The Role of Syntactic Planning in Compositional Image Captioning

Emanuele Bugliarello and **Desmond Elliott**
Department of Computer Science
University of Copenhagen
{emanuele,de}@di.ku.dk

Abstract

Image captioning has focused on generalizing to images drawn from the same distribution as the training set, and not to the more challenging problem of generalizing to different distributions of images. Recently, Nikolaus et al. (2019) introduced a dataset to assess compositional generalization in image captioning, where models are evaluated on their ability to describe images with unseen adjective–noun and noun–verb compositions. In this work, we investigate different methods to improve compositional generalization by planning the syntactic structure of a caption. Our experiments show that jointly modeling tokens and syntactic tags enhances generalization in both RNN- and Transformer-based models, while also improving performance on standard metrics.

1 Introduction

Image captioning is a core task in multimodal NLP, where the aim is to automatically describe the content of an image in natural language. To succeed in this task, a model first needs to recognize and understand the properties of the image. Then, it needs to generate well-formed sentences, requiring both a syntactic and a semantic knowledge of the language (Hossain et al., 2019). Deep learning techniques are the standard approach to tackling this problem: images are represented by visual features extracted from Convolutional Neural Networks (*e.g.* He et al. 2016), and sentences are generated by conditioning Recurrent Neural Networks (*e.g.* Hochreiter and Schmidhuber 1997), or Transformers (Vaswani et al., 2017) on the extracted visual features.

While deep neural networks achieve impressive performance in a variety of applications, including image captioning, their ability to demonstrate compositionality, defined as the algebraic potential to understand and produce novel combinations from known components (Loula et al., 2018), has been questioned. Semantic compositionality of language in neural networks has attracted interest in the community (Irsoy and Cardie, 2014; Lake and Baroni, 2018; Baroni, 2019) as compositionality is conjectured to be a core feature not only of language but also of human thought (Fodor and Lepore, 2002).

In image captioning, improving compositional generalization is a fundamental step towards generalizable systems that can be employed in daily life. To this end, Nikolaus et al. (2019) recently introduced a compositional generalization dataset where models need to describe images that depict unseen compositions of primitive concepts. For example, models are trained to describe images with "white" entities and all types of "dog" concepts but never the adjective–noun composition of "white dog." In their dataset, models are evaluated on their ability to caption images depicting the unseen composition of held out concepts. Their study suggests that RNN-based captioning models do not compositionally generalize, and that this is primarily attributable to the language generation component.

In this paper, we study the potential for syntax to improve compositional generalization in image captioning by combining syntactic planning and language generation in a single model. Our study is inspired by the traditional Natural Language Generation (NLG) framework (Reiter and Dale, 1997), where NLG is split into three distinct steps: text planning, sentence planning, and linguistic realization. While state-of-the-art captioning models typically proceed directly from visual features to sentence generation, we hypothesize that a model that plans the structure of a sentence as an intermediate step will improve compositional generalization. A model with a planning step can learn the high-level structure of sentences, making it less prone to overfitting the training data.

Specifically, we explore three methods for integrating syntactic planning into captioning in our

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 593–607
April 19 - 23, 2021.

experiments: (a) pre-generation of syntactic tags from the image, (b) interleaved generation of syntactic tags and words (Nădejde et al., 2017), and (c) multi-task learning with a shared encoder that predicts syntactic tags or words (Currey and Heafield, 2019). We do so while also empirically investigating four different levels of syntactic granularity.

The main findings of our experiments are that:

- jointly modeling syntactic tags and tokens leads to improvements in Transformer-based (Cornia et al., 2020) and RNN-based (Anderson et al., 2018) image captioning models;
- although the effectiveness of each syntactic tag set varies across our explored approaches, the widely-used *chunking* tag set never outperforms syntactic tags with finer granularity;
- compositional generalization is affected by directly mapping from image representation to tokens because performance can be improved by interleaving a dummy tag with no meaning;
- interleaving syntactic tags with tokens leads to a loss in performance for retrieval systems.

Finally, we also propose an attention-driven image–sentence ranking model, which makes it possible to adaptively combine syntax within the re-scoring approach of Nikolaus et al. (2019) to further improve compositional generalization in image captioning.

2 Planning Image Captions

Natural language generation has traditionally been framed in terms of six basic sub-tasks: content determination, discourse planning, sentence aggregation, lexicalization, referring expression generation and linguistic realization (Reiter and Dale, 1997). Within this framework, a three-stage pipeline has emerged (Reiter, 1994):

- **Text Planning:** combining content determination and discourse planning.
- **Sentence Planning:** combining sentence aggregation, lexicalization and referring expression generation to determine the structure of the selected input to be included in the output.
- **Linguistic Realization:** this stage involves syntactic, morphological and orthographic processing to produce the final sentence.

Early methods for image captioning drew inspiration from this framework; for example, the MIDGE system (Mitchell et al., 2012) features explicit steps for content determination, given detected objects, and sentence aggregation based on local and full phrase-structure tree construction, and TREETALK composes tree fragments using integer linear programming (Kuznetsova et al., 2014). More recently, Wang et al. (2017) propose a two-stage algorithm where the skeleton sentence of the caption (main objects and their relationships) is first generated, and then the attributes for each object are generated if they are worth mentioning. In contrast, the majority of neural network models are based on the encoder-decoder framework (Sutskever et al., 2014) of learning a direct mapping from different granularities of visual representations (Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018) to language model decoders based on RNNs (Vinyals et al., 2015) or Transformers (Guo et al., 2020; Cornia et al., 2020).

2.1 Motivation

In this paper, we explore whether image captioning models can be improved by explicitly modeling sentence planning as an intermediate step between content determination and linguistic realization. In particular, we study the use of syntactic tags in enriching the sentence planning step to improve compositional generalization. In the compositional image captioning task, models are tasked with describing images that depict unseen combinations of adjective–noun and noun–verb constructions (see Nikolaus et al. 2019 for a more detailed description of this task). Nikolaus et al. (2019) presented a model that improves generalization with a jointly trained discriminative re-ranker, whereas here, we investigate the role of sentence planning via syntax.

From a psycholinguistic perspective (Griffin and Bock, 2000; Coco and Keller, 2012), there is evidence that humans make plans about how to describe the visual world: they first decide what to talk about (analogous to content determination), then they decide what they will say (a sentence planning phase), and finally, they produce an utterance (linguistic realization). We hypothesize that, analogously to humans, neural network decoders will also find it useful to make such sentence plans.

From a machine learning perspective, the use of syntactic structure can mitigate the bias introduced by the maximum likelihood training of neural network image captioning models. Recall that in the context of image captioning, the optimization objective consists of maximizing the likelihood:

$$\mathcal{L} = \prod_{t=1}^{T} \mathbb{P}(y_t | y_{<t}, v), \tag{1}$$

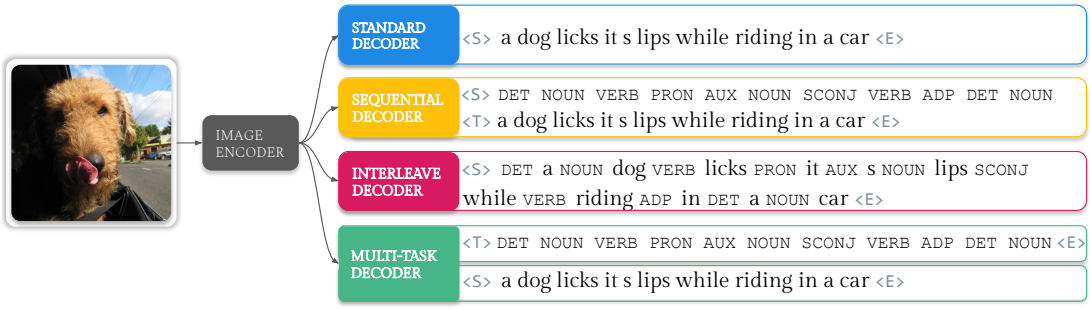

Figure 1: Approaches to syntactically plan image captioning (with POS tags). STANDARD captioning systems directly generate a sequence of surface forms (i.e. words). SEQUENTIAL generates a sequence of syntactic tags, followed by a sequence of surface forms. INTERLEAVE alternates syntactic tags and surface forms. MULTI-TASK generates either a sequence of syntactic tags or a sequence of surface forms from a shared decoder.

where v denotes the visual features (either a single vector or a set of vectors extracted from an image).

In a standard generation task, a model learns to predict the next token based on what it has observed so far. This is especially limiting when it is evaluated on unseen combinations of adjective–noun and noun–verb constructions in the compositional generalization task (i.e. data points that fall outside the training distribution). In fact, models are not explicitly asked to learn *word classes* nor how to connect them to form novel combinations. Whereas, if a system also models syntax, it can assign higher probability to "white dog" if it expects to generate a sequence with an adjective followed by a noun.

2.2 Planning Approaches

We investigate three approaches to jointly modeling tokens and syntactic tags: syntax-driven sequential caption planning (SEQUENTIAL), syntax-interleaved caption generation (INTERLEAVE), and syntax and caption generation via multi-task learning (MULTI-TASK). See Figure 1 for an overview.

SEQUENTIAL: Our first approach closely follows the traditional NLG pipeline and it is related to the *text planning* stage defined above, although limited to sentence-level rather than to a full discourse. Here, a model plans, through syntactic tags, the order of the information to be presented. Specifically, the model is required to generate a sequence whose first T outputs represent the underlying syntactic structure of the sentence before subsequently generating the corresponding T surface forms.

INTERLEAVE: Our second approach consists of interleaving syntactic tags and tokens during generation, which means a syntactic tag and its realization are next to each other, removing the pressure for a model to successfully track long-range dependencies between tags and tokens. Moreover, this allows for a more flexible planning, where the model can adapt the sentence structure based on the previously generated tags and tokens. In particular, a model can break bi-gram dependencies and learn narrower distributions over the next word based on the current syntactic tag. For instance, if we consider part-of-speech tags, the model learns that only a subset of the vocabulary corresponds to nouns, and another subset to adjectives, and so on.

MULTI-TASK: Our last approach is based on multi-task learning, where a model produces either a sequence of tokens (main task) or syntactic tags (secondary task). We draw on the simple and effective approach of Currey and Heafield (2019), proposed for neural machine translation (NMT). In the NMT framework, the source sentence was prepended a task-specific tag, which led the decoder to either predict the translation of the source sentence or the syntactic tags of the source sentence. We adapt this to image captioning by setting the first token to either a start-of-syntax token (<T>) or start-of-sentence token (<S>) and then generating tags or tokens, respectively. Compared to the other approaches, MULTI-TASK allows the model to learn both types of forms at the same position. While this approach does not double sequence length, it doubles the number of sequences per training epoch.

2.3 Syntactic Granularity

In addition to the three approaches of realizing sentence planning, we investigate the effects of different syntactic tags from a coarse to fine granularity. We experiment with the following tags:

- **CHUNK:** Also known as *shallow parsing*, chunks are syntactic tags that model phrasal

structure in a sentence, such as noun phrases (NP) and verb phrases (VP).

- **POS:** Part-of-speech tags are specific lexical categories to which words are assigned, based on their syntactic context and role, such as nouns (N) and adjectives (ADJ).
- **DEP:** Dependency-based grammars model the structure as well as the semantic dependencies and relationships between words in a sentence. In this study, we consider the dependency labels assigned to each word, such as adjectival modifiers (amod), which denote any adjective that modifies the meaning of a noun.
- **CCG:** Combinatory categorial grammar (Steedman and Baldridge, 2006) is based on combinatory logic and provides a transparent interface between surface syntax and the underlying semantic representation. For example, the syntactic category assigned to "sees" is "(S\NP)/NP", denoting it as a transitive verb that will be followed by a noun phrase.

We also study the merit of breaking bi-gram dependencies for the INTERLEAVE approach by tagging each word with a synthetic tag <IDLE>. We hypothesize this approach would not give any benefits in any metric, as attention-based models can simply learn to ignore these pseudo-tags.

3 Experimental Setup

Data We use training and evaluation sets such that paradigmatic gaps exist in the training set. That is, for a concept pair $\{c_i, c_j\}$, the validation $\mathcal{D}_{val}$ and test $\mathcal{D}_{test}$ sets only contain images in which at least one of the captions contains the pair of concepts, while the complementary set – where concepts c_i and c_j can only be observed independently – is used for training $\mathcal{D}_{train}$. Following Nikolaus et al. (2019), we select the same 24 adjective–noun and verb–noun concept pairs, and split the English COCO dataset (Lin et al., 2014) into four sets, each containing six held out concept pairs.

Pre-processing We first lower-case and strip away punctuation from the captions. We then use StanfordNLP (Qi et al., 2018) to tokenize and lemmatize the captions, and to extract universal POS tags and syntactic dependency relations. For IOB-based chunking, we train a classifier-based tagger on *CoNLL2000* data (Tjong Kim Sang and Buchholz, 2000) using NLTK (Bird et al., 2009). Finally, we use the A* CCG parsing model by Yoshikawa et al. (2017) with ELMo embeddings (Peters et al., 2018) to extract CCG tags. Visual features are extracted from 36 regions of interest in each image using Bottom-Up attention (Anderson et al., 2018) trained on Visual Genome (Krishna et al., 2017).

Evaluation Following Nikolaus et al. (2019), we evaluate compositional generalization with Recall@K. Given K generated captions for each of the M images in an evaluation set, $\{\langle s_1^1, \ldots, s_K^1\rangle, \ldots, \langle s_1^M, \ldots, s_K^M\rangle\}$, the recall of the concept pairs is given by:

$$\text{Recall@K} = \frac{|\{\langle s_k^m\rangle | \exists k : s_k^m \in \mathcal{C}\}|}{M}, \quad (2)$$

where s_k^m denotes the k-th generated caption for image m and $\mathcal{C}$ is the set of captions which contain the expected concept pair and in which the adjective or the verb is a dependent of the noun.

In addition, we use `pycocoeval` to score models on the common image captioning metrics: METEOR (M; Denkowski and Lavie 2014), SPICE (S; Anderson et al. 2016), CIDER (C; Vedantam et al. 2015), and BLEU (B; Papineni et al. 2002); and the recent multi-reference BERTSCORE (BS; Yi et al. 2020). In particular, we report the average recall across all concept pairs, the average across the four splits for each score in `pycocoeval`, and the average across all captions for BERTSCORE.

Models We evaluate three models:

- **BUTD**: Bottom-Up and Top-Down attention (Anderson et al., 2018), a strong and widely-employed RNN-based captioning system.
- **BUTR**: Bottom-Up and Top-Down attention with Ranking (Nikolaus et al., 2019), an RNN-based, multi-task model trained for image captioning and image–sentence ranking that achieves state-of-the-art performance in the compositional generalization task.[1]
- **$\mathcal{M}^2$-TRM**: Meshed-Memory Transformer (Cornia et al., 2020), a recently proposed Transformer-based architecture that achieves state-of-the-art performance in image captioning on the COCO dataset.

Implementation details We follow Nikolaus et al. (2019) and Cornia et al. (2020) to train their systems. Model selection is performed using early stopping, which is determined when the BLEU score of the generated captions in the validation set

[1] We denote as BUTR the model that uses the re-ranking module (BUTR+RR in Nikolaus et al. 2019) as it was shown to be essential to improve compositional generalization.

	Model	R@5	M	S	C	B	BS
	BUTD	9.5	25.2	18.6	92.7	32.3	41.7
SEQUENTIAL	+IDLE	8.7	23.7	17.8	87.6	30.0	38.8
	+CHUNK	10.9	24.7	18.2	89.2	31.2	41.2
	+POS	9.5	24.1	17.5	86.1	30.1	40.7
	+DEP	11.1	24.6	17.8	89.7	30.8	41.0
	+CCG	10.6	24.5	18.0	88.4	30.4	41.0
INTERLEAVE	+IDLE	10.5	25.3	18.8	94.3	32.3	41.7
	+CHUNK	9.7	25.2	18.7	93.4	32.5	41.7
	+POS	**11.8**	25.4	18.8	94.4	**32.7**	41.7
	+DEP	10.8	25.2	18.7	93.0	31.9	41.6
	+CCG	10.5	25.4	**19.0**	94.6	**32.7**	41.9
MULTI-TASK	+IDLE	9.8	25.5	18.7	94.5	**32.7**	41.8
	+CHUNK	10.3	25.5	**19.0**	94.5	32.4	41.9
	+POS	10.3	25.4	18.8	93.8	32.6	41.8
	+DEP	11.4	25.5	18.9	93.9	**32.7**	41.9
	+CCG	10.8	**25.7**	**19.0**	**95.6**	**32.7**	**42.0**

Table 1: Average validation results for our approaches to integrating syntactic planning into image captioning evaluated across four types of syntactic forms.

does not increase for five consecutive epochs.[2] We use the default hyperparameters and do not fine-tune them when tasking the models with syntax generation. For full experimental details, refer to App. A. Our code and data are publicly available.[3]

4 Syntax Awareness

In Table 1, we first report the performance of BUTD when jointly modeling different types of syntactic tags and each approach to sentence planning.

Syntax helps compositional image captioning Table 1 clearly shows that, regardless of the level of granularity, syntactic planning enhances compositional generalization in image captioning (R@5). Moreover, CHUNK – one of the most widely-used tag sets for syntax-aware image captioning (*e.g.* Kuznetsova et al. 2012; Yang and Liu 2020) – is outperformed by tag sets with finer granularity (e.g. DEP) in *every* approach, motivating further research into incorporating them in image captioning.

Looking at the results for the SEQUENTIAL approach, we see that, with the exception of POS tags, syntactic planning increases the ability of the model to recall novel concept pairs, with gains of at least +1.1 R@5 points. We then hypothesize that syntax-based sequential planning is effective if the tags convey information about words in relation to each other, e.g. CCG tags as opposed to POS tags.

When the model INTERLEAVEs syntactic tags and words, there is an improvement of at least +1.0 R@5, except for CHUNK. Moreover, POS tags lead to the highest gain of +2.3 R@5.

Finally, the MULTI-TASK approach also leads to significant gains in compositional generalization, with DEP (original setup of Currey and Heafield 2019) giving the highest R@5, corroborating the effectiveness of our porting into image captioning.

Generalization across categories We further investigate the role of syntactic planning for the different unseen composition categories defined by Nikolaus et al. (2019). Figure 2 illustrates how our different combinations of approaches and syntactic tags deal with color and size, type of the objects (animate and inanimate) and type of the verbs (transitive and intransitive). We see that DEP tags consistently improve upon BUTD for color and size concept pairs, regardless of the planning approach, making them a robust tag set for future research. INTERLEAVE+POS also leads to gains for all color and size categories, with up to +10 R@5 for colors of inanimate objects. Conversely, all the variants perform worse than the baseline for the sizes of animate objects. However, this drop is not substantial because BUTD already performs poorly.

Towards neural NLG pipelines While the SEQUENTIAL approach closely follows the traditional NLG pipeline, it consistently degrades performance in standard metrics for image captioning. On the other hand, both INTERLEAVE and MULTI-TASK lead to higher performance in compositional generalization and other metrics. In particular, when BUTD is trained to predict either words or CCG tags in the MULTI-TASK approach, the generated captions achieve the highest average scores, including a substantial gain of +2.9 CIDEr points. These results indicate that neural models require novel ways of sentence planning; and that effectively doing so *consistently* leads to the same or better performance in every considered metric.

Grounding the need for planning Overall, Table 1 provides empirical support that an explicit planning step improves compositional generalization in image captioning. In fact, even breaking bi-grams with the <IDLE> tag in the INTERLEAVE approach improves performance: the standard approach of directly mapping image representations to tokens is sub-optimal because the model learns to generate *n*-grams seen during training.

[2] Whenever present, syntactic tags are stripped away when computing evaluation metrics such as BLEU scores.

[3] https://github.com/e-bug/syncap.

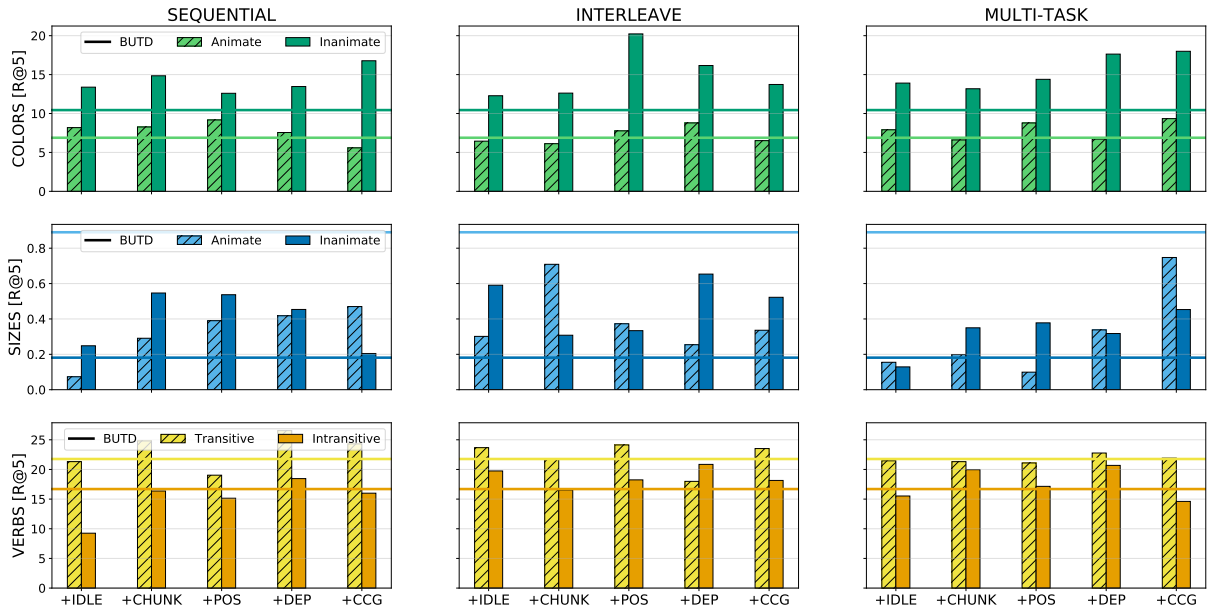

Figure 2: R@5 of unseen compositions by planning approaches (columns) and composition categories (rows).

	Captioning Evaluation						Text Retrieval			Image Retrieval		
Model	**R@5**	**M**	**S**	**C**	**B**	**BS**	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
BUTR	15.0	26.2	19.9	88.6	28.9	41.8	21.3	45.4	57.8	14.7	35.1	47.0
+POS	12.0	25.7	19.4	85.4	27.4	41.4	17.4	39.3	50.6	12.4	31.0	42.0
BUTR$_{mean}$ +POS	14.2	25.9	19.7	87.4	28.3	42.9	23.3	47.7	60.0	17.1	38.6	50.2
BUTR$_{weight}$	14.9	**26.4**	**20.2**	88.8	28.5	**43.2**	**26.0**	**51.7**	**63.6**	**18.6**	**40.9**	**52.8**
+POS	**16.4**	**26.4**	20.0	**89.8**	**29.1**	43.1	24.5	48.6	60.6	18.0	40.2	52.0

Table 2: Average validation results when interleaving syntactic and lexical forms in BUTR and our variants.

Given its superior performance in the recall of novel compositions of concepts, we adopt INTERLEAVE+POS throughout the remainder of this paper to jointly model syntactic tags and words. For clarity of exposition, we refer to this approach as POS.

4.1 Adaptive Re-Ranking for Syntax

Recall that the best-performing model for compositional image captioning re-ranks its generated captions given the image (BUTR; Nikolaus et al. 2019). Here, we study how to combine the benefits of syntactic planning and their re-ranking approach.

The BUTD model, investigated above, is a two-layer LSTM (Hochreiter and Schmidhuber, 1997) in which the first LSTM encodes the sequence of words, and the second LSTM integrates visual features through an attention mechanism to generate the output sequence (Anderson et al., 2018). The state-of-the-art BUTR model extends this with an image–sentence ranking network that projects images and captions into a joint visual-semantic embedding. The sentence representation used by the ranking network is a learned projection of the final hidden state of the first LSTM: $\mathbf{s} = \mathbf{W}\mathbf{h}_T^l$.

Ranking performance Table 2 shows that the image–sentence retrieval performance of BUTR decreases when interleaving POS tags with words. Given the previous formulation of BUTR and its connection to BUTD, we conclude that jointly modeling syntactic tags and words leads to decreased performance in the generation and ranking tasks.

Adaptively attending to tags We explore two approaches to combining the improvements of interleaved syntactic tagging with ranking:

- *mean*: The model creates a mean representation over the hidden states of the first LSTM.
- *weight*: The model forms a weighted pooling of the hidden states of the first LSTM layer, whose weights are learned through a linear layer. This is a simple form of attention mechanism and it is equivalent to the one used

Model	R@5	M	S	C	B	BS
BUTD	9.5	25.2	18.6	92.7	32.3	41.7
+POS	11.8	25.4	18.8	94.4	32.7	41.7
BUTR$_{weight}$	14.9	26.4	20.2	88.8	28.5	43.2
+POS	**16.4**	26.4	20.0	89.8	29.1	43.1
$\mathcal{M}^2$-TRM	10.6	27.9	21.6	**114.0**	**37.2**	44.4
+POS	13.2	**28.0**	**21.7**	113.8	35.4	**44.9**

Model	R@5	M	S	C	B	BS
BUTD	9.2	25.4	18.6	94.4	32.4	41.8
+POS	11.1	25.4	18.7	96.3	32.9	41.8
BUTR$_{weight}$	13.5	26.4	20.1	91.0	28.6	43.3
+POS	**15.4**	26.3	20.0	91.0	28.7	43.2
$\mathcal{M}^2$-TRM	10.1	27.8	21.5	**115.7**	**36.5**	44.5
+POS	12.1	**28.0**	**21.6**	**115.7**	35.0	**44.9**

Table 3: Average validation (left) and test (right) results when interleaving syntactic (POS) and lexical forms.

by Nikolaus et al. (2019) to represent image features in the shared embedding space:

$$\begin{aligned} \omega_t &= \mathbf{W}_\alpha \mathbf{h}_t^l, \\ \boldsymbol{\alpha} &= \text{softmax}\left(\omega\right), \\ \mathbf{s} &= \mathbf{W} \sum_{t=1}^{T} \alpha_t \mathbf{h}_t^l. \end{aligned} \tag{3}$$

Table 2 shows that the weighting mechanism in the ranking model effectively disentangles syntactic tags and tokens, resulting in +1.5 RECALL points over BUTR, with small improvements to the other metrics. Compared to BUTR, BUTR$_{weight}$ also improves the retrieval performance of the ranking module. Adding POS tags still decreases retrieval performance but, compared to BUTR+POS, the difference is now halved for text retrieval and only 0.7 points for image retrieval. Overall, our BUTR$_{weight}$ is a more general and robust approach to jointly training a captioning system and a discriminative image–sentence ranker.

5 Results and Discussion

We now report the final performance of three image captioning models that integrate the syntactic planning (INTERLEAVE+POS) with word generation.

Model-agnostic improvements We start by investigating whether the compositionality given by syntactic planning generalizes across architectures. Table 3 reports average validation and test scores for the BUTD, BUTR$_{weight}$ and $\mathcal{M}^2$-TRM models. We find that interleaving POS tags and tokens consistently leads to +2 RECALL points in each model without affecting the performance on other metrics, with the exception of decreased BLEU score of $\mathcal{M}^2$-TRM. In this case, $\mathcal{M}^2$-TRM +POS generates captions that are abnormally truncated, ending with bi-grams such as "of a," "on a" and "to a".[4] Furthermore, we can clearly see that despite

[4]This is known as reward hacking (Amodei et al., 2016) which arises in models with a reinforcement learning-based optimization phase. Investigating whether proposed approaches to mitigate this problem (Liu et al., 2017; Li et al., 2019, *inter alia*) are also effective in our setup is left as future work.

Model	Color		Size		Verb	
	A	I	A	I	T	I
BUTD	6.9	10.4	**0.9**	0.2	21.8	16.7
+POS	7.8	20.2	0.4	0.3	24.1	18.2
BUTR$_{weight}$	15.0	24.9	0.8	1.6	26.2	20.8
+POS	**16.2**	**30.4**	**0.9**	**2.5**	26.8	**21.9**
$\mathcal{M}^2$-TRM	5.2	11.9	0.2	0.2	29.1	16.6
+POS	7.9	17.9	0.1	0.2	**32.3**	20.6

Table 4: Average validation R@5 scores for different categories of held out pairs. Color and size adjectives are split into **A**nimate or **I**nanimate objects; Verbs are split into **T**ransitive and **I**ntransitive verbs.

$\mathcal{M}^2$-TRM outperforming the RNN-based models in every standard metric, it is only +1 RECALL point better than BUTD at compositional generalization. Hence, syntactic planning is an effective strategy to compositional generalization, *regardless* of the language model used.

Generalization across categories Table 4 lists the R@5 scores for different categories of held out pairs. Differently from the results reported by Nikolaus et al. (2019), we find the performance of BUTD for noun–verb concept pairs to be much higher thanks to a larger beam size (equal to the one used for BUTR in our experiments). Moreover, the performance from interleaving POS across different categories of held out pairs shows that improvements are consistent across categories and models, with the exception of size modifiers of animate objects, where all models perform poorly. This was also found by Nikolaus et al. (2019) and it is likely due to the need for real-world knowledge (i.e. does this image depict a "big dog" compared to *all other* "dogs"?). For a full breakdown of the R@5 generalization performance for each held out pair by each model, see Table 9 in App. B.

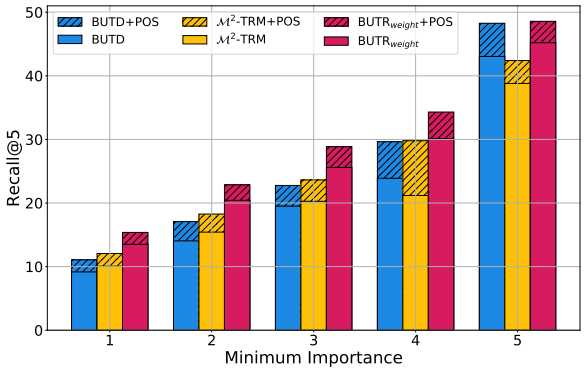

Figure 3: Average test R@5 as a function of the minimum importance of a concept pair for each model.

Model	ASL	Types	TTR$_1$	TTR$_2$	%Novel	Cov	Loc$_5$
BUTD	8.6	463	0.16	0.37	69.2	0.12	0.74
+POS	8.6	466	0.16	0.37	70.3	0.12	0.75
BUTR$_{weight}$	**10.3**	**783**	**0.20**	**0.49**	**97.2**	**0.20**	0.78
+POS	**10.3**	778	**0.20**	**0.49**	96.7	**0.20**	0.78
$\mathcal{M}^2$-TRM	9.1	580	0.14	0.33	80.1	0.15	**0.83**
+POS	9.5	601	0.13	0.33	88.4	0.15	**0.83**

Table 5: Average validation set scores for diversity metrics as defined in van Miltenburg et al. (2018).

Performance by minimum importance Given that annotators of the COCO dataset were given a relatively open task to describe images, captioning systems should exhibit higher recall of concept pairs when more annotators use them in the descriptions. As shown in Figure 3, this behavior is seen in each model, with increasing gains given by jointly modeling lexical and syntactic forms. In particular, we observe that the $\mathcal{M}^2$-TRM model recalls fewer pairs than BUTD when they are considered more relevant (more annotators use them in describing an image), and that interleaving POS tags partially solves its limitations. Moreover, as agreement among annotators increases, we also see that BUTD+POS is as effective as BUTR$_{weight}$, corroborating the effectiveness of our model-agnostic approach against a more complex, multi-task model.

Captions diversity Table 5 reports the average scores for caption diversity (van Miltenburg et al., 2018) in the validation data. Comparing BUTD (RNN-based) and $\mathcal{M}^2$-TRM (Transformer-based) models, we see that the output vocabulary of the $\mathcal{M}^2$-TRM-based model spans many more word types, resulting in $+11\%$ novel captions. However, $\mathcal{M}^2$-TRM has lower mean segmented type-token ratios (TTRs), contrasting the conclusion of van Miltenburg et al. (2018) that the number of novel descriptions is strongly correlated with the TTR (while this correlation is maintained with the number word types). The models that jointly model syntactic tags and tokens lead to a higher number of types in both models and a substantial $+8\%$ in novel captions for $\mathcal{M}^2$-TRM, without affecting other metrics. Clearly, BUTR$_{weight}$ leads to longer sentences, more types, higher TTRs and the highest percentage of novel captions. We can also see that BUTR$_{weight}$ achieves the highest coverage (defined as the percentage of learnable words it can recall), while $\mathcal{M}^2$-TRM has the highest local recall score, being able to better recall the content words that are important to describe a given image.

Accuracy of syntactic forms We verify that the models can correctly predict syntactic tags, regardless of their granularity and the approach used to jointly modeling them with tokens. Indeed, the accuracy of the generated syntactic tags, measured as the ratio of *sequences* matching the annotations from StanfordNLP, by BUTD is high, ranging between 95% and 99%. See App. B for details.

Qualitative examples Figure 4 shows generated captions. Compared to standard BUTD, all syntax-aware approaches allow the model to recall more unseen concept pairs, while also improving the overall quality of the captions. In addition, when looking at the captions generated by all three models, both with and without interleaving POS tags, we find the integration of syntactic tags to clearly **improve the quality** of the generated caption. See Figure 5 in App. B for more examples.

6 Related Work

Compositional image captioning Nikolaus et al. (2019) studies compositional generalization in image captioning with combinations of unseen adjective–noun and verb–noun pairs, whose constituents are observed at training time but not their combination, thus introducing a *paradigmatic gap* in the training data. Nikolaus et al. (2019) showed how to improve compositional generalization by jointly training an image–sentence ranking model with a captioning model. Other work has also investigated generalization to unseen combinations of visual concepts as a classification task (Misra et al., 2017; Kato et al., 2018), triplet prediction (Atzmon et al., 2016), or unseen objects (Lu et al., 2018). Here, we improve generalization by jointly modeling syntactic tags and tokens, and we show

BUTD: there is a woman that is on the floor
BUTD + SEQUENTIAL: a woman doing a trick on a bicycle
BUTD + INTERLEAVE: a woman riding a bike on a wooden floor
BUTD + MULTI-TASK: a woman riding a bike on a wooden surface

BUTD: a woman with a child sitting on a bench
BUTD + POS: **a girl that is standing** on a skateboard

BUTR$_{weight}$: a girl and child playing with a toy in a backyard
BUTR$_{weight}$ + POS: **a girl doing a trick on a skateboard** on a brick walkway

$\mathcal{M}^2$-TRM: a girl doing a trick on a skateboard with
$\mathcal{M}^2$-TRM + POS: a little girl **standing on a skateboard** in the

Figure 4: **Top:** Captions generated by BUTD with different approaches to integrating syntactic tags (based on best R@5). **Bottom:** Captions generated by BUTD, BUTR$_{weight}$, and $\mathcal{M}^2$-TRM, and each model when interleaving POS tags. Syntax-aware approaches generate **higher quality** captions with more unseen concept pairs.

how to combine this with the improvements gained from a jointly-trained ranking model.

Joint syntactic and semantic representations While little work has investigated the interaction of jointly modeling semantics and various syntactic forms in captioning models, a few studies have exploited syntax in image and video captioning. Zhao et al. (2018) propose a multi-task system to jointly train the task of image captioning with two additional tasks: multi-object classification and syntax generation. The same LSTM decoder is used to generate captions *and* CCG tags by mapping the hidden representations to either word or tag vocabularies through two different output layers. Dai et al. (2018) propose a two-stage sequential pipeline where a sequence of noun-phrases is first selected from a fixed pool, which are then patched together via predetermined connecting phrases. This method, however, is unlikely to realize any benefits for compositional generalization because it uses the top-50 noun-phrases and $1,000$ connecting phrases from the training set. Our INTERLEAVE approach can be used to address these limitations in their "phrase pool" and "connecting" modules to produce unseen compositions. Deshpande et al. (2019) rely on sequences of POS tags to produce diverse captions. Similarly to our SEQUENTIAL approach, their model first predicts a sequence of POS tags conditioned on the input image. However, the authors limit the POS sequences to $1,024$ templates obtained through quantization of the training set. During inference, the model samples k POS tag sequences and uses them to condition a greedy decoder for captions generation. Hou et al. (2019) take yet another approach to jointly learn POS tags and surface forms in the framework of video captioning. They introduce a model that resembles our INTERLEAVE approach but with two main differences: (i) the t-th tag is not conditioned on previous tags, and (ii) the t-th word is only conditioned on the t-th tag and the video.

7 Conclusion

We investigated a variety of approaches along with the use of syntactic tag sets to achieve compositional generalization in image captioning via sentence planning. Our results support the claim that combining syntactic planning and language generation *consistently* improves the generalization capability of RNN- and Transformer-based image captioning models, especially for inanimate color–noun combinations. While this approach penalizes image–sentence ranking models, we showed that this can be overcome with an adaptive mechanism, resulting in state-of-the-art performance on the compositional generalization task. We believe our results will lead to further exploration of syntax-aware captioning models given their potential to better generalize, both in terms of under-researched syntactic granularity (e.g. CCG) and more expressive alternatives to modeling syntactic structure. Another direction for future work is to focus on size–noun compositions, which rely on the successful integration of real-world knowledge.

Acknowledgments

We are grateful to the anonymous reviewers, Mitja Nikolaus, Mert Kiliçkaya and members of the CoAStaL NLP group for their useful comments and discussions. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety. *ArXiv*, abs/1606.06565.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086.

Yuval Atzmon, Jonathan Berant, Vahid Kezami, Amir Globerson, and Gal Chechik. 2016. Learning to generalize to new compositions in image understanding. *ArXiv*, abs/1608.07639.

Marco Baroni. 2019. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791).

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 794–803, Stockholmsmässan, Stockholm Sweden. PMLR.

Moreno I Coco and Frank Keller. 2012. Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive science*, 36(7):1204–1223.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10575–10584.

Anna Currey and Kenneth Heafield. 2019. Incorporating source syntax into transformer-based neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 24–33, Florence, Italy. Association for Computational Linguistics.

Bo Dai, Sanja Fidler, and Dahua Lin. 2018. A neural compositional paradigm for image captioning. In *Advances in Neural Information Processing Systems*, page 656–666. Curran Associates, Inc.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G Schwing, and David Forsyth. 2019. Fast, diverse and accurate image captioning guided by part-of-speech. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10687–10696.

Jerry A. Fodor and Ernie Lepore. 2002. *The compositionality papers*. Oxford University Press.

Zenzi M Griffin and Kathryn Bock. 2000. What the eyes say about speaking. *Psychological science*, 11(4):274–279.

Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. 2020. Normalized and geometry-aware self-attention network for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10324–10333.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*, 51(6).

Jingyi Hou, Xinxiao Wu, Wentian Zhao, Jiebo Luo, and Yunde Jia. 2019. Joint syntax representation learning and visual cue translation for video captioning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8917–8926.

Ozan Irsoy and Claire Cardie. 2014. Deep recursive neural networks for compositionality in language. In *Advances in Neural Information Processing Systems*, volume 27, pages 2096–2104. Curran Associates, Inc.

Keizo Kato, Yin Li, and Abhinav Gupta. 2018. Compositional learning for human object interaction. In *European Conference on Computer Vision*, pages 247–264. Springer.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *ArXiv*, abs/1412.6980.

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

Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 359–368, Jeju Island, Korea. Association for Computational Linguistics.

Polina Kuznetsova, Vicente Ordonez, Tamara L. Berg, and Yejin Choi. 2014. Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics*, 2:351–362.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2873–2882, Stockholmsmässan, Stockholm Sweden. PMLR.

Nannan Li, Zhenzhong Chen, and Shan Liu. 2019. Meta learning for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8626–8633.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 873–881.

João Loula, Marco Baroni, and Brenden Lake. 2018. Rearranging the familiar: Testing compositional generalization in recurrent networks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 108–114, Brussels, Belgium. Association for Computational Linguistics.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7219–7228.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ishan Misra, Abhinav Gupta, and Martial Hebert. 2017. From red wine to red tomato: Composition with context. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1160–1169.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756, Avignon, France. Association for Computational Linguistics.

Maria Nădejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. 2017. Predicting target language CCG supertags improves neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 68–79, Copenhagen, Denmark. Association for Computational Linguistics.

Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. 2019. Compositional generalization in image captioning. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 87–98, Hong Kong, China. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Ehud Reiter. 1994. Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Workshop on Natural Language Generation*.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

M. Steedman and J. Baldridge. 2006. Combinatory categorial grammar. In Keith Brown, editor, *Encyclopedia of Language & Linguistics (Second Edition)*, second edition edition, pages 610 – 621. Elsevier, Oxford.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 3104–3112. Curran Associates, Inc.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W Cottrell. 2017. Skeleton key: Image captioning by skeleton-attribute decomposition. In *2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7378–7387.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2048–2057, Lille, France. PMLR.

Zhenyu Yang and Qiao Liu. 2020. Att-bm-som: A framework of effectively choosing image information and optimizing syntax for image captioning. *IEEE Access*, 8:50565–50573.

Yanzhi Yi, Hangyu Deng, and Jinglu Hu. 2020. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 985–994, Online. Association for Computational Linguistics.

Masashi Yoshikawa, Hiroshi Noji, and Yuji Matsumoto. 2017. A* CCG parsing with a supertag and dependency factored model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 277–287, Vancouver, Canada. Association for Computational Linguistics.

Wei Zhao, Benyou Wang, Jianbo Ye, Min Yang, Zhou Zhao, Ruotian Luo, and Yu Qiao. 2018. A multi-task learning approach for image captioning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1205–1211. International Joint Conferences on Artificial Intelligence Organization.

A Experimental Setup

Data In order to evaluate the compositional generalization of a model, we use training and evaluation sets such that paradigmatic gaps are observed in the training set. That is, for a concept pair $\{c_i, c_j\}$, the validation $\mathcal{D}_{val}$ and test $\mathcal{D}_{test}$ sets only contain images in which at least one of the captions contains the pair of concepts, while the complementary set – where concepts c_i and c_j can only be observed independently – is used for training $\mathcal{D}_{train}$. Specifically, following Nikolaus et al. (2019), we select the same 24 adjective–noun and verb–noun concept pairs, and split the English COCO dataset (Lin et al., 2014) into four sets, each containing six held out concept pairs (training and validation instances are drawn from `train2014`, while test instances from `val2014`). Table 6 lists the sizes (in number of images) of each split.[5] For more details, we refer the reader to Nikolaus et al. (2019).

Training details Following Nikolaus et al. (2019) and Cornia et al. (2020), each system is trained with teacher forcing. Model selection is performed using early stopping, which is determined when the BLEU score of the generated captions in the validation set does not increase for five consecutive epochs.[6] All models are trained using the Adam optimizer (Kingma and Ba, 2014): BUTD and BUTR use an initial learning rate of $1e-4$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and gradients are clipped when they exceed 10.0. For the GradNorm optimizer (Chen et al., 2018) used in BUTR, the initial learning rate is 0.01 and the asymmetry is 2.5, although we find it beneficial to tune the latter when generating syntax.[7] Moreover, we find that taking the absolute value of the GradNorm weights for each loss in the renormalization step (given that our loss functions are by definition positive) leads to more stable multi-task training. $\mathcal{M}^2$-TRM first uses an initial learning rate of 1, $\beta_1 = 0.9$ and $\beta_2 = 0.98$, with a warm-up equal to $10,000$ iterations (Vaswani et al., 2017), and it is then fixed to $5e-6$ during CIDEr-D optimization. A batch size of 50 is used when training $\mathcal{M}^2$-TRM, while BUTD and BUTR are trained with batch sizes of 100. When adding syntactic forms, due to memory limitations, a batch size of 50 is always used (see Table 7 for a comparison of the planning approaches). All models are trained on one NVIDIA TitanX GPU in a shared cluster. See Table 8 for running times on the second held out dataset.

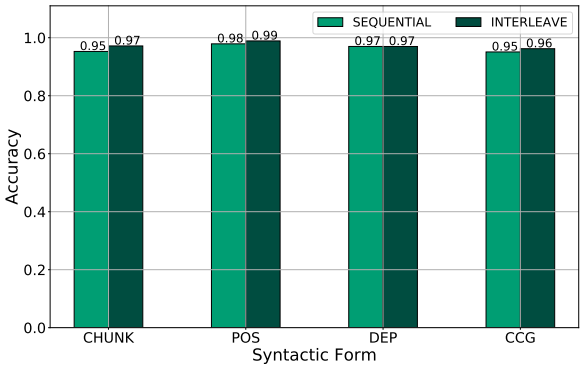

Figure 6: Validation accuracy of syntactic tags generated by BUTD using SEQUENTIAL and INTERLEAVE.

Inference At evaluation time, a maximum caption length of 20 is used when generating lexical forms only, and of 40 when also syntactic tags are generated. Notably, we use the default hyperparameters provided by the respective authors and do not fine-tune them when tasking the models with syntax generation. Differently from Nikolaus et al. (2019), rather than using a beam of 100 for BUTR only, we let all systems generate captions using such beam size as we found it to significantly improve compositional generalization of BUTD in our validation sets.[8]

B Further Analysis

Accuracy of syntactic forms We verify that a model can correctly predict syntactic forms, regardless of their granularity and the approach used to jointly modeling them with lexical forms. Figure 6 shows that, indeed, the accuracy of the generated syntactic tags, measured as the ratio of *sequences* matching the annotations from StanfordNLP, by BUTD is high, ranging between 95% and 99%. Note that we only evaluate accuracy of the SEQUENTIAL and INTERLEAVE approaches as there is no close relationship between syntactic and lexical sequences in the MULTI-TASK approach.

Qualitative examples Figure 5 shows more generated captions for images in the validation sets.

[5] Note that the size of each set is slightly different from the one in Nikolaus et al. (2019) as they used different tools for tokenizing and parsing, while we use a single framework to maximize its performance when parsing the captions (used to identify concept pair candidates).

[6] When present, syntactic forms are stripped away when computing evaluation metrics such as BLEU scores.

[7] Searched over the minimal grid: $\alpha \in \{1.5, 2.0, 2.5, 4.0\}$.

[8] BUTD R@5: 7.8 (beam size 5), 9.5 (beam size 100).

	Held out pairs	$\mathcal{D}_{train}$	$\mathcal{D}_{val}$	$\mathcal{D}_{test}$
1	black cat, big bird, red bus, small plane, eat man, lie woman	79,847	2,936	1,349
2	brown dog, small cat, white truck, big plane, ride woman, fly bird	79,823	2,960	1,355
3	white horse, big cat, blue bus, small table, hold child, stand bird	79,922	2,861	1,459
4	black bird, small dog, white boat, big truck, eat horse, stand child	79,627	3,156	1,501

Table 6: Held out word pairs in each dataset split. Dataset sizes are measured in number of images.

Approach	T	d(S_t, W_t)	N
SEQUENTIAL	2×	T	1×
INTERLEAVE	2×	1	1×
MULTI-TASK	1×	–	2×

Table 7: Comparison of the planning approaches studied in this paper in terms of the sequence lengths used to train the captioning models (**T**), the distance between the syntactic tag and its corresponding word d(S_t, W_t), and the number of training examples per epoch (**N**).

Model	Training time [hour]
BUTD	6
+POS	9
BUTR$_{weight}$	18
+POS	17
$\mathcal{M}^2$-TRM	383
+POS	414

Table 8: Training times on held out dataset 2.

Concept pair	BUTD	+POS	BUTR$_{weight}$	+POS	$\mathcal{M}^2$-TRM	+POS
black cat	16.1	15.3	26.0	29.1	5.0	10.9
eat man	19.0	24.5	26.5	26.5	23.8	29.1
small plane	0.2	0.0	0.0	0.4	0.6	0.0
red bus	16.1	24.0	51.8	54.2	13.3	24.9
big bird	2.4	0.9	0.9	0.9	0.5	0.0
lie woman	17.5	18.5	27.1	20.8	14.9	19.8
ride woman	27.3	30.2	23.5	27.4	28.5	30.8
white truck	17.6	30.6	23.3	29.4	9.4	12.7
fly bird	21.8	25.5	26.3	32.5	29.2	35.8
small cat	1.2	0.4	1.2	1.6	0.4	0.0
brown dog	1.7	0.5	4.8	7.8	0.9	1.2
big plane	0.5	0.6	3.7	8.7	0.3	0.0
stand bird	18.4	18.6	16.9	24.0	16.8	19.5
white horse	4.9	10.9	16.9	21.3	10.1	10.1
small table	0.0	0.0	0.0	0.0	0.0	0.0
big cat	0.0	0.0	0.0	0.0	0.0	0.0
blue bus	6.5	22.7	21.3	33.2	23.5	32.1
hold child	17.8	19.8	10.3	13.8	14.8	13.0
big truck	0.0	0.7	2.9	0.7	0.0	1.0
white boat	1.6	3.5	3.3	4.9	1.6	1.9
small dog	0.0	0.1	1.2	0.9	0.0	0.3
eat horse	23.0	22.1	44.6	39.4	49.3	56.3
stand child	9.1	10.4	13.0	10.3	5.7	7.4
black bird	4.9	4.4	12.3	6.4	4.9	9.3

Table 9: R@5 for each of the held out concept pairs in the validation sets.

BUTD: a man standing in front of a pizza
BUTD + SEQUENTIAL: a couple of people that are standing around a pizza
BUTD + INTERLEAVE: a man *sitting* at a table with a pizza
BUTD + MULTI-TASK: a man *sitting* at a table with a pizza

BUTD: a baby laying on a bed with a teddy bear
BUTD + SEQUENTIAL: a couple of kids sitting on a bed
BUTD + INTERLEAVE: a large teddy bear sitting on top of a bed
BUTD + MULTI-TASK: two children sitting on a bed with a laptop

---

BUTD: a man sitting on a bed using a laptop computer
BUTD + POS: a person sitting on a bed with a laptop

BUTR$_{weight}$: a woman sitting on a bed and using a laptop
BUTR$_{weight}$ + POS: a woman sitting on a bed with a laptop and a laptop

$\mathcal{M}^2$-TRM: a man sitting on a couch using a laptop computer
$\mathcal{M}^2$-TRM + POS: a woman sitting on a bed using a laptop computer

---

BUTD: a plane sitting on top of a lush green field
BUTD + POS: a red and white plane in an open field

BUTR$_{weight}$: a red and white plane taking off on a field
BUTR$_{weight}$ + POS: a red and white plane on a lush green field

$\mathcal{M}^2$-TRM: an airplane is on the runway in a field
$\mathcal{M}^2$-TRM + POS: a red and white plane sitting on a runway

---

BUTD: a couple of sheep standing on top of a lush green field
BUTD + POS: a couple of animals standing in the grass

BUTR$_{weight}$: two small lambs standing in a green field
BUTR$_{weight}$ + POS: two baby bears standing in a grassy field

$\mathcal{M}^2$-TRM: a mother sheep and a baby sheep in a field
$\mathcal{M}^2$-TRM + POS: a baby sheep standing next to an adult sheep in

Figure 5: More examples of generated captions from the validation sets. While syntax-aware approaches generate more accurate captions overall, they are sometimes worse than the standard system (second example). Moreover, the last example shows how most systems confuse a kitty and a puppy with two sheep, lambs or bears.