

# Compositional Multimodal Understanding of Actions

## Alternative: Grounded Compositional Understanding of Actions

Anonymous EACL submission

### Abstract

Humans can rapidly understand new concepts, relying upon and combining context information with basic concepts from their existing knowledge. Compared to humans, neural network models trained over increasingly large datasets perform impressively well on a wide range of tasks, but they often fail to compositionally generalize to unseen concepts. In this study, we investigate compositionality and systematic generalization in a perceptually grounded setting by using a dataset of everyday household activities. This dataset depicts sequences of activities in pursuit of a wide variety of goals, *e.g. preparing celery, washing plates*. Each activity is represented with crowd-sourced utterances that describe different steps of the activity alongside with egocentric video frames and audio features. We evaluate several unimodal and multimodal baselines on future utterance prediction and action anticipation tasks, that respectively aim at describing and predicting an activity involving novel compositions of seen concepts. The models that exploit visual and audio signals do indeed improve over text-only model when they are evaluated on the long tail of rare complex concepts.

### 1 Introduction

As a long-standing problem in language, compositionality has been widely studied for many years. It deals with describing the relationship between an unbounded number of sentences and a vast set of meanings from a finite set of rules (Frege, 1963). Therefore, compositionality aims to address the problem of finding ways to define the meaning of an entire sentence as a function of the meaning of its constituents and the rules that are used to put those constituents together. In that regard, compositionality and systematic generalization have been used to characterize symbolic computation and human cognition (Fodor and Pylyshyn, 1988; Smolensky and Legendre, 2006). Humans demonstrate compositional ability in different domains

such as natural language understanding, and visual scene understanding. As put by (Lake, 2019), “Once a person learns the meaning of a new verb ‘dax’, he or she can immediately understand the meaning of ‘dax twice’ and ‘sing and dax’.” Similarly, it has been shown that humans can learn a new object shape and understand its compositions using previously learned colors or concepts (Johnson et al., 2017; Higgins et al., 2018).

Neural networks have been recently shown to perform well on many different tasks that require effective generalization abilities (LeCun et al., 2015). The compositionality and systematicity of neural networks have been long debated whether – and to what extent – neural networks display compositional generalization (Fodor and Pylyshyn, 1988; Christiansen et al., 1994; van der Velde et al., 2004; Brakel and Frank, 2009; Frank et al., 2014). Moreover, deep neural networks have been commonly criticized for requiring a very large number of training examples to succeed and argued to lack compositional abilities (Lake et al., 2017). Discussions around neural networks’ inability to capture the compositional structure of the underlying problem, thus failing to generalize compositionally have recently sparked a lot of interest in the machine learning community (Johnson et al., 2017; Lake and Baroni, 2018; Bastings et al., 2018; Loula et al., 2018; Hupkes et al., 2019; Russin et al., 2019; de Vries et al., 2019; Keysers et al., 2020).

Although recent studies towards understanding and improving the compositional generalization abilities of neural networks have garnered a lot of interest in the research community, little work has been done on the role of multimodal and grounded language processing. Lake (2019) investigated picking up new concepts and applying them in test time by coupling previously learned concepts with new concepts. Surís et al. (2020) investigated language modeling for acquiring new words and predicting new compositions by learning text rep-

representations from visual context for understanding instructional videos. Other existing works are centered around designing conceptual benchmark datasets specifically constructed for testing compositionality, *e.g.* (de Vries et al., 2019; Keysers et al., 2020; Vani et al., 2021). These studies have demonstrated that many state-of-the-art deep models fail to capture the compositional structures in the underlying tasks and cannot generalize well even on really simple textual data.

Our motivation in this study is to test linguistic compositionality and systematic generalization in a perceptually grounded setting and to understand whether leveraging visual and auditory cues can contribute to systematic generalization capabilities of deep models. Towards this goal, we turn our attention to multimodal how-to instructions as they provide a good test bed for our needs. More concretely, we make the following contributions:

- We curate a new benchmark: Epic-Kitchens-100-Systematicity (EK-100-SYS) for future utterance prediction and action anticipation tasks, which can be used to analyze compositional generalization in a grounded setup.
- We implement several neural models, and through them we analyze whether multimodality helps linguistic compositionality in the context of the proposed tasks.

## 2 Problem Formulation

### 2.1 Future Utterance Prediction Task

Predicting what comes next plays a central role in cognition (Bar, 2007; Clark, 2015) and also has been attributed as an interesting training scheme from cognitive perspective (Baroni, 2020). We study this task on a multimodal dataset that we curated containing videos of people performing everyday household activities, *e.g.* *preparing celery*. Each video in EK-100-SYS consists of a number of short clips (microsegments) that define sub-tasks of an activity, and each sub-task is described by a textual description, *e.g.* “pick up plate”, “put plate in sink”, “turn on water”, and “wash plate”, annotated by the actors after the recording. In the following, we formulate the future utterance prediction task as a language generation problem.

Let  $\mathcal{S} = (\mathbf{X}, \mathbf{V}, \mathbf{A})$  denote a triplet representing a short video clip with  $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^K$  being a sequence of  $K$  utterances, which describe a household activity and grounded with visual and audio signals, denoted by  $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^K$  and

$\mathbf{A} = \{\mathbf{a}_i\}_{i=1}^K$ , respectively. Our proposed future utterance prediction task involves generating the  $(K + 1)^{th}$  utterance,  $\mathbf{y} = \mathbf{x}_{K+1}$ , following the preceding  $K$  utterances and multimodal cues. The training data consists of a set of sequences of microsegments,  $\{(\mathcal{S}, \mathbf{y})\}$ .

During the training, our goal is to minimize the negative log-likelihood of the generated next utterance, where the multimodal models are conditioned on additional modalities such as image, or audio. Given the microsegment  $\mathcal{S}$  and the model parameters  $\theta$ , our objective is to minimize the negative log-likelihood of all next utterance tokens  $\mathbf{y} = \{y_i\}_{i=1}^m$ :

$$\log p(\mathbf{y}|\mathcal{S}; \theta) = - \sum_{i=1}^m \log p(y_i|\mathcal{S}; \theta) \quad (1)$$

### 2.2 Action Anticipation Task

We formulate action anticipation as a classification task, where we study the problem of predicting the next action with the target verbs and nouns. The main difference between this task and the future utterance prediction task is that utterance prediction is a natural language generation task. In particular, in the action anticipation task, the goal is to predict the next action by leveraging the previously observed actions. Differing from other action anticipation tasks, our setup allows us to formulate action anticipation in a compositional manner by predicting the verb and noun separately.

More formally, let  $\mathcal{S} = (\mathbf{X}, \mathbf{V}, \mathbf{A})$  denote a triplet representing a video clip with  $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^K$  representing a sequence of  $K$  utterances, which describe a household activity and grounded with visual and audio signals, denoted by  $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^K$  and  $\mathbf{A} = \{\mathbf{a}_i\}_{i=1}^K$ , respectively. Our action anticipation task involves predicting the verb/noun in the  $(K + 1)^{th}$  utterance,  $\mathbf{y} = \mathbf{x}_{K+1}^C$ , following the preceding  $K$  utterances and multimodal cues where  $C$  denotes the verb or noun class.

## 3 EK-100-SYS Dataset

We use the EPIC-Kitchens-100 dataset (EK-100) as the starting point for our experiments (Damen et al., 2020). EK-100 contains first-person videos of unscripted daily kitchen activities in natural household environments. Each video  $\mathbf{V}$  is split into a sequence of shorter clips  $\mathbf{v}_1, \dots, \mathbf{v}_k$ , which have manually-annotated English narrations of the activities within clips denoted by  $\mathbf{x}_1, \dots, \mathbf{x}_k$ . The clips





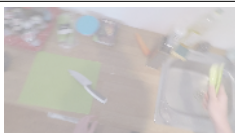
	Inputs (key frames and utterances)			Targets (future utterance)
Training				
	take celery	throw things into garbage bin	open fridge	put celery back into fridge
				
	pour sesame oil	close sesame oil	pick up onion	cut onion in half
				
	pick up containers	wash colander	put down colander	wash container
Evaluation				
	wash celery	close tap	put down celery	cut celery

Figure 1: Overview of our systematicity setup for future utterance prediction task using microsegments from our EK-100-SYS dataset. During training time the model has already been exposed to the primitives ‘wash’, ‘close’, ‘put down’, and ‘celery’ but not the “cut celery” composition and the goal is to be able to generalize to novel compositions of primitive elements in test time.

also have audio tracks  $\mathbf{a}_1, \dots, \mathbf{a}_k$ , which only contains sounds, *e.g.* a knife cutting onion or a person opening the fridge. We denote a sequence of video clips – audio tracks – narrations as an *instance*.

Recall from Section 2.1 that our aim is to model sequences of video clips. Therefore, we select instances from the EK-100 dataset with a window of  $K = 4$  clips: the first 3 clips are used for context, while the final clip is used for prediction. This results in 22,136 instances that can be used for our experiments. In Fig. 1, we provide some examples, along with representative keyframes and their corresponding narrations.

### 3.1 Systematicity Splits (EK-100-SYS)

Given a dataset of video sequences, our main focus is to study how well models compositionally generalize to unseen combinations of concepts. In the same vein of compositional captioning (Nikolaus et al., 2019) or novel compositions of object properties (Ruis et al., 2020), we create a dataset where the distribution of the individual concepts is similar across the dataset, but the compositions of those concepts is different. Consider the example given in Fig. 1. The model has already seen the nouns CELERY, TAP and verbs WASH, CLOSE, PUT

DOWN, CUT but it has not seen the combination of CUT CELERY during training. The model has to compositionally generalize to this new instance.

To obtain such splits of the dataset, we followed the *Maximum Compound Divergence* heuristic to create similar distributions of individual concepts (atoms) but different distributions of combinations of concepts (Keysers et al., 2020). We use the 97 verb classes and 300 noun classes in the EK-100 dataset as the atoms. In particular, we assign each sample to a split based on the atomic and compound divergence (similarity) based on weighted distributions using Chernoff coefficient  $C_\alpha(P\|Q) = \sum_k p_k^\alpha q_k^{1-\alpha} \in [0, 1]$  (Chung et al., 1989). To make atom distributions similar in train and test, we use  $\alpha = 0.5$  for atom divergence. Here, we set  $\alpha = 0.1$  to reflect that it is more important for a certain compound to be found in  $P$  (train) rather than the probabilities in  $P$  (train) and  $Q$  (test) match exactly. Following this logic, we define compound divergence, and atom divergence for a train set  $U$  and test set  $W$  as follows:

$$\mathcal{D}_C(U\|W) = 1 - C_{0.1}(\mathcal{F}_C(U) \parallel \mathcal{F}_C(W))$$

$$\mathcal{D}_A(U\|W) = 1 - C_{0.5}(\mathcal{F}_A(U) \parallel \mathcal{F}_A(W))$$

where  $\mathcal{F}_A(T)$  denotes frequency distribution of



atoms, and  $\mathcal{F}_C(T)$  denotes the distribution of compounds for a given set  $T$  and  $D_A$  and  $D_C$  denote atom and compound divergences, respectively. We calculated divergence scores for each data sample until the atomic divergence of train and test set  $D_A < 0.02$  and compound divergence of train and test set  $D_C < 0.6$ , which represents a sweet spot in terms of target distributions of atoms and compounds in the train and test sets (see Fig. 4 in the Appendices). Finally, we randomly divide this test set into two sets with similar distributions, one for validation and the other for testing.

The resulting EK-100-SYS dataset has 8,766 instances, which are split into 4,407 training, 2,184 validation, and 2,175 test instances. We use these splits to train and evaluate our models on the future utterance prediction and action anticipation tasks.

## 4 Models for Future Utterance Prediction

In this section, we provide details of the models tested for the future utterance prediction task. We benchmark a unimodal text-only model, along with several multimodal models to assess the importance of different modalities in systematic generalization.

### 4.1 Text-only Unimodal Baseline (L)

Our first baseline is a text-only model to assess potential biases in the dataset (Thomason et al., 2019). This model is a 1-layer attention-based encoder-decoder model based on Long-Short Term Memory networks (LSTM) (Hochreiter and Schmidhuber, 1997) with a hidden size of 256 units. The decoder learns a time-step dependent context over the encoder hidden states (Bahdanau et al., 2014). The model is trained using only the textual utterances  $\mathbf{x}_{1:K}$  from the microsegment as the input, and the next utterance  $\mathbf{x}_{K+1}$  as the target, *i.e.* to predict  $p(\mathbf{x}_{K+1}|\mathbf{x}_{1:K})$ . The model uses 200D word embeddings, over a vocabulary of unique words in training samples where vocabularies are not shared between the encoder and the decoder. We use the same textual encoder and decoder in all multimodal baselines (See Appendix A in the supplementary material for the number of trainable parameters and the vocabulary sizes).

### 4.2 Multimodal Baselines

We also evaluate several multimodal baselines that operate over combinations of the textual, visual, and audio modalities, as illustrated in Fig. 2.

#### 4.2.1 Vision and Language (VL)

Our Vision and Language baseline encodes both textual and visual context for the future utterance prediction task. In particular, the model encodes the textual utterances  $\mathbf{x}_{1:K}$  of each action from microsegments and the keyframe images  $\mathbf{v}_{1:K}$  to predict the next utterance  $\mathbf{x}_{K+1}$ , *i.e.*  $p(\mathbf{y} = \mathbf{x}_{K+1}|\mathbf{x}_{1:K}, \mathbf{v}_{1:K})$ . This model is adapted from a model that parses a visual scene and learns cross-modal self-attention (Tsai et al., 2019) over textual inputs and visual data.

The visual inputs are encoded using pre-trained CNN, and the textual inputs are encoded using an LSTM. More specifically, for the visual modality, we extracted two types of features: one type represents global visual features, and the other represents object-level features. For the global features, we used a pre-trained ResNet50 model (He et al., 2016) with ImageNet weights (Russakovsky et al., 2015). Object-level features were extracted using a pre-trained Faster-RCNN object detector (Ren et al., 2017) with a ResNet-101 backbone (He et al., 2016) which is pre-trained on MSCOCO (Lin et al., 2014) and finetuned on EK-100. We extract visual features from 5 objects for each keyframe. The resulting representation of a visual keyframe is the concatenation of the global and the object-level features. This concatenated vector is projected into a lower-dimensional space using a 1D convolution. The textual inputs are encoded using an LSTM with 200D word embeddings and a 256D hidden layer.

The visual and textual modalities are then encoded by a cross-modal self-attention mechanism, CM. In this model, we consider two modalities  $\alpha$  and  $\beta$ , sequences of each modalities are denoted as  $X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$  and  $X_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$ , respectively and  $T_{(\cdot)}$  denotes sequence length and  $d_{(\cdot)}$  denotes feature dimension. In this model,  $\alpha$  is the language modality, and  $\beta$  is the visual modality. In the cross-modal attention, the textual features are the *keys*, and the visual features are the *queries* and *values*, for aligning visual features to textual features. Let the Query be defined as  $Q_\alpha = X_\alpha W_{Q_\alpha}$ , the Keys as  $K_\beta = X_\beta W_{K_\beta}$ , and the Values as  $V_\beta = X_\beta W_{V_\beta}$ , where  $W_{Q_\alpha} \in \mathbb{R}^{d_\alpha \times d_k}$ ,  $W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$  and  $W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$  are learnable weights. The cross-modal self-attention from  $\beta$  to  $\alpha$  is formulated as a latent adaptation  $Y_\alpha \in \mathbb{R}^{T_\alpha \times d_v}$ :

$$\begin{aligned} Y_\alpha &= \text{CM}_{\beta \rightarrow \alpha}(X_\alpha, X_\beta) \\ &= \text{softmax} \left( \frac{Q_\alpha K_\beta^\top}{\sqrt{d_k}} \right) V_\beta \end{aligned} \quad (2)$$

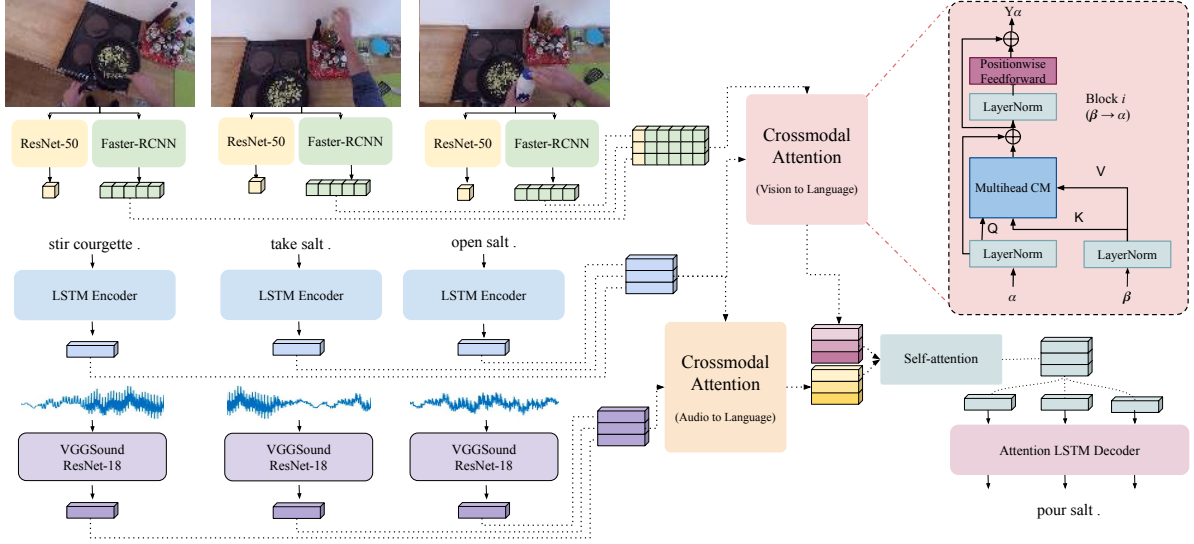


Figure 2: Overview of multimodal Audio, Vision and Language baseline model which incorporates global image features, object level image features, audio features as well as textual features using two crossmodal self-attention blocks with an LSTM decoder to predict next utterances.

The output  $Y_\alpha$  has the same length as  $Q_\alpha$ , but it is represented in the feature space of  $V_\beta$ . This enables the model to fuse different modalities, learning an alignment between the visual and textual features. Finally, there is a self-attention layer (Vaswani et al., 2017) over the aligned vision and language features, which are the input to an attention-based LSTM decoder that generates the next utterance.

#### 4.2.2 Audio and Language (AL)

The Audio and Language baseline has the same structure as the Vision and Language baseline. The key difference is that we represent the additional context using audio features instead of visual features. In particular, the model encodes both the textual utterances  $\mathbf{x}_{1:K}$  and the accompanying audio data  $\mathbf{a}_{1:K}$  to predict the next utterance  $\mathbf{x}_{K+1}$ , i.e.  $p(\mathbf{x}_{K+1}|\mathbf{x}_{1:K}, \mathbf{a}_{1:K})$ . The audio features are 512D vectors extracted using VGGSound (Chen et al., 2020), which is pre-trained on 200K videos from YouTube videos totaling up to 550 hours of audio data. Here, the model learns a cross-modal attention block over the audio and the textual features, analogously to using the visual and textual features, as inputs to an LSTM-based decoder.

#### 4.2.3 Object and Language (OL)

The Object and Language baseline uses the same architecture as Vision and Language baseline, but we represent the visual context using the tags of the detected objects, instead of the CNN visual features to more explicitly encode the visual content. In this

model, we embed object tags as a secondary set of textual features to our model along with the input utterances. Here, the object tags are represented as 292-dimensional one-hot encoded vectors (based on the number of unique tags) and projected to 256D with a simple linear layer. In this case, the cross-modal attention mechanism aligns object tag features with language features.

#### 4.2.4 Audio, Vision and Language (AVL)

In the Audio, Vision, and Language (AVL) baseline, we leverage the audio, visual, and textual data using two cross-modal self-attention blocks. We use textual utterances  $\mathbf{x}_{1:K}$  of each action along with the visual features  $\mathbf{v}_{1:K}$  from the keyframes, and the VGGSound audio features  $\mathbf{a}_{1:K}$  to guess the future utterance  $\mathbf{x}_{K+1}$ , i.e.  $p(y = \mathbf{x}_{K+1}|\mathbf{x}_{1:K}, \mathbf{v}_{1:K}, \mathbf{a}_{1:K})$ . In this model, the input to the self-attention layer before the decoder is the concatenation of the audio-aligned textual features from the audio-textual cross-modal block with the visual-aligned textual features from the visual-textual cross-modal block.

#### 4.2.5 Object, Audio and Language (OAL)

We perform an extra experiment to determine whether adding an extra modality to the OL baseline model improves its performance by coupling the object tags with audio features. Here, we include the extracted audio features from each microsegment to the OL model and train accordingly.

#### 4.2.6 Pretrained Audio-Vision and Language (PAVL)

In order to understand the importance of pretraining on a large scale aligned audio, visual and linguistic data, we also evaluate Merlot Reserve (Zellers et al., 2022). Merlot Reserve learns multimodal video representations over video frames, text, and audio. We extract multimodal audio-vision and language features through its pre-trained encoder while considering the same decoder network as the other baseline models. We report our experimental results with this pre-trained transformer baseline.

### 5 Models for Action Anticipation

Different than the future utterance prediction task which is formulated as a conditional language generation problem, action anticipation task requires (grounded) language understanding. We adapt the models described in the previous section for this task by slightly modifying their architectures. In particular, we replace the final layer in these models with two new fully connected layers, and finetune the pretrained models by considering a classification objective that involves predicting either the VERB or the NOUN in the anticipated action.

## 6 Experimental Setup

### 6.1 Evaluation Metrics

For evaluation, we used BLEU (Papineni et al., 2002), Exact Match (EM), and Categorical Accuracy (CA) metrics. For BLEU, we use NLTK toolkit and report unigram BLEU scores. For EM, we calculate an accuracy score between the generated text sequence and the groundtruth. CA uses the verb and noun categories in EK-100 and calculates the categorization accuracy based on noun category match between the predicted sequence and groundtruth, e.g. the verbs *slice*, *dice*, and *chop* fall into the same verb category *cut*, and the nouns *mozzarella*, *paneer* and *parmesan* are grouped into the same noun category *cheese*.

### 6.2 Training Details

We used SGD optimizer with 0.9 momentum with an initial learning rate of  $1e-1$  and a batch size of 128 and used ReduceOnPlateau learning rate scheduler to reduce the learning rate during training when validation loss metric plateaus. To train the models for future utterance prediction, we employed cross-entropy loss, initialized network weights via uniform distribution and used a dropout

rate of 0.3 for both the encoder and the decoder. We used an early stopping strategy and stopped the training if validation BLEU did not improve after a certain threshold (patience = 50). We clipped gradients and set the gradient threshold to 0.1, and used a 4-head multihead attention mechanism in the crossmodal self attention block in all our multimodal models. As a preprocessing step, we replace multiword tokens with a single word. For example, each occurrence of “olive oil” is replaced with “olive\_oil”. While training the models for action anticipation, we again used early stopping, but stopped the training if validation loss did not improve after a certain threshold (patience = 5).

## 7 Results

### 7.1 Future Utterance Prediction

Table 1 shows the results of the future utterance prediction experiments. As can be seen, all of the multimodal models outperform the language-only baseline. Models that use visual features (VL) or object tag features (OL) improve the performance approximately by 5 BLEU and 2 EM points compared to the language-only model. Using additional audio features (AL) brings the largest improvement in performance; BLEU and EM increase by 9 points and 4 points, respectively. Finally, the combination of audio, visual, and language features performs the best across all metrics. All these results clearly demonstrate the contribution of considering additional modalities to achieving better performances in the future utterance prediction task.

To further investigate the behaviour of the VL model, we examine the role of different layers in the crossmodal attention block in predicting which object regions correspond to the nouns in the target utterance. A model succeeds in this prediction task if the attention weight to the expected object region is maximized. In Table 2, we report these object level attention accuracies. We observe that the middle layers and final layer maximize attention to the expected object region. This indicates that to achieve better generalization, the VL model needs to use the semantic information encoded in the latter layers of the crossmodal attention block.

In Fig. 3, we provide a qualitative comparison of the baseline models. In the SYS setup, *put\_down knife* (third row), or *rinse pan* compounds have never been observed by the models during training time. In both of these examples, text-only unimodal model fails to generalize to novel compo-

Inputs	BLEU val	BLEU test	EM val	EM test	CA val	CA test
L	10.88 $\pm$ 0.7	10.55 $\pm$ 1.0	0.91 $\pm$ 0.4	0.61 $\pm$ 0.4	2.18 $\pm$ 0.6	2.16 $\pm$ 0.9
VL	19.49 $\pm$ 1.2	19.43 $\pm$ 0.7	4.60 $\pm$ 0.8	4.62 $\pm$ 0.4	8.62 $\pm$ 0.6	8.38 $\pm$ 0.3
AL	24.38 $\pm$ 0.9	25.21 $\pm$ 1.3	6.54 $\pm$ 0.4	<b>6.29</b> $\pm$ 0.5	12.46 $\pm$ 0.7	11.67 $\pm$ 1.1
OL	24.36 $\pm$ 0.8	<u>25.60</u> $\pm$ 1.4	6.13 $\pm$ 0.4	<u>6.17</u> $\pm$ 0.3	12.28 $\pm$ 0.7	<b>12.52</b> $\pm$ 0.4
AVL	19.53 $\pm$ 0.1	20.10 $\pm$ 0.1	4.70 $\pm$ 0.1	4.72 $\pm$ 0.2	8.77 $\pm$ 0.3	9.13 $\pm$ 0.5
OAL	24.97 $\pm$ 0.4	<b>25.55</b> $\pm$ 0.5	6.17 $\pm$ 0.4	5.72 $\pm$ 0.6	12.39 $\pm$ 0.7	<u>12.06</u> $\pm$ 1.0

Table 1: Performance of models on future utterance prediction. Using audio, visual, or object features always improves performance compared to the language-only unimodal baseline. We report the mean and the standard deviation across 3 runs.







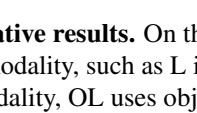
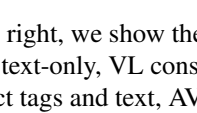
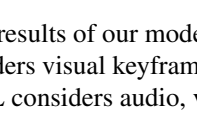
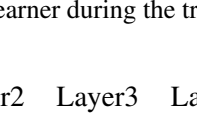
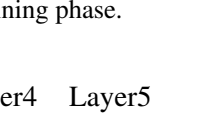
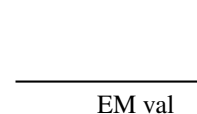
Inputs (utterances and auxiliary modalities)			Prediction (future utterance)	
Image				GT : place bowl
				L : put fridge
Text	clean bowl .	open dishwasher .	open drawer .	OL : close bin
				VL : wash bowl
Image				AL : close bowl
				AVL : place bowl
Text	turn_on tap .	rinse chopsticks .	take fork .	OAL : place bowl
				GT : rinse fork
Image				L : put_down bowl
				OL : rinse fork
Text	turn_on tap .	rinse chopsticks .	take fork .	VL : put fork
				AL : put fork
Image				AVL : place bowl
				OAL : rinse fork

Figure 3: **Qualitative results.** On the right, we show the results of our models. Each model considers a different combination of modality, such as L is text-only, VL considers visual keyframes and language, AL considers audio and language modality, OL uses object tags and text, AVL considers audio, visual and language modalities, OAL uses audio modality in addition to object tags and text. The targets, *rinse fork* and *place bowl* have never been observed by the learner during the training phase.

Layer1	Layer2	Layer3	Layer4	Layer5
8.6	9.8	10.4	8.8	9.8

Table 2: Object-level attention accuracy at different layers of the crossmodal attention block in the VL model.

sitions whereas AVL baseline predicts the target composition correctly for both examples.

## 7.2 Action Anticipation

We report the results of the baseline models in Table 3 on the action anticipation task, considering the SYS setup. We present the results for both train, val and test sets to demonstrate that SYS setup poses a much challenging case for the models. All model performances seem to degrade consistently when val and test sets are considered. The obtained performances are inline with those for the future utterance prediction task results. Even though there is not one particular model that outperforms the remaining models, multimodality seems to improve the overall performance of models’ compositional generalization ability on the action anticipation task.

	EM val	EM test	CA val	CA test
L	2.32 $\pm$ 0.6	1.99 $\pm$ 0.3	6.34 $\pm$ 0.7	6.02 $\pm$ 0.9
VL	2.16 $\pm$ 0.2	2.02 $\pm$ 0.0	7.45 $\pm$ 0.3	7.12 $\pm$ 0.6
AL	5.03 $\pm$ 0.5	4.45 $\pm$ 0.9	12.04 $\pm$ 1.4	11.36 $\pm$ 2.6
OL	4.74 $\pm$ 1.0	<b>5.31</b> $\pm$ 0.5	13.3 $\pm$ 1.6	<b>14.5</b> $\pm$ 1.2
AVL	3.23 $\pm$ 0.3	2.84 $\pm$ 0.6	9.95 $\pm$ 0.5	8.76 $\pm$ 0.9
OAL	6.09 $\pm$ 0.7	<u>5.14</u> $\pm$ 1.1	13.26 $\pm$ 0.4	<u>12.48</u> $\pm$ 1.3

Table 3: **Quantitative comparison of baselines for action anticipation task for predicting compound action (nouns and verbs).** We report mean across three runs. Best and second best performing results are bolded and underlined, respectively.

## 8 Related Work

**Compositionality.** Much recently, the compositionality problem has been investigated in various settings. Baroni (2020) analyzed the capacity of artificial neural networks in linguistic compositionality. Lake et al. (2019) examined systematicity and compositionality with a human-like number of examples. Nikolaus et al. (2019) investigated compositional generalization, in terms of a model’s performance to composing unseen combinations



of concepts when describing images. Dasgupta et al. (2018) explored compositionality in sentence embeddings for understanding how words combine, for generalizing to unencountered words and phrases. Ettinger et al. (2018) examined compositionality in sentence vector representations by probing for compositional information in embeddings using a number of existing sentence composition models. Andreas (2019) analyzed measuring compositionality in the aspect of representation learning *e.g.* a learned embedding. Bahdanau et al. (2019) investigated systematic generalization in a VQA-like setting. Lake (2019) examined compositionality in terms of systematic generalization for a meta-learning setting. Nye et al. (2019) considered learning entire rule systems from examples instead of learning to predict the correct output given a novel input. Hill et al. (2019) studied the emergence of systematic generalization in a situated agent setting where the agent learns to perform tasks based on textual instructions and visual observations. Ruis et al. (2020) proposed gSCAN dataset by extending SCAN dataset to a 2d grid world setting for situated language understanding using grounded instructions. Wu et al. (2020) suggested a transformer-based method for analogical reasoning in language acquisition setup coupled with visual data to pick up novel words.

Among the prior work, the closest work to ours is (Surís et al., 2020) in which the authors investigated compositionality and generalization in a novel word acquisition setting from narrated videos. They suggested to train a model using a masking strategy with a reference set where the model learns to map the masked word or words in the target tokens using the reference set examples in the same episode which is a sequence of image-text pairs. In our study, we focus on systematic generalization to novel compositions where models need to generalize to unseen compositions hence learns to learn primitive elements and concepts in the training set to generalize to novel compositions whereas in (Surís et al., 2020), the models learn to map target words from the reference set examples consisting of image-text pairs. In other words, here we introduce a setup where models need to generalize to novel compositions even the primitive elements need not be presented during inference, hence proposing a setup where models need to learn not only primitive elements but how they should learn

to form novel compositions (*e.g.* see Fig.1).

**Visually Grounded Reasoning.** Neural reasoning models have been shown to generalize well across biased dataset splits, towards addressing this problem, recently Johnson et al. (2017) proposed CLEVR-CoGenT dataset derived from CLEVR dataset to test models’ ability for compositional generalization on visual reasoning tasks. Anderson et al. (2018) investigated models’ ability to generalize to the previously unseen scene for visually-grounded natural language navigation in real buildings. Shridhar et al. (2020) studied models’ ability to translate grounded instructions to robot actions to accomplish household tasks in unseen environments. Related to our work Seo et al. (2020) predicted future utterances from multimodal data, using instructional videos and their transcribed speech as text where the goal is to rank groundtruth utterance among candidates whereas we focus on predicting future utterances by generating utterances autoregressively to assess compositional generalization abilities of models trained on aligned textual and visual data.

## 9 Conclusion

In this paper, we presented an investigation of linguistic compositionality and systematic generalization in a perceptually grounded setting. We showed how a multimodal how-to instructions dataset can be utilized as a challenging test bed to assess compositional generalization. For this purpose, we designed the future utterance prediction and action anticipation tasks and followed a methodical approach in generating the training, validation and test sets in our systematicity split. We experimented with several baseline models and investigated models’ ability to generalize to novel compositions and showed how multimodal data can contribute towards solving systematic generalization problem. We hope that our work will stimulate further research along these directions. That being said, we must admit that the textual utterances that we consider in our work are too simplistic and clearly does not capture true complexities of language. Hence, extending this work to a more natural source of language data will be quite interesting.



## Limitations

Mandatory Limitations section. (Appears after the conclusion. Is not counted towards the page limit.)

## Ethics Statement

We curate our EK-100-SYS dataset using the video clips from the published EPIC-Kitchens-100 dataset (Damen et al., 2020), which is publicly available.

Scientific work published at EACL 2023 must comply with the ACL Ethics Policy. We encourage all authors to include an explicit ethics statement on the broader impact of the work, or other ethical considerations after the conclusion but before the references. The ethics statement will not count toward the page limit (8 pages for long, 4 pages for short papers).

## 10 To do List

- Appendix will be revisited. There are a few things inconsistencies in the text. Also referral to Appendix in the main text should be checked.
- Remove RND from anywhere.
- make sure figures and tables are referred in the text and appendix.
- Mandatory Limitations section. (Appears after the conclusion. Is not counted towards the page limit.)
- Address the remaining comments.
- Ethics statement is encouraged and not counted towards page limit.

## References

- Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE CVPR*, pages 3674–3683.
- Jacob Andreas. 2019. [Measuring compositionality in representation learning](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries,

and Aaron C. Courville. 2019. Systematic generalization: What is required and can it be learned? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Moshe Bar. 2007. The proactive brain: using analogies and associations to generate predictions. *Trends in cognitive sciences*, 11(7):280–289.

Marco Baroni. 2020. Linguistic generalization and compositionality in modern artificial neural networks. *Phil. Trans. R. Soc. B*, 375(1791):20190307.

Jasmijn Bastings, Marco Baroni, Jason Weston, Kyunghyun Cho, and Douwe Kiela. 2018. [Jump to better conclusions: SCAN both left and right](#). In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, Black-boxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 47–55. Association for Computational Linguistics.

Philémon Brakel and Stefan Frank. 2009. Strong systematicity in sentence processing by simple recurrent networks. In *31th Annual Conference of the Cognitive Science Society (COGSCI-2009)*, pages 1599–1604. Cognitive Science Society.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE.

M Christiansen, Nick Chater, et al. 1994. Generalization and connectionist language learning. *Mind and Language*, 9(3).

JK Chung, PL Kannappan, CT Ng, and PK Sahoo. 1989. Measures of distance between probability distributions. *Journal of mathematical analysis and applications*, 138(1):280–292.

Andy Clark. 2015. *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2020. [Rescaling egocentric vision](#). *CoRR*, abs/2006.13256.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. 2018. [Evaluating compositionality in sentence embeddings](#). *CoRR*, abs/1802.04302.

Harm de Vries, Dzmitry Bahdanau, Shikhar Murty, Aaron C. Courville, and Philippe Beaudoin. 2019. [CLOSURE: assessing systematic generalization of CLEVR models](#). In *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop, Vancouver, Canada, December 13, 2019*.

712	Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. <a href="#">Assessing composition in sentence vector representations</a> . In <i>Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018</i> , pages 1790–1801. Association for Computational Linguistics.	767
713		768
714		769
715		
716		770
717		771
718		772
719	Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. <i>Cognition</i> , 28(1-2):3–71.	773
720		774
721		775
722	Stefan L Frank, P Calvo, and J Symons. 2014. Getting real about systematicity. <i>The architecture of cognition: Rethinking Fodor and Pylyshyn’s systematicity challenge</i> , pages 147–164.	776
723		777
724		778
725		779
726	Gottlob Frege. 1963. Compound thoughts. <i>Mind</i> , 72(285):1–17.	780
727		781
728	Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In <i>Proceedings of the IEEE conference on CVPR</i> , pages 770–778.	782
729		783
730		
731		
732	Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P. Burgess, Matko Bosnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. 2018. <a href="#">SCAN: learning hierarchical compositional visual concepts</a> . In <i>6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings</i> . OpenReview.net.	784
733		785
734		786
735		787
736		788
737		789
738		790
739		
740		
741	Felix Hill, Andrew K. Lampinen, Rosalia Schneider, Stephen Clark, Matthew Botvinick, James L. McClelland, and Adam Santoro. 2019. <a href="#">Emergent systematic generalization in a situated agent</a> . <i>CoRR</i> , abs/1910.00571.	791
742		792
743		793
744		794
745		
746	Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. <i>Neural computation</i> , 9(8):1735–1780.	795
747		796
748		
749	Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2019. <a href="#">The compositionality of neural networks: integrating symbolism and connectionism</a> . <i>CoRR</i> , abs/1908.08351.	797
750		798
751		799
752		800
753	Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. <a href="#">CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning</a> . In <i>2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017</i> , pages 1988–1997. IEEE Computer Society.	801
754		802
755		803
756		804
757		
758		
759		
760		
761	Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. <a href="#">Measuring compositional generalization: A comprehensive</a>	805
762		806
763		807
764		808
765		809
766		810
		811
		812
		813
		814
		815
		816
		817
		818
		819
		820
		821
		822

- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. [Faster R-CNN: Towards real-time object detection with region proposal networks](#). *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149.
- Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M. Lake. 2020. [A benchmark for systematic generalization in grounded language understanding](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.
- Jake Russin, Jason Jo, Randall C. O’Reilly, and Yoshua Bengio. 2019. [Compositional generalization in a deep seq2seq model by separating syntax and semantics](#). *CoRR*, abs/1904.09708.
- Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. 2020. [Look before you speak: Visually contextualized utterances](#). *CoRR*, abs/2012.05710.
- Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. [ALFRED: A benchmark for interpreting grounded instructions for everyday tasks](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10737–10746. IEEE.
- Paul Smolensky and Géraldine Legendre. 2006. *The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture)*, Vol. 1. MIT press.
- Dídac Surís, Dave Epstein, Heng Ji, Shih-Fu Chang, and Carl Vondrick. 2020. [Learning to learn words from visual scenes](#). In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, pages 434–452. Springer.
- Jesse Thomason, Daniel Gordon, and Yonatan Bisk. 2019. [Shifting the baseline: Single modality performance on visual navigation & QA](#). 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)*, pages 1977–1983, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.
- Frank van der Velde, Gwendid T van der Voort van der Kleij, and Marc de Kamps. 2004. Lack of combinatorial productivity in language processing with simple recurrent networks. *Connection Science*, 16(1):21–46.
- Ankit Vani, Max Schwarzer, Yuchen Lu, Eeshan Dhekane, and Aaron Courville. 2021. [Iterated learning for emergent systematicity in VQA](#). In *International Conference on Learning Representations*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Bo Wu, Haoyu Qin, Alireza Zareian, Carl Vondrick, and Shih-Fu Chang. 2020. [Analogical reasoning for visually grounded language acquisition](#). *CoRR*, abs/2007.11668.
- Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. *arXiv preprint arXiv:2201.02639*.



## A Experimental Setup

**Systematicity Split (EK-100-SYS).** Fig. 4 illustrates the atomic and the compound distributions over the constructed training, validation and test splits of our proposed systematicity setup. As can be seen, while these splits have similar distributions over atoms, training and val/test splits do differ in terms of compounds.

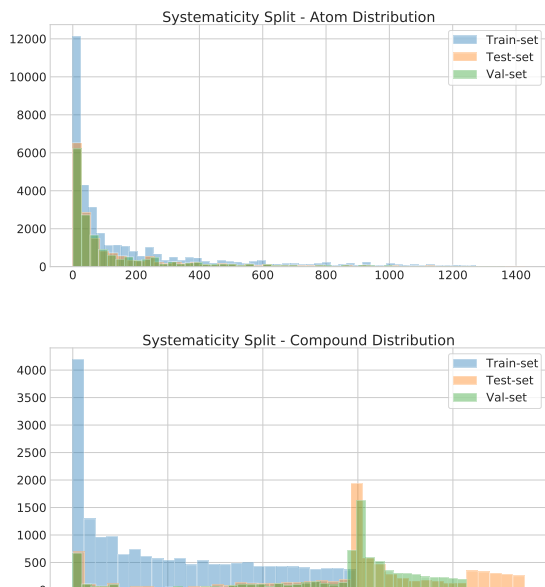


Figure 4: For train/val/test splits for systematicity split setup, plot at the top demonstrates the distribution of atoms while plot at the bottom shows the distribution of compounds.

**Choosing Keyframes from Videos.** In our experimental setup, we choose and use representative images from each microsegment. While choosing such a representative image for each video sequence, we follow a simple heuristics-based strategy. In particular, we run an object detector on the video frames and select the frames containing the maximum number of object proposals captured by the object detector as the representative frames.

**Model Sizes and Training Time.** In Table 4, we present the number of trainable parameters and training time for all of our baseline models for the future utterance prediction task. All of our models are implemented with PyTorch and trained with Nvidia 1080Ti GPUs.

Model	Parameters	Training Time
L	2,827,218	10 mins
OL	6,833,874	35 mins
VL	7,301,074	18 mins
AL	6,907,858	31 mins
AVL	15,331,026	37 mins
OAL	14,863,826	39 mins

Table 4: Model sizes and their training times along with the vocabulary sizes considered in our experiments.

## B Further Analysis

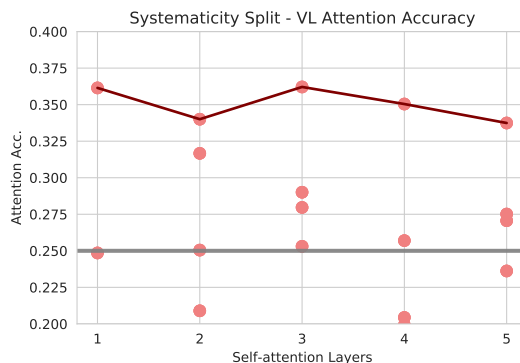


Figure 5: Attention accuracies over layers and heads. Systematicity split performance for three multiple runs and 4 heads.

**How does object-level attention change over RND and SYS splits?** In Fig. 6, on two sample microsegments from RND and SYS splits, we illustrate how attention weights assigned to object regions change for the proposed VL model. Here we visualize the weights specifically at the layers which is found to be the most successful in our analysis. As conjectured in our main paper, in the RND setup, the model prefers to depend more on the global video frame than the object regions and thus it gives lower attention scores to the extracted object regions. On the other hand, in the SYS setup, the model favors the auxiliary object information more and assigns higher attention scores to these object regions.

**What visual and auditory features encode?** We carry out an experimental analysis to investigate what kind information raw visual and auditory features encode in our next utterance prediction task using a t-SNE visualization. In this analysis, we consider the most commonly observed compounds



### How models generalize for isolated VERB and Noun prediction in action anticipation task?

	EM val	EM test	CA val	CA test
L	19.05 $\pm$ 0.1	18.97 $\pm$ 0.2	38.55 $\pm$ 0.4	<b>40.21</b> $\pm$ 0.7
VL	21.15 $\pm$ 0.3	<b>19.69</b> $\pm$ 0.5	38.64 $\pm$ 1.0	37.67 $\pm$ 1.2
AL	19.45 $\pm$ 0.6	19.44 $\pm$ 0.9	37.02 $\pm$ 1.4	37.02 $\pm$ 2.2
AVL	20.52 $\pm$ 0.4	19.58 $\pm$ 0.5	38.75 $\pm$ 1.4	38.7 $\pm$ 1.1
OL	18.94 $\pm$ 0.6	18.79 $\pm$ 0.2	38.51 $\pm$ 1.1	<u>38.98</u> $\pm$ 1.1
OAL	18.82 $\pm$ 1.2	<u>19.59</u> $\pm$ 0.7	36.09 $\pm$ 1.1	37.17 $\pm$ 1.8

Table 5: **Quantitative comparison of baselines for action anticipation task for predicting target verbs.** We report mean across three runs. Best and second best performing results are bolded and underlined, respectively.

	EM val	EM test	CA val	CA test
L	10.15 $\pm$ 2.3	10.04 $\pm$ 2.4	15.42 $\pm$ 2	15.86 $\pm$ 2.2
VL	11.87 $\pm$ 0.5	12.2 $\pm$ 1.1	19.94 $\pm$ 0.2	21.17 $\pm$ 0.8
AL	23.13 $\pm$ 0.6	23.48 $\pm$ 0.8	30.65 $\pm$ 1	31.52 $\pm$ 1.6
AVL	16.34 $\pm$ 0.1	16.16 $\pm$ 0.6	24.98 $\pm$ 0.3	25.34 $\pm$ 0.4
OL	27.08 $\pm$ 0.8	<u>28.12</u> $\pm$ 0.6	35.49 $\pm$ 0.5	<u>36.49</u> $\pm$ 0.1
OAL	28.81 $\pm$ 0.4	<b>31.28</b> $\pm$ 1.1	35.34 $\pm$ 0.7	<b>38.03</b> $\pm$ 1

Table 6: **Quantitative comparison of baselines in the action anticipation task for predicting target nouns.** We report mean across three runs. Best and second best performing results are bolded and underlined, respectively.