Grounded Video Description

Luowei Zhou
University of Michigan
luozhou@umich.edu

Yannis Kalantidis
Facebook Research
yannisk@fb.com

Xinlei Chen
Facebook AI Research
xinleic@fb.com

Jason J. Corso
University of Michigan
jjcorso@umich.edu

Marcus Rohrbach
Facebook AI Research
mrf@fb.com

Abstract

Video description is one of the most challenging problems in vision and language understanding due to the large variability both on the video and language side. Models, hence, typically shortcut the difficulty in recognition and generate plausible sentences that are based on priors but are not necessarily grounded in the video. In this work, we explicitly link the sentence to the evidence in the video by annotating each noun phrase in a sentence with the corresponding bounding box in one of the frames of a video. Our novel dataset, ActivityNet-Entities, is based on the challenging ActivityNet Captions dataset and augments it with 158k bounding box annotations, each grounding a noun phrase. This allows training video description models with this data, and importantly, evaluate how grounded or "true" such model are to the video they describe. To generate grounded captions, we propose a novel video description model which is able to exploit these bounding box annotations. We demonstrate the effectiveness of our model on our ActivityNet-Entities, but also show how it can be applied to image description on the Flickr30k Entities dataset. We set new state-of-the-art performance on video description, video paragraph description, and image description and demonstrate our generated sentences are more explainable through grounding.

1. Introduction

Image and video description models are frequently not well grounded [17] which can increase their bias [11] and lead to hallucination of objects [28], *i.e*. the model mentions objects which are not in the image or video *e.g*. because they might have appeared in the training data in similar contexts. This makes models less accountable and trustworthy, which is important if we hope that such models will eventually assist people in need [3, 31]. Additionally, grounded models

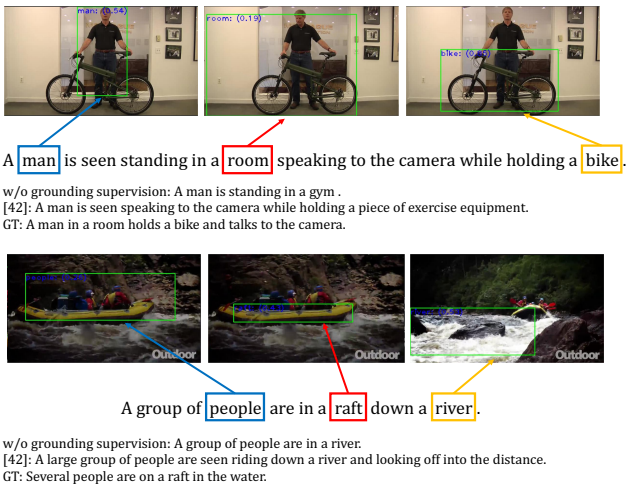


Figure 1. Word-level grounded video descriptions generated by our model on two segments from our novel ActivityNet-Entities dataset. Under the generated descriptions, we present the descriptions generated by our model without explicit bounding box supervision (denoted by "w/o grounding supervision"), the descriptions generated by [47] and the ground-truth descriptions (GT).

can help to explain the model's decisions to humans and allow humans to diagnose them [24]. While researchers have started to discover and study these problems for image description[1] [17, 11, 28, 24], these challenges are even more pronounced for video description due to the increased difficulty and diversity, both on the visual and the language side.

Fig. 1 illustrates this problem. A video description approach (without grounding supervision) generated the sentence "A man standing in a gym" which correctly mentions "a man" but hallucinates "gym" which is not visible in the video. Although a man is in the video it is not clear if the

[1]We use "description" rather than "captioning" as "captioning" is frequently used to refer to transcribing the speech in the video rather than *describing* the content.

model looked at the bounding box of the man to say this word [11, 28]. For the sentence "A man [...] is playing the piano" in Fig. 2, it is important to understand that which "man" in the image "A man" is referring to, to determine if a model is correctly grounded. Such understanding is crucial for many applications when trying to build accountable systems or when generating the next sentence or responding to a follow up question of a blind person: *e.g*. answering "Is *he* looking at me?" requires an understanding which of the people in the image the model talked about.

The goal of our research is to build such grounded systems. As one important step in this direction, we collect ActivityNet-Entities (short as ANet-Entities) which grounds or links noun phrases in sentences with bounding boxes in the video frames. It is based on ActivityNet Captions [13], one of the largest benchmarks in video description. When annotating objects or noun phrases we specifically annotate the bounding box which corresponds to the instance referred to in the sentence rather than all instances of the same object category, *e.g*. in Fig. 2, for the noun phrase "the man" in the video description, we only annotate the sitting man and not the standing man or the woman, although they are all from the object category "person". We note that annotations are sparse, in the sense that we only annotate a single frame of the video for each noun phrase. ANet-Entities has a total number of 51.8k annotated video segments/sentences with 157.8k labeled bounding boxes, more details can be found in Sec. 3.

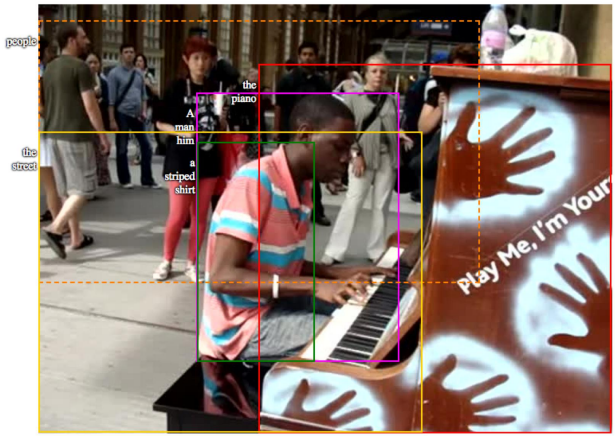
Our new dataset allows us to introduce a novel grounding-based video description model that learns to jointly generate words and refine the grounding of the objects generated in the description. We explore how this explicit supervision can benefit the description generation compared to unsupervised methods that might also utilize region features but do not penalize grounding.

Our contributions are threefold. First, we collect our large-scale ActivityNet-Entities dataset, which grounds video descriptions to bounding boxes on the level of noun phrases. Our dataset allows both, *teaching* models to explicitly rely on the corresponding evidence in the video frame when generating words and *evaluating* how well models are doing in grounding individual words or phrases they generated. Second, we propose a grounded video description framework which is able to learn from the bounding box supervision in ActivityNet-Entities and we demonstrate its superiority over baselines and prior work in generating grounded video descriptions. Third, we show the applicability of the proposed model to image captioning, again showing improvements in the generated captions and the quality of grounding on the Flickr30k Entities [26] dataset.

2. Related Work

Video & Image Description. Early work on automatic caption generation mainly includes template-based approaches [6, 15, 21], where predefined templates with slots are first generated and then filled in with detected visual evidences. Although these works tend to lead to well-grounded methods, they are restricted by their template-based nature. More recently, neural network and attention-based methods have started to dominate major captioning benchmarks. Visual attention usually comes in the form of temporal attention [39] (or spatial-attention [37] in the image domain), semantic attention [16, 40, 41, 46] or both [22]. The recent unprecedented success in object detection [27, 9] has regained the community's interests on detecting fine-grained visual clues while incorporating them into end-to-end networks [19, 30, 2, 18]. Description methods which are based on object detectors [19, 43, 2, 18, 6, 15] tackle the captioning problem in two stages. They first use off-the-shelf or fine-tuned object detectors to propose object proposals/detections as for the visual recognition heavy-lifting. Then, in the second stage, they either attend to the object regions dynamically [19, 43, 2] or classify the regions into labels and fill into pre-defined/generated sentence templates [18, 6, 15]. However, directly generating proposals from off-the-shelf detectors causes the proposals to bias towards classes in the source dataset (*i.e*. for object detection) v.s. contents in the target dataset (*i.e*. for description). One solution is to fine-tune the detector specifically for a dataset [18] but this requires exhaustive object annotations that are difficult to obtain, especially for videos. Instead of fine-tuning a general detector, we transfer the object classification knowledge from off-the-shelf object detectors to our model and then fine-tune this representation as part of our generation model with sparse box annotations. With a focus on co-reference resolution and identifying people, [30] proposes a framework that can refer to particular character instances and do visual co-reference resolution between video clips. However, their method is restricted to identifying human characters whereas we study more general the grounding of objects.

Attention Supervision. As fine-grained grounding becomes a potential incentive for next-generation vision-language systems, to what degree it can benefit remains an open question. On one hand, negative voices come from recent work on VQA [5, 44], where the authors point out the attention model does not attend to same regions as humans and adding attention supervision barely helps the performance. On the other hand, adding supervision to feature map attention [17, 42] is demonstrated to be beneficial. We notice in our experiments that directly guiding the region attention with supervision [18] does not necessary lead to performance gain. We hypothesize that this might be due to the lack of object context information. A self-attention [32]



A man in a striped shirt is playing the piano on the street while people watch him.

Figure 2. An annotated example from our dataset. The dashed box ("people") indicates a group of objects.

based context encoding is introduced in our attention model accordingly, which allows message passing across all regions in the sampled video frames.

3. ActivityNet-Entities Dataset

In order to train and test models capable of explicit grounding-based video description, one requires both language and grounding supervision. Although Flickr30k Entities [26] contains such annotations for images, no large-scale description datasets with object localization annotation exists for videos. The large-scale ActivityNet Captions dataset [13] contains dense language annotations for about 20k videos from ActivityNet [4] but lacks grounding annotations. Leveraging the language annotations from the ActivityNet Captions dataset [13], we further collected entity-level bounding box annotations and created the ActivityNet-Entities (ANet-Entities) dataset, a rich dataset that can be used for video description with explicit grounding. With 15k videos and more than 158k annotated bounding boxes, ActivityNet-Entities is the largest annotated dataset of its kind to the best of our knowledge.

When it comes to videos, region-level annotations come with a number of unique challenges. A video contains more information than can fit in a single frame, and video descriptions reflect that. They may reference objects that appear in a disjoint set of frames, as well as multiple persons and motions. To be more precise and produce finer-grained annotations, we annotate *noun phrases* (NP) (defined below) rather than simple object labels. Moreover, one would ideally have dense region annotations at every frame, but the annotation cost in this case would be prohibitive for even small datasets. Therefore in practice, video datasets are sparsely annotated at the region level [8]. Favouring scale over density, we choose to annotate segments as sparsely as possible and annotate every noun phrase only in one frame inside each segment.

Dataset	Domain	# Vid/Img	# Sent	# Obj	# BBoxes
Flickr30k Entities [26]	Images	32k	160k	480	276k
MPII-MD [30]	Video	$\ll$1k	$\ll$1k	4	2.6k
YouCook2 [45]	Video	2k	15k	67	135k
ActivityNet Humans [38]	Video	5.3k	5.3k	1	63k
ActivityNet-Entities (ours)	**Video**	**15k**	**52k**	**432**	**158k**
–train		10k	35k	432	105k
–val		2.5k	8.6k	427	26.5k
–test		2.5k	8.5k	421	26.1k

Table 1. Comparison of video description datasets with noun phrase or word-level grounding annotations. Our ActivityNet-Entities and ActivityNet Humans [38] dataset are both based on ActivityNet [4], but ActivityNet Humans provides only person bounding boxes for a small subset of videos. YouCook2 is restricted to cooking and only has box annotations for the val and the test splits.

Noun Phrases. Following [26], we define noun phrases as short, non-recursive phrases that refer to a specific region in the image, able to be enclosed within a bounding box. They can contain a single instance or a group of instances and may include adjectives, determiners, pronouns or prepositions. For granularity, we further encourage the annotators to split complex NPs into their simplest form (*e.g.* "the man in a white shirt with a heart" can be split into three NPs: "the man", "a white shirt", and "a heart").

3.1. Annotation Process

We uniformly sampled 10 frames from each video segment and presented them to the annotators together with the corresponding description. We asked them to identify all concrete NPs from the segment description and then draw bounding boxes around them in *one* frame of the video where the target NPs can be clearly observed. Further instructions were provided including guidelines for resolving co-references within a sentence, *i.e.* boxes may correspond to multiple NPs in the sentence (*e.g.*, a single box could refer to both "the man" and "him") or when to use *multi-instance boxes* (*e.g.* "crowd", "a group of people" or "seven cats"). An annotated example is shown in Fig. 2. It is noteworthy that 10% of the final annotations refer to multi-instance boxes. We trained annotators, and deployed a rigid quality control by daily inspection and feedback. All annotations were verified in a second round. *The full list of instructions provided to the annotators, validation process, as well as screen-shots of the annotation interface can be found in the Appendix.*

3.2. Dataset Statistics and Analysis

As the test set annotations for the ActivityNet Captions dataset are not public, we only annotate the segments in the training (train) and validation (val) splits. This brings the total number of annotated videos in ActivityNet-Entities to 14,281. In terms of segments, we ended up with about 52k

video segments with at least one NP annotation and 158k NP bounding boxes in total.

Respecting the original protocol, we keep as our training set the corresponding split from the ActivityNet Captions dataset. We further randomly & evenly split the original val set into our val set and our test set. We use all available bounding boxes for training our models, *i.e.*, including multi-instance boxes. Complete stats and comparisons with other related datasets can be found in Tab. 1.

From Noun Phrases to Objects Labels. Although we chose to annotate noun phrases, in this work, we model sentence generation as a word-level task. We follow the convention in [18] to determine the list of object classes and convert the NP label for box to a single-word object label. First, we select all nouns and pronouns from the NP annotations using the Stanford Parser [20]. The frequency of these words in the train and val splits are computed and a threshold determines whether each word is a object class. For ANet-Entities, we set the frequency threshold to be 50 which produces 432 object classes.[2]

4. Description with Grounding Supervision

In this section we describe the proposed grounded video description framework, shown in Fig. 3. The framework consists of three modules: grounding, region attention and language generation. The grounding module detects visual clues from the video, the region attention dynamically attends on the visual clues to form a high-level impression of the visual content and provides it to the language generation module for language decoding. We illustrate three options for incorporating the object-level supervision: region classification, object grounding (localization), and supervised attention.

4.1. Overview

We formulate the problem as a joint optimization over the language and grounding tasks. The overall loss function consists of four parts:

$$L = L_{sent} + \lambda_c L_{cls} + \lambda_\beta L_{grd} + \lambda_\alpha L_{attn}, \qquad (1)$$

where L_{sent} denotes the teacher-forcing language generation cross-entropy loss, commonly used for language generation tasks (details in Sec. 4.3). L_{cls} and L_{grd} are cross-entropy losses that correspond to the grounding module for region classification and supervised object grounding (localization), respectively (Sec. 4.2). Finally, L_{attn} corresponds to the cross entropy region attention loss which is presented in Sec. 4.4. The three grounding-related losses are weighted by coefficients λ_c, λ_β, and λ_α which we selected on the dataset val split.

[2]We will release ActivityNet-Entities publicly upon acceptance, with the complete set of NP-based annotations as well as the object-based ones.

We denote the input video (segment) as V and the target/generated sentence description (words) as S. We uniformly sample F frames from each video as $\{v_1, v_2, \ldots, v_F\}$ and define N_f object regions on sampled frame f. Hence, we can assemble a bag of regions $R = [R_1, \ldots, R_F] = [r_1, r_2, \ldots, r_N] \in \mathbb{R}^{d \times N}$ to represent the video, where r_i $(i = 1, 2, \ldots, N)$ are feature embeddings for the regions and $N = \sum_{f=1}^{F} N_f$ is the total number of regions. We represent words in S with one-hot vectors which are further encoded to word embeddings $y_t \in \mathbb{R}^e$ where $t \in 1, 2, \ldots, T$, where T indicates the sentence length and e is the embedding size.

4.2. Grounding Module

Let $\mathcal{K}$ be a set of visually-groundable object class labels $\{c_1, c_2, \ldots, c_\mathcal{K}\}$, short as object classes, where $\mathcal{K}$ is the total number of classes. Given a bag of object regions from all sampled frames, the grounding module estimates the class probability distribution for each region.

We define a set of object classifiers as $W_c = [w_1, w_2, \ldots, w_\mathcal{K}] \in \mathbb{R}^{d \times \mathcal{K}}$ and the learnable scalar biases as $B = [b_1, b_2, \ldots, b_\mathcal{K}]$. So, a naive way to estimate the class probabilities for all regions (embeddings) $R = [r_1, r_2, \ldots, r_N]$ is through dot-product:

$$M_s(R) = \text{Softmax}(W_c^\top R + B\mathbb{1}^\top), \qquad (2)$$

where $\mathbb{1}$ is a vector with all ones, $W_c^\top R$ is followed by a ReLU and a Dropout layer, and M_s is the *region-class similarity matrix* as it captures the similarity between regions and object classes. For clarify, we omit the ReLU and Dropout layer after the linear embedding layer throughout Sec. 4 unless otherwise specified. The Softmax operator is applied along the object class dimension of M_s to ensure the class probabilities for each region sum up to 1.

We transfer detection knowledge from an off-the-shelf detector that is pre-trained on a general source dataset (*i.e.*, Visual Genome, or VG [14]) to our object classifiers. We find the nearest neighbor for each of the $\mathcal{K}$ object classes from the VG object classes according to their distances in the embedding space (glove vectors [25]). We then initialize W_c and B with the corresponding classifier, *i.e.*, the weights and biases, from the last linear layer of the detector.

On the other hand, we represent the spatial and temporal configuration of the region as a 5-D tuple, including 4 values for the normalized spatial location and 1 value for the normalized frame index. Then, the 5-D feature is projected to a $d_s = 300$-D location embedding for all the regions $M_l \in \mathbb{R}^{300 \times N}$. Finally, we concatenate all three components: i) region feature, ii) region-class similarity matrix, and iii) location embedding together and project into a lower dimension space (m-D):

$$\tilde{R} = W_g[\, R \mid M_s(R) \mid M_l \,], \qquad (3)$$

where $[\cdot|\cdot]$ indicates a row-wise concatenation and $W_g \in$

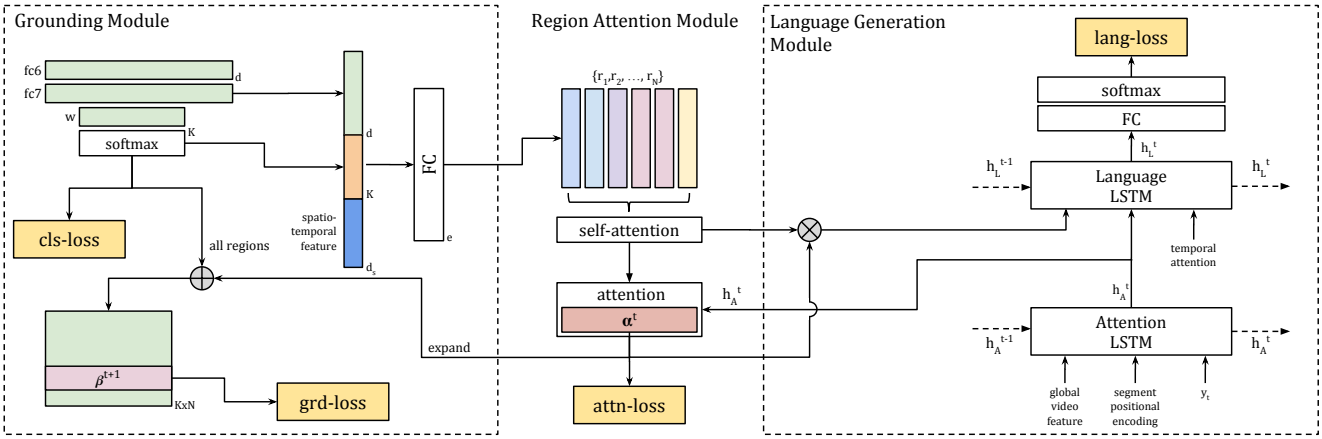


Figure 3. The proposed framework consists of three parts: the grounding module (left), the region attention module (middle) and the language generation module (right). Given a region proposal, we first represent it with a grounding-aware region encoding. The language model then dynamically attends on the region encodings when generating each word. Three losses are imposed on the attention weights (attn-loss), grounding weights (grd-loss), and the region classification probabilities (cls-loss). For clarity, the details of the temporal attention are omitted.

$\mathbb{R}^{m\times(d+K+d_s)}$ is the embedding weights. We name $\tilde{R}$ the *grounding-aware region encoding*. To further model the relations between regions, we deploy a self-attention layer over $\tilde{R}$ which allows message passing across any arbitrary regions in the sampled video frames [32, 47]. The final region encoding is fed into the region attention module (See Fig. 3, middle).

So far the object classifier discriminates classes without the prior knowledge about the semantic context, *i.e.*, the information the language model has captured. To incorporate semantics, we condition the class probabilities on the sentence encoding from Attention LSTM. A memory-efficient approach is treating attention weights α^t as this semantic prior, as formulated below:

$$M_s^t(R, \alpha^t) = \text{Softmax}(W_c^\top R + B\mathbb{1}^\top + \mathbb{1}{\alpha^t}^\top), \quad (4)$$

where the region attention weights α^t are determined by Eq. 7. Note that here the Softmax operator is applied row-wise to ensure the probabilities on regions sum up to 1. To learn a reasonable object classifier, we can either deploy a region classification task on $M_s(R)$ or a sentence-conditioned grounding task on $M_s^t(R, \alpha^t)$, with the word-level grounding annotations from Sec. 3.

Region Classification. We first define a positive region as a region that has a over 0.5 overlapping (IoU) with an arbitrary ground-truth (GT) box. If a region matches to multiple GT boxes, the one with the largest IoU is the final matched GT box. Then we classify the positive region, say region i to the same class label as in the GT box, say class c_j. The normalized class probability distribution is hence $M_s[:, i]$ and the cross-entropy on class c_j is as the follow:

$$L_{cls} = -\log M_s[j, i]. \quad (5)$$

The final L_{cls} is the average of losses on all positive regions.

Object Grounding. Given a visually-groundable word s_{t+1} at time step $t+1$ and the encoding of all the previous words, we aim to localize s_{t+1} in the video as one or a few of the region proposals. Supposing s_{t+1} corresponds to class c_j and the GT box is r_{GT}, we regress the confidence score of regions $M_s^t[j, :] = \beta^{t+1} = [\beta_1^{t+1}, \beta_2^{t+1}, \ldots, \beta_N^{t+1}]$ to the indicators of positive/negative regions $\gamma^t = [\gamma_1^t, \gamma_2^t, \ldots, \gamma_N^t]$, where $\gamma_i^t = 1$ when the region r_i has a over 0.5 overlapping with the GT box r_{GT} and otherwise 0. The grounding loss for word s_{t+1} is defined as:

$$L_{grd} = -\sum_{i=1}^{N} \gamma_i^t \log \beta_i^{t+1}. \quad (6)$$

4.3. Language Generation Module

For language generation, we adapt the language model from [18] for video inputs, *i.e.* extend it to incorporate temporal information. The model consists of two LSTMs: the first one for encoding the global video feature and the word embedding y_t into the hidden state $h_A^t \in \mathbb{R}^m$ and the second LSTM for language generation (see Fig. 3, right). The language model dynamically attends on videos frames or regions for visual clues to generate a description. We refer to the attention on video frames as temporal attention and the one on regions as region attention.

The temporal attention takes in a sequence of frame-wise feature vectors and determines by the hidden state how significant each frame should contribute to generate a description word. We deploy a similar module as in [47], except for we replace the self-attention context encoder with Bi-directional GRU (Bi-GRU) which yields superior results.

4.4. Region Attention Module

Unlike the temporal attention that works on a frame level, the region attention [2, 18] focuses on more fine-grained details in the video, *i.e.*, object regions [27]. Denote the grounding-aware region encoding defined in Eq. 3 as $\tilde{R} = [\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_N]$. At time t of the caption generation, the attention weight over region i is formulated as:

$$\alpha_i^t = w_\alpha^\top \tanh(W_r \tilde{r}_i + W_h h_A^t), \quad \alpha^t := \text{Softmax}(\alpha^t), \quad (7)$$

where $W_r \in \mathbb{R}^{m \times d}$, $W_h \in \mathbb{R}^{m \times m}$, $w_\alpha \in \mathbb{R}^m$, and $\alpha^t = [\alpha_1^t, \alpha_2^t, \ldots, \alpha_N^t]$. The region attention encoding is then $R\alpha^t$ and along with the temporal attention encoding, fed into the language LSTM. In Sec. 5.2, we show that the two attentions are complementary in a sense that the temporal attention captures the coarse-level details while the region attention captures more fine-grained details.

Supervised Attention. We want to encourage the language model to attend on the correct region when generating a visually-groundable word. As this effectively assists the language model in learning to attend to the correct region, we call this *attention supervision*. The corresponding loss, L_{attn}, is simply defined by replacing β_i^{t+1} in Eq. 6 with α_i^t:

$$L_{attn} = -\sum_{i=1}^{N} \gamma_i^t \log \alpha_i^t. \quad (8)$$

The difference between the grounding supervision and the attention supervision is that, in the former task, the target object c_j is known beforehand, while the attention module is not aware of which object to seek in the scene. In practice, restricting the grounding region candidates within the target frame (w/ GT box) during training, *i.e.*, only consider the N_f proposals on the frame f with the GT box, gives decent grounding accuracy during inference. Note that the final loss on L_{grd} or L_{attn} is the average of losses on all visually-groundable words.

5. Experiments

Datasets. We conduct most experiments and ablation studies on the newly-collected ActivityNet-Entities dataset on video description given the set of temporal segments (*i.e.* using the ground-truth events from [13]) and video paragraph description [35]. We also demonstrate our framework can easily be applied to image description and evaluate it on the Flickr30k Entities dataset [26]. Note that we did not apply our method to COCO captioning as there is no exact match between words in COCO captions and object annotations in COCO (limited to only 80). We use the same process described in Sec. 3.2 to convert NPs to object labels. Since Flickr30k Entities contains more captions, labels that occur at least 100 times are taken as object labels, resulting in 480 object classes [18].

Pre-processing. For ANet-Entities, we truncate captions longer than 20 words and build a vocabulary on words with at least 3 occurrences. For Flickr30k Entities, since the captions are generally shorter and have a larger amount, we truncate captions longer than 16 words and build a vocabulary based on words that occur at least 5 times.

5.1. Compared Methods and Metrics

Compared methods. The state-of-the-art (SotA) video description methods on ActivityNet Captions include Masked Transformer and Bi-LSTM+TempoAttn [47]. We re-train the models on our dataset splits with the original settings. For a fair comparison, we use exactly the same frame-wise feature from this work for our temporal attention module. For video paragraph description, we compare our methods against the SotA method MFT [35] with the evaluation script provided by the authors [35]. For image captioning, we compare against two SotA methods, Neural Baby Talk (NBT) [18] and BUTD [2]. For a fair comparison, we provide the same region proposal and features for both the baseline BUTD and our method, *i.e.*, from Faster R-CNN pre-trained on Visual Genome (VG). NBT is specially tailored for each dataset (*e.g.*, detector fine-tuning), so we retain the same feature as in the paper, *i.e.*, from ResNet pre-trained on ImageNet. All our experiments are performed three times and the average scores are reported.

Metrics. For the region classification task, we compute the top-1 classification accuracy (*Cls.* in the tables) for positive regions. To measure the object grounding and attention correctness, we compute the localization accuracy (*Grd.* and *Attn.* in the tables). During inference, the region with the highest attention weight (α_i) or grounding weight (β_j) is compared against the GT box. If the IoU is over 0.5, then the region is correct. Note that the object localization accuracy is computed over GT sentences following [29, 45]. We also study the attention accuracy on generated sentences, denoted by Prec. and Recall in the tables. This metric only considers correctly-predicted objects. If multiple instances of the same object exist in the target sentence, we only consider matching the first instance. Due to the sparsity of the annotation, *i.e.*, each object only annotated in one frame, we only consider proposals in the frame of the GT box when computing the localization accuracy. For all the metrics, we average the scores across over object classes. For evaluating the sentence quality, we use standard language evaluation metrics, including Bleu@1, Bleu@4, METEOR, CIDEr, and SPICE, and the official evaluation script[3].

5.2. Implementation Details

Region proposal and feature. We uniformly sample 10 frames per video segment (an event in ANet-Entities) and extract region features. For each frame, we use a Faster RCNN model [27] with a ResNeXt-101 backbone [34] for region proposal and feature extraction (fc6 layer output).

[3] https://github.com/ranjaykrishna/densevid_eval

Method	λ_α	λ_β	λ_c	B@1	B@4	M	C	S	Attn.	Grd.	Prec.	Recall	Cls.
Unsup. (w/o SelfAttn)	0	0	0	23.2	2.28	10.9	45.6	**15.0**	14.7	21.0	7.63	7.70	6.87
Unsup.	0	0	0	23.0	2.27	10.7	44.6	13.8	2.27	19.4	2.33	2.90	6.10
Sup. Attn.	0.05	0	0	23.7	2.56	**11.1**	47.0	14.9	33.5	36.8	10.5	15.2	0.42
Sup. Grd.	0	0.5	0	23.5	2.50	11.0	46.8	14.7	31.5	42.9	**13.7**	**18.2**	0.06
Sup. Cls.	0	0	0.1	23.3	2.43	10.9	45.7	14.1	2.48	25.4	3.30	3.53	**14.9**
Sup. Attn.+Grd.	0.5	0.5	0	**23.8**	2.44	**11.1**	46.1	14.8	**35.2**	41.2	11.5	14.8	0
Sup. Attn.+Cls.	0.05	0	0.1	**23.9**	**2.59**	**11.2**	**47.5**	**15.1**	34.0	41.2	11.6	15.6	**14.3**
Sup. Grd. +Cls.	0	0.05	0.1	**23.8**	**2.59**	**11.1**	**47.5**	**15.0**	26.7	**45.3**	**11.9**	14.3	13.9
Sup. Attn.+Grd.+Cls.	0.1	0.1	0.1	**23.8**	2.57	**11.1**	46.9	**15.0**	**35.4**	**44.4**	11.3	**16.4**	12.3

Table 2. Results on ANet-Entities val set. w/o SelfAttn indicates self-attention is not used for region feature encoding. B@1 stands for Bleu@1, B@4 stands for Bleu@4, M stands for METEOR, C stands for CIDEr, S stands for SPICE. Attn. and Grd. are the object localization accuracies for attention and grounding on GT sentences. Prec. and Recall are the object localization accuracies for attention on generated sentences. Cls. indicates classification accuracy. All the accuracies are in %. The top two scores on each metric are bold.

Method	B@1	B@4	M	C	S	Attn.	Grd.	Prec.	Recall	Cls.
Masked Transformer [47]	22.9	**2.41**	10.6	**46.1**	13.7	–	–	–	–	–
Bi-LSTM+TempoAttn [47]	22.8	2.17	10.2	42.2	11.8	–	–	–	–	–
Our Unsup. (w/o SelfAttn)	23.1	2.16	10.8	44.9	**14.9**	15.9	22.2	7.43	7.00	6.5
Our Sup. Attn.+Cls.	**23.6**	2.35	**11.0**	45.5	14.7	**34.6**	**43.2**	**11.3**	**15.9**	**14.6**

Table 3. Results on ANet-Entities test set. The top one score for each metric is bold.

The Faster RCNN model is pretrained on the Visual Genonme dataset [14]. More model and training details are in the Appendix.

Feature map and attention. The temporal feature map is essentially a stack of frame-wise appearance and motion features from [47, 36]. The spatial feature map in image description is the output of the conv4 layer from ResNet-101 [18, 10]. Note that an average pooling on the temporal or spatial feature map gives the global feature. In video description, we augment the global feature with segment positional information (*i.e.*, total number of segments, segment index, start time and end time) , which is empirically important.

Hyper-parameters. The margin loss coefficient Δ is set to 0.5, $\lambda_\alpha \in \{0.05, 0.1, 0.5\}$, $\lambda_\beta \in \{0.05, 0.1, 0.5\}$, and $\lambda_c \in \{0.1, 0.5, 1\}$ vary in the experiments as a result of model validation. We set $\lambda_\alpha = \lambda_\beta$ when they are both non-zero considering the two losses have a similar functionality. The region encoding size $d = 2048$, word embedding size $e = 512$ and RNN encoding size $m = 1024$ for all methods. Other hyper-parameters in the language module are the same as in [18]. We use a 2-layer 6-head Transformer encoder as the self-attention module [47].

5.3. Results on ActivityNet-Entities

5.3.1 Video Event Description

Although dense video description [14] further entails localizing the segments to describe on the temporal axis, in this paper we focus on the language generation part and assume the temporal boundaries for events events are given beforehand. We name this task Video Event Description. Results on the validation and test splits of our ActivityNet-Entities dataset are shown in Tab. 2 and Tab. 3, respectively. Given the selected set of region proposals, the localization upper bound on the val/test sets is 82.5%/83.4%, respectively.

In general, methods with some form of grounding supervision work consistently better than the methods without. Moreover, combining multiple losses, *i.e.* stronger supervision, leads to higher performance. On the val set, the best variant of supervised methods (*i.e.*, Sup. Grd.+Cls.) ourperforms the best variant of unsupervised methods (*i.e.*, Unsup. (w/o SelfAttn)) by a relative 1-13% on all the metrics. On the test set, the gaps are small for Bleu@1, METEOR, CIDEr, and SPICE (within $\pm$ 2%), but the supervised method has a 8.8% relative improvement on Bleu@4.

The results in Tab. 3 show that adding box supervision dramatically improves the grounding accuracy from 22.2% to 43.2%. Hence, our supervised models can better localize the objects mentioned which can be seen as an improvement in their ability to explain or justify their own description. The attention accuracy also improves greatly from 15.9% to 34.6% on GT sentences and 7.43%/7.00% to 11.3%/15.9% on generated sentences, implying that the supervised models learn to attend on more relevant objects during language generation than the unsupervised models. However, grounding loss alone fails with respect to classification accuracy (see Tab. 2), and therefore the classification loss is required in that case. Conversely, the classification loss alone can implicitly learn grounding and maintains a fair grounding accuracy.

Comparison to existing methods. We show that our best model sets the new SotA on Bleu@1, METEOR and SPICE on ActivityNet Captions dataset with relative gains of 2.8%, 3.9% and 6.8% over the previous best [47]. We observe

Approach	Judgments in %	Δ
About Equal	38.9	
Our Unsup. (w/o SelfAttn) is better	27.5	6.1
Our Sup. Attn.+Cls. is better	**33.6**	

Table 4. Human evaluation of sentence quality on test set of ActivityNet-Entities (for automatic evaluation see Tab. 3 in main paper). Our supervised approach vs. our unsupervised baseline.

Approach	Judgments in %	Δ
About Equal	34.9	
Masked Transformer [47] is better	29.3	6.5
Our Sup. Attn.+Cls. is better	**35.8**	

Table 5. Human evaluation of sentence quality on test set of ActivityNet-Entities (for automatic evaluation see Tab. 3 in main paper). Our supervised approach vs. Masked Transformer [47]

slightly inferior results on Bleu@4 and CIDEr (-2.8% and -1.4%, respectively) but after examining the generated sentences (see Appendix) we see that [47] generates repeated words way more often. This may increase the aforementioned evaluation metrics, but the generated descriptions are of lower quality. Another noteworthy observation is that the self-attention context encoder (on top of $\tilde{R}$) brings consistent improvements on methods with grounding supervision, but hurts the performance of methods without, *i.e.*, "Unsup.". We hypothesize that the extra context and region interaction introduced by the self-attention confuses the region attention module and without any grounding supervision makes it fail to properly attend to the right region, something that leads to a huge attention accuracy drop from 14.7% to merely 2.27%.

Human Evaluation. Automatic metrics to evaluate generated sentences, such as Bleu [23], Meteor [7], Cider [33], or Spice [1], have frequently shown to be unreliable and not consistent with human judgments, especially for video description when there is only a single reference [31]. Hence, we conducted a human evaluation to evaluate the sentence quality on the same test set of ActivityNet-Entities. We study the two most interesting comparisons: i) our supervised approach (Sup. Attn.+Cls.) v.s. our unsupervised baseline (Unsup. (w/o SelfAttn)) and ii) our supervised approach (Sup. Attn.+Cls.) v.s. the state-of-the-art approach Masked Transformer [47]. We randomly sampled 329 video segments and presented the video segments and descriptions to the judges. The human judges have to choose that either one of the sentence is better or that both sentences are about equal (could be equally good or bad). From Tab. 4, we observe that, while they frequently produce captions with similar quality, our Sup. Attn.+Cls. works better than the unsupervised baseline (with a significant gap of 6.1%). In Tab. 5 we can see that our Sup. Attn.+Cls. approach works better than the Masked Transformer [47] with a significant gap of 6.5%. We believe these results are a strong indication that our approach is not only better grounded but also generates better sentences, both compared to baselines and prior work [47]. See also our qualitative results in Figs. 6 and 7.

Method	B@1	B@4	M	C	S
Region Attn.	23.2	2.55	10.9	43.5	14.5
Tempo. Attn.	23.5	2.45	11.0	44.3	14.0
Both	**23.9**	**2.59**	**11.2**	**47.5**	**15.1**

Table 6. Ablation study on the two attention modules. Experiments are based on our best model and conducted on the val set.

Method	B@1	B@4	M	C
MFT [35]	45.5	9.78	14.6	20.4
Our Unsup. (w/o SelfAttn)	49.8	10.5	15.6	21.6
Our Sup. Attn.+Cls.	**49.9**	**10.7**	**16.1**	**22.2**

Table 7. Results on video paragraph description. All scores are on our test set.

Temporal attention & region attention. We conduct ablation studies on the two attention modules to study the impact of each component on the overall performance (see Tab. 6). Each module alone performs similarly and the combination of two performs the best, which indicates the two attention modules are complementary. Note that the region attention module takes in a lower sampling rate input than the region attention module, so we expect it can be further improved if having a higher sampling rate and the context (other events in the video). We leave this for future studies.

5.3.2 Video Paragraph Description

Besides measuring the quality of each individual description, we also evaluate the coherence among sentences within a video. Sentences from the same video are stitched together chronologically into one paragraph and evaluated against the GT paragraph. The authors of the SoTA method [35] kindly provided us with their result file and evaluation script, but as they were unable to provide us with their splits, we evaluated both methods on *our* test split. Even though we are under an unfair disadvantage, *i.e.*, the authors' val split might contain videos from our test split, we still outperform SotA method by a large margin, with relative improvements of 8.9-10% on all the metrics (see Tab. 7. The results are even more surprising given that we generate description for each event separately, without conditioning on previously-generated sentences. We hypothesize that the temporal attention module can effectively model the event context through the Bi-GRU context encoder and the context information benefits the coherence of consecutive sentences.

Method	VG	Box	B@1	B@4	M	C	S	Attn.	Grd.	Prec.	Recall	Cls.
ATT-FCN* [41]			64.7	19.9	18.5	–	–	–	–	–	–	–
NBT* [18]		✓	69.0	27.1	21.7	57.5	15.6	–	–	–	–	–
BUTD [2]	✓		69.4	**27.3**	21.7	56.6	16.0	24.5	32.3	13.5	13.7	1.89
Our Unsup. (w/o SelfAttn)	✓		69.5	27.0	22.1	60.1	16.1	21.7	25.6	11.9	11.8	18.4
Our Sup. Attn.+Grd.+Cls.	✓	✓	**69.9**	**27.3**	**22.5**	**62.3**	**16.5**	**41.8**	**51.2**	**25.0**	**26.6**	**19.9**

Table 8. Results on Flickr30k Entities test set. * indicates the results are obtained from the original papers. "VG" indicates region features are from VG pre-training. The top one score for each metric is bold.

5.4. Results on Flickr30k Entities

We show the overall results on image description in Tab. 8 (test). Due to space limit, we place the validation results in the Appendix. The method with the best validation CIDEr score is the full model (Sup. Attn.+Grd.+Cls.). The upper bounds on the val/test sets are 90.0%/88.5%, respectively. The major findings are as follows. The supervised method outperforms the unsupervised baseline by a relative 0.6-3.6% over all the metrics. Our best model sets new SotA for all the five metrics with relative gains from 0.6-10%. In the meantime, the object localization and region classification accuracies are all significantly boosted, showing that our captions can be better visually explained and understood.

6. Conclusion

We collected ActivityNet-Entities, a novel dataset that allows the study of video description and grounding. In this work, we show how to leverage the noun phrase annotations to generate grounded video descriptions. We also use our dataset to evaluate how well the generated sentences are grounded. We believe our large-scale annotations will also allow for more in-depth analysis which have previously only been able on images, *e.g.* about hallucination [28] and bias [11] as well as studying co-reference resolution. Besides, we showed in our comprehensive experiments on video and image description, how the box supervision can improve the accuracy and the explainability of the generated captions by not only generating sentences but also the corresponding regions in the image. Our model sets the new state-of-the-art performance when evaluated for video paragraph description and has a significant increase in grounding metrics without loosing to competitive existing methods according to video description metrics on ActivityNet-Entities. We also adapted our model to image description and evaluated it on the Flickr30k Entities dataset where our model outperforms existing methods, both with respect to description quality and grounding accuracy.

Acknowledgement. The technical work was performed during Luowei's summer intern at Facebook AI Research. This work is also partly supported by DARPA FA8750-17-2-0112 and NSF IIS 1522904. This article solely reflects the opinions and conclusions of its authors but not the DARPA or NSF.

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. 8

[2] 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 vqa. *arXiv preprint arXiv:1707.07998*, 2017. 2, 6, 9, 14

[3] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. Vizwiz: Nearly real-time answers to visual questions. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 333–342, New York, NY, USA, 2010. ACM. 1

[4] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 3

[5] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra. 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. 2

[6] P. Das, C. Xu, R. F. Doell, and J. J. Corso. 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*, pages 2634–2641, 2013. 2

[7] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 8

[8] C. Gu, C. Sun, S. Vijayanarasimhan, C. Pantofaru, D. A. Ross, G. Toderici, Y. Li, S. Ricco, R. Sukthankar, C. Schmid, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE international conference on computer vision*, 2018. 3

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 2

[10] 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*, pages 770–778, 2016. 7

[11] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach. Women also snowboard: Overcoming bias in

captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787, 2018. 1, 2, 9

[12] Y. Jiang, V. Natarajan, X. Chen, M. Rohrbach, D. Batra, and D. Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018. 14

[13] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 2, 3, 6

[14] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 4, 7, 14

[15] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013. 2

[16] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7492–7500, 2018. 2

[17] C. Liu, J. Mao, F. Sha, and A. L. Yuille. Attention correctness in neural image captioning. In *AAAI*, pages 4176–4182, 2017. 1, 2

[18] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7219–7228, 2018. 2, 4, 5, 6, 7, 9

[19] C.-Y. Ma, A. Kadav, I. Melvin, Z. Kira, G. AlRegib, and H. P. Graf. Attend and interact: Higher-order object interactions for video understanding. *arXiv preprint arXiv:1711.06330*, 2017. 2

[20] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014. 4

[21] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics, 2012. 2

[22] Y. Pan, T. Yao, H. Li, and T. Mei. Video captioning with transferred semantic attributes. In *CVPR*, volume 2, page 3, 2017. 2

[23] 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*, pages 311–318. Association for Computational Linguistics, 2002. 8

[24] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. 2018. 1

[25] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 4

[26] 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 IEEE international conference on computer vision*, pages 2641–2649, 2015. 2, 3, 6, 12

[27] 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*, pages 91–99, 2015. 2, 6, 14

[28] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018. 1, 2, 9

[29] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016. 6

[30] A. Rohrbach, M. Rohrbach, S. Tang, S. J. Oh, and B. Schiele. Generating descriptions with grounded and co-referenced people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3

[31] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele. Movie description. 2017. 1, 8

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 2, 5

[33] 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*, pages 4566–4575, 2015. 8

[34] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5987–5995. IEEE, 2017. 6, 14

[35] Y. Xiong, B. Dai, and D. Lin. Move forward and tell: A progressive generator of video descriptions. *ECCV*, 2018. 6, 8

[36] Y. Xiong, L. Wang, Z. Wang, B. Zhang, H. Song, W. Li, D. Lin, Y. Qiao, L. Van Gool, and X. Tang. Cuhk & ethz & siat submission to activitynet challenge 2016. *arXiv preprint arXiv:1608.00797*, 2016. 7

[37] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 2

[38] M. Yamaguchi, K. Saito, Y. Ushiku, and T. Harada. Spatio-temporal person retrieval via natural language queries. *arXiv preprint arXiv:1704.07945*, 2017. 3

[39] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015. 2

[40] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. In *IEEE International Conference on Computer Vision, ICCV*, pages 22–29, 2017. 2

[41] 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*, pages 4651–4659, 2016. 2, 9

[42] Y. Yu, J. Choi, Y. Kim, K. Yoo, S.-H. Lee, and G. Kim. Supervising neural attention models for video captioning by human gaze data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu, Hawaii*, pages 2680–29, 2017. 2

[43] M. Zanfir, E. Marinoiu, and C. Sminchisescu. Spatio-temporal attention models for grounded video captioning. In *Asian Conference on Computer Vision*, pages 104–119, 2016. 2

[44] Y. Zhang, J. C. Niebles, and A. Soto. Interpretable visual question answering by visual grounding from attention supervision mining. *arXiv preprint arXiv:1808.00265*, 2018. 2

[45] L. Zhou, N. Louis, and J. J. Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. *BMVC*, 2018. 3, 6

[46] L. Zhou, C. Xu, P. Koch, and J. J. Corso. Watch what you just said: Image captioning with text-conditional attention. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 305–313. ACM, 2017. 2

[47] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. 1, 5, 6, 7, 8

A. Appendix

This Appendix provides additional details, evaluations, and qualitative results.

- In Sec. A.1, we provide more details on our dataset including the annotation interface and examples of our dataset, which are shown in Figs. 4, 5.
- In Sec. A.2, we provide additional ablations and results on our ActivityNet-Entities dataset, including qualitative results, which are shown in Figs. 6, 7.
- In Sec. A.3, we provide additional ablations and results on the Flickr30kEntities dataset, including qualitative results, which are shown in Fig. 8.
- In Sec. A.4, we provide more implementation details (*e.g.*, training details).

A.1. Dataset

Definition of a noun phrase. Following the convention from Flickr30k Entities dataset [26], we define noun phrase as:

- short (avg. 2.23 words), non-recursive phrases (*e.g.*, the complex NP "the man in a white shirt with a heart" is split into three: "the man", "a white shirt", and "a heart")
- refer to a specific region in the image so as to be annotated as a bounding box.
- could be
  - a single instance (*e.g.*, a cat),
  - multiple distinct instances (*e.g.* two men),
  - a group of instances (*e.g.*, a group of people),
  - a region or scene (*e.g.*, grass/field/kitchen/town),
  - a pronoun, *e.g.*, it, him, they.
- could include
  - adjectives (*e.g.*, a *white* shirt),
  - determiners (*e.g.*, *A* piece of exercise equipment),
  - prepositions (*e.g.* the woman *on the right*)
  - other noun phrases, if they refer to the identical bounding concept & bounding box (*e.g.*, a group of people, a shirt of red color)

Annotator instructions

Further instructions include:

- Each word from the caption can appear in at most one NP. "A man in a white shirt" and "a white shirt" should not be annotated at the same time.
- Annotate multiple boxes for the same NP if the NP refers to multiple instances.
  - If there are more than 5 instances/boxes (*e.g.*, six cats or many young children), mark all instances as a single box and mark as "a group of objects".
  - Annotate 5 or fewer instances with a single box if the instances are difficult to separate, *e.g.* if they are strongly occluding each other.
- We don't annotation a NP if it's abstract and not presented in the scene (*e.g.*, "the camera" in "A man is speaking to the camera")
- One box can correspond to multiple NPs in the sentence (*e.g.*, "the man" and "him"), *i.e.*, we annotate co-references within one sentence.

See Fig. 4 for more examples.

Annotation interface. We show a screen shot of the interface in Fig. 5.

Validation process. We deployed a rigid quality control process during annotations. We were in daily contact with the annotators, encouraged them to flag all examples that were unclear and inspected a sample of the annotations daily, providing them with feedback on possible spotted annotation errors or guideline violations. We also had a post-annotation verification process where all the annotations are verified by human annotators.

Dataset statistics. The average number of annotated boxes per video segment is 2.56 and the standard deviation is 2.04. The average number of object labels per box is 1.17 and the standard deviation is 0.47. The top ten frequent objects are "man", "he", "people", "they", "she", "woman", "girl", "person", "it", and "boy". Note that the statistics are on object boxes, *i.e.*, after pre-processing.

List of objects. Tab. 10 lists all the 432 object classes which we use in our approach. We threshold at 50 occurrences. Note that the annotations in ActivityNet-Entities also contain the full noun phrases w/o thresholds.

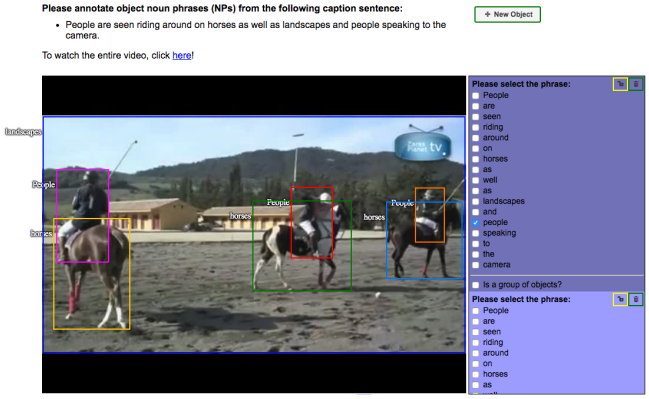
A.2. Results on ActivityNet-Entities

Qualitative examples. See Figs. 6 and 7 for the qualitative results by our methods and the Masked Transformer on ANet-Entities val set. We visualize the proposal with the highest attention weight in the corresponding frame. In (a), the supervised model correctly attends on "man" and "Christmas tree" in the video when generating the corresponding words. The unsupervised model mistakenly predicts "Two boys". In (b), both "man" and "woman" are correctly grounded. In (c), both "man" and "saxophone" are correctly grounded by our supervised model while Masked Transformer hallucinates a "bed". In (d), all the visually-groundable objects (*i.e.*, "people", "beach", "horses") are correctly localized. The caption generated by Masked Transformer is incomplete. In (e), surprisingly, not only major objects "woman" and "court" are localized, but also the small object "ball" is attended with a high accuracy. Masked Transformer incorrectly predicts the gender of the person. In (f), the unsupervised model overlooks most of the visual details (*e.g.*, "bottle", "glass"). The output of the supervised model is much more accurate, despite that the glass is grounded to a picture of a glass rather than the target object (*i.e.*, the real glass). In (g), the Masked Transformer outputs a unnatural caption "A group of people are in a raft and a man in red raft raft raft raft raft" containing consecutive repeated words "raft".

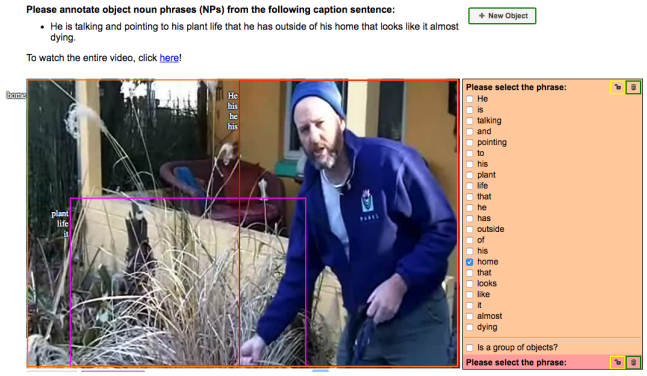
A.3. Results on Flickr30k Entities

See Tab. 9 for the results on Flickr30k Entities val set. Note that the results on the test set can be found in the main paper

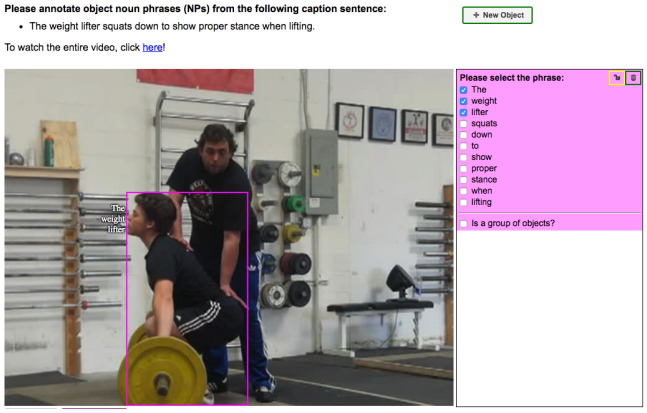
(a) "Teams" refers to more than 5 instances and hence should be annotated as a group.



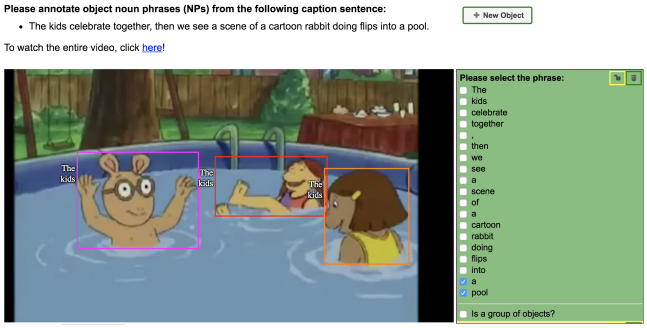
(b) "People" and "horses" can be clearly separated and the # of instances each is ≤ 5. So, annotate them all.



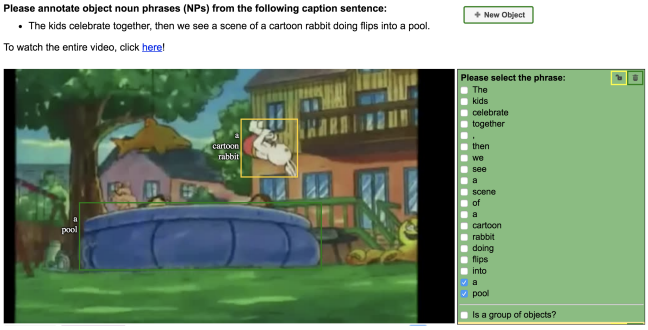
(c) "plant life" and "it" refer to the same box and "He", "'his", "he", "his" all refer to the same box.



(d) Only annotate the NP mentioned in the sentence, in this case, "The weight lifter". "proper stance" is a NP but not annotated because it is abstract/not an object in the scene.



(e) Note that (e) and (f) refer to the same video segment. See the caption of (f) for more details.



(f) "The kids" is annotated in a different frame as "a cartoon rabbit" and "a pool", since they can not be observed in the same frame.

Figure 4. Examples of our ActivityNet-Entities annotations in the annotation interface.

in Tab. 6. The proposal upper bound for attention and grounding is 90.0%. For supervised methods, we perform a light hyper-parameter search and notice the setting $\lambda_\alpha = 0.1$, $\lambda_\beta = 0.1$ and $\lambda_c = 1$ generally works well. The supervised methods outperform the unsupervised baseline by a decent amount in all the metrics with only one exceptions: Sup. Cls., which has a slightly inferior result in CIDEr. The best supervised method outperforms the best unsupervised baseline by a relative 0.9-4.8% over all the metrics.

Qualitative examples. See Fig. 8 for the qualitative results by our methods and the BUTD on Flickr30k Entities val set. We visualize the proposal with the highest attention weight as the green box. The corresponding attention weight and the most confident object prediction of the proposal are displayed as the blue text inside the green box. In (a), the supervised model correctly attends on

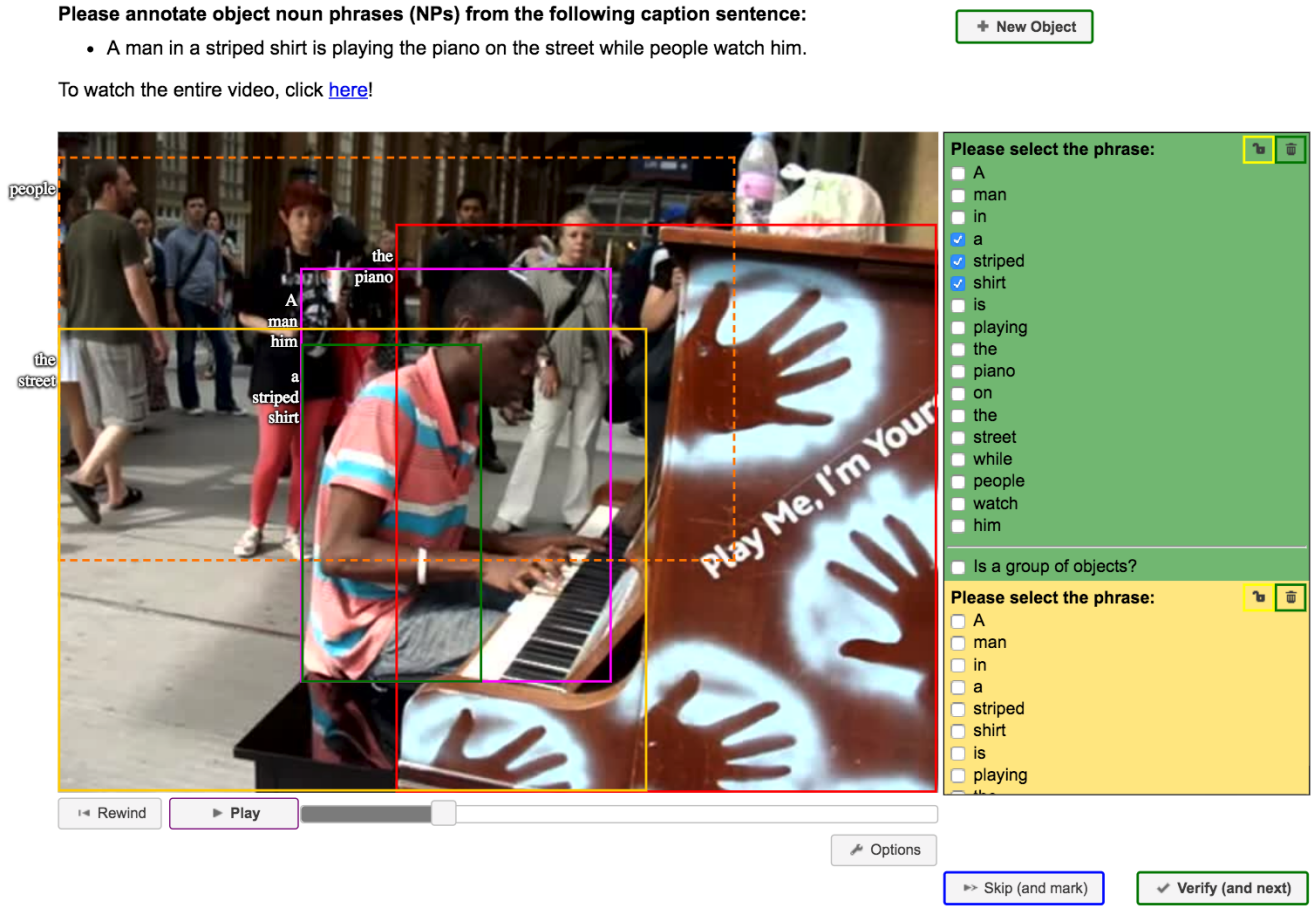


Figure 5. A screen shot of our annotation interface. The "verify (and next)" button indicates the annotation is under the verification mode, where the initial annotation is loaded and could be revised.

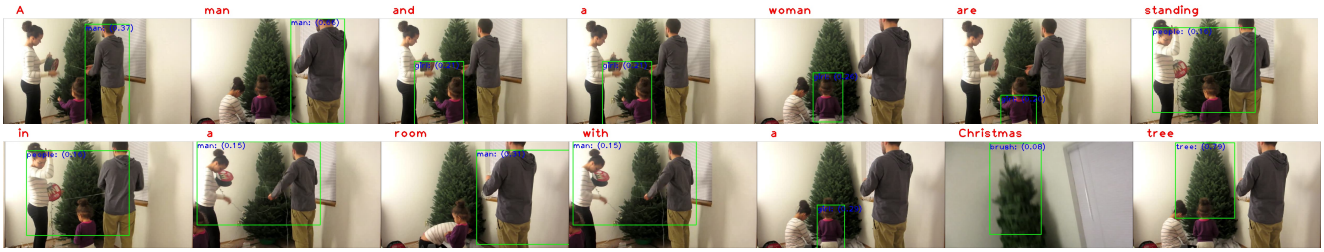
"man", "dog" and "snow" in the image when generating the corresponding words. The unsupervised model misses the word "snow" and BUTD misses the word "man". In (b), the supervised model successfully incorporates the detected visual clues (*i.e.*, "women", "building") into the description. We also show a negative example in (c), where interestingly, the back of the chair looks like a laptop, which confuses our grounding module. The supervised model hallucinates a "laptop" in the scene.

A.4. Implementation Details

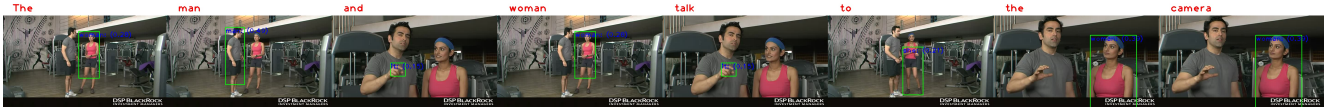
Region proposal and feature. We uniformly sample 10 frames per video segment (an event in ANet-Entities) and extract region features. For each frame, we use a Faster RCNN model [27] with a ResNeXt-101 FPN backbone [34] for region proposal and feature extraction. The Faster RCNN model is pretrained on the Visual Genonme dataset [14]. We use the same train-val-test split pre-processed by Anderson *et al*. [2] for joint object detection (1600 classes) and attribute classification. In order for a proposal to be considered valid, its confident score has to be greater than 0.2. And we limit the number of regions per image to a fixed 100 [12]. We take the output of the fc6 layer as the feature representation for each region, and fine-tune the fc7 layer with $0.1\times$ learning rate during model training.

Training details. We optimize the training with Adam (params: 0.9, 0.999). The learning rate is set at 5e-4 in general and 5e-5 for fine-tuning, *i.e.*, fc7 layer and object classifiers, decayed 0.8 every 3 epochs. The batch size is 240 for all the methods. We implement the model in PyTorch based on NBT[4] and train on 8x V100 GPUs. The training is limited to 40 epochs and the model with the best validation CIDEr score is selected for testing.

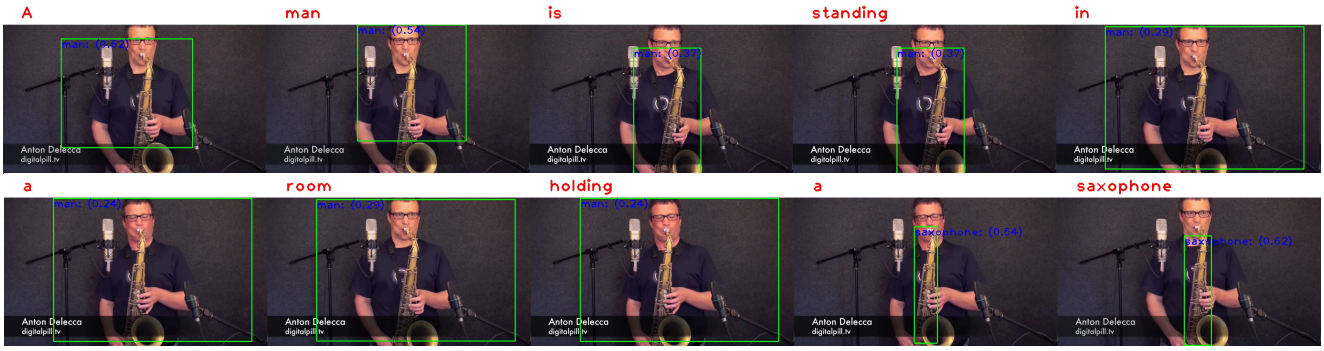
[4]https://github.com/jiasenlu/NeuralBabyTalk



(a) **Sup.**: A man and a woman are standing in a room with a Christmas tree;
Unsup.: Two boys are seen standing around a room holding a tree and speaking to one another;
Masked Trans.: They are standing in front of the christmas tree;
GT: Then, a man and a woman set up a Christmas tree.



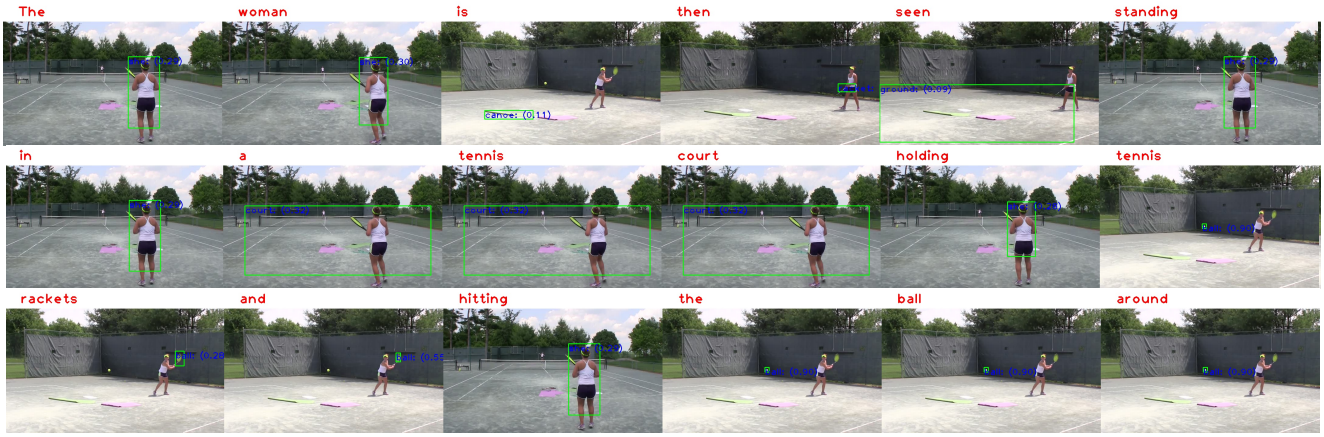
(b) **Sup.**: The man and woman talk to the camera;
Unsup.: The man in the blue shirt is talking to the camera;
Masked Trans.: The man continues speaking while the woman speaks to the camera;
GT: The man and woman continue speaking to the camera.



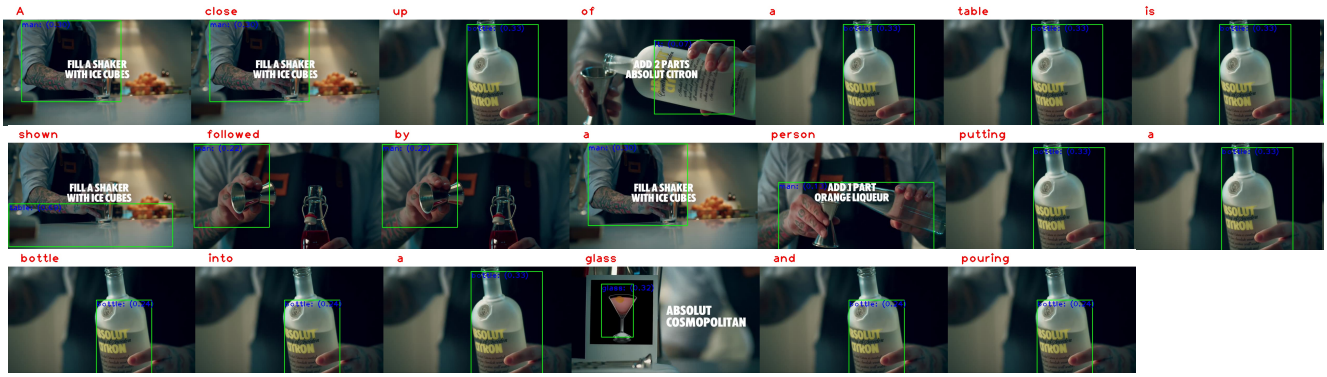
(c) **Sup.**: A man is standing in a room holding a saxophone;
Unsup.: A man is playing a saxophone;
Masked Trans.: A man is seated on a bed;
GT: We see a man playing a saxophone in front of microphones.

(d) **Sup.**: The people ride around the beach and ride around on the horses;
Unsup.: The people ride around the beach and ride around;
Masked Trans.: The camera pans around the area and the girl leading the horse and the woman leading the;
GT: We see four people on horses on the beach.

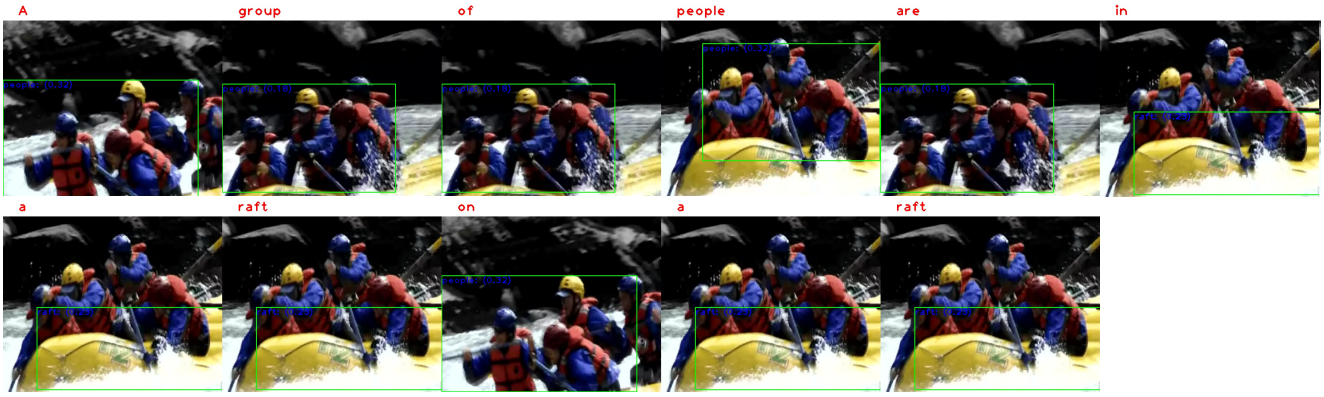
Figure 6. Qualitative results on ANet-Entities val set. The red text at each frame indicates the generated word. The green box indicates the proposal with the highest attention weight. The blue text inside the green box corresponds to i) the object class with the highest probability and ii) the attention weight. Better zoomed and viewed in color. See Sec. A.2 for discussion.



(e) **Sup.**: The woman is then seen standing in a tennis court holding tennis rackets and hitting the ball around;
Unsup.: The woman serves the ball with a tennis racket;
Masked Trans.: We see a man playing tennis in a court;
GT: Two women are on a tennis court, showing the technique to posing and hitting the ball.



(f) **Sup.**: A close up of a table is shown followed by a person putting a bottle into a glass and pouring;
Unsup.: A man is standing in a kitchen;
Masked Trans.: A person is seen speaking to the camera and leads into him pouring liquids into a glass;
GT: Several shots of alcohol as well as ingredients are shown followed by a person pouring a mixture into a glass and cutting up limes.

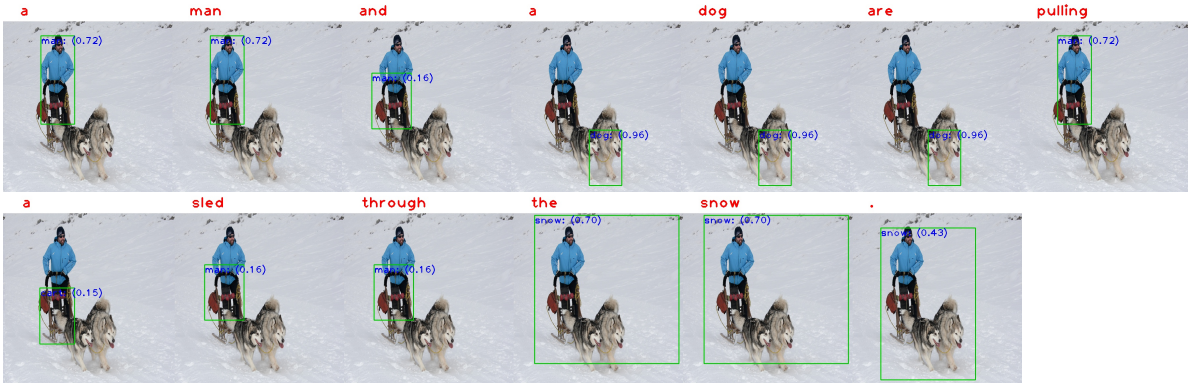


(g) **Sup.**: A group of people are in a raft on a raft;
Unsup.: A group of people are in a raft;
Masked Trans.: A group of people are in a raft and a man in red raft raft raft raft raft;
GT: People are going down a river in a raft.

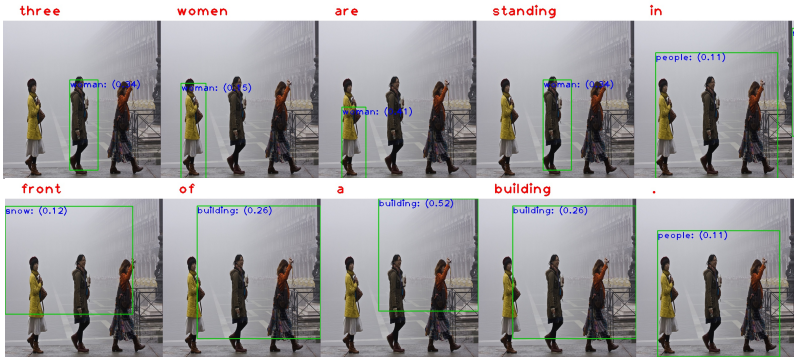
Figure 7. (Continued) Qualitative results on ANet-Entities val set. See the caption in Fig. 6 for more details.

Method	B@1	B@4	M	C	S	Attn.	Grd.	Prec.	Recall	Cls.
Unsup. (w/o SelfAttn)	70.0	27.5	22.0	60.4	15.9	22.1	26.0	14.5	14.5	18.1
Unsup.	69.3	26.8	22.1	59.4	15.7	4.10	16.4	5.63	6.00	1.40
Sup. Attn.	**71.0**	**28.2**	**22.7**	63.0	**16.7**	**43.7**	45.8	26.2	28.2	7.74
Sup. Grd.	70.1	27.6	22.5	**63.1**	16.1	38.6	49.2	26.1	27.3	0.04
Sup. Cls. (w/o SelfAttn)	70.1	27.6	22.0	60.2	15.8	20.9	32.2	13.9	14.0	**20.3**
Sup. Attn.+Grd.	70.2	27.6	22.5	62.3	**16.3**	**42.8**	49.5	**28.4**	**30.1**	0
Sup. Attn.+Cls.	70.0	27.9	22.6	62.4	**16.3**	42.1	46.6	26.9	28.3	**20.3**
Sup. Grd. +Cls.	70.4	28.0	**22.7**	62.8	**16.3**	29.0	**51.2**	24.6	25.6	20.2
Sup. Attn.+Grd.+Cls.	**70.6**	**28.1**	22.6	**63.3**	**16.3**	41.4	**50.9**	**27.9**	**29.4**	20.0

Table 9. Results on Flickr30k Entities val set. The top two scores on each metric are bold.



(a) **Sup.**: A man and a dog are pulling a sled through the snow;
Unsup.: A man in a blue jacket is pulling a dog on a sled;
BUTD: Two dogs are playing in the snow;
GT (5): A bearded man wearing a blue jacket rides his snow sled pulled by his two dogs / Man in blue coat is being pulled in a dog sled by two dogs / A man in a blue coat is propelled on his sled by two dogs / A man us using his two dogs to sled across the snow / Two Huskies pull a sled with a man in a blue jacket.



(b) **Sup.**: Three women are standing in front of a building;
Unsup.: Three women in costumes are standing on a stage with a large wall in the background;
BUTD: Three women in yellow and white dresses are walking down a street;
GT (5): Three woman are crossing the street and on is wearing a yellow coat / Three ladies enjoying a stroll on a cold, foggy day / A woman in a yellow jacket following two other women / Three women in jackets walk across the street / Three women are crossing a street.

(c) **Sup.**: A man in a gray jacket is sitting in a chair with a laptop in the background;
Unsup.: A man in a brown jacket is sitting in a chair at a table;
BUTD: A man in a brown jacket is sitting in a chair with a woman in a brown jacket in a;
GT (5): Several chairs lined against a wall, with children sitting in them / A group of children sitting in chairs with monitors over them / Children are sitting in chairs under some television sets / Pre-teen students attend a computer class / Kids conversing and learning in class.

Figure 8. Qualitative results on Flickr30k Entities val set. Better zoomed and viewed in color. See Sec. A.3 for discussion.

__background__	egg	nail	kid	snowboard	hoop	roller	pasta
bagpipe	stilt	metal	butter	cheerleader	puck	kitchen	stage
coach	paper	dog	surfboard	landscape	scene	guitar	trophy
bull	dough	tooth	object	eye	scissors	grass	stone
rod	costume	pipe	ocean	sweater	ring	drum	swimmer
disc	oven	shop	person	camera	city	accordion	stand
dish	braid	shot	edge	vehicle	horse	ramp	road
chair	pinata	kite	bottle	raft	basketball	bridge	swimming
carpet	bunch	text	camel	themselves	monkey	wall	image
animal	group	barbell	photo	calf	top	soap	playground
gymnast	harmonica	biker	polish	teen	paint	pot	brush
mower	platform	shoe	cup	door	leash	pole	female
bike	window	ground	sky	plant	store	dancer	log
curler	soccer	tire	lake	glass	beard	table	area
ingredient	coffee	title	bench	flag	gear	boat	tennis
woman	someone	winner	color	adult	shorts	bathroom	lot
string	sword	bush	pile	baby	gym	teammate	suit
wave	food	wood	location	hole	wax	instrument	opponent
gun	material	tape	ski	circle	park	blower	head
item	number	hockey	skier	word	part	beer	himself
sand	band	piano	couple	room	herself	stadium	t-shirt
saxophone	they	goalie	dart	car	chef	board	cloth
team	foot	pumpkin	sumo	athlete	target	website	line
sidewalk	silver	hip	game	blade	instruction	arena	ear
razor	bread	plate	dryer	roof	tree	referee	he
clothes	name	cube	background	cat	bed	fire	hair
bicycle	slide	beam	vacuum	wrestler	friend	worker	slope
fence	arrow	hedge	judge	closing	iron	child	potato
sign	rock	bat	lady	male	coat	bmx	bucket
jump	side	bar	furniture	dress	scuba	instructor	cake
street	everyone	artist	shoulder	court	rag	tank	piece
video	weight	bag	towel	goal	clip	hat	pin
paddle	series	she	gift	clothing	runner	rope	intro
uniform	fish	river	javelin	machine	mountain	balance	home
supplies	gymnasium	view	glove	rubik	microphone	canoe	ax
net	logo	set	rider	tile	angle	it	face
exercise	girl	frame	audience	toddler	snow	surface	pit
body	living	individual	crowd	beach	couch	player	cream
trampoline	flower	parking	people	product	equipment	cone	lemon
leg	container	racket	back	sandwich	chest	violin	floor
surfer	house	close	sponge	mat	contact	helmet	fencing
water	hill	arm	mirror	tattoo	lip	shirt	field
studio	wallpaper	reporter	diving	ladder	tool	paw	other
sink	dirt	its	slice	bumper	spectator	bowl	oar
path	toy	score	leaf	end	track	member	picture
box	cookie	finger	bottom	baton	flute	belly	frisbee
boy	guy	teens	tube	man	cigarette	vegetable	lens
stair	card	pants	ice	tomato	mouth	pan	pool
bow	yard	opening	skateboarder	neck	letter	wheel	building
credit	skateboard	screen	christmas	liquid	darts	ball	lane
smoke	thing	outfit	knife	light	pair	drink	phone
trainer	swing	toothbrush	hose	counter	knee	hand	mask
shovel	castle	news	bowling	volleyball	class	fruit	jacket
kayak	cheese	tub	diver	truck	lawn	student	stick

Table 10. List of objects in ActivityNet-Entities, including the "__background__" class.