitative experiments. We will release all code and trained models.

2 Related Work

Visual language modeling: Machine learning models have leveraged large text datasets to create strong language models that achieve state-of-the-art results on a variety of tasks [10,38,39]. To improve the representation, a series of papers have tightly integrated vision as well [26,48,30,40,27,63,52,7,49,50,2]. However, since these approaches directly learn the embedding, they often require large amounts of data, poorly generalize to new compositions, and cannot adapt to an open-world vocabulary. In this paper, we introduce a meta-learning framework that instead learns the language acquisition process itself. Our approach outperforms established vision and language models by a significant margin. Since our goal is word acquisition, we evaluate both our method and baselines on language modeling directly.

Compositional models: Due to the diversity of the visual world, there has been extensive work in computer vision on learning compositional representations for objects and attributes [34,20,33,36] as well as for objects and actions [22,36,58]. Compositions have also been studied in natural language processing [9,12]. Our paper builds on this foundation. The most related is [24], which also develops a meta-learning framework for compositional generalization. However, unlike [24], our approach works for realistic language and natural images.

Out-of-vocabulary words: This paper is related but different to models of out-of-vocabulary words (OOV) [45,25,18,23,44,43,19,45]. Unlike this paper, most of them require extra training, or gradient updates on new words. We compare to the most competitive approach [45], which reduces to regular BERT in our setting, as a baseline. Moreover, we incorporate OOV words not just as an input to the system, but also as output. Previous work on captioning [28,31,61,59] produces words never seen in the ground truth captions. However, they use pretrained object recognition systems to obtain labels and use them to caption the new words. Our paper is different because we instead learn the word acquisition process from vision and text data. Finally, unlike [4], our approach does not require any side information or external information, and instead acquires new words using their surrounding textual and visual context.

Few-shot learning: Our paper builds on foundational work in few-shot learning, which aims to generalize with little or no labeled data. Past work has explored a variety of tasks, including image classification [60,51,47], translating between a language pair never seen explicitly during training [21] or understanding text from a completely new language [5,1], among others. In contrast, our

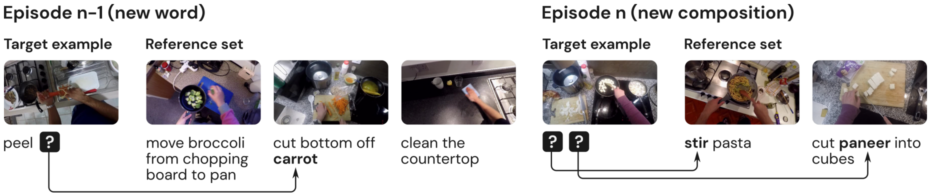


Fig. 3: **Episodes for Meta-Learning**: We illustrate two examples of training episodes. Each episode consists of several pairs of image and text. During learning, we mask out one or more words, indicated by a ?, and train the model to reconstruct it by pointing to ground truth (in **bold**) among other examples within the episode. By controlling the generalization gaps within an episode, we can explicitly train the model to generalize and learn new words and new compositions. For example, the left episode requires the model to learn how to acquire a new word ("carrot"), and the right episode requires the model to combine known words to form a novel composition ("stir paneer").

approach is designed to acquire *language* from minimal examples. Moreover, our approach is not limited to just few-shot learning. Our method also learns a more robust underlying representation, such as for compositional generalization.

Learning to learn: Meta-learning is a rapidly growing area of investigation. Different approaches include learning to quickly learn new tasks by finding a good initialization [14,29], learning efficient optimization policies [46,6,3,41,29], learning to select the correct policy or oracle in what is also known as hierarchical learning [15,19], and others [11,32]. In this paper, we apply meta-learning to acquire new words and compositions from visual scenes.

3 Learning to Learn Words

We present a framework that learns how to acquire words from visual context. In this section, we formulate the problem as a meta-learning task and propose a model that leverages self-attention based transformers to learn from episodes.

3.1 Episodes

We aim to learn the word acquisition process. Our key insight is that we can construct training episodes that demonstrate language acquisition, which provides the data to meta-learn this process. We create training *episodes*, each of which contain multiple *examples* of text-image pairs. During meta-learning, we sample episodes and train the model to acquire words from examples within each episode. Figure 3 illustrates some episodes and their constituent examples.

To build an episode, we first sample a *target example*, which is an image and text pair, and mask some of its word tokens. We then sample *reference examples*, some of which contain tokens masked in the target. We build episodes that require overcoming substantial generalization gaps, allowing us to explicitly meta-learn the model to acquire robust word representations. Some episodes may contain new words, requiring the model to learn a policy for acquiring the word from reference examples and using it to describe the target scene in the episode.

Other episodes may contain familiar words but novel compositions in the target. In both cases, the model will need to generalize to target examples by using the reference examples in the episode. Since we train our model on a distribution of episodes instead of a distribution of examples, and each episode contains new scenes, words, and compositions, the learned policy will be robust at generalizing to testing episodes from the same distribution. By propagating the gradient from the target scene back to other examples in the episode, we can directly train the model to learn a word acquisition process.

3.2 Model

Let an episode be the set $e_k = \{v_1, \ldots, v_i, w_{i+1}, \ldots, w_j\}$ where v_i is an image and w_i is a word token in the episode. We present a model that receives an episode e_k, and train the model to reconstruct one or more masked words w_i by pointing to other examples within the same episode. Since the model must predict a masked word by drawing upon other examples within the same episode, it will learn a policy to acquire words from one example and use them for another example.

Transformers on Episodes: To parameterize our model, we need a representation that is able to capture pairwise relationships between each example in the episode. We propose to use a stack of transformers based on self-attention [56], which is able to receive multiple image and text pairs, and learn rich contextual outputs for each input [10]. The input to the model is the episode $\{v_1, \ldots, w_j\}$, and the stack of transformers will produce hidden representations $\{h_1, \ldots, h_j\}$ for each image and word in the episode.

Transformer Architecture: We input each image and word into the transformer stack. One transformer consists of a multi-head attention block followed by a linear projection, which outputs a hidden representation at each location, and is passed in series to the next transformer layer. Let $H^z \in \mathbb{R}^{d \times j}$ be the d dimensional hidden vectors at layer z. The transformer first computes vectors for queries $Q = W_q^z H^z$, keys $K = W_k^z H^z$, and values $V = W_v^t H^z$ where each $W_* \in \mathbb{R}^{d \times d}$ is a matrix of learned parameters. Using these queries, keys, and values, the transformer computes the next layer representation by attending to all elements in the previous layer:

$$H^{z+1} = SV \quad \text{where} \quad S = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right). \tag{1}$$

In practice, the transformer uses multi-head attention, which repeats Equation 1 once for each head, and concatenates the results. The network produces a final representation $\{h_1^Z, \ldots, h_i^Z\}$ for a stack of Z transformers.

Input Encoding: Before inputting each word and image into the transformer, we encode them with a fixed-length vector representation. To embed input words, we use an $N \times d$ word embedding matrix ϕ_w, where N is the size of the vocabulary considered by the tokenizer. To embed visual regions, we use a convolutional network $\phi_v(\cdot)$ over images. We use ResNet-18 initialized on

ImageNet [42,17]. Visual regions can be the entire image in addition to any region proposals. Note that the region proposals only contain spatial information without any category information.

To augment the input encoding with both information about the modality and the positional information (word index for text, relative position of region proposal), we translate the encoding by a learned vector:

$$
\begin{aligned}
\phi_{\text{img}}(v_i) &= \phi_v(v_i) + \phi_{\text{loc}}(v_i) + \phi_{\text{mod}}(\text{IMG}) + \phi_{\text{id}}(v_i) \\
\phi_{\text{txt}}(w_j) &= \phi_w^T w_j + \phi_{\text{pos}}(w_j) + \phi_{\text{mod}}(\text{TXT}) + \phi_{\text{id}}(w_j)
\end{aligned}
\tag{2}
$$

where ϕ_{loc} encodes the spatial position of v_i, ϕ_{pos} encodes the word position of w_j, ϕ_{mod} encodes the modality and ϕ_{id} encodes the example index.

Please see Appendix C for all implementation details of the model architecture. Code will be released.

3.3 Learning Objectives

To train the model, we mask input elements from the episode, and train the model to reconstruct them. We use three different complementary loss terms.

Pointing to Words: We train the model to "point" to other words within the same episode. Let w_i be the target word that we wish to predict, which is masked out. Furthermore, let $w_{i'}$ be the same word which appears in a reference example in the episode ($i' \neq i$). To fill in the masked position w_i, we would like the model to point to $w_{i'}$, and not any other word in the reference set.

We estimate similarity between the ith element and the jth element in the episode. Pointing to the right word within the episode corresponds to maximizing the similarity between the masked position and the true reference position, which we implement as a cross-entropy loss:

$$
\mathcal{L}_{\text{point}} = -\log\left(\frac{A_{ii'}}{\sum_k A_{ik}}\right) \quad \text{where} \quad \log A_{ij} = f(h_i)^T f(h_j) \tag{3}
$$

where A is the similarity matrix and $f(h_i) \in \mathbb{R}^d$ is a linear projection of the hidden representation for the ith element. Minimizing the above loss over a large number of episodes will cause the neural network to produce a policy such that a novel reference word $w_{i'}$ is correctly routed to the right position in the target example within the episode.

Other similarity matrices are possible. The similarity matrix A will cause the model to fill in a masked word by pointing to another contextual representation. However, we can also define a similarity matrix that points to the input word embedding instead. To do this, the matrix is defined as $\log A_{ij} = f(h_i)^T \phi_w(w_j)$. This prevents the model from solely relying on the context and forces it to specifically attend to the reference word, which our experiments will show helps generalizing to new words.

Word Cloze: We additionally train the model to reconstruct words by directly predicting them. Given the contextual representation of the masked word

h_i, the model predicts the missing word by multiplying its contextual representation with the word embedding matrix, $\hat{w}_i = \phi_w^T h_i$. We then train with cross-entropy loss between the predicted word $\hat{w}_i$ and true word w_i, which we write as $\mathcal{L}_{\text{cloze}}$. This objective is the same as in the original BERT [10].

Visual Cloze: In addition to training the word representations, we train the visual representations on a cloze task. However, whereas the word cloze task requires predicting the missing word, generating missing pixels is challenging. Instead, we impose a metric loss such that a linear projection of h_i is closer to $\phi_v(v_i)$ than $\phi_v(v_{k \neq i})$. We use the tripet loss [57] with cosine similarity and a margin of one. We write this loss as $\mathcal{L}_{\text{vision}}$. This loss is similar to the visual loss used in state-of-the-art visual language models [7].

Combination: Since each objective is complementary, we train the model by optimizing the neural network parameters to minimize the sum of losses:

$$\min_{\Omega} \mathbb{E}\left[\mathcal{L}_{\text{point}} + \alpha \mathcal{L}_{\text{cloze}} + \beta \mathcal{L}_{\text{vision}}\right] \tag{4}$$

where $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$ are scalar hyper-parameters to balance each loss term, and Ω are all the learned parameters. We sample an episode, compute the gradients with back-propagation, and update the model parameters by stochastic gradient descent.

3.4 Information Flow

To use the episode, information needs to flow from reference examples to the target example. Since the transformer computes attention between elements, we can control how information flows in the model by constraining the attention. We implement this as a mask on the attention: $H^{z+1} = (S \odot M)V$ where M_{ij} is a binary mask to indicate whether information can flow from element j to i. Several masks M are possible. Figure 4 visualizes them.

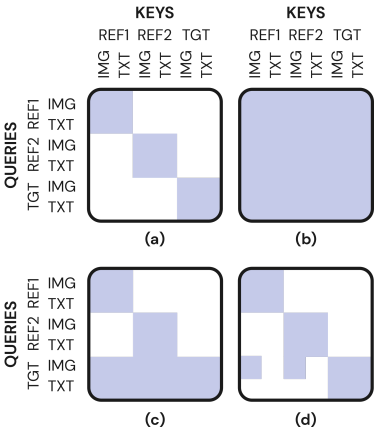


Fig. 4: We visualize some attention masks for how information can flow through our model, where white locations are masked.

(a) Isolated attention: By setting $M_{ij} = 1$ iff i and j belong to the same example in the episode, examples can only attend within themselves. This is equivalent to running each example separately through the model, and optimizing the model with a metric learning loss.

(b) Full attention: By unconditionally setting $M_{ij} = 1$, attention is fully connected and every element can attend to all other elements.

(c) Target-to-reference attention: We can constrain the attention to only allow the target elements to attend to the reference elements, and prevent the reference elements from communicating across each other. To do this, $M_{ij} = 1$ iff i and j are from the same example or i is a target element.

(d) Attention via vision: We can also constrain the attention to only transfer information through vision. Here, $M_{ij} = 1$ if i and j are from the same example, and also $M_{ij} = 1$ if i and j are both images and i is a target element. Otherwise, $M_{ij} = 0$. Information is first propagated from reference text nodes to visual nodes, then propagated from the visual nodes to the target text node.

Some attention mechanisms are more computationally efficient because they do not require computing representations for all pairwise relationships. For full attention, computation scales quadratically with the number of examples. However, for the other attention mechanisms, computation scales linearly with the number of examples, allowing us to efficiently operate on large episodes.

3.5 Inference

After learning, we obtain a policy that can acquire words from an episode consisting of vision and language pairs. Since the model produces words by pointing to them, which is a non-parametric mechanism, the model is consequently able to acquire words that were absent from the training set. As image and text pairs are encountered, they are simply inserted into the reference set. When we ultimately input a target example, the model is able to use new words to describe it by pulling from other examples in the reference set.

Moreover, the model is not restricted to only producing words from the reference set. Since the model is also trained on a cloze task, the underlying model is able to perform any standard language modeling task. In this setting, we only give the model a target example without a reference set. As our experiments will show, the meta-learning objective also improves these language modeling tasks.

4 Experiments

The goal of our experiments is to analyze the language acquisition process that is learned by our model. Therefore, we train the model on vision-and-language datasets, without any language pretraining. We call our approach **EXPERT**.[2]

4.1 Datasets

We use two datasets with natural images and realistic textual descriptions.

EPIC-Kitchens is a large dataset consisting of 39,594 video clips across 32 homes. Each clip has a short text narration, which spans 314 verbs and 678 nouns, as well as other word types. EPIC-Kitchens is challenging due to the complexities of unscripted video. We use object region proposals on EPIC-Kitchens, but discard any class labels for image regions. We sample frames from videos and feed them to our models along with the corresponding narration. Since we aim to analyze generalization in language acquisition, we create a train-test split such that some words and compositions will only appear at test time. We list the full train-test split in the Appendix A.

[2] Episodic Cross-Modal Pointing for Encoder Representations from Transformers

		Ratio		
		1:1	**2:1**	**Cost**
	Chance	13.5	8.7	
	BERT (scratch) [10]	36.5	26.3	-
	BERT+Vision [7]	63.4	57.5	
EXPERT	Isolated attention	69.0	57.8	$O(n)$
EXPERT	Tgt-to-ref attention	71.0	63.2	$O(n)$
EXPERT	Via-vision attention	72.7	64.5	$O(n)$
EXPERT	+ Input pointing	**75.0**	**67.4**	$O(n)$
EXPERT	Full attention	**76.6**	**68.4**	$O(n^2)$
	BERT (pretrained) [10]	53.4	48.8	

Table 1: **Acquiring New Words on EPIC-Kitchens:** We test our model's ability to acquire new words at test time by pointing. The difficulty of this task varies with the number of distractor examples in the reference set. We show **top-1 accuracy** results on both 1:1 and 2:1 ratios of distractors to positives. The rightmost column shows computational cost of the attention variant used.

Flickr30k contains 31, 600 images with five descriptions each. The language in Flickr30k is more varied and syntactically complex than in EPIC-Kitchens, but comes from manual descriptive annotations rather than incidental speech. Images in Flickr30k are not frames from a video, so they do not present the same amount of visual challenges in motion blur, clutter, etc., but they cover a wider range of scene and object categories. We again use region proposals without their labels and create a train-test split that withholds some words and compositions.

Our approach does not require additional image regions as input beyond the full image, and our experiments show that our method outperforms baselines similarly even when trained only with the full image as input, without other cropped regions (see supplementary material).

4.2 Baselines

We compare to established, state-of-the-art models in vision and language, as well as to ablated versions of our approach.

BERT is a language model that recently obtained state-of-the-art performance across several natural language processing tasks [10]. We consider two variants. Firstly, we download the pre-trained model, which is trained on three billion words, then fine-tune it on our training set. Secondly, we train BERT from scratch on our data. We use BERT as a strong language-only baseline.

BERT+Vision refers to the family of visually grounded language models [26,2,50,63,27,48,35,7], which adds visual pre-training to BERT. We experimented with several of them on our tasks, and we report the one that performs the best [7]. Same as our model, this baseline does not use language pretraining.

We also compare several different attention mechanisms. **Tgt-to-ref attention**, **Via-vision attention**, and **Full attention** indicate the choice of attention mask; the base one is the **Isolated attention**. **Input pointing** indicates the choice of pointing to the input encodings in addition to contextual encodings. Unless otherwise noted, EXPERT refers to the variant trained with via-vision attention.

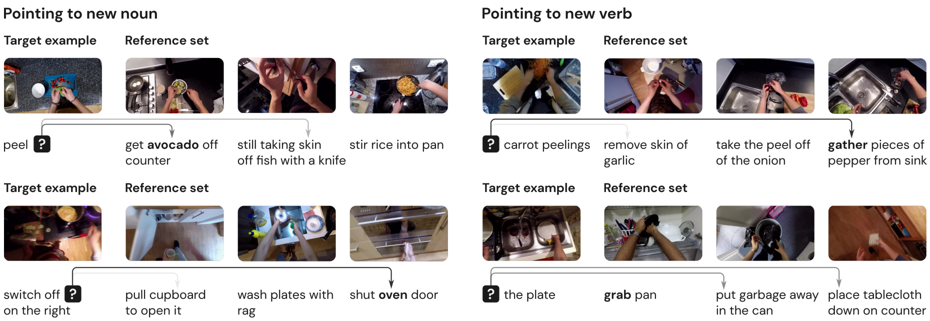
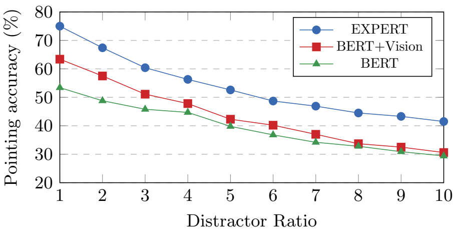


Fig. 5: **Word Acquisition:** We show examples where the model acquires new words. ? in the target example indicates the masked out new word. **Bold** words in the reference set are ground truth. The model makes predictions by pointing into the reference set, and the weight of each pointer is visualized by the shade of the arrows shown (weight $< 3\%$ is omitted). In bottom right, we show an error where the model predicts the plate is being placed, where the ground truth is "grabbed".

Fig. 6: **Word Acquisition versus Distractors:** As more distractors are added (testing on EPIC-Kitchens), the problem becomes more difficult, causing performance for all models to go down. However, EXPERT decreases at a lower rate than baselines.



4.3 Acquisition of New Words

Our model learns the word acquisition *process*. We evaluate this learned process at how well it acquires new words not encountered in the training set. At test time, we feed the model an episode containing many examples, which contain previously unseen words. Our model has learned a strong word acquisition policy if it can learn a representation for the new words, and correctly use them to fill in the right masked words in the target example.

Specifically, we pass each example in an episode forward through the model and store hidden representations at each location. We then compute hidden representation similarity between the masked location in the target example and every example in the reference set. We experimented with a few similarity metrics, and found dot-product similarity performs the best, as it is a natural extension of the attention mechanism that transformers are composed of.

We compare our meta-learned representations to state-of-the-art vision and language representations, i.e. BERT and BERT with Vision. When testing, baselines use the same pointing mechanism (similarity score between hidden representations) and reference set as our model. Baselines achieve strong performance since they are trained to learn contextual representations that have meaningful similarities under the same dot-product metric used in our evaluation.

Method	Ratio 1:1	Ratio 2:1
Chance	3.4	2.3
BERT (scratch)	31.6	25.7
BERT with Vision [7]	32.1	26.8
EXPERT	**69.3**	**60.9**
BERT (pretrained)	**69.4**	**60.8**

Table 2: **Acquiring New Words on Flickr30k:** We run the same experiment as Table 1 (**top-1 accuracy** pointing to new words), except on the Flickr30k dataset, which has more complex textual data. As before, we show results on 1:1 and 2:1 ratios of distractors to positives. By learning the acquisition policy, our model obtains competitive performance with orders of magnitude less training data.

We show results on this experiment in Table 1. Our complete model obtains the best performance in word acquisition on both EPIC-Kitchens and Flickr30k. In the case of EPIC-Kitchens, where linguistic information is scarce and sentence structure simpler, meta-learning a strong lexical acquisition policy is particularly important for learning new words. Our model outperforms the strongest baselines (including those pretrained on enormous text corpora) by up to 13% in this setting. Isolating attention to be only within examples in an episode harms accuracy significantly, suggesting that the interaction between examples is key for performance. Moreover, by constraining this interaction to pass through the visual modality, the computational cost is linear in number of examples with only a minor drop in accuracy. This allows our approach to efficiently scale to episodes with more examples.

Figure 5 shows qualitative examples where the model must acquire novel language by learning from its reference set, and use it to describe another scene with both nouns and verbs. In the bottom right of the figure, an incorrect example is shown, in which EXPERT points to *place* and *put* instead of *grab*. However, both incorrect options are plausible guesses given only the static image and textual context "plate". This example suggests that video information would further improve EXPERT's performance.

Figure 6 shows that, even as the size of the reference set (and thus the difficulty of language acquisition) increases, the performance of our model remains relatively strong. EXPERT outperforms baselines by 18% with one distractor example, and by 36% with ten. This shows that our model remains relatively robust compared to baselines.

In Flickr30k, visual scenes are manually described in text by annotators rather than transcribed from incidental speech, so they present a significant challenge in their complexity of syntactic structure and diversity of subject matter. In this setting, our model significantly outperforms all baselines that train from scratch on Flickr30k, with an increase in accuracy of up to 37% (Table 2). Since text is more prominent, a state-of-the-art language model pretrained on huge (>3 billion token) text datasets performs well, but EXPERT achieves the same accuracy while requiring several orders of magnitude less training data.

4.4 Acquisition of Familiar Words

By learning a policy for word acquisition, the model also jointly learns a representation for the familiar words in the training set. Since the representation is

Table 3: **Acquiring Familiar Words:** We report **top-5 accuracy** on masked language modeling of words which appear in training. Our model outperforms all other baselines.

Method	EPIC-Kitchens Verbs	EPIC-Kitchens Nouns	EPIC-Kitchens All	Flickr30k Verbs	Flickr30k Nouns	Flickr30k All
Chance	0.1	0.1	0.1	< 0.1	< 0.1	< 0.1
BERT (scratch) [10]	68.2	48.9	57.9	64.8	69.4	66.2
BERT with Vision [7]	77.3	63.2	65.6	65.1	70.2	66.5
EXPERT	**81.9**	**73.0**	**74.9**	**69.1**	**79.8**	**72.0**
BERT (pretrained) [10]	71.4	51.5	59.8	**69.5**	**79.4**	**72.2**

Table 4: **Compositionality:** We show top-5 accuracy at predicting masked compositions of seen nouns and verbs. Both the verb and the noun must be correctly predicted. EXPERT achieves the best performance on both datasets.

Method	EPIC-Kitchens Seen	EPIC-Kitchens New	EPIC-Kitchens Diff	Flickr30k Seen	Flickr30k New	Flickr30k Diff
Chance	< 0.1	< 0.1	-	< 0.1	< 0.1	-
BERT (scratch) [10]	34.3	17.7	16.6	43.4	39.4	4.0
BERT with Vision [7]	56.1	37.6	18.5	45.0	42.0	3.0
EXPERT	**63.5**	**53.0**	**10.5**	**48.7**	**47.1**	**1.6**
BERT (pretrained) [10]	39.8	20.7	19.1	**48.8**	**47.2**	**1.6**

trained to facilitate the acquisition process, we expect these embeddings to also be robust at standard language modeling tasks. We directly evaluate them on the standard cloze test [53], which all models (including baselines) are trained to complete.

Table 5 shows performance on language modeling. The results suggest that visual information helps learn a more robust language model. Moreover, our approach, which learns the process in addition to the embeddings, outperforms all baselines by between 4 and 9 percent across both datasets. While a fully pretrained BERT model also obtains strong performance on Flickr30k, our model is able to match its accuracy with orders of magnitude less training data.

Our results suggest that learning a process for word acquisition also collaterally improves standard vision and language modeling. We hypothesize this happens because learning acquisition provides an incentive for the model to generalize, which acts as a regularization for the underlying word embeddings.

4.5 Compositionality

Since natural language is compositional, we quantify how well the representations generalize to novel combinations of verbs and nouns that were absent from the training set. We again use the cloze task to evaluate models, but require the model to predict both a verb and a noun instead of only one word.

We report results on compositions in Table 6 for both datasets. We breakdown results by whether the compositions were seen or not during training. Note that, for all approaches, there is a substantial performance gap between seen and novel compositions. However, since our model is explicitly trained for generalization, the gap is significantly smaller (nearly twice as small). Moreover, our approach also shows substantial gains over baselines for both seen and novel compositions, improving by seven and sixteen points respectively. Additionally, our approach is able to exceed or match the performance of pretrained BERT, even though our model is trained on three orders of magnitude less training data.

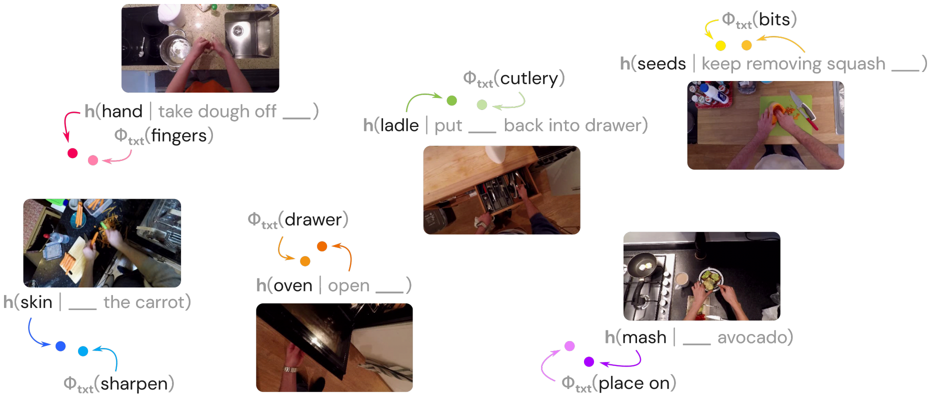


Fig. 7: **Embedding New Words with EXPERT:** We give EXPERT sentences with unfamiliar language at test time. We show the hidden vectors $h(\textit{new word} \mid \textit{context}, \textit{image})$ it produces, conditioned on visual and linguistic context, and their nearest neighbors in word embedding space $\phi_{txt}(\textit{known word})$. EXPERT can use its learned vision-and-language policy to embed new words near other words that are similar in object category, affordances, and semantic properties.

4.6 Analysis

In this section, we analyze *why* EXPERT obtains better performance.

How are new words embedded in EXPERT? Figure 7 shows how EXPERT represents new words in its embedding space at test time. We run sentences which contain previously unseen words through our model. Then, we calculate the nearest neighbor of generated hidden representations of these unseen words in the learned word embedding matrix. Our model learns a representation space such that new words are embedded near semantically similar words (dependent on context), even though we use no such supervisory signal in training.

Does EXPERT use vision? We take our complete model, trained with both text and images, and withhold images at test time. Performance drops to nearly chance, showing that EXPERT uses visual information to predict words and disambiguate between similar language contexts.

What visual information does EXPERT use? To study this, we withhold one visual region at a time from the episode and find the regions that cause the largest decrease in prediction confidence. Figure 8 visualizes these regions, showing that removing the object that corresponds to the target word causes the largest drop in performance. This suggests that the model is correlating these words with the right visual region, without direct supervision.

How does information flow through EXPERT? Our model makes predictions by attending to other elements within its episode. To analyze the learned attention, we take the variant of EXPERT trained with full pairwise attention and measure changes in accuracy as we disable query-key interactions one by one. Figure 9 shows which connections are most important for performance.

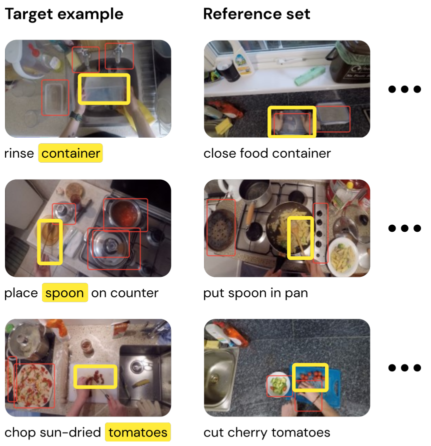


Fig. 8: **Visualizing the Attention:** We probe how the model uses visual information. We remove various objects from input images in an episode, and evaluate the model's confidence in predicting masked words. Removing image regions with a yellow box causes the greatest drop in confidence (other regions are shown in red). The most important visual regions for the prediction task contain an instance of the target word. These results suggest that our model learns some spatial localization of words automatically.

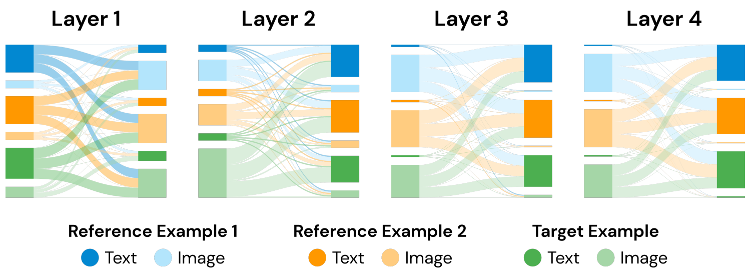


Fig. 9: **Visualizing the Learned Process:** We visualize how information flows through the learned word acquisition process. The width of the pipe indicates the importance of the connection, as estimated by how much performance drops if removed. In the first layer, information tends to flow from the textual nodes to the image nodes. In subsequent layers, information tends to flow from image nodes back to text nodes.

This reveals a strong dependence on cross-modal attention, where information flows from text to image in the first layer, and back to text in the last layer.

How does EXPERT disambiguate multiple new words? We evaluate our model on episodes that contain five new words in the reference set, only one of which matches the target token. Our model obtains an accuracy of 56% in this scenario, while randomly picking one of the novel words would give 20%. This shows that our model is able to discriminate between many new words in an episode. We also evaluate the fine-tuned BERT model in this same setting, where it obtains a 37% accuracy, significantly worse than our model. This suggests that vision is important to disambiguate new words.

5 Discussion

We believe the language acquisition process is too complex to hand-craft. In this paper, we instead propose to meta-learn a policy for word acquisition from visual scenes. Compared to established baselines across two datasets, our experiments show significant gains at acquiring novel words, generalizing to novel compositions, and learning more robust word representations. Visualizations and analysis reveal that the learned policy leverages both the visual scene and linguistic context.

Acknowledgements: We thank Alireza Zareian, Bobby Wu, Spencer Whitehead, Parita Pooj and Boyuan Chen for helpful discussion. Funding for this research was provided by DARPA GAILA HR00111990058. We thank NVidia for GPU donations.

References

1. Adams, O., Makarucha, A., Neubig, G., Bird, S., Cohn, T.: Cross-lingual word embeddings for low-resource language modeling. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 937–947. Association for Computational Linguistics, Valencia, Spain (Apr 2017), https://www.aclweb.org/anthology/E17-1088
2. Alberti, C., Ling, J., Collins, M., Reitter, D.: Fusion of Detected Objects in Text for Visual Question Answering (B2T2) (aug 2019), http://arxiv.org/abs/1908.05054
3. Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M.W., Pfau, D., Schaul, T., Shillingford, B., De Freitas, N.: Learning to learn by gradient descent by gradient descent. In: Advances in neural information processing systems. pp. 3981–3989 (2016)
4. Anne Hendricks, L., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., Darrell, T.: Deep compositional captioning: Describing novel object categories without paired training data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–10 (2016)
5. Artetxe, M., Schwenk, H.: Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. Tech. rep. (2019)
6. Bengio, S., Bengio, Y., Cloutier, J., Gecsei, J.: On the optimization of a synaptic learning rule
7. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: Learning UNiversal Image-TExt Representations. Tech. rep. (2019)
8. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In: The European Conference on Computer Vision (ECCV) (2018), http://youtu.be/Dj6Y3H0ubDw.
9. Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S.J., Goodman, N.D.: Evaluating compositionality in sentence embeddings. arXiv preprint arXiv:1802.04302 (2018)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Tech. rep., https://github.com/tensorflow/tensor2tensor
11. Duan, Y., Schulman, J., Chen, X., Bartlett, P.L., Sutskever, I., Abbeel, P.: Rl2: Fast reinforcement learning via slow reinforcement learning. arXiv preprint arXiv:1611.02779 (2016)
12. Ettinger, A., Elgohary, A., Phillips, C., Resnik, P.: Assessing composition in sentence vector representations. arXiv preprint arXiv:1809.03992 (2018)

13. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. In: European conference on computer vision. pp. 15–29. Springer (2010)
14. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1126–1135. JMLR. org (2017)
15. Frans, K., Ho, J., Chen, X., Abbeel, P., Schulman, J.: Meta learning shared hierarchies. arXiv preprint arXiv:1710.09767 (2017)
16. Gandhi, K., Lake, B.M.: Mutual exclusivity as a challenge for neural networks. arXiv preprint arXiv:1906.10197 (2019)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
18. Herbelot, A., Baroni, M.: High-risk learning: acquiring new word vectors from tiny data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 304–309 (2017)
19. Hu, Z., Chen, T., Chang, K.W., Sun, Y.: Few-shot representation learning for out-of-vocabulary words. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4102–4112 (2019)
20. Johnson, J., Fei-Fei, L., Hariharan, B., Zitnick, C.L., Van Der Maaten, L., Girshick, R.: CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017), `https://arxiv.org/pdf/1612.06890.pdf`
21. Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J.: Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics **5**, 339–351 (2017)
22. Kato, K., Li, Y., Gupta, A.: Compositional Learning for Human Object Interaction. In: ECCV. vol. 11218 LNCS, pp. 247–264 (2018), `http://openaccess.thecvf.com/content{_}ECCV{_}2018/papers/Keizo{_}Kato{_}Compositional{_}Learning{_}of{_}ECCV{_}2018{_}paper.pdf`
23. Khodak, M., Saunshi, N., Liang, Y., Ma, T., Stewart, B.M., Arora, S.: A la carte embedding: Cheap but effective induction of semantic feature vectors. In: 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018. pp. 12–22. Association for Computational Linguistics (ACL) (2018)
24. Lake, B.M.: Compositional generalization through meta sequence-to-sequence learning. In: NeurIPS (2019)
25. Lazaridou, A., Marelli, M., Baroni, M.: Multimodal word meaning induction from minimal exposure to natural text. Cognitive science **41**, 677–705 (2017)
26. Li, G., Duan, N., Fang, Y., Jiang, D., Zhou, M.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. ArXiv **abs/1908.06066** (2019)
27. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: VisualBERT: A Simple and Performant Baseline for Vision and Language. Tech. rep. (2019), `http://arxiv.org/abs/1908.03557`
28. Li, Y., Yao, T., Pan, Y., Chao, H., Mei, T.: Pointing novel objects in image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12497–12506 (2019)
29. Li, Z., Zhou, F., Chen, F., Li, H.: Meta-sgd: Learning to learn quickly for few-shot learning. arXiv preprint arXiv:1707.09835 (2017)

30. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In: Neural Information Processing Systems (NeurIPS), 2019 (2019), http://arxiv.org/abs/1908.02265
31. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7219–7228 (2018)
32. Mishra, N., Rohaninejad, M., Chen, X., Abbeel, P.: A simple neural attentive meta-learner. arXiv preprint arXiv:1707.03141 (2017)
33. Misra, I., Gupta, A., Hebert, M.: From red wine to red tomato: Composition with context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1792–1801 (2017)
34. Nagarajan, T., Grauman, K.: Attributes as Operators: Factorizing Unseen Attribute-Object Compositions. European Conference on Computer Vision (ECCV) (2018), https://arxiv.org/pdf/1803.09851.pdf
35. Nangia, N., Bowman, S.R.: Human vs. muppet: A conservative estimate of human performance on the glue benchmark. arXiv preprint arXiv:1905.10425 (2019)
36. Nikolaus, M., Abdou, M., Lamm, M., Aralikatte, R., Elliott, D.: Compositional generalization in image captioning. In: CoNLL (2018)
37. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
38. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
39. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training
40. Rahman, W., Hasan, M.K., Zadeh, A., Morency, L.P., Hoque, M.E.: M-BERT: Injecting Multimodal Information in the BERT Structure. Tech. rep. (2019), http://arxiv.org/abs/1908.05787
41. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning (2016)
42. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
43. Schick, T., Schütze, H.: Attentive mimicking: Better word embeddings by attending to informative contexts. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 489–494 (2019)
44. Schick, T., Schütze, H.: Learning semantic representations for novel words: Leveraging both form and context. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6965–6973 (2019)
45. Schick, T., Schütze, H.: Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. arXiv preprint arXiv:1904.06707 (2019)
46. Schmidhuber, J.: Evolutionary Principles in Self-Referential Learning. On Learning now to Learn: The Meta-Meta-Meta...-Hook. Diploma thesis, Technische Universitat Munchen, Germany (14 May 1987), http://www.idsia.ch/~juergen/diploma.html
47. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4077–4087. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/6996-prototypical-networks-for-few-shot-learning.pdf

48. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: Pre-training of Generic Visual-Linguistic Representations. Tech. rep. (2019), http://arxiv.org/abs/1908.08530
49. Sun, C., Baradel, F., Murphy, K., Schmid, C.: Contrastive Bidirectional Transformer for Temporal Representation Learning. Tech. rep. (2019)
50. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: VideoBERT: A Joint Model for Video and Language Representation Learning (apr 2019), http://arxiv.org/abs/1904.01766
51. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
52. Tan, H., Bansal, M.: LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (aug 2019), http://arxiv.org/abs/1908.07490
53. Taylor, W.L.: “cloze procedure”: A new tool for measuring readability. Journalism Bulletin **30**(4), 415–433 (1953)
54. Tincoff, R., Jusczyk, P.W.: Some beginnings of word comprehension in 6-month-olds. Psychological science **10**(2), 172–175 (1999)
55. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-sne. Journal of Machine Learning Research **9**, 2579–2605 (2008)
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (2017)
57. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research **10**(Feb), 207–244 (2009)
58. Wray, M., Larlus, D., Csurka, G., Damen, D.: Fine-grained action retrieval through multiple parts-of-speech embeddings. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
59. Wu, Y., Zhu, L., Jiang, L., Yang, Y.: Decoupled novel object captioner. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 1029–1037 (2018)
60. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning - the good, the bad and the ugly. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
61. Yao, T., Pan, Y., Li, Y., Mei, T.: Incorporating copying mechanism in image captioning for learning novel objects. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6580–6588 (2017)
62. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
63. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J.J., Gao, J.: Unified Vision-Language Pre-Training for Image Captioning and VQA. Tech. rep. (2019), https://github.com/LuoweiZhou/VLP.

Appendix

A EPIC-Kitchens Train-Test Split

The EPIC-Kitchens dataset does not provide action narrations for its test set, which we need for evaluation. We therefore create our own train-test split from the dataset following their conventions. We aim to generate an 80-20 train-test split. We select a small subset of verbs and nouns to remove entirely from the training set, as well as verb-noun compositions. We remove around 4% of nouns and verbs and withhold them for testing, and around 10% of all compositions. These are not uniformly distributed, and removing a verb or noun entirely from the training set often results in withholding a disproportionate amount of data. Therefore, with these parameters, we end up with around a 81-19 split. We will release this dataset splits for others to build on our this work.

Below, we list the words and compositions that are withheld from the training set, but present during test.

List of new nouns:

1. avocado
2. counter
3. hand
4. ladle
5. nesquik
6. nut
7. oven
8. peeler
9. salmon
10. sandwich
11. seed
12. onion
13. corn

List of new verbs:

1. fry
2. gather
3. grab
4. mash
5. skin
6. watch

List of new verb/noun compositions:

Since there are around 400 new compositions, we include only the 20 most common here.

1. close oven
2. cut peach
3. dry hand
4. fry in pan
5. grab plate
6. open cupboard
7. open oven
8. pick up sponge
9. put onion
10. put in oven
11. put spoon
12. remove garlic
13. rinse hand
14. skin carrot
15. stir pasta
16. take plate
17. wash hand
18. wash knife
19. wipe counter
20. wipe hand

Since the Flickr30k dataset has a larger vocabulary than EPIC-Kitchens and thus a larger train-test split, we only list here the new verbs and nouns, and not the new compositions.

Flickr30k new nouns: 3d, a, A&M, ad, Adidas, Africa, AIDS, airborne, Airways, Alaska, America, american, Americans, Amsterdam, Angeles, Asia, ax, B, badminton, bale, barrier, Batman, batman, Bay, be, Beijing, Berlin, big, Birmingham, Blue, Boston, bottle, Brazil, Bridge, Britain, British, british, Bush, c, California, Calvin, Canada, canadian, Canyon, card, Carolina, case, catholic, cause, cd, cello, Celtics, Chicago, China, chinatown, Chinese, chinese, christmas, Circus, Claus, Clause, clipper, co, cocktail, colleague, Coney, content, convertible, courtyard, Cruz, cup, cycle, dance, David, Deere, desk, Dior, Disney, dj, Domino, dragster, drum, dryer, DS, dump, east, Easter, eastern, Eiffel, Elmo, elmo, England, Europe, fellow, Florida, Ford, France, Francisco, Gate, gi, Giants, glove, go, God, Golden, golden, graduate, Grand, Great, Haiti, halloween, hedge, Heineken, helicopter, hi, hide, highchair, hispanic, Hollister, Hollywood, hoop, Houston, iMac, India, Indian, Indians, information, ingredient, Iowa, Island, Israel, Italy, Jackson, jam, Japan, Jesus, Jim, Joe, John, Klein, knoll, La, Lakers, Las, latex, layup, legged, liberty, library, Lincoln, locomotive, London, loom, Los, Lynyrd, ma, mariachi, Mets, Mexico, Miami, Michael, Michigan, Mickey, mickey, Mike, Miller, mind, Morgan, Mr, Mrs, Music, muslim, Navy, New, new, NFL, Nintendo, no, nun, NY, NYC, o, Obama, officer, Oklahoma, old, op, oriental, ox, Oxford, p, Pabst, Pacific, Paris, Patrick, Paul, pavement, pc, Penn, pew, piercing, pig, plane, pot, punk, rafting, rain, razor, Red, Renaissance, repairman, research, robe, Rodgers, runway, RV, S, Salvation, San, saris, scrimmage, scrubs, Seattle, second, shooting, shrine, Skynyrd, something, South, Sox, Spain, Spanish, Square, SquarePants, St, St, start, States, Statue, Story, style, superman, surfs, T, t, tech, Texas, texas, the, Thomas, Times, today, tooth, Toronto, Toy, trinket, tv, type, U, UFC, UK, ultimate, Unicef, United, up, US, USA, v, Vegas, Verizon, Volvo, vw, W, Wall, Wars, Washington, Wells, West, White, Wii, wind, windsurfer, Winnie, winter, Wonder, wonder, x, Yankees, yard, yo, yong, York, york.

Flickr30k new verbs: address, am, amused, applaud, armed, baked, bar, be, bearded, blend, boat, bound, broken, build, button, cling, complect, confused, cooked, costumed, covered, crashing, crowded, crumble, darkened, decorated, deflated, do, dyed, fallen, fenced, file, flying, go, goggle, graze, haircut, handicapped, inflated, injured, juggle, listening, lit, living, looking, magnify, measure, mixed, motorize, mounted, muzzle, muzzled, numbered, oncoming, opposing, organized, patterned, pierced, populated, proclaim, puzzle, restrain, rim, saddle, scratch, seated, secure, sell, shaved, skinned, slump, solder, spotted, sprawl, streaked, strike, stuffed, suited, sweeping, tanned, tattooed, tile, tiled, train, uniformed, up, wheeled, woode, wooded.

B Language Model Results

We show EXPERT's outputs when given a sentence containing a new composition of verb and noun. The verb and noun are masked, and we ask EXPERT to make language model predictions at these locations (as in the standard BERT cloze task setting). Results are shown in Figure 12.

In addition, we provide detailed language modeling performance figures in Tables 5 and 6, breaking accuracy down by model variant.

C Implementation Details

For all our experiments we use four transformers ($Z = 4$) and four heads. All the models are trained with the Adam optimizer, with a learning rate of 3×10^{-5} and $\epsilon = 0.0001$. In our experiments, optimization typically takes one week on a single GPU.

In training, we mask out text tokens $\frac{1}{3}$ of the time and image tokens $\frac{1}{6}$ of the time. Following [10], a masked text token gets assigned a special `[MASK]` token 80% of the time, a random word token 10%, and remains unchanged 10%.

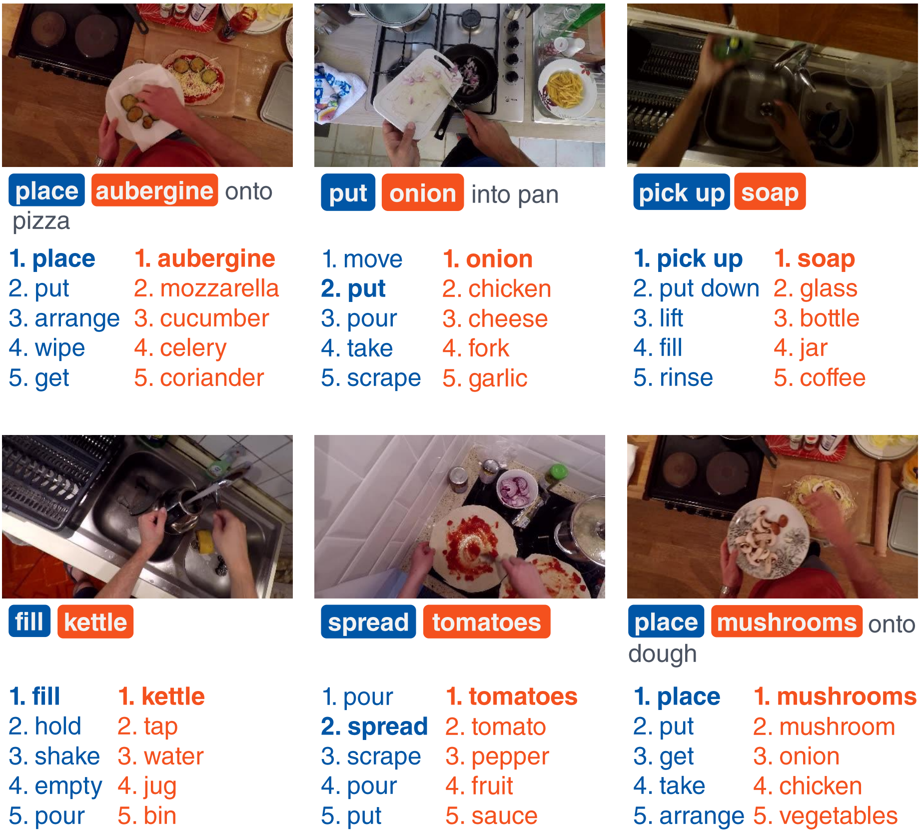


Fig. 10: **Predictions of new compositions:** We show some examples of our model's ability to generalize to new compositions, given only the images shown (and their bounding boxes), as well as the unmasked words in the sentence. We show the top five predictions for each word. As in the paper, we indicate masked words with the dark box, and their predictions below it.

Similarly, we zero out image tokens 90% of the time and leave them unaltered 10%.

We construct episodes by first randomly sampling a target example from the training set. We then randomly select between 0 and $k_+ = 2$ text tokens as targets, which will be masked with probability 1. For each one of these tokens, we randomly add to the episode another example in the training set whose text contains the token. We then randomly add between 0 and $k_{-} = 2$ negative examples (distractors) to the episode, which do not contain any of the target tokens in their text.

We randomly shuffle examples in an episode before feeding them to the model. Then, we combine them with indicator tokens to demarcate examples: `[IMG]` $v_1^1, \ldots, v_{I_1}^1$ `[SEP]` $\ldots$ `[SEP]` $v_1^k, \ldots, v_{I_k}^k$ `[TXT]` $w_1^1, \ldots, w_{J_1}^1$ `[SEP]` $\ldots$ `[SEP]`

	Method	Verbs	Nouns	All PoS
	Chance	0.1	0.1	0.1
	BERT (scratch) [10]	68.2	48.9	57.9
	BERT (pretrained) [10]	71.4	51.5	59.8
	BERT with Vision [7]	77.3	63.2	65.6
EXPERT	Isolated attn	81.2	74.2	66.8
EXPERT	Target-to-ref attn	80.9	69.6	69.8
EXPERT	Via-vision attn	81.9	73.0	74.9
EXPERT	+ Input pointing	80.1	73.2	67.0
EXPERT	Full attn	79.4	68.7	67.3

Table 5: **Acquiring Familiar Words on EPIC-Kitchens:** We report top-5 accuracy at predicting words seen in training for new visual instances.

$w_1^k, \ldots, w_{J_k}^k$ `[SEP]`. We denote example index with the superscript and $k = \text{rand}(0, k_-) + \text{rand}(0, k_+) + 1$ is the number of examples in the episode. I_i and J_i are the number of image and text tokens in the i$^{\text{th}}$ example of the episode.

We use the PyTorch framework [37] to implement our model. We base much of our code on a modified version of the Hugging Face PyTorch transformer repository. [3] In particular, we implement our own model class that extends `BertPreTrainedModel`. We write our own input encoding class that extends `BertEmbeddings` and adds visual embedding functionality as well as all the ϕ functions that are added to the word and image region embeddings. We use `hidden_size=384, intermediate_size=1536, num_attention_heads=4, num_hidden_layers=4, max_position_embeddings=512, type_vocab_size=64, vocab_size=32000` (and all other parameters default) to configure all models except the pretrained BERT one, which uses the original weights from `BERT-small`.

When we give an example to the model (whether individually or as part of an episode), we include all bounding boxes provided in the EPIC-Kitchens and Flickr30k datasets as well as the whole scene image, though our method can generalize to all formats of image input (*e.g.* multiple whole image frames, to provide temporal information and help action disambiguation). In practice, we filter out all bounding boxes of width or height < 10px since they do not provide useful information to the model, and resize both the entire image and bounding boxes to 112×112.

Does EXPERT require bounding boxes? We test EXPERT on EPIC-Kitchens without any additional image regions (i.e., only with the full image) to quantify the importance of providing grouped image regions. This model experiences a slight performance decrease in language acquisition, performing 4% worse with a 1:1 distractor ratio, and 2% worse with a 2:1 ratio. However, the baseline BERT with Vision model has larger decreases on all metrics when tested without additional image regions, so our model still outperforms it. When testing on masked language modeling (as in Sections 4.4 and 4.5 in the main

[3] https://huggingface.co/transformers/

	Method	Seen	New	Difference
	Chance	~0	~0	-
	BERT (scratch) [10]	34.3	17.7	16.6
	BERT (pretrained) [10]	39.8	20.7	19.1
	BERT with Vision [7]	56.1	37.6	18.5
EXPERT	Isolated attn	65.0	51.6	13.4
	Target-to-ref attn	61.7	45.7	16.0
	Via-vision attn	63.5	53.0	10.5
	+ Input pointing	62.7	48.3	15.4
	Full attn	59.2	44.4	14.8

Table 6: **Compositionality on EPIC-Kitchens:** We show top-5 accuracy at predicting masked compositions of seen nouns and verbs – both the verb and the noun must be correctly predicted.

Test	BERT with Vision	EXPERT
Acquiring Familiar Words		
Seen verbs	68.5	71.7
Seen nouns	44.6	59.3
Seen compositions	35.7	40.9
New compositions	30.0	34.4
Acquiring New Words		
Nouns 1:1	56.2	78.6
Verbs 1:1	60.5	72.1
Nouns 2:1	50.7	71.5
Verbs 2:1	44.1	64.5

Table 7: **Results on EPIC-Kitchens without Bounding Boxes:** We show accuracy on all evaluation metrics used in the paper, but withhold ground truth object bounding boxes at test time. EXPERT continues to outperform the competitive vision and language baseline.

paper), EXPERT performs 10% worse on verbs, 14% worse on nouns, 24% on seen compositions, and 29% on new compositions. However, the baseline model again has larger decreases in performance. These experiments show that while EXPERT improves when provided more visual information, it can acquire new language even when provided just with one image and perform stronger than baselines.

On Flickr30k, when testing with only the full image as model input in addition to text, performance decreases by *at most 7%*. Therefore, EXPERT does not require using image regions at test time. Performance decreases by a similar amount on vision and language baselines, such that EXPERT still outperforms them.

D Flickr Visualizations

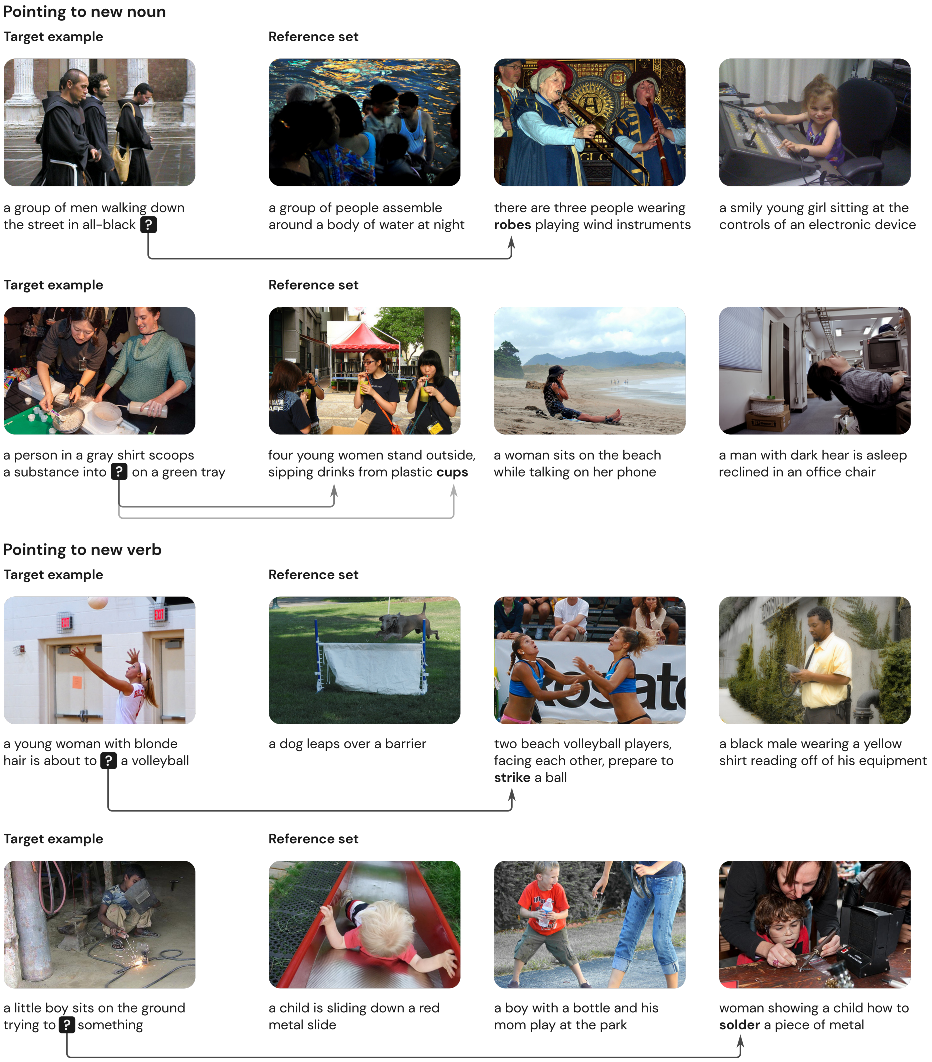


Fig. 11: **Acquiring New Words on Flickr30k:** We show examples where the model acquires new words. ? in the target example indicates the masked out new word. **Bold** words in the reference set are ground truth. The model makes predictions by pointing into the reference set, and the weight of each pointer is visualized by the shade of the arrows shown (weight $< 3\%$ is omitted).

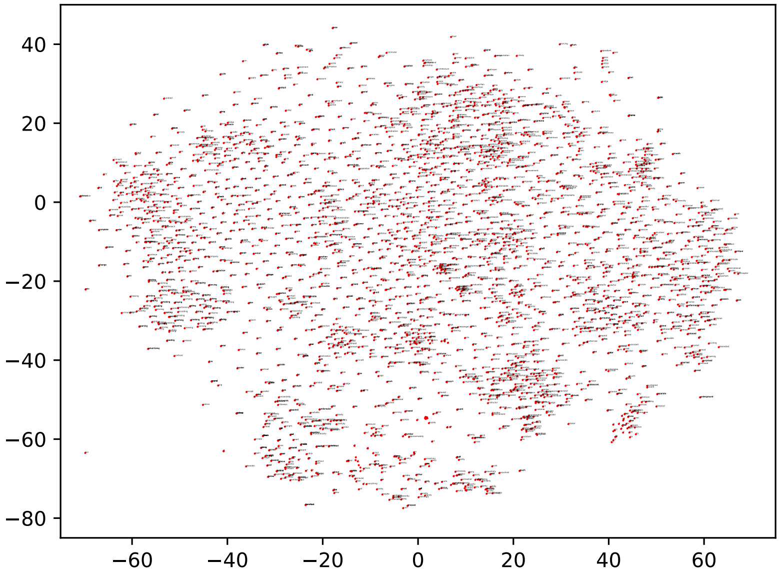


Fig. 12: **t-SNE 2D projection of Flickr embeddings:** We show a 2D projection computed using the t-SNE algorithm [55] of the word embedding matrix for EXPERT trained on the Flickr dataset. Each dot corresponds to the word shown to its top right. Please zoom in to view in detail.