SEM-POS: Grammatically and Semantically Correct Video Captioning

Asmar Nadeem[1], Adrian Hilton[1], Robert Dawes[2], Graham Thomas[2], Armin Mustafa[1]

Abstract

Generating grammatically and semantically correct captions in video captioning is a challenging task. The captions generated from the existing methods are either word-by-word that do not align with grammatical structure or miss key information from the input videos. To address these issues, we introduce a novel global-local fusion network, with a Global-Local Fusion Block (GLFB) that encodes and fuses features from different parts of speech (POS) components with visual-spatial features. We use novel combinations of different POS components - 'determinant + subject', 'auxiliary verb', 'verb', and 'determinant + object' for supervision of the POS blocks - Det + Subject, Aux Verb, Verb, and Det + Object respectively. The novel global-local fusion network together with POS blocks helps align the visual features with language description to generate grammatically and semantically correct captions. Extensive qualitative and quantitative experiments on benchmark MSVD and MSRVTT datasets demonstrate that the proposed approach generates more grammatically and semantically correct captions compared to the existing methods, achieving the new state-of-the-art. Ablations on the POS blocks and the GLFB demonstrate the impact of the contributions on the proposed method.

1. Introduction

Video captioning is a challenging task as it aims to describe the contents of a video in a natural language like humans do [50]. This finds application in media production [3], visual retrieval [19, 26, 43, 53, 62], visual question answering [4], etc. Existing methods for video captioning can broadly be classified into two categories: (1) learning visual representation for the whole caption in an end-to-end manner leaving the fine-grained local features unaddressed [1,34,36,38], which generates captions with key information missing in the description; and (2) incorporating features at the level of objects and motion [33,51,58,60,64], generating word-by-word descriptions which do not align with the grammatical structure. In this paper, we aim to address both issues in the existing methods by proposing a novel global-local fusion network that consists of- first a variety of POS components of a sentence; and second the Global-Local Fusion Block (GLFB) which fuses the local and global, visual and language features in the network. Prior knowledge of what to predict in a template-based sentence [13, 23, 24], gives POS blocks the ability to focus on the most suitable visual features to produce grammatically correct captions.

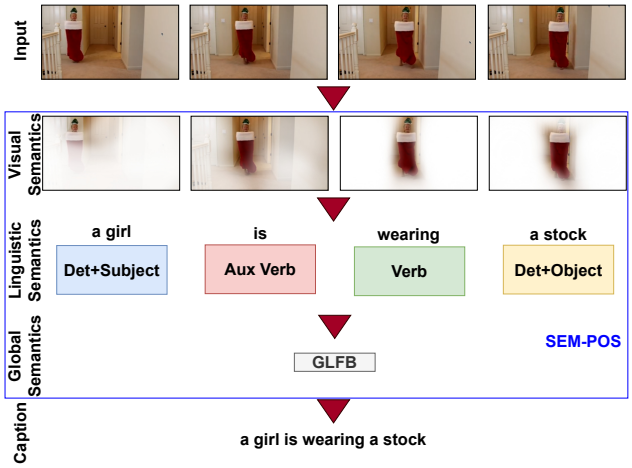

Figure 1. SEM-POS aligns visual and linguistic semantics for video captioning task. Det + Subject, Aux Verb, Verb, and Det + Object are the four POS blocks used in the proposed approach.

Existing methods [36,38,57] use a single type of visual feature i.e., spatial or temporal, as an input to the entire model which leads to reduced semantic correctness. The proposed method inputs both spatial and temporal input features, which improves the semantics of the generated caption. The output visual semantics from each POS block are input to the GLFB block, which fuses the local and global visual information before fusing it with the language semantics in the Caption block, unlike previous methods [1,34] that generate only a global caption from a video. This helps to align the visual semantics of a sentence with natural language semantics and grammatical structure. Existing methods [1,33,58,60] mainly use global verbs and nouns in the form

[1]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, United Kingdom. [2]BBC Research and Development, United Kingdom. Correspondence to: Asmar Nadeem <asmar.nadeem@surrey.ac.uk>.

of actions and objects, generating word-by-word descriptions. [41] attempts to explore the local semantics in the input video, however, it does not align the visual and language semantics, leading to limited performance. In this paper, in addition to the visual features from the POS blocks, we also input language semantics (supervised by 'determinant + subject', 'auxiliary verb', 'verb', and 'determinant + object') to the GLFB block, which aligns visual semantics with linguistics and generating grammatically correct captions. The input, output, and pipeline from the proposed network are shown in Figure 1. We also mask input spatial as well as temporal features in end-to-end training which also results in an improvement in the performance.

The proposed method is inspired by the previous work POS-CG [51] and SAAT [64]. However, these methods do not fuse language and visual information and they assume the subjects as well as the objects to be physical entities and the verbs to be motion-based actions which may not be true semantically. This paper aims to answer three main questions previously unaddressed by state-of-the-art methods: 1) How to align visual and linguistic semantics to generate grammatically correct captions? (2) How to add fine-grained local features to increase semantic correctness? and (3) How to fuse local and global features together for simultaneous grammatical and semantic correctness? To answer these questions, we have proposed the following novel contributions:

- A novel global-local fusion network with GLFB that encodes and fuses fine-grained local visual features from different POS blocks with global caption language features for semantically correct captions.
- We propose four POS blocks: Det+Subject, Aux Verb, Verb and Det+Object with spatial and temporal features as input. Comprehensive ablation studies on the selection of POS blocks and input features show the generation of captions closer to the natural language.
- An extensive evaluation of the proposed method against state-of-the-art on two benchmark datasets, MSVD [8] and MSRVTT [55], on five metrics, demonstrates improved grammatical and semantic caption generation.

2. Related Work

This section reviews video captioning techniques, and visual and linguistic features fusion to support the contributions of the paper.

2.1. Video Captioning

Video representation: Video captioning [2, 50] is a challenging task in video representation. Earlier, CNN-based methods [42, 47] have been applied to learn video representations which lack to model the long-range dependencies but some [16, 46] are still popular for their spatial and temporal representations of the videos. These video features, in a global fashion, combined with RNN [9, 11, 18, 32, 59] and LSTM [12, 45], have been employed for the video captioning task but visual-linguistic alignment and fine-grained features are not explored.

Recently, multi-modal methods [44, 56] have become popular using text and audio information for video understanding tasks which result in text or audio-dependent models lacking visual information. [15] uses a cross-attention mechanism to attend to different modalities and other, visions only, methods [45, 61] use attention mechanisms for extracting linguistic semantics, globally, from the visual features. The proposed method in this paper exploits the power of LSTMs, inspired by [12], for aligning and fusing visual and linguistic semantics. We also employ an attention mechanism [61] to make the linguistic components focus on the relevant visual features for video captioning.

Transformer for Captioning: Transformer [20, 48] is increasingly being used for object detection [7], classification [14], scene understanding [63] and image captioning [30] tasks. It gained its popularity from the natural language-based models like BERT [10] and roBERTa [31], a version of BERT [10], where it demonstrates significant improvement in the performance when using a huge corpus of web text data. These methods are recently adopted for vision-based tasks while using a billion images to give a state-of-the-art performance. Our method generates more grammatically correct captions than these methods (see Section 4.1). [28, 29, 40] are not fair to be compared with as [28] has used an ensemble model together with a transformer decoder and rest have a pre-training stage for their extremely large transformer-based models using ten to hundred times more computing than ours.

2.2. Masking Techniques

Previous methods [17, 25, 54] use masking on image frames to generalize the input to the model which improves the performance of the model. Pre-training tasks for language transformers [10, 31] require language masking, and similarly, image frames are masked to recover the masked portion of the image. To the best of our knowledge, we are the first to mask spatial features as well as temporal features for an end-to-end video captioning task, which results in an improvement in the performance (see Section 4.2).

2.3. Visual and Linguistic Alignment and Fusion

Existing methods [33, 51, 60, 64] generate captions with POS in the decoder block without using separate POS blocks. We are the first to explore separate POS blocks in the encoder part to align visual semantics with linguistic POS components in a supervised manner. [21] predicts the label of the POS for every word as part of the caption generation using a softmax layer, without any POS blocks. The

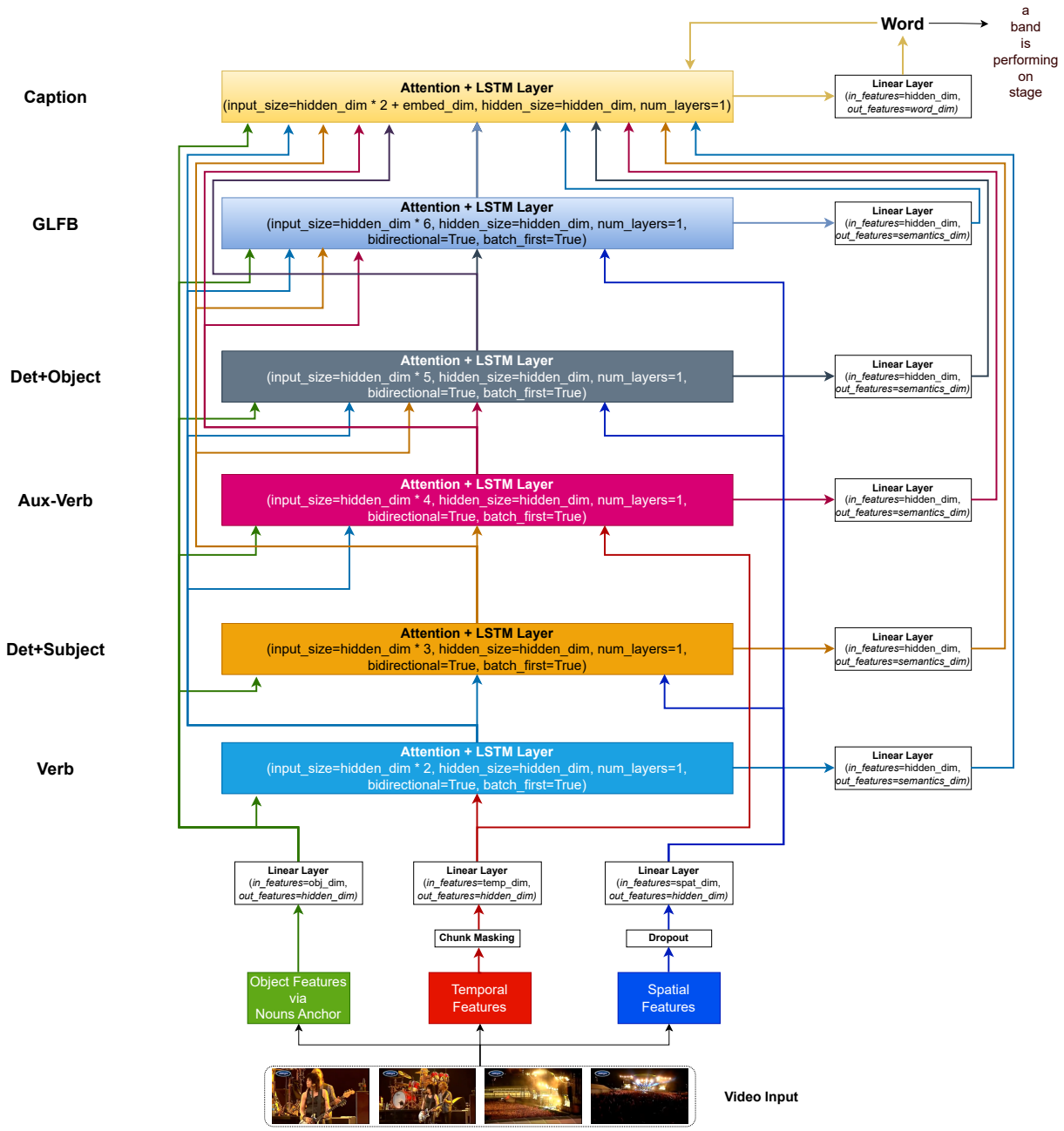

Figure 2. Illustration of our SEM-POS where POS blocks in Section 3.3 help align visual-linguistic semantics and attain fine-grained features. Also, GLFB (see Section 3.4) performs global-local feature fusion to generate grammatically and semantically correct captions.

proposed method aligns the visual and language components at both fine-grained local and global levels by proposing to use of four POS blocks and GLFB.

3. Method

This section gives an overview of the proposed method, which is followed by the problem formulation in Section 3.1 and a detailed explanation for the POS blocks in the subsequent sections.

An overview of the proposed network is shown in Figure 2. The input to the system is a video and the output is a caption that describes the video. Spatial and temporal features are extracted from the input video along with noun features via the nouns anchor block to use as input to the network. We use Deformable DETR [58,65] transformer architecture for the nouns anchor block which takes object features as an input and is supervised by the nouns in the ground-truth caption. Either spatial or temporal features along with object semantics from the nouns anchor block are given as in-

puts to the LSTM-based POS blocks using different masking mechanisms. We have introduced random spatial features masking and chunk-wise temporal features masking as Visual Features Masking (VFM) in Section 3.2 for an end-to-end video captioning task. In this work, we proposed 4 POS blocks - Verb, Det + Subject, Aux Verb, and Det + Object. The visual features from the POS blocks are given as input to the GLFB. In addition to this, word embeddings exploiting linguistic information are obtained by supervising POS blocks with respective POS components. This helps in achieving the alignment between visual and linguistic features at both the fine-grained local and global levels, unlike previous methods. The fine-grained local and global visual and linguistic features are then fed into the Caption block to generate the final caption.

3.1. Problem Formulation

Given a video-captioning task, we aim to align the visual semantics with the linguistic semantics using the POS blocks. These POS components are inherently present in any grammatically structured caption and these blocks help to exploit fine-grained local visual features of a video to align linguistic and visual semantics. These aligned semantics on the fine-grained local level are then used as input to the GLFB to fuse local and global representations for improved semantic correctness.

For supervised training of each POS block, we extract the respective POS component from the ground-truth caption $\mathcal{C}$. Let $\mathcal{R}$ be all the reference captions for a given video $\mathcal{V}$, ground-truth $\mathcal{G}_{POS}$ for the POS blocks (see Section 3.3) are defined as:

$$\mathcal{G}_{POS} = \{\{S^d, V, A, O^d\} \in \mathcal{C} \,|\, \forall \mathcal{C} \in \mathcal{R}\}, \qquad (1)$$

where S^d, V, A, O^d are the ground-truth labels for Det + Subject, Verb, Aux Verb and Det + Object blocks respectively. These labels when converted to word embeddings help align the visual semantics with their linguistic counterparts in a supervised setting. Unlike existing methods that align visual and linguistic semantics on a caption or a nouns-verb level only, we perform a fine-grained local alignment of visual and linguistic semantics for four POS blocks. Our overall objective is as follows:

$$\mathcal{O}_{OVERALL} = \mathcal{O}_{POS} + \mathcal{O}_{GLFB} + \mathcal{O}_{VFM}, \qquad (2)$$

where $\mathcal{O}_{POS}$, $\mathcal{O}_{GLFB}$ and $\mathcal{O}_{VFM}$ are POS, GLFB and VFM objectives, respectively.

3.2. Visual Features Masking (VFM)

Random image masking improves the results for image detection and classification tasks in the previous methods [17,54]. A random patch of the image is masked as an input to the model and the model is forced to recover the masked patch along with the overall image. In our approach, we randomly mask 30% of the spatial features and a chunk of the temporal features, approximately 15%, before the features are input to the POS blocks. The masking percentages are selected through experimentation. When a chunk is masked in the temporal features of a video, it masks the information in the temporal dimension in all the sampled frames. This helps the proposed network model to generalize well across the entire data distribution. It also forces the spatial and temporal features to interact to find the masked information and this improves the overall performance (see Section 4.2).

3.3. POS Blocks

We propose to use four POS blocks in the proposed network, two of these four blocks are a combination of two POS components i.e. 'determinant + subject' and 'determinant + object'. The other two components are 'verb' and 'auxiliary verb'. We performed a detailed analysis of the ground-truth captions in benchmark datasets to analyze the presence of the individual parts of speech components in each caption. Figure 3 shows the percentage of the presence of each POS component introduced in our model in the captions for MSVD and MSRVTT benchmark datasets. It is evident from the figure that the four POS blocks used in the proposed method are the largely represented POS components in the captions and also, adding blocks in the network for other less represented POS components affects the performance adversely. An extensive ablation (see Section 4.2) demonstrates the effectiveness of using these four POS blocks within the network to generate grammatically and semantically correct captions.

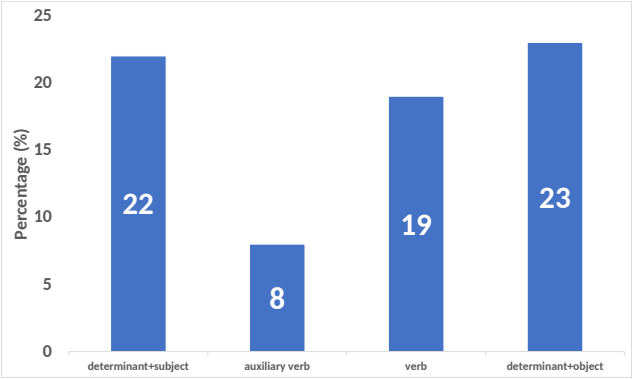

Figure 3. POS components' representation in the captions from the MSVD and MSRVTT benchmark datasets.

Verb. This is the first block in the pipeline which takes the temporal features t_f as well as the nouns features n_f (potential candidates for 'subject' and 'object') via the nouns anchor as input and it aligns the t_f and n_f with the linguistic verb. These features are concatenated (cat) together and passed through an LSTM layer, in bi-directional mode, to get the visual verb semantics v_f. Finally, a fully connected (fc) layer is used to predict the linguistic 'verb' embedding

v_p:

$$v_f = lstm_b(cat(t_f, \mathcal{A}_{nv}[n_f])), \tag{3}$$

hence, $v_p = fc(v_f)$. $\mathcal{A}_{nv}$ is a learnable attention for n_f.
For training of the Verb block, we get 'verb' from the ground-truth caption using pre-trained roBERTa [31] model and use the same model for ground-truth embedding v_g. The distance with the predicted embedding v_p is minimized as part of the model learning. Let $\mathcal{L}_v$ be the Verb block loss:

$$\mathcal{L}_v = Dist.(v_p, v_g) \tag{4}$$

Det+Subject. In this block, we use a combination of two POS components, i.e., 'determinant + subject'. This block uses the spatial visual features s_f, n_f and v_f (see Eqn. 3) to determine the 'determinant + subject' of the caption. Similar to the Verb block, attention is used to combine the information from the inputs. n_f and v_f are used as inputs via learnable attentions $\mathcal{A}_{ns}$ and $\mathcal{A}_{vs}$ respectively. s_f are then concatenated with these attended features and are passed through an LSTM layer similar to the Verb block to get 'determinant + subject' visual features ds_f. Let ds_p be the predicted 'determinant + subject' linguistic embedding after passing ds_f through the fully connected layer,

$$ds_f = lstm_b(cat(s_f, \mathcal{A}_{vs}[v_f], \mathcal{A}_{ns}[n_f])), \tag{5}$$

hence, $ds_p = fc(ds_f)$. This approach creates a dynamic template of captions being filled with outputs from the POS blocks as these become available both during training and testing.
For labels, we get 'determinant' and 'subject' from the ground-truth caption using the pre-trained roBERTa [31] model. We combine these together with a space in-between and use the same model for ground-truth embedding ds_g. The distance with the predicted embedding ds_p is minimized as a part of the model learning. Let $\mathcal{L}_{ds}$ be the Det+Subject block loss:

$$\mathcal{L}_{ds} = Dist.(ds_p, ds_g) \tag{6}$$

Aux Verb. Aux verb block uses t_f, noun features n_f, v_f (see Eqn. 3) and ds_f (see Eqn. 5) to predict 'auxiliary verb' for the caption. It also gives an insight into the singular or plural nature of the 'subject', which is used as an input to the GLFB. n_f, v_f and ds_f are used as the inputs via learnable attentions $\mathcal{A}_{na}$, $\mathcal{A}_{va}$ and $\mathcal{A}_{dsa}$ respectively. t_f is then concatenated with these attended features and is passed through an LSTM layer like in previous blocks to get 'auxiliary verb' visual semantics a_f. a_p is obtained after passing a_f through a fully connected layer:

$$a_f = lstm_b(cat(t_f, \mathcal{A}_{va}[v_f], \mathcal{A}_{na}[n_f], \mathcal{A}_{dsa}[ds_f])), \tag{7}$$

hence, $a_p = fc(a_f)$. For ground-truth labels, we extract the 'auxiliary verb' from the ground-truth caption and also, its embedding a_g using pre-trained roBERTa [31] model. Then, the distance with a_p is minimized as part of the model training. Let $\mathcal{L}_a$ be the Aux Verb block loss:

$$\mathcal{L}_a = Dist.(a_p, a_g) \tag{8}$$

Det+Object. Last, in the POS blocks, the Det+Object block is a combination of the 'determinant' and 'object'. It exploits linguistically aligned visual features from the previous POS blocks along with s_f. Like previous POS blocks, learnable attentions $\mathcal{A}_{no}$, $\mathcal{A}_{vo}$, $\mathcal{A}_{dso}$ and $\mathcal{A}_{ao}$ are used to exploit information in n_f, v_f, ds_f and a_f respectively. In similar manner, do_p is obtained from do_f by:

$$\begin{aligned} do_f = lstm_b(cat(s_f, \mathcal{A}_{vo}[v_f], \mathcal{A}_{no}[n_f], \\ \mathcal{A}_{dso}[ds_f], \mathcal{A}_{ao}[a_f])), \end{aligned} \tag{9}$$

hence, $do_p = fc(do_f)$. For training, we get the 'determinant' and 'object' from the ground-truth caption using pre-trained roBERTa [31] model, combine these together with a space in-between and then, use the same model for ground-truth embedding do_g. The distance with the predicted embedding do_p is minimized as part of the model learning. Let $\mathcal{L}_{do}$ be the Det+Object block loss:

$$\mathcal{L}_{do} = Dist.(do_p, do_g) \tag{10}$$

3.4. Global Local Fusion Block (GLFB)

The proposed GLFB is used to achieve fusion between the global and local features, by using the fine-grained local and linguistically aligned visual features from the POS blocks and the visual-spatial features. Visual features g_f are the output of an LSTM layer, given a set of inputs using the attention mechanism. To predict a linguistic GLFB representation g_p, g_f is passed through a fully connected layer:

$$\begin{aligned} g_f = lstm_b(cat(s_f, \mathcal{A}_{vg}[v_f], \mathcal{A}_{ng}[n_f], \\ \mathcal{A}_{dsg}[ds_f], \mathcal{A}_{ag}[a_f], \mathcal{A}_{dog}[do_f])), \end{aligned} \tag{11}$$

hence, $g_p = fc(g_f)$. For training of this block, the whole ground-truth caption is first converted to lower-case letters and then, passed through roBERTa [31] to get ground-truth embedding g_g. Then, the distance with the predicted embedding g_p is minimized as part of the supervised learning. Let $\mathcal{L}_g$ be the GLFB loss, then:

$$\mathcal{L}_g = Dist.(g_p, g_g) \tag{12}$$

3.5. Attention Mechanism

Our approach employs an attention mechanism in all blocks to exploit the inputs as can be seen in previous sections. Here, we generically define the attention α_g:

$$\alpha_g = w_g \, tanh(\mathcal{W}_g \, X + \mathcal{U}_g \, Y + b_g), \tag{13}$$

where $w_g, \mathcal{W}_g, \mathcal{U}_g$ and b_g are the training parameters. X is either spatial or temporal features, depending on whatever is the input to each block, in all blocks, except the Caption block (see Section 3.6) where it is the previous word in the generated caption. Y is the visual semantics generated in the POS blocks and GLFB.

3.6. Overall Training Objective

The overall training objective adds the POS blocks and GLFB together along with the Caption block. The Caption block uses an LSTM layer, followed by a fully connected layer, with inputs of both visual and linguistic semantics from the POS blocks and GLFB to generate a caption, word by word, using beam search, as an output. Overall learning loss $\mathcal{L}_{all}$ is defined as:

$$\mathcal{L}_{all} = \mathcal{L}_c + \mathcal{L}_v + \mathcal{L}_{ds} + \mathcal{L}_a + \mathcal{L}_{do} + \mathcal{L}_g, \tag{14}$$

where $\mathcal{L}_c$ is the cross entropy loss for the Caption block:

$$\mathcal{L}_c = -\sum_{n=1}^{N} E(w_n) log P_n \tag{15}$$

$E(w_n)$ is the one-hot encoding of the word w_n and P_n, the output of the fully connected layer, is the probability distribution over the word vocabulary. It is worth mentioning that loss for the VFM objective is an integral part of $\mathcal{L}_{all}$ as the model has to predict the right embedding with masked input. Also, we have done some experiments by tuning the loss weights but the effect on the overall performance was found to be minimal.

The objective of the Caption block is to generate the final video caption which is both grammatically and semantically correct.

4. Experimental Results and Evaluations

Datasets. Similar to the existing video captioning approaches [1,33,36,38,51,58,60,64], we evaluate our model on MSRVTT [55] and MSVD [8] benchmark datasets. These datasets have a total of 10,000 and 1,970 videos respectively. We use the same train, validation, and test split as the existing methods.

Metrics. For evaluation with the existing methods, we use the following set of video captioning metrics: BLEU@4 (B@4) [35], METEOR (M) [5], ROUGE-L (R) [27] and CIDEr (C) [49]. B@4 is based on n-gram (n=4) matches between the predicted and ground-truth captions. In contrast, M is a recall-based metric that uses word-to-word alignment between predicted and ground-truth captions. The final score is calculated between the best-scoring ground truth and the predicted caption. R, another recall-based metric, measures the longest common set of shared words with similar order between predicted and ground-truth captions. C is studied to be robust in the condition where the semantic meaning of the caption remains intact [6]. It measures the average cosine similarity between predicted and ground-truth captions. Also, we are the first to use GPT-2 [37] pre-trained model for measuring the Grammatical correctness Score (GS) of the captions generated by our model in comparison to state-of-the-art which we implement for the GS metric, demonstrating improved performance. GS metric measures the grammatical correctness without using the ground-truth captions unlike the other four metrics (see Section 4.2 and 4.3).

Implementation. For text, we have used *spaCy*[1] with roBERTa [31], a version of BERT [10], to extract POS components along with nouns from the ground-truth captions. We again use pre-trained roBERTa [31] for text embedding which gives an embedding of size 768 for 'determinant + subject', 'verb', 'auxiliary verb', 'determinant + object', and the whole caption. Following the existing methods, we use InceptionResNetV2 [46] to extract the spatial features and C3D [16] to extract the temporal features. These features are projected to 512 sizes before being input into the network. We train for epochs 25 and use a learning rate of 0.00015, batch size 16, ADAM optimizer [22], 16 samples per video as well as a hidden state size of 512 for the Caption block. Our model has 76M parameters and 0.045s inference time. Apart from that, we use Yolov7 [52] for extracting object features for the noun anchor. The whole implementation is performed using one NVIDIA GeForce RTX 3090 and PyTorch.

4.1. Results and Comparison with Existing Methods

We do a quantitative evaluation of the benchmark datasets for several state-of-the-art methods and also, demonstrate qualitative results of the proposed method.

Quantitative Evaluation: Table 1 gives a quantitative comparative performance of the proposed model against the state-of-the-art methods which do not employ large-scale transformer pre-training. The proposed SEM-POS network outperforms all these methods on both MSVD [8] and MSRVTT [55] benchmarks. Improvement is around 3% and 4% on CIDEr for MSRVTT and MSVD benchmarks respectively. In terms of GS metric, our method gives an improvement of around 10% and 16% for MSRVTT and MSVD respectively.

This is due to the fact that our method explores fine-grained local features in the four POS blocks before generating a global representation, giving it a better generalization ability compared to the other approaches. Visual features aligned with linguistics in POS blocks are learned across the entire data distribution for the video captioning task, which contributes to this improved performance. POS blocks help to align the linguistic and visual semantics at

[1] https://spacy.io/

Method	Language Components	MSVD					MSRVTT				
		B@4 ↑	M ↑	R ↑	C ↑	GS ↓	B@4 ↑	M ↑	R ↑	C ↑	GS ↓
OA-BTG [60]	Nouns (Physical objects)	56.9	36.2	-	90.6	852.6	41.4	28.2	-	46.9	337.5
MARN [36]	Word by word (Global)	48.6	35.1	71.9	92.2	822.2	40.4	28.1	60.7	47.1	328.1
POS-CG [51]	Subject-verb-object	52.5	34.1	71.3	88.7	801.5	42.0	28.2	61.6	48.7	306.4
GRU-EVE [1]	Nouns-verb	47.9	35.0	71.5	78.1	-	38.3	28.4	60.7	48.1	-
STG-KD [33]	Nouns-verb	52.2	36.9	73.9	93.0	802.9	40.5	28.3	60.9	47.1	300.3
SAAT [64]	Subject-verb-object	46.5	33.5	69.4	81.0	806.1	40.5	28.2	60.9	49.1	282.6
SGN [38]	Word by word (Global)	52.8	35.5	72.9	94.3	830.3	40.8	28.3	60.8	49.5	316.4
GL-RG [57]	Word by word (Global)	55.5	37.8	74.7	94.3	792.5	45.5	30.1	62.6	51.2	247.1
HMN [58]	Nouns-verb	59.2	37.7	75.1	104.0	728.7	43.5	29.0	62.7	51.5	214.7
SEM-POS (Ours)	Subject-aux verb-verb-object	**60.1**	**38.5**	**76.0**	**108.3**	**607.1**	**45.2**	**30.7**	**64.1**	**53.1**	**192.6**

Table 1. Comparison against state-of-the-art methods. The best results on each metric are in bold on MSVD and MSRVTT benchmarks.

the fine-grained local level and then the GLFB visibly fuses it with global representation. Both these contributions of POS blocks and GLFB help generate grammatically and semantically correct sentences. Our method also outperforms these methods [28, 29, 40] (which have a large-scale transformer-based pre-training stage) on the GS metric, improving on their scores of 655.2, 643.7, and 612.8 (MSVD) and 224.7, 212.9 and 201.6 (MSRVTT) respectively.

Qualitative Results: In Figure 4, we show qualitative results for our approach. It can be seen that generated captions adequately represent not only the ground-truth captions but also the fine-grained visual features. For example, in the images in the first two rows, on the right-hand side, the proposed caption is better than ground truth and provides more grammatically and semantically correct information about 'man', 'riding', 'road' (single) and 'football' (single) instead of 'person', 'on', 'streets' (plural) and 'sports' (plural) respectively, which is derived from the fine-grained local features in the POS blocks.

We also compare the qualitative results of our method with the state-of-the-art method HMN [58]. Row 1 right, row 2 left and row 3 right show that our method generates captions that are grammatically and semantically more correct than HMN [58]. The rest of the results also show that our method generates captions as well as state-of-the-art.

Figure 4. Qualitative Results- SEM-POS predictions show fine-grained visual details and global-local features' fusion.

4.2. Ablation Results

Effectiveness of POS blocks and GLFB. This section supports our first two contributions of using POS blocks and GLFB in the network and the results are shown in Table 2. Ablation is performed on the proposed method by removing the POS blocks one by one from the proposed network and evaluating the performance. The first four rows in the table demonstrate these results. As it is evident from the table the Verb block contributes significantly to the performance of the network. Det+Subject and Det+Object blocks contribute equally to the MSRVTT [55] benchmark but Det+Object shows only a minor contribution for the MSVD [8] benchmark dataset as it has a significant number of ground-truth captions without an object.

'w/o GLFB' shows results without the GLFB (see Section 3.4). It demonstrates the importance of combining fine-grained local and global features for video captioning. Three rows 'w Adverb', 'w Adjective', and 'w Conjunction' show that the performance is affected adversely as these POS components are not widely represented in the captions. Rows 'w Subject only' and 'w Object only' are Subject and Object blocks without Determinant and these show a decrease in performance.

Row 'w All' show the results for adding Adverb, Adjective, and Conjunction blocks to our model. It is evident that adding blocks that are not represented enough in the POS components, decreases the performance significantly.

Effectiveness of VFM: We also evaluate the performance without Visual Feature Masking (VFM) which is introduced

Block	MSVD					MSRVTT				
	B@4 ↑	M ↑	R ↑	C ↑	GS ↓	B@4 ↑	M ↑	R ↑	C ↑	GS ↓
w/o Det + Subject	56.4	36.6	72.0	100.1	695.1	41.2	28.0	60.0	49.8	244.9
w/o Verb	55.9	34.9	70.8	98.1	733.1	40.8	27.6	58.3	49.1	267.3
w/o Aux Verb	57.7	37.1	73.9	101.2	688.3	42.5	28.2	60.9	50.6	244.2
w/o Det + Object	57.8	37.1	73.9	102.3	642.4	41.6	28.9	62.1	51.7	221.1
w/o GLFB	56.7	36.0	73.1	101.2	638.1	41.9	28.2	61.9	50.1	246.3
w/o VFM	59.6	38.1	75.3	106.5	628.7	44.6	29.6	63.4	52.5	224.1
w Adverb	56.5	36.8	71.2	99.4	710.1	40.7	28.1	59.5	49.6	264.9
w Adjective	56.2	34.5	70.8	98.1	762.3	40.8	27.8	57.5	49.3	273.6
w Conjunction	56.4	35.3	72.9	97.5	785.4	40.4	27.7	59.1	49.0	275.7
w Subject only	58.4	37.9	74.7	103.3	640.1	43.8	28.9	62.6	51.5	206.1
w Object only	58.9	37.8	74.9	103.7	632.5	44.0	29.1	61.6	51.7	209.5
w All	57.9	35.9	73.1	100.5	657.3	41.4	27.5	61.2	48.4	293.6
SEM-POS (Ours)	**60.1**	**38.5**	**76.0**	**108.3**	**607.1**	**45.2**	**30.7**	**64.1**	**53.1**	**192.6**

Table 2. Ablation results for the effectiveness of POS, GLFB, and VFM on MSVD and MSRVTT benchmarks.

in Section 3.2. As seen from 'w/o VFM' in Table 2, masking spatial and temporal features increases the performance to a considerable extent. We use different techniques of masking for spatial and temporal features. For spatial features, we find random masking across the whole frame features to be more effective than masking particular patches in a frame. We have introduced chunk-wise masking for temporal features as random masking is counterproductive in this case. In chunk-wise masking, we mask a chunk of values that are adjacent to each other in the temporal dimension and spatially the same for each frame.

4.3. Grammatical Score Evaluation

In video captioning, in addition to semantic correctness, it is important to check whether the generated captions are grammatically correct. Existing methods only evaluate the semantic correctness of the generated captions, however, in this section, we evaluate the Grammatical Score (GS) of the captions generated from the proposed method against state-of-the-art methods. [39] performed a comprehensive study to use GPT-2 [37] for measuring the grammatical correctness of English sentences. We employ the same technique to measure the average GS of generated captions on both MSVD [8] and MSRVTT [55] datasets as part of our ablation study in Table 2 and we compare the GS against state-of-the-art methods in Table 1. It is based on the perplexity measurement for each caption and the lower score means that the caption has fewer grammatical errors. The results demonstrate that the proposed method with POS blocks and GLFB blocks gives the best performance and generated more descriptive and grammatically correct captions compared to the existing methods achieving the new state-of-the-art.

4.4. Limitations

The performance of each POS block is limited to the number of samples present in the datasets. For e.g., due to a lack of 'auxiliary verb' samples in the dataset, this does not contribute significantly to the final performance. Similarly, other POS, like adverbs and adjectives, are far less represented in the captions, and adding these blocks, affects the performance adversely. Also, the proposed method does not handle multiple subjects/objects like existing methods. We aim to address these limitations in our future work.

5. Conclusion

This paper introduces a novel global-local feature fusion network that fuses local and global, language and visual features for video captioning. Four POS and a GLFB block are introduced within the network to create a coherent representation to generate grammatically and semantically correct captions, which is proved by the performance of the proposed method on MSVD [8] and MSRVTT [55] benchmarks for five different metrics against state-of-the-art methods. The application of Visual Features Masking (VFM) within the novel network helps our model to not only achieve better performance but also generalize well across the data distribution, as seen in the ablation results. Qualitative results show the ability of our model to learn fine-grained details in the video and we believe that the insights from this work will further open up research in video-language alignment and understanding not only for semantic correctness but also for grammatical correctness.

6. Acknowledgement

This research was partly supported by the British Broadcasting Corporation Research and Development (BBC R&D), Engineering and Physical Sciences Research Council (EPSRC) Grant EP/V038087/1 "BBC Prosperity Partnership: Future Personalised Object-Based Media Experiences Delivered at Scale Anywhere".

References

[1] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. Spatio-Temporal Dynamics and Semantic Attribute Enriched Visual Encoding for Video Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12487–12496, 2019. 1, 6, 7

[2] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video Description: A Survey of Methods, Datasets, and Evaluation Metrics. *ACM Computing Surveys (CSUR)*, 52(6):1–37, 2019. 2

[3] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised Learning from Narrated Instruction Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4575–4583, 2016. 1

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. 1

[5] Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, 2005. 6

[6] Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. Curious Case of Language Generation Evaluation Metrics: A Cautionary Tale. *arXiv preprint arXiv:2010.13588*, 2020. 6

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end Object Detection with Transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2

[8] David Chen and William B Dolan. Collecting Highly Parallel Data for Paraphrase Evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 2, 6, 7, 8

[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv preprint arXiv:1406.1078*, 2014. 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 6

[11] Jeffrey L Elman. Finding Structure in Time. *Cognitive Science*, 14(2):179–211, 1990. 2

[12] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video Captioning with Attention-based LSTM and Semantic Consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055, 2017. 2

[13] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2Text: Recognizing and Describing Arbitrary Activities using Semantic Hierarchies and Zero-Shot Recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2712–2719, 2013. 1

[14] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in Transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021. 2

[15] Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention combining Global Knowledge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 221–231, 2017. 2

[16] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and Imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 2, 6

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders are Scalable Vision Learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 4

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. 2

[19] Richang Hong, Yang Yang, Meng Wang, and Xian-Sheng Hua. Learning Visual Semantic Relationships for Efficient Visual Retrieval. *IEEE Transactions on Big Data*, 1(4):152–161, 2015. 1

[20] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in Vision: A Survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 2

[21] Dong-Jin Kim, Tae-Hyun Oh, Jinsoo Choi, and In So Kweon. Dense relational image captioning via multi-task triple-stream networks. *IEEE TPAMI*, 44(11):7348–7362, 2021. 2

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. 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. 1

[24] Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond Mooney, Kate Saenko, and Sergio Guadarrama. Generating Natural-Language Video Descriptions using Text-Mined Knowledge. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013. 1

[25] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical Encoder for Video + Language Omni-Representation Pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 2

[26] Xuelong Li, Bin Zhao, and Xiaoqiang Lu. A General Framework for Edited Video and Raw Video Summarization. *IEEE*

Transactions on Image Processing, 26(8):3652–3664, 2017. 1

[27] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[28] Ke Lin, Zhuoxin Gan, and Liwei Wang. Augmented partial mutual learning with frame masking for video captioning. In *AAAI*, volume 35, pages 2047–2055, 2021. 2, 7

[29] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *CVPR*, pages 17949–17958, 2022. 2, 7

[30] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. CPTR: Full Transformer Network for Image Captioning. *arXiv preprint arXiv:2101.10804*, 2021. 2

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019. 2, 5, 6

[32] Xiang Long, Chuang Gan, and Gerard De Melo. Video Captioning with Multi-Faceted Attention. *Transactions of the Association for Computational Linguistics*, 6:173–184, 2018. 2

[33] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-Temporal Graph for Video Captioning with Knowledge Distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10870–10879, 2020. 1, 2, 6, 7

[34] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly Modeling Embedding and Translation to Bridge Video and Language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4594–4602, 2016. 1

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[36] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-Attended Recurrent Network for Video Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8347–8356, 2019. 1, 6, 7

[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9, 2019. 6, 8

[38] Hobin Ryu, Sunghun Kang, Haeyong Kang, and Chang D Yoo. Semantic Grouping Network for Video Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2514–2522, 2021. 1, 6, 7

[39] Scribendi Inc. Comparing bert and gpt-2 as language models to score the grammatical correctness of a sentence, 2020. 8

[40] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *CVPR*, pages 17959–17968, 2022. 2, 7

[41] Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. Weakly Supervised Dense Video Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1916–1924, 2017. 2

[42] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[43] Jingkuan Song, Lianli Gao, Li Liu, Xiaofeng Zhu, and Nicu Sebe. Quantization-Based Hashing: a General Framework for Scalable Image and Video Retrieval. *Pattern Recognition*, 75:175–187, 2018.

[44] Jingkuan Song, Yuyu Guo, Lianli Gao, Xuelong Li, Alan Hanjalic, and Heng Tao Shen. From Deterministic to Generative: Multimodal Stochastic RNNs for Video Captioning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(10):3047–3058, 2018. 2

[45] Jingkuan Song, Zhao Guo, Lianli Gao, Wu Liu, Dongxiang Zhang, and Heng Tao Shen. Hierarchical LSTM with Adjusted Temporal Attention for Video Captioning. *arXiv preprint arXiv:1706.01231*, 2017. 2

[46] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 2, 6

[47] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015. 2

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[49] 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*, pages 4566–4575, 2015. 6

[50] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to Sequence-Video to Text. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4534–4542, 2015. 1, 2

[51] Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, Jingwen Wang, and Wei Liu. Controllable Video Captioning with Pos Sequence Guidance Based on Gated Fusion Network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2650, 2019. 1, 2, 6, 7

[52] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable Bag-of-Freebies Sets new State-of-the-Art for Real-Time Object Detectors. *arXiv preprint arXiv:2207.02696*, 2022. 6

[53] Jingdong Wang, Ting Zhang, Nicu Sebe, Heng Tao Shen, et al. A Survey on Learning to Hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):769–790, 2017. 1

[54] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SIMMIM: A Simple Framework for Masked Image Modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 2, 4

[55] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT:A Large Video Description Dataset for Bridging Video and Language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016. 2, 6, 7, 8

[56] Jun Xu, Ting Yao, Yongdong Zhang, and Tao Mei. Learning Multimodal Attention LSTM Networks for Video Captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 537–545, 2017. 2

[57] Liqi Yan, Qifan Wang, Yiming Cui, Fuli Feng, Xiaojun Quan, Xiangyu Zhang, and Dongfang Liu. Gl-rg: Global-local representation granularity for video captioning. *arXiv preprint arXiv:2205.10706*, 2022. 1, 7

[58] Hanhua Ye, Guorong Li, Yuankai Qi, Shuhui Wang, Qingming Huang, and Ming-Hsuan Yang. Hierarchical Modular Network for Video Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17939–17948, 2022. 1, 3, 6, 7

[59] Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. Generation for User Generated Videos. In *European Conference on Computer Vision*, pages 609–625. Springer, 2016. 2

[60] Junchao Zhang and Yuxin Peng. Object-Aware Aggregation With Bidirectional Temporal Graph for Video Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8327–8336, 2019. 1, 2, 6, 7

[61] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. CAM-RNN: Co-Attention Model based RNN for Video Captioning. *IEEE Transactions on Image Processing*, 28(11):5552–5565, 2019. 2

[62] Bin Zhao, Xuelong Li, Xiaoqiang Lu, and Zhigang Wang. A CNN–RNN Architecture for Multi-Label Weather Recognition. *Neurocomputing*, 322:47–57, 2018. 1

[63] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 16259–16268, 2021. 2

[64] Qi Zheng, Chaoyue Wang, and Dacheng Tao. Syntax-Aware Action Targeting for Video Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13096–13105, 2020. 1, 2, 6, 7

[65] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for end-to-end Object Detection. *arXiv preprint arXiv:2010.04159*, 2020. 3