Adversarial Inference for Multi-Sentence Video Description

Jae Sung Park[1], Marcus Rohrbach[2], Trevor Darrell[1], Anna Rohrbach[1]
[1] University of California, Berkeley, [2] Facebook AI Research

Abstract

While significant progress has been made in the image captioning task, video description is still in its infancy due to the complex nature of video data. Generating multi-sentence descriptions for long videos is even more challenging. Among the main issues are the fluency and coherence of the generated descriptions, and their relevance to the video. Recently, reinforcement and adversarial learning based methods have been explored to improve the image captioning models; however, both types of methods suffer from a number of issues, e.g. *poor readability and high redundancy for RL and stability issues for GANs. In this work, we instead propose to apply adversarial techniques during inference, designing a discriminator which encourages better multi-sentence video description. In addition, we find that a multi-discriminator "hybrid" design, where each discriminator targets one aspect of a description, leads to the best results. Specifically, we decouple the discriminator to evaluate on three criteria: 1) visual relevance to the video, 2) language diversity and fluency, and 3) coherence across sentences. Our approach results in more accurate, diverse, and coherent multi-sentence video descriptions, as shown by automatic as well as human evaluation on the popular ActivityNet Captions dataset.*

Transformer: A man is seen riding on a board with a **kite on a board.** The people are seen *riding around* on *the water while the camera* follows movements. The people continue *riding around the water while the camera* captures them from the angles.

VideoStory: A person is seen riding a board **on a board** and begins moving *along the water*. The person continues riding *along the water* and ends by several more people riding **along** the board. The camera pans around the water and ends with one another person on the board.

MoveForwardTell: A large group of people are seen riding *along the water* **on the water.** A person is seen riding *on the water* and moving *along the water*. A person is seen **speaking to the camera** and leads into him riding around *on the water.*

Our *Adversarial Inference*: A large group of people are seen standing on a large field with one another and leads into them riding around on a large <u>body of water</u>. The person is <u>parasailing</u> on the water. The person continues riding along the water as well as the camera panning around.

Ground Truth: A group is standing on the sand and waves at the camera. They are shown parasailing in the ocean water. They take turns, several people floating on the water.

Figure 1: Comparison of the state-of-the-art video description approaches, Transformer [76], VideoStory [13], MoveForwardTell [67], and our proposed *Adversarial Inference*. Our approach generates more interesting and accurate descriptions with less redundancy. Video from ActivityNet Captions [3, 29] with three segments (left to right); red/bold indicates content errors, blue/italic indicates repetitive patterns, underscore highlights more interesting phrases.

1. Introduction

Being able to automatically generate a natural language description for a video has fascinated researchers since the early 2000s [27]. Despite the high interest in this task and ongoing emergence of new datasets [13, 29, 75] and approaches [67, 69, 76], it remains a highly challenging problem. Consider the outputs of the three recent video description methods on an example video from the ActivityNet Captions dataset [3, 29] in Figure 1. We notice that there are multiple issues with these descriptions, in addition to the errors with respect to the video content: there are semantic inconsistencies and lack of diversity within sentences, as well as redundancies across sentences. There are multiple challenges towards more accurate and natural video description. One of the issues is the size of the available training data, which, despite the recent progress, is limited. Besides, video representations are more complex than *e.g.* image representations, and require modeling temporal structure jointly with the semantics of the content. Moreover, describing videos with multiple sentences, requires correctly recognizing a sequence of events in a video, maintaining linguistic coherence and avoiding redundancy.

Another important factor is the target metric used in the description models. Most works still exclusively rely on the automatic metrics, e.g. METEOR [31], despite the evidence that they *are not consistent* with human judgments [24, 57]. Further, some recent works propose to explicitly optimize for the sentence metrics using reinforcement learning based methods [35, 46]. These techniques have become quite widespread, both for image and video description [1, 67]. Despite getting higher scores, reinforcement learning based methods have been shown to lead to *unwanted artifacts*, such as ungrammatical sentence endings [15], increased object hallucination rates [47] and lack of diverse content [36]. Overall, while informative, sentence metrics should not be the only way of evaluating the description approaches.

Some works aim to overcome this issue by using the adversarial learning [9, 53]. While Generative Adversarial Networks [14] have achieved impressive results for image and even video generation [21, 43, 63, 77], their success in language generation has been limited [55, 71]. The main issue is the difficulty of achieving stable training due to the discrete output space [4, 5]. Another reported issue is lack of coherence, especially for long text generation [20]. Still, the idea of *learning* to distinguish the "good" natural descriptions from the "bad" fake ones, is very compelling.

Rather than learning with adversarial training, we propose a simpler approach, *Adversarial Inference* for video description, which relies on a discriminator to improve the description quality. Specifically, we are interested in the task of multi-sentence video description [48, 70], *i.e.* the output of our model is a paragraph that describes a video. We assume that the ground-truth temporal segments are given, *i.e.* we do not address the event detection task, but focus on obtaining a coherent multi-sentence description. We first design a strong baseline generator model trained with the maximum likelihood objective, which relies on a previous sentence as context, similar to [13, 67]. We also introduce object-level features in the form of object detections [1] to better represent people and objects in video. We then make the following contributions:

(1) We propose the *Adversarial Inference* for video description, where we progressively sample sentence candidates for each clip, and select the best ones based on a discriminator's score. Prior work has explored sampling with log probabilities [12], while we show that a specifically trained discriminator leads to better results in terms of correctness, coherence, and diversity (see Figure 1).

(2) Specifically, we propose the "hybrid discriminator", which combines three specialized discriminators: one measures the language characteristics of a sentence, the second assesses its relevance to a video segment, and the third measures its coherence with the previous sentence. Prior work has considered a "single discriminator" for adversarial training to capture both the linguistic characteristics and visual relevance [53, 9]. We show that our "hybrid discriminator" outperforms the "single discriminator" design.

(3) We compare our proposed approach to multiple baselines on a number of metrics, including automatic sentence scores, diversity and repetition scores, person correctness scores, and, most importantly, human judgments. We show that our *Adversarial Inference* approach leads to more accurate and diverse multi-sentence descriptions, outperforming GAN and RL based approaches in a human evaluation.

2. Related Work

We review existing approaches to video description, including recent work based on reinforcement and adversarial learning. We then discuss related works that also sample and re-score sentence descriptions, and some that aim to design alternatives to automatic evaluation metrics.

Video description. Over the past years there has been an increased interest in video description generation, notably with the broader adoption of the deep learning techniques. S2VT [58] was among the first approaches based on LSTMs [19, 11]; some of the later ones include [38, 49, 52, 68, 72, 73]. Most recently, a number of approaches to video description have been proposed, such as replacing LSTM with a Transformer Network [76], introducing a reconstruction objective [59], using bidirectional attention fusion for context modeling [61], and others [7, 13, 33].

While most works focus on "video in - one sentence out" task, some aim to generate a multi-sentence paragraph for a video [48, 54, 70]. Recently, [69] propose a fine-grained video captioning model for generating detailed sports narratives, and [67] propose the Move Forward and Tell approach, which localizes events and progressively decides when to generate the next sentence. This is related to the task of dense captioning [29], where videos are annotated with multiple localized sentences but the task does not require to produce a single coherent paragraph for the video.

Reinforcement learning for caption generation. Most deep language generation models rely on Cross-Entropy loss and during training are given a previous ground-truth word. This is known to cause an exposure bias [42], as at test time the models need to condition on the predicted words. To overcome this issue, a number of reinforcement learning (RL) actor-critic [28] approaches have been proposed [45, 46, 74]. [35] propose a policy gradient optimization method to directly optimize for language metrics, like CIDEr [57], using Monte Carlo rollouts. [46] propose a Self-Critical Sequence Training (SCST) method based on REINFORCE [66], and instead of estimating a baseline, use the test-time inference algorithm (greedy decoding).

Recent works adopt similar techniques to video description. [40] extend the approach of [42] by using a mixed loss (both cross-entropy and RL) and correcting CIDEr with an entailment penalty. [65] propose a hierarchical reinforcement learning approach, where a Manager generates subgoals, a Worker performs low-level actions, and a Critic determines whether the goal is achieved. Finally, [32] propose a multitask RL approach, built off [46], with an additional attribute prediction loss.

GANs for caption generation. Instead of optimizing for hand-designed metrics, some recent works aim to learn what the "good" captions should be like using adversarial training. The first works to apply Generative Adversarial Networks (GANs) [14] to image captioning are [53] and [9]. [53] train a discriminator to distinguish natural human captions from fake generated captions, focusing on caption diversity and image relevance. To sample captions they rely on Gumbel-Softmax approximation [22]. [9] instead rely on policy gradient, and their discriminator focuses on caption naturalness and image relevance. Some works have applied adversarial learning to generate paragraph descriptions for images/image sequences. [34] propose a joint training approach which incorporates multi-level adversarial discriminators, one for sentence level and another for coherent topic transition at a paragraph level. [64] rely on adversarial reward learning to train a visual storytelling policy. [60] use a multi-modal discriminator and a paragraph level language-style discriminator for their adversarial training. Their multi-modal discriminator resembles the standard discriminator design of [9, 53]. In contrast, we decouple the multi-modal discriminator into two specialized discriminators, Visual and Language, and use a Pairwise discriminator for sentence pairs' coherence. Importantly, none of these works rely on their trained discriminators during inference.

Two recent image captioning works propose using discriminator scores instead of language metrics in the SCST model [6, 36]. We implement a GAN baseline based on this idea, and compare it to our approach.

Caption sampling and re-scoring. A few prior works explore caption sampling and re-scoring during inference [2, 18, 56]. Specifically, [18] aim to obtain more image-grounded bird explanations, while [2, 56] aim to generate discriminative captions for a given distractor image. While our approach is similar, our goal is different, as we work with video rather than images, and aim to improve multi-sentence description with respect to multiple properties.

Alternatives to automatic metrics. There is a growing interest in alternative ways of measuring the description quality, than *e.g*. [39, 31, 57]. [8] train a general critic network to learn to score captions, providing various types of corrupted captions as negatives. [51] use a composite metric, a classifier trained on the automatic scores as input. In contrast, we do not aim to build a general evaluation tool, but propose to improve the video description quality with our Adversarial Inference for a given generator.

3. Generation with Adversarial Inference

In this section, we present our approach to multi-sentence description generation based on our *Adversarial Inference* method. We first introduce our baseline generator G and then discuss our discriminator D. The task of D is to score the descriptions generated by G for a given video. This includes, among others, to measure whether the multi-sentence descriptions are (1) correct with respect to the video, (2) fluent within individual sentences, and (3) form a coherent story across sentences. Instead of assigning all three tasks to a *single* discriminator, we propose to compose D out of three separate discriminators, each focusing on one of the above tasks. We denote this design a *hybrid* discriminator (see Figure 3).

While prior works mostly rely on discriminators for joint adversarial training [9, 53], we argue that using them during inference is a more robust way of improving over the original generator. In our *Adversarial Inference*, the pre-trained generator G presents D with the sentence candidates by sampling from its probability distribution. In its turn, our *hybrid* discriminator D selects the best sentence relying on the combination of its sub-discriminators. The overview of our approach is shown in Figure 2.

3.1. Baseline Multi-Sentence Generator: G

Given L clips $[v^1, v^2, ..., v^L]$ from a video v, the task of G is to generate L sentences $[s^1, s^2, ..., s^L]$, where each sentence s^i matches the content of the corresponding clip v^i. As the clips belong to the same video and are thus contextually dependent, our goal is to not only generate a sentence that matches its visual content, but to obtain a coherent and diverse sequence of sentences, *i.e*. a natural paragraph.

Our generator follows a standard LSTM decoder [11, 19] to generate individual sentences s^i with encoded representation of v^i as our visual context. Typically, for each step m, the LSTM hidden state h^i_m expects an input vector that encodes the visual features from v^i as well as the previous word w^i_{m-1}. For our visual context, we use motion, RGB images, and object detections as features for each video clip, and follow the settings from [62, 67] to obtain a single vector representation of each feature using a temporal attention mechanism [68][1]. The three vectors are concatenated to get the visual input $\bar{v}^i_m$. To encourage coherence among consecutive sentences, we additionally append the last hidden state of the previous sentence h^{i-1} as input to the LSTM decoder [13, 67]. The final input to the LSTM decoder for clip v^i at time step m is defined as follows:

$$\begin{aligned} h^i_m = LSTM(\bar{v}^i_m, w^i_{m-1}, h^{i-1}), \\ \text{with} \quad h^0 = 0, \end{aligned} \tag{1}$$

We follow the standard Maximum Likelihood Estimation (MLE) training for G, *i.e*. we maximize the likelihood of each word w^i_m given the current LSTM hidden state h^i_m.

3.2. Discriminator: D

The task of a discriminator D is to score a sentence s w.r.t. a video v as $D(s|v) \in (0, 1)$, where 1 indicates a positive match, while 0 is a negative match. Most prior

[1] For details, please, see the supplemental material.

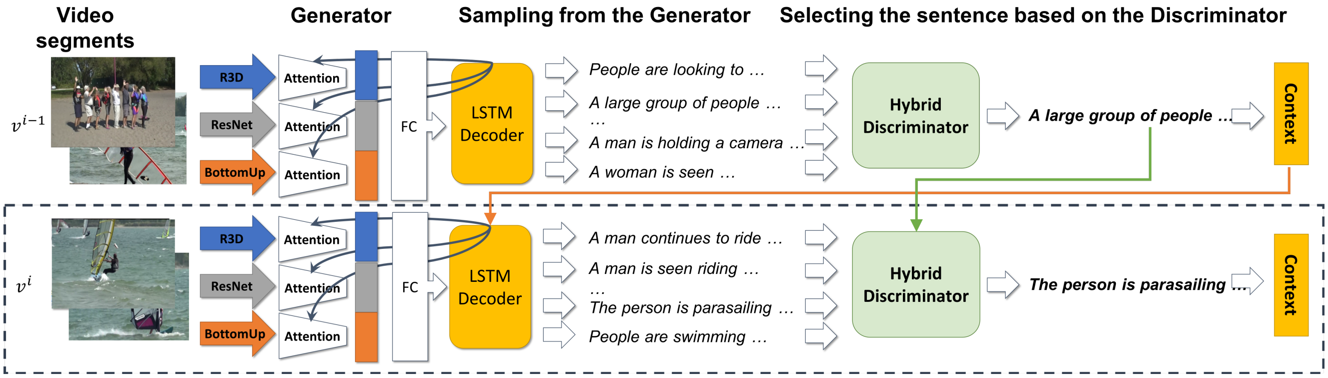


Figure 2: The overview of our Adversarial Inference approach. The Generator progressively samples candidate sentences for each clip, using the previous sentence as context. The Hybrid Discriminator scores the candidate sentences, and chooses the best one based on its visual relevance, linguistic characteristics and coherence to the previous sentence (details in Figure 3).

works that perform adversarial training for image captioning [6, 9, 36, 53], rely on the following "single discriminator" design. D is trained to distinguish human ground-truth sentences as positives vs. sentences generated by G and mismatched ground truth sentences (from a different video) as negatives. The latter aim to direct the discriminator's attention to the sentences' visual relevance.

For a given generator G, the discriminator D is trained with the following objective:

$$\max \frac{1}{N} \sum_{j=1}^{N} L_D(v^j), \tag{2}$$

where N is the number of training videos. For a video v^j a respective term is defined as:

$$\begin{aligned} L_D(v^j) = \mathbb{E}_{s \in S_{v^j}} [\log(D(s|v^j))] \quad + \\ \mu \cdot \mathbb{E}_{s \in S_G} [\log(1 - D(s|v^j))] \quad + \\ \nu \cdot \mathbb{E}_{s \in S_{\backslash v^j}} [\log(1 - D(s|v^j))], \end{aligned} \tag{3}$$

where S_{v^j} is the set of ground truth descriptions for v^j, S_G are generated samples from G, $S_{\backslash v^j}$ are ground truth descriptions from *other* videos, μ, ν are hyper-parameters.

3.2.1 Hybrid Discriminator

In the "single discriminator" design, the discriminator is given multiple tasks at once, *i.e.* to detect generated "fakes", which requires looking at linguistic characteristics, such as diversity or language structure, as well the mismatched "fakes", which requires looking at sentence semantics and relate it to the visual features. Moreover, for multi-sentence description, we would also like to detect cases where a sentence is inconsistent or redundant to a previous sentence.

To obtain these properties, we argue it is important to decouple the different tasks and allocate an individual discriminator for each one. In the following we introduce our visual, language and pairwise discriminators, which jointly constitute our *hybrid discriminator* (see Figure 3). We use the objective defined above for all three, however, the types of negatives vary by discriminator.

Visual Discriminator. The v isual discriminator D_V determines whether a sentence s^i refers to concepts present in a video clip v^i, regardless of fluency and grammatical structure of the sentence. We believe that as the pre-trained generator already produces video relevant sentences, we should not include the generated samples as negatives for D_V. Instead, we use the mismatched ground truth as well as mismatched generated sentences as our two types of negatives. While randomly mismatched negatives may be easier to distinguish, hard negatives, *e.g.* sentences from videos with the same activity as a given video, require stronger visual discriminative abilities. To improve our discriminator, we introduce such hard negatives, after training D_V for 2 epochs.

Note, that if we use an LSTM to encode our sentence inputs to D_V, it may exploit the language characteristics to distinguish the generated mismatched sentences, instead of looking at their semantics. To mitigate this issue, we replace the LSTM encoding with a bag of words (BOW) representation, *i.e.* each sentence is represented as a vocabulary-sized binary vector. The BOW is further embedded via a linear layer, and thus we obtain our final sentence encoding ω^i.

Similar to G, D_V also considers multiple visual features, *i.e.* we aggregate features from different misaligned modalities (video, image, objects). We individually encode each feature f using temporal attention based on the entire sentence representation ω^i. The obtained vector representations $\hat{v}^i_f$ are then fused with the sentence representation ω^i, using Multimodal Low-rank Bilinear pooling (MLB) [25], which is known to be effective in tasks like multi-modal retrieval or VQA. The score for visual feature f and sentence

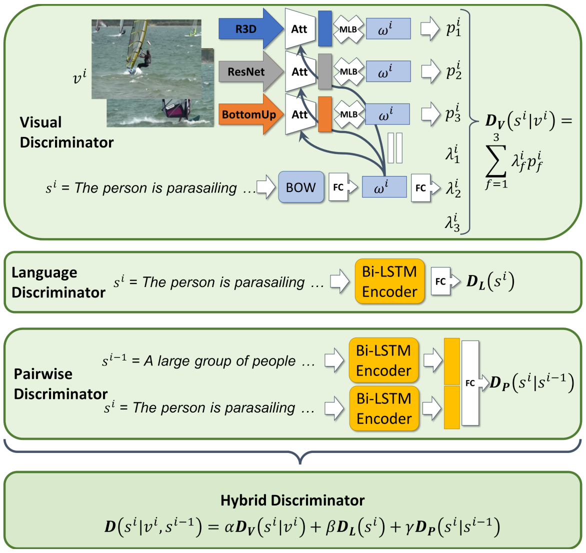


Figure 3: An overview of our Hybrid Discriminator. We score a sentence s^i for a given video clip v^i and a previous sentence s^{i-1}.

representation ω^i is obtained as follows:

$$p_f^i = \sigma(\tanh(U^T \hat{v}_f^i) \odot \tanh(V^T \omega^i)), \tag{4}$$

where σ is a sigmoid, producing values in $(0, 1)$, $\odot$ is the Hadamard product, U, V are linear layers. Instead of concatenating features $\hat{v}_f^i$ as done in the generator, here we determine the scores p_f^i between the sentence and each modality, and learn to weigh them adaptively based on the sentence. The intuition is that some sentences are more likely to require video features ("a man is jumping"), while others may require *e.g.* object features ("a man is wearing a red shirt"). Following [37], we assign weights λ_f^i to each modality based on the sentence representation ω^i:

$$\lambda_f^i = \frac{e^{a_f^T \omega^i}}{\sum_j e^{a_j^T \omega^i}}, \tag{5}$$

where a_j are learned parameters. Finally, the D_V score is the sum of the scores p_f^i weighted by λ_f^i:

$$D_V(s^i|v^i) = \sum_f \lambda_f^i p_f^i. \tag{6}$$

Language Discriminator. Language discriminator D_L focuses on language structure of an individual sentence s^i, independent of its visual relevance. Here we want to ensure fluency as well as diversity of sentence structure that is lacking in G. The ActivityNet Captions [29] dataset, that we experiment with, has long (over 13 words on average) and diverse descriptions with varied grammatical structures. In initial experiments we observed that a simple discriminator is able to point out a obvious mismatches based on diversity of the real vs. fake sentences, but fails to capture fluency or repeating N-grams. To address this, in addition to generated sentences from G, D_L is given negative inputs with a mixture of randomly shuffled words or with repeated phrases within a sentence.

To obtain a D_L score, we encode a sentence s^i with a bidirectional LSTM, concatenate both last hidden states, denoted as $\bar{h}^i$, followed by a fully connected layer and a sigmoid layer:

$$D_L(s^i) = \sigma(W_L \bar{h}^i + b_L). \tag{7}$$

Pairwise Discriminator. Pairwise discriminator D_P evaluates whether two consecutive sentences s^{i-1} and s^i are coherent yet diverse in content. Specifically, D_P scores s^i based on s^{i-1}. To ensure coherence, we include "shuffled" sentences as negatives, *i.e.* the order of sentences in a paragraph is randomly changed. We also design negatives with a pair of identical sentences ($s^i = s^{i-1}$) and optionally cutting off the endings (*e.g.* "a person enters and takes a chair" and "a person enters") to avoid repeating contents.

Similar to D_L above, we encode both sentences with a bidirectional LSTM and obtain $\bar{h}^{i-1}$ and $\bar{h}^i$. We concatenate the two vectors and compute the D_P score as follows:

$$D_P(s^i|s^{i-1}) = \sigma(W_P[\bar{h}^{i-1}, \bar{h}^i] + b_P). \tag{8}$$

Note, that the first sentence of a video description paragraph is not assigned a pairwise score, as there is no previous sentence.

3.3. Adversarial Inference

In adversarial training for caption generation, G and D are first pre-trained and then jointly updated, where the discriminator improves the generator by providing feedback to the quality of sampled sentences. To deal with the issue of non-differentiable discrete sampling in joint training, several solutions have been proposed, such as Reinforcement Learning with variants of policy gradient methods or Gumbel softmax relaxation [6, 9, 53]. While certain improvement has been shown, as we discussed in Section 1, GAN training can be very unstable.

Motivated by the difficulties of joint training, we present our *Adversarial Inference* method, which uses the discriminator D during inference of the generator G. We show that our approach outperforms a jointly trained GAN model, most importantly, in human evaluation (see Section 4).

During inference, the generator typically uses greedy max decoding or beam search to generate a sentence based on the maximum probability of each word. One alternative to this is sampling sentences based on log probability [12]. Instead, we use our Hybrid Discriminator to score the sampled sentences. Note, that we generate sentences *progressively*, *i.e.* we provide the hidden state representation of the previous best sentence as context to sample the next sentence (see Figure 2). Formally, for a video clip v^i, a previous best sentence s_*^{i-1} and K sampled sentences $s_1^i, s_2^i, ...s_K^i$ from the generator G, the scores from our hybrid discriminator can be used to compare the sentences and select the best one:

$$s_*^i = s^i_{\text{argmax}_{j=1..K} D(s_j^i|v^i, s_*^{i-1})}, \tag{9}$$

where s_j^i is the j^{th} sampled sentence. The final discriminator score is defined as:

$$\begin{aligned} D(s_j^i|v^i, s_*^{i-1}) = \alpha \cdot D_V(s_j^i|v^i) \quad + \\ \beta \cdot D_L(s_j^i) + \gamma \cdot D_P(s_j^i|s_*^{i-1}), \end{aligned} \tag{10}$$

where α, β, γ are hyper-parameters.

4. Experiments

We benchmark our approach for multi-sentence video description on the ActivityNet Captions dataset [29] and compare our Adversarial Inference to GAN and other baselines, as well as to state-of-the-art models.

4.1. Experimental Setup

Dataset. The ActivityNet Captions dataset contains 10,009 videos for training and 4,917 videos for validation with two reference descriptions for each[2]. Similar to prior work [76, 13], we use the validation videos with the 2nd reference for development, while the 1st reference is used for evaluation. While the original task defined on ActivityNet Captions involves both event localization and description, we run our experiments with ground truth video intervals. Our goal is to show that our approach leads to more correct, diverse and coherent multi-sentence video descriptions.

Visual Processing. Each video clip is encoded with 2048-dim ResNet-152 features [17] pre-trained on ImageNet [10] (denoted as ResNet) and 8192-dim ResNext-101 features [16] pre-trained on the Kinetics dataset [23] (denoted as R3D). We extract both ResNet and R3D features at every 16 frames and use a temporal resolution of 16 frames for R3D. The features are uniformly divided into 10 segments as in [62, 67], and mean pooled within each segment to represent the clip as 10 sequential features. We also run the Faster R-CNN detector [44] from [1] trained on Visual Genome [30], on 3 frames (at the beginning, middle and end of a clip) and detect top 16 objects per frame.We encode the predicted object labels with bag of words weighted by detection confidences (denoted as BottomUp). Thus, a visual representation for each clip consists of 10 R3D features, 10 ResNet features, and 3 BottomUp features.

Language Processing. The sentences are "cut" at a maximum length of 30 words. The LSTM cells' dimensionality is fixed to 512. The discriminators' word embeddings are initialized with 300-dim Glove embeddings [41].

Training and Inference. We train the generator and discriminators with cross entropy objectives using the ADAM optimizer [26] with a learning rate of $5e^{-4}$. One batch consists of multiple clips and captions from the same video, and the batch size is fixed to 16 when training all models. The weights for all the discriminators' negative inputs (μ, ν in Eq. 3), are set to 0.5. The weights for our hybrid discriminator are set as α = 0.8, β = 0.2, γ = 1.0. Sampling temperature during discriminator training is 1.0; during inference we sample $K = 100$ sentences with temperature 0.2. When training the discriminators, a specific type of a negative example is randomly chosen for a video, *i.e.* a batch consists of a combination of different types of negatives.

Baselines and SoTA. We compare our Adversarial Inference (denoted MLE+HybridDis) to: our baseline generator (MLE); multiple inference procedures, *i.e.* beam search with size 3 (MLE+BS3), sampling with log probabilities (MLE+LP) and inference with the single discriminator (MLE+SingleDis); Self Critical Sequence Tranining [46] which optimizes for CIDEr (SCST); GAN models built off [6, 36] with a single discriminator[3], with and without a cross entropy (CE) loss (GAN, GAN w/o CE). Finally, we also compare to the following state-of-the-art methods: Transformer [76], VideoStory [13] and MoveForwardTell [67], whose predictions we obtained from the authors.

4.2. Results

Automatic Evaluation. Following [67], we conduct our evaluation at paragraph-level. We include standard metrics, *i.e.* METEOR [31], BLEU@4 [39] and CIDEr-D [57]. However, these alone are not sufficient to get a holistic view of the description quality, since the scores fail to capture content diversity or detect repetition of phrases and sentence structures. To see if our approach improves on these properties, we report Div-1 and Div-2 scores [53], that measure a ratio of unique N-grams (N=1,2) to the total number of words, and RE-4 [67], that captures a degree of N-gram repetition (N=4) in a description[4]. We compute these scores at video (paragraph) level, and report the average score over all videos. Finally, we want to capture the degree of "discriminativeness" among the descriptions of videos with similar content. ActivitiyNet [3] includes 200 activity

[2] The two references are not aligned to the same time intervals, and even may have a different number of sentences.

[3] We have tried incorporating our hybrid discriminator in GAN training, however, we have not observed a large difference, likely due to a large space of training hyper-parameters which is challenging to explore.

[4] For Div-1,2 higher is better, while for RE-4 lower is better.

Method	METEOR	Per video BLEU@4	CIDEr-D	Overall Vocab Size	Sent Length	Per act. RE-4 ↓	Per video Div-1 ↑	Div-2 ↑	RE-4 ↓
MLE	16.70	9.95	20.32	1749	13.83	0.38	0.55	0.74	0.08
GAN w/o CE	16.49	9.76	20.24	2174	13.67	0.35	0.56	0.74	0.07
GAN	16.69	10.02	21.07	1930	13.60	0.36	0.56	0.74	0.07
SCST	15.80	10.82	20.89	941	12.13	0.52	0.47	0.65	0.11
MLE + BS3	16.22	10.79	21.81	1374	12.92	0.48	0.55	0.71	0.11
MLE + LP	17.51	8.70	12.23	1601	18.68	0.48	0.48	0.69	0.12
MLE + SingleDis	16.29	9.25	18.17	2291	13.98	0.37	0.59	0.75	0.07
MLE + SingleDis w/ Pair	16.16	9.32	18.72	2375	13.75	0.37	0.60	0.77	0.06
(Ours) MLE + HybridDis w/o Vis	16.33	8.92	17.29	2462	14.43	0.34	0.59	0.76	0.06
(Ours) MLE + HybridDis w/o Lang	16.44	9.37	19.44	2697	13.77	0.30	0.59	0.78	0.05
(Ours) MLE + HybridDis w/o Pair	16.60	9.56	19.39	2390	13.86	0.32	0.58	0.76	0.06
(Ours) MLE + HybridDis	16.48	9.91	20.60	2346	13.38	0.32	0.59	0.77	0.06
Human	-	-	-	8352	14.27	0.04	0.71	0.85	0.01
SoTA models									
VideoStory [13]	16.26	7.66	14.53	1269	16.73	0.37	0.51	0.72	0.09
Transformer [76]	16.15	10.29	21.72	1819	12.42	0.34	0.53	0.73	0.07
MoveForwardTell [67]	14.67	10.03	19.49	1926	11.46	0.53	0.55	0.66	0.18

Table 1: Comparison to video description baselines and SoTA models. Statistics over generated descriptions include N-gram Diversity (Div-1,2, higher better) and Repetition (RE-4, lower better) per video and per activity. See Section 4.2 for details.

labels, and the videos with the same activity have similar visual content. We thus also report RE-4 per activity by combining all sentences associated with each activity, and averaging the score over all activities.

We compare our model to baselines in Table 1 (top). The best performing models in standard metrics do not include our adversarial inference procedure nor the jointly trained GAN models. This is somewhat expected, as prior work shows that adversarial training does worse in these metrics than the MLE baseline [9, 53]. We note that adding a CE loss benefits GAN training, leading to more fluent descriptions (GAN w/o CE vs. GAN). We also observe that the METEOR score, popular in video description literature, is strongly correlated with sentence length.

We see that our Adversarial Inference leads to more diverse descriptions with less repetition than the baselines, including GANs. Our MLE+HybridDis model outperforms the MLE+SingleDis in every metric, supporting our hybrid discriminator design. Furthermore, MLE + SingleDis w/ Pair scores higher than the SingleDis but lower than our HybridDis. This shows that a *decoupled* Visual discriminator is important for our task. Note that the SCST has the lowest diversity and highest repetition among all baselines. Our MLE+HybridDis model also improves over baselines in terms of repetition score "per activity", suggesting that it obtains more video relevant and less generic descriptions.

To show the importance of all three discriminators, we provide ablation experiments by taking out each component, respectively (w/o Vis, w/o Lang, w/o Pair). Our HybridDis performs the worst when without its visual component and the combination of three discriminators outperforms each of the ablations on the standard metrics. In Figure 4, we show a qualitative result obtained by the ablated models vs. our full model. Removing the Visual discriminator leads to incorrect mention of "pushing a puck", as the visual error is not penalized as needed. Model without the Language discriminator results in somewhat implausible constructs ("stuck in the column") and incorrectly mentions "holding a small child". Removing the Pairwise discriminator leads to incoherently including a "woman" while missing the salient ending event (kids leaving).

Human Evaluation. The most reliable way to evaluate the description quality is with human judges. We run our evaluation on Amazon Mechanical Turk (AMT)[5] with a set of 200 random videos. To make the task easier for humans we compare two systems at a time, rather than judging multiple systems at once. We design a set of experiments, where each system is being compared to the MLE baseline. The human judges can select that one description is better than another or that both as similar. We ask 3 human judges to score each pair of sentences, so that we can compute a ma-

[5] https://www.mturk.com

HybridDis w/o Vis: A little girl is seen riding around a bumper car pushing a puck around a large set of bumper cars. The girl continues to move around the bumper car while the camera follows around. The girl smiles and walks away.

HybridDis w/o Lang: A little girl is sitting in a bumper car holding a small child in a red shirt. The girl in the red shirt gets stuck on the column. The girl walks away.

HybridDis w/o Pair: A small group of people are seen riding around in bumper cars and bumping into one another. The girl continues riding around the bumper car while the other people around the bumper car. The woman laughs and the girl smiles.

HybridDis: A small group of people are seen riding around in bumper cars and bumping into one another. The girl continues riding around the bumper car while others watch on the side. The girl finishes and walks away.

Ground Truth: Kids are sitting in bumper cars. They drive them and crash into each other. They stop and the kids get out.

Figure 4: Comparison of ablated models vs. our full model (discussion in text). Content errors are highlighted in red.

Method	Better than MLE	Worse than MLE	Delta
SCST	22.0	62.0	-40.0
GAN	32.5	30.0	+2.5
MLE + BS3	27.0	31.0	-4.0
MLE + LP	32.5	34.0	-1.5
MLE + SingleDis	29.0	30.0	-1.0
(Ours) MLE + HybridDis w/o Pair	42.0	36.5	+5.5
(Ours) MLE + HybridDis	38.0	31.5	**+6.5**

Table 2: Human evaluation of multi-sentence video descriptions, see text for details.

jority vote (*i.e*. at least 2 out of 3 agree on a judgment), see results in Table 2. Our proposed approach improves over all other inference procedures, as well as over GAN and SCST. We see that the GAN is rather competitive, but still overall not scored as high as our approach. Notably, SCST is scored rather low, which we attribute to its grammatical issues and high redundancy in the descriptions.

Comparison to SoTA. We compare our approach to multiple state-of-the-art methods using the same automatic metrics as above. As can be seen from Table 1 (bottom), our MLE + HybridDis model performs on par with the state-of-the-art on standard metrics and wins in diversity metrics. We provide a qualitative comparison to the state-of-the-art models in Figure 1 and in the supplemental material.

Person Correctness. Most video descriptions in the ActivityNet Captions dataset discuss people and their actions. To get additional insights into correctness of the generated descriptions, we evaluate the "person words" correctness. Specifically, we compare (a) the exact person words (e.g. *girl*, *guys*) and (b) only gender with plurality (e.g. *female-single*, *male-plural*) between the references and the predicted descriptions, and report the *F1* score in Table 3 (this is similar to [50], who evaluate character correctness in movie descriptions). Interestingly, our MLE baseline already outperforms the state-of-the-art in terms of person correctness, likely due to the additional object-level features [1]. SCST leads to a significant decrease in person word correctness, while our Adversarial Inference improves it.

Method	Exact word	Gender+ plurality
VideoStory [13]	44.9	64.1
Transformer [76]	45.8	66.0
MoveForwardTell [67]	42.6	64.1
MLE	48.8	67.5
SCST	44.0	63.3
GAN	48.9	67.5
(Ours) MLE + HybridDis	**49.1**	**67.9**

Table 3: Correctness of person-specific words, F1 score.

5. Conclusion

The focus of prior work on video description generation has so far been on training better generators and improving the input representation. In contrast, in this work we advocate an orthogonal direction to improve the quality of video descriptions: We propose the concept *Adversarial Inference* for video description where a trained discriminator selects the best from a set of sampled sentences. This allows to make the final decision on what is the best sample *a posteriori* by relying on strong trained discriminators, which look at the video and the generated sentences to make a decision. More specifically, we introduce a *hybrid discriminator* which consists of three individual experts: one for language, one for relating the sentence to the video, and one pairwise, across sentences. In our experimental study, humans prefer sentences selected by our *hybrid discriminator* used in *Adversarial Inference* better than the default greedy decoding. Beam search, sampling with log probability as well as previous approaches to improve the generator (SCST and GAN) are judged not as good as our sentences. We include further qualitative results which demonstrate the strength of our approach in supplemental materials.

Acknowledgements. The work of Trevor Darrell and Anna Rohrbach was in part supported by the DARPA XAI program, the Berkeley Artificial Intelligence Research (BAIR) Lab, and the Berkeley DeepDrive (BDD) Lab.

Supplemental Material

Here we provide implementation details for our approach and baseline models (Section A), and include qualitative comparison of our approach to ablations, baselines and state-of-the-art methods (Section B).

A. Implementation Details

Processing the Visual Feature. First, we detail how we obtain the visual input $\bar{v}_m^i$ in Equation 1. Unlike image captioning that relies on static features, video description requires a dynamic multimodal fusion over different visual features, such as e.g. a stream of RGBs and motion features. In addition to video and image-level features, we introduce object detections extracted for a subset of frames. Different features may be temporally misaligned (*i.e.* extracted over different sets of frames). We address this as follows. Suppose, a visual feature f extracted from v^i is represented as a sequence of T_f segments: $v_f^i = [v_{f,1}^i, v_{f,2}^i, ... v_{f,T_f}^i]$ [62, 67]. The previous hidden state h_{m-1}^i is used to predict temporal attention [68] over these segments, which then results in a single feature vector $\hat{v}_{m,f}^i$. We concatenate the resulting vectors from all features as our final visual input to the decoder: $\bar{v}_m^i = [\hat{v}_{m,1}^i, \hat{v}_{m,2}^i, ..., \hat{v}_{m,f}^i, ...]$.

Self-Critical Sequence Training. Self-Critical Sequence Training [46] (SCST)[6] is a variant of REINFORCE [66] where the inference algorithm is used as a baseline. Suppose we have a generator model G_θ with parameters θ; a complete sequence $x^s = (x_1^s, ... x_T^s)$ is sampled using the probability distribution $p_\theta(x_t^s | x_1^s, ... x_{t-1}^s)$ at each time step t. To reduce the variance during training and explore beyond the current best policy, SCST decodes another sequence $\hat{x}$ with the inference algorithm (greedy decoding) and aims to improve x^s over $\hat{x}$ based on a reward r such as a CIDER metric [57]. The gradient function for the model is calculated as:

$$\nabla_\theta L_{G_\theta}(\theta) = \sum_{t=1}^{T} (r(x^s) - r(\hat{x})) \nabla_\theta \log p_\theta(x_t^s | x_{1:t-1}^s). \tag{11}$$

GANs for Captioning. GANs for image captioning [9, 53] are typically trained with the following procedure due to their instability in early training stages: 1) pre-train the generator G_θ optimizing MLE objective, 2) pre-train discriminator D_η by sampling sentences from pre-trained G_θ, and 3) jointly update G_θ and D_η iteratively with a different objective for G_θ to deal with non-differentiable sampling. Cross Entropy loss is used to pre-train G_θ and D_η, where D_η is trained with negative samples as in Equation 3, with $\mu = 0.5, \nu = 0.5$. After both G_θ and D_η have been pre-trained, we follow [6, 36] and jointly train them using SCST but replacing reward r with an output of a standard ("single") discriminator $D_\eta(V, x^s)$, where V is a given video segment and x^s is a sampled description. We find that it is best to update G_θ for 5 steps for each update of D_η. The gradient for the above GAN model is:

$$\nabla_\theta L_{G_\theta}(\theta) = \sum_{t=1}^{T} (D_\eta(V, x^s) - D_\eta(V, \hat{x})) \nabla_\theta \log p_\theta(x_t^s | x_{1:t-1}^s). \tag{12}$$

Due to instability of adversarial training, we additionally include a cross entropy (CE) loss that ensures that the generator will explore an output space in a more stable manner and maintain its language model [40]. The final objective of G_θ is a mixed loss function, a weighted combination of Cross-Entropy Loss (L_{CE}) optimizing the maximum-likelihood training objective and Adversarial Loss (L_{GAN}) with its gradient function defined in Equation 12:

$$L_{\text{MIX}} = \lambda L_{\text{GAN}} + (1 - \lambda) L_{\text{CE}}, \tag{13}$$

where we use $\lambda = 0.995$. We compare this mixed objective to not using the CE loss in Table 1 of the main paper.

Adversarial Inference. Suppose each word w_i in a vocabulary of size K can be sampled with a probability $p(w_i)$. One can additionally modify the probability distribution during sampling with a temperature parameter τ:

$$p_\tau(w_i) = \frac{p(w_i)^{1/\tau}}{\sum_{j=1}^{K} p(w_j)^{1/\tau}}. \tag{14}$$

Based on Equation 14, $\tau = 1$ is a default sampling procedure. Setting $\tau < 1$ shifts the distribution to favor larger probabilities, making the overall distribution more "peaky". We explore parameter τ for both discriminator training, τ_T, and adversarial inference, τ_I. We obtain more fluent captions by setting $\tau_I < 1$ during inference, however we find it is best to set $\tau_T = 1$ during discriminator training so that it learns to distinguish natural and fake descriptions. In our adversarial inference procedure, we sample $K = 100$ sentences with $\tau_I = 0.2$ for each for each video segment. One can see the effect of different temperatures during inference in Figure 5.

B. Qualitative Examples

In this section, we provide qualitative examples comparing our Adversarial Inference method to its ablations, other baselines and state-of-the-art models.

[6]Our SCST model is based on the implementation of https://github.com/ruotianluo/self-critical.pytorch

τ	Descriptions
1.0	- Man is then seen doing various interview on a magazine and showing how the ending credits. - A young man in the red shirt is demonstrating how to do a straight clap and cover his movements and then a man in white shorts why. - The two then begin dribbling each others and legs around each others and show how to properly play some board one another.
0.8	- Still images of the screen that are shown and we see the ending screen. - We see a man on a black background and a man is seen standing before sanding a bar and holding a photo in a gym. - The man in white uniform that does some flips and punches.
0.5	- A man is seen speaking to the camera and leads into several clips of people performing a photo. - A man is seen speaking to the camera and leads into various clips of people performing martial arts. - The man then moves the camera around and leads into them performing various martial arts moves.
0.2	- A man is seen speaking to the camera and leads into several clips of people performing various sports on the screen. - The man then begins talking to the camera and leads into him performing various martial arts moves in a gym. - The two men begin fighting one another and ends with one another and the man in the middle of the circle.
Ground Truth	- A man is seen speaking to the camera and pans out into more men standing behind him. - The first man then begins performing martial arts moves while speaking to the camera - He continues moving around and looking to the camera

(a)

τ	Descriptions
1.0	- A woman walks past the camera gradually. - One of the women toss the board and wave to the other side of his knee side while people walks up and down. - We see a girl standing on a blue mat with girls trophies. - Several shots of different people standing waiting on the top and sharply are displayed on one side of them.
0.8	- People stand on a street holding a cord and warming up. - The women are seen speaking to the camera and show shots of the group of people getting ready to perform. - The group continues to play playing with the girls who are smiling while the camera captures them. - People are in a large auditorium and people doing their hands on the sides and they read their medals through the sky.
0.5	- A woman in a yellow shirt is dancing with the people in the background. - The people continue dancing around and then the woman help the woman in the yellow shirt. - The girls hands and the girls are standing up on the ground. - The camera pans around a large group of people dancing on a city and people enjoying them.
0.2	- A woman is seen walking down a street and leads into several shots of people playing with one another. - The girls continue to play the game and ends with the people walking around and ends with the crowd clapping. - The girls then begin to dance and hold up their arms up and down. - The people continue to ride around and waving to the camera and then a group of people are shown running in the city.
Ground Truth	- People are dancing having fun at a party. - A race starts and people are running. - Cheerleaders are standing on the side of the road. - People are dancing on the grass.

(b)

Figure 5: Sampling multi-sentence descriptions with different temperatures. The sentences are sampled from a pre-trained generator with temperatures $\{1.0, 0.8, 0.5, 0.2\}$. Each sentence corresponds to a clip in a video. Note that higher temperatures tend to lead to more diverse vocabulary with the cost of decreased fluency.

B.1. Comparison to Model Ablations and GAN

Figure 6 shows a few qualitative examples comparing ground truth descriptions to the ones generated by the following methods: MLE, SCST (with CIDEr), GAN, MLE+SingleDis (Single Disc), and our MLE+HybridDis (Ours). We highlight errors, *e.g*. objects not present in video, in bold/red, and repeating phrases in italic/blue. Overall, our approach leads to more correct, more fluent, and less repetitive multi-sentence descriptions than the baselines. In (a), our prediction is preferable to all the baselines w.r.t. the sentence fluency. While all models recognize the presence of a baby and a person eating an ice cream, the baselines fail to describe the scene in a coherent way, but our approach summarizes the visual information correctly. Our model also generates more diverse descriptions specific to what is happening in the video, often mentioning more interesting and informative words/phrases, such as "trimming the hedges" in (b) or "their experience" in (c). MLE and SCST mention less visually specific information, and generate more generic descriptions, such as "holding a piece of wood". In an attempt to explore diverse phrases, the single discriminator is more prone to hallucinating non-existing objects, *e.g*. "monkey bars" in (b). Finally, our model outperforms the baselines in terms of lower redundancy across sentences. As seen in (c), our approach delivers more diverse content for each clip, while all others more frequently generate "speaking/talking to the camera", a very common phrase in the dataset.

We provide additional examples comparing our approach to SCST and GAN in Figure 7, further illustrating how adversarial inference improves over adversarial training in terms of correctness and fluency. Again, our approach leads to mentioning important concepts, such as *e.g*. "tai chi". SCST results in ungrammatical sentence endings (*e.g*. "a game of", "begins to the camera").

We also show the effect of our Pairwise Discriminator in Figure 8. As we see, an additional consistency score between sentences helps us obtain less redundant and sometimes more correct predictions (*e.g*. in (a) the hybrid w/o pair never mentions dropping the weights).

B.2. Comparison to State-of-the-Art

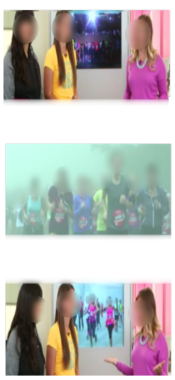
Figure 9 provides a comparison of descriptions obtained by our approach to three recent video description models (VideoStory [13], Transformer [76], MoveForwardTell [67]). While the state-of-the-art models are often able to capture the relevant visual information, they are still prone to issues like repetition, lack of diverse and precise content as well as content errors. In particular, VideoStory and MoveForwardTell suffer from the dominant language prior and repeatedly mention "the camera", making the stories less informative and specific to the events in the video. Despite having less repeating contents and high scores in language metrics, the Transformer model is prone to produce incoherent phrases *e.g*. "a man is a bikini" or "putting sunscreen on the beach water", and ungrammatical endings, *e.g*. "and a" in (a). On the other hand, our model captures the visual content more precisely, *e.g*. in the top example it refers to the subject as a "girl", pointing out that the girl is "laying on a bed", correctly recognizing "sand castles", etc. Besides, unlike prior work, our approach mentions important video relevant concepts (*e.g*. "choppy waters", "rapids", "afloat" in (b); "synchronized", "stepper" in (c)). Overall, we see more diversity and less repetitiveness, along with more accurate description of video content. We note that there is still a large room for improvement w.r.t.

	MLE	SCST	GAN	Single Disc	Ours	Ground Truth
1	A baby is sitting on a chair eating **a baby.**	A **man** is sitting **on** a table with a baby **in a cup**.	A baby is sitting in a chair eating **a baby eating *a* baby.**	A baby is sitting on a chair with **a baby baby in a baby** and a baby eating ice cream **cone**.	A baby is sitting **on** a table eating ice cream.	A woman is seen scooping up a spoonful of ice cream and taking a bite with a baby in front of her.
2	The baby licks **the baby** and the girl licks the ice cream.	The **man** is sitting **on** the ice cream.	A baby is sitting in a chair.	The baby enjoys the ice cream **cone**.	The baby licks the ice cream **cone**.	The woman continues to tease the baby with the ice cream giving him little bites here and there as well as taking bites for herself and laughing.

(a)

	MLE	SCST	GAN	Single Disc	Ours	Ground Truth
1	A man is standing in a yard holding a large **piece of wood**.	A man is standing in a yard.	A man is standing outside holding a large tree.	A man is standing outside talking.	A man is standing next to a man wearing a blue shirt and a hat is standing next to a tree.	A man starts up a gas powered hedger and hands it to someone standing on a platform.
2	A man in a blue shirt is standing next to a tree.	A man is seen standing on a **piece of wood**.	A man is seen standing in a yard holding a **piece of wood**.	A man is seen standing in a yard while a man watches him.	The man in the blue shirt is trimming the hedges.	The man on the platform trims the top of a hedge with the hedge trimmer.
3	He is then shown **doing** a little girl in front of a **large crowd**.	The man continues to **the camera and down the man** and the man continues to speak to the camera.	He then goes back to the **monkey bars.**	He then goes back to the **monkey bars** and then walks away.	The man in the **red** shirt is standing near the ladder.	The platform is towed with a tractor alongside the hedge while the man continues to trim it.

(b)



	MLE	SCST	GAN	Single Disc	Ours	Ground Truth
1	A woman is seen speaking to the camera and *leads into her speaking to the camera.*	A woman is seen *speaking to the camera* and leads into a woman *speaking to the camera.*	A woman is seen speaking to the camera and *leads into her speaking to the camera.*	A woman is seen hosting a news segment that leads into a woman *speaking to the camera* and leads into her *speaking to the camera.*	A woman is seen hosting a news segment that leads into a woman *speaking to the camera.*	Three women are seen speaking to the camera and answering questions that the hosts ask.
2	The woman then speaks to the camera while several people are seen speaking to the camera and *leads into her speaking to the camera.*	The woman is *talking to the camera* and showing the woman.	The woman talks to the camera while several people are shown in front of the camera.	The woman continues *speaking to the camera* while more shots of people speaking and leads into several shots of people *speaking to the camera.*	The woman continues *speaking to the camera* while showing **off her new york** and news reporter.	Clips are shown of people running down the street while the women continue to speak.
3	The woman continues speaking to the camera and *leads into her speaking to the camera and leads into her speaking to the camera.*	The woman is *talking to the camera* and then the woman in the news.	The woman continues speaking to the camera and *leads into her speaking to the camera.*	The women are interviewed in the end of the race	The women are interviewed by the camera and then they are shown talking about their experience.	The girls talk more and more while people are still shown running down the road

(c)

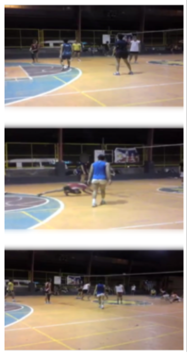
Figure 6: Comparison of our approach to MLE baseline, SCST, GAN, and Adversarial Inference with Single Discriminator. Red/bold indicates content errors, blue/italic indicates repetitive patterns.

the human ground-truth descriptions.

B.3. Failure Analysis

Finally, we analyze failures of our approach. As shown in the previous examples, our model is not free of errors, *e.g*. it hallucinates an ice cream "cone" (Figure 6 (a)), incorrectly mentions "showing off her new york" (Figure 6 (c)), predicts "man" instead of a woman (Figure 7 (b)) and "woman" instead of a child (Figure 9 (a)) or "lifting" instead of "dropping" (Figure 8 (a)), etc. It is also still prone to some repetition (*e.g*. Figure 7 (a), (b), Figure 9 (a)). Overall, however, our captions improve over those of the baselines, as supported by our human evaluation.

We include a few additional failure cases in Figure 10, showcasing difficult examples from the ActivityNet Captions dataset. In particular, fine-grained activities that involve small objects are hard, *e.g*. our model confuses applying makeup with inserting a contact lens in Figure 10 (a), incorrectly mentions a "hair dryer" and "scissors" in Figure 10 (b), and "vegetables" and "potatos" in Figure 10

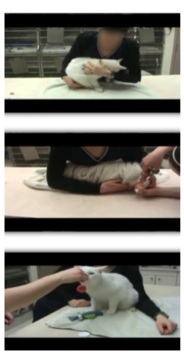


MLE	SCST	GAN	Ours	Ground Truth
A group of people are seen standing around a court playing a game of **soccer.**	A group of people are seen standing on a court playing a game **of**.	A group of people are seen standing around a court playing a game of **soccer.**	A group of people are seen standing around a court playing a game of volleyball.	A small group of people are seen wandering around a gym hitting a ball.
The men *continue playing the game* and the camera pans around the area and the other team mates back.	The people *continue playing the ball* around and the man in the end and the other team **around** the ball.	The people continue playing the game **of the game** and the game ends with the people watching the game.	The people *continue playing the game* and ends with the camera panning around the area.	The people hit the ball back and fourth while others watch on the side.
The people *continue playing the game* and the man continues to play the game **of the game being played**.	The people *continue playing the ball* and hitting the ball.	The players continue to play the game **of the game**.	The people *continue playing the game* and the people hit the ball back and fourth.	The people chase after the ball and continue to hit it up into the air.

(a)

MLE	SCST	GAN	Ours	Ground Truth
A **man** is standing in a room in a **gym.**	A **man** is seen standing in a room and begins to the **camera.**	A **man** is seen standing in a large field looking back to the camera.	A **man** is standing in a **gym** and then the **man** demonstrates how to do **a** martial arts moves.	A woman is seen standing outside with her feet together and looking off into the distance.
He is standing in a room.	The **man** is standing in a room and begins to the **camera.**	**He** then demonstrates how to *properly perform moves* **moves** in the end.	She begins to demonstrate how to properly execute the moves *side to side side to side*.	The woman then begins moving slowly around the area while moving her hands back and fourth.
She is doing various moves.	The woman is standing in the room and begins to the **camera.**	She then demonstrates how to *properly perform moves* **moves**.	She then demonstrates how to *properly perform* tai chi and demonstrate how to *properly perform* moves.	She continues moving her body around and looking off into the distance.

(b)

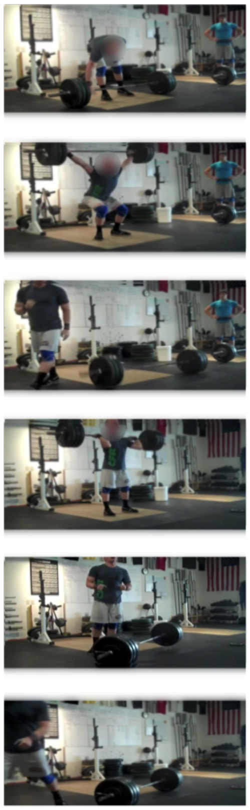


MLE	SCST	GAN	Ours	Ground Truth
A person is seen holding a cat and laying **out of a cat** on a bed.	A **man** is seen sitting on a **couch** and leads into a woman speaking to the camera.	A person is seen holding a cat on the **floor** and leads into her holding a cat claws.	A person is seen holding a cat on a table and leads into **him** cutting a cat's claws.	Woman is standing holding a cat and put her on top of a table and giving her affection.
The woman then grabs a cat and begins cutting the cat's claws.	The woman is then seen holding a cat **on the cat and begins to the cat.**	The person then puts the cats claws on the side while **the cat** attempts to cut the cat's claws.	The person continues cutting the cat's nails while the camera captures **him** from various angles.	Woman holds a nail clipper and wrap a cat in a towel to cut her nails while other woman is holding the cat.
The woman is then seen cutting the cat's claws and the **cat** is still **cutting** the cat's claws.	The woman is seen *speaking to the camera* and leads into her cutting the **cat** and *speaking to the camera.*	The woman then begins to brush the cat's nails **with a cat.**	The woman continues to brush the cats nails and ends by still speaking to the camera.	Woman is throwing a ball of yarn to the cat.

(c)

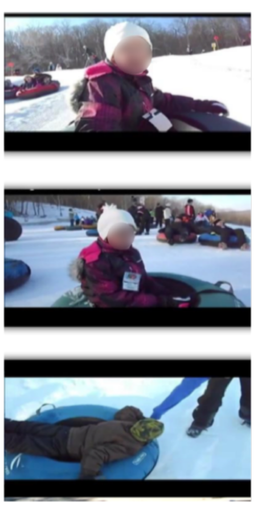
Figure 7: Comparison of our approach to MLE baseline, SCST, and GAN. Red/bold indicates content errors, blue/italic indicates repetitive patterns.

(c). The other methods are also struggling on these challenging videos, by either making errors or lacking detail, showing that there is still a long way to go towards solving multi-sentence video description in the wild.



Hybrid w/o Pair	Ours	Ground Truth
A man is seen bending over to lift a weight lifting above his head.	A man is seen bending over to a set of weights while lifting weights.	A man walks to a barbell and grips the handle.
The man lifts up a barbell and **walks away**.	The man lifts up over his head and drops it down.	The man then lifts the weight over his head and stands up.
He **lifts** a weight over his head.	He bends down and **lifts** it over his head.	The man drops the weight, pumps his fist, and walks off.
He lifts it over his head.	He lifts it over his head.	The man returns and lifts the weight to his shoulders then over his head.
The man **lifts** the weight over his head.	The man drops the weight.	The man drops the weight and laughs and pumps his fist.
He **lifts** it over his head.	He **lifts** the weight over his head.	The man then walks off camera.

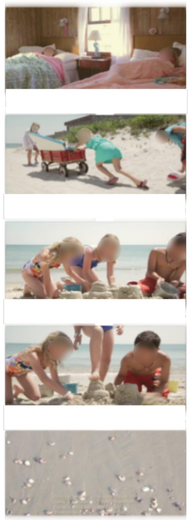
(a)



Hybrid w/o Pair	Ours	Ground Truth
A woman is seen sitting in a tube with a camera following behind them.	A woman is seen sitting in a tube with a camera following behind her.	A person is seen riding in a tube and looking at the camera.
The man in the red jacket is pushed down the hill.	The people continue riding down the hill and ends with the camera panning around the area.	More people are seen riding down a snowy hill on tubes as well as laughing into the camera.
The man in the **red hat** is pushed down the hill.	They are then shown sledding down the hill together.	More children play on the hill and pull one another along down the mountain.

(b)

Figure 8: Effect of Pairwise Discriminator term in our approach. Red/bold indicates content errors. While both models in a) are not perfectly aligned with ground truth descriptions, the one without pairwise discriminator keeps repeating *lifts a weight* and fails to mention that the man *drops the weight*. Similarly in b), the model without pairwise discriminator mentions that man is *pushed down the hill* twice in a row, while ours avoids generating similar descriptions but more diverse phrases within the paragraph such as *continue riding down the hill* and *shown sledding down the hill together* .

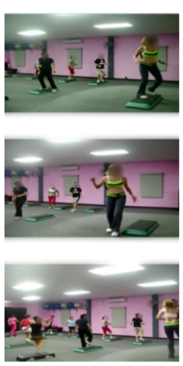


Video Story	Transformer	MFT	Ours	Ground Truth
A **woman** is seen **standing in a chair** and leads into her holding **a box** and looking off into the distance.	A **woman** is seen sitting on a **beach** followed by a **woman putting sunscreen on a beach**.	A **man** is seen sitting on a **chair** and *speaking to the camera*.	A girl is seen sitting on a bed and leads into her laying on a bed.	Little blonde kid is waking up and throw a eddy to her sister in the other bed and stands in front of a drawer looking for clothes.
The **woman** then begins **climbing the dog** while the **woman** *continues to speak to the camera*.	A **man** is **a bikini** and a **man** is walking in *a beach a beach*.	A person is seen riding down a road while others watch on the side.	The girl continues to run around the area while the camera captures her movements.	Kids are running in the outside and pulling a cart into a sandy beach.
The camera pans around the beach and the **woman** *continues to speak to the camera*.	A **man** in a **blue shirt** is putting **sunscreen on the beach water**.	A **man** is seen kneeling down on the ground and *speaking to the camera*.	*The camera pans around the beach* and leads into her laying down on a large beach and **laying** sand castles.	The two girls and a kid are doing a sandcastle on seashore.
The **woman** *continues to speak to the camera* and ends with her hands up in the water.	People are in the beach and a **man** and a **woman** are in the beach **and a.**	A group of people are seen standing around a beach and leads into people *speaking to the camera*.	The **woman** continues to dig up the sand castle and smiling to the camera.	The kids step on the sandcastle and destroy it and walks into the shore and sunbathe on top of towels.
The **woman** is shown again with the **woman** who is now shown **of the woman who is shown again and the woman begins to do it with the woman.**	The people continue to ride around the beach and end by *the water* and *the water.*	A **man** is seen *speaking to the camera* and leads into him *speaking to the camera.*	*The camera pans all around* the beach and leads into several shots of *the camera panning around* the water.	Kid gives a seashell to the girls and walks in the seashore jumping and laughing and then the credits appears.

(a)

Video Story	Transformer	MFT	Ours	Ground Truth
People are rafting down a river .	The people raft down a river **river raft.**	People are paddling down the river.	A group of rafters **don rafts** down a river in choppy waters.	These people are sitting in the inflatable red/white boat and they're floating along the waves.
The camera pans around a person riding a **rock** and leads into a person riding down a river in the water and the *camera zooms in on the water*.	The people then paddle *down the river* while paddling *down the river* raft.	People are *paddling down the river*.	They are going down the rapids trying to stay afloat.	They all work together and paddle themselves through the water and most of them are smiling.
The person continues to *paddle around the water* and begins to *paddle around the water* and the *camera zooms in on the water*.	The rafters continue to paddle through the rapids and end by holding their paddles.	People are *paddling down the river* **in a river**.	The rafters are then shown paddling through the water as the camera follows their movements.	The water splashes onto the front of the camera while they're in the water and they're wearing helmets along with water gear.

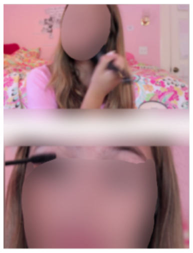
(b)



Video Story	Transformer	MFT	Ours	Ground Truth
A group of people are seen standing in a room with a large group of people *moving around* and *moving around*.	We see people in a room.	A group of people are seen standing in a room with a **man** speaking to the camera.	A group of women are in a gym doing a synchronized move up and down on a stair stepper.	A number of women exercise together using a stepping type of implement
The people continue dancing around on the floor while the camera pans around and moving around the **rope**.	They are dancing in a room.	A group of people are inside a gym.	They are doing the same dance in a synchronized manner.	The camera pans slightly to the right.
The people *continue to dance around* and *continue to dance around* and end by holding the pose.	The people continue dancing around the room.	A group of people are seen standing in a room with a **man** speaking to the camera.	They are using **a** synchronized steppers to move.	The camera pans back slightly to the left.

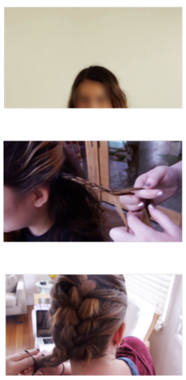
(c)

Figure 9: Comparison of our approach to state-of-the-art video description approaches (VideoStory [13], Transformer [76], MoveForwardTell [67]). Red/bold indicates content errors, blue/italic indicates repetitive patterns.



Video Story	Transformer	MFT	Ours	Ground Truth
A woman is seen looking at the camera and leads into her holding a brush and looking into her.	A woman is seen speaking to the camera and leads into her putting makeup **makeup** on her face.	A woman is seen speaking to the camera and leads into her brushing her face.	A young girl is seen speaking to the camera and leads into her holding up a **contact lens**.	A girl is shown in several shots putting makeup on and leads into her putting makeup for her chin and lips.
The woman then puts her face **on her face** and begins to **apply her face down**.	She then puts eyeliner on the eyelids and puts it on her eye **eye**.	She puts mascara on her eyelashes.	She then puts mascara on her eyelashes and places it on her **eye**.	She then puts mascara on while continuously smiling into the camera and followed by more makeup being put on.

(a)



Video Story	Transformer	MFT	Ours	Ground Truth
A woman is seen speaking to the camera and leads into her holding a **brush**.	We see a girl with makeup on her face.	The woman **finishes** and looks at the camera.	A woman is seen sitting in a chair and leads into her holding up a **hair dryer**.	The top of a woman's head is seen.
The woman continues braiding her hair and *ends by smiling to the camera*.	A woman is seen **putting makeup on her face** and leads into her **putting makeup** on her hair.	We see the **ending title screen.**	She then takes a pair of **scissors** and begins to **blow dry** her hair into a ponytail.	She is shown getting her hair braided by another woman.
The woman continues to **use** the hair and *ends by smiling to the camera*.	The woman continues to **put makeup on her face** and ends with her hair and smiling **to**.	A **man** is seen speaking to the camera and leads into clips of **him running down the streets**.	She then brushes her hair and shows off the finished result.	The woman tucks the braid in showing how she keeps it clipped with bobby pins.

(b)

Video Story	Transformer	MFT	Ours	Ground Truth
A woman is seen *speaking to the camera* and leads into a woman *speaking to the camera* and leads into her pouring ingredients into a pan.	A woman stands in a kitchen **kitchen kitchen kitchen**.	The woman mixes the ingredients together and pours them into a bowl.	A woman is seen speaking to the camera and leads into her pouring ingredients into a bowl filled with **vegetables**.	A woman in a kitchen talks to a camera while cooking spaghetti and preparing a complete spaghetti dish including sauce.
She then mixes the ingredients into a bowl and mixes them into a pan.	A woman is seen speaking to the camera while holding **a** ingredients and leads into her holding **a**.	A woman is seen speaking to the camera while holding up a **bottle of water**	The woman then begins to put the ingredients in a **kitchen aid** and then takes a sip of the ingredients and then proceeds to **bake the pan in the oven**.	A woman in a red shirt boils spaghetti on a stove top and taste tests it for doneness.
She then *pours the mixture into* the **water** and she puts it on the counter .	She woman puts the ingredients into a bowl and pours it into a bowl it.	A woman is seen *speaking to the camera* and leads into her *speaking to the camera*.	The woman then begins to **peel the potatoes** and the woman mixes ingredients together in the kitchen.	The woman moves the spaghetti to a sink and pours it into a white bowl.
She then *pours the mixture into* a pan and pours it into a pan	She then mixes the ingredients together with the **salad** and ends by presenting it to the camera.	The woman mixes the ingredients together and pours them into a bowl.	She adds some more vegetables and mixes it together.	The woman pours sauce over the spaghetti puts spices on top along with shredding cheese on top of it before.

(c)

Figure 10: Failure cases of our approach and state-of-the-art video description approaches (VideoStory [13], Transformer [76], MoveForwardTell [67]). Red/bold indicates content errors, blue/italic indicates repetitive patterns.

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 (CVPR)*, 2018.

[2] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[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.

[4] Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. Language gans falling short. *arXiv preprint arXiv:1811.02549*, 2018.

[5] Tong Che, Yanran Li, Ruixiang Zhang, R Devon Hjelm, Wenjie Li, Yangqiu Song, and Yoshua Bengio. Maximum-likelihood augmented discrete generative adversarial networks. *arXiv preprint arXiv:1702.07983*, 2017.

[6] Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, Fuming Ma, and Qi Ju. Improving image captioning with conditional generative adversarial nets. *arXiv:1805.07112*, 2018.

[7] Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. Less is more: Picking informative frames for video captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[8] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. Learning to evaluate image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5804–5812, 2018.

[9] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[11] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[12] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017.

[13] Spandana Gella, Mike Lewis, and Marcus Rohrbach. A dataset for telling the stories of social media videos. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–974, 2018.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[15] Tszhang Guo, Shiyu Chang, Mo Yu, and Kun Bai. Improving reinforcement learning based image captioning with natural language prior. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[16] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, pages 18–22, 2018.

[17] 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.

[18] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

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

[20] Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. Learning to write with cooperative discriminators. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[24] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2016.

[25] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, JungWoo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision (IJCV)*, 50(2):171–184, 2002.

[28] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1008–1014, 2000.

[29] 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.

[30] 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.

[31] Michael Denkowski Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, page 376, 2014.

[32] Lijun Li and Boqing Gong. End-to-end video captioning with multitask reinforcement learning. *arXiv preprint arXiv:1803.07950*, 2018.

[33] 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 (CVPR)*, pages 7492–7500, 2018.

[34] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[35] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 3, page 3, 2017.

[36] Igor Melnyk, Tom Sercu, Pierre L Dognin, Jarret Ross, and Youssef Mroueh. Improved image captioning with adversarial semantic alignment. *arXiv:1805.00063*, 2018.

[37] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.

[38] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[40] Ramakanth Pasunuru and Mohit Bansal. Reinforced video captioning with entailment rewards. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[41] 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.

[42] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[43] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

[44] 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.

[45] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[46] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[47] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[48] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *Proceedings of the German Confeence on Pattern Recognition (GCPR)*, 2014.

[49] Anna Rohrbach, Marcus Rohrbach, and Bernt Schiele. The long-short story of movie description. In *Proceedings of the German Confeence on Pattern Recognition (GCPR)*, 2015.

[50] 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.

[51] Naeha Sharif, Lyndon White, Mohammed Bennamoun, and Syed Afaq Ali Shah. Learning-based composite metrics for improved caption evaluation. In *Proceedings of ACL 2018, Student Research Workshop*, pages 14–20, 2018.

[52] Rakshith Shetty and Jorma Laaksonen. Frame- and segment-level features and candidate pool evaluation for video caption generation. In *Proceedings of the ACM international conference on Multimedia (MM)*, pages 1073–1076, 2016.

[53] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[54] Andrew Shin, Katsunori Ohnishi, and Tatsuya Harada. Beyond caption to narrative: Video captioning with multiple sentences. In *Proceedings of the IEEE IEEE International Conference on Image Processing (ICIP)*, 2016.

[55] Sandeep Subramanian, Sai Rajeswar, Francis Dutil, Chris Pal, and Aaron Courville. Adversarial generation of natural language. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 241–251, 2017.

[56] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2017.

[57] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[58] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence – video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[59] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7622–7631, 2018.

[60] Jing Wang, Jianlong Fu, Jinhui Tang, Zechao Li, and Tao Mei. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2018.

[61] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7190–7198, 2018.

[62] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[63] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[64] Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[65] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4213–4222, 2018.

[66] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[67] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[68] 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 (ICCV)*, 2015.

[69] Huanyu Yu, Shuo Cheng, Bingbing Ni, Minsi Wang, Jian Zhang, and Xiaokang Yang. Fine-grained video captioning for sports narrative. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6006–6015, 2018.

[70] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[71] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pages 2852–2858, 2017.

[72] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[73] Mihai Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Spatio-temporal attention models for grounded video captioning. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2016.

[74] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. Actor-critic sequence training for image captioning. In *Advances in Neural Information Processing Systems (NIPS Workshops)*, 2017.

[75] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2018.

[76] 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 (CVPR)*, pages 8739–8748, 2018.

[77] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.