Adversarial Inference for Multi-Sentence Video Description

Jae Sung Park\textsuperscript{1}, Marcus Rohrbach\textsuperscript{2}, Trevor Darrell\textsuperscript{1}, Anna Rohrbach\textsuperscript{1}
\textsuperscript{1}University of California, Berkeley, \textsuperscript{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.\textsuperscript{1}

1. Introduction

Being able to automatically generate a natural language description for a video has fascinated researchers since the early 2000s\cite{27}. Despite the high interest in this task and ongoing emergence of new datasets\cite{13,29,75} and approaches\cite{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\cite{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\cite{31}, despite the evidence that they are not consistent with human judgments\cite{24,57}. Further, some recent works propose to explicitly optimize for the sentence metrics using reinforcement learning based methods\cite{35,46}. These techniques have become quite widespread, both for image and video description\cite{1,67}.

\textsuperscript{1}https://github.com/jamespark3922/adv-inf.
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]². The three vectors are concatenated to get the visual input $v^i_m$. To encourage coherence among consecutive sentences, we additionally append the last hidden state of the previous sentence $h^i_{m−1}$ 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:

$$h^i_m = LSTM(v^i_m, w^i_{m−1}, h^i_{m−1}), \quad \text{with} \quad h^0 = 0, \quad (1)$$

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

²For details, please, see the supplemental material.
positive match, while 0 is a negative match. Most prior 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:

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

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

$$L_D(v^j) = \mathbb{E}_{s \in S_{v^j}} \log(D(s|v^j)) + \mu \cdot \mathbb{E}_{s \in S_G} \log(1 - D(s|v^j)) + \nu \cdot \mathbb{E}_{s \in S_{\backslash v^j}} \log(1 - D(s|v^j)),$$

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, $\mu, \nu$ 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 visual discriminator $D_V$ determines whether a sentence $s^i$ refers to concepts present in a video clip $v^j$, 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 $\omega^i$.

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 $\omega^i$. The obtained vector representations $\hat{\omega}^j$ are then fused with the sentence representation $\omega^i$, 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 representation $\omega^s$ is obtained as follows:

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

where $\sigma$ is a sigmoid, producing values in $(0, 1)$, $\odot$ is the Hadamard product, $U$, $V$ are linear layers. Instead of concatenating features $\hat{s}_f^i$ as done in the generator, here we determine the scores $p_f^s$ between the sentence and each modality, and learn to weight 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 $\lambda_f^i$ to each modality based on the sentence representation $\omega^i$:

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

where $a_f$ are learned parameters. Finally, the $D_V$ score is the sum of the scores $p_f^s$ weighted by $\lambda_f^i$:

$$D_V(s^i|v^i) = \sum_f \lambda_f^i p_f^i, \quad (6)$$

**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 $\tilde{h}_i$, followed by a fully connected layer and a sigmoid layer:

$$D_L(s^i) = \sigma(W_L \tilde{h}_i + b_L). \quad (7)$$

**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 $\tilde{h}^{i-1}$ and $\tilde{h}_i$. We concatenate the two vectors and compute the $D_P$ score as follows:

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

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...
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 \( s^i_{\text{s}} \) and \( K \) sampled sentences \( s^i_1, s^i_2, ..., s^i_K \) from the generator \( G \), the scores from our hybrid discriminator can be used to compare the sentences and select the best one:

\[
 s^i_\ast = \arg\max_{j=1,...,K} D(s^i_j | v^i, s^i_{\text{s}}^{-1}) 
\]

where \( s^i_j \) is the \( j \)-th sampled sentence. The final discriminator score is defined as:

\[
 D(s^i_j | v^i, s^i_{\text{s}}^{-1}) = \alpha \cdot D_V(s^i_j | v^i) + \beta \cdot D_L(s^i_j) + \gamma \cdot D_P(s^i_j | s^i_{\text{s}}^{-1}), 
\]

where \( \alpha, \beta, \gamma \) 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\(^3\). 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 description.

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 (\( \mu, \nu \) in Eq. 3), are set to 0.5. The weights for our hybrid discriminator are set as \( \alpha = 0.8, \beta = 0.2, \gamma = 1.0 \). Sampling temperature during discriminator training is 1.0; during inference we sample \( K = 100 \) sentences with temperature 0.1. 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 Training [46] which optimizes for CIDEr (SCST); GAN models built off [6, 36] with a single discriminator\(^4\), 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\(^5\). We compute these scores at video (paragraph) level, and report the average

\(^3\)The two references are not aligned to the same time intervals, and even may have a different number of sentences.

\(^4\)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.

\(^5\)For Div-1,2 higher is better, while for RE-4 lower is better.
<table>
<thead>
<tr>
<th>Method</th>
<th>Per video</th>
<th>Overall</th>
<th>Per act.</th>
<th>Per video</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>METEOR</td>
<td>BLEU@4</td>
<td>CIDEr-D</td>
<td>RE-4</td>
</tr>
<tr>
<td>MLE</td>
<td>16.70</td>
<td>9.95</td>
<td>20.32</td>
<td>1749</td>
</tr>
<tr>
<td>GAN w/o CE</td>
<td>16.49</td>
<td>9.76</td>
<td>20.24</td>
<td>2174</td>
</tr>
<tr>
<td>GAN</td>
<td>16.69</td>
<td>10.02</td>
<td>21.07</td>
<td>1930</td>
</tr>
<tr>
<td>SCST</td>
<td>15.80</td>
<td>10.82</td>
<td>20.89</td>
<td>941</td>
</tr>
<tr>
<td>MLE + BS3</td>
<td>16.22</td>
<td>10.79</td>
<td>21.81</td>
<td>1374</td>
</tr>
<tr>
<td>MLE + LP</td>
<td>17.51</td>
<td>8.70</td>
<td>12.23</td>
<td>1601</td>
</tr>
<tr>
<td>MLE + SingleDis</td>
<td>16.29</td>
<td>9.25</td>
<td>18.17</td>
<td>2291</td>
</tr>
<tr>
<td>MLE + SingleDis w/ Pair</td>
<td>16.16</td>
<td>9.32</td>
<td>18.72</td>
<td>2375</td>
</tr>
<tr>
<td>(Ours) MLE + HybridDis w/o Vis</td>
<td>16.33</td>
<td>8.92</td>
<td>17.29</td>
<td>2462</td>
</tr>
<tr>
<td>(Ours) MLE + HybridDis w/o Lang</td>
<td>16.44</td>
<td>9.37</td>
<td>19.44</td>
<td>2697</td>
</tr>
<tr>
<td>(Ours) MLE + HybridDis w/o Pair</td>
<td>16.60</td>
<td>9.56</td>
<td>19.39</td>
<td>2390</td>
</tr>
<tr>
<td>(Ours) MLE + HybridDis</td>
<td>16.48</td>
<td>9.91</td>
<td>20.60</td>
<td>2346</td>
</tr>
<tr>
<td>Human</td>
<td></td>
<td></td>
<td></td>
<td>8352</td>
</tr>
</tbody>
</table>

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.

score over all videos. Finally, we want to capture the degree of “discriminativeness” among the descriptions of videos with similar content. ActivityNet [3] includes 200 activity 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 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)\(^6\) 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

---

\(^6\)https://www.mturk.com
**Table 2:** Human evaluation of multi-sentence video descriptions, see text for details.

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

**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.

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.

### 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.
References


