Thinking Hallucination for Video Captioning

Nasib Ullah[1][*] and Partha Pratim Mohanta[2]

[1] LIVIA, Dept. of Systems Engineering, ÉTS, Montreal, Canada
nasibullah.nasibullah.1@ens.etsmtl.ca
[2] ECSU, Indian Statistical Institute, Kolkata, India
ppmohanta@isical.ac.in

Abstract. With the advent of rich visual representations and pre-trained language models, video captioning has seen continuous improvement over time. Despite the performance improvement, video captioning models are prone to hallucination. Hallucination refers to the generation of highly pathological descriptions that are detached from the source material. In video captioning, there are two kinds of hallucination: object and action hallucination. Instead of endeavoring to learn better representations of a video, in this work, we investigate the fundamental sources of the hallucination problem. We identify three main factors: (i) inadequate visual features extracted from pre-trained models, (ii) improper influences of source and target contexts during multi-modal fusion, and (iii) exposure bias in the training strategy. To alleviate these problems, we propose two robust solutions: (a) the introduction of auxiliary heads trained in multi-label settings on top of the extracted visual features and (b) the addition of context gates, which dynamically select the features during fusion. The standard evaluation metrics for video captioning measures similarity with ground truth captions and do not adequately capture object and action relevance. To this end, we propose a new metric, COAHA (caption object and action hallucination assessment), which assesses the degree of hallucination. Our method achieves state-of-the-art performance on the MSR-Video to Text (MSR-VTT) and the Microsoft Research Video Description Corpus (MSVD) datasets, especially by a massive margin in CIDEr score.

1 Introduction

Video captioning is the translation of a video into a natural language description. It has many potential applications, including video retrieval, human-computer interface, assisting visually challenged, and many more. The encoder-decoder based sequence to sequence architecture (initially proposed for machine translation) has helped video captioning exceptionally. Pre-trained vision models extract the frame features at the encoder section, and the decoder is a conditional language model. Recent improvements happened in broadly three areas, better visual feature extraction [23,20], better and support of external pre-trained language models [21,47] and, better strategy [9,29] to sample informative frames.

---

[*] This work was done while Nasib Ullah was a PLP at ECSU, ISI Kolkata, India.
The code is available at: https://github.com/nasib-ullah/THVC

Object Hallucination

MARN: A man is riding a horse.
BLEU4: 76.0, METEOR: 42.4, ROUGE_L: 83.3, COAHA: 8.12

Action Hallucination

MARN: A woman is drawing a paper.
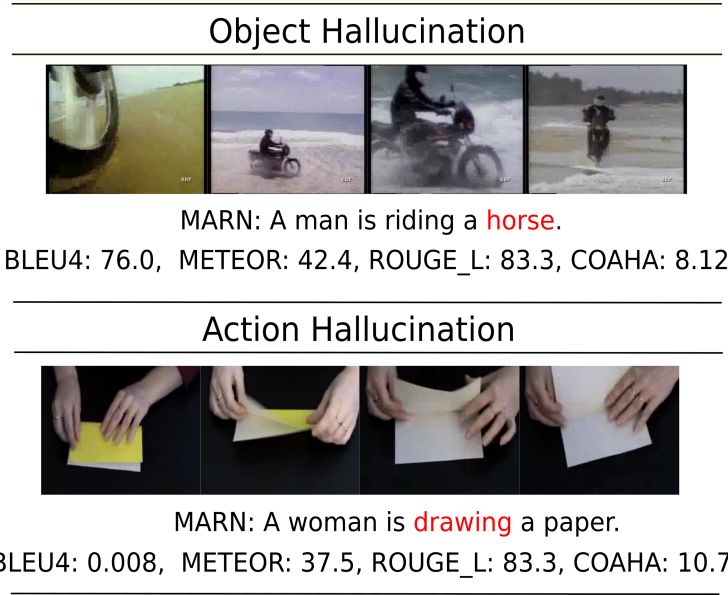BLEU4: 0.008, METEOR: 37.5, ROUGE_L: 83.3, COAHA: 10.7

Fig. 1. Video Captioning models suffer from two types of hallucination, Object hallucination (above) and action hallucination (below). Unfortunately, the standard metrics fail to adequately capture hallucination, whereas our proposed COAHA metric is more relevant to hallucination. The results are generated using the MARN [23] model. For COAHA, lower is better.

Despite the improvements, a significant problem remains that is known as a hallucination [27,38,24] in the literature. Unlike image captioning, there are two types of hallucination in the case of video captioning: object hallucination occurs when the model describes objects that are not present in the video, and action hallucination occurs when the model describes an action that has not been performed in the video as depicted in Fig. 1. Another evidence to understand that hallucination is a significant problem can be seen from the token position vs. model confidence plot in Fig. 2. Generally, the object and action words occur towards the middle of the caption. Thus, we can see that the model has low confidence while generating tokens related to objects and actions.

In this work, we point out three fundamental sources of hallucination for video captioning: (i) inadequate visual features from pre-trained models, (ii) improper influence of features during intra-modal and multi-modal fusion, and (iii) exposure bias in the model training strategy. We propose two robust and straightforward techniques that alleviate the problem. The exposure bias problem acts mainly under domain shift [38,19], so we will not be focusing on that. Standard evaluation metrics only measure similarity with ground truth caption and may not fully capture the object and action relevance. To this end, we propose a new metric COAHA (caption object and action hallucination assessment)
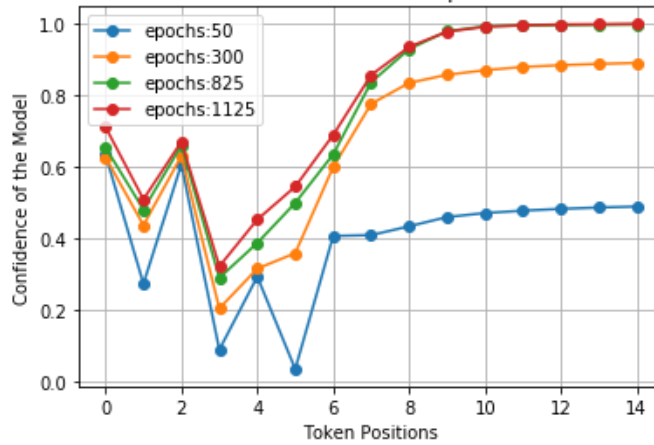
Fig. 2. The plot of the model confidence at the different token positions. The plot is done by averaging confidence over validation set on MSVD data and using the model MARN [23].

to assess the hallucination rate. The CHAIR metric [27] has been proposed for image captioning. However, it considers only object hallucination and relies on object segmentation annotation unavailable for video captioning datasets. Unlike CHAIR, our metric considers both object and action hallucination and leverage existing captions instead of relying on segmentation annotations.

Finally, we show that our solutions outperform the state-of-the-art models in standard metrics, especially by a large margin in CIDEr-D on Microsoft Research Video Description Corpus (MSVD) and MSR-Video to Text (MSR-VTT) datasets.

2   Related Work

2.1   Video Captioning.

The main breakthrough in video captioning happened with the encoder-decoder based sequence to sequence architecture. Although MP-LSTM [35] applied this first, they did not capture the temporal information between frames. S2VT [34] and SA-LSTM [43] capture the temporal dependencies between frames. The former shares a single LSTM network for both encoder and decoder, whereas the later uses attention weights over frames and 3D HOG features. Recent methods have improved upon SA-LSTM [43]. For example, RecNet [2] has added reconstruction loss based on backward flow from decoder output to the video feature construction. MARN [23] has used external memory to capture various visual contexts corresponding to each word. Both MARN [23] and M3 [39] utilize motion and appearance features. However, unlike MARN [23], M3 [39] uses hetero-

geneous memory to model long-term visual-text dependency. OA-BTG [45] and STG-KD [20] both use object features along with appearance and motion features. OA-BTG [45] leverage the temporal trajectory features of salient objects in the video. In contrast, STG-KD [20] uses a spatio-temporal graph network for object interaction features and a Transformer network for the language model instead of recurrent neural networks. ORG-TRL [47] also uses object interaction features, but unlike STG-KD [20], it utilizes an external language model which guides the standard recurrent neural network based language model. Another line of work focuses on a better sampling strategy to choose informative video frames. PickNet [9] samples informative frames based on reward-based objectives, whereas SGN [29] uses partially decoded information to sample frames. Regardless of the improvements, hallucination is still one of the major problems in existing models. In this work, we will be focusing on significant sources of hallucination, how to mitigate them, and finally, a metric to measure the rate of hallucination.

2.2 Hallucination Problems in Natural Language Generation.

Koehn et al. [16] pointed out the problem of hallucination among the six challenges in neural machine translation. Müller et al. [19] and Wang et al. [38] linked the problem of hallucination with exposure bias. Wang et al. [38] have used minimum risk training (MRT) instead of Maximum likelihood (MLE) to reduce the exposure bias and concluded the improvement mostly happens under domain shift. Lee et al. [17] has claimed that NMT models hallucinate under source perturbation. Tu et al. [31] has used the context gates to control the source and target contribution, reducing the hallucination for neural machine translation. In the case of image captioning, Anna et al. [27] discussed object hallucination and proposed CHAIR metric to assess the degree of object hallucination. Unlike neural machine translation and image captioning, video captioning is more complex because of the spatio-temporal nature of the input data. Although state-of-art video captioning models suffer from hallucination, there is no pre-attempt. To the best of our knowledge, this is the first attempt to address hallucination for video captioning.

3 Methodology

As shown in Fig. 3, the components of our proposed system are (a) Visual Encoder responsible for the extraction of different visual features from the input video, (b) Auxiliary heads to reduce the hallucination that occurs from inadequate visual features, (c) Context gates to dynamically control the importance of features during intra-modal and multi-modal fusion and, (d) Decoder which is a conditional language model. We utilize running visual and language memory along with these components to act as an external short-term memory. We have used multi-label Auxiliary head loss and Coherent loss along with the standard cross-entropy loss for training.
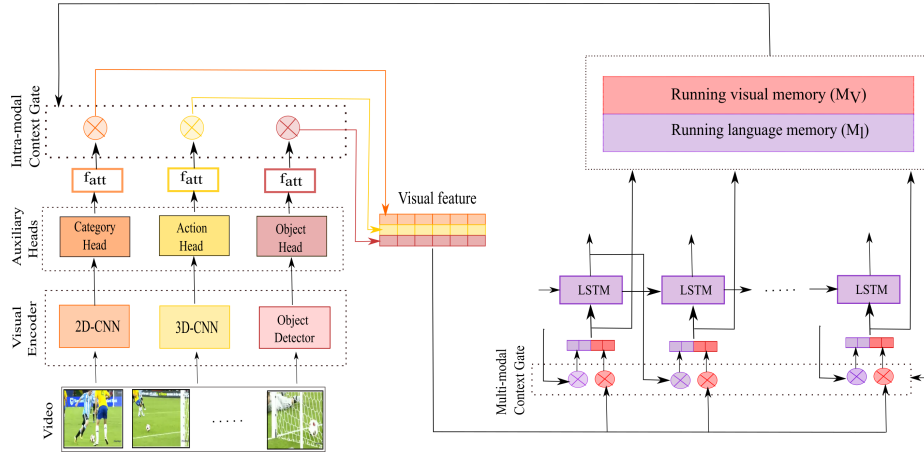
Fig. 3. The architecture of our proposed model. The components are (1) Visual Encoder for visual feature extraction, (2) auxiliary heads, and (3) context gates to reduce hallucination, and finally, (4) Decoder for sentence generation.

3.1 Visual Encoder

From a given input video, we uniformly sample N frames $\{f_i\}_{i=1}^N$ and clips $\{c_i\}_{i=1}^N$, where each c_i is a sequence of frames around f_i. We extract appearance features $\{a_i\}_{i=1}^N$ and motion features $\{m_i\}_{i=1}^N$ using pre-trained 2D CNN ϕ^a [10] and 3D CNN ϕ^m [13] respectively, where $a_i = \phi^a(f_i)$ and $m_i = \phi^m(c_i)$. Apart from appearance and motion, we also extract object features $\{o_i\}_{i=1}^N$ using a pre-trained object detection module ϕ^o [26] where $o_i = \phi^o(f_i)$. For each frame, we extract salient objects based on the objectiveness threshold v and average the features of those salient objects. Appearance ($\{a_i\}_{i=1}^N$) and motion ($\{m_i\}_{i=1}^N$) feature helps to understand the global context and the motion information of the video. In contrast, the object features ($\{o_i\}_{i=1}^N$) are more localized and helps to understand fine-grained information.

3.2 Auxiliary Heads: Category, Action, and Object

Unlike image captioning, the sequence of frames in video captioning does make the training of the feature extraction module computationally extensive. Existing methods use pre-trained models for different visual feature extraction and consider these features as the input to the model. However, the pre-trained models are trained on different data distribution, and extracted features are not adequate, which leads to the hallucination problem. We hypothesize that inadequate object and appearance features are responsible for object hallucination, whereas inadequate motion features lead to action hallucination.

   We propose specialized classification heads on top of each type of visual feature, and they are trained in multi-label settings. The supervision is generated by

leveraging existing captions, and the features from the heads are more suitable and semantically relevant. We include object head ψ_{Obj}^h on top of the object features $\{o_i\}_{i=1}^N$, where the training objective is to predict visual objects present in the caption for that input video. Similarly, action head ψ_{Act}^h applies to motion features $\{m_i\}_{i=1}^N$ to predict actions present in the caption. The ground truth for both the action head and object head are extracted using Named Entity Recognition [12,7] and Parts-of-speech tagging [37] methods from NLP literature. To refine the appearance features $\{a_i\}_{i=1}^N$, we add categorical head ψ_{Cat}^h to predict the video category, which is available for the MSR-VTT dataset but not for MSVD. Although we have used three specific objectives, this framework can be applied with other objectives, such as predicting attributes on top of the appearance feature. Finally, we use the features $\{a_i^r\}_{i=1}^N$, $\{m_i^r\}_{i=1}^N$ and $\{o_i^r\}_{i=1}^N$ from the specific auxiliary head as the input to the model, where $a_i^r = \psi_{Cat}^h(a_i)$, $m_i^r = \psi_{Act}^h(m_i)$ and $o_i^r = \psi_{Obj}^h(o_i)$. The Fig. 4 shows the modeling of the action head.

3.3   Context Gates: Balancing Intra-Modal and Multi-Modal Fusion

There are two types of fusion of features in the existing video captioning framework: intra-modal fusion and inter-modal or multi-modal fusion. In intra-modal fusion, different visual (appearance, motion, and object) features are merged, and thus one feature might dominate other vital features. For example, motion features might dominate appearance and object features, which lead to object hallucination, and action hallucination occurs due to dominance of object and appearance features over motion features. Similarly, in the case of multi-modal fusion at the decoder, the dominant language prior leads to hallucination [36,31], whereas the dominant source context leads to a not fluent caption.

We have included two separate context gate units for intra-modal and multi-modal fusion of features, which mitigate the hallucination. The gates are responsible for dynamically selecting the importance of each type of feature. Tu et al. [31] have used the context gate before to balance the source and target context in neural machine translation. Nevertheless, unlike Tu et al. [31], we have used separate gates for intra-modal and multi-modal fusion. Also, instead of depending on the decoder memory, our gates are conditioned on running visual (M_v) and language (M_l) memory as shown in Fig. 3. Empirical studies have shown better performance with separate running memories rather than decoder memory. One possible reason might be the decoder memory is contaminated by a mixture of source and target features and noisy LSTM gates.
At the encoder, the combined visual feature $v_{m,t}$ is,

$$v_{m,t} = [CG_a^E \circ a_t^r; CG_m^E \circ m_t^r; CG_o^E \circ o_t^r] \tag{1}$$

where,

$$a_t^r = \sum_{i=1}^N \alpha_{i,t} a_{i,t}^r, m_t^r = \sum_{i=1}^N \alpha_{i,t} m_{i,t}^r, o_t^r = \sum_{i=1}^N \alpha_{i,t} o_{i,t}^r \tag{2}$$
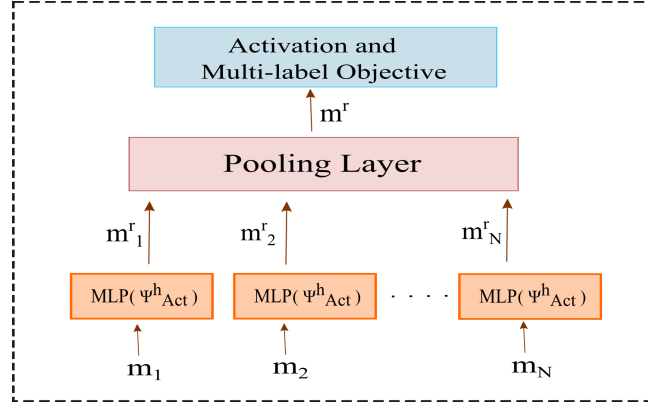
Fig. 4. Diagram of action auxiliary head. $\{m_i\}_{i=1}^N$ are the motion features from visual encoder. The MLP is shared over clip features. The same architecture follows for object and categorical auxiliary heads.

$$CG_{a/m/o}^E = f(a_t^r, m_t^r, o_t^r, M_v, M_l) \tag{3}$$

$CG_{a/m/o}^E$ are the context gates corresponding to appearance, motion, and object features, respectively, and $\alpha_{i,t}$ are the attention weights [43]. We share the attention network over the object, motion, and appearance features for the regularization purpose.

At the decoder, the multi-modal feature C_t is,

$$C_t = [CG_S^D \circ v_{m,t}; CG_T^D \circ E(y_{t-1})] \tag{4}$$

where,

$$CG_{S/T}^D = f(v_{m,t}, E(y_{t-1}), M_v, M_l) \tag{5}$$

$CG_{S/T}^D$ are the context gates corresponding to source and target context, respectively, and $E(y_{t-1})$ is the embedding feature of the previous token. f is realized using MLP, [;] denotes concatenation, and $\circ$ denotes Hadamard product.

3.4 Decoder

The decoder is a language model conditioned on merged visual and language prior features C_t. We have used the LSTM network because of its superior sequential modeling performance. At t time steps during the sentence generation, the hidden memory of LSTM can be expressed as,

$$h_t = LSTM(C_t, h_{t-1}) \tag{6}$$

where h_{t-1} is the hidden memory at previous time steps, and C_t is the combined multi-modal feature as shown in Equation 4. Although transformer-based [32]

models have outperformed LSTM in most NLP benchmarks, LSTM and transformer models performed almost equally in our case. Finally, the probability distribution over words is generated using a fully connected layer followed by a softmax layer.

$$P(y_t|V, y_1, y_2, .., y_{t-1}) = softmax(V_h h_t + b_h) \tag{7}$$

where V_h and b_h are learnable parameters.

3.5   Parameter Learning

Along with the standard cross-entropy loss, our model is trained using two additional losses, Auxiliary head loss, and Coherent loss.

Attention Based Decoder. Negative log-likelihood (or cross-entropy) is the standard objective function for the video captioning models. The loss for a mini-batch is

$$L_{CE} = -\sum_{i=1}^{B}\sum_{t=1}^{T} \log p(y_t|V, y_1, y_2, .., y_{t-1}; \theta) \tag{8}$$

Where θ is learnable parameters, V is the video feature, y_t is the tth word in the sentence of length T, and B is the mini-batch size.

Auxiliary Heads. Auxiliary heads are trained in multi-label settings. For all three heads, we have used binary cross-entropy loss over sigmoid activation. The total auxiliary loss is,

$$L_{AH} = \lambda_C L_{AH}^C + \lambda_A L_{AH}^A + \lambda_O L_{AH}^O \tag{9}$$

where λ_C, λ_A and λ_O are hyperparameters corresponding to Category head loss L_{AH}^C, Action head loss L_{AH}^A, and Object head loss L_{AH}^C, respectively. Each loss is calculated as

$$L_{AH}^C = \Upsilon(Y^C, \psi_{Cat}^h(a_j\,_{j=1}^N)), L_{AH}^A = \Upsilon(Y^A, \psi_{Act}^h(m_j\,_{j=1}^N))$$

$$L_{AH}^O = \Upsilon(Y^O, \psi_{Obj}^h(o_j\,_{j=1}^N))$$

$$where,$$

$$\Upsilon(y, p) = \sum_{i=1}^{B}\sum_{m=1}^{M} [y_{i,m} \log p_{i,m} + (1 - y_{i,m}) \log(1 - p_{i,m})] \tag{10}$$

Y^C, Y^A and Y^O are category, action and object ground truth for the given batch (B) of data.

Coherent Loss. The consecutive frames in a video are highly repetitive. So encoding of consecutive frames should be similar. We apply the coherent loss to restrict the embedding of consecutive frames to be similar. Coherent loss [23] has been used before to regularize attention weights, but unlike Pei et al. [23], we apply coherent loss on appearance, motion, and object features. It is debatable whether to apply coherent loss on motion features, but empirically we got a good result, and the possible reason might be shorter clip length. The total coherent loss for a mini-batch is,

$$L_{CL} = \lambda_{fcl} L_{CL}^a + \lambda_{mcl} L_{CL}^m + \lambda_{ocl} L_{CL}^o + \lambda_{acl} L_{CL}^\alpha \tag{11}$$

where $\lambda_{fcl}, \lambda_{mcl}, \lambda_{ocl}$ and λ_{acl} are hyperparameters corresponding to appearance coherent loss L_{CL}^a, motion coherent loss L_{CL}^m, object coherent loss L_{CL}^o and attention coherent loss L_{CL}^α respectively.

The individual coherent losses are calculated as, $L_{CL}^a = \Phi(a_i^r)$, $L_{CL}^m = \Phi(m_i^r)$, $L_{CL}^o = \Phi(o_i^r)$ and $L_{CL}^\alpha = \Phi(\alpha_i)$ where,

$$\Phi(f) = \sum_{i=1}^B \sum_{t=1}^T \sum_{n=2}^N |f_{n,t}^{(i)} - f_{n-1,t}^{(i)}| \tag{12}$$

The overall loss function of our model is

$$L = L_{CE} + L_{AH} + L_{CL} \tag{13}$$

3.6 The COAHA Metric

In order to measure the degree of hallucination, we propose COAHA (caption object and action hallucination assessment) metric. The metric is defined for per instance and is given by,

$$COAHA = OH + AH \tag{14}$$

where,

$$OH = \frac{\sum_{h_i \in H_O} d_i^h}{T}, \quad AH = \frac{\sum_{a_i \in H_A} d_i^a}{T} \tag{15}$$

OH measures the degree of object hallucination, and AH measures the degree of action hallucination. T is the average ground truth caption length and N_O and N_A are the set of objects and actions mentioned in the video. H_O is the set of hallucinated objects, and H_A is the set of hallucinated actions. d_i^h measures the average semantic distance between the current hallucinated object and all objects in N_O. Similarly, d_i^a measures the average semantic dissimilarity of hallucinated action. The higher d_i^h or higher cardinality of H_O will lead to a higher value of OH, and the same follows for AH. So, if the hallucinated object or action is a synonym, then the semantic distance would be small, which will lead to a low hallucination value. Finally, we have taken the addition of OH and AH. We have applied the same approach mentioned in the Auxiliary Head section

to extract the list of objects and actions from each video (from the ground truth caption). For measuring the semantic distance (d_i^h and d_i^a) we have used pre-trained Fasttext [6] embedding model. Mathematically,

$$d_i^h = \frac{1}{|N_O|} \sum_{w_k \in N_O} \vartheta(w_k, h_i) \qquad (16)$$

where, ϑ is the cosine similarity on top of the Fasttext [6] embedding. Similarly the formulae follows for d_i^a.

4   Experimental Results

We evaluate our proposed model on two most popular benchmark datasets: MSVD and MSR-VTT. To compare with the state-of-the-art results, we have used the four most popular metrics: CIDEr [33], METEOR [3], ROUGE-L [18] and BLEU-4 [22].

4.1   Datasets

MSVD. Microsoft Video Description (MSVD) dataset [8] consists of 1970 open domain videos of single activity and is described by 40 captions generated by Amazon Mechanical Turk. For a fair comparison, we have followed the standard split [35,43] of 1200 videos for training, 100 for validation, and 670 for testing.
MSR-VTT. MSR Video-to-Text (MSR-VTT) dataset [42] is a large scale benchmark datasets with 10k videos and 20 categories. Each video is annotated with 20 captions with an average length of 20 seconds. We have followed the standard benchmark split [42] of 6513 for training, 497 for validation, and 2990 for testing.

4.2   Implementation Details

Feature Extraction and Decoding. We uniformly sample 28 frames per video, and sentences longer than 30 words are truncated. The 1024-D appearance features of each frame are extracted by ViTL [10] pre-trained on Imagenet [28]. For the 2048-D motion features, we use C3D [13] with ResNeXt-101 [41] and pre-trained on the Kinetic-400 dataset. To extract the object features, we apply Faster-RCNN [26] with ResNet-101 and FPN backbone pre-trained on MSCOCO [40]. Appearance, motion, and objects features are embedded to 512-D before sending to the next stage. At the decoder LSTM, we use a hidden layer size of 512, and the word embedding size is fixed to 512.The vocabulary is built by words with at least 5 occurrences. During testing, we use greedy decoding to generate the sentences.
Auxiliary Heads and Context Gates Related Details. The running visual and language memory size is set to 512. We use the Spacy [14] module pre-trained with transformers [32] for object and action extraction from ground

Models	MSVD				MSR-VTT			
	B@4	M	R	C	B@4	M	R	C
SA-LSTM [43]	45.3	31.9	64.2	76.2	36.3	25.5	58.3	39.9
h-RNN [44]	44.3	31.1	-	62.1	-	-	-	-
hLSTMat [30]	53.0	33.6	-	73.8	38.3	26.3	-	-
RecNet [2]	52.3	34.1	69.8	80.3	39.1	26.6	59.3	42.7
M3 [39]	52.8	33.3	-	-	38.1	26.6	-	-
PickNet [9]	52.3	33.3	69.6	76.5	41.3	27.7	59.8	44.1
MARN [23]	48.6	35.1	71.9	92.2	40.4	28.1	60.7	47.1
GRU-EVE [1]	47.9	35.0	71.5	78.1	38.3	28.4	60.7	48.1
POS+CG [37]	52.5	34.1	71.3	88.7	42.0	28.2	61.6	48.7
OA-BTG [45]	**56.9**	36.2	-	90.6	41.4	28.2	-	46.9
STG-KD [20]	52.2	**36.9**	73.9	93.0	40.5	28.3	60.9	47.1
SAAT [46]	46.5	33.5	69.4	81.0	40.5	28.2	60.9	49.1
ORG-TRL [47]	54.3	36.4	73.9	95.2	**43.6**	28.8	**62.1**	50.9
SGN [29]	52.8	35.5	72.9	94.3	40.8	28.3	60.8	49.5
Ours	53.3	36.5	**74.0**	**99.9**	41.1	**28.9**	61.9	**51.7**

Table 1. Performance comparison on MSVD and MSR-VTT benchmarks. B4, M, R, and C denote BLEU-4, METEOR, ROUGE_L, and CIDEr, respectively.

truth captions. We have also used Porter Stemmer [5] to work with the root form of words. Finally, to measure the semantic distance between words, we have used the pre-trained Fasttext [6] model from Gensim [25].

Other Details. We apply Adam [15] with a fixed learning rate of 1e-4 and a gradient clip value of 5. Our model is trained for 900 epochs with a batch size of 100. The coherent loss weights λ_{acl}, λ_{fcl}, λ_{mcl}, and λ_{ocl} are set as 0.01, 0.1, 0.01, and 0.1, respectively. All the experiments are done in a single Titan X GPU.

4.3   Quantitative Results

To proclaim the effectiveness of our approach, we compare the performance of our model with state-of-the-art models. The quantitative results in Table 1 show that our model performs significantly better than other methods, especially in CIDEr. CIDEr is specially designed for captioning tasks and is more similar to human judgment than other metrics. Compared to OA-BTG [45] and ORG-TRL [47] our model lags in the BLEU-4 score, which is more suitable and designed for machine translation evaluation. Also, ORG-TRL [47] and OA-BTG [45] have utilized object interaction and object trajectory features, respectively, whereas we have only taken mean localized object features for simplicity.

4.4   Ablation Studies

We perform the quantitative evaluation to investigate the effectiveness of the components we proposed. We conduct ablation studies that start with a primary

Methods						
A.Heads	C.Gates	B@4	M	R	C	COAHA
×	×	50.1	34.8	72.9	91.1	10.57
✓	×	52.6	36.2	73.4	96.6	7.90
×	✓	52.5	36.3	73.6	97.1	7.96
✓	✓	53.3	36.5	74.0	99.9	7.02

Table 2. Ablation studies of the context gates and auxiliary heads on MSVD benchmark. C.Gates denotes context gates, and A.Heads denotes auxiliary heads.

decoder and add auxiliary heads and context gates incrementally. The Table 2 shows the results.

Caption	OH	AH	COAHA
A man is riding a motorcycle	0	0	0
A man is riding a **horse**	8.12	0	8.12
A man is riding a **car**	5.10	0	5.10
A **woman** is riding a motorcycle	4.67	0	4.67
An **animal** is riding a motorcycle	9.58	0	9.58
A man is **playing** with a motorcycle	0	9.52	9.52
A man is **eating** a motorcycle	0	10.4	10.4
A man is **playing** with a **toy**	6.45	9.52	15.97

Table 3. Semantic significance of COAHA metric. OH and AH represent object and action hallucination scores respectively. For OH, AH and COAHA, lower is better.

4.5   Validation of Context Gates

To validate that the context gates are helping in dynamically selecting the feature importance, we have shown the values of context gates during multi-modal fusion. The Fig. 5(a) shows that the source context gate value is high during the generation of visual object words. In contrast, the target context gate value is high during the generation of non-visual words.

4.6   Effect of Exposure Bias on Hallucination

In the case of machine translation, it is proven [38,19] that exposure bias affects mainly under domain shift. In order to validate that exposure bias does not affect in same domain cases, we have shown the model confidence (Fig. 5(b)) and
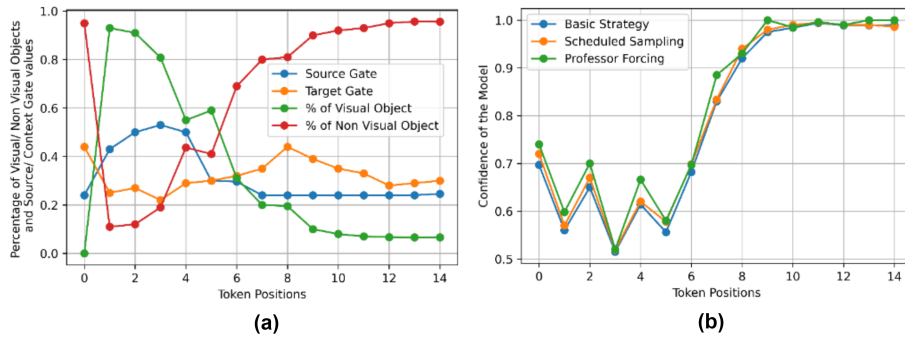
Fig. 5. (a) Contribution of context gates during the generation of visual and non-visual words. (b) The model confidence vs. token positions under different training strategies.

Training Strategy	COAHA
Teacher Forcing	7.02
Scheduled Sampling	7.01
Professor Forcing	6.98

Table 4. COAHA scores at different training strategies.

COAHA values (Table 4) at different training strategies. Although the scheduled sampling [4] and the professor forcing [11] are designed to mitigate exposure bias, they do not affect hallucination significantly.

4.7   Significance of COAHA

In order to show the semantic significance of our proposed metric, we have calculated OH (Object hallucination rate), AH (Action hallucination rate), and COAHA for different semantically perturbed captions, as shown in Table 3.

4.8   Qualitative Results

We have compared the captions generated by our model and by Pei et al. [23]. From Fig. 6, we can see that our model is less prone to hallucination and better identify objects and actions in the video.

5   Conclusion

This work shows that hallucination is a significant problem for video captioning and points out three significant hallucination sources in the existing framework.
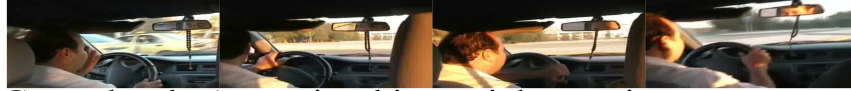
Groundtruth: A man rides a bike through beach.
MARN: A man is riding a horse.
Ours: A man is riding a motorcycle.
(a)

Groundtruth: A man is taking a right turn in a car.
MARN: A car is driving a car.
Ours: A man is driving a car.
(b)

Groundtruth: A man is eating spaghatti out of a large bowl.
MARN: A man is pouring a chicken.
Ours: A man is eating a bowl of food.
(c)

Groundtruth: Someone is folding a piece of paper.
MARN: A woman is drawing a paper.
Ours: A man is folding a piece of paper.
(d)

Fig. 6. Captions generated by our model and MARN [23].

We propose two robust and straightforward strategies to mitigate hallucination. We have got new state-of-the-art on two popular benchmark datasets by applying our solutions, especially by a significant margin in CIDEr score. Furthermore, we have shown that exposure bias is not a significant issue in same domain cases. Finally, we propose a new metric to measure the hallucination rate that the standard metrics cannot capture. Qualitative and quantitative experimental results show our model's effectiveness in overcoming the hallucination problem in video captioning.

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 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pages 12487–12496 (2019)
2. Bairui, W., Lin, M., Wei, Z., Wei, L.: Reconstruction network for video captioning. In Proceed- ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
3. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005. pp. 65–72. Association for Computational Linguistics (2005), `https://aclanthology.org/W05-0909/`
4. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015,Montreal, Quebec, Canada (2015)
5. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly (2009), `http://www.oreilly.de/catalog/9780596516499/index.html`
6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguistics (2017)
7. Chen, D., Manning, C.D.: A fast and accurate dependency parser using neural networks. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, Qatar (2014)
8. Chen, D.L., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA. pp. 190–200. The Association for Computer Linguistics (2011), `https://aclanthology.org/P11-1020/`
9. 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. (2018)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria (2021)
11. Goyal, A., Lamb, A., Zhang, Y., Zhang, S., Courville, A.C., Bengio, Y.: Professor forcing: A new algorithm for training recurrent networks. Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, Barcelona, Spain (2016)
12. Guillaume, L., Miguel, B., Sandeep, S., Kazuya, K., Chris, D.: Neural architectures for named entity recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2016)
13. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)

14. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), to appear

15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA (2015)

16. Koehn, P., Knowles, R.: Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017. pp. 28–39. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/w17-3204, `https://doi.org/10.18653/v1/w17-3204`

17. Lee, K., Firat, O., Agarwal, A., Fannjiang, C., Sussillo, D.: Hallucinations in neural machine translation. In NIPS 2018 Interpretability and Robustness for Audio, Speech and Language Workshop (2018)

18. Lin, C.: Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop, volume 8. Barcelona, Spain (2004)

19. Müller, M., Rios, A., Sennrich, R.: Domain robustness in neural machine translation. In: Proceedings of the 14th Conference of the Association for Machine Translation in the Americas, AMTA 2020, Virtual, October 6-9, 2020. pp. 151–164. Association for Machine Translation in the Americas (2020), `https://aclanthology.org/2020.amta-research.14/`

20. Pan, B., Cai, H., Huang, D., Lee, K., Gaidon, A., Adeli, E., Niebles, J.C.: Spatiotemporal graph for video captioning with knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)

21. Pan, P., Xu, Z., Yang, Y., Wu, F., Zhuang, Y.: Hierarchical recurrent neural encoder for video representation with application to captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016)

22. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. pp. 311–318. ACL (2002). https://doi.org/10.3115/1073083.1073135, `https://aclanthology.org/P02-1040/`

23. Pei, W., Zhang, J., Wang, X., Ke, L., Shen, X., Tai, Y.: Memory-attended recurrent network for video captioning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 8347–8356. Computer Vision Foundation / IEEE (2019)

24. Raunak, V., Menezes, A., Junczys-Dowmunt, M.: The curious case of hallucinations in neural machine translation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021. pp. 1172–1183. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.naacl-main.92, `https://doi.org/10.18653/v1/2021.naacl-main.92`

25. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), `http://is.muni.cz/publication/884893/en`

26. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. (2017)

27. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object halluci-
nation in image captioning. In: Proceedings of the 2018 Conference on Empirical
Methods in Natural Language Processing, Brussels, Belgium, October 31 - Novem-
ber 4, 2018. pp. 4035–4045. Association for Computational Linguistics (2018).
https://doi.org/10.18653/v1/d18-1437, `https://doi.org/10.18653/v1/d18-1437`
28. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z.,
Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale
visual recognition challenge. International Journal of Computer Vision, (2015)
29. Ryu, H., Kang, S., Kang, H., Yoo, C.D.: Semantic grouping network for video
captioning. Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021
(2021)
30. Song, J., Gao, L., Guo, Z., Liu, W., Zhang, D., Shen, H.T.: Hierarchical lstm
with adjusted temporal attention for video captioning. In Proceedings of the 26th
International Joint Conference on Artificial Intelligence (2017)
31. Tu, Z., Liu, Y., Lu, Z., Liu, X., Li, H.: Context gates for neural machine translation.
Trans. Assoc. Comput. Linguistics **5**, 87–99 (2017), `https://transacl.org/ojs/`
`index.php/tacl/article/view/948`
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser,
L., Polosukhin, I.: Attention is all you need. In Advances in Neural Information
Processing Systems (2017)
33. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description
evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition,
CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 4566–4575. IEEE Computer
Society (2015). https://doi.org/10.1109/CVPR.2015.7299087, `https://doi.org/`
`10.1109/CVPR.2015.7299087`
34. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R.J., Darrell, T., Saenko, K.:
Sequence to sequence - video to text. In: 2015 IEEE International Conference on
Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 4534–
4542. IEEE Computer Society (2015). https://doi.org/10.1109/ICCV.2015.515,
`https://doi.org/10.1109/ICCV.2015.515`
35. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R.J., Saenko, K.:
Translating videos to natural language using deep recurrent neural networks. In:
NAACL HLT 2015, The 2015 Conference of the North American Chapter of
the Association for Computational Linguistics: Human Language Technologies,
Denver, Colorado, USA, May 31 - June 5, 2015. pp. 1494–1504. The Associa-
tion for Computational Linguistics (2015). https://doi.org/10.3115/v1/n15-1173,
`https://doi.org/10.3115/v1/n15-1173`
36. Voita, E., Sennrich, R., Titov, I.: Analyzing the source and target contributions to
predictions in neural machine translation. In: Proceedings of the 59th Annual Meet-
ing of the Association for Computational Linguistics and the 11th International
Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume
1: Long Papers), Virtual Event, August 1-6, 2021. pp. 1126–1140. Association for
Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.91,
`https://doi.org/10.18653/v1/2021.acl-long.91`
37. Wang, B., Ma, L., Zhang, W., Jiang, W., Wang, J., Liu, W.: Controllable video
captioning with POS sequence guidance based on gated fusion network. In The
IEEE International Conference on Computer Vision (ICCV) (2019)
38. Wang, C., Sennrich, R.: On exposure bias, hallucination and domain shift
in neural machine translation. In: Proceedings of the 58th Annual Meeting
of the Association for Computational Linguistics, ACL 2020, Online, July 5-
10, 2020. pp. 3544–3552. Association for Computational Linguistics (2020).

https://doi.org/10.18653/v1/2020.acl-main.326, `https://doi.org/10.18653/v1/2020.acl-main.326`

39. Wang, J., Wang, W., Huang, Y., Wang, L., Tan, T.: M3: Multimodal memory modelling for video captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
40. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019)
41. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pages 5987–5995 (2017)
42. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: A large video description dataset for bridging video and language. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 5288–5296. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.571, `https://doi.org/10.1109/CVPR.2016.571`
43. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C.J., Larochelle, H., Courville, A.C.: Describing videos by exploiting temporal structure. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 4507–4515. IEEE Computer Society (2015). https://doi.org/10.1109/ICCV.2015.512, `https://doi.org/10.1109/ICCV.2015.512`
44. Yu, H., Wang, J., Huang, Z., Yang, Y., Xu, W.: Video paragraph captioning using hierarchical recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
45. Zhang, J., Peng, Y.: Object-aware aggregation with bidirectional temporal graph for video captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
46. 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 (2020)
47. Ziqi, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., Zha, Z.: Object relational graph with teacher- recommended learning for video captioning. In Proceed- ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)