Grounded Video Description

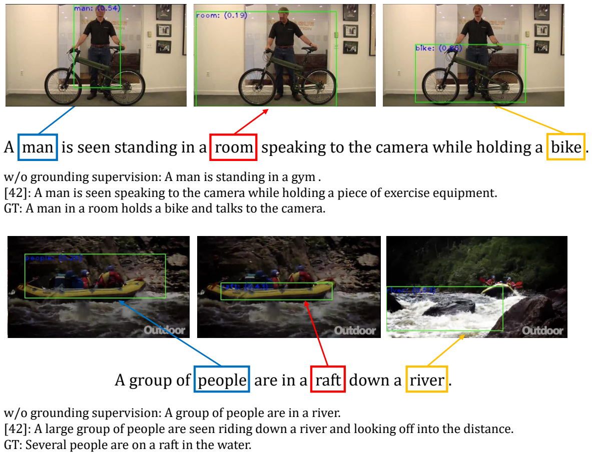
Luowei Zhou[1,2], Yannis Kalantidis[1], Xinlei Chen[1], Jason J. Corso[2], Marcus Rohrbach[1]
[1] Facebook AI, [2] University of Michigan
github.com/facebookresearch/grounded-video-description

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 dataset, ActivityNet-Entities, augments the challenging ActivityNet Captions dataset 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 dataset, but also show how it can be applied to image description on the Flickr30k Entities dataset. We achieve state-of-the-art performance on video description, video paragraph description, and image description and demonstrate our generated sentences are better grounded in the video.

1. Introduction

Image and video description models are frequently not well grounded [14] which can increase their bias [9] and lead to hallucination of objects [24], *i.e.* the model mentions objects which are not in the image or video *e.g.* because they might have appeared in similar contexts during training. This makes models less accountable and trustworthy, which is important if we hope such models will eventually assist people in need [2, 27]. Additionally, grounded models can help to explain the model's decisions to humans and allow humans to diagnose them [20]. While researchers have started to discover and study these problems for image description [14, 9, 24, 20],[1] they are even more pronounced for video description due to the increased difficulty and diversity, both on the visual and the language side.



A man is seen standing in a room speaking to the camera while holding a bike.

w/o grounding supervision: A man is standing in a gym .
[42]: A man is seen speaking to the camera while holding a piece of exercise equipment.
GT: A man in a room holds a bike and talks to the camera.

A group of people are in a raft down a river.

w/o grounding supervision: A group of people are in a river.
[42]: A large group of people are seen riding down a river and looking off into the distance.
GT: Several people are on a raft in the water.

Figure 1: Word-level grounded video descriptions generated by our model on two segments from our ActivityNet-Entities dataset. We also provide the descriptions generated by our model without explicit bounding box supervision, the descriptions generated by [42] and the ground-truth descriptions (GT) for comparison.

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 model looked at the bounding box of the man to say this word [9, 24]. 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 col-

[1] We use *description* instead of *captioning* as *captioning* is often used to refer to transcribing the speech in the video, not *describing* the content.

lect 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 [10], 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 three-fold. 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 [22] dataset.

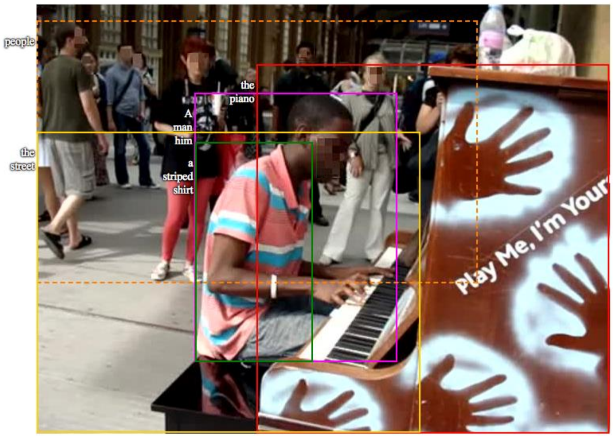
2. Related Work

Video & Image Description. Early work on automatic caption generation mainly includes template-based approaches [5, 12, 18], 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 [34] (or spatial-attention [32] in the image domain), semantic attention [13, 35, 36, 41] or both [19]. The recent unprecedented success in object detection [23, 7] has regained the community's interests on detecting fine-grained visual clues while incorporating them into end-to-end networks [16, 26, 1, 15]. Description methods which are based on object detectors [16, 38, 1, 15, 5, 12] 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 [16, 38, 1] or classify the regions into labels and fill into pre-defined/generated sentence templates [15, 5, 12]. 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 [15] 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, [26] 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, for VQA [4, 39] the authors point out that 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 [14, 37] was found to be beneficial. We noticed in our preliminary experiments that directly guiding the region attention with supervision [15] does not necessary lead to improvements in automatic sentence metrics. We hypothesize that this might be due to the lack of object context information and we thus introduce a self-attention [28] based context encoding in our attention model, which allows information 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 [22] contains such annotations for images, no large-scale description datasets with object localization annotation exists for videos. The large-scale ActivityNet Captions dataset [10] contains dense language annotations for about 20k videos from ActivityNet [3] but lacks grounding annotations. Leveraging the language annotations from



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.

Dataset	Domain	# Vid/Img	# Sent	# Obj	# BBoxes
Flickr30k Entities [22]	Image	32k	160k	480	276k
MPII-MD [26]	Video	≪1k	≪1k	4	2.6k
YouCook2 [40]	Video	2k	15k	67	135k
ActivityNet Humans [33]	Video	5.3k	30k	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 [33] dataset are both based on ActivityNet [3], but ActivityNet Humans provides boxes only for person on a small subset of videos. YouCook2 is restricted to cooking and only has box annotations for the val and the test splits.

the ActivityNet Captions dataset [10], we collected entity-level bounding box annotations and created the ActivityNet-Entities (ANet-Entities) dataset[2], 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 typically sparsely annotated at the region level [6]. 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.

Noun Phrases. Following [22], 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 sentence. We asked the annotators to identify all concrete NPs from the sentence describing the video segment 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

[2] ActivityNet-Entities is released at https://github.com/facebookresearch/ActivityNet-Entities.

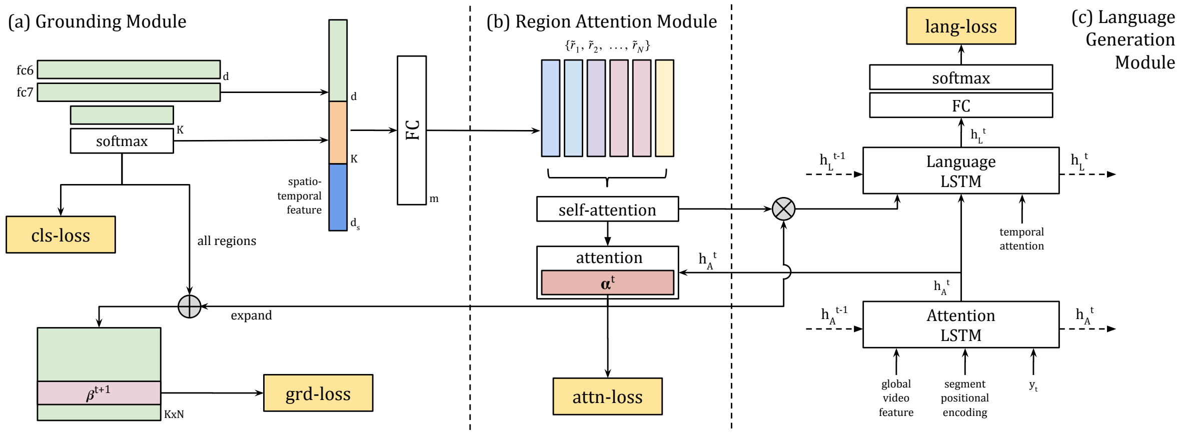


Figure 3: The proposed framework consists of three parts: the grounding module (a), the region attention module (b) and the language generation module (c). Region proposals are first represented with grounding-aware region encodings. The language model then dynamically attends on the region encodings to generate each word. 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.

chose to annotate noun phrases, in this work, we model sentence generation as a word-level task. We follow the convention in [15] 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 [17]. The frequency of these words in the train and val splits are computed and a threshold determines whether each word is an object class. For ANet-Entities, we set the frequency threshold to be 50 which produces 432 object classes.

4. Description with Grounding Supervision

In this section we describe the proposed grounded video description framework (see 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 feeds it to the language generation module for 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:

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

where L_{sent} denotes the teacher-forcing language generation cross-entropy loss, commonly used for language generation tasks (details in Sec. 4.2). L_{attn} corresponds to the cross entropy region attention loss which is presented 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. 4.4). The three grounding-related losses are weighted by coefficients λ_α, λ_c, and λ_β which we selected on the dataset validation split.

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 set 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 $N = \sum_{f=1}^{F} N_f$ is the total number of regions. We overload the notation here and use r_i ($i \in \{1, 2, \ldots, N\}$) to also represent region feature embeddings, as indicated by fc6 in Fig. 3. 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. Language Generation Module

For language generation, we adapt the language model from [15] 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 $h_A^t \in \mathbb{R}^m$ where m is the dimension and the second one for language generation (see Fig. 3c). The language model dynamically attends on videos frames or regions for visual clues to generate words. 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 [42], except that we replace the self-attention context encoder with Bi-directional GRU (Bi-GRU) which yields superior results. We train with cross-entropy loss L_{sent}.

4.3. Region Attention Module

Unlike temporal attention that works on a frame level, the region attention [1, 15] focuses on more fine-grained details in the video, *i.e.*, object regions [23]. We denote the region encoding as $\tilde{R} = [\tilde{r}_1, \tilde{r}_2, \dots, \tilde{r}_N]$, more details are defined later in Eq. 5. At time t of the caption generation, the attention weight over region i is formulated as:

$$\alpha_i^t = w_\alpha^\top \tanh(W_r \tilde{r}_i + W_h h_A^t), \quad \alpha^t := \text{Softmax}(\alpha^t), \tag{2}$$

where $W_r \in \mathbb{R}^{m \times d}$, $W_h \in \mathbb{R}^{m \times m}$, $w_\alpha \in \mathbb{R}^m$, and $\alpha^t = [\alpha_1^t, \alpha_2^t, \dots, \alpha_N^t]$. The region attention encoding is then $\tilde{R}\alpha^t$ and along with the temporal attention encoding, fed into the language LSTM.

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*. Denote the indicators of positive/negative regions as $\gamma^t = [\gamma_1^t, \gamma_2^t, \dots, \gamma_N^t]$, where $\gamma_i^t = 1$ when the region r_i has over 0.5 IoU with the GT box r_{GT} and otherwise 0. We regress α^t to γ^t and hence the attention loss for object word s_t can be defined as:

$$L_{attn} = -\sum_{i=1}^{N} \gamma_i^t \log \alpha_i^t. \tag{3}$$

4.4. Grounding Module

Assume we have a set of visually-groundable object class labels $\{c_1, c_2, \dots, c_\mathcal{K}\}$, short as object classes, where $\mathcal{K}$ is the total number of classes. Given a set 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 $W_c = [w_1, w_2, \dots, w_\mathcal{K}] \in \mathbb{R}^{d \times \mathcal{K}}$ and the learnable scalar biases as $B = [b_1, b_2, \dots, b_\mathcal{K}]$. So, a naive way to estimate the class probabilities for all regions (embeddings) $R = [r_1, r_2, \dots, r_N]$ is through dot-product:

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

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 clarity, 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 (VG) [11], 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 [21]). 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):

$$\tilde{R} = W_g[\, R \mid M_s(R) \mid M_l \,], \tag{5}$$

where $[\cdot|\cdot]$ indicates a row-wise concatenation and $W_g \in \mathbb{R}^{m \times (d + \mathcal{K} + d_s)}$ are the embedding weights. We name $\tilde{R}$ the *grounding-aware region encoding*, corresponding to the right portion of Fig. 3a. To further model the relations between regions, we deploy a self-attention layer over $\tilde{R}$ [28, 42]. The final region encoding is fed into the region attention module (see Fig. 3b).

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 the Attention LSTM. A memory-efficient approach is treating attention weights α^t as this semantic prior, as formulated below:

$$M_s^t(R, \alpha^t) = \text{Softmax}(W_c^\top R + B\mathbb{1}^\top + \mathbb{1}{\alpha^t}^\top), \tag{6}$$

where the region attention weights α^t are determined by Eq. 2. 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 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. Next, we describe them both.

Region Classification. We first define a positive region as a region that has over 0.5 intersection over union (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 loss on class c_j is

$$L_{cls} = -\log M_s[j, i]. \tag{7}$$

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, 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 indicators γ^t as defined in Sec. 4.3. The grounding loss for word s_{t+1} is defined as:

$$L_{grd} = -\sum_{i=1}^{N} \gamma_i^t \log \beta_i^{t+1}. \tag{8}$$

Note that the final loss on L_{attn} or L_{grd} is the average of losses on all visually-groundable words. The difference between the attention supervision and the grounding supervision is that, in the latter task, the target object c_j is known beforehand, while the attention module is not aware of which object to seek in the scene.

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 [10]) and video paragraph description [30]. We also demonstrate our framework can easily be applied to image description and evaluate it on the Flickr30k Entities dataset [22]. 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 [15].

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 it is a larger corpus, 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 [42]. 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 [30] with the evaluation script provided by the authors [30]. For image captioning, we compare against two SotA methods, Neural Baby Talk (NBT) [15] and BUTD [1]. 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. To measure the object grounding and attention correctness, we first compute the localization accuracy (*Grd.* and *Attn.* in the tables) over GT sentences following [25, 40]. Given an unseen video, we feed the GT sentence into the model and measure the localization accuracy at each annotated object word. We compare the region with the highest attention weight (α_i) or grounding weight (β_j) against the GT box. An object word is correctly localized if the IoU is over 0.5. We also study the attention accuracy on generated sentences, denoted by $F1_{all}$ and $F1_{loc}$ in the tables. In $F1_{all}$, a region prediction is considered correct if the object word is correctly predicted and also correctly localized. We also compute $F1_{loc}$, which only considers correctly-predicted object words. See Appendix for details. 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 all the localization accuracies. For the region classification task, we compute the top-1 classification accuracy (*Cls.* in the tables) for positive regions. For all metrics, we average the scores across object classes. To evaluate the sentence quality, we use standard language evaluation metrics, including Bleu@1, Bleu@4, METEOR, CIDEr, and SPICE, and the official evaluation script[3]. We additionally perform human evaluation to judge the sentence quality.

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 R-CNN detector [23] with ResNeXt-101 backbone [29] for region proposal and feature extraction (fc6). The detector is pretrained on Visual Genome [11]. 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 [42, 31]. The spatial feature map is the conv4 layer output from a ResNet-101 [15, 8] model. 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. Coefficients $\lambda_\alpha \in \{0.05, 0.1, 0.5\}$, $\lambda_\beta \in \{0.05, 0.1, 0.5\}$, 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 $e = 512$

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

Method	λ_α	λ_β	λ_c	B@1	B@4	M	C	S	Attn.	Grd.	F1$_{all}$	F1$_{loc}$	Cls.
Unsup. (w/o SelfAttn)	0	0	0	23.2	2.28	10.9	45.6	**15.0**	14.9	21.3	3.70	12.7	6.89
Unsup.	0	0	0	23.0	2.27	10.7	44.6	13.8	2.42	19.7	0.28	1.13	6.06
Sup. Attn.	0.05	0	0	23.7	2.56	**11.1**	47.0	14.9	34.0	37.5	6.72	22.7	0.42
Sup. Grd.	0	0.5	0	23.5	2.50	11.0	46.8	14.7	31.9	43.2	6.04	21.2	0.07
Sup. Cls.	0	0	0.1	23.3	2.43	10.9	45.7	14.1	2.59	25.8	0.35	1.43	**14.9**
Sup. Attn.+Grd.	0.5	0.5	0	**23.8**	2.44	**11.1**	46.1	14.8	**35.1**	40.6	6.79	23.0	0
Sup. Attn.+Cls.	0.05	0	0.1	**23.9**	**2.59**	**11.2**	**47.5**	**15.1**	34.5	41.6	**7.11**	**24.1**	**14.2**
Sup. Grd. +Cls.	0	0.05	0.1	**23.8**	**2.59**	**11.1**	**47.5**	**15.0**	27.1	**45.7**	4.79	17.6	13.8
Sup. Attn.+Grd.+Cls.	0.1	0.1	0.1	**23.8**	2.57	**11.1**	46.9	**15.0**	**35.7**	**44.9**	**7.10**	**23.8**	12.2

Table 2: Results on ANet-Entities val set. "w/o SelfAttn" indicates self-attention is not used for region feature encoding. Notations: B@1 - Bleu@1, B@4 - Bleu@4, M - METEOR, C - CIDEr, S - SPICE. Attn. and Grd. are the object localization accuracies for attention and grounding on GT sentences. F1$_{all}$ and F1$_{loc}$ are the object localization accuracies for attention on generated sentences. Cls. is classification accuracy. All accuracies are in %. Top two scores on each metric are in bold.

Method	B@1	B@4	M	C	S	Attn.	Grd.	F1$_{all}$	F1$_{loc}$	Cls.
Masked Transformer [42]	22.9	**2.41**	10.6	**46.1**	13.7	–	–	–	–	–
Bi-LSTM+TempoAttn [42]	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**	16.1	22.3	3.73	11.7	6.41
Our Sup. Attn.+Cls. (GVD)	**23.6**	2.35	**11.0**	45.5	14.7	**34.7**	**43.5**	**7.59**	**25.0**	**14.5**

(a) Results on ANet-Entities test set.

	vs. Unsupervised		vs. [42]	
	Judgments		Judgments	
Method	%	Δ	%	Δ
About Equal	34.9		38.9	
Other is better	29.3	6.5	27.5	6.1
GVD is better	**35.8**		**33.6**	

(b) Human evaluation of sentences.

Table 3: (a) Results on ANet-Entities test set. The top one score for each metric is in bold. (b) Human evaluation of sentence quality. We present results for our supervised approach vs. our unsupervised baseline and vs. Masked Transformer [42].

and RNN encoding size $m = 1024$ for all methods. Other hyper-parameters in the language module are the same as in [15]. We use a 2-layer 6-head Transformer encoder as the self-attention module [42].

5.3. Results on ActivityNet-Entities

5.3.1 Video Event Description

Although dense video description [11] 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 are given. 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. 3a, 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. Attn.+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 ± 2%), but the supervised method has a 8.8% relative improvement on Bleu@4.

The results in Tab. 3a show that adding box supervision dramatically improves the grounding accuracy from 22.3% to 43.5%. 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 on both GT and generated sentences, implying that the supervised models learn to attend on more relevant objects during language generation. 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 refer to our best model (Sup. Attn.+Cls.) as GVD (Grounded Visual Description) and show that it sets the new SotA on ActivityNet Captions for the Bleu@1, METEOR and SPICE metrics, with relative gains of 2.8%, 3.9% and 6.8%, respectively over the previous best [42]. We observe 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 [42] 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

Method	VG	Box	B@1	B@4	M	C	S	Attn.	Grd.	F1$_{all}$	F1$_{loc}$	Cls.
ATT-FCN* [36]			64.7	19.9	18.5	–	–	–	–	–	–	–
NBT* [15]		✓	69.0	27.1	21.7	57.5	15.6	–	–	–	–	–
BUTD [1]	✓		69.4	**27.3**	21.7	56.6	16.0	24.2	32.3	4.53	13.0	1.84
Our Unsup. (w/o SelfAttn)	✓		69.2	26.9	22.1	60.1	16.1	21.4	25.5	3.88	11.7	17.9
Our GVD model	✓	✓	**69.9**	**27.3**	**22.5**	**62.3**	**16.5**	**41.4**	**50.9**	**7.55**	**22.2**	**19.2**

Table 4: Results on Flickr30k Entities test set. * indicates the results are obtained from the original papers. GVD refers to our Sup. Attn.+Grd.+Cls. model. "VG" indicates region features are from VG pre-training. The top one score is in bold.

Method	B@1	B@4	M	C
MFT [30]	45.5	9.78	14.6	20.4
Our Unsup. (w/o SelfAttn)	49.8	10.5	15.6	21.6
Our GVD	**49.9**	**10.7**	**16.1**	**22.2**

Table 5: Results of video paragraph description on test set.

to properly attend to the right region, something that leads to a huge attention accuracy drop from 14.9% to 2.42%.

Human Evaluation. Automatic metrics for evaluating generated sentences have frequently shown to be unreliable and not consistent with human judgments, especially for video description when there is only a single reference [27]. Hence, we conducted a human evaluation to evaluate the sentence quality on the test set of ActivityNet-Entities. We randomly sampled 329 video segments and presented the segments and descriptions to the judges. From Tab. 3b, we observe that, while they frequently produce captions with similar quality, our GVD works better than the unsupervised baseline (with a significant gap of 6.1%). We can also see that our GVD approach works better than the Masked Transformer [42] 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 [42].

5.3.2 Video Paragraph Description

Besides measuring the quality of each individual description, we also evaluate the coherence among sentences within a video as in [30]. We obtained the result file and evaluation script from [30] and evaluated both methods on *our* test split. The results are shown in Tab. 5 and show that we outperform the SotA method of [30] by a large margin. 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 context benefits the coherence of consecutive sentences.

5.4. Results on Flickr30k Entities

We show the overall results on image description in Tab. 4 (test) and the results on the validation set are in the Appendix. The method with the best validation CIDEr score is the full model (Sup. Attn.+Grd.+Cls.), which we further refer to as the GVD model in the table. The upper bounds on the val/test sets are 90.0%/88.5%, respectively. We see that the supervised method outperforms the unsupervised baseline by a relative 1-3.7% over all the metrics. Our GVD model sets new SotA for all the five metrics with relative gains up to 10%. In the meantime, object localization and region classification accuracies are significantly boosted, showing that our captions can be better visually explained and understood.

6. Conclusion

In this work, we collected ActivityNet-Entities, a novel dataset that allows joint study of video description and grounding. 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 [24] and bias [9] 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 description by not only generating sentences but also pointing to the corresponding regions in the video frames or image. According to automatic metrics and human evaluation, on ActivityNet-Entities our model performs state-of-the-art w.r.t. description quality, both when evaluated per sentence or on paragraph level with a significant increase in grounding performance. We also adapted our model to image description and evaluated it on the Flickr30k Entities dataset where our model outperforms existing methods, both w.r.t. description quality and grounding accuracy.

Acknowledgement. The technical work was performed during Luowei's summer internship at Facebook AI Research. Luowei Zhou and Jason Corso were partly supported by DARPA FA8750-17-2-0125 and NSF IIS 1522904 as part of their affiliation with University of Michigan. This article solely reflects the opinions and conclusions of its authors but not the DARPA or NSF.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 2, 5, 6, 8

[2] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom 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

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan 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. 2, 3

[4] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv 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

[5] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason 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

[6] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia 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

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[9] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision*, pages 771–787, 2018. 1, 8

[10] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 706–715, 2017. 2, 3, 6

[11] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David 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. 5, 6, 7

[12] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013. 2

[13] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao 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

[14] Chenxi Liu, Junhua Mao, Fei Sha, and Alan L Yuille. Attention correctness in neural image captioning. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pages 4176–4182, 2017. 1, 2

[15] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi 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, 8

[16] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6790–6800, 2018. 2

[17] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David 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

[18] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal 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

[19] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 3, 2017. 2

[20] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[21] Jeffrey Pennington, Richard Socher, and Christopher 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. 5

[22] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Pro-*

ceedings of the IEEE international conference on computer vision, pages 2641–2649, 2015. 2, 3, 6

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian 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, 5, 6

[24] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018. 1, 8

[25] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016. 6

[26] Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, and Bernt 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

[27] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision (IJCV)*, 2017. 1, 8

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 2, 5

[29] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming 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

[30] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. *Proceedings of the European Conference on Computer Vision*, 2018. 6, 8

[31] Yuanjun Xiong, Limin Wang, Zhe Wang, Bowen Zhang, Hang Song, Wei Li, Dahua Lin, Yu Qiao, Luc Van Gool, and Xiaoou Tang. Cuhk & ethz & siat submission to activitynet challenge 2016. *arXiv preprint arXiv:1608.00797*, 2016. 6

[32] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 2

[33] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1453–1462, 2017. 3

[34] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015. 2

[35] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *IEEE International Conference on Computer Vision, ICCV*, pages 22–29, 2017. 2

[36] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 2, 8

[37] Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee 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

[38] Mihai Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Spatio-temporal attention models for grounded video captioning. In *Asian Conference on Computer Vision*, pages 104–119, 2016. 2

[39] Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 349–357. IEEE, 2019. 2

[40] Luowei Zhou, Nathan Louis, and Jason J Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 3, 6

[41] Luowei Zhou, Chenliang Xu, Parker Koch, and Jason 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

[42] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming 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