Supplementary Material to: 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

Here we provide implementation details for our approach and some of the baselines (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 $\bar{v}^i_m$ in Equation 1 of the main paper. Unlike image captioning that relies on static features, video description requires a dynamic multimodal fusion over different visual features, such as e.g. stream of RGBs and motion. 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: $v^i_f = [v^i_{f,1}, v^i_{f,2}, ...v^i_{f,T_f}]$ [9, 11]. The previous hidden state h^i_{m-1} is used to predict temporal attention [12] over these segments, which then results in a single feature vector $\hat{v}^i_{m,f}$. We concatenate the resulting vectors from all features as our final visual input to the decoder: $\bar{v}^i_m = [\hat{v}^i_{m,1}, \hat{v}^i_{m,2}, ..., \hat{v}^i_{m,f}, ...]$.

Self-Critical Sequence Training. Self-Critical Sequence Training [6] (SCST)[1] is a variant of REINFORCE [10] where the inference algorithm is used as a baseline. Suppose we have a generator model G_θ with parameters θ; a complete sequence $x^s = (x^s_1, ...x^s_T)$ is sampled using the probability distribution $p_\theta(x^s_t|x^s_1, ...x^s_{t-1})$ 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 [8]. The gradient function for the model is calculated as:

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

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

GANs for Captioning. GANs for image captioning [2, 7] 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 of the main paper, with $\alpha = 0.5, \beta = 0.5$. After both G_θ and D_η have been pre-trained, we follow [1, 4] 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:

$$\nabla_\theta L_{G_\theta}(\theta) = \sum_{t=1}^{T} (D_\eta(V, x^s) - D_\eta(V, \hat{x}))\nabla_\theta \log p_\theta(x^s_t|x^s_{1:t-1}). \tag{2}$$

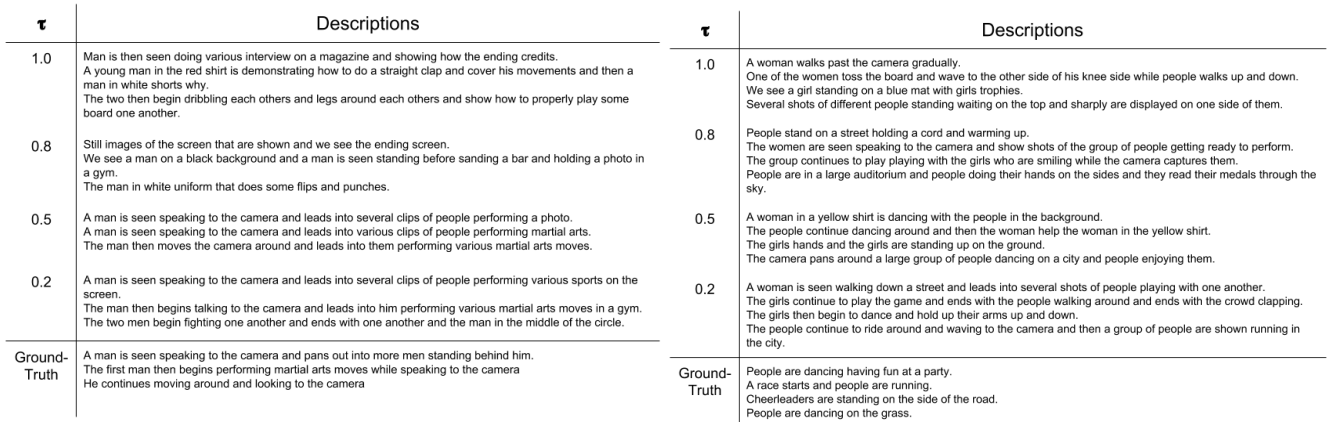
Due to instability of adversarial training, we additionally include cross entropy (CE) loss that ensures that the generator will explore an output space in a more stable manner and maintain its language model [5]. The final objective of G_θ is a mixed loss function, a weighted combination of Cross-Entropy Loss (L_{CE}) optimizing the maximum-likelihood training objective and Adversarial Loss (L_{GAN}) with its gradient function defined in Equation 2:

$$L_{\text{MIX}} = \lambda L_{\text{GAN}} + (1 - \lambda) L_{\text{CE}}, \tag{3}$$

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:

$$p_\tau(w_i) = \frac{p(w_i)^{1/\tau}}{\sum_{j=1}^{K} p(w_j)^{1/\tau}}. \tag{4}$$



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 1: 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.

Based on Equation 4, $\tau = 1$ 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 1.

B. Qualitative Examples

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

B.1. Comparison to Model Ablations and GAN

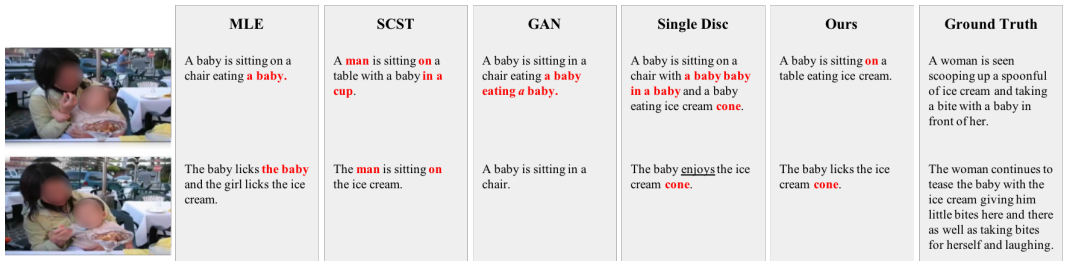
Figure 2 shows a few qualitative examples comparing ground truth descriptions to the ones generated by the following methods: MLE, SCST (with CIDEr), GAN, MLE+SingleDis (Single Disc), and our MLE+HybridDis (Ours). We highlight errors, *e.g*. objects not present in video, in bold/red, and repeating phrases 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 preferred 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 over-report "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 3, 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 4. 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).

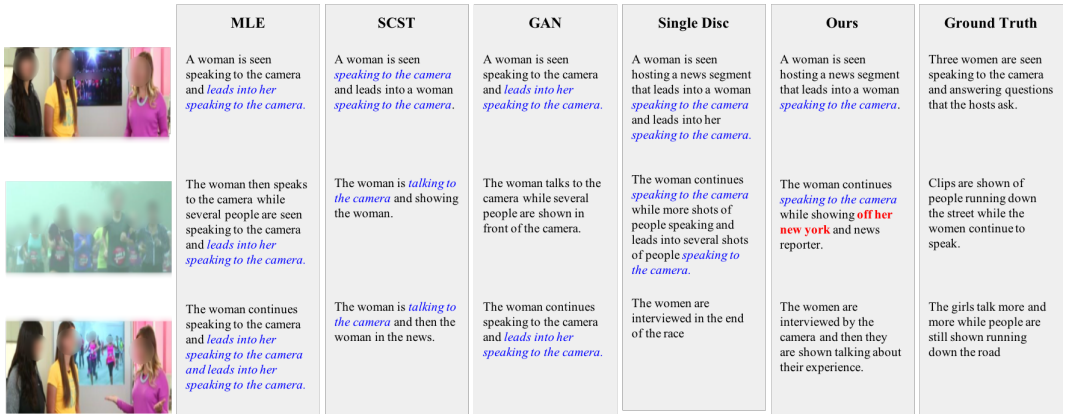
B.2. Comparison to State-of-the-Art

Figure 5 provides a comparison of descriptions obtained by our approach to three recent video description models (VideoStory [3], Transformer [13], MoveForwardTell [11]).



(a)

(b)



(c)

Figure 2: 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.

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

	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.

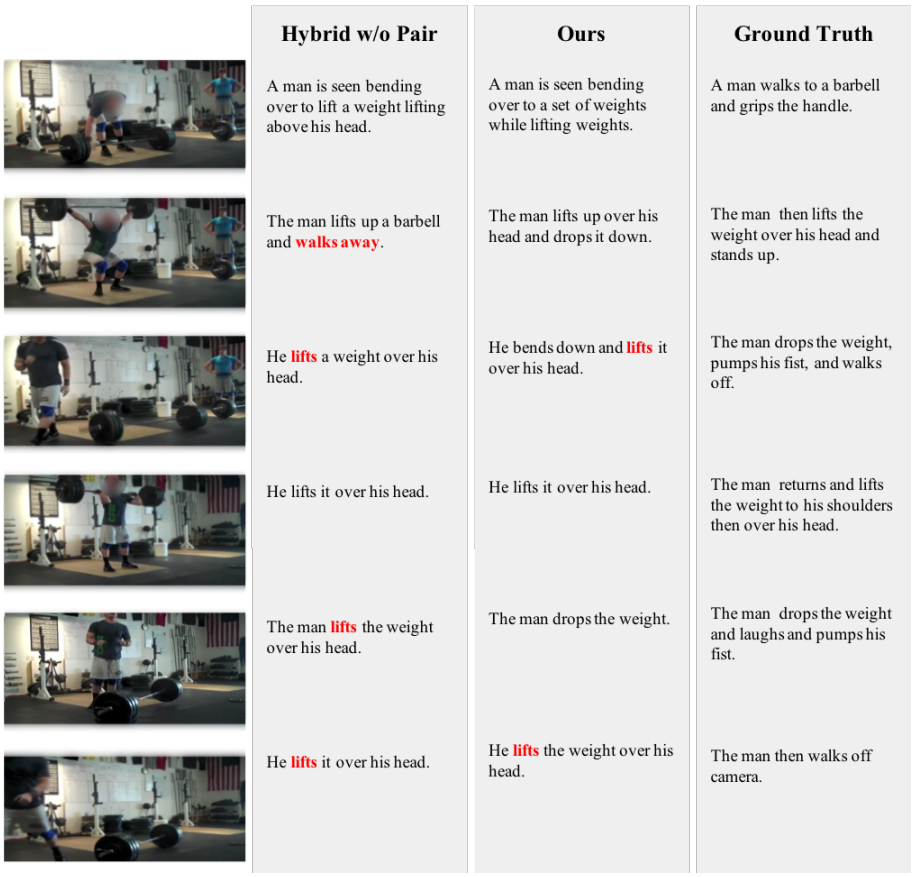
(b)

	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.

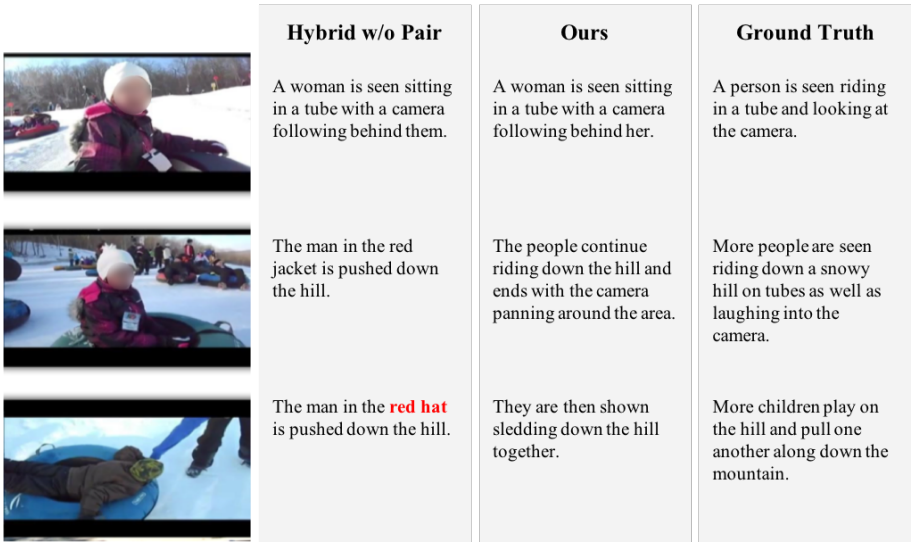
(c)

Figure 3: Comparison of our approach to MLE baseline, SCST, and GAN. Red/bold indicates content errors, blue/italic indicates repetitive patterns.

still a large room for improvement w.r.t. the human ground-truth descriptions.

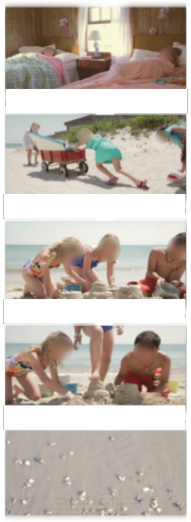


(a)



(b)

Figure 4: 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* .

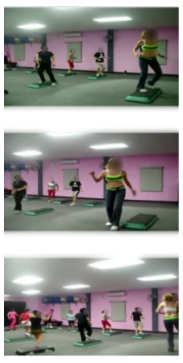


Video Story	Transformer	MFT	Ours	Ground Truth
A **woman** is seen **standing in a chair** and leads into her holding **a box** and looking off into the distance.	A **woman** is seen sitting on a **beach** followed by a **woman putting sunscreen on a beach.**	A **man** is seen sitting on a **chair** and *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.

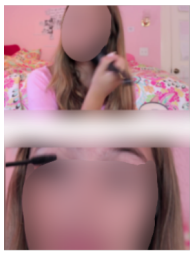
(b)



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.

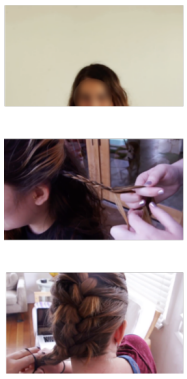
(c)

Figure 5: Comparison of our approach to state-of-the-art video description approaches (VideoStory [3], Transformer [13], MoveForwardTell [11]). 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 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 6: Failure cases of our approach and state-of-the-art video description approaches (VideoStory [3], Transformer [13], MoveForwardTell [11]). Red/bold indicates content errors, blue/italic indicates repetitive patterns.

B.3. Failure Analysis

This brings us to the last section, where we analyze the typical 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 2 (a)), incorrectly mentions "showing off her new york" (Figure 2 (c)), predicts "man" instead of a woman (Figure 3 (b)) and "woman" instead of a child (Figure 5 (a)) or lifting instead of dropping

(Figure 4 (a)), etc. It is also still prone to some repetition (*e.g*. Figure 3 (a), (b), Figure 5 (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 6, 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 6 (a), incorrectly mentions a "hair dryer" and "scissors" in Figure 6 (b), and "vegetables" and "potatos" in Figure 6 (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.

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.