Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions

Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara
University of Modena and Reggio Emilia
{name.surname}@unimore.it

Abstract

Current captioning approaches can describe images using black-box architectures whose behavior is hardly controllable and explainable from the exterior. As an image can be described in infinite ways depending on the goal and the context at hand, a higher degree of controllability is needed to apply captioning algorithms in complex scenarios. In this paper, we introduce a novel framework for image captioning which can generate diverse descriptions by allowing both grounding and controllability. Given a control signal in the form of a sequence or set of image regions, we generate the corresponding caption through a recurrent architecture which predicts textual chunks explicitly grounded on regions, following the constraints of the given control. Experiments are conducted on Flickr30k Entities and on COCO Entities, an extended version of COCO in which we add grounding annotations collected in a semi-automatic manner. Results demonstrate that our method achieves state of the art performances on controllable image captioning, in terms of caption quality and diversity. Code and annotations are publicly available at: https://github.com/aimagelab/show-control-and-tell.

1. Introduction

Image captioning brings vision and language together in a generative way. As a fundamental step towards machine intelligence, this task has been recently gaining much attention thanks to the spread of Deep Learning architectures which can effectively describe images in natural language [42, 18, 46, 43]. Image captioning approaches are usually capable of learning a correspondence between an input image and a probability distribution over time, from which captions can be sampled either using a greedy decoding strategy [43], or more sophisticated techniques like beam search and its variants [1].

As the two main components of captioning architectures are the image encoding stage and the language model, researchers have focused on improving both phases, which resulted in the emergence of attentive models [46] on one side, and of more sophisticated interactions with the language model on the other [25, 5]. Recently, attentive models have been improved by replacing the attention over a grid of features with attention over image regions [3, 44, 50]. In these models, the generative process attends a set of regions which are softly selected while generating the caption.

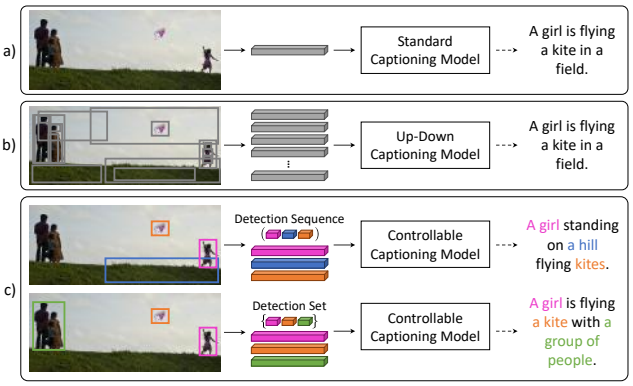

Figure 1: Comparison between (a) captioning models with global visual feature [43], (b) attentive models which integrate features from image regions [3] and (c) our *Show, Control and Tell*. Our method can produce multiple captions for a given image, depending on a control signal which can be either a sequence or a set of image regions. Moreover, chunks of the generated sentences are explicitly grounded on regions.

Despite these advancements, captioning models still lack controllability and explainability – *i.e.*, their behavior can hardly be influenced and explained. As an example, in the case of attention-driven models, the architecture implicitly selects which regions to focus on at each timestep, but it cannot be supervised from the exterior. While an image can be described in multiple ways, such an architecture provides no way of controlling which regions are described and what importance is given to each region. This lack of controllability creates a distance between human and machine

intelligence, as humans can manage the variety of ways in which an image can be described, and select the most appropriate one depending on the task and the context at hand. Most importantly, this also limits the applicability of captioning algorithms to complex scenarios in which some control over the generation process is needed. As an example, a captioning-based driver assistance system would need to focus on dangerous objects on the road to alert the driver, rather than describing the presence of trees and cars when a risky situation is detected. Eventually, such systems would also need to be explainable, so that their behavior could be easily interpreted in case of failures.

In this paper, we introduce *Show, Control and Tell*, that explicitly addresses these shortcomings (Fig. 1). It can generate diverse natural language captions depending on a control signal which can be given either as a sequence or as a set of image regions which need to be described. As such, our method is capable of describing the same image by focusing on different regions and in a different order, following the given conditioning. Our model is built on a recurrent architecture which considers the decomposition of a sentence into noun chunks and models the relationship between image regions and textual chunks, so that the generation process can be explicitly grounded on image regions. To the best of our knowledge, this is the first captioning framework controllable from image regions.

Contributions. Our contributions are as follows:

- We propose a novel framework for image captioning which is controllable from the exterior, and which can produce natural language captions explicitly grounded on a sequence or a set of image regions.
- The model explicitly considers the hierarchical structure of a sentence by predicting a sequence of noun chunks. Also, it takes into account the distinction between visual and textual words, thus providing an additional grounding at the word level.
- We evaluate the model with respect to a set of carefully designed baselines, on Flickr30k Entities and on COCO, which we semi-automatically augment with grounding image regions for training and evaluation purposes.
- Our proposed method achieves state of the art results for controllable image captioning on Flick30k and COCO both in terms of diversity and caption quality, even when compared with methods which focus on diversity.

2. Related work

A large number of models has been proposed for image captioning [37, 47, 24, 23, 17, 25]. Generally, all integrate recurrent neural networks as language models, and a representation of the image which might be given by the output of one or more layer of a CNN [43, 10, 37, 24], or by a time-varying vector extracted with an attention mechanism [46, 48, 24, 7, 3] selected either from a grid over CNN features, or integrating image regions eventually extracted from a detector [32, 3]. Attentive models provided a first way of grounding words to parts of the image, although with a blurry indication which was rarely semantically significant. Regarding the training strategies, notable advances have been made by using Reinforcement Learning to train non-differentiable captioning metrics [35, 23, 37]. In this work, we propose an extended version of this approach which deals with multiple output distributions and rewards the alignment of the caption to the control signal.

Recently, more principled approaches have been proposed for grounding a caption on the image [34, 38, 15, 16]: DenseCap [17] generates descriptions for specific image regions. Further, the Neural Baby Talk approach [25] extends the attentive model in a two-step design in which a word-level sentence template is firstly generated and then filled by object detectors with concepts found in the image. We instead decompose the caption at the level of noun chunks, and explicitly ground each of them to a region. This approach has the additional benefit of providing an explicability method at the chunk level.

Another related line of work is that of generating diverse descriptions. Some works have extended the beam-search algorithm to sample multiple captions from the same distribution [41, 1], while different GAN-based approaches have also appeared [8, 39, 45]. Most of these improve on diversity, but suffer on accuracy and do not provide controllability over the generation process. Others have conditioned the generation with a specific style or sentiment [27, 28, 11]. Our work is mostly related to [9], which uses a control input as a sequence of part-of-speech tags. This approach, while generating diversity, is hardly employable to effectively control the generation of the sentence; in contrast, we use image regions as a controllability method.

3. Method

Sentences are natural language structures which are hierarchical by nature [26]. At the lowest level, a sentence might be thought as a sequence of words: in the case of a sentence describing an image, we can further distinguish between *visual* words, which describe something visually present in the image, and *textual* words, that refer to entities which are not present in the image [25]. Analyzing further the syntactic dependencies between words, we can recover a higher abstraction level in which words can be organized into a tree-like structure: in a dependency tree [12, 14, 6], each word is linked together with its modifiers (Fig. 2).

Given a dependency tree, nouns can be grouped with their modifiers, thus building *noun chunks*. For instance, the caption depicted in Fig. 2 can be decomposed into a sequence of different noun chunks: “a young boy”, “a cap”,

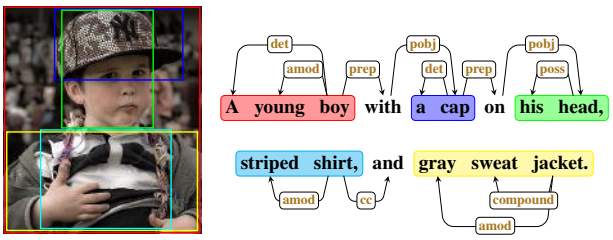

Figure 2: Example of a dependency tree for a caption. Noun chunks are marked with rounded boxes; chunks corresponding to image regions are depicted using the same color.

"his head", "striped shirt", and "gray and sweat jacket". As noun chunks, just like words, can be visually grounded into image regions, a caption can also be mapped to a sequence of regions, each corresponding to a noun chunk. A chunk might also be associated with multiple image regions of the same class if more than one possible mapping exists.

The number of ways in which an image can be described results in different sequences of chunks, linked together to form a fluent sentence. Therefore, captions also differ in terms of the set of considered regions, the order in which they are described, and their mapping to chunks given by the linguistic abilities of the annotator.

Following these premises, we define a model which can recover the variety of ways in which an image can be described, given a control input expressed as a sequence or set of image regions. We begin by presenting the former case, and then show how our model deals with the latter scenario.

3.1. Generating controllable captions

Given an image $\boldsymbol{I}$ and an ordered sequence of set of regions $\boldsymbol{R} = (\boldsymbol{r}_0, \boldsymbol{r}_1, ..., \boldsymbol{r}_N)$[1], the goal of our captioning model is to generate a sentence $\boldsymbol{y} = (y_0, y_1, ..., y_T)$ which in turns describes all the regions in $\boldsymbol{R}$ while maintaining the fluency of language.

Our model is conditioned on both the input image $\boldsymbol{I}$ and the sequence of region sets $\boldsymbol{R}$, which acts as a control signal, and jointly predicts two output distributions which correspond to the word-level and chunk-level representation of the sentence: the probability of generating a word at a given time, *i.e.* $p(y_t|\boldsymbol{R}, \boldsymbol{I}; \boldsymbol{\theta})$, and that of switching from one chunk to another, *i.e.* $p(g_t|\boldsymbol{R}, \boldsymbol{I}; \boldsymbol{\theta})$, where g_t is a boolean chunk-shifting gate. During the generation, the model maintains a pointer to the current region set $\boldsymbol{r}_i$ and can shift to the next element in $\boldsymbol{R}$ by means of the gate g_t.

To generate the output caption, we employ a recurrent neural network with adaptive attention. At each timestep, we compute the hidden state $\boldsymbol{h}_t$ according to the previous hidden state $\boldsymbol{h}_{t-1}$, the current image region set $\boldsymbol{r}_t$ and the current word w_t, such that $\boldsymbol{h}_t = \text{RNN}(w_t, \boldsymbol{r}_t, \boldsymbol{h}_{t-1})$. At training time, $\boldsymbol{r}_t$ and w_t are the ground-truth region set and word corresponding to timestep t; at test time, w_t is sampled from the first distribution predicted by the model, while the choice of the next image region is driven by the values of the chunk-shifting gate sampled from the second distribution:

$$\boldsymbol{r}_{t+1} \leftarrow \boldsymbol{R}[i], \quad \text{where } i = \min\left(\sum_{k=1}^{t} g_k, N\right), \; g_k \in \{0, 1\} \tag{1}$$

where $\{g_k\}_k$ is the sequence of sampled gate values, and N is the number of region sets in $\boldsymbol{R}$.

Chunk-shifting gate. We compute $p(g_t|\boldsymbol{R})$ via an adaptive mechanism in which the LSTM computes a compatibility function between its internal state and a latent representation which models the state of the memory at the end of a chunk. The compatibility score is compared to that of attending one of the regions in $\boldsymbol{r}_t$, and the result is used as an indicator to switch to the next region set in $\boldsymbol{R}$.

The LSTM is firstly extended to obtain a chunk sentinel $\boldsymbol{s}_t^c$, which models a component extracted from the memory encoding the state of the LSTM at the end of a chunk. The sentinel is computed as:

$$\boldsymbol{l}_t^c = \sigma(\boldsymbol{W}_{ig}\boldsymbol{x}_t + \boldsymbol{W}_{hg}\boldsymbol{h}_{t-1}) \tag{2}$$

$$\boldsymbol{s}_t^c = \boldsymbol{l}_t^c \odot \tanh(\boldsymbol{m}_t) \tag{3}$$

where $\boldsymbol{W}_{ig} \in \mathbb{R}^{d \times k}$, $\boldsymbol{W}_{hg} \in \mathbb{R}^{d \times d}$ are learnable weights, $\boldsymbol{m}_t \in \mathbb{R}^d$ is the LSTM cell memory and $\boldsymbol{x}_t \in \mathbb{R}^k$ is the input of the LSTM at time t; $\odot$ represents the Hadamard element-wise product and σ the sigmoid logistic function.

We then compute a compatibility score between the internal state $\boldsymbol{h}_t$ and the sentinel vector through a single-layer neural network; analogously, we compute a compatibility function between $\boldsymbol{h}_t$ and the regions in $\boldsymbol{r}_t$.

$$z_t^c = \boldsymbol{w}_h^T \tanh(\boldsymbol{W}_{sg}\boldsymbol{s}_t^c + \boldsymbol{W}_g\boldsymbol{h}_t) \tag{4}$$

$$\boldsymbol{z}_t^r = \boldsymbol{w}_h^T \tanh(\boldsymbol{W}_{sr}\boldsymbol{r}_t + (\boldsymbol{W}_g\boldsymbol{h}_t)\mathbb{1}^T) \tag{5}$$

where n is the number of regions in $\boldsymbol{r}_t$, $\mathbb{1} \in \mathbb{R}^n$ is a vector with all elements set to 1, $\boldsymbol{w}_h^T$ is a row vector, and all $\boldsymbol{W}_*$, $\boldsymbol{w}_*$ are learnable parameters. Notice that the representation extracted from the internal state is shared between all compatibility scores, as if the region set and the sentinel vector were part of the same attentive distribution. Contrarily to an attentive mechanism, however, there is no value extraction.

The probability of shifting from one chunk to the next one is defined as the probability of attending the sentinel vector $\boldsymbol{s}_t^c$ in a distribution over $\boldsymbol{s}_t^c$ and the regions in $\boldsymbol{r}_t$:

$$p(g_t = 1|\boldsymbol{R}) = \frac{\exp z_t^c}{\exp z_t^c + \sum_{i=1}^{n} \exp \boldsymbol{z}_{ti}^r} \tag{6}$$

where $\boldsymbol{z}_{ti}^r$ indicates the i-th element in $\boldsymbol{z}_t^r$, and we dropped the dependency between n and t for clarity. At test time, the

[1]For generality, we will always consider sequences of sets of regions, to deal with the case in which a chunk in the target sentence can be associated to multiple regions in training and evaluation data.

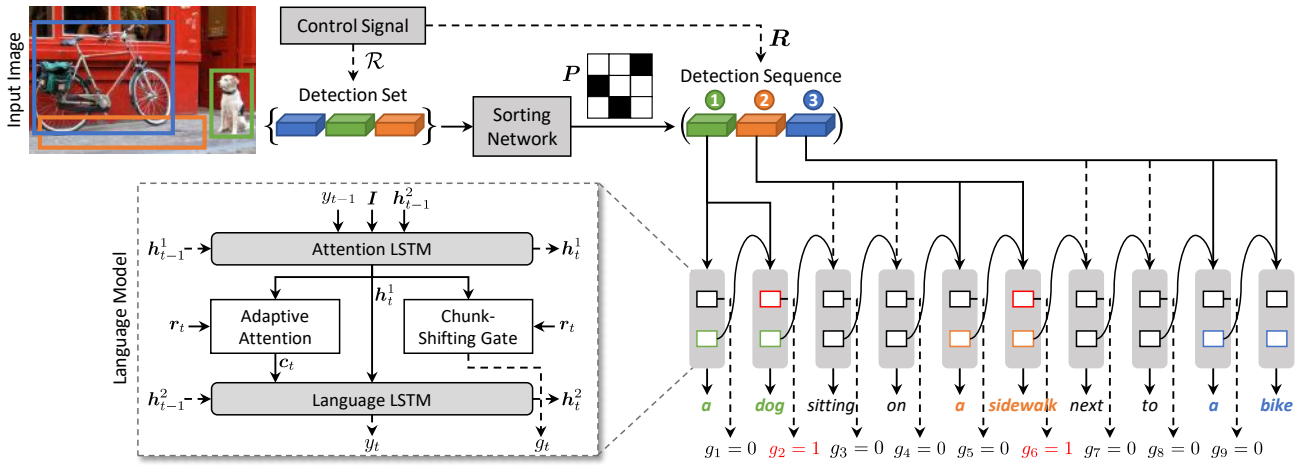

Figure 3: Overview of the approach. Given an image and a control signal, the figure shows the process to generate the controlled caption and the architecture of the language model.

value of gate $g_t \in \{0, 1\}$ is then sampled from $p(g_t|\boldsymbol{R})$ and drives the shifting to the next region set in $\boldsymbol{R}$.

Adaptive attention with visual sentinel. While the chunk-shifting gate predicts the end of a chunk, thus linking the generation process with the control signal given by $\boldsymbol{R}$, once $\boldsymbol{r}_t$ has been selected a second mechanism is needed to attend its regions and distinguish between visual and textual words. To this end, we build an adaptive attention mechanism with a visual sentinel [24].

The visual sentinel vector models a component of the memory to which the model can fall back when it chooses to not attend a region in $\boldsymbol{r}_t$. Analogously to Eq. 2, it is defined as:

$$\boldsymbol{l}_t^v = \sigma(\boldsymbol{W}_{is}\boldsymbol{x}_t + \boldsymbol{W}_{hs}\boldsymbol{h}_{t-1}) \tag{7}$$

$$\boldsymbol{s}_t^v = \boldsymbol{l}_t^v \odot \tanh(\boldsymbol{m}_t) \tag{8}$$

where $\boldsymbol{W}_{is} \in \mathbb{R}^{d\times k}$ and $\boldsymbol{W}_{hs} \in \mathbb{R}^{d\times d}$ are matrices of learnable weights. An attentive distribution is then generated over the regions in $\boldsymbol{r}_t$ and the visual sentinel vector $\boldsymbol{s}_t^v$:

$$\boldsymbol{\alpha}_t = \text{softmax}([\boldsymbol{z}_t^r; \boldsymbol{w}_h^T \tanh(\boldsymbol{W}_{ss}\boldsymbol{s}_t^v + \boldsymbol{W}_g\boldsymbol{h}_t)]) \tag{9}$$

where $[\cdot]$ indicates concatenation. Based on the attention distribution, we obtain a context vector which can be fed to the LSTM as a representation of what the network is attending:

$$\boldsymbol{c}_t = \sum_{i=1}^{n+1} \boldsymbol{\alpha}_{ti}[\boldsymbol{r}_t; \boldsymbol{s}_t^v] \tag{10}$$

Notice that the context vector will be, mostly, an approximation of one of the regions in $\boldsymbol{r}_t$ or the visual sentinel. However, $\boldsymbol{r}_t$ will vary at different timestep according to the chunk-shifting mechanism, thus following the control input. The model can alternate the generation of visual and textual words by means of the visual sentinel.

3.2. Objective

The captioning model is trained using a loss function which considers the two output distributions of the model. Given the target ground-truth caption $\boldsymbol{y}_{1:T}^*$, the ground-truth region sets $\boldsymbol{r}_{1:T}^*$ and chunk-shifting gate values corresponding to each timestep $g_{1:T}^*$, we train both distributions by means of a cross-entropy loss. The relationship between target region sets and gate values will be further expanded in the implementation details. The loss function for a sample is defined as:

$$\begin{aligned} L(\theta) = -\sum_{t=1}^{T} \Big(& \log \overbrace{p(y_t^*|\boldsymbol{r}_{1:t}^*, \boldsymbol{y}_{1:t-1}^*)}^{\text{Word-level probability}} + \\ & + g_t^* \log p(g_t = 1|\boldsymbol{r}_{1:t}^*, \boldsymbol{y}_{1:t-1}^*) + \\ & + (1 - g_t^*) \log(1 - \underbrace{p(g_t = 1|\boldsymbol{r}_{1:t}^*, \boldsymbol{y}_{1:t-1}^*)}_{\text{Chunk-level probability}}) \Big) \end{aligned} \tag{11}$$

Following previous works [35, 37, 3], after a pre-training step using cross-entropy, we further optimize the sequence generation using Reinforcement Learning. Specifically, we use the self-critical sequence training approach [37], which baselines the REINFORCE algorithm with the reward obtained under the inference model at test time.

Given the nature of our model, we extend the approach to work on multiple output distributions. At each timestep, we sample from both $p(y_t|\boldsymbol{R})$ and $p(g_t|\boldsymbol{R})$ to obtain the next word w_{t+1} and region set $\boldsymbol{r}_{t+1}$. Once a EOS tag is reached, we compute the reward of the sampled sentence $\boldsymbol{w}^s$ and backpropagate with respect to both the sampled word sequence $\boldsymbol{w}^s$ and the sequence of chunk-shifting gates $\boldsymbol{g}^s$. The final gradient expression is thus:

$$\nabla_\theta L(\theta) = -(r(\boldsymbol{w}^s) - b)(\nabla_\theta \log p(\boldsymbol{w}^s) + \nabla_\theta \log p(\boldsymbol{g}^s)) \tag{12}$$

where $b = r(\hat{\boldsymbol{w}})$ is the reward of the sentence obtained using the inference procedure (*i.e.* by sampling the word and gate value with maximum probability). We then build a reward function which jointly considers the quality of the caption and its alignment with the control signal $\boldsymbol{R}$.

Rewarding caption quality. To reward the overall quality of the generated caption, we use image captioning metrics as a reward. Following previous works [3], we employ the CIDEr metric (specifically, the CIDEr-D score) which has been shown to correlate better with human judgment [40].

Rewarding the alignment. While captioning metrics can reward the semantic quality of the sentence, none of them can evaluate the alignment with respect to the control input[2]. Therefore, we introduce an alignment score based on the Needleman-Wunsch algorithm [30].

Given a predicted caption $\boldsymbol{y}$ and its target counterpart $\boldsymbol{y}^*$, we extract all nouns from both sentences, and evaluate the alignment between them, recalling the relationships between noun chunks and region sets. We use the following scoring system: the reward for matching two nouns is equal to the cosine similarity between their word embeddings; a gap gets a negative reward equal to the minimum similarity value, *i.e.* -1. Once the optimal alignment is computed, we normalize its score, $al(\boldsymbol{y}, \boldsymbol{y}^*)$ with respect to the length of the sequences. The alignment score is thus defined as:

$$\text{NW}(\boldsymbol{y}, \boldsymbol{y}^*) = \frac{al(\boldsymbol{y}, \boldsymbol{y}^*)}{\max(\#\boldsymbol{y}, \#\boldsymbol{y}^*)} \tag{13}$$

where $\#\boldsymbol{y}$ and $\#\boldsymbol{y}^*$ represent the number of nouns contained in $\boldsymbol{y}$ and $\boldsymbol{y}^*$, respectively. Notice that $\text{NW}(\cdot, \cdot) \in [-1, 1]$. The final reward that we employ is a weighted version of CIDEr-D and the alignment score.

3.3. Controllability through a set of detections

The proposed architecture, so far, can generate a caption controlled by a sequence of region sets $\boldsymbol{R}$. To deal with the case in which the control signal is unsorted, *i.e.* a set of regions sets, we build a *sorting network* which can arrange the control signal in a candidate order, learning from data. The resulting sequence can then be given to the captioning network to produce the output caption (Fig. 3).

To this aim, we train a network which can learn a permutation, taking inspiration from Sinkhorn networks [29]. As shown in [29], the non-differentiable parameterization of a permutation can be approximated in terms of a differentiable relaxation, the so-called Sinkhorn operator. While a permutation matrix has exactly one entry of 1 in each row and each column, the Sinkhorn operator iteratively normalizes rows and columns of any matrix to obtain a "soft" permutation matrix, *i.e.* a real-valued matrix close to a permutation one.

Given a set of region sets $\mathcal{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_N\}$, we learn a mapping from $\mathcal{R}$ to its sorted version $\boldsymbol{R}^*$. Firstly, we pass each element in $\mathcal{R}$ through a fully-connected network which processes every item of a region set independently and produces a single output feature vector with length N. By concatenating together the feature vectors obtained for all region sets, we thus get a $N \times N$ matrix, which is then passed to the Sinkhorn operator to obtain the soft permutation matrix $\boldsymbol{P}$. The network is then trained by minimizing the mean square error between the scrambled input and its reconstructed version obtained by applying the soft permutation matrix to the sorted ground-truth, *i.e.* $\boldsymbol{P}^T\boldsymbol{R}^*$.

At test time, we take the soft permutation matrix and apply the Hungarian algorithm [20] to obtain the final permutation matrix, which is then used to get the sorted version of $\mathcal{R}$ for the captioning network.

3.4. Implementation details

Language model and image features. We use a language model with two LSTM layers (Fig. 3): the input of the bottom layer is the concatenation of the embedding of the current word, the image descriptor, as well as the hidden state of the second layer. This layer predicts the context vector via the visual sentinel as well as the chunk-gate. The second layer, instead, takes as input the context vector and the hidden state of the bottom layer and predicts the next word.

To represent image regions, we use Faster R-CNN [36] with ResNet-101 [13]. In particular, we employ the model finetuned on the Visual Genome dataset [19] provided by [3]. As image descriptor, following the same work [3], we average the feature vectors of all the detections.

The hidden size of the LSTM layers is set to 1000, and that of attention layers to 512, while the input word embedding size is set to 1000.

Ground-truth chunk-shifting gate sequences. Given a sentence where each word of a noun chunk is associated to a region set, we build the chunk-shifting gate sequence $\{g_t^*\}_t$ by setting g_t^* to 1 on the last word of every noun chunk, and 0 otherwise. The region set sequence $\{\boldsymbol{r}_t^*\}_t$ is built accordingly, by replicating the same region set until the end of a noun chunk, and then using the region set of the next chunk. To compute the alignment score and for extracting dependencies, we use the spaCy NLP toolkit[3]. We use GloVe [33] as word vectors.

Sorting network. To represent regions, we use Faster R-CNN vectors, the normalized position and size and the GloVe embedding of the region class. Additional details on architectures and training can be found in the Supplementary material.

[2] Although METEOR creates an alignment with respect to the reference caption, this is done for each unigram, thus mixing semantic and alignment errors.

[3] https://spacy.io/

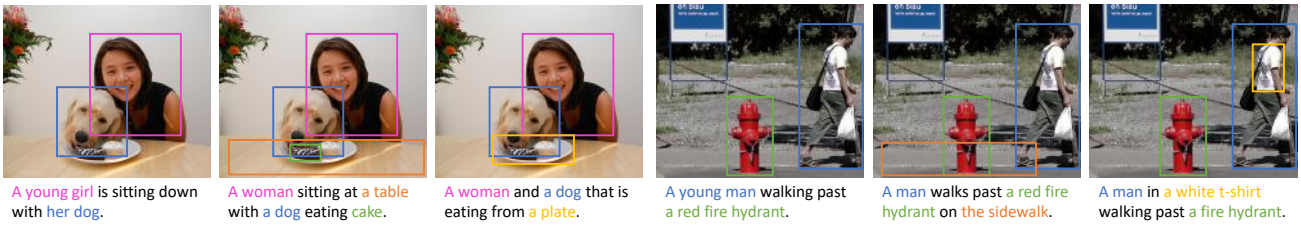

Figure 4: Sample captions and corresponding visual groundings from the COCO Entities dataset. Different colors show a correspondence between visual chunks and image regions.

COCO Entities (ours)	Train	Validation	Test
Nb. of captions	545,202	7,818	7,797
Nb. of noun chunks	1,518,667	20,787	20,596
Nb. of noun chunks per caption	2.79	2.66	2.64
Nb. of unique classes	1,330	725	730
Flickr30k Entities	Train	Validation	Test
Nb. of captions	144,256	5,053	4,982
Nb. of noun chunks	416,018	14,626	14,556
Nb. of noun chunks per caption	2.88	2.89	2.92
Nb. of unique classes	1,026	465	458

Table 1: Statistics on our COCO Entities dataset, in comparison with those of Flick30k Entities.

4. Experiments

4.1. Datasets

We experiment with two datasets: Flickr30k Entities, which already contains the associations between chunks and image regions, and COCO, which we annotate semi-automatically. Table 1 summarizes the datasets we use.

Flickr30k Entities [34]. Based on Flickr30k [49], it contains $31,000$ images annotated with five sentences each. Entity mentions in the caption are linked with one or more corresponding bounding boxes in the image. Overall, $276,000$ manually annotated bounding boxes are available. In our experiments, we automatically associate each bounding box with the image region with maximum IoU among those detected by the object detector. We use the splits provided by Karpathy *et al.* [18].

COCO Entities. Microsoft COCO [22] contains more than $120,000$ images, each of them annotated with around five crowd-sourced captions. Here, we again follow the splits defined by [18] and automatically associate noun chunks with image regions extracted from the detector [36].

We firstly build an index associating each noun of the dataset with the five most similar class names, using word vectors. Then, each noun chunk in a caption is associated by using either its name or the base form of its name, with the first class found in the index which is available in the image. This association process, as confirmed by an extensive manual verification step, is generally reliable and produces few false positive associations. Naturally, it can result in region sets with more than one element (as in Flickr30k), and noun chunks with an empty region set. In this case, we fill empty training region sets with the most probable detections of the image and let the adaptive attention mechanism learn the corresponding association; in validation and testing, we drop those captions. Some examples of the additional annotations extracted from COCO are shown in Fig. 4.

4.2. Experimental setting

The experimental settings we employ is different from that of standard image captioning. In our scenario, indeed, the sequence of set of regions is a second input to the model which shall be consider when selecting the ground-truth sentences to compare against. Also, we employ additional metrics beyond the standard ones like BLEU-4 [31], METEOR [4], ROUGE [21], CIDEr [40] and SPICE [2].

When evaluating the controllability with respect to a sequence, for each ground-truth regions-image input $(\boldsymbol{R}, \boldsymbol{I})$, we evaluate against all captions in the dataset which share the same pair. Also, we employ the alignment score (NW) to evaluate how the model follows the control input.

Similarly, when evaluating the controllability with respect to a set of regions, given a set-image pair $(\mathcal{R}, \boldsymbol{I})$, we evaluate against all ground-truth captions which have the same input. To assess how the predicted caption covers the control signal, we also define a soft intersection-over-union (IoU) measure between the ground-truth set of nouns and its predicted counterpart, recalling the relationships between region sets and noun chunks. Firstly, we compute the optimal assignment between the two set of nouns, using distances between word vectors and the Hungarian algorithm [20], and define an intersection score between the two sets as the sum of assignment profits. Then, recalling that set union can be expressed in function of an intersection, we define the IoU measure as follows:

$$\mathrm{IoU}(\boldsymbol{y}, \boldsymbol{y}^*) = \frac{\mathrm{I}(\boldsymbol{y}, \boldsymbol{y}^*)}{\#\boldsymbol{y} + \#\boldsymbol{y}^* - \mathrm{I}(\boldsymbol{y}, \boldsymbol{y}^*)} \tag{14}$$

where $\mathrm{I}(\cdot, \cdot)$ is the intersection score, and the $\#$ operator represents the cardinality of the two sets of nouns.

Method	Cross-Entropy Loss						CIDEr Optimization						CIDEr + NW Optimization					
	B-4	M	R	C	S	NW	B-4	M	R	C	S	NW	B-4	M	R	C	S	NW
FC-2K† [37]	10.4	17.3	36.8	98.3	25.2	0.257	12.3	18.5	39.6	117.5	26.9	0.273	-	-	-	-	-	-
Up-Down† [3]	12.9	19.3	40.0	119.9	29.3	0.296	14.2	20.0	42.1	133.9	30.0	0.310	-	-	-	-	-	-
Neural Baby Talk† [25]	12.9	19.2	40.4	120.2	29.5	0.305	-	-	-	-	-	-	-	-	-	-	-	-
Controllable LSTM	11.4	18.1	38.5	106.8	27.6	0.275	12.8	18.9	40.9	123.0	28.5	0.290	12.9	19.3	41.3	124.0	28.9	0.341
Controllable Up-Down	17.3	23.0	46.7	161.0	39.1	0.396	17.4	22.9	47.1	168.5	39.0	0.397	17.9	23.6	48.2	171.3	40.7	0.443
Ours *w/* single sentinel	20.0	23.9	51.1	183.3	43.9	0.480	21.7	25.3	54.5	202.6	47.6	0.606	21.3	25.3	54.5	201.1	48.1	0.648
Ours *w/o* visual sentinel	20.8	**24.4**	52.4	191.2	45.1	**0.508**	22.2	25.4	55.0	206.2	47.6	0.607	21.5	25.1	54.7	202.2	48.1	0.639
Ours	**20.9**	**24.4**	**52.5**	**193.0**	**45.3**	**0.508**	**22.5**	**25.6**	**55.1**	**210.1**	**48.1**	**0.615**	**22.3**	**25.6**	**55.3**	**209.7**	**48.5**	**0.649**

Table 2: Controllability via a sequence of regions, on test portion of COCO Entities. NW refers to the visual chunk alignment measure defined in Sec. 3.2. The † marker indicates non-controllable methods.

Figure 5: Sample results of controllability via a sequence of regions. Different colors and numbers show the control sequence and the associations between chunks and regions.

Method	B-4	M	R	C	S	NW
Neural Baby Talk† [25]	8.5	13.5	31.7	53.9	17.9	0.090
Controllable LSTM	6.5	12.6	30.2	43.5	15.8	0.124
Controllable Up-Down	10.4	15.2	35.2	69.5	21.7	0.190
Ours *w/* single sentinel	10.7	16.1	38.1	76.5	22.8	0.260
Ours *w/o* visual sentinel	11.1	15.5	37.2	74.7	22.4	0.244
Ours	**12.5**	**16.8**	**38.9**	**84.0**	**23.5**	**0.263**

Table 3: Controllability via a sequence of regions, on the test portion of Flickr30K Entities.

4.3. Baselines

Controllable LSTM. We start from a model without attention: an LSTM language model with a single visual feature vector. Then, we generate a sequential control input by feeding a flattened version of $\boldsymbol{R}$ to a second LSTM and taking the last hidden state, which is concatenated to the visual feature vector. The structure of the language model resembles that of [3], without attention.

Controllable Up-Down. In this case, we employ the full Up-Down model from [3], which creates an attentive distribution over image regions and make it controllable by feeding only the regions selected in $\boldsymbol{R}$ and ignoring the rest. This baseline is not sequentially controllable.

Ours without visual sentinel. To investigate the role of the visual sentinel and its interaction with the gate sentinel, in this baseline we ablate our model by removing the visual sentinel. The resulting baseline, therefore, lacks a mechanism to distinguish between visual and textual words.

Ours with single sentinel. Again, we ablate our model by merging the visual and chunk sentinel: a single sentinel is used for both roles, in place of $\boldsymbol{s}_t^c$ and $\boldsymbol{s}_t^v$.

As further baselines, we also compare against non-controllable captioning approaches, like FC-2K [37], Up-Down [3], and Neural Baby Talk [25].

4.4. Quantitative results

Controllability through a sequence of detections. Firstly, we show the performance of our model when providing the full control signal as a sequence of region sets. Table 2 shows results on COCO Entities, in comparison with the aforementioned approaches. We can see that our method achieves state of the art results on all automatic evaluation metrics, outperforming all baselines both in terms of overall caption quality and in terms of alignment with the control signal. Using the cross-entropy pre-training, we outperform the Controllable LSTM and Controllable Up-Down by 32.0 on CIDEr and 0.112 on NW. Optimizing the model with CIDEr and NW further increases the alignment quality while maintaining outperforming results on all metrics, leading to a final 0.649 on NW, which outperforms the Controllable Up-Down baseline by a 0.25. Recalling that NW ranges from -1 to 1, this improvement amounts to a 12.5% of the full metric range.

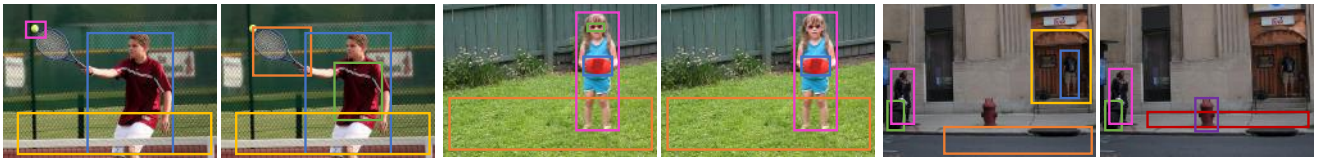

Figure 6: Sample results of controllability via a set of regions. Different colors show the control set and the associations between chunks and regions.

Method	B-4	M	R	C	S	IoU
FC-2K† [37]	12.5	18.5	39.6	116.5	26.6	61.0
Up-Down† [3]	14.4	20.0	42.2	132.8	29.7	63.2
Neural Baby Talk† [25]	13.1	19.2	40.5	119.1	29.2	62.6
Controllable LSTM	12.9	19.3	41.3	123.4	28.7	64.2
Controllable Up-Down	**18.1**	23.6	48.4	170.5	40.4	71.6
Ours *w/* single sentinel	17.4	23.6	48.4	168.4	43.7	75.4
Ours *w/o* visual sentinel	17.6	23.4	48.5	168.9	43.6	75.3
Ours	18.0	**23.8**	**48.9**	**173.3**	**44.1**	**75.5**

Table 4: Controllability via a set of regions, on the test portion of COCO Entities.

In Table 3, we instead show the results of the same experiments on Flickr30k Entities, using CIDEr+NW optimization for all controllable methods. Also on this manually annotated dataset, our method outperforms all the compared approaches by a significant margin, both in terms of caption quality and alignment with the control signal.

Controllability through a set of detections. We then assess the performance of our model when controlled with a set of detections. Tables 4 and 5 show the performance of our method in this setting, respectively on COCO Entities and Flickr30k Entities. We notice that the proposed approach outperforms all baselines and compared approaches in terms of IoU, thus testifying that we are capable of respecting the control signal more effectively. This is also combined with better captioning metrics, which indicate higher semantic quality.

Diversity evaluation. Finally, we also assess the diversity of the generated captions, comparing with the most recent approaches that focus on diversity. In particular, the variational autoencoder proposed in [45] and the approach of [9], which allows diversity and controllability by feeding PoS sequences. To test our method on a significant number of diverse captions, given an image we take all regions which are found in control region sets, and take the permutations which result in captions with higher log-probability. This approach is fairly similar to the sampling strategy used in [9], even if ours considers region sets. Then, we follow the experimental approach defined in [45, 9]: each ground-truth sentence is evaluated against the generated caption with the maximum score for each metric. Higher scores, thus, indicate that the method is capable of sampling high accuracy captions. Results are reported in Table 6, where to guarantee the fairness of the comparison, we run this experiments on the full COCO test split. As it can be seen, our method can generate significantly diverse captions.

Method	B-4	M	R	C	S	IoU
Neural Baby Talk† [25]	8.6	13.5	31.9	53.8	17.8	49.9
Controllable LSTM	6.4	12.5	30.2	42.9	15.6	50.8
Controllable Up-Down	10.5	15.2	35.5	69.5	21.6	54.8
Ours *w/* single sentinel	9.5	15.2	35.8	65.6	21.2	**55.0**
Ours *w/o* visual sentinel	9.8	14.8	35.0	64.2	20.9	54.3
Ours	**10.9**	**15.8**	**36.2**	**70.4**	**21.8**	**55.0**

Table 5: Controllability via a set of regions, on the test portion of Flickr30K Entities.

Method	Samples	B-4	M	R	C	S
AG-CVAE [45]	20	**47.1**	30.9	63.8	130.8	24.4
POS [9]	20	44.9	36.5	67.8	146.8	27.7
Ours	20	44.8	**36.6**	**68.9**	**156.5**	**30.9**

Table 6: Diversity performance on the test portion of COCO.

5. Conclusion

We presented *Show, Control and Tell*, a framework for generating controllable and grounded captions through regions. Our work is motivated by the need of bringing captioning systems to more complex scenarios. The approach considers the decomposition of a sentence into noun chunks, and grounds chunks to image regions following a control signal. Experimental results, conducted on Flickr30k and on COCO Entities, validate the effectiveness of our approach in terms of controllability and diversity.

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016. 1, 2

[2] P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *Proceedings of the European Conference on Computer Vision*, 2016. 6

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 4, 5, 7, 8

[4] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005. 6

[5] L. Baraldi, C. Grana, and R. Cucchiara. Hierarchical boundary-aware neural encoder for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[6] D. Chen and C. Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014. 2

[7] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(2):48, 2018. 2

[8] B. Dai, S. Fidler, R. Urtasun, and D. Lin. Towards Diverse and Natural Image Descriptions via a Conditional GAN. In *Proceedings of the International Conference on Computer Vision*, 2017. 2

[9] A. Deshpande, J. Aneja, L. Wang, A. Schwing, and D. A. Forsyth. Diverse and controllable image captioning with part-of-speech guidance. *arXiv preprint arXiv:1805.12589*, 2018. 2, 8

[10] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[11] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[12] Y. Goldberg and J. Nivre. Training deterministic parsers with non-deterministic oracles. *Transactions of the Association of Computational Linguistics*, 1:403–414, 2013. 2

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 5

[14] M. Honnibal, Y. Goldberg, and M. Johnson. A non-monotonic arc-eager transition system for dependency parsing. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 2013. 2

[15] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[16] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[17] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[18] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1, 6

[19] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 5

[20] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5, 6, 11

[21] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL Workshop*, volume 8, 2004. 6

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*, 2014. 6

[23] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the International Conference on Computer Vision*, 2017. 2

[24] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 4

[25] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural Baby Talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 7, 8

[26] C. D. Manning, C. D. Manning, and H. Schütze. *Foundations of statistical natural language processing*. MIT press, 1999. 2

[27] A. Mathews, L. Xie, and X. He. SemStyle: Learning to Generate Stylised Image Captions using Unaligned Text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[28] A. P. Mathews, L. Xie, and X. He. SentiCap: Generating Image Descriptions with Sentiments. In *Proceedings of the Conference on Articial Intelligence*, 2016. 2

[29] G. Mena, D. Belanger, S. Linderman, and J. Snoek. Learning latent permutations with gumbel-sinkhorn networks. *Proceedings of the International Conference on Learning Representations*, 2018. 5

[30] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970. 5

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002. 6

[32] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek. Areas of attention for image captioning. In *Proceedings of the International Conference on Computer Vision*, 2017. 2

[33] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014. 5

[34] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the International Conference on Computer Vision*, 2015. 2, 6

[35] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. In *Proceedings of the International Conference on Learning Representations*, 2015. 2, 4

[36] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. 5, 6

[37] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 4, 7, 8

[38] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *Proceedings of the European Conference on Computer Vision*, 2016. 2

[39] R. Shetty, M. Rohrbach, L. A. Hendricks, M. Fritz, and B. Schiele. Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training. In *Proceedings of the International Conference on Computer Vision*, 2017. 2

[40] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 5, 6

[41] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. J. Crandall, and D. Batra. Diverse Beam Search for Improved Description of Complex Scenes. In *Proceedings of the Conference on Articial Intelligence*, 2018. 2

[42] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1

[43] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663, 2017. 1, 2

[44] J. Wang, P. Madhyastha, and L. Specia. Object Counts! Bringing Explicit Detections Back into Image Captioning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018. 1

[45] L. Wang, A. Schwing, and S. Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems*, 2017. 2, 8

[46] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*, 2015. 1, 2

[47] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov. Review networks for caption generation. In *Advances in Neural Information Processing Systems*, 2016. 2

[48] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[49] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6

[50] Y. Zhang, J. C. Niebles, and A. Soto. Interpretable Visual Question Answering by Visual Grounding from Attention Supervision Mining. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2018. 1

A. Sorting network

We provide additional details on the architecture and training strategy of the sorting network. For the ease of the reader, a schema is reported in Fig. 7. Given a scrambled sequence of N region sets, each region is encoded through a fully connected network which returns a N-dimensional descriptor. The fully connected network employs visual, textual and geometric features: the Faster R-CNN vector of the detection (2048-d), the GloVe embedding of the region class (300-d) and the normalized position and size of the bounding-box (4-d). The visual vector is processed by two layers (512-d, 128-d), while the textual feature is processed by a single layer (128-d). The outputs of the visual and textual branches are then concatenated with the geometric features and fed through another fully connected layer (256-d). A final layer produces the resulting N-dimensional descriptors. All layers have ReLU activations, except for the last fully-connected which has a $\tanh$ activation. In case the region set contains more than one detection, we average-pool the resulting N-dimensional descriptors to obtain a single feature vector for a region set.

Once the feature vectors of the scrambled sequence are concatenated, we get a $N \times N$ matrix, which is then converted into a "soft" permutation matrix $\boldsymbol{P}$ through the Sinkhorn operator. The operator processes a N-dimensional square matrix $\boldsymbol{X}$ by applying L consecutive row-wise and column-wise normalization, as follows:

$$S^0(\boldsymbol{X}) = \exp(\boldsymbol{X}) \tag{15}$$

$$S^l(\boldsymbol{X}) = \mathcal{T}_c(\mathcal{T}_r(S^{l-1}(\boldsymbol{X}))) \tag{16}$$

$$\boldsymbol{P} := S^L(\boldsymbol{X}) \tag{17}$$

where $\mathcal{T}_r(\boldsymbol{X}) = \boldsymbol{X} \oslash (\boldsymbol{X}\mathbf{1}_N\mathbf{1}_N^T)$, and $\mathcal{T}_c(\boldsymbol{X}) = \boldsymbol{X} \oslash (\mathbf{1}_N\mathbf{1}_N^T\boldsymbol{X})$ are the row-wise and column-wise normalization operators, with $\oslash$ denoting element-wise division, $\mathbf{1}_N$ a column vector of N ones. At test time, once L normalizations ($L = 20$ in our experiments) have been performed, the resulting "soft" permutation matrix can be converted into a permutation matrix via the Hungarian algorithm [20].

At training time, instead, we measure the mean square error between the scrambled sequence and its reconstructed version obtained by applying the soft permutation matrix to the sorted ground-truth sequence $\boldsymbol{R}^*$, *i.e.* $\boldsymbol{P}^T\boldsymbol{R}^*$. On the implementation side, all tensors are appropriately masked to deal with variable-length sequences and sets. We set the maximum length of input scrambled sequences to 10.

In Table 7 we evaluate the quality of the rankings in terms of accuracy (proportion of completely correct rankings) and Kendall's Tau (correlation between GT and predicted ranking, between -1 and 1). We compare with a predefined local ranking (sorting detections with their probability), a predefined global ranking based on detection classes, and compare the Sinkorn network with a SVM Rank model trained on the same features. As it can be seen, the Sinkhorn network performs better than other baselines and can generate accurate rankings.

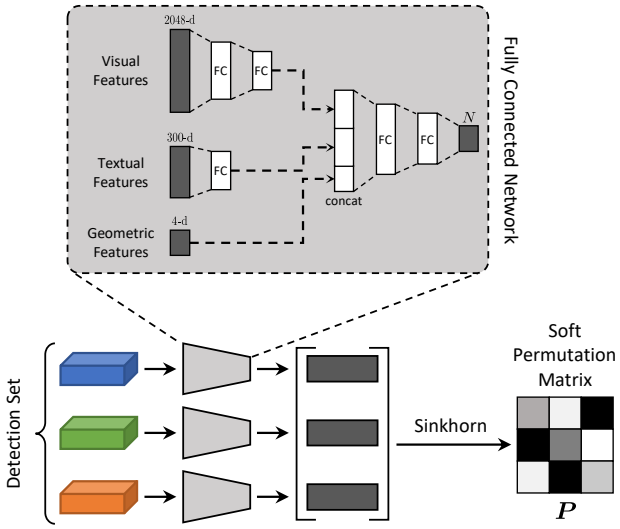

Figure 7: Schema of the sorting network.

	COCO Entities		Flickr Entities	
	Accuracy	Kendall's Tau	Accuracy	Kendall's Tau
Predefined local (det. prob.)	36.2%	0.145	40.0%	0.249
Predefined global (det. class)	59.1%	0.525	58.4%	0.565
SVM Rank	54.6%	0.448	49.5%	0.418
Sinkhorn Network	**67.1%**	**0.613**	**65.2%**	**0.633**

Table 7: Sorting network: experimental evaluation.

B. Training details

We used a weight of 0.2 for the word loss and 0.8 for the two chunk-level terms in Eq. 11. To train both the captioning model and the sorting network, we use the Adam optimizer with an initial learning rate of 5×10^{-4} decreased by a factor of 0.8 every epoch. For the captioning model, we run the reinforcement learning training with a fixed learning rate of 5×10^{-5}. We use a batch size of 100 for all our experiments. During caption decoding, we employ for all experiments the beam search strategy with a beam size of 5: similarly to what has been done when training with Reinforcement Learning, we sample from both output distribution to select the most probable sequence of actions. We use early stopping on validation CIDEr for the captioning network, and validation accuracy of the predicted permutations for the sorting network.

C. The COCO Entities dataset

In Fig. 8, we report additional examples of the semi-automatic annotation procedure used to collect COCO Entities. As in the main paper, we use different colors to visu-

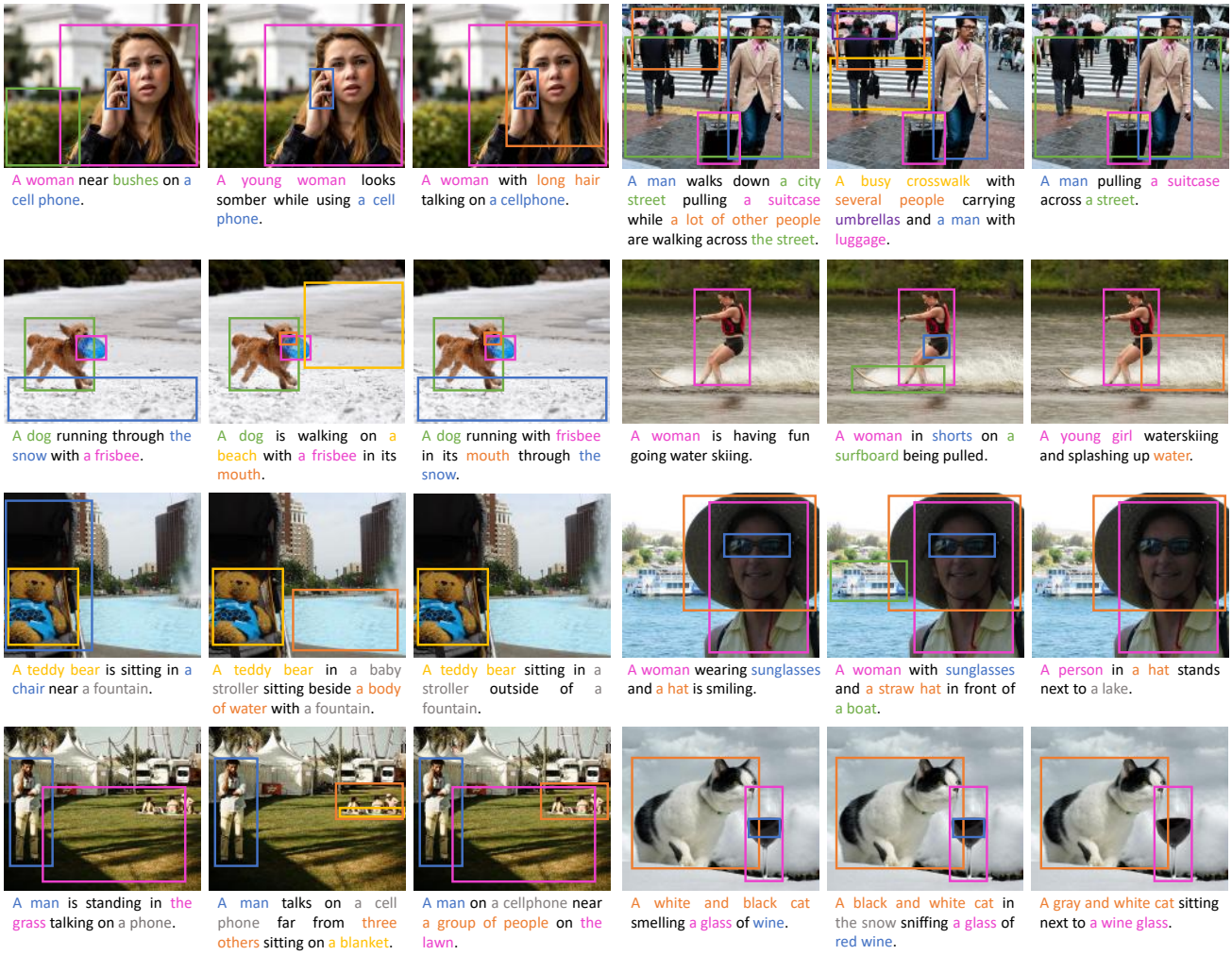

Figure 8: Additional sample captions and corresponding visual groundings from the COCO Entities dataset. Different colors show a correspondence between textual chunks and image regions. Gray color indicates noun chunks for which a visual grounding could not be found, either for missing detections or for errors in the noun-class association.

alize the correspondences between noun chunks and image regions. For the ease of visualization, we display a single region for chunk, even though multiple associations are possible. In this case, the region set would contain more than one element.

In the last two rows, we also report samples in which at least one noun chunk could not be assigned to any detection. Recall that in this case, at training time, we use the most probable detections of the image and let the adaptive attention mechanism learn the corresponding association: we found that this procedure, overall, increases the final accuracy of the network rather than feeding empty region sets. Captions with missing associations are dropped in validation and testing.

D. Additional experimental results

Tables 8, 9 and 10 report additional experimental results which have not been reported in the main paper for space constraints. In particular, Table 8 integrates Table 3 of the main paper by evaluating the controllability via a sequence of region sets on Flickr30K, when training with cross-entropy only, and when optimizing with CIDEr and CIDEr+NW. Analogously, Tables 9 and 10 analyze the controllabilty via a set of regions, on both Flickr30K and COCO Entities and with all training strategies.

We observe that the CIDEr+NW fine-tuning approach is effective on all settings, and that our model outperforms by a clear margin the baselines both when controlled via a sequence and when controlled by a scrambled set of regions, regardless of the careful choice of the baselines. The per-

Method	Cross-Entropy Loss						CIDEr Optimization						CIDEr + NW Optimization					
	B-4	M	R	C	S	NW	B-4	M	R	C	S	NW	B-4	M	R	C	S	NW
Controllable LSTM	6.5	12.0	29.6	40.4	15.7	0.078	6.7	12.1	30.0	45.5	15.8	0.079	6.5	12.6	30.2	43.5	15.8	0.124
Controllable Up-Down	10.1	15.2	34.9	69.2	21.6	**0.158**	10.1	14.8	35.0	69.3	21.2	0.148	10.4	15.2	35.2	69.5	21.7	0.190
Ours *w/* single sentinel	11.0	**15.5**	36.3	71.7	22.6	0.134	11.2	15.8	37.9	77.9	22.9	0.199	10.7	16.1	38.1	76.5	22.8	0.260
Ours *w/o* visual sentinel	10.8	14.9	35.4	69.3	22.2	0.142	11.1	15.5	36.8	75.0	22.2	0.197	11.1	15.5	37.2	74.7	22.4	0.244
Ours	**11.3**	15.4	**36.9**	**74.5**	**23.4**	0.152	**12.4**	**16.6**	**38.8**	**83.7**	**23.5**	**0.221**	**12.5**	**16.8**	**38.9**	**84.0**	**23.5**	**0.263**

Table 8: Controllability via a sequence of regions, on the test portion of Flickr30K Entities.

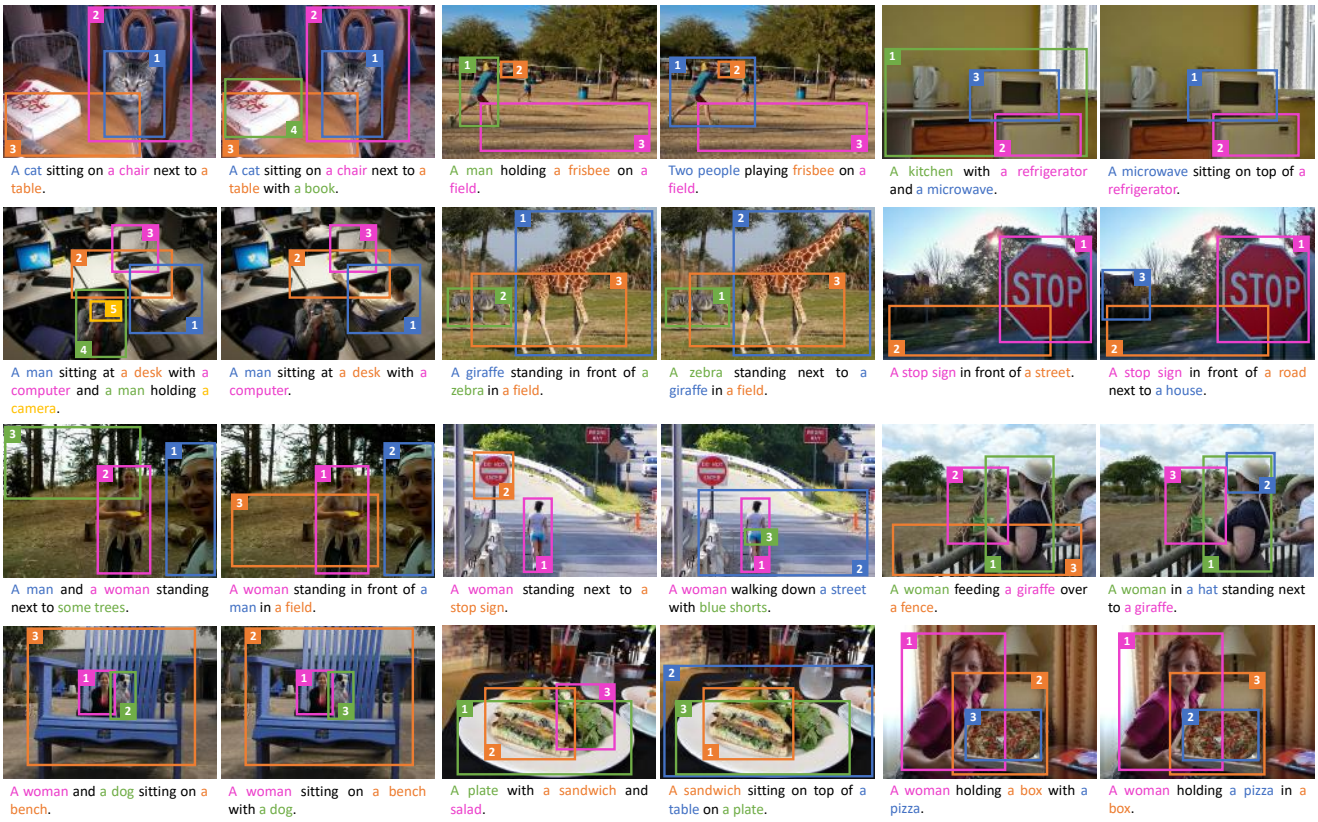

Figure 9: Additional sample results of controllability via a sequence of regions. Different colors and numbers show the control sequence and the associations between chunks and regions.

formance of the Controllable LSTM baseline is constantly significantly lower than that of the Controllable Up-Down, thus indicating both the importance of an attention mechanism and that of having a good representation of the control signal. The Controllable Up-Down baseline, however, shows lower performance when compared to our approach, in both sequence- and set-controlled scenarios.

E. Additional qualitative results

Finally, Fig. 9 and 10 report other qualitative results on COCO Entities. As in the main paper, the same image is reported multiple times with different control inputs: our method generates multiple captions for the same image, and can accurately follow the control input.

Method	Cross-Entropy Loss						CIDEr Optimization						CIDEr + NW Optimization					
	B-4	M	R	C	S	IoU	B-4	M	R	C	S	IoU	B-4	M	R	C	S	IoU
Controllable LSTM	11.5	18.1	38.5	105.8	27.1	60.7	12.9	18.9	40.9	122.0	28.2	62.0	12.9	19.3	41.3	123.4	28.7	0.642
Controllable Up-Down	17.5	23.0	46.9	160.6	38.8	69.2	17.7	22.9	47.3	167.6	38.7	69.4	**18.1**	23.6	48.4	170.5	40.4	71.6
Ours *w/* single sentinel	16.9	22.6	46.9	159.6	40.9	70.2	17.9	23.7	48.7	171.1	43.5	74.4	17.4	23.6	48.4	168.4	43.7	75.4
Ours *w/o* visual sentinel	**17.7**	23.1	47.9	166.6	**42.1**	71.3	18.1	23.7	48.9	172.5	43.3	74.2	17.6	23.4	48.5	168.9	43.6	75.3
Ours	**17.7**	**23.2**	**48.0**	**168.3**	**42.1**	**71.4**	**18.5**	**23.9**	**49.0**	**176.7**	**43.8**	**74.5**	18.0	**23.8**	**48.9**	**173.3**	**44.1**	**75.5**

Table 9: Controllability via a set of regions, on the test portion of COCO Entities.

Method	Cross-Entropy Loss						CIDEr Optimization						CIDEr + NW Optimization					
	B-4	M	R	C	S	IoU	B-4	M	R	C	S	IoU	B-4	M	R	C	S	IoU
Controllable LSTM	6.7	12.0	29.8	41.0	15.6	48.8	6.8	12.1	30.2	45.4	15.6	49.0	6.4	12.5	30.2	42.9	15.6	50.8
Controllable Up-Down	**10.1**	**15.2**	35.1	**68.8**	21.5	**53.6**	10.2	14.8	35.3	69.1	21.1	52.9	10.5	15.2	35.5	69.5	21.6	54.8
Ours *w/* single sentinel	**10.1**	**15.2**	**35.5**	67.5	21.7	52.5	10.1	15.3	36.1	68.9	21.7	53.5	9.5	15.2	35.8	65.6	21.2	**55.0**
Ours *w/o* visual sentinel	9.7	14.5	34.4	63.1	21.0	52.2	9.9	14.7	34.8	65.5	20.8	52.9	9.8	14.8	35.0	64.2	20.9	54.3
Ours	9.9	14.9	35.3	67.3	**22.2**	52.7	**10.8**	**15.7**	**36.4**	**71.3**	**22.0**	**53.9**	**10.9**	**15.8**	**36.2**	**70.4**	**21.8**	**55.0**

Table 10: Controllability via a set of regions, on the test portion of Flickr30K Entities.

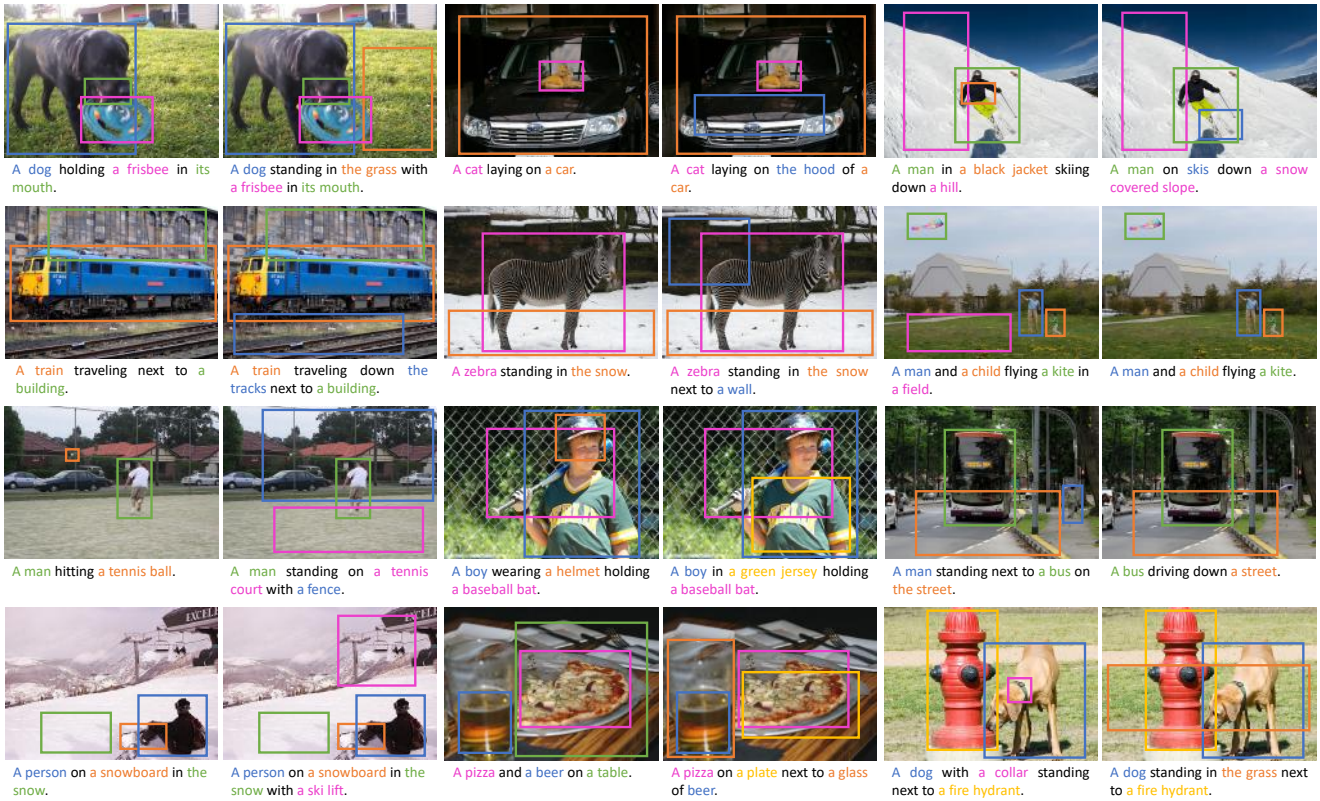

Figure 10: Additional sample results of controllability via a set of regions. Different colors show the control set and the associations between chunks and regions.