

# Learning to Generate Grounded Visual Captions without Localization Supervision

Chih-Yao Ma<sup>1,3</sup>, Yannis Kalantidis<sup>\*2</sup>, Ghassan AlRegib<sup>1</sup>,

Peter Vajda<sup>3</sup>, Marcus Rohrbach<sup>3</sup>, Zsolt Kira<sup>1</sup>

<sup>1</sup>Georgia Tech, <sup>2</sup>NAVER LABS Europe, <sup>3</sup>Facebook

{cyma, alregib, zkira}@gatech.edu, yannis.kalantidis@naverlabs.com, {cyma, vajdap, mrf}@fb.com

**Abstract.** When automatically generating a sentence description for an image or video, it often remains unclear how well the generated caption is grounded, that is whether the model uses the correct image regions to output particular words, or if the model is hallucinating based on priors in the dataset and/or the language model. The most common way of relating image regions with words in caption models is through an attention mechanism over the regions that are used as input to predict the next word. The model must therefore learn to predict the attentional weights without knowing the word it should localize. This is difficult to train without grounding supervision since recurrent models can propagate past information and there is no explicit signal to force the captioning model to properly ground the individual decoded words. In this work, we help the model to achieve this via a novel cyclical training regimen that forces the model to localize each word in the image *after* the sentence decoder generates it, and then reconstruct the sentence from the localized image region(s) to match the ground-truth. Our proposed framework only requires learning one extra fully-connected layer (the localizer), a layer that can be removed at test time. We show that our model significantly improves grounding accuracy without relying on grounding supervision or introducing extra computation during inference, for both image and video captioning tasks. Code is available at <https://github.com/chihyaoma/cyclical-visual-captioning>.

**Keywords:** image captioning, video captioning, self-supervised learning, visual grounding

## 1 Introduction

Vision and language tasks such as visual captioning or question answering, combine linguistic descriptions with data from real-world scenes. Deep learning models have achieved great success for such tasks, driven in part by the development of attention mechanisms that focus on various objects in the scene while generating captions. The resulting models, however, are known to have poor grounding performance [20], leading to undesirable behaviors (such as object hallucinations [27]), despite having high captioning accuracy. That is, they often do not

---

\* Work done while at Facebook.

correctly associate generated words with the appropriate image regions (*e.g.*, objects) in the scene, resulting in models that lack interpretability.

Several existing approaches have tried to improve the grounding of captioning models. One class of methods generate sentence *templates* with slot locations explicitly tied to specific image regions. These slots are then filled in by visual concepts identified by off-the-shelf object detectors [21]. Other methods have developed specific grounding or attention modules that aim to *attend* to the correct region(s) for generating visually groundable word. Such methods, however, rely on explicit supervision for optimizing the grounding or attention modules [20,47] and require bounding box annotations for each visually groundable word.

In this work, we propose a novel cyclical training regimen that is able to significantly improve grounding performance without any grounding annotations. An important insight of our work is that current models use attention mechanisms conditioned on the hidden features of recurrent modules such as LSTMs, which leads to effective models with high accuracy but entangle grounding and decoding. Since LSTMs are effective at propagating information across the decoding process, the network does not necessarily need to associate particular decoded words with their corresponding image region(s). However, for a captioning model to be visually grounded, the model has to predict attentional weights without knowing the word to localize.

Based on this insight, we develop a cyclical training regimen to force the network to ground individual decoded words: *decoding* → *localization* → *reconstruction*. Specifically, the model of the decoding stage can be any state-of-the-art captioning model; in this work, we follow GVD [47] to extend the widely used Up-Down model [2]. At the localization stage, each word generated by the first decoding stage is localized through a *localizer*, and the resulting grounded image region(s) are then used to reconstruct the ground-truth caption in the final stage. Both decoding and reconstruction stages are trained using a standard cross-entropy loss. Important to our method, both stages share the same decoder, thereby causing the localization stage to guide the decoder to improve its attention mechanism. Our method is simple and only adds a fully-connected layer to perform localization. During inference, we only use the (shared) decoder, thus we do not add any computational cost.

To compare with the state-of-the-art [47], we evaluate our proposed method on the challenging Flickr30k Entities image captioning dataset [25] and the ActivityNet-Entities video captioning dataset [47] on both captioning and grounding. In addition to the existing grounding metrics that calculate the grounding accuracy for each object class [47], we further include a grounding metric that compute grounding accuracy for each generated sentence. We achieve around 18% relative improvements averaged over grounding metrics in terms of bridging the gap between the unsupervised baseline and supervised methods on Flickr30k Entities and around 34% on ActivityNet-Entities. We further find that our method can even outperform the supervised method on infrequent words, owing to its self-supervised nature. In addition, we also conduct human eval-

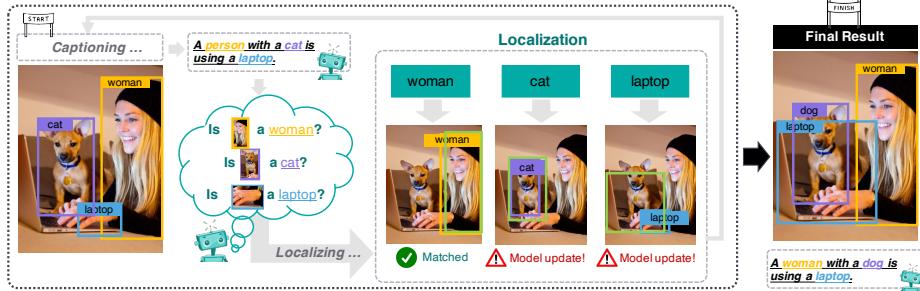


Fig. 1: Visual captioning models are often not visually-grounded. As humans, after generating a caption, we perform localization to check whether the generated caption is visually-grounded. If the localized image region is incorrect, we will correct it. Same goes for training a model. We would like to update the model accordingly. However, without the ground-truth grounding annotation, how does the model know the localized region is incorrect? How can the model then be updated? To overcome this issue, we propose to perform *localization* and *reconstruction* to regularize the captioning model to be visually-grounded without relying on the grounding annotations.

uation on visual grounding to further verify the improvement of the proposed method.

**Contributions summary.** We propose object re-localization as a form of self-supervision for grounded visual captioning and present a cyclical training regimen that re-generates sentences after re-localizing the objects conditioned on each word, implicitly imposing grounding consistency. We evaluate our proposed approach on both image and video captioning tasks. We show that the proposed training regime can boost grounding accuracy over a state-of-the-art baseline, enabling grounded models to be trained without bounding box annotations, while retaining high captioning quality across two datasets and various experimental settings.

## 2 Related work

**Visual captioning.** Neural models for visual captioning have received significant attention recently [2, 22, 21, 8, 38, 30, 37, 29, 33, 24]. Most current state-of-the-art models contain attention mechanisms, allowing the process to focus on subsets of the image when generating the next word. These attention mechanisms can be defined over spatial locations [39], semantic metadata [19, 43, 44, 49] or a predefined set of regions extracted via a region proposal network [22, 45, 2, 21, 6, 18]. In the latter case, off-the-shelf object detectors are first used to extract object proposals [26, 11] and the captioning model then learns to dynamically attend over them when generating the caption.

**Visual grounding.** Although attention mechanisms are generally shown to improve captioning quality and metrics, it has also been shown that they don't

really focus on the same regions as a human would [5]. This make models less trustworthy and interpretable, and therefore creating *grounded* image captioning models, *i.e.*, models that accurately link generated words or phrases to specific regions of the image, has recently been an active research area. A number of approaches have been proposed, *e.g.*, for grounding phrases or objects from image descriptions [28,14,42,7,47,46], grounding visual explanations [12], visual co-reference resolution for actors in video [29], or improving grounding via human supervision [31]. Recently, Zhou et al. [47] presented a model with self-attention based context encoding and direct grounding supervision that achieves state-of-the-art results in both the image and video tasks. They exploit ground-truth bounding box annotations to significantly improve the visual grounding accuracy. In contrast, we focus on reinforcing the visual grounding capability of the existing captioning model via a cyclical training regimen without using bounding box annotations and present a method that can increase grounding accuracy while maintaining comparable captioning performance with state-of-the-arts. Another closely related work is [28] where the authors focus on grounding or localizing a given textual phrase in an image. This creates a critical difference, as the ground-truth caption is provided during both training and test time and renders the training regimen proposed in [28] not applicable to visual captioning tasks.

**Cyclical training.** Cycle consistency [41,51,10,4] has been used recently in a wide range of domains, including machine translation [10], unpaired image-to-image translation [51], visual question answering [32], question answering [34], image captioning [4], video captioning [40,9], captioning and drawing [15] as well as domain adaptation [13]. While the cyclical training regime has been explored vastly in both vision and language domains, it has not yet been used for enforcing the *visual grounding* capability of a captioning model.

### 3 Method

**Notation.** For the visual captioning task we have pairs of target sentences and images (or videos). Each image (or video) is represented by spatial feature map(s) extracted by a ResNet-101 model and a bag of regions obtained from Faster-RCNN [26] as  $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N] \in \mathbb{R}^{d \times N}$ . The target sentence is represented as a sequence of one-hot vectors  $\mathbf{y}_t^* \in \mathbb{R}^s$ , where  $T$  is the sentence length,  $t \in 1, 2, \dots, T$ , and  $s$  is the dictionary size.

#### 3.1 Baseline

We reimplemented the model used in GVD [47] without self-attention for region feature encoding [22,35] as our baseline<sup>1</sup>. It is an extension of the state-of-the-art Up-Down [2] model with the *grounding-aware region encoding* (see Appendix D).

---

<sup>1</sup> We removed self-attention because we found that removing it slightly improved both captioning and grounding accuracy in our implementation.

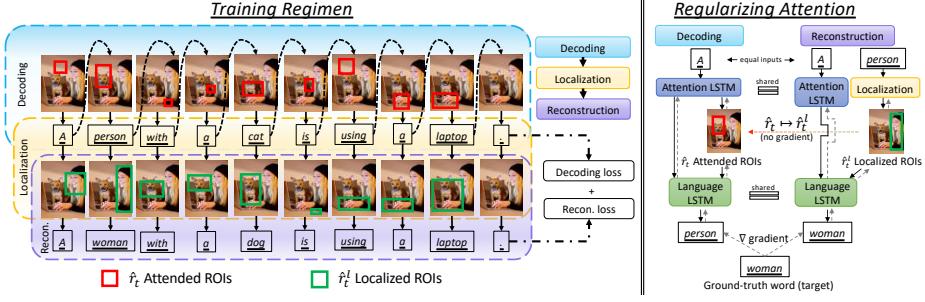


Fig. 2: (Left) Proposed cyclical training regimen: *decoding*  $\rightarrow$  *localization*  $\rightarrow$  *reconstruction*. The decoder attends to the image regions and sequentially generates each of the output words. The localizer then uses the generated words as input to locate the image regions. Finally, the shared decoder during the reconstruction stage uses the localized image regions to regenerate a sentence that matches with the ground-truth sentence. (Right) Because of the shared Attention LSTM, Language LSTM, and equal word inputs, this training regimen regularizes the attention mechanism inside the Attention LSTM so that the attended ROIs get closer to the less biased and better localized ROIs  $\hat{r}_t \mapsto \hat{r}_t^l$ .

Specifically, our baseline model uses two LSTM modules: Attention LSTM and Language LSTM. The Attention LSTM identifies which visual representation in the image is needed for the Language LSTM to generate the next word. It encodes the global image feature  $v_g$ , previous hidden state output of the Language LSTM  $h_{t-1}^L$ , and the previous word embedding  $e_{t-1}$  into the hidden state  $h_t^A$ .

$$h_t^A = LSTM_{Attn}([v_g; h_{t-1}^L; e_{t-1}]), \quad e_{t-1} = W_e y_{t-1}, \quad (1)$$

where  $[;]$  denotes concatenation, and  $W_e$  are learned parameters. We omit the Attention LSTM input hidden and cell states to avoid notational clutter in the exposition.

The Language LSTM uses the hidden state  $h_t^A$  from the Attention LSTM to dynamically attend on the bag of regions  $R$  for obtaining visual representations of the image  $\hat{r}_t$  to generate a word  $y_t$ .

$$z_{t,n} = W_{aa} \tanh(W_a h_t^A + r_n), \quad \alpha_t = \text{softmax}(z_t), \quad \hat{r}_t = R \alpha, \quad (2)$$

where  $W_{aa}$  and  $W_a$  are learned parameters. The conditional probability distribution over possible output words  $y_t$  is computed as:

$$h_t^L = LSTM_{Lang}([\hat{r}_t, h_t^A]), \quad p(y_t | y_{1:t-1}) = \text{softmax}(W_o h_t^L), \quad (3)$$

where  $y_{1:t-1}$  is a sequence of outputs  $(y_1, \dots, y_{t-1})$ . We refer the Language LSTM and the output logit layer as the complete language decoder.

### 3.2 Overview

Our goal is to enforce the generated caption to be visually grounded, *i.e.*, attended image regions correspond specifically to individual words being generated, *without* ground-truth grounding supervision. Towards this end, we propose a novel cyclical training regimen that is comprised of *decoding*, *localization*, and *reconstruction* stages, as illustrated in Figure 2.

The intuition behind our method is that the baseline network is not forced to generate a correct correspondence between the attended objects and generated words, since the LSTMs can learn priors in the data instead of looking at the image or propagate information forward which can subsequently be used to generate corresponding words in future time steps. The proposed cyclical training regimen, in contrast, aims at enforcing visual grounding to the model by requiring the language decoder (Eq. 3) to rely on the localized image regions  $\hat{\mathbf{r}}_t^l$  to reconstruct the ground-truth sentence, where the localization is conditioned *only* on the generated word from the decoding stage. Our cyclical method can therefore be done without using any annotations of the grounding itself.

Specifically, let  $\mathbf{y}_t^d = \mathcal{D}^d(\hat{\mathbf{r}}_t; \theta_d)$  be the initial language decoder with parameters  $\theta_d$  (Eq. 3), trained to sequentially generate words  $\mathbf{y}_t^d$ . Let  $\hat{\mathbf{r}}_t^l = \mathcal{G}(\mathbf{y}_t^d, \mathbf{R}; \theta_g)$  define a *localizer* unit with parameters  $\theta_g$ , that learns to map (ground) each generated word  $\mathbf{y}_t^d$  to region(s) in the image  $\mathbf{R}$ . Finally, let  $\mathbf{y}_t^l = \mathcal{D}^l(\hat{\mathbf{r}}_t^l; \theta_l)$  be a second decoder, that is required to reconstruct the ground-truth caption using the localized region(s), instead of the attention computed by the decoder itself. We define the cycle:

$$\mathbf{y}_t^l = \mathcal{D}^l(\mathcal{G}(\mathcal{D}^d(\hat{\mathbf{r}}_t; \theta_d), \mathbf{R}; \theta_g); \theta_l), \quad \theta_d = \theta_l, \quad (4)$$

where  $\mathcal{D}^d$  and  $\mathcal{D}^l$  share parameters. Although parameters are shared, the inputs for the two language decoders differ, leading to unique LSTM hidden state values during a run. Note that the Attention LSTMs and logit layers in the two stages also share parameters, though they are omitted for clarity.

Through cyclical joint training, both  $\mathcal{D}^d$  and  $\mathcal{D}^l$  are required to generate the same ground-truth sentence. They are both optimized to maximize the likelihood of the correct caption:

$$\theta^* = \arg \max_{\theta_d} \sum \log p(\mathbf{y}_t^d; \theta_d) + \arg \max_{\theta_l} \sum \log p(\mathbf{y}_t^l; \theta_l), \quad (5)$$

During training, the localizer regularizes the region attention of the reconstructor and the effect is further propagated to the baseline network in the decoding stage, since the parameters of Attention LSTM and Language LSTM are shared for both decoding and reconstruction stages. Note that the gradient from reconstruction loss will not backprop to the decoder  $\mathcal{D}^d$  in the first decoding stage since the generated words used as input to the localizer are leafs in the computational graph. The network is implicitly regularized to update its attention mechanism to match with the localized image regions  $\hat{\mathbf{r}}_t \mapsto \hat{\mathbf{r}}_t^l$ . In Sec. 4.4, we demonstrate that the localized image regions  $\hat{\mathbf{r}}_t^l$  indeed have higher attention accuracy than  $\hat{\mathbf{r}}_t$  when using ground-truth words as inputs for the localizer, which

drives the attention mechanism and helps the attended region  $\hat{\mathbf{r}}_t$  to be more visually grounded.

### 3.3 Cyclical Training

We now describe each stage of our cyclical model, also illustrated in Figure 3.

**Decoding.** We first use the baseline model presented in Sec. 3.1 to generate a sequence of words  $\mathbf{y} = [\mathbf{y}_1^d, \mathbf{y}_2^d, \dots, \mathbf{y}_T^d]$ , where  $T$  is the ground-truth sentence length.

**Localization.** Following the decoding process, a localizer  $\mathcal{G}$  is then learned to localize the image regions from each generated word  $\mathbf{y}_t$ .

$$\mathbf{e}_t = \mathbf{W}_e \mathbf{y}_t^d, \quad z_{t,n}^l = (\mathbf{W}_l \mathbf{e}_t)^\top \mathbf{r}_n \quad \text{and} \quad \beta_t = \text{softmax}(\mathbf{z}_t^l), \quad (6)$$

where  $\mathbf{e}_t$  is the embedding for the word generated during decoding stage at step  $t$ ,  $\mathbf{r}_n$  is the image representation of a region proposal, and  $\mathbf{W}_e$  and  $\mathbf{W}_l$  are the learned parameters. Based on the localized weights  $\beta_t$ , the localized region representation can be obtained by  $\hat{\mathbf{r}}_t^l = \mathbf{R}\beta$ . Our localizer essentially is a linear layer, and we have experimented with non-linear layers but found it performed worse (see Table 9 in the Appendix).

**Reconstruction.** Finally, the shared language decoder  $\mathcal{D}^l$  relies on the localized region representation  $\hat{\mathbf{r}}_t^l$  to generate the next word. The probability over possible output words is:

$$\mathbf{h}_t^L = LSTM_{Lang}([\hat{\mathbf{r}}_t^l; \mathbf{h}_t^A]), \quad p(\mathbf{y}_t^l | \mathbf{y}_{1:t-1}^l) = \text{softmax}(\mathbf{W}_o \mathbf{h}_t^L), \quad (7)$$

Given the target ground truth caption  $\mathbf{y}_{1:T}^*$  and our proposed captioning model parameterized with  $\theta$ , we minimize the following cross-entropy losses:

$$\mathcal{L}_{CE}(\theta) = \underbrace{-\lambda_1 \sum_{t=1}^T \log(p_\theta(\mathbf{y}_t^* | \mathbf{y}_{1:t-1}^*)) \mathbb{1}_{(\mathbf{y}_t^* = \mathbf{y}_t^d)}}_{\text{decoding loss}} - \underbrace{\lambda_2 \sum_{t=1}^T \log(p_\theta(\mathbf{y}_t^* | \mathbf{y}_{1:t-1}^*)) \mathbb{1}_{(\mathbf{y}_t^* = \mathbf{y}_t^l)}}_{\text{reconstruction loss}} \quad (8)$$

where  $\lambda_1$  and  $\lambda_2$  are weighting coefficient selected on the validation split.

Note that before entering the proposed cyclical training regimen, the decoder was first pre-trained until convergence to make sure the generated caption, which used as inputs to the localizer, are reasonable. The cyclical training regimen mainly serves as a regularization method.

## 4 Experiments

**Datasets.** We use the Flickr30k Entities image dataset [25] and the ActivityNet-Entities video dataset [47] to provide a comparison with the state-of-the-art [47]. Flickr30k Entities contains 275k annotated bounding boxes from 31k images associated with natural language phrases. Each image is annotated with 5 crowd-sourced captions. ActivityNet-Entities contains 15k videos with 158k spatially annotated bounding boxes from 52k video segments.

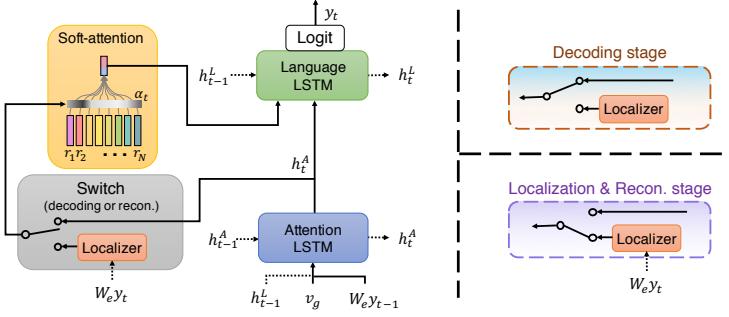


Fig. 3: Proposed model architecture (left) and how the model operates during decoding, localization, and reconstruction stages (right). During the decoding stage, the soft-attention module uses the hidden state of the Attention LSTM to compute attention weights on image regions. During the localization and reconstruction stage, the soft-attention module instead uses the generated word from decoding stage to compute attention weights on image regions.

#### 4.1 Evaluation Metrics

**Captioning evaluation metrics.** We measure captioning performance using common language metrics, *i.e.*, BLEU [23], METEOR [3], CIDEr [36], and SPICE [1].

**Grounding evaluation metrics.** Following the grounding evaluation from GVD [47], we measure the attention accuracy on generated sentences, denoted by  $F1_{all}$  and  $F1_{loc}$ . In  $F1_{all}$ , a region prediction is considered correct if the object word<sup>2</sup> is correctly predicted and also correctly localized. We also compute  $F1_{loc}$ , which only considers correctly-predicted object words. Please see illustration of the grounding metrics in Appendix A.

In the original formulation, the precision and recall for the two F1 metrics are computed **for each object class**, and it is set to zero if an object class has never been predicted. The scores are computed for each object class and averaged over the total number of classes. Such metrics are extremely stringent as captioning models are generally biased toward certain words in the vocabulary, given the long-tailed distribution of words. In fact, both the baseline and proposed method only generate about 45% of the annotated object words within the val set in Flickr30k Entities. The grounding accuracy of the other 55% of the classes are therefore zero, making the averaged grounding accuracy seemingly low.

**Measuring grounding per generated sentence.** Instead of evaluating grounding on each object class (which might be less intuitive), we include a new grounding evaluation metric *per sentence* to directly reflect the grounding measurement of each generated sentence. The metrics are computed against a pool of object words and their ground-truth bounding boxes (GT bbox) collected across five

<sup>2</sup> The object words are words in the sentences that are annotated with corresponding image regions.

Method	Grounding supervision	Captioning Evaluation					Grounding Evaluation		
		B@1	B@4	M	C	S	F1 <sub>all</sub>	F1 <sub>loc</sub>	F1 <sub>all_per_sent</sub>
ATT-FCN [44]		64.7	19.9	18.5	-	-	-	-	-
NBT [21]		69.0	27.1	21.7	57.5	15.6	-	-	-
Up-Down [2]		69.4	27.3	21.7	56.6	16.0	4.14	12.3	-
GVD (w/o SelfAttn) [47]		69.2	26.9	22.1	60.1	16.1	3.97	11.6	-
GVD [47]	✓	69.9	27.3	22.5	62.3	16.5	7.77	22.2	-
Baseline*	✓	69.0	26.8	22.4	61.1	16.8	8.44 (+100%)	22.78 (+100%)	27.37 (+100%)
Baseline*		69.1	26.0	22.1	59.6	16.3	4.08 (+0%)	11.83 (+0%)	13.20 (+0%)
Cyclical*		<b>69.9</b>	<b>27.4</b>	<b>22.3</b>	<b>61.4</b>	<b>16.6</b>	<b>4.98</b> (+21%)	<b>13.53</b> (+16%)	<b>15.03</b> (+13%)
								<b>35.54</b> (+12%)	

Table 1: Performance comparison on the Flickr30k Entities **test set**. \*: our results are averaged **across five runs**. Only numbers reported by multiple runs are considered to be bolded<sup>3</sup>.

GT captions on Flickr30k Entities (and one GT caption on ActivityNet-Entities). We use the same Precall, Recall, Prec<sub>loc</sub>, and Rec<sub>loc</sub> as defined previously, but their scores are averaged on each of the generated sentence. As a result, the F1<sub>loc\_per\_sent</sub> measures the F1 score only on the generated words. The model will not be punished if some object words are not generated, but it also needs to maintain diversity to achieve high captioning performance.

#### 4.2 Implementation and Training Details

**Region proposal and spatial features.** Following GVD [47], we extracted 100 region proposals from each image (video frame) and encode them via the *grounding-aware region encoding*.

**Training.** We train the model with ADAM optimizer [16]. The initial learning rate is set to  $1e - 4$ . Learning rates automatically drop by 10x when the CIDEr score is saturated. The batch size is 32 for Flickr30k Entities and 96 for ActivityNet-Entities. We learn the word embedding layer from scratch for fair comparisons with existing work [47]. Please see the Appendix D for additional training and implementation details.

#### 4.3 Captioning and Grounding Performance Comparison

**Flickr30k Entities.** We first compare the proposed method with our baseline with or without grounding supervision on the Flickr30k Entities test set (see Table 1). To train the supervised baseline, we train the attention mechanism as well as add the region classification task using the ground-truth grounding annotation, similar to GVD [47]. We train the proposed baselines and our method on the training set and choose the best performing checkpoints based on their CIDEr score on the val set. Unlike previous work, our experimental results are reported by averaging across five runs on the test set. We report only the mean of the five runs to keep the table uncluttered.

<sup>3</sup> Note that the since supervised methods are used as upper bound, their numbers are not bolded.

Method	Grounding supervision	Captioning Evaluation						Grounding Evaluation		
		B@1	B@4	M	C	S	F1 <sub>all</sub>	F1 <sub>loc</sub>	F1 <sub>all_per_sent</sub>	F1 <sub>loc_per_sent</sub>
GVD [47]		23.0	2.27	10.7	44.6	13.8	0.28	1.13	-	-
GVD (w/o SelfAttn) [47]		23.2	2.28	10.9	45.6	15.0	3.70	12.7	-	-
GVD [47]	✓	23.9	2.59	11.2	47.5	15.1	7.11	24.1	-	-
Baseline*	✓	23.1	2.13	10.7	45.0	14.6	7.30 (+100%)	25.02 (+100%)	17.88 (+100%)	60.23 (+100%)
Baseline*		23.2	2.22	10.8	45.9	<b>15.1</b>	3.75 (+0%)	12.00 (+0%)	9.41 (+0%)	31.68 (+0%)
Cyclical*		<b>23.7</b>	<b>2.45</b>	<b>11.1</b>	<b>46.4</b>	14.8	<b>4.71</b> (+26%)	<b>15.84</b> (+29%)	<b>11.73</b> (+38%)	<b>41.56</b> (+43%)

Table 2: Performance comparison on the ActivityNet-Entities **val set**. \*: our results are averaged **across five runs**. Only numbers reported by multiple runs are considered to be bolded.

When compared to the existing state of the arts, our proposed baselines achieve comparable captioning evaluation performances and better grounding accuracy. Using the resulting supervised baseline as the upper bound, our proposed method with cyclical training statistically achieves around 15 to 20% relative grounding accuracy improvements for both  $F1_{all}$  and  $F1_{loc}$  and 10 to 15% for  $F1_{all\_per\_sent}$  and  $F1_{loc\_per\_sent}$  without utilizing any grounding annotations or additional computation during inference.

**ActivityNet-Entities.** We adapt our proposed baselines and method to the ActivityNet-Entities video dataset (see Table 2 for results on the validation set and Table 12 in the Appendix for results on the test set). We can see that our baseline again achieved comparable performance to the state of the arts. The proposed method then significantly improved the grounding accuracy around 25% to 30% relative grounding accuracy improvements for both  $F1_{all}$  and  $F1_{loc}$  and around 40% for  $F1_{all\_per\_sent}$  and  $F1_{loc\_per\_sent}$ . Interestingly, we observed that baseline with grounding supervision does not improve the captioning accuracy, different from the observation in previous work [47].

#### 4.4 Quantitative Analysis

**Are localized image regions better than attended image regions *during training*?** Given our intuition described in Sec. 3, we expect the decoder to be regularized to update its attention mechanism to match with the localized image regions  $\hat{r}_t \mapsto \hat{r}_t^l$  during training. This indicates that the localized image regions should be more accurate than the attended image regions by the decoder and drives the update on attention mechanism. To verify this, we compute the attention accuracy for both decoder and localizer over ground-truth sentences following [28,48]. The attention accuracy for localizer is 20.4% and is higher than the 19.3% from the decoder at the end of training, which confirms our hypothesis on how cyclical training helps the captioning model to be more grounded.

**Grounding performance when using a *better* object detector.** In Table 1 and 2 we showed that our proposed method significantly improved the grounding accuracy for both image and video captioning. These experimental settings follow the widely used procedure for visual captioning systems: extract regional proposal features and generate visual captions by attending to those extracted

#	supervision	Grounding		Captioning		Eval.	Grounding	Eval.
		M	C	S	F1 <sub>all</sub>	F1 <sub>loc</sub>	F1 <sub>loc_per_sent</sub>	
<b>Unrealistically perfect object detector</b>								
Baseline	✓	25.3	76.5	22.3	23.19	52.83	90.76	
Baseline		25.2	76.3	22.0	20.82	48.74	77.81	
Cyclical		<b>25.8</b>	<b>80.2</b>	<b>22.7</b>	<b>25.27</b>	<b>54.54</b>	<b>81.56</b>	
<b>Grounding-biased object detector</b>								
Baseline	✓	21.3	53.3	15.5	8.23	23.95	66.96	
Baseline		<b>21.2</b>	<b>52.4</b>	<b>15.4</b>	<b>5.95</b>	<b>17.51</b>	<b>42.84</b>	
Cyclical		<b>21.2</b>	<b>52.0</b>	<b>15.4</b>	<b>6.87</b>	<b>19.65</b>	<b>50.25</b>	

Table 3: Comparison when using *better* object detector on Flickr30k Entities **test** set (see Table 11 for complete version).

visual features. One might ask, what if we have a better object detector that can extract robust visual representation that are better aligned with the word embeddings? Will visual grounding still an issue for captioning?

To answer this, we ran two sets of experiments (Table 3): (1) **Perfect object detector**: we replace the ROIs by ground-truth bbox and represent the new ROIs by learning embedding features directly from ground-truth object words associated with each ground-truth bbox. This experiment gives an estimate of the captioning and grounding performance if we have (almost) perfect ROI representations (though unrealistic). We can see that the fully-supervised method achieves an F1<sub>all</sub> of only 23%, which further confirms the difficulty of the metric and the necessity of our grounding metric on a per sentence level (note that F1<sub>loc\_per\_sent</sub> shows 90%). We can also see that baseline (unsup.) still leaves room for improvement on grounding performance. Surprisingly, our method improved both captioning and grounding accuracy and surpasses the fully-supervised baseline except on the F1<sub>loc\_per\_sent</sub>. We find that it is because the baseline (sup.) overfits to the training set, while ours is regularized from the cyclical training. Also, our generated object words are more diverse, which is important for F1<sub>all</sub> and F1<sub>loc</sub>. (2) **Grounding-biased object detector**: we extract ROI features from an object detector pre-trained on Flickr30k. Thus, the ROI features and their associated object predictions are biased toward the annotated object words but do not generalize to predict diverse captions compared to the original object detector trained from Visual Genome, resulting in lower captioning performance. We can see that our proposed method still successfully improves grounding and maintains captioning performance in this experiment setting as well.

**How does the number of annotations affect grounding performance?** In Figure 4, we present the average F1-score on the Flickr30k Entities val set when grouping classes according to their frequency of appearance in the training set<sup>4</sup>. We see that, unsurprisingly, the largest difference in grounding accuracy between the supervised and our proposed cyclical training is for the 50 most frequently appearing object classes, where enough training data exists. As the number of

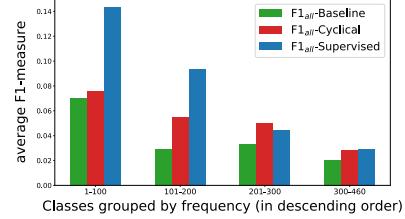


Fig. 4: Average F1<sub>all</sub>-score per class as a function of class frequency.

<sup>4</sup> We group the 460 object classes in 10 groups, sorted by the number of annotated bounding boxes.

#	Captioning Eval.				Grounding Eval. F1 <sub>loc</sub>
	M	C	S	F1 <sub>all</sub>	
Baseline (Unsup.)	<b>22.3</b>	62.1	16.0	4.18	11.9
Cyclical	22.2	<b>62.2</b>	<b>16.2</b>	<b>5.63</b>	<b>14.6</b>
- Attention consistency	<b>22.3</b>	61.8	<b>16.2</b>	4.19	11.3
- Localizer using $\mathbf{h}^A$	22.2	61.8	16.1	4.58	11.3

Table 4: Model ablation study on the Flickr30k Entities val set.

Method	Human Grounding Eval.	
	%	
About equal	47.1	
Cyclical is better	28.1	
Baseline is better	24.8	

Table 5: Human evaluation on grounding on the Flickr30k Entities val set.

annotated boxes decreases, however, the difference in performance diminishes, and cyclical training appears to be more robust. Overall, we see that the supervised method is biased towards frequently appearing objects, while grounding performance for the proposed approach is more balanced among classes.

**Should we explicitly make attended image regions to be similar to localized image regions?** One possible way to regularize the attention mechanism of the decoder is to explicitly optimize  $\hat{\mathbf{r}}_t \mapsto \hat{\mathbf{r}}_t^l$  via KL divergence over two soft-attention weights  $\alpha_t$  and  $\beta_t$ . The experimental results are shown in Table 4 (*Attention consistency*). We use a single run unsupervised baseline with a fix random seed as baseline model for ablation study. We can see that when explicitly forcing the attended regions to be similar to the localized regions, both the captioning performance and the grounding accuracy remain similar to the baseline (unsup.). We conjecture that this is due to the noisy localized regions at the initial training stage. When forcing the attended regions to be similar to noisy localized regions, the Language LSTM will eventually learn to not rely on the attended region at each step for generating sequence of words. To verify, we increase the weight for attention consistency loss and observed that it has lower grounding accuracy ( $F1_{all} = 3.2$ ), but the captioning will reach similar performance while taking 1.5x longer to reach convergence.

**Is using only the generated word for localization necessary?** Our proposed localizer (Eq. 6 and Figure 3) relies on purely the word embedding representation to locate the image regions. This forces the localizer to rely only on the word embedding without biasing it with the memorized information from the Attention LSTM. As shown in the Table 4 (localizer using  $\mathbf{h}^A$ ), although this achieves comparable captioning performance, it has lower grounding accuracy improvement compared to our proposed method.

#### 4.5 Human Evaluation on Grounding

We conduct a human evaluation on the perceptual quality of the grounding. We asked 10 human subjects (not familiar with the proposed method) to pick the best among two grounded regions (by baseline and Cyclical) for each word. The subjects have three options to choose from: 1) grounded region A is better, 2) grounded region B is better, and 3) they are about the same (see Figure 8 for illustration). Each of the human subjects were given 25 images, each with a varying number of groundable words. Each image was presented to two different human subjects in order to be able to measure inter-rater agreement. For this

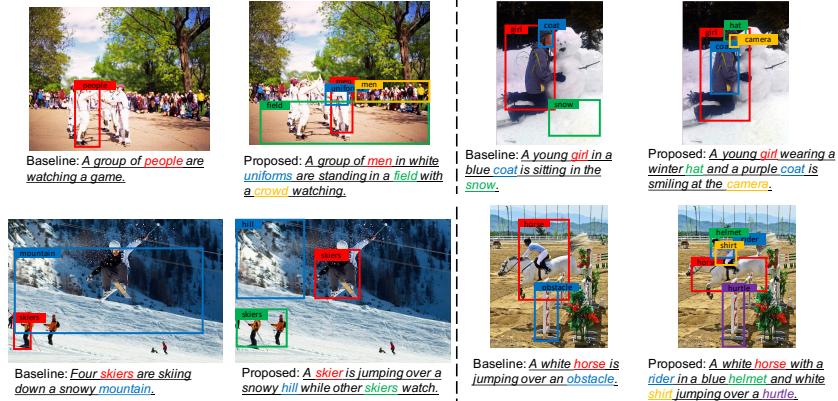


Fig. 5: Generated captions and corresponding visual grounding regions with comparison between baseline (left) and proposed approach (right). Our proposed method is able to generate more descriptive sentences while selecting the correct regions for generating the corresponding words.

study, we define a word to be groundable if it is either a noun or verb. The image regions are selected based on the region with the maximum attention weight in  $\alpha_t$  for each word. The order of approaches was randomized for each sentence.

Our experiment on the Flickr30k Entities val set is shown in Table 5: 28.1% of words are more grounded by Cyclical, 24.8% of words are more grounded by baseline, and 47.1% of words are similarly grounded. We also measured inter-rater agreement between each pair of human subjects: 72.7% of ratings are the same, 4.9% of ratings are the opposite, and 22.4% of ratings could be ambiguous (*e.g.*, one chose A is better, the other chose they are about the same).

We would also like to note that the grounded words judged to be similar largely consisted of very easy or impossible cases. For example, words like *mountain*, *water*, *street*, *etc*, are typically rated to be “about the same” since they usually have many possible boxes and is very easy for both models to ground the words correctly. On the other hand, for visually ungroundable cases, *e.g.*, *stand* appears a lot and the subject would choose *about the same* since the image does not cover the fact that the person’s feet are on the ground. We see that the human study results follow the grounding results presented in the paper and show an improvement in grounding accuracy for the proposed method over a strong baseline. The improvement is achieved without grounding annotations or extra computation at test time.

#### 4.6 Qualitative Analysis

We additionally present some qualitative results comparing the baseline (Unsup.) and the proposed method in Figure 5. To avoid clutter, only the visually-relevant object words in the generated sentences are highlighted with colors. Each highlighted word has a corresponding image region annotated on the original image.

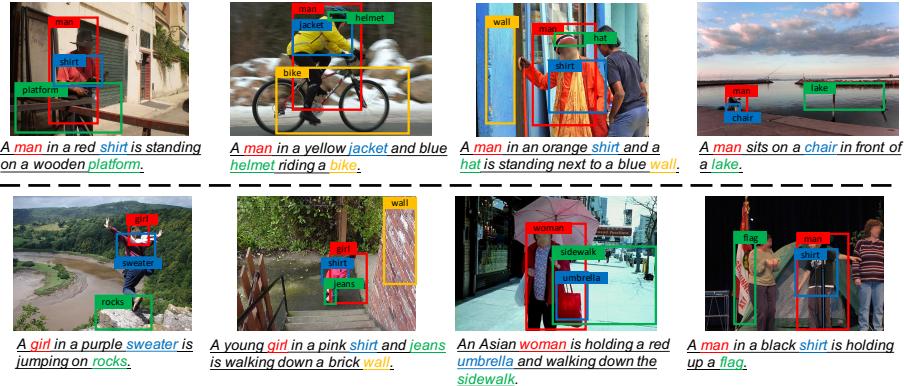


Fig. 6: Examples with correct grounding (top) as well as failure cases (bottom).

The image regions are selected based on the region with the maximum attention weight in  $\alpha_t$ . We can see that our proposed method significantly outperforms the baseline (Unsup.) in terms of both the quality of the generated sentence and grounding accuracy. In Figure 6, we show a number of correct and incorrect examples of our proposed method. We observe that while the model is able to generate grounded captions for the images, it may sometimes overlook the semantic meaning of the generated sentences (*e.g.*, “A young girl [...] walking down a brick wall”), or the spatial relationship between the objects (“A man [...] is holding up a flag”).

## 5 Conclusion

Working from the intuition that typical attentional mechanisms in the visual captioning task are not forced to ground generated words since recurrent models can propagate past information, we devise a novel cyclical training regime to explicitly force the model to ground each word without grounding annotations. Our method only adds a fully-connected layer during training, which can be removed during inference, and we show thorough quantitative and qualitative results demonstrating around 20% or 30% relative improvements in visual grounding accuracy over existing methods for image and video captioning tasks.

## Acknowledgments

Chih-Yao Ma and Zsolt Kira were partly supported by DARPA’s Lifelong Learning Machines (L2M) program, under Cooperative Agreement HR0011-18-2-0019, as part of their affiliation with Georgia Tech. We thank Chia-Jung Hsu for her valuable and artistic help on the figures.

## Appendix

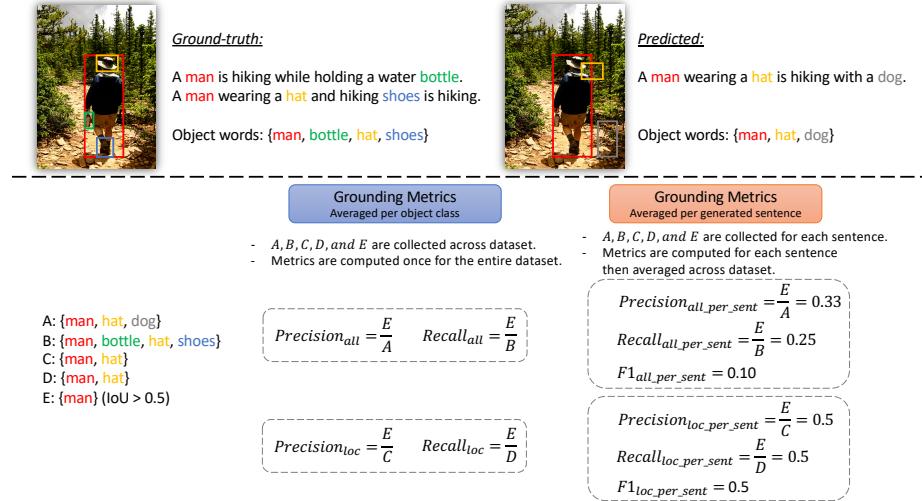


Fig. 7: Illustration of Grounding metrics.

## A Grounding Evaluation Metrics illustrated

To help better understand the grounding evaluation metrics used in this work, we illustrated the grounding evaluation metrics in Figure 7.

We define the number of object words in the generated sentences as A, the number of object words in the GT sentences as B, the number of correctly predicted object words in the generated sentences as C and the counterpart in the GT sentences as D, and the number of correctly predicted and localized words as E. A region prediction is considered correct if the object word is correctly predicted and also correctly localized (*i.e.*, IoU with GT box  $> 0.5$ ). We then compute two version of the precision and recall as  $Prec_{all} = \frac{E}{A}$ ,  $Rec_{all} = \frac{E}{B}$ ,  $Prec_{loc} = \frac{E}{C}$ , and  $Rec_{loc} = \frac{E}{D}$ .

The original grounding evaluation metric proposed in GVD [47] average the grounding for each object class. We additionally calculate the grounding accuracy for each generated sentence as demonstrated in the figure. From this example, we can see that while  $Precision_{all}$  counts *dog* as a wrong prediction for the *dog* object class, the  $Precision_{loc}$  only cares if *man* and *hat* are predicted and correctly localizer (IoU  $> 0.5$ ).

Method	Captioning Evaluation					Grounding Evaluation			
	B@1	B@4	M	C	S	F1 <sub>all</sub>	F1 <sub>loc</sub>	F1 <sub>all_per_sent</sub>	F1 <sub>loc_per_sent</sub>
Baseline	69.1	26.0	22.1	59.6	16.3	4.08	11.83	13.20	31.83
Cyclical	69.4	27.4	22.3	61.4	<b>16.6</b>	4.98	13.53	15.03	35.54
Cyclical (zero-loss)	69.7	27.0	22.2	60.1	16.5	<b>5.14</b>	<b>14.32</b>	15.36	36.33
Cyclical (zero-representation)	<b>69.9</b>	<b>27.5</b>	<b>22.4</b>	<b>62.0</b>	<b>16.6</b>	5.13	13.99	<b>16.30</b>	<b>38.45</b>

Table 6: Performance comparison on the Flickr30k Entities **test set**. All results are averaged **across five runs**.

Method	Captioning Evaluation					Grounding Evaluation			
	B@1	B@4	M	C	S	F1 <sub>all</sub>	F1 <sub>loc</sub>	F1 <sub>all_per_sent</sub>	F1 <sub>loc_per_sent</sub>
Baseline	23.2	2.22	10.8	45.9	<b>15.1</b>	3.75	12.00	9.41	31.68
Cyclical	23.7	2.45	11.1	46.4	14.8	<b>4.68</b>	<b>15.84</b>	<b>12.60</b>	<b>44.04</b>
Cyclical (zero-representation)	<b>23.9</b>	<b>2.58</b>	<b>11.2</b>	<b>46.6</b>	14.8	4.48	15.01	11.53	40.30

Table 7: Performance comparison on the ActivityNet-Entities **val set**. All results are averaged **across five runs**.

## B Additional Quantitative Analysis

**Can words that are not visually-groundable be handled differently?** In the proposed method, all the words are handled the same regardless of whether they are visually-groundable or not, *i.e.*, the localizer is required to use *all* generated words at each step in a sentence to localize regions in the image. Yet, typically words that are nouns or verbs are more likely to be grounded, and words like "a", "the", *etc*, are not visually-groundable.

We explored a few method variants to handle nouns and verbs differently. Mainly, we explored with two variants.

- **Cyclical (zero-loss):** the reconstruction loss is only computed when the target word is either a noun or a verb.
- **Cyclical (zero-representation):** the localized region representation will be invalid (set to zero) if the target word is neither nouns nor verbs.

The experimental results are shown in Table 6, 7, and 8. For the first variant, Cyclical (zero-loss), we observed that the captioning performance stays the same while grounding accuracy has a small improvement. On the other hand, for the second variant, Cyclical (zero-representation), we can see that all captioning scores are improved over baseline with CIDEr improved 2.4 (see Table 6). We can also see that grounding accuracy on per sentence basis further improved as well. We then conducted further experiments on both ActivityNet-Entities and Flickr30k Entities with *unrealistically perfect object detector* (see Table 7 and 8), but the improvements however are not consistent. In summary: on the Flickr30k Entities test set, we observed that CIDEr is better and grounding per sentence better, on the ActivityNet-Entities val set, the captioning performances are about the same but grounding accuracy became worse, and on the Flickr30k

Method	Captioning Evaluation						Grounding Evaluation		
	B@1	B@4	M	C	S	F1 <sub>all</sub>	F1 <sub>loc</sub>	F1 <sub>all_per_sent</sub>	F1 <sub>loc_per_sent</sub>
<b>Unrealistically perfect object detector</b>									
Baseline	75.1	32.1	25.2	76.3	22.0	20.82	48.74	43.21	77.81
Cyclical	<b>76.7</b>	<b>32.8</b>	<b>25.8</b>	<b>80.2</b>	<b>22.7</b>	25.27	54.54	46.98	81.56
Cyclical (zero-representation)	75.8	32.2	25.6	79.0	22.4	<b>25.65</b>	<b>55.81</b>	<b>48.99</b>	<b>85.99</b>

Table 8: Grounding performance when using better object detector on the Flickr30k Entities **test** set (results are averaged three runs).

Method	Captioning Evaluation						Grounding Evaluation		
	B@1	B@4	M	C	S	F1 <sub>all</sub>	F1 <sub>loc</sub>	F1 <sub>all_per_sent</sub>	F1 <sub>loc_per_sent</sub>
Cyclical	<b>69.4</b>	<b>27.4</b>	<b>22.3</b>	<b>61.4</b>	<b>16.6</b>	<b>4.98</b>	<b>13.53</b>	<b>15.03</b>	<b>35.54</b>
Cyclical (MLP Localizer)	69.2	26.4	22.0	58.7	16.2	4.40	12.77	13.97	33.40

Table 9: Performance comparison on the Flickr30k Entities **test set** using FC or MLP as the localizer. All results are averaged **across five runs**.

Entities test set with unrealistically perfect object detector, captioning performances are slightly worse but grounding accuracy improved. We thus keep the most general variant "Cyclical" which treats all words equally.

**Will a non-linear localizer performs better?** In practice, our localizer is a single fully-connected layer. It is possible to replace it with a non-linear layer, *e.g.*, multi-layer perceptron (MLP). We however observed that both captioning and grounding accuracy reduced if a MLP is used as the localizer (see Table 9).

**Weighting between decoding and reconstruction losses.** The weighting between the two losses was chosen with a grid search on the val set. We report the experimental results on Flickr30k Entities val set in Table 10. We can see that when comparing to the baseline, all different loss weightings consistently improved both captioning and grounding accuracy. Unless further specified, we use default (0.5, 0.5) weighting for the two losses, except (0.6, 0.4) for the final result on Flickr30k Entities test set in Table 1.

## C Additional Qualitative Results

In Figure 9, 10, 11, 12, 13, 14, 15, and 16, we illustrated the sequence of attended image region when generating each word for a complete image description. At each step, only the top-1 attended image region is shown. This is the same as how the grounding accuracy is measured. Please see the description for Figure 9 - 16 for further discussions on the qualitative results.

## D Additional Implementation Details

**Region proposal features.** We use a Faster-RCNN model [26] pre-trained on Visual Genome [17] for region proposal and feature extraction. In practice,

$(\lambda_1, \lambda_2)$	Captioning Evaluation						Grounding Evaluation			
	B@1	B@4	M	C	S	F1 <sub>all</sub>	F1 <sub>loc</sub>	F1 <sub>all_per_sent</sub>	F1 <sub>loc_per_sent</sub>	
baseline	69.7	26.7	22.3	61.1	16.1	4.61	13.11	12.41	30.61	
(0.8, 0.2)	70.3	27.9	22.4	62.2	16.5	4.96	13.95	13.95	33.49	
(0.6, 0.4)	70.4	28.0	22.4	62.7	16.3	5.04	13.92	14.46	34.95	
(0.5, 0.5)	70.2	27.9	22.5	62.3	16.5	4.93	13.70	14.28	34.62	
(0.4, 0.6)	69.8	27.6	22.5	62.3	16.4	4.97	13.67	14.97	36.31	
(0.2, 0.8)	69.5	27.7	22.3	61.4	16.1	5.07	14.05	15.41	37.63	

Table 10: Performance comparison on the Flickr30k Entities **val** set with different weightings on decoding and reconstruction losses. All results are averaged across five runs.

Method	Captioning Evaluation						Grounding Evaluation			
	Grounding supervision	B@1	B@4	M	C	S	F1 <sub>all</sub>	F1 <sub>loc</sub>	F1 <sub>all_per_sent</sub>	F1 <sub>loc_per_sent</sub>
<b>Unrealistically perfect object detector</b>										
Baseline	✓	75.6	32.0	25.3	75.6	22.3	23.19 (+100%)	52.83 (+100%)	51.43 (+100%)	90.76 (+100%)
Baseline		75.1	32.1	25.2	76.3	22.0	20.82 (+0%)	48.74 (+0%)	43.21 (+0%)	77.81 (+0%)
Cyclical		<b>76.7</b>	<b>32.8</b>	<b>25.8</b>	<b>80.2</b>	<b>22.7</b>	<b>25.27</b> (+188%)	<b>54.54</b> (+142%)	<b>46.98</b> (+46%)	<b>81.56</b> (+29%)
<b>Grounding-biased object detector</b>										
Baseline	✓	65.9	23.4	21.3	53.3	15.5	8.23 (+100%)	23.95 (+100%)	28.06 (+100%)	66.96 (+100%)
Baseline		<b>66.1</b>	<b>23.5</b>	<b>21.2</b>	<b>52.4</b>	<b>15.4</b>	5.95 (+0%)	17.51 (+0%)	18.11 (+0%)	42.84 (+0%)
Cyclical		65.5	23.3	<b>21.2</b>	52.0	15.4	<b>6.87</b> (+40%)	<b>19.65</b> (+33%)	<b>20.82</b> (+27%)	<b>50.25</b> (+31%)

Table 11: Grounding performance when using better object detector on the Flickr30k Entities **test** set (results are averaged three runs). Fully-supervised method is used as upper bound, thus its numbers are not bolded.

besides the region proposal features, we also use the Conv features (*conv4*) extracted from an ImageNet pre-trained ResNet-101. Following GVD [47], the region proposals are represented using the *grounding-aware region encoding*, which is the concatenation of i) region feature, ii) region-class similarity matrix, and iii) location embedding.

For region-class similarity matrix, we define a set of object classifiers as  $\mathbf{W}_c$ , and the region-class similarity matrix can be computed as  $M_s = \text{softmax}(\mathbf{W}_c^\top \mathbf{R})$ , which captures the similarity between regions and object classes. We omit the ReLU and Dropout layer after the linear embedding layer for clarity. We initialize  $\mathbf{W}_c$  using the weight from the last linear layer of an object classifiers pre-trained on the Visual Genome dataset [17].

For location embedding, we use 4 values for the normalized spatial location. The 4-D feature is then projected to a  $d_s = 300$ -D location embedding for all the regions.

**Software and hardware configuration.** Our code is implemented in PyTorch. All experiments were ran on the 1080Ti, 2080Ti, and Titan Xp GPUs.

**Network architecture.** The embedding dimension for encoding the sentences is 512. We use a dropout layer with ratio 0.5 after the embedding layer. The hidden

Method	Grounding supervision	Captioning Evaluation					Grounding Evaluation	
		B@1	B@4	M	C	S	F1 <sub>all</sub>	F1 <sub>loc</sub>
Masked Transformer [50]		22.9	2.41	10.6	46.1	13.7	-	-
Bi-LSTM+TempoAttn [50]		22.8	2.17	10.2	42.2	11.8	-	-
GVD (w/o SelfAttn) [47]		23.1	2.16	<b>10.8</b>	44.9	14.9	3.73	11.7
GVD [47]	✓	23.6	2.35	11.0	45.5	14.7	7.59	25.0
Baseline	✓	23.1	2.28	10.8	45.6	14.7	7.66 (+100%)	25.7 (+100%)
Baseline		23.2	2.17	<b>10.8</b>	46.2	<b>15.0</b>	3.60 (+0%)	12.3 (+0%)
Cyclical		<b>23.4</b>	<b>2.43</b>	<b>10.8</b>	<b>46.6</b>	14.3	4.70 (+27%)	15.6 (+29%)

Table 12: Performance comparison on the ActivityNet-Entities **test set**. Grounding evaluation metrics on per generated sentences are not available on the test server.

Method	Captioning Evaluation						Grounding Evaluation			
	B@1	B@4	M	C	S	F1 <sub>all</sub>	F1 <sub>loc</sub>	F1 <sub>all_per_sent</sub>	F1 <sub>loc_per_sent</sub>	
Baseline	69.1±0.6	26.0±0.6	22.1±0.3	59.6±0.6	16.3±0.2	4.08±0.40	11.83±1.27	13.20±0.60	31.83±1.36	
Cyclical	<b>69.4±0.4</b>	<b>27.4±0.1</b>	<b>22.3±0.2</b>	<b>61.4±0.8</b>	<b>16.6±0.2</b>	<b>4.98±0.48</b>	<b>13.53±0.84</b>	<b>15.03±0.81</b>	<b>35.54±2.10</b>	

Table 13: Mean and standard deviation on the Flickr30k Entities **test set**. All results are averaged **across five runs**.

state size of the Attention and Language LSTM are 1024. The dimension of other learnable matrices are:  $\mathbf{W}_e \in \mathbb{R}^{d_v \times 512}$ ,  $\mathbf{W}_a \in \mathbb{R}^{1024 \times 512}$ ,  $\mathbf{W}_{aa} \in \mathbb{R}^{512 \times 1}$ ,  $\mathbf{W}_o \in \mathbb{R}^{1024 \times d_v}$ ,  $\mathbf{W}_l \in \mathbb{R}^{512 \times 512}$ , where the vocabulary size  $d_v$  is 8639 for Flickr30k Entities and 4905 for ActivityNet-Entities.

**Training details.** We train the model with ADAM optimizer [16]. The initial learning rate is set to  $1e-4$ . Learning rates automatically drop by 10x when the CIDEr score is saturated. The batch size is 32 for Flickr30k Entities and 96 for ActivityNet-Entities. We learn the word embedding layer from scratch for fair comparisons with existing work [47]. The hyper-parameters  $\lambda_1$  and  $\lambda_2$  are set to 0.5 after hyper-parameter search between 0 and 1.

**Flickr30k Entities.** Images are randomly cropped to  $512 \times 512$  during training, and resized to  $512 \times 512$  during inference. Before entering the proposed cyclical training regimen, the decoder was pre-trained for about 35 epochs. The total training epoch with the cyclical training regimen is around 80 epochs. The total training time takes about 1 day.

**ActivityNet-Entities.** Before entering the proposed cyclical training regimen, the decoder was pre-trained for about 50 epochs. The total training epoch with the cyclical training regimen is around 75 epochs. The total training time takes about 1 day.

Method	Captioning Evaluation						Grounding Evaluation			
	B@1	B@4	M	C	S	F1 <sub>all</sub>	F1 <sub>loc</sub>	F1 <sub>all_per_sent</sub>	F1 <sub>loc_per_sent</sub>	
Baseline	23.2±0.5	2.22±0.2	10.8±0.3	45.9±1.5	<b>15.1±0.2</b>	3.75±0.16	12.00±0.76	9.41±0.26	31.68±0.93	
Cyclical	<b>23.7±0.13</b>	<b>2.45±0.1</b>	<b>11.1±0.1</b>	<b>46.4±0.6</b>	14.8±0.2	<b>4.71±0.41</b>	<b>15.84±1.56</b>	<b>11.73±0.22</b>	<b>41.56±0.75</b>	

Table 14: Mean and standard deviation on the ActivityNet-Entities **val** set. All results are averaged **across five runs**.

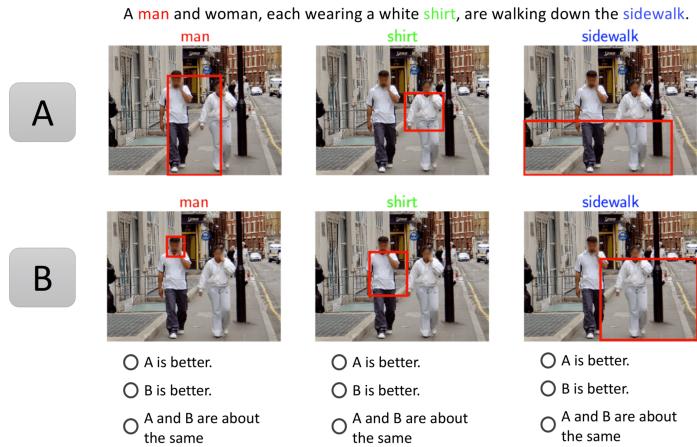


Fig. 8: Demonstration of our human evaluation study on grounding. Each human subject is required to rate which method (A or B) has a better grounding on each highlighted word.



Fig. 9: *A group of men in white uniforms are standing in a field with a crowd watching.* We can see that our proposed method attends to the sensible image regions for generating visually-groundable words, e.g., *man*, *uniforms*, *field*, and *crowd*. Interestingly, when generating *standing*, the model pays its attention on the image region with a foot on the ground.

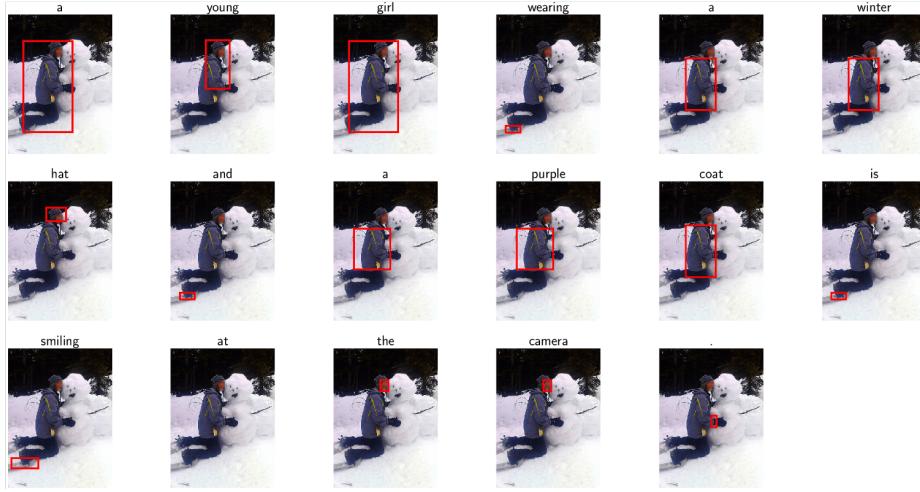


Fig. 10: *A young girl wearing a winter hat and a purple coat is smiling at the camera.* The proposed method is able to select the corresponding image regions to generate *girl*, *hat*, and *coat* correctly. We have also observed that the model tends to localize the person's face when generating *camera*.

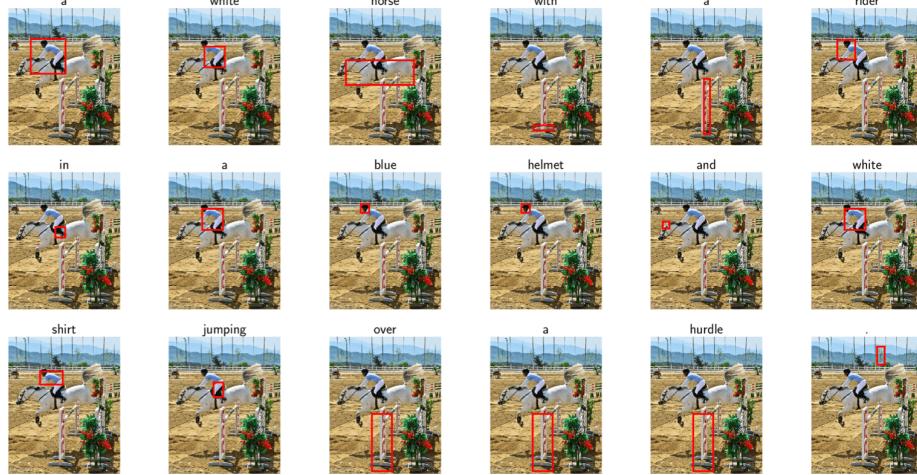


Fig. 11: *A white horse with a rider in a blue helmet and white shirt jumping over a hurdle.* While the model is able to correctly locate objects such as *horse*, *rider*, *helmet*, *shirt*, and *hurdle*, it mistakenly describes the rider as wearing a blue helmet, while it's actually black, and with white shirt while it's blue.

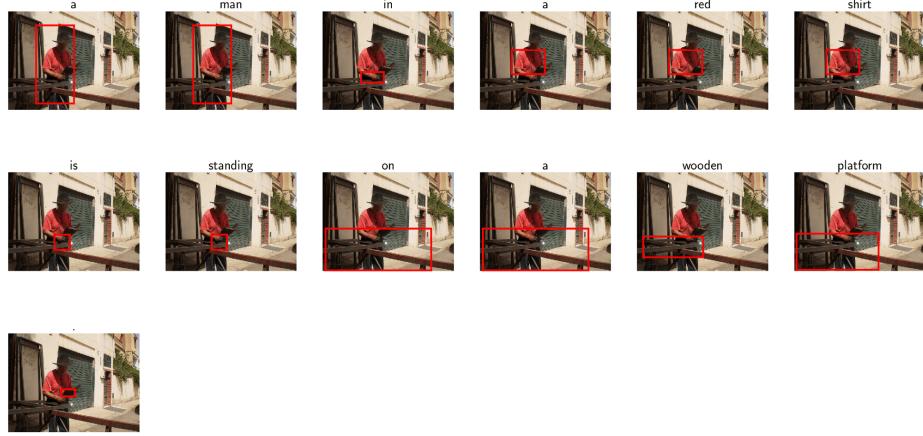


Fig. 12: *A man in a red shirt is standing on a wooden platform.* Our method correctly attends on the correct regions for generating *man*, *shirt*, and *platform*.



Fig. 13: *A man in a yellow jacket and blue helmet riding a bike.* The proposed method correctly generates a descriptive sentence while precisely attending to the image regions for each visually-groundable words: *man*, *jacket*, *helmet*, and *bike*.

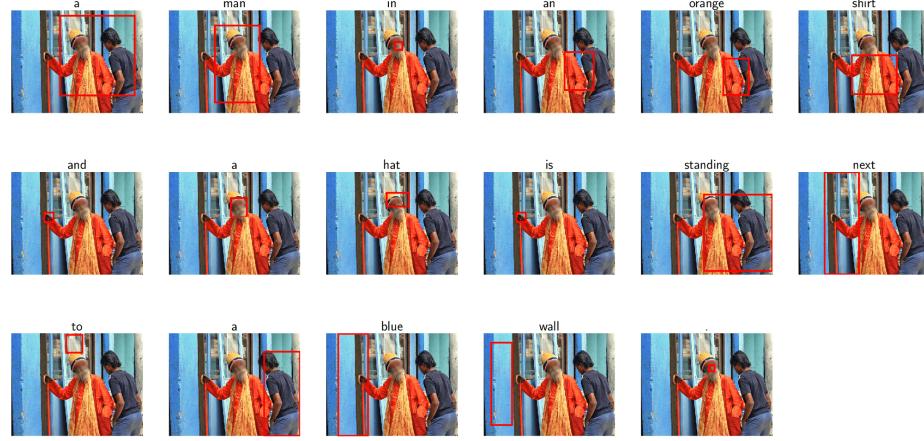


Fig. 14: *A man in an orange shirt and a hat is standing next to a blue wall.* While our method is able to ground the generated sentence on the objects like: *man, shirt, hat, and wall*, it completely ignores the person standing next to the man in the orange cloth.

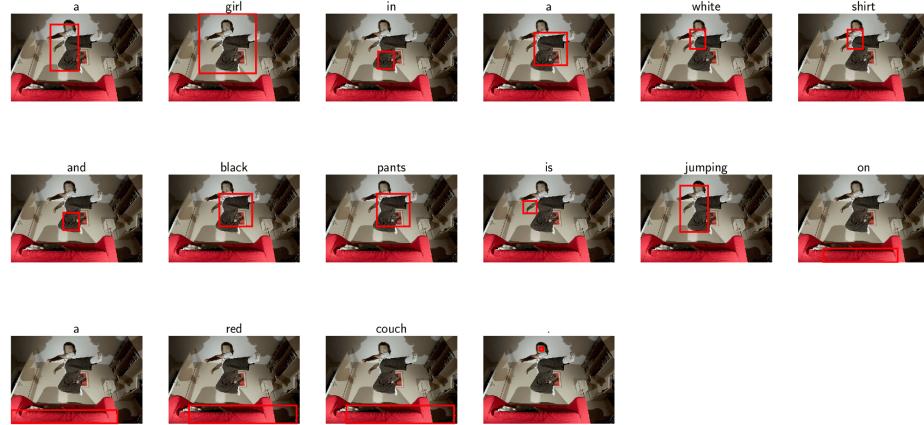


Fig. 15: *A girl in a white shirt and black pants is jumping on a red couch.* Our method is able to ground the generated descriptive sentence with the correct grounding on: *girl, shirt, pants, and couch*.



Fig. 16: *A man in a blue robe walks down a cobblestone street.* Our method grounds the visually-relevant words like: *man*, *robe*, and *street*. We can also see that it is able to locate the foot on ground for *walks*.

## References

- Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: European Conference on Computer Vision. pp. 382–398. Springer (2016) 8
- Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. 6 (2018) 2, 3, 4, 9
- Banerjee, S., Lavie, A.: 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. vol. 29, pp. 65–72 (2005) 8
- Chen, X., Lawrence Zitnick, C.: Mind’s eye: A recurrent visual representation for image caption generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 2422–2431 (2015) 4
- Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D.: Human attention in visual question answering: Do humans and deep networks look at the same regions? Computer Vision and Image Understanding **163**, 90–100 (2017) 4
- Das, P., Xu, C., Doell, R.F., Corso, J.J.: A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2634–2641 (2013) 3
- Deng, C., Wu, Q., Wu, Q., Hu, F., Lyu, F., Tan, M.: Visual grounding via accumulated attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7746–7755 (2018) 4
- Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual

- recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2625–2634 (2015) 3
9. Duan, X., Huang, W., Gan, C., Wang, J., Zhu, W., Huang, J.: Weakly supervised dense event captioning in videos. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 3063–3073 (2018) 4
  10. He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.Y., Ma, W.Y.: Dual learning for machine translation. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 820–828 (2016) 4
  11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Computer Vision (ICCV), 2017 IEEE International Conference on. pp. 2980–2988. IEEE (2017) 3
  12. Hendricks, L.A., Hu, R., Darrell, T., Akata, Z.: Grounding visual explanations. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 264–279 (2018) 4
  13. Hosseini-Asl, E., Zhou, Y., Xiong, C., Socher, R.: Augmented cyclic adversarial learning for low resource domain adaptation. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019) 4
  14. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4555–4564 (2016) 4
  15. Huang, Q., Zhang, P., Wu, D., Zhang, L.: Turbo learning for captionbot and drawingbot. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 6456–6466 (2018) 4
  16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015) 9, 19
  17. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017) 17, 18
  18. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(12), 2891–2903 (2013) 3
  19. Li, Y., Yao, T., Pan, Y., Chao, H., Mei, T.: Jointly localizing and describing events for dense video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7492–7500 (2018) 3
  20. Liu, C., Mao, J., Sha, F., Yuille, A.: Attention correctness in neural image captioning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017) 1, 2
  21. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7219–7228 (2018) 2, 3, 9
  22. Ma, C.Y., Kadav, A., Melvin, I., Kira, Z., AlRegib, G., Graf, H.P.: Attend and interact: Higher-order object interactions for video understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 3, 4
  23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002) 8
  24. Park, J.S., Rohrbach, M., Darrell, T., Rohrbach, A.: Adversarial inference for multi-sentence video description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3

25. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2641–2649 (2015) [2](#), [7](#)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 91–99 (2015) [3](#), [4](#), [17](#)
27. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4035–4045 (2018) [1](#)
28. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: European Conference on Computer Vision (ECCV). pp. 817–834. Springer (2016) [4](#), [10](#)
29. Rohrbach, A., Rohrbach, M., Tang, S., Joon Oh, S., Schiele, B.: Generating descriptions with grounded and co-referenced people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4979–4989 (2017) [3](#), [4](#)
30. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie description. International Journal of Computer Vision **123**(1), 94–120 (2017) [3](#)
31. Selvaraju, R.R., Lee, S., Shen, Y., Jin, H., Batra, D., Parikh, D.: Taking a hint: Leveraging explanations to make vision and language models more grounded. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019) [4](#)
32. Shah, M., Chen, X., Rohrbach, M., Parikh, D.: Cycle-consistency for robust visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) [4](#)
33. Shetty, R., Rohrbach, M., Anne Hendricks, L., Fritz, M., Schiele, B.: Speaking the same language: Matching machine to human captions by adversarial training. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4135–4144 (2017) [3](#)
34. Tang, D., Duan, N., Yan, Z., Zhang, Z., Sun, Y., Liu, S., Lv, Y., Zhou, M.: Learning to collaborate for question answering and asking. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). vol. 1, pp. 1564–1574 (2018) [4](#)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 5998–6008 (2017) [4](#)
36. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 4566–4575 (2015) [8](#)
37. Venugopalan, S., Anne Hendricks, L., Rohrbach, M., Mooney, R., Darrell, T., Saenko, K.: Captioning images with diverse objects. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5753–5761 (2017) [3](#)
38. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4534–4542 (2015) [3](#)

39. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 3156–3164 (2015) [3](#)
40. Wang, B., Ma, L., Zhang, W., Liu, W.: Reconstruction network for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7622–7631 (2018) [4](#)
41. Wang, F., Huang, Q., Guibas, L.J.: Image co-segmentation via consistent functional maps. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2013) [4](#)
42. Xiao, F., Sigal, L., Jae Lee, Y.: Weakly-supervised visual grounding of phrases with linguistic structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5945–5954 (2017) [4](#)
43. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 22–29 (2017) [3](#)
44. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4651–4659 (2016) [3, 9](#)
45. Zanfir, M., Marinou, E., Sminchisescu, C.: Spatio-temporal attention models for grounded video captioning. In: Asian Conference on Computer Vision. pp. 104–119 (2016) [3](#)
46. Zhang, Y., Niebles, J.C., Soto, A.: Interpretable visual question answering by visual grounding from attention supervision mining. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 349–357. IEEE (2019) [4](#)
47. Zhou, L., Kalantidis, Y., Chen, X., Corso, J.J., Rohrbach, M.: Grounded video description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) [2, 4, 7, 8, 9, 10, 15, 18, 19](#)
48. Zhou, L., Louis, N., Corso, J.J.: Weakly-supervised video object grounding from text by loss weighting and object interaction. In: British Machine Vision Conference (BMVC) (2018) [10](#)
49. Zhou, L., Xu, C., Koch, P., Corso, J.J.: Watch what you just said: Image captioning with text-conditional attention. In: Proceedings of the on Thematic Workshops of ACM Multimedia 2017. pp. 305–313. ACM (2017) [3](#)
50. Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8739–8748 (2018) [19](#)
51. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2223–2232 (2017) [4](#)