

Diverse Video Captioning by Adaptive Spatio-temporal Attention

Zohreh Ghaderi¹, Leonard Salewski¹, and Hendrik P. A. Lensch¹

University of Tübingen, Tübingen, Germany¹

{zohreh.ghaderi, leonard.salewski, hendrik.lensch}@uni-tuebingen.de

Abstract. To generate proper captions for videos, the inference needs to identify relevant concepts and pay attention to the spatial relationships between them as well as to the temporal development in the clip. Our end-to-end encoder-decoder video captioning framework incorporates two transformer-based architectures, an adapted transformer for a single joint spatio-temporal video analysis as well as a self-attention-based decoder for advanced text generation. Furthermore, we introduce an adaptive frame selection scheme to reduce the number of required incoming frames while maintaining the relevant content when training both transformers. Additionally, we estimate semantic concepts relevant for video captioning by aggregating all ground truth captions of each sample. Our approach achieves state-of-the-art results on the MSVD, as well as on the large-scale MSR-VTT and the VATEX benchmark datasets considering multiple Natural Language Generation (NLG) metrics. Additional evaluations on diversity scores highlight the expressiveness and diversity in the structure of our generated captions.

Keywords: Video Captioning, Transformer, Diversity Scores

1 Introduction

The interplay between visual and text information has recently captivated scientists in the field of computer vision research. Generating a caption for a short video is a simple task for most people, but a tough one for a machine. In particular, video captioning can be seen as a sequence-to-sequence task [40] similar to machine translation. Video captioning frameworks aim to learn a high-level understating of the video and then convert it into text.

Since deep learning has revolutionized almost all computer vision sub-fields, it also plays a notable role in video description generation. Usually, two main Deep Neural Networks are involved, the encoder analyses visual content and the decoder generates text [17,40,47,53,46]. The employed networks often are a variety of 2D-CNN and 3D-CNNs. They extract visual features and local motion information between successive frames. Furthermore, a Faster RCNN object recognition (FRCNN) [37] can be used to obtain fine-grained spatial information.

Attention mechanisms are adopted to let the model build relations between local and global temporal and spatio-temporal information. This information is

subsequently fed to a recurrent neural network such as an LSTM or a GRU in order to produce grammatically correct sentences [1,52,57,32]. The temporal processing is, however, somehow limited as it either involves global aggregation with no temporal resolution or is based on 3D-CNNs with a rather small temporal footprint. Transformer-based encoder-decoder architectures, on the other hand, can inherently establish relations between all components in a sequence independent of their positions [44]. After their breakthrough on language tasks, lately, transformers have been successfully applied to diverse vision applications, mainly for pure classification [3,29].

In this work, we present VASTA, an end-to-end encoder-decoder framework for the task of video captioning where transformers perform detailed visual spatio-temporal analysis of the input as well as generating the caption output.

Our encoder architecture is adopted from the Video Swin Transformer [30], which has been shown to be able to interpret non-local temporal dependencies in video-based action recognition. The task of video captioning, however, requires even more than just spatio-temporal analysis. It needs to extract all semantically relevant concepts [38,34], which will be key for the downstream text generation. We identify all relevant concepts in the captions of the training data sets and fine-tune the Swin Transformer to explicitly predict those before handing latent information to the BERT generator.

As end-to-end training of two transformers, in particular for video processing, is quite involved, we introduce an adaptive frame sampling (AFS) strategy that identifies informative keyframes for caption generation.

Our transformer-based encoder-decoder architecture harnesses the power of transformers for both the visual analysis as well as for the language generating part, rendering quite faithful descriptions to a broad range of videos. In summary, our contributions are:

- a) a simple transformer-based video captioning approach, where a *single* encoder extracts all necessary spatio-temporal information. Unlike other recent works we do not employ disjoint 2D analysis (e.g. object detection) and 3D analysis (e.g. 3D convolution).
- b) adaptive frame sampling for selecting more informative frames
- c) visually grounded semantic context vectors derived from all captions of each sample provide high-quality semantic guidance to the decoder
- d) state-of-the-art results on three datasets (MSVD, MSR-VTT and VATEX)
- e) significantly increased diversity in the predicted captions.

2 Related Work

Existing video captioning approaches can be grouped according to the techniques used for visual analysis and text generation. See Aafaq et al. [2] for a detailed survey.

2.1 Classical Models

Classical models mainly concentrate on detecting objects, actors and events in order to fill the SVO (SVOP) structure [41]. Detection of objects and humans was accomplished by model-based shape matching including HOG, HOF and MbH, [15,48,49]. The analysis part is typically weak on interpreting dynamics and the output of these models is limited due to their explicit sentence structures [25,24,27,20,26]. Classical models have recently been outperformed by models based on deep learning.

Spatio-temporal Analysis with CNNs On the encoder side, a more fine-grained analysis of the spatio-temporal aspects has been enabled with the advent of 3D-convolutions and the corresponding C3D network [42]. Li et al. [53] present a 3D-CNN network for fine local motion analysis and employ soft-attention [4] for adaptive aggregation to obtain a global feature for video captioning. Methods like [1,13,50,39] combine both 2D and 3D-CNNs with attention mechanisms to obtain stronger spatial-temporal and global features. As the video captions often reflect some temporal relation between specific objects, the use of explicit object detectors [32,54,58,57] can improve the generated descriptions. Recently, the work on MGCMP [10] and CoSB [43] illustrates that extracting fine-grained spatial information followed by propagation across time frames could provide visual features as good as other methods that use external object detector features as long as their relation is adequately realized in the temporal domain. While temporally resolved features are necessary to analyse the dynamics, global aggregates can provide the proper semantic context for generating captions. In [38,34,8], semantic attributes are learned inside a CNN-RNN framework. In contrast to the work of Gan et al. [18] we condition our decoder on the semantic context vector once instead of feeding it to the decoder at every step. Still, the self-attention operation of our decoder allows it to be accessed whenever it is needed. It has been shown that selecting informative frames aids in video action recognition [19] as this reduces the overall processing cost and lets the system focus on the relevant parts only. The frame selection could be trained to optimize the input for the downstream task as in [12,19] but this would introduce further complexity to controlling the entire pipeline. In contrast to [12,19] our method is simpler and does not require to be learned.

Transformer-based Models Following the success of transformers [44] in text-related sequence-to-sequence tasks like translation, they have recently also been applied to vision tasks and in particular to video classification. A key ingredient of transformers is the multi-head self-attention mechanism where each head individually can attend and combine different parts of the sequence. This way, a transformer can explore both long-term and short-term dependencies in the same operation. For the task of action recognition, the ViViT transformer [3] chops the video cube into separate spatio-temporal blocks, applying multi-head self-attention. To keep the complexity at bay, factorized attention alternates between

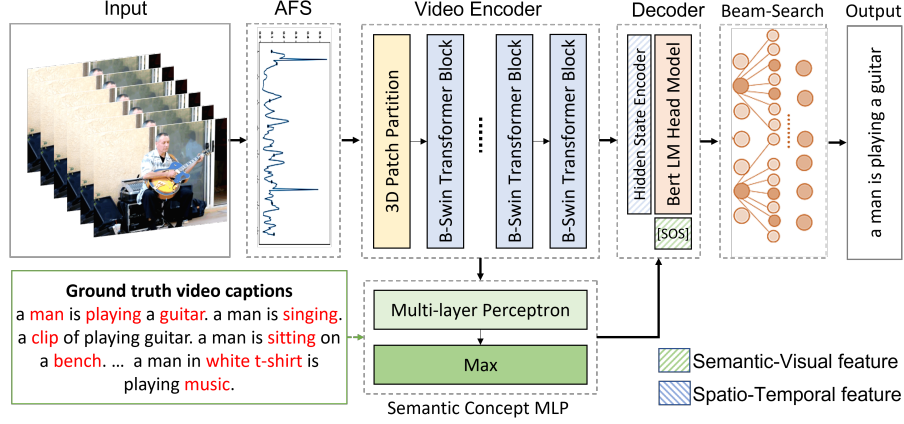


Fig. 1: VASTA (diverse Video captioning by Adaptive Spatio-Temporal Attention) The most informative 32 frames are selected by adaptive frame sampling (AFS) as input to a Swin transformer [30] for spatio-temporal video analysis. The output tokens of the encoder are used twice, once, to predict a semantic concept vector aggregating the entire sequence as start-of-sequence token for the BERT decoder, and, second, for cross-attention in the decoder. A beam search on the most likely words predicts the final caption.

relating temporal or spatially-aligned blocks. The video Swin Transformer [30] overcomes the problem of hard partition boundaries by shifting the block boundaries by half a block in every other attention layer. Instead for action recognition, we use the Swin transformer for video captioning. On a different task, Zoha et al. [59] employ transformer in video dense captioning with long sequences and multiple events to be described. The concept of cross-attention can easily fuse the information coming in from different feature extractors. TVT [9] uses a transformer instead of a CNN-RNN network for the video captioning task. They use attentive-fusion blocks to integrate image and motion information. The sparse boundary-aware transformer method [21] explicitly performs cross-attention between the image domain and extracted motion features. In addition, a specific scoring scheme tries to tune the multi-head attention to ignore redundant information in subsequent frames. Unlike these works we do not require special fusion, as we use a single joint encoder of the input video.

3 Model Architecture

The composed architecture of our VASTA model, visualized in Figure 1, is based on an encoder-decoder transformer [44]. First, an adaptive selection method is used to find informative frames in the whole video length. Thereafter, the model encodes the selected video frames into a contextualised but temporally resolved embedding. We modify the Swin Transformer block [30], which was originally

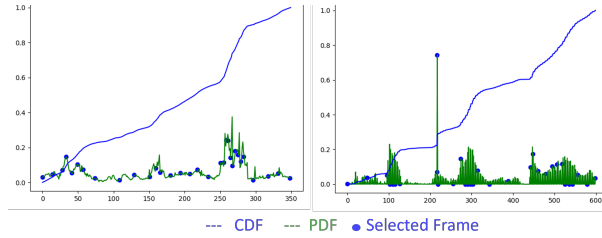


Fig. 2: Adaptive frame selection visualized for two videos from the MSR-VTT. x-axis: video length, y-axis: PDF and CDF driven from LPIPS sampling.

designed for action recognition and classification tasks, to interpret the input video. The last hidden layer of this encoder is passed to the decoder. Though compressed, the output of the encoder still contains a temporal sequence. This allows the BERT [16] decoder to cross-attend to that sequence when generating the output. Besides the direct encoder-decoder connection, the encoder output is further used to predict a globally aggregated semantic context vector that is to condition the language generator.

3.1 Adaptive Frame Selection

In videos, not all frames contribute the same information to the final caption. Some frames are rather similar to the previous ones while some contain dynamics or show new objects. In most video captioning approaches, frames are selected with fixed uniform intervals [13,42,57].

Our adaptive frame selection (AFS) performs importance sampling on the input frames based on local frame similarity. First, the similarity for each pair of two consecutive frames is computed by the LPIPS score [55]. As indicated in Figure 2, we consider this similarity as a probability density function (PDF) f , normalizing it over all frames. Computing and inverting the cumulative density function (CDF) F with

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt, \text{ for } x \in \mathbb{R} \quad (1)$$

one can sample N frames i according to f starting with a uniform distribution $j = \{0, \frac{1}{N}, \frac{2}{N}, \dots, \frac{N-1}{N}\}$, $i = \text{round}(F^{-1}(j))$.

We select the frame i by rounding to the nearest integer. The resulting sequence forms the input to the 3D patch partition of the encoder.

3.2 Encoder

The Swin architecture is a hierarchical transformer that is able to act as general-purpose backbone in computer vision tasks [29]. The vision input is initially split into non-overlapping patches, followed by **Shifting** these **Windows** by half the patch size in every other self-attention step to avoid artifacts from the discrete

partitioning. Consequently, the swin video transformer [30] operates on shifted 3D partitions to cover both the spatial and the temporal domain. The encoder backbone is the Swin-B variant with last the hidden layer size $(BS \times 16 \times 1024)$.

3.3 Semantic Concept Network

The second step of our pipeline includes predicting a semantic concept vector of the entire video that is used as the start-of-sequence token in the decoder to condition its generation on it. The training signal for this concept vector aggregates *all* ground truth captions of a video to provide a high-quality signal for the caption generation. The concepts are defined by the K most frequent words found in the captions of the entire data set. Predicting the semantic concept vector of a video is learned as a binary multi-class classification task indicating which of the frequent words are relevant to describe the video. For generating the ground truth classification vectors, we first select nouns, verbs and adverbs¹ from all captions of the training videos (see Figure 1). Each video is labeled with the K -dimensional vector L by:

$$L_k = \begin{cases} 1, & \text{if word } k \text{ occurs in any caption} \\ 0, & \text{otherwise} \end{cases}$$

An MLP acts separately with shared weights on each of the encoder outputs. A single max-pooling layer merges them to one token. Afterwards a two-layer MLP with RELU activation predicts the K concepts. For training, the binary cross-entropy loss is minimized. In essence, the probability of each word in the concept dictionary is used as a semantic feature. Introducing the semantic concept vector provides an aggregated constant signal for each video while the decoder is trained by generating individual captions.

3.4 Decoder

Our decoder generates the language output word by word while self-attending to all already generated tokens. During training, masking ensures that BERT cannot access the future tokens while during inference it is auto-regressive.

In our architecture, we pass the semantic feature vector as start-of-sequence token to the decoder for generating the first output word. The self-attention to the first input token conditions the predicted caption on the semantic concepts which have been identified by the semantics MLP. In order to couple the hidden states of the decoder to the output of the encoder, cross-modal fusion is necessary. The necessary functionality is incorporated by extending the decoder with multi-head cross-attention [44]. In 13 layers, the architecture alternates between multi-head self-attention on the language tokens, cross-attention between the language tokens and the Swin output tokens, and a feed-forward layer followed by normalization. All these steps are bridged by a residual connection. A final

¹ categorizing and POS tagging using NLTK (<https://www.nltk.org/>)

linear layer projects the decoder internal hidden states to the size of the BERT vocabulary, followed by a softmax to produce word probabilities for each token. The sentence is finally generated by applying beam search [36] to obtain the most likely combination of words.

4 Experiments

To show the effectiveness of our proposed architecture, we train our model on three common video captioning data sets and achieve high-ranking results. Details on the architecture, training and optimizer [31] settings are described in the supplementary.

4.1 Datasets and Metrics

Our model is trained on MSR-VTT [51], MSVD [7] and VATEX [50]. **MSVD** [7] includes 1970 videos with an average length of 10 seconds. It provides up to 45 captions per video which are randomly sampled during training. **MSR-VTT** [51] contains a wide variety of open domain videos of 20 seconds average and 20 captions per video. **VATEX** [50] is a large-scale video description dataset. It is lexically richer, as each video has 10 unique sentences and every caption is unique in the whole corpus.

Similarity Metrics All captions in all datasets have been annotated by humans. We utilize the MS COCO Caption Evaluation [11] protocol on both datasets and evaluate standard natural language generation metrics (NLG) as done by previous works in this field. These metrics include BLEU(4-gram)(B4) [33], METEOR(M) [5], CIDEr(C) [45] and ROUGE-L(R) [28]. Additionally, we evaluate on BERTScore [56], a more modern evaluation metric, that has shown great alignment with human judgement. Notably, it does not share the same brittleness as the n -gram based metrics [33].

Diversity Metrics In contrast to previous work we further measure the diversity of the generated captions. Zhu et al. [60] introduced Self-BLEU (SB) to evaluate the diversity of generated sentences. It measures the BLEU value of each prediction wrt. the remaining predictions and averages the obtained BLEU values. A lower value indicates higher diversity, as on average a prediction is less similar to all other predictions. Furthermore, Dai et al. [14] proposed the concepts of Novel Captions (N), Unique Caption (U) and Vocab Usage (V) to evaluate diversity of the generated caption. Novel Caption shows the percentage of generated captions which have not been seen in the training data; Unique Caption denotes the percentage of distinct captions among all generated captions; Vocab Usage indicates the percentage of words that are used to generate captions from the whole vocabulary.

4.2 Quantitative Results

We present quantitative results in Tables 1 and 2 and highlight that besides explanation quality also explanation diversity is important (Table 3). We ab-

	Method	Model	Year	B4 \uparrow	M \uparrow	C \uparrow	R \uparrow	BERT-S \uparrow
	Shared Enc-Dec [50]	A	2019	28.4	21.7	45.1	47.0	-
	NITS-VC [39]	A	2020	20.0	18.0	24.0	42.0	-
	ORG-TRL [57]	A	2020	32.1	22.2	49.7	48.9	-
	VASTA (Kinetics-backbone)	T	2022	36.25	25.32	65.07	51.88	90.76

Table 1: Natural Language Generation (NLG) and BERT scores for VATEX.

late the components of our model in Table 4 and discuss qualitative examples in Section 4.4.

Comparison to Related Approaches On the very large VATEX data set our generated captions show significant performance improvements on all scores (see Table 1). Similarly, on MSR-VTT and the even smaller MSVD we obtain high-ranking, most often top-scoring results with slightly less improvements (see Table 2). This indicates that fine tuning of the encoder and decoder transformers benefits from the additional training data.

Thus, instead of just fine-tuning the full pipeline starting with the backbone trained on Kinetics [22] for each individual data set, we trained once end-to-end on VateX and then fine-tuned for MSVD and MSR-VTT. Through this transfer learning VASTA improves in general, particularly the CIDEr score on MSR-VTT and MSVD by a big margin. The performance on METEOR and CIDEr is relevant as both consider semantic relatedness. METEOR excepts synonyms and it exhibits a higher correlation with human judgment on captions and explanations [23]. The CIDEr score has been particularly designed for measuring the quality of descriptions for visual content. While the NLG scores are all based on n -grams the BERTScore is more semantically robust and agrees even better with human assessments. On both data sets, our model achieves the highest BERTScore.

Caption Diversity Albeit their wide-spread use, NLG metrics only assess a few aspects of the generated captions. Maximising the scores on existing NLG metrics as presented in Table 2 can for example be achieved with focusing on the most prevalent sentence structure found in the ground truth captions. However, we are interested in captions that are the most “human-like”. Thus, we compute these diversity metrics on MSR-VTT, MSVD and VATEX and compare our model to those competitors where we have access to the generated captions for re-evaluation. As seen in Table 3, our model not only predicts highly accurate captions but also manages to predict a highly diverse set of captions. VASTA generates the most distinct captions and does not overfit to the training data, i.e. generates novel captions for some of the test videos. Our model by far exploits most of the training vocabulary. Further analysis on the diversity of sentence structures is given in the supplementary.

Method	Model	Year	MSR-VTT					MSVD				
			B4	M	C	R	BERT-S	B4	M	C	R	BERT-S
Att-TVT [9]	T	2018	40.12	27.86	47.72	59.63	-	53.21	35.23	86.76	-	-
GRU-EVE [1]	A	2019	38.3	28.4	48.1	60.7	-	47.9	35.0	78.1	71.5	-
OA-BTG [54]	A	2019	41.4	28.2	46.9	-	-	-	36.9	36.2	90.6	-
STG [32]	A	2020	40.5	28.3	47.10	60.9	-	52.2	36.9	93.0	73.9	-
STATS [13]	A	2020	40.1	27.5	43.4	60.4	-	52.6	33.5	80.2	69.5	-
SAAT [58]	A	2020	40.5	28.2	49.1	60.9	82.50	46.5	33.5	81.0	69.4	-
ORG-TRL [57]	A	2020	43.6	28.8	50.9	62.1	-	54.3	36.4	95.2	73.9	-
SAVCSS [8]	A	2020	43.8	28.9	51.4	62.4	90.00	61.8	37.8	103	76.8	91.25
DSD-3 DS-SEM [38]	A	2020	45.2	29.9	51.1	64.2	-	50.1	34.7	76	73.1	-
SBAT [21]	T	2020	42.9	28.9	51.6	61.5	-	53.1	35.3	89.5	72.3	-
SemSynAN [34]†	A	2021	46.4	30.4	51.9	64.7	82.13	64.4	41.9	111.5	79.5	82.67
MGCMP [10]	A	2021	41.7	28.9	51.4	62.1	-	55.8	36.9	98.5	74.5	-
CoSB [43]	T	2022	41.4	27.8	46.5	61.0	-	50.7	35.3	97.8	72.1	-
VASTA (Kinetics-backbone)	T	2022	43.4	30.2	55.0	62.5	90.10	56.1	39.1	106.4	74.5	92.00
VASTA (Vatex-backbone)	T	2022	44.21	30.24	56.08	62.9	90.17	59.2	40.65	119.7	76.7	92.21

Table 2: Natural Language Generation (NLG) and BERT scores for the MSR-VTT and MSVD datasets (T: Transformer, A: Attention). Darker blue indicates higher scores. For both data sets our approach improves the BERTScore and produces high-ranking NLG scores. †: BERT-score is computed on reproduced captions by the released code.

Method	MSR-VTT				MSVD				VATEX			
	SB ↓	N ↑	U ↑	V ↑	SB ↓	N ↑	U ↑	V ↑	SB ↓	N ↑	U ↑	V ↑
SAVCSS [8]	95.19	44.61	33.44	1.88	84.32	51.34	42.08	2.07	-	-	-	-
SAAT [58]	99.99	40.46	20.06	1.33	-	-	-	-	-	-	-	-
SemSynAN [34]	96.47	42.84	18.92	1.57	83.00	47.16	37.61	2.19	-	-	-	-
VASTA (Kinetics-backbone)	92.94	45.98	34.74	2.93	81.88	30.49	42.89	3.48	86.18	97.29	85.80	7.04
VASTA (Vatex-backbone)	92.70	45.51	34.21	3.00	76.90	42.75	52.16	3.94	-	-	-	-

Table 3: Diversity of the generated captions.

Analyzing the performance of SemSynAN [34] (a model which has strong similarity metrics) where the number of captions per video is limited to just five to train a predictor for the most common syntactic POS structures (see Supplementary) reveals that this comes at the cost of reduced caption diversity, sentence quality and video-caption match. Thus, we found that its diversity is much lower.

4.3 Ablation Study

The results of Table 2 have been achieved by carefully designing our adaptive spatio-temporal encoder-decoder framework. As demonstrated by the ablation results in Table 4, introducing adaptive frame sampling (AFS) helps improve the image-description related CIDEr score while adding the semantic concept vector further improves on the more translation-related scores (BLEU-4, METEOR, ROUGE-L, CIDEr). Similarly, both AFS and the semantic concept prediction improves the diversity score. Thus, more informative and more precise encoder predictions support higher quality and more diverse language output. In the

Method	MSR-VTT								VATEX							
	B4 ↑	M ↑	C ↑	R ↑	SB ↓	N ↑	U ↑	V ↑	B4 ↑	M ↑	C ↑	R ↑	SB ↓	N ↑	U ↑	V ↑
UFS-SB	43.21	29.55	52.91	62.1	96.48	37.09	19.23	1.90	35.31	25.05	63.82	51.27	87.73	97.62	81.41	6.70
AFS-SB	43.07	29.72	55.08	62.02	93.93	38.29	27.95	2.46	35.64	25.43	64.98	51.53	88.57	97.51	83.33	6.50
UFS-SBS	43.51	29.75	53.59	62.27	94.82	42.44	26.48	2.36	35.68	25.42	65.63	51.58	88.40	97.35	83.38	6.43
AFS-SBS	43.43	30.24	55.00	62.54	92.94	45.98	34.74	2.93	36.25	25.32	65.04	51.88	86.18	97.29	85.80	7.04

Table 4: Influence of the individual components in VASTA. Applying both, AFS and semantic vectors, yields favorable scores. UFS: uniform frame selection, AFS: Adaptive frame selection, SB: Swin BERT, SBS: Swin BERT Semantics.

supplemental we demonstrate how the results depend on the chosen decoder model, by replacing Bert by GPTNeo Causal LM [6,35]. There, we also study different inference methods (beam search, greedy, top-k, top-p).



Fig. 3: Adaptive Frame Selection (AFS). Uniform sampling (top) keeps frames with repetitive non-informative content (cf. frames with foliage). In contrast, adaptive sampling enhances the diversity of input frames by selecting those with activity (cf. frames with people walking). Ground truth: “Two men walking around a forest by a lake.”

AFS Results Figure 3 exemplifies the effect of our adaptive frame selection approach. The video transformer can only take in 32 frames. Driven by the frame differences, the adaptive frame selection samples more diverse frames than simple uniform subsampling, increasing the chance of selecting informative time steps. An irregular temporal sampling pattern on the other hand leads to a non-uniform play-back speed. Still, AFS consistently improves the CIDEr result (Table 4), indicating that the gain in input information has a more positive effect than potential introducing temporal disturbance in the Swin transformer inference.

Semantic Concept Vectors The accuracy (BCE score) of the multi-class extraction task for the semantic concept vectors is 0.88 (0.12) on the training set and 0.85 (0.15) on the test set. This indicates, that this training step generalizes

	BCE	B4 \uparrow	M \uparrow	C \uparrow	R \uparrow
10%-best	55.67	41.25	109.4	75.44	
10%-worst	31.50	21.97	22.76	49.09	

Table 5: Dependency on the quality of the predicted semantic vector. Sorting all test samples of the MSR-VTT wrt. the classification accuracy of the proposed semantic vector, a strong correlation with the evaluation scores is revealed.

well. Introducing the semantic concept vectors improves the overall performance, as one can see by the strong correlation between classification accuracy and resulting NLG scores in Table 5. Bad examples most often occur in conjunction with misclassification of the main actors or concepts in the video. In these cases, often the content of the video is not well represented by the most common 768 concepts.



Reference: a group of people are dancing and singing
Our: a group of people are dancing and singing
SymsynAN: a group of people are dancing



Reference: a dog is playing on a trampoline
Our: a dog is playing on a trampoline
SymsynAN: a dog is playing with a dog



Reference: a person is making a paper aircraft
Our: a person is making a paper airplane
SymsynAN: a person is folding paper

Fig. 4: Examples for the top-performing videos in the test set.

4.4 Qualitative Results

In Figure 4, representative videos and the generated captions are shown. We list examples of the top 1%-percentile on the METEOR score. For positive examples, the content of the video is fully recognized leading to a description that matches one of the reference captions almost exactly. In the bad examples (see supplementary) the content is often misinterpreted or the video is so abstract that there could be many diverse explanations.

Spatio-Temporal Attention The video in Figure 5 on the top features a complex temporal interaction between two actors (*monkey and dog*). The generated caption correctly reflects both spatial detail (*dog's tail*) as well as multiple



Fig. 5: Spatio-temporal inference gathers information from different segments.

temporal stages (*grabbing the tail* and *running away*). Similarly, the temporal domain is also respected in the second example. Different parts of the video contribute to different sections in the generated captions (*woman talking about food* and *cooking in pot*). These examples demonstrate that high-quality detection and tracking from the Swin transformer across multiple frames goes hand-in-hand with the powerful language skills of the fine-tuned generator in our proposed framework.

5 Limitations and Discussion

The introduced VASTA architecture performs quite well according to the commonly used evaluation metrics. Even though our model has the best diversity scores, looking at some samples of generated captions, they are often rather general and might miss some important detail about the video. This suggests, further research in the diversity aspect is important.

As indicated in Section 4.4, the current extraction of concepts for video captioning might need further improvement, potentially by the use of a larger training data set. Compared to the very good performance of the produced language, the visual analysis is not yet on par. The training and the evaluation are done on three data sets MSR-VTT, MSVD and VATEX, which come with their own distributions of scenes, people, objects, and actions. Any marginalization of specific social groups present in the data sets will likely also be present in our trained encoder-decoder framework. In general, our approach to automated video analysis and captioning might furthermore be trained and applied in other contexts. While we think that the application to the presented data is not problematic, ethical issues can quickly arise in surveillance or military applications.

6 Conclusion

We presented VASTA, a video captioning encoder-decoder approach which incorporates the processed visual tokens by multi-layer multi-head cross-attention.

While the Swin tokens extracts separate spatio-temporal information, we introduce also a globally aggregating semantic concept vector that initializes the sentence generation in the BERT module. By proposing a content-based adaptive frame selection sampling, we can assure that the most informative frames are selected while maintaining efficient training.

This transformer-based video captioning framework introduces a new architecture that produces plausible captions that are state-of-the-art on the MSR-VTT, MSVD and VATEX benchmark data sets. Our evaluation highlights that the commonly used NLG metrics only address some of the aspects necessary to fully assess the quality of video descriptions. We demonstrate that our method produces highly diverse captions. We hope that this work will inspire further research for a better, broader assessment of the performance of caption generation algorithms.

7 Acknowledgements

This work has been supported by the German Research Foundation: EXC 2064/1 – Project number 390727645, the CRC 1233 - Project number 276693517, as well as by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Zohreh Ghaderi and Leonard Salewski.

References

1. Aafaq, N., Akhtar, N., Liu, W., Gilani, S.Z., Mian, A.: Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12487–12496 (2019)
2. Aafaq, N., Mian, A., Liu, W., Gilani, S.Z., Shah, M.: Video description: A survey of methods, datasets, and evaluation metrics. *ACM Comput. Surv.* **52**(6) (Oct 2019). <https://doi.org/10.1145/3355390>, <https://doi.org/10.1145/3355390>
3. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691* (2021)
4. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014)
5. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
6. Black, S., Gao, L., Wang, P., Leahy, C., Biderman, S.: Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. If you use this software, please cite it using these metadata **58** (2021)
7. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 190–200 (2011)

8. Chen, H., Lin, K., Maye, A., Li, J., Hu, X.: A semantics-assisted video captioning model trained with scheduled sampling. *Frontiers in Robotics and AI* **7** (2020)
9. Chen, M., Li, Y., Zhang, Z., Huang, S.: Tvt: Two-view transformer network for video captioning. In: *Asian Conference on Machine Learning*. pp. 847–862. PMLR (2018)
10. Chen, S., Jiang, Y.G.: Motion guided region message passing for video captioning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 1543–1552 (2021)
11. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015)
12. Chen, Y., Wang, S., Zhang, W., Huang, Q.: Less is more: Picking informative frames for video captioning. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 358–373 (2018)
13. Cherian, A., Wang, J., Hori, C., Marks, T.: Spatio-temporal ranked-attention networks for video captioning. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 1617–1626 (2020)
14. Dai, B., Fidler, S., Lin, D.: A neural compositional paradigm for image captioning. *NIPS* **31**, 658–668 (2018)
15. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: *European conference on computer vision*. pp. 428–441. Springer (2006)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018)
17. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2625–2634 (2015)
18. Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., Deng, L.: Semantic compositional networks for visual captioning. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 5630–5639 (2017)
19. Gowda, S.N., Rohrbach, M., Sevilla-Lara, L.: Smart frame selection for action recognition. *arXiv preprint arXiv:2012.10671* (2020)
20. Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., Saenko, K.: Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2712–2719 (2013)
21. Jin, T., Huang, S., Chen, M., Li, Y., Zhang, Z.: Sbat: Video captioning with sparse boundary-aware transformer. *arXiv preprint arXiv:2007.11888* (2020)
22. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017)
23. Kayser, M., Camburu, O.M., Salewski, L., Emde, C., Do, V., Akata, Z., Lukasiewicz, T.: e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 1244–1254 (2021)
24. Khan, M.U.G., Zhang, L., Gotoh, Y.: Human focused video description. In: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. pp. 1480–1487. IEEE (2011)

25. Kojima, A., Tamura, T., Fukunaga, K.: Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision* **50**(2), 171–184 (2002)
26. Krishnamoorthy, N., Malkarnenkar, G., Mooney, R., Saenko, K., Guadarrama, S.: Generating natural-language video descriptions using text-mined knowledge. In: *Twenty-Seventh AAAI Conference on Artificial Intelligence* (2013)
27. Lee, M.W., Hakeem, A., Haering, N., Zhu, S.C.: Save: A framework for semantic annotation of visual events. In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. pp. 1–8. IEEE (2008)
28. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: *Text summarization branches out*. pp. 74–81 (2004)
29. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030* (2021)
30. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. *arXiv preprint arXiv:2106.13230* (2021)
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017)
32. Pan, B., Cai, H., Huang, D.A., Lee, K.H., Gaidon, A., Adeli, E., Niebles, J.C.: Spatio-temporal graph for video captioning with knowledge distillation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10870–10879 (2020)
33. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. pp. 311–318 (2002)
34. Perez-Martin, J., Bustos, B., Pérez, J.: Improving video captioning with temporal composition of a visual-syntactic embedding. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 3039–3049 (2021)
35. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
36. Reddy, D.R., et al.: Speech understanding systems: A summary of results of the five-year research effort. Department of Computer Science, CMU, Pittsburgh, PA (1977)
37. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **28** (2015)
38. Shekhar, C.C., et al.: Domain-specific semantics guided approach to video captioning. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 1587–1596 (2020)
39. Singh, A., Singh, T.D., Bandyopadhyay, S.: Nits-vc system for vatex video captioning challenge 2020. *arXiv preprint arXiv:2006.04058* (2020)
40. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: *Advances in neural information processing systems*. pp. 3104–3112 (2014)
41. Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., Mooney, R.: Integrating language and vision to generate natural language descriptions of videos in the wild. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. pp. 1218–1227 (2014)
42. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: *Proceedings of the IEEE international conference on computer vision*. pp. 4489–4497 (2015)

43. Vaidya, J., Subramaniam, A., Mittal, A.: Co-segmentation aided two-stream architecture for video captioning. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 2774–2784 (2022)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: *Advances in neural information processing systems*. pp. 5998–6008 (2017)
45. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4566–4575 (2015)
46. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: *Proceedings of the IEEE international conference on computer vision*. pp. 4534–4542 (2015)
47. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729* (2014)
48. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: *Bmvc 2009-british machine vision conference*. pp. 124–1. BMVA Press (2009)
49. Wang, X., Han, T.X., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: *2009 IEEE 12th international conference on computer vision*. pp. 32–39. IEEE (2009)
50. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: VateX: A large-scale, high-quality multilingual dataset for video-and-language research. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 4581–4591 (2019)
51. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 5288–5296 (2016)
52. Yan, C., Tu, Y., Wang, X., Zhang, Y., Hao, X., Zhang, Y., Dai, Q.: Stat: Spatial-temporal attention mechanism for video captioning. *IEEE transactions on multimedia* **22**(1), 229–241 (2019)
53. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: *Proceedings of the IEEE international conference on computer vision*. pp. 4507–4515 (2015)
54. Zhang, J., Peng, Y.: Object-aware aggregation with bidirectional temporal graph for video captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8327–8336 (2019)
55. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *CVPR* (2018)
56. Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K.Q., Artzi, Y.: BertScore: Evaluating text generation with bert. In: *International Conference on Learning Representations* (2020), <https://openreview.net/forum?id=SkeHuCVFDr>
57. Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., Zha, Z.J.: Object relational graph with teacher-recommended learning for video captioning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 13278–13288 (2020)
58. Zheng, Q., Wang, C., Tao, D.: Syntax-aware action targeting for video captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13096–13105 (2020)
59. Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. In: *CVPR*. pp. 8739–8748 (2018)

60. Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., Yu, Y.: Texygen: A benchmarking platform for text generation models. In: ACM SIGIR. pp. 1097–1100 (2018)

8 Supplementary: Diverse Video Captioning by Adaptive Spatio-temporal Attention.

This supplementary presents model details in Section 8.1, and provides additional ablation studies on our decoder model in Section 8.2. Also, an analysis on the diversity of generated captions is shown in Section 8.4 and demonstrates additional qualitative results in Section 8.3. Training detail of our model is explained in Section 8.5.

8.1 Architecture

In this section, we show technical sketches of the three subnetworks in our VASTA framework and present additional details of the network structures and configurations. Also, our code is available at <https://github.com/zohrehghaderi/VASTA>.

Data Preparation The input videos are read with the mmcv library² choosing the NCTHW input shape format. Each input is resized and cropped to 224×224 resolution followed by normalization with mean of $[123.675, 116.28, 103.53]$ and standard deviation of $[58.395, 57.12, 57.375]$.

The adaptive frame selection (AFS, see Section 3.1 of the main paper) extracts the 32 most informative frames based on LPIPS [55] similarity scores and passes the video with shape $32 \times 224 \times 224 \times 3$ to the Swin encoder.

Data set splits MSVD [7]. Following [7,53] the data set is split into 1200 samples for training, 100 samples for validation and the remaining 670 samples for testing. **MSR-VTT [51].** Following the official setting [51], the data set is split as follows: 6513, 497 and 2990 videos for training, validation and test. **VATEX [50].** This data set officially includes 25991, 3000, 6000 videos for train, validation and test. Unfortunately, roughly 10% of the original set are no longer available for download. Thus, our evaluation in paper is on 23303 videos for training, 2690 videos for validation and 5398 videos for test (see Table table 6).

	MSR-VTT [51]	MSVD [7]	VATEX [50]
Train	6513	1200	23303
Val	497	100	2690
Test	2990	670	5398
Total	10000	1970	31391

Table 6: Train, validation and test splits of the utilized data sets.

² <https://github.com/open-mmlab/mmcv>

Swin Encoder The encoder backbone to our VASTA model is the Swin network, more precisely the Swin-B variant [29]. While the original pipeline is designed for action classification in our context, video description, we modify the network as shown in Figure 6.a, generating 16 output tokens that are reshaped to fit the expected input size of the BERT decoder. More details on the specific configuration and parameters of the Swin transformer [29] are listed in Table 9.

Semantic Context Network The network to extract the semantic context vector is three-layer MLP that is shown in Figure 6.b. The MLP operates on each Swin token individually and then joins them using max-pooling. The resulting vector yields the probability of the most frequent 768 words.

The predicted semantic feature is passed as the start-of-sequence ([SOS]) token to the BERT decoder where it replaces the BERT-embedding layer for the first time step.

BERT Decoder The architecture of the BERT decoder is shown in Figure 7. It follows the traditional BERT architecture [16] with an additional cross-attention layer in the 12 transformer blocks which relates the BERT inference to the encoder output sequence. The specific configuration of the layers and attention heads is given in Table 9.

8.2 Ablation Study

Different Inference Methods. As illustrated in Table 7, we compare frequently used inference methods applied to our VASTA model. Beam search with 3 beams achieves the best results on the MSVD and MSR-VTT data sets.

Method	MSR-VTT				MSVD			
	B4 ↑	M ↑	C ↑	R ↑	B4 ↑	M ↑	C ↑	R ↑
Top-p p=0.95	42.28	29.53	53.00	62.28	54.56	38.52	102.5	74.15
Top-k k=20	19.03	22.18	28.50	46.39	30.96	29.58	57.50	61.48
Top-k k=3	29.59	25.85	39.12	54.16	41.78	33.79	74.73	67.04
Greedy	42.28	29.53	53.00	62.28	54.56	38.52	102.5	74.15
Beam Search b=3	43.43	30.24	55.00	62.54	56.14	39.09	106.3	74.47

Table 7: Influence of the difference inference method on VASTA model.

Ablation on the decoder model. To study the effect of different pre-trainings of our language decoder we try two different decoder models. GPT [35] is an auto-regressive language model whose aim is to predict the next word based

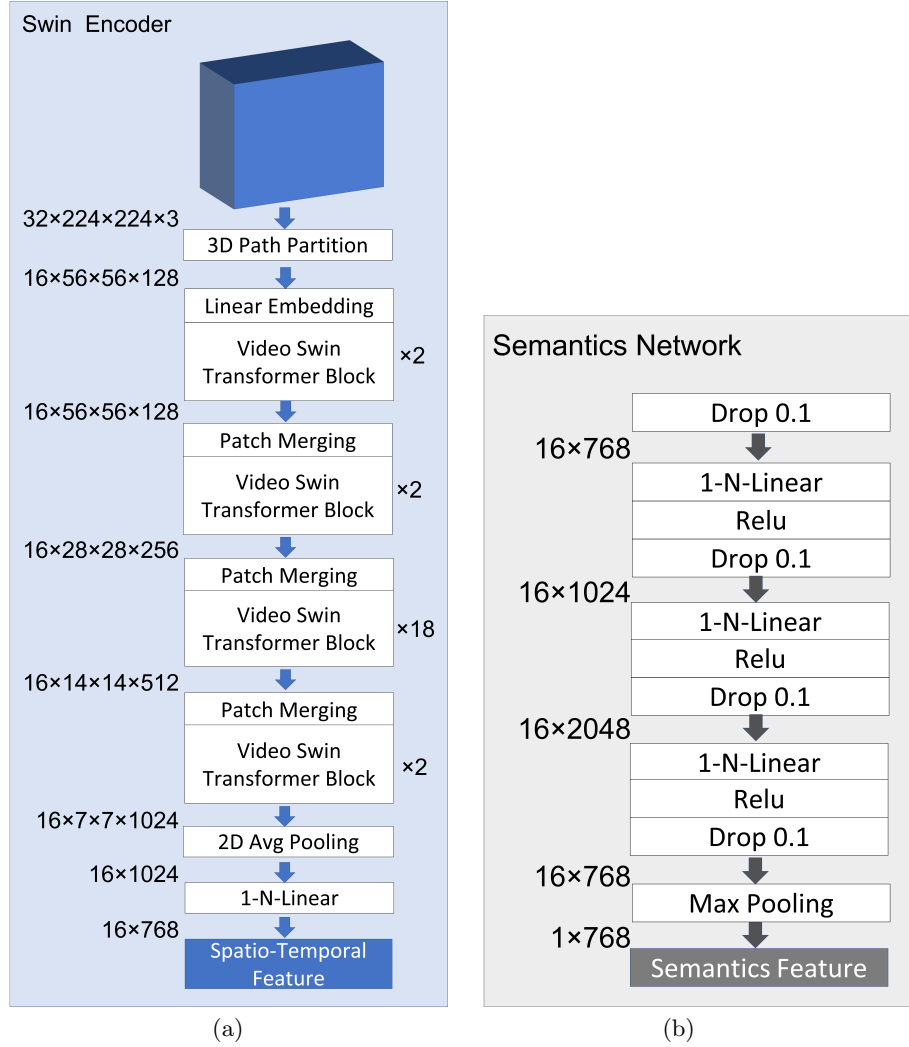


Fig. 6: a) Detailed architecture of our encoder model based on the Video Swin Transformer [30]. b) Detailed architecture of our semantics model. Each token of the Swin transformer output is processed individually by a three-layer MLP before fusing all 16 tokens by a max layer.

on all of previous words. Huge amounts of data are used to train the large number of parameters. Additionally, the size of the GPT model with its 125M parameters limits the ability to combine it with a large Swin network. Thus we use the GPTNeo Causal Language Modelling [6] model, which is similar to GPT2 except that GPTNeo uses local attention in every other layer. We compare the performance when replacing the BERT decoder by GPTNeo in Table 8.

Decoder	MSR-VTT				MSVD				VATEX			
	B4 ↑	M ↑	C ↑	R ↑	B4 ↑	M ↑	C ↑	R ↑	B4 ↑	M ↑	C ↑	R ↑
BERT	43.43	30.24	55.00	62.54	56.14	39.09	106.3	74.47	34.96	25.46	51.33	64.33
GPT-Neo	40.13	28.09	46.91	57.25	50.83	33.80	81.40	65.89	27.25	24.50	47.01	38.79

Table 8: Comparison of BERT and GPTNeo [6] as decoders for our model VASTA. BERT clearly outperforms GPTNeo as a decoder.

We hypothesize, that BERT is trained on mask filling whose aim is to predict a masked word bases on previous words and upcoming words in the sentences. Thus, it is leading to improved sentence understanding which is related to video understanding for caption generation.

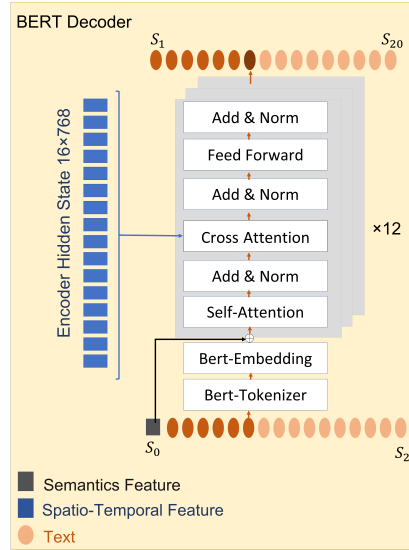


Fig. 7: Detailed architecture of our BERT-based decoder model. The decoder considers the visual input once by using the semantic vector as [SOS] token and second by cross-attending to it in the 12 transformer layers. Note, that the conditioning on the semantic vector bypasses the embedding step.

Swin Encoder		Semantics Network		BERT Decoder	
Type	Size	Type	Size	Type	Size
Patch-size	(2,4,4)	Linear	1024	Hidden-size	768
Depths	[2,2,18,2]	Activation	RELU	Num-hidden-layers	12
Embed-dim	128	Drop out train	0.5	Num-attention-heads	12
Num-heads	[4,8,16,32]	Linear	2048	Intermediate-size	2872
Window-size	(8,7,7)	Activation	RELU	Hidden-act	Gelu
MLP-ratio	4	Drop out train	0.5	Hidden-dropout-prob	0.3
qkv-bias	True	Max pooling	1D	Attention-probs-drop-out-prob	0.3
qk-scale	None	Drop out fine tunig	0.1	Max-position-embeddings	512
Drop-rate	0			Type-vocab-size	2
Attn-drop-rate	0			Initializer-rang	0.02
Drop-path-rate	0.2			Layer-norm-eps	1e-12
Path-norm	True			Vocab size	30522
				Position-embedding-type	absolute
				Pad-token-id	0

Table 9: Parameter sets and configurations for the three subnetworks in our pipeline: the Swin encoder, the semantics network and the BERT decoder.

8.3 Qualitative Examples

To better assess the quality of the captions generated by our model, we present a few more generated examples. Figure 8, Figure 9 and Figure 15 feature some videos of the MSVD and the MSR-VTT data sets, respectively, where introducing the semantic context vector into our pipeline (see Section 4.3) improved the CIDEr score significantly compared to the version of our model without the semantic vector.

Additional top-performing videos based on METEOR and BLUE-4 and their corresponding generated captions are demonstrated in Figure 13 and Figure 14.

AFS Examples In addition, Figure 10 shows representative frames in two videos of MSVD and MSR-VTT where the adaptive frame selection makes a major difference. *Video xxHx6s-DbUo-162-165* is about a man running on the road. Adaptive frame sampling is able to pick up the informative frames about the running movement. Thus, it generates a proper caption. Also, in *Video 7153* AFS selects frames, which contain the wrestling movement, that are key to correctly describing the “wrestling [...] competition” instead of making a broad statement about “playing sports”.

8.4 Caption Diversity

Following the discussion on caption diversity in Section 4.4, we visualize the distribution of generated Part-of-Speech tagging (POS) structures for the captions by our model and by SemSynAN [34] in Figure 11 for the MSR-VTT and Figure 12 for the MSVD data set. While some sentence structures are more frequent than others it can be seen that our approach generates more diverse sentences.

As explained in Section 4.2 of our paper some related works operate on a limited subset of the data sets. As we show in Table 10 working on a subset limits the diversity of the sentences and the size of the vocabulary. Our vocabulary on MSVD is about twelve-thousand words. In contrast, the vocabulary in SemSynAN [34] has only half the size, which will also restrict the diversity of the captions generated by the SemSynAN model.

Data set	VASTA (Ours)	SemSynAN [34]
MSVD	9636	~6K
MSR-VTT	23081	~12K

Table 10: Vocabulary sizes on MSVD and MSR-VTT for our model and SemSynAN [34] showing that our model operates on a much larger set of words.



Video hJFBXhtxKIc-286-291

Reference: a man pours cooked pasta from a plastic container into a bowl CIDEr

AFS-Swin-Bert: a man cooking his kichen 35.04

AFS-Swin-Bert-Semantics: a man is pouring pasta out of a plastic container 75.86

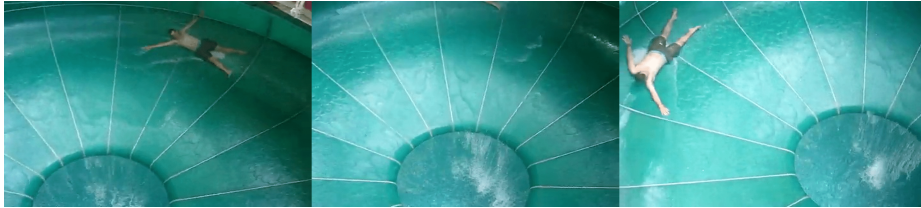


Video qypmR401Gwk-0-10

Reference: a gazelle is fighting with a baboon CIDEr

AFS-Swin-Bert: two zebras are fighting 23.74

AFS-Swin-Bert-Semantics: a cheetah is chasing a gazelle 57.89



Video idXJu0BQRvo-2-6

Reference: a boy is sliding around a water slide CIDEr

AFS-Swin-Bert: a dog is swimming in a pool 26.64

AFS-Swin-Bert-Semantics: a man is sliding down a water slide 208.0



Video zpgW7m7-LZw-2-15

Reference: a little boy is playing golf CIDEr

AFS-Swin-Bert: a man is riding a bicycle 0.005

AFS-Swin-Bert-Semantics: a boy is playing 157.5

Fig. 8: Effect of considering the semantic context vector in caption generation. In these videos from the MSVD collection the CIDEr score has been effectively increased.

8.5 Training Details

We initialize our encoder with the weights of a Swin-B network weight with layer number = $\{2, 2, 18, 2\}$ [29], which was trained on the Kinetics400 data set [22]. For all videos, we employ AFS to select $N = 32$ frames and follow the default Swin-B normalization, cropping and resizing operations to transform them to a dimension of $3 \times 224 \times 224$. The input to the Swin network is $T \times W \times H \times 3$ (in our case $(32, 224, 224, 3)$), which is divided to $3 \times \frac{T}{2} \times \frac{W}{4} \times \frac{H}{4}$ 3D tokens. These 3D tokens are hierarchically merged to analyse their temporal relations. The output of our encoder is the last hidden layer of the Swin-B architecture variant. It has the size of $\text{Size} = (\frac{T}{2}, \frac{W}{32}, \frac{H}{32}, 1024)$. After average 2D-pooling on (W, H) , followed by a single linear layer, we produce sixteen output tokens of size 768, encoding the visual state to be fed to the decoder. The same output is also taken to produce a semantic context vector.

Our decoder backbone is initialized by pre-trained BERT [16] on which we add a language modelling head. During training and inference we employ causal masks to prevent attention to future tokens. Following prior work [10,9,57], sentences longer than 20 words are truncated. All words are converted to lower case and the decoder predicts lower case tokens. Its dropout probability is set to 0.3. A two-layer MLP with RELU activation is trained for $K = 768$ concepts (layer sizes $(1024, 2048, 768)$). In a first step, we train the semantic concept MLP network using the pre-trained Swin-B and BCE loss.

We train our model end-to-end network to minimize the overall loss \mathcal{L} where the training loss for decoder is the cross-entropy to the randomly selected training caption and the loss for the full network including the semantic vector MLP is given by

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{BCE}, \text{ with } \lambda = 0.1. \quad (2)$$

To train our models, we use the AdamW [31] optimizer with default settings, with a learning rate of 0.00001 and an effective batch size of 8. Additionally, we use gradient clipping of 0.05. The final model is selected based on the best harmonic mean of the METEOR [5] and BLEU-4 [33] scores to avoid optimizing for a single metric.



Video 7021

Reference: a baseball batter hits the ball	CIDEr
AFS-Swin-Bert: there is a man is playing with a ball	29.45
AFS-Swin-Bert-Semantics: a man is playing baseball in a field	115.1



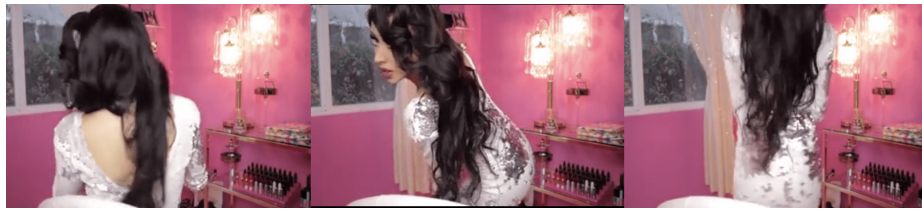
Video 7053

Reference: a man playing with a lion	CIDEr
AFS-Swin-Bert: a lion is playing with a lion	155.6
AFS-Swin-Bert-Semantics: a man is playing with a lion	300.9



Video 7060

Reference: two men are giving an introduction	CIDEr
AFS-Swin-Bert: a man is talking to another man	3.20
AFS-Swin-Bert-Semantics: a person is explaining something	101.1



Video 7265

Reference: a girl in her room shows off her new hairstyle	CIDEr
AFS-Swin-Bert: a woman is showing how to wash her hair	31.37
AFS-Swin-Bert-Semantics: a woman is showing off her hair	73.79

Fig. 9: Effect of considering the semantic context vector in caption generation. In these videos from the MSR-VTT collection the CIDEr score has been effectively increased.

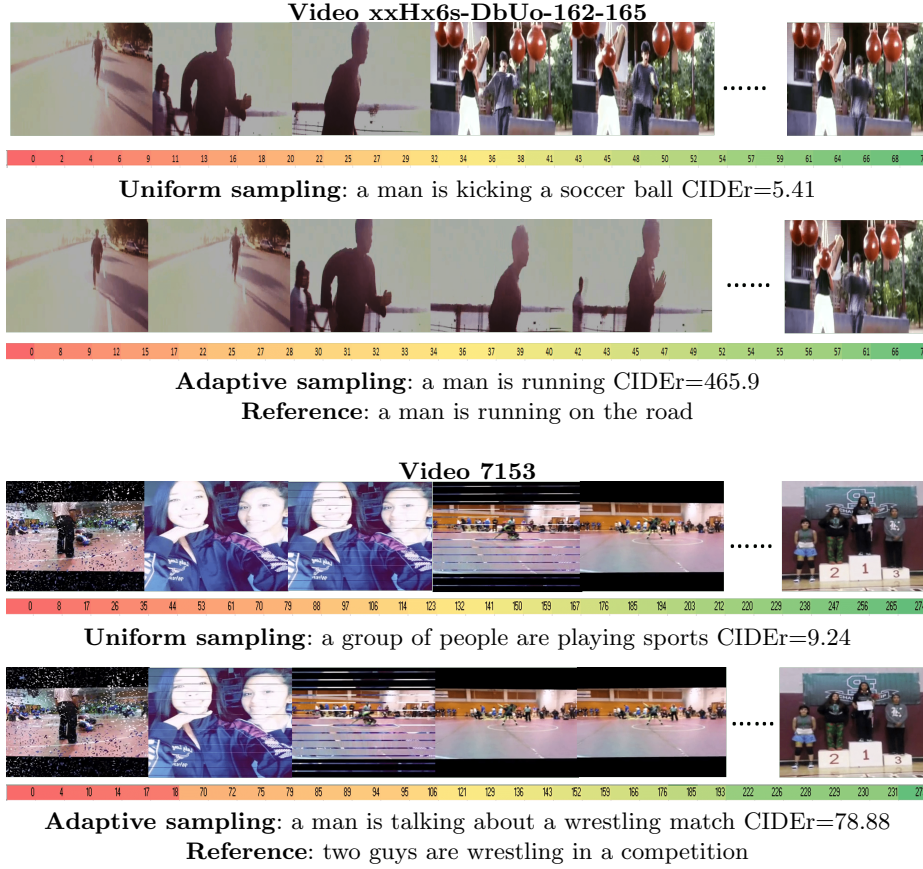


Fig. 10: Effect of the Adaptive Frame Selection. In these example videos uniform sampling (top) wastes some of the 32 input frames for repetitive non-informative content. Our adaptive frame selection prefers those frames with strong differences to the previous one. Often, more diverse frames are selected helping generate better captions.

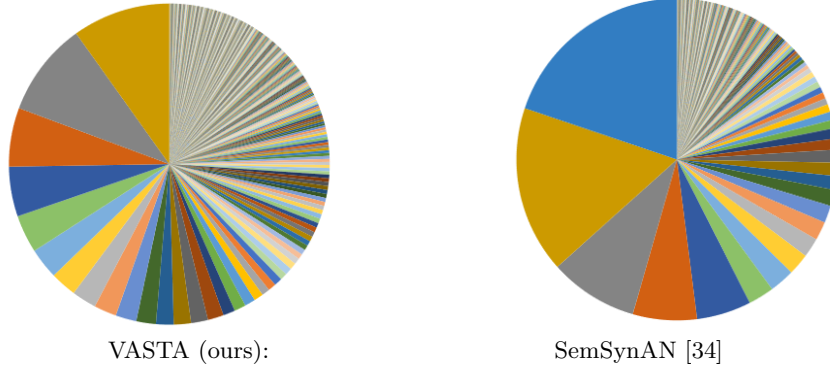


Fig. 11: Frequency of different POS structures on MSR-VTT Data set. More different segments indicate a higher diversity in captions. Note, that SemSynAN only uses 4 different POS structures for more than 50% of its generated captions.

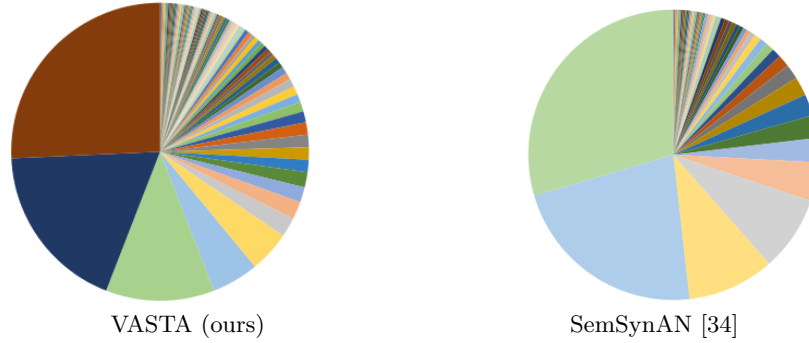


Fig. 12: Frequency of POS structures on MSVD Data set. More different segments indicate a higher diversity in captions. Note, that SemSynAN only uses 6 different POS structures for more than 75% of its generated captions.

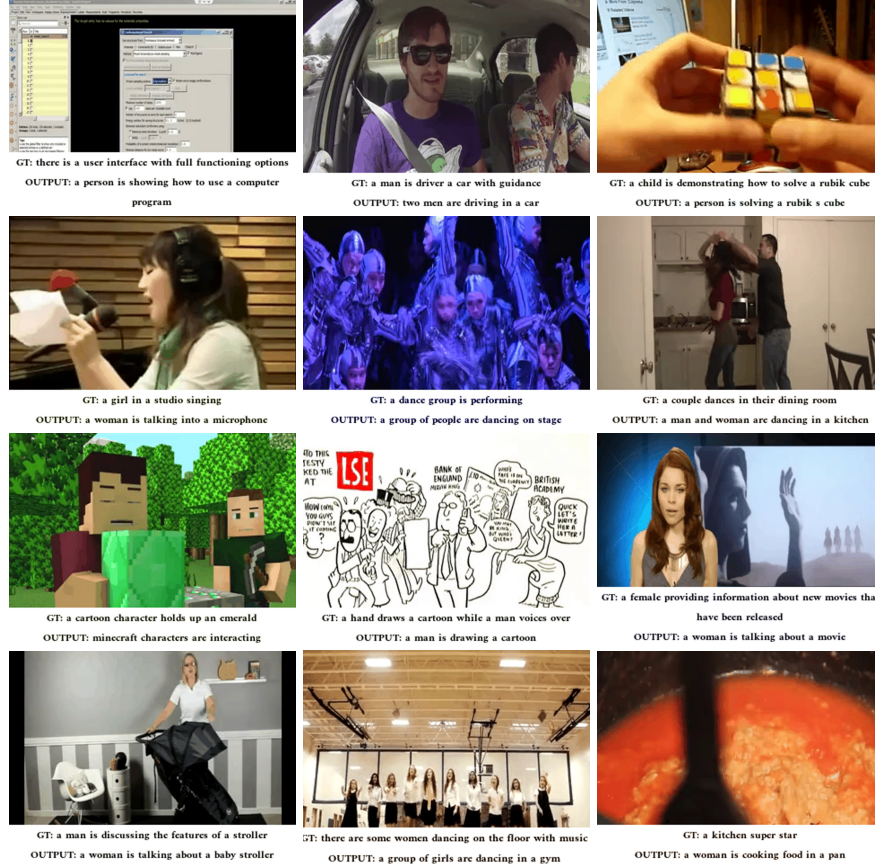


Fig. 13: Based on METEOR and BLEU-4 we present a selection of the 10 percent best-performing videos in the MSR-VTT data set. GT refers to the ground truth reference.

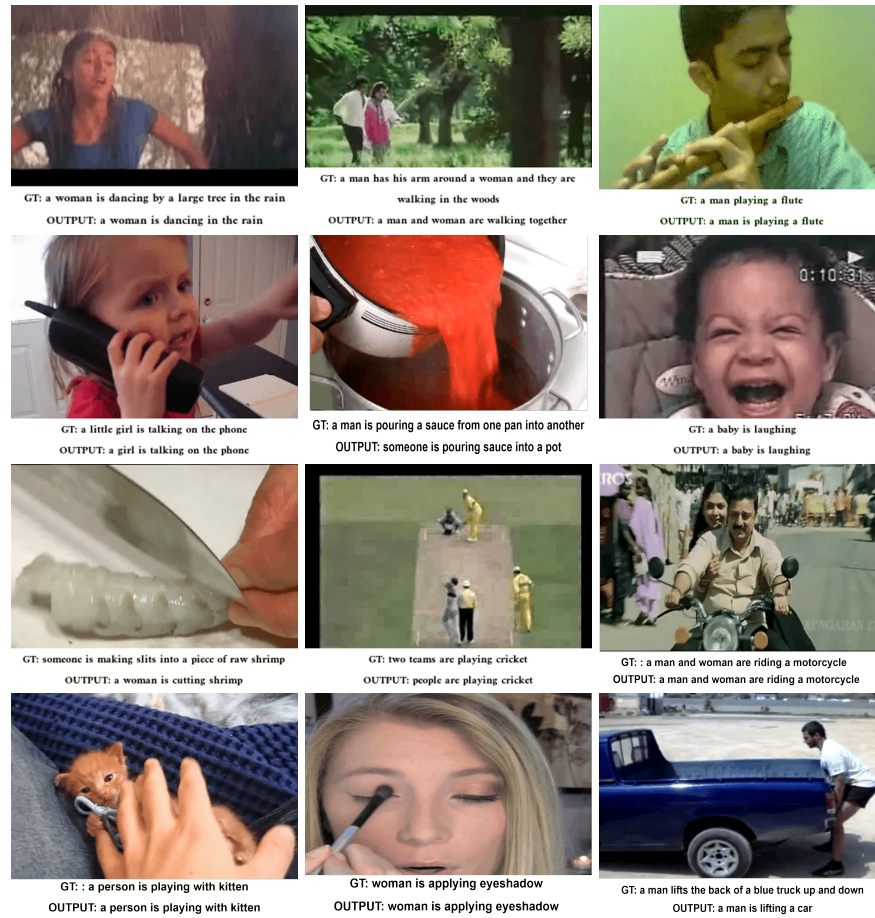


Fig. 14: Based on METEOR and BLEU-4 we present a selection of the 10 percent best-performing videos in the MSVD data set. GT refers to the ground truth reference.



GT: a man with harness around his lower waist section is showing how to tie a special knot in a large rope

OUTPUT: a young man demonstrates how to tie a knot in a rope



GT: a group of people are practicing and playing curling inside the practice ring

OUTPUT: a group of people are playing a game of curling



GT: a little that appears to be playing with sand on a shore, but the background is rocky

OUTPUT: a little girl is sitting on the beach and making a sandcastle



GT: at the athletic event officials were measuring the distance that long jumpers accomplished

OUTPUT: a boy runs down a track and jumps into a pile of sand



GT: a man is running on a treadmill while being show from the buttocks down

OUTPUT: a person is running on a treadmill in slow motion



GT: a toddler is swinging back in forth in a baby swing at the park

OUTPUT: a little boy is swinging back and forth in a swing

Fig. 15: Based on METEOR and BLEU-4 we present a selection of the 10 percent best-performing videos in the VATEX data set. GT refers to the ground truth reference.