Leveraging Weighted Fine-Grained Cross-Graph Attention for Visual and Semantic Enhanced Video Captioning Network

Deepali Verma, Arya Haldar, Tanima Dutta

Department of Computer Science and Engineering, IIT(BHU) Varanasi
{deepaliverma.rs.cse19, arya.haldar.cse17, tanima.cse}@itbhu.ac.in

Abstract

Video captioning has become a broad and interesting research area. Attention-based encoder-decoder methods are extensively used for caption generation. However, these methods mostly utilize the visual attentive feature to highlight the video regions while overlooked the semantic features of the available captions. These semantic features contain significant information that helps to generate highly informative human description-like captions. Therefore, we propose a novel visual and semantic enhanced video captioning network, named as VSVCap, that efficiently utilizes multiple ground truth captions. We aim to generate captions that are visually and semantically enhanced by exploiting both video and text modalities. To achieve this, we propose a fine-grained cross-graph attention mechanism that captures detailed graph embedding correspondence between visual graphs and textual knowledge graphs. We have performed intra- and inter-node level matching and structure level reasoning between the weighted regional graph and knowledge graph. The proposed network achieves promising results on three benchmark datasets: YouTube2Text, MSR-VTT, and VATEX. The experimental results show that our network accurately captures all key objects, relationships, and semantically enhanced events of a video to generate human annotation-like captions.

Introduction

Video captioning aims to automatically generate natural language descriptions for a video. It is highly utilized in real-world problems, such as video understanding (Schroff, Kalenichenko, and Philbin 2015; Zhang et al. 2020a), assisting visually impaired people (Mocanu, Tapu, and Zaharia 2019), and visual intelligence (Kojima, Tamura, and Fukunaga 2002) in robotics. Video captioning (VC) lies in the intersection of two main domains, *i.e.*, computer vision (CV) and natural language processing (NLP). Natural languages contain semantic information to better describe the visual content of an image or a video. Generating captions based on visual and spatio-temporal features is a challenging task as videos contain complex regional structural patterns and features that may convey multiple different summaries (Zhang et al. 2020b). Recent studies (Venugopalan et al. 2015; Donahue et al. 2015; Zhao, Li, and Lu 2019) on VC have shown that the approaches can effectively learn the generic representation using visual features. These generic captions are rewritten from the training set that lacks to generate visually and semantically enhanced captions.

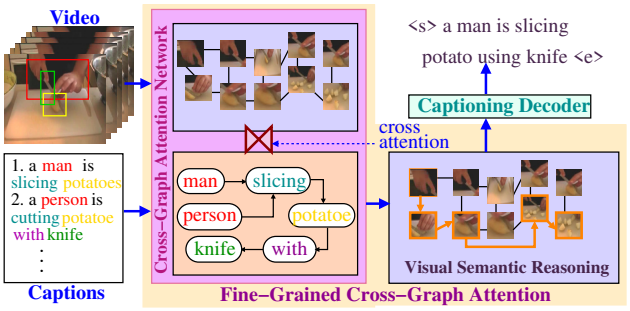


Figure 1: Cross-Graph attention focuses on node-level matching. This model proposes fine-grained cross-graph attention network that combines cross-graph attention with structural and semantic reasoning to generate human-like captions based on semantic information. The global visual data capture key objects (man, potato, knife) and facts ($< man, slicing, potato >$) to generate captions.

To improve the captioning task, the attentive models (Zhao, Li, and Lu 2019; Gao et al. 2017; Yan et al. 2019) are explored over the region of interest (ROI). However, these models ignore the saliency information present in a video since the visual regions are often ambiguous and over-sampled, which makes the captioning task more challenging. Even some works (Lin, Gan, and Wang 2020; Liu et al. 2019) have focused on the ROI and visual grounding by decomposing expressions into logic structures. Yet, these methods are not able to preserve the semantic structure of salient regions to generate human-like descriptions. Therefore, it is better to leverage the semantic structures by using a textual knowledge graph since 'text' is the defacto source, laden with 'knowledge'. On the other hand, a graph helps to structure the knowledge in a better way, where nodes represent entities and edges denote relationships between nodes (Nickel et al. 2015). A knowledge graph can structure the information and highlight the semantic facts present in an image or video as 'SPO' triples ($< subject, predicate, object >$). These structured graphs,

known as semantic knowledge graphs (or textual knowledge graph), can effectively organize human thoughts and clarify their intended meaning.

Motivated by the aforementioned discussion, we propose a Visual and Semantic enhanced Video Captioning network, named as VSVCap. Our main objective is to design a model that can identify relevant semantic information from a video and decode it into a corresponding description. This paper proposes a fine-grained cross-graph attention mechanism that leverage weighted graphs for region-word alignment & reasoning. We have utilized the human-annotated facts that contain semantic information of a video. However, the annotated facts contain redundant information to describe the video. Therefore, we have proposed a weighted knowledge graph that helps to filter the multiple descriptions based on their weights. In contrast, visual information is complex since it contains multiple objects and relationships. We generate a weighted visual graph that efficiently structures the regional information. However, the weighted visual graph is dense, where each entity is connected to every other entity. These connections describe relations between two or more entities in which some relations are relevant while some are irrelevant, which creates overhead. To capture the salient regions and structured information of the two modalities, we propose a fine-grained cross-graph attention mechanism, as shown in Figure 1. Further, inspired by the performance of bottom-up attention (Anderson et al. 2018), we use a captioning decoder for generating visual and semantic enhanced captions. In summary, our contributions are listed as follows:

• We propose a novel visual and semantic enhanced video captioning network that encodes the video into a weighted graph and corresponding textual graph of the available descriptions. We leverage both modalities to decode the word based on the graph attentive features to generate human description-like captions.

• We propose a fine-grained cross-graph attention mechanism that helps to bridge the gap between visual and semantic representations by preserving their intra- and inter-node correspondence, followed by structure-level reasoning that preserves the internal structure of the graphs. Moreover, we employ visual semantic reasoning to obtain the global contextual information.

• We generate a semantic knowledge graph that utilizes multiple grounded captions of a video to generate a semantic graph based on facts present in human annotations. To efficiently use this knowledge graph, we prioritize the relevancy of each relation between entities by assigning the weights.

• The experimental results of three benchmark datasets, i.e., YouTube2Text, MSR-VTT, and VATEX, demonstrate that the proposed method focuses on the key semantics of a video and preserves the grounded facts lie in a video using the semantic knowledge graph that helps to correctly pair the objects based on their relationship.

Related Work

• **Video Captioning.** In the early stage, template-based approaches (Guadarrama et al. 2013) were widely studied, which first detect sentence templates with grammar rules and align them with video content. However, these methods have fixed syntactic structures. With the deep neural networks (DNNs), the sequence learning methods (Yao et al. 2015; Pan et al. 2017; Rohrbach et al. 2013; Venugopalan et al. 2015; Pan et al. 2016; Baraldi, Grana, and Cucchiara 2017) that use encoder-decoder framework has become popular, which utilized global features to generate captions. These are improved by attention-based methods (Zhao, Li, and Lu 2019; Xiao and Shi 2020; Gao et al. 2017; Yan et al. 2019; Xu et al. 2018; Deng et al. 2022; Zheng et al. 2022; Deng et al. 2022) that highlight the relevant regions but lack to utilize structured semantic features. SAVCSS (Chen et al. 2020a) proposed semantic-based captions by incorporating POS features, whereas DSD-3DS-SEM (Shekhar et al. 2020) suggested semantic-guided approach using domain data. POS-based approaches (Wang et al. 2019a; Hou et al. 2019; Zheng, Wang, and Tao 2020) utilized part-of-speech to generate grammatically correct captions, but these syntactically correct captions are not semantically aligned. The authors designed objects and graph-based models in (Junchao and Peng 2019; Aafaq et al. 2019a; Zhang et al. 2020b; Chen and Jiang 2021; Li et al. 2022b) that generate graphs using either object or grid-based regions. But, none of them leverage semantic KG for better quality features.

• **Graph-based Cross-Attention.** The main challenge in graph-based attention is to map visual and semantic nodes or align them while preserving their internal structure. In DNNs, authors in (Zanfir and Sminchisescu 2018) suggested graph matching under node and pairwise constraints. To match the visual and textual graph, method (Xiong et al. 2019) employed a framework for matching between movie segments and synopsis paragraphs. The authors in (Liu et al. 2020) proposed a graph-structured network for an image to text matching. The networks (Li et al. 2019; Jiang et al. 2022) used cross-graph attention for graph matching to compute the similarity between nodes. The method (Xie et al. 2020) utilized contextual alignment enhanced cross-graph attention network for cross-lingual entity alignment by jointly learning the embedding of a different knowledge graph. The network (He et al. 2021) employed a cross-graph attention model to learn modality-specific shared semantic features for fine-grain image-text retrieval.

The aforediscussed methods use cross-graph attention for node-level matching. However, only one-to-one mapping constraint is explored and fails to work for many-to-one or one-to-many mapping. Motivated by this, we propose a fine-grained cross-graph attention mechanism that performs intra- and inter-node correspondence that can efficiently handle many-to-one or one-to-one mapping constraints.

Proposed Methodology

We propose Visual and Semantic enhanced Video Captioning (VSVCap) network that comprises of four modules; (a) **Global Encoders** takes a video as an input and generate frame and segment representations; (b) **Visual and Semantic Graph Construction** takes the visual region and textual annotations as input and generate the respective weighted graphs; (c) **Fine-Grained Cross-Graph Attention** filters out the intra- and inter-graph similar nodes and preserve the saliency information of both modalities; and (d) **Captioning**

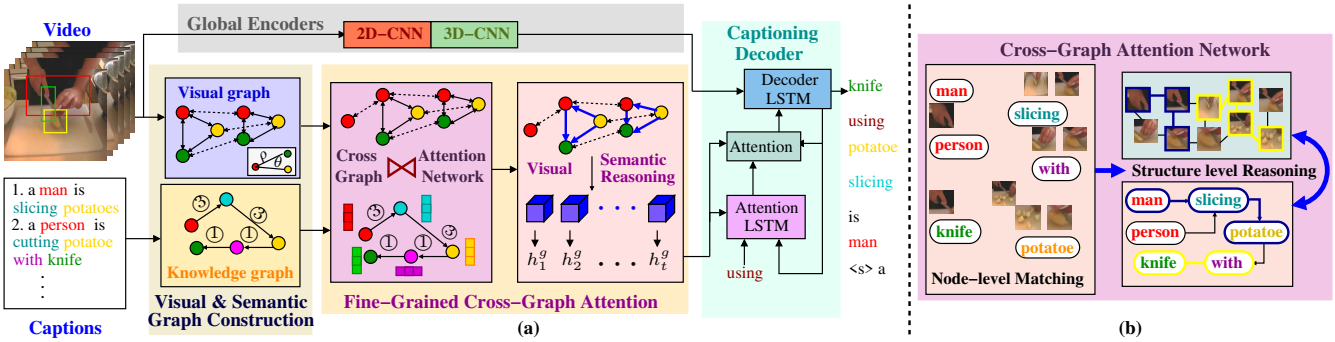


Figure 2: (a) Overview of the proposed network. Encode the appearance (2D-CNN), motion (3D-CNN) as global features, and region features as local information. We use global encoder to generate global features. Then, visual and semantic graph construction generates visual regional graph of video and textual knowledge graph using facts in annotations. We propose a fine-grained cross-graph attention to perform mapping between the intra- and inter-graph nodes followed by visual semantic reasoning to obtain the overall contextual representation, which is fed to the captioning decoder for caption generation. (b) Visualization of the fine-grained cross-graph attention using node-level matching and structure-level reasoning.

Decoder exploits the attentive features to predict the next word of the decoded caption; as shown in Figure 2.

• **Global Encoders.** Given a video V as input, T frames $\{f_i\}_{i=1}^{N}$ and clips $\{c_i\}_{i=1}^{N}$ are sampled uniformly. We extract the appearance and motion features represented by $v_a \in \mathbb{R}^{T\times\zeta}$ and $v_m \in \mathbb{R}^{T\times\zeta}$, respectively. T is the video temporal length and ζ is the feature dimension. The features are extracted from the pre-trained 2D-CNN Ψ^a and 3D-CNN Ψ^m. The global representation is the concatenation of two visual features as follows:

$$F = \mathsf{W}_v[v_a; v_m], \tag{1}$$

where $v_a = \Psi^a(f), v_m = \Psi^m(c)$. W_v is the trainable embedding and $[\cdot;\cdot]$ concatenation operation.

• **Visual and Semantic Graph Construction.** Networks (Pan et al. 2020; Venugopalan et al. 2015) use feature-based models to learn appropriate features from a video. However, these strategies are not able to capture the internal structure of the video regions and textual annotations. Therefore, we use a graph-based approach that provides flexible ways to capture middle-level elements and their relationships. In this module, our objective is to construct graph-based representations for video and text annotations to explore their internal structures and key components.

○○ **Weighted Semantic Knowledge Graph.** Textual modality is an important cue for the applications that use cross-modality features as it contains significant semantic information in the form of entities, attributes, and their relationship. In (Ali et al. 2022; Chen et al. 2020c), only natural language (NL) annotations are utilized to directly encode the text. It is semantically correct, but ignores the internal structure of the sentence. Therefore, we substitute NL annotations with knowledge graph to preserve the semantics and structure of information from unstructured data (text). In the proposed KG, we make directed and weighted KG that uses transitivity property for graph construction to preserve SPO mapping. We use available multi-captions per video (Mahon et al. 2020) as redundant information to assign the edge weights (discussed in supplementary). Formally, we construct a weighted textual (semantic) knowledge graph $G^w = (V^w, \mathfrak{E}^w, A^w)$ for each text, where V^w is a set of nodes and $\mathfrak{E}^w$ is set of edges in a graph. Let $E = \{e_1, \cdots, e_{N_e}\}$ be the set of all entities and $\mathcal{R} = \{r_1, \cdots, r_{N_r}\}$ is set of all relationships in a KG. There are N_e entities and N_r types of relationships extract from a graph. Let e_i is the subject node, e_j is the object node and r_k is relationship type. We model each possible triples $x(i,j,k) = (e_i, r_k, e_j)$ that has a random variable $\omega \in \mathbb{R}^n$. For simplicity, all possible triples in $E \times \mathcal{R} \times E$ can be grouped in an adjacency matrix $A^w \in \mathbb{R}^{N_y \times N_y}$, where $N_y = N_e + N_r + N_e$ and entries are:

$$A^w(i,j) = \begin{cases} \omega & \text{if the relation } (e_i, r_j) \text{ exists,} \\ \omega & \text{if the relation } (r_i, e_j) \text{ exists,} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where ω denotes the frequency of co-occurrence of the SPO pairs in given captions. We filter and prioritize the semantic graph based on the edge weights. The edge weight is denoted by an adjacency matrix A^w, which shows the frequency of similar pairwise nodes. We use rule-based parser (Qi et al. 2019) to parse the sentences. This not only generates a dependency parse of a sentence that identifies the subject (object), predicate (relation), and quantifier (attribute) in a sentence but also preserve semantic dependencies that hold true between them. Next, we generate a region graph from each video for semantically enhanced visual graph.

○○ **Weighted Visual graph.** In this module, our aim is to generate a weighted visual graph for a video. We take N frames from each video and extract object features $v_o \in \mathbb{R}^{T\times N_o\times\zeta}$ derived from Faster-RCNN Ψ^o on each frame. We use these object features to construct the visual graph $G^r = (V^r, \mathfrak{E}^r, A^r)$, where $V^r = \{1, \cdots, o\}$ is the set of objects, $\mathfrak{E}^r$ is the set of edges lies between objects, and A^r is the adjacency matrix. $|V|$ and $|E|$ represent the number of objects and edges in a graph. To efficiently capture the relationship between the entities, we generate a weighted visual graph, where we represent each frame as a fully connected dense graph. We compute edge-weights of adjacency matrix A^r using polar coordinate based on the bounding box (x_1, y_1, x_2, y_2) of pair-wise regions, as mentioned

in (Norcliffe-Brown, Vafeias, and Parisot 2018; Liu et al. 2020). We use geodesic polar coordinates (Monti et al. 2017) to model the internal structure of each frame, which disentangle the distance and orientation of each pair-wise region. The polar coordinates are given by:

$$(w, h) = (x_2 - x_1\ ,\ y_2 - y_1), \tag{3}$$

$$(x_c, y_c) = (x_1 + \gamma_1 * w\ ,\ y_1 + \gamma_2 * h), \tag{4}$$

$$\theta = \mathbf{tan}^{-1}(y_c x_c), \quad \rho = \sqrt{x_c^2 + y_c^2}, \quad A^r = \rho \,||\, \theta, \tag{5}$$

where $||$ is the stacking operation. $\{w, h, x_c, y_c\}$ denote width, height, and center coordinates of the bounding boxes. Further, we process the polar coordinates using gaussian kernels that have a learnable mean and variance. These visual and knowledge graphs extract semantic information in the form of relationships from both video and the available annotations. We fed these obtained multimodal graphs in a fine-grained cross-graph attention module to efficiently leverage the visual and textual information.

• **Fine-grained Cross-Graph Attention.** We observe that the object interaction is often described in the text in more efficiently manner. Therefore, in this module, our objective is to highlight the video regions by leveraging the available textual saliency features. Previous methods (Ali et al. 2022; Liu et al. 2020; Li et al. 2022a) utilize graph matching technique to incorporate the semantic information for the retrieval task. However, these methods perform node-level matching that have one-to-one constraint for nodes and may fails to capture one-to-many or many-to-one node correspondence. Therefore, we propose a fine-grained cross-graph attention mechanism that utilizes intra- and inter-modality graph-based attention. It can handle one-to-many, many-to-one, and one-to-one mapping constraints. In short, we perform the fine-grained graph embedding correspondence between the region graph and the semantic knowledge graph. For fine-grained embedding, we perform a node-level matching followed by structure-level reasoning. Given a weighted semantic knowledge graph $G^w = (V^w, \mathfrak{E}^w, A^w)$ and a weighted visual graph $G^r = (V^r, \mathfrak{E}^r, A^r)$ as inputs. Our aim is to generate a semantic saliency enhanced graph as an output. The weights in the graph define the priority or relevancy of the nodes and its connected edges.

○○ **Node-level Matching (NLM).** In this matching, we corresponds the nodes of weighted textual KG and weighted visual graph. In order to efficiently employ the textual KG, we initially perform the intra-node correspondence. Since, the nodes are semantically dependent and sometimes they belong to same entity or objects, such as chipmunk, rabbit, and squirrel denote the same class. Therefore, we compute the similarity between the intra-node representations of V^w, which is given by:

$$s_{ij} = \frac{\mathbf{exp}(\phi(e_i)\phi(e_j)^\top)}{\sum_{j=0}^{N_y} \mathbf{exp}(\phi(e_i)\phi(e_j)^\top)}, \tag{6}$$

where $\phi(\cdot)$ is the embedding layer and e is the node of the knowledge graph. N_y is the number of nodes. s_{ij} is the similarity between the $i-th$ node and $j-th$ node. Finally, we calculate weighted adjacency matrix by:

$$A_e^w = ||S^w \circ A^w||_2, \tag{7}$$

where $S^w \in \mathbb{R}^{N_y \times N_y}$ is the similarity matrix and $A^w \in \mathbb{R}^{N_y \times N_y}$ denotes sparse adjacency matrix. $\circ$ denote Hadamard product. In contrast, to find the similarity between inter-graph, we propose a cross-graph attention mechanism. The node representation of semantic knowledge graph is defined as $V^w \in \mathbb{R}^{N_y \times d}$, and the node representation of video visual graph as $V^r \in \mathbb{R}^{N_o \times d}$. Here, N_y, N_o and d denote the nodes of knowledge graph, visual graph, and feature dimension, respectively. The matching is performed by calculating the similarity between the nodes using dot product operation, which is given by:

$$\mathcal{P}^w = \left[\frac{\mathbf{exp}(V_i^r (V_j^w)^\top)}{\sum_{j=0}^{N_y} \mathbf{exp}(V_i^r (V_j^w)^\top)}\right] V_j^w, \tag{8}$$

$$\mathcal{P}^r = \left[\frac{\mathbf{exp}(V_i^w (V_j^r)^\top)}{\sum_{j=0}^{N_o} \mathbf{exp}(V_i^w (V_j^r)^\top)}\right] V_j^r. \tag{9}$$

We extract an updated visual graph from the heterogeneous graph using node-level and structure-level matching with associate nodes and structure-level correspondence by propagating associated nodes to neighbors.

○○ **Structure-level reasoning (SLR).** Node level correspondence may focus on semantically similar nodes present in a graphs but not able to preserve the relationship between the nodes which are necessary for connecting the components of a video and the given text. Therefore, we utilize structure-level reasoning to incorporate the relationships present in textual semantics over the visual graphs. Each node is connected to its correlated neighbor nodes. To calculate structure-level reasoning, we take node level matching vectors as input and propagate these vectors to their neighbors along with its graph edges. The matching vector of each node is updated by integrating neighborhood matching vectors using GCN. We use K kernels to integrate the neighborhood matching vectors. The convolved features $\dot{a}_i$ are obtained using GCN layer, which is calculated by:

$$\dot{a}_i^w = \left[\sum_{j \in \mathcal{N}_i} A_e^w W_1 \mathcal{P}_j^w; \cdots; \sum_{j \in \mathcal{N}_i} A_e^w W_K \mathcal{P}_j^w\right], \tag{10}$$

$$\dot{a}_i^r = \left[\sum_{j \in \mathcal{N}_i} A^r W_1 \mathcal{P}_j^r; \cdots; \sum_{j \in \mathcal{N}_i} A^r W_K \mathcal{P}_j^r\right], \tag{11}$$

where $\{A_e^w, A^r\}$ is the weighted adjacency matrix of knowledge graph and visual graph. $\mathcal{N}_i$ is the neighborhood of $i-th$ node. W_k are learnable parameters for $K-th$ kernel. The GCN layer reflects the correspondence between the connected nodes that preserve the localized structure. The convolved vectors are fed into the fully connected layers. We further concatenate the convolved vectors to jointly consider the fine-grained graph correspondence $\mathcal{C}$, which is given by:

$$\mathcal{C} = [\mathbf{ReLU}(\mathsf{W}_{hr} \dot{a}_i^r); \mathbf{ReLU}(\mathsf{W}_{hw} \dot{a}_i^w)], \tag{12}$$

where $\{\mathsf{W}_{hr}, \mathsf{W}_{hw}\}$ are the learnable parameters. We pass the concatenated graph for global contextual reasoning in the visual semantic reasoning (VSR).

○○ **Visual Semantic Reasoning (VSR).** This module aims to perform the global reasoning to select the discriminative

saliency information based on the visually and semantically structured information. It also helps to filter out the irrelevant information from the visual region to obtain the final representation from all semantically enhanced region information. We have incorporated the sequence of enhanced semantic and visual regions $\mathcal{C} = \{c_1, \cdots, c_r\}$ into LSTM for the global semantic reasoning. LSTM helps to select information based on the current region data to update its hidden state in the following way:

$$h_t^g, c_t^g = \mathbf{LSTM_g}(c_t), \tag{13}$$

where $\{h_t^g, c_t^g\}$ are related to the hidden and cell states of the global semantic reasoning LSTM. We fed the global reasoning data to the decoder module for generating captions.

• **Captioning Decoder.** Our final step is to generate visual and semantic enhanced captions in word-by-word manner based on the video representations. LSTM has a capability of modeling the long-term dependencies. Therefore, we use LSTMs to decode video representations into a caption. We use the global and semantically enhanced region graph-based features to generate the video descriptions using bottom-up attention mechanism (Anderson et al. 2018). For each timestep t, the concatenation of word embedding $\mathbb{E}(y_t)$ and the global reasoning data (h_t^g) are passed to the attention LSTM. We calculate the word embedding using embedding matrix $\mathbb{E}$ and the $t-th$ word in a caption, which is represented by y_t.

$$h_t^{att}, c_t^{att} = \mathbf{LSTM_{att}}([\mathbb{E}(y_t); h_t^g; h_{t-1}^d]), \tag{14}$$

$$a_{i,t} = \mathsf{W}_a^\top \, \mathbf{tanh}(\mathsf{W}_{va} f_i + \mathsf{W}_{vh} h_t^{att}), \tag{15}$$

$$\alpha_t = \mathbf{softmax}(a_{i,t}), \quad \widehat{f}_t = \sum_{i=1}^{k} \alpha_{i,t} f_i, \tag{16}$$

$$h_t^d, c_t^d = \mathbf{LSTM_d}([\widehat{f}_t; h_t^{att}; h_t^g]), \tag{17}$$

where $\{h_t^{att}, c_t^{att}, h_t^d, c_t^d\}$ are the hidden and cell state of attention and decoder LSTM. The appearance and motion features $\{v_a, v_m\}$ are used to attend the saliency representations based on the previous data. f_i is the $i-th$ element of the global vector F (using Eq. (1)). $\{\mathbf{softmax}(\cdot), \mathbf{tanh}(\cdot)\}$ denote softmax and tanh activation. We fed the attended features with global contextual reasoning (h_t^g) to decoder LSTM for the caption generation. Further, we decode the obtained hidden state h_t^d to generate word probability $\mathbf{p}$:

$$\mathbf{p}(y_t|y_{1:t-1}) = \mathbf{softmax}(\mathsf{W}_p h_t^d + F + \mathsf{b}_p), \tag{18}$$

where $\{\mathsf{W}_p, \mathsf{b}_p\}$ are the learnable parameters. We predict y_t i.e., $t-th$ word in presence of $y_{1:t-1}$ previous words.

• **Loss Functions.** We train our model in a end-to-end manner by learning the captioning decoder. Given a video V and its corresponding annotations $\mathcal{Y}$ of length L, we compute the overall loss for $\mathcal{N}$ samples. Our main objective is to optimize the captioning loss by utilizing cross-entropy. The loss function is formulated by:

$$\mathbb{L}_{cap} = -\sum_{i=1}^{\mathcal{N}} \mathbf{log}\, \mathbb{P}(\mathcal{Y}^i|V^i) + ||\theta||_2^2, \tag{19}$$

$$\mathbb{P}(\mathcal{Y}|V) = \prod_{t=1}^{L} p(y_t|y_{1:t-1}, V). \tag{20}$$

We use L2 regularization as a penalty for DNN learnable parameters to prevent our network from overfitting. L2 forces DNN parameters toward taking relatively small values.

Experimental Results

• **Datasets and Evaluation Metrics.** We briefly discuss the popular datasets and evaluation metrics for VC task. **Microsoft Video description corpus** (YouTube2Text) (Guadarrama et al. 2013) is a benchmark VC dataset which contains 1970 short YouTube clips. There are about 8000 clips description pairs and each video has approximate 10-20 descriptions. We split the data in 1200 clips for training, 100 clips for validation, and 670 clips for testing (Rohrbach et al. 2013). **Microsoft research-video to text** (MSR-VTT) (Xu et al. 2016) is large-scale VC dataset, which consists of 10,000 videos. Each video clip is of average 10-30 seconds duration and has approximately 20 annotation descriptions. It splits into 6513 training, 497 validation, and 2990 testing videos. **VATEX** (Wang et al. 2019b) is a multilingual video description dataset that contains 41,250 videos and 825,000 captions in both Chinese and English. We only use English corpora and follow the specified split: 25,991 videos for training, 3,000 videos for validation, and 6,000 testing videos.

We choose the following popular evaluation metrics: CIDEr (C) (Vedantam, Lawrence Zitnick, and Parikh 2015), BLEU@4 (B4) (Papineni et al. 2002), METEOR (M) (Denkowski and Lavie 2014), and ROUGE-L (R) (Lin 2004). These metrics are highly correlated with human perception. We adopt Microsoft COCO evaluation tool [1] to implement these metrics. All results are calculated in $\%$ and '-' indicates that the authors do not mention the result.

• **Implementation Details.** In this section, we discuss the details of the preprocessing, training and inference.

○○ **Features and Word Processing.** We utilize Inception-ResNetV2 pretrained on ILSVRC-2012 dataset as 2D-CNN features to extract appearance information from 26 frames each of $299 \times 299 \times 3$ spatial size. For 3D-CNN, we use I3D (Carreira and Zisserman 2017) pretrained on Kinetics dataset (Kay et al. 2017) to extract spatio-temporal features from a video. In addition, we adopt Faster-RCNN (Ren et al. 2015) to extract 36 region features for each frame. The dataset sentences are preprocessed by removing punctuation, converting all captions to lowercase, and then tokenize it. We fix the length of sentence as 26 words, where we truncate the words if size is greater than 26; otherwise, pad zeros. We also add 'bos' and 'eos' tokens at the beginning and end of each caption, respectively. We use 'GloVe' embedding (Pennington, Socher, and Manning 2014) to initialize word embedding representation.

○○ **Training Details.** The model is optimized by Adam (Kingma and Ba 2015) with a learning rate of 10^{-4} which is divided by 10 on every 10 epochs. The exponential decay rate of first and second moment estimate was set to β_1 = 0.9 and β_2 = 0.999. In our experiments, we consider $\gamma_1 = 0.5$ and $\gamma_2 = 0.5$. The dimensionality of word embedding for each word in a sentence is initialized with 300 di-

[1] https://github.com/tylin/coco-caption

Models	YouTube2Text				MSR-VTT			
	B4	M	R	C	B4	M	R	C
Base	42.7	24.1	57.0	62.3	30.6	26.5	58.9	41.6
Base+GC	57.3	36.1	72.8	83.2	42.9	28.0	60.7	48.5
Base+GC+ FCGA	58.6	37.5	74.6	90.0	44.3	28.6	62.8	50.2
Base+GC+ FCGA+VR	**60.9**	**38.2**	**75.3**	**92.11**	**46.1**	**29.9**	**63.8**	**53.7**

Table 1: Effect of different modules on YouTube2Text and MSR-VTT datasets.

Models	YouTube2Text				MSR-VTT			
	B4	M	R	C	B4	M	R	C
UD+UW	52.1	31.2	70.9	85.3	43.6	27.9	59.3	48.6
UD+W	58.3	36.3	72.0	90.6	44.1	28.9	61.0	50.7
D+UW	55.1	35.0	71.9	89.3	44.2	27.5	61.4	49.9
D+W	**60.9**	**38.2**	**75.3**	**92.11**	**46.1**	**29.9**	**63.8**	**53.7**

Table 2: Analysis of different semantic graph paradigms on YouTube2Text and MSR-VTT datasets.

mension using pretrained 'GloVe' (Pennington, Socher, and Manning 2014). We set the hidden size of the LSTM as 512. We train our network for 30 epochs. We utilize NVIDIA Tesla V100 GPU for training the model.

○○ **Inference Details.** We remove cross-graph attention module since the captions are unavailable during the inference time (Zhou et al. 2020). Therefore, we only forward the visual graph (VG) to visual-semantic reasoning (VSR) and captioning decoder. We observe that our method still achieves significant performance. We use beam search for final caption generation. The beam search size is set to 5, as proposed in (Wang et al. 2019a). The sentence generation will terminate when the 'eos' token generates.

• **Ablation Study.** In this section, we examine the efficacy of our architecture to demonstrate its effectiveness against a set of ablated models.

○○ **Effect of different modules.** We validate VSVCap modules on different variants: (a) Base: uses an encoder-decoder architecture in which we use 2D-, 3D-CNN, and object-detection, (b) Base+GC: adds the graph construction module with base architecture. (c) Base+GC+FCGA: adds both GC and fine-grained cross-graph attention (FCGA) to the base, (d) Base+GC+FCGA+VR: is proposed architecture that utilizes all the modules together(VR is Visual Semantic Reasoning). We examine the combination of proposed modules outperforms the variant methods, as shown in Table 1.

○○ **Analysis of Semantic Graph Paradigms.** Table. 2 demonstrate the comparison between different graph paradigms, such as, directed (D), undirected (UD), weighted (W), and unweighted (UW). Our experiment shows that the directed and weighted semantic graph outperforms other graph paradigms in fine-grained cross-graph attention. We also noted that UD graphs reduce the performance of the overall network due to wide information.

○○ **Number of kernels in GCN.** Table 3 conducts various experiments over the number of kernels required in GCN layer (Eq. 10). We conclude that 8 layers are adequate for extracting salient neighborhood features in our network.

No. of kernels	YouTube2Text				MSR-VTT			
	B4	M	R	C	B4	M	R	C
2	55.0	32.8	70.6	81.2	39.1	27.0	56.3	48.4
4	58.1	35.4	72.8	88.8	42.6	28.3	59.1	50.2
8	**60.9**	**38.2**	**75.3**	**92.11**	**46.1**	**29.9**	**63.8**	**53.7**
16	60.8	38.2	75.0	90.6	46.2	29.0	63.5	50.2

Table 3: Impact of different number of kernels in GCN layer on YouTube2Text and MSR-VTT datasets.

Method	Relevancy	Coherence
Seq2Seq (Yao et al. 2015)	0.5043	0.6938
POS-CG (Wang et al. 2019a)	0.7858	0.8053
VSVCap	**0.9095**	**0.8821**

Table 4: Results of Human Evaluation.

• **Human Evaluation.** We asked the human evaluators to rate the predicted captions on two aspects, *i.e.*, relevancy and coherence of a generated sentence, as suggested in (Aafaq et al. 2019b; Zhao, Wu, and Luo 2021). Relevance states how much generated caption is related to the given captions. Coherence scores on the basis of fluency, logic, and readability for the predicted captions. Each rating aspect is graded in a range of 0 to 1 (higher scores are better). We randomly choose 500 captions generated by SOTA methods from the testing set. Table 4 shows average scores.

• **Qualitative Analysis.** Figure 3 shows the qualitative results for VSVCap. Seq2Seq (Yao et al. 2015) generate monotonous captions that bear rewording of training corpus. Our captions comply with a visual and semantic reasoning graph and generate enhanced captions that contain key components and relationships present in a video. In the first sample of Figure 3, our model not only recognizes 'man', 'violin' and 'woman' but also accurately pairs the relationship between 'man' with 'playing violin' as well as 'woman' with 'singing'. Similarly, in the second sample, model accurately identifies 'man', 'chair', and 'phone' objects present in a video with correct relations of 'talking' and 'sitting'. Generated captions contain pivotal objects and correct relationship that generate better captions than Seq2Seq. It can be observed that our model not only accurately recognizes key semantics but also generate more accurate and semantically enhanced captions for all type videos.

• **Quantitative Analysis.** To demonstrate the effectiveness of the proposed approach, we show comparative results between the VSVCap and SOTA methods on YouTube2Text, MSR-VTT, and VATEX datasets, as shown in Table 5. We limit the feature extractors within IRV2, I3D, and object detection as much as possible for fair comparison. We use cross-entropy loss as training strategy for variants. It is observed that our model achieves the better performance across 'B4' and 'M' due to the fine-grained cross-graph attention mechanism. Our model outperforms traditional encoder-decoder method (Pan et al. 2016) that utilized global features. We use graph-based methods for the feature refine-

Methods	YouTube2Text				MSR-VTT				VATEX			
	B4	M	R	C	B4	M	R	C	B4	M	R	C
HRNE (Pan et al. 2016)	46.7	33.9	-	-	-	-	-	-				
RecNet (Wang et al. 2018)	52.3	34.1	69.8	80.3	39.1	26.6	59.3	42.7	-	-	-	-
DS-RNN (Xu et al. 2018)	53.0	34.7	65.9	79.4	42.3	29.4	62.3	46.1	-	-	-	-
CAM-RNN (Zhao, Li, and Lu 2019)	42.4	33.4	-	54.3	36.2	27.9	-	38.8	-	-	-	-
POS-CG (Wang et al. 2019a)	52.5	34.1	72.1	88.7	42.0	28.2	61.6	48.7	-	-	-	-
OA-BTG (Junchao and Peng 2019)	56.9	36.2	-	90.6	41.4	28.2	-	46.9	-	-	-	-
GRU-EVE (Aafaq et al. 2019a)	47.9	35.0	71.5	78.1	38.3	28.4	60.7	48.1	-	-	-	-
POS+VCT (Hou et al. 2019)	52.8	36.1	71.8	87.8	42.3	29.7	62.8	49.1	-	-	-	-
SAAT (Zheng, Wang, and Tao 2020)	46.5	33.5	69.4	81.0	39.9	27.7	61.2	51.0	-	-	-	-
TDAM (Xiao and Shi 2020)	54.0	36.1	72.3	85.8	44.7	28.9	62.3	48.9	-	-	-	-
PMI-CAP (Chen et al. 2020b)	54.6	36.4	-	95.11	42.1	28.7	-	49.45	-	-	-	-
ORG-TRL (Zhang et al. 2020b)	54.3	36.4	73.9	**95.2**	43.6	28.8	62.1	50.9	32.1	22.2	48.9	49.7
SibNet (Liu, Ren, and Yuan 2020)	54.2	34.8	71.7	88.2	40.9	27.5	60.2	47.5	-	-	-	-
DSD-3DS SEM (Shekhar et al. 2020)	50.1	34.7	73.1	76.0	45.2	**29.9**	**64.2**	51.1	-	-	-	-
SMAN (Zheng et al. 2022)	52.5	35.0	72.4	94.5	41.3	28.7	62.1	**53.8**	-	-	-	-
SGHAN (Deng et al. 2022)	54.3	35.3	72.2	91.3	39.7	28.3	60.4	49.0	-	-	-	-
VSVCap	**60.9**	**38.2**	**75.3**	92.11	**46.1**	**29.9**	63.8	53.7	**39.6**	**26.5**	**50.3**	**70.2**

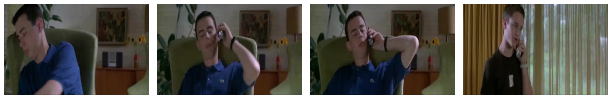
Table 5: Comparison between the proposed model and state-of-art methods on YouTube2Text, MSR-VTT, and VATEX datasets.

GT:{'a man is playing a violin on a stage', 'the man played the violin while the lady sang' }

Objects: man, violin, woman Verb: playing, singing

Seq2Seq: 'a man is playing violin'

VSVCap:'a man is playing violin and woman is singing on stage'



GT: {'two boys are talking on the phone', 'a young man is sitting in a chair talking on the phone' }

Objects: phone, man, chair Verb: talking, sitting

Seq2Seq: 'a man is talking on phone'

VSVCap: 'a boy is sitting on chair and talking on phone'

Figure 3: Qualitative results of the VSVCap. The first row shows the extracted frame. Objects and Verb shows the respective object and verbs present in a video. GT denote the ground-truth captions.

ment and find that the use of knowledge graph in captioning significantly improves the performance compared to graph-based method (Junchao and Peng 2019). Network (Zhang et al. 2020b) uses additional teacher recommended and reinforcement learning concepts, but our methods achieve comparable results without any additional learning. It shows the effectiveness of our method on YouTube2Text dataset. We also compare VSVCap with attention-based VC models (Xu et al. 2018; Zhao, Li, and Lu 2019; Xiao and Shi 2020; Zheng et al. 2022; Xiao and Shi 2020) and found that our fine-grained cross-graph attention achieves comparable results. Our method further achieves SOTA performance on B4 and C metrics in comparison with POS-based methods (Zheng, Wang, and Tao 2020; Hou et al. 2019; Wang et al. 2019a; Deng et al. 2022) and multimodal approaches (Chen et al. 2020b; Aafaq et al. 2019a) due to semantic knowledge incorporation and structure-level reasoning in FCGA. DSD-3DS-SEM (Shekhar et al. 2020) shows comparable performance on MSR-VTT dataset on M and R metrics due to the usage of domain specific knowledge. Also, our network shows SOTA performance on VATEX dataset.

Conclusion

This paper introduces a novel video captioning approach to generate human-like captions based on visual and semantic information. It uses a graph construction module to generate weighted visual regions and semantic knowledge graphs for a given video and its available captions. We matched these graphs to get the highlighted visual graphs that reflect the importance of each relation between nodes based on their weights. The obtained region graph helps to decompose the complex visual graph into a global reasoning graph that contains important semantic video components and structured them based on their relationship. The visual semantic reasoning graph is finally passed to captioning decoder to decode the graph and generate visually and semantically enhanced captions. The proposed network is superior to existing methods in generating highly appropriate captions with more grounded and structured information.

Acknowledgements

We would like to express our sincere gratitude to SERB MTR/2021/604 for supporting this research work.

References

Aafaq, N.; Akhtar, N.; Liu, W.; Gilani, S. Z.; and Mian, A. 2019a. Spatio-temporal dynamics and semantic attribute en-

riched visual encoding for video captioning. In *Proc. of CVPR*, 12487–12496.

Aafaq, N.; Mian, A.; Liu, W.; Gilani, S. Z.; and Shah, M. 2019b. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, 52(6): 1–37.

Ali, A.; Schwartz, I.; Hazan, T.; and Wolf, L. 2022. Video and Text Matching with Conditioned Embeddings. In *Proc. of CVPR*, 1565–1574.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. of CVPR*, 6077–6086.

Baraldi, L.; Grana, C.; and Cucchiara, R. 2017. Hierarchical boundary-aware neural encoder for video captioning. In *Proc. of CVPR*, 1657–1666.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. of CVPR*, 6299–6308.

Chen, H.; Lin, K.; Maye, A.; Li, J.; and Hu, X. 2020a. A Semantics-Assisted Video Captioning Model Trained with Scheduled Sampling. *Frontiers in Robotics and AI*, 7: 1–11.

Chen, S.; Jiang, W.; Liu, W.; and Jiang, Y.-G. 2020b. Learning modality interaction for temporal sentence localization and event captioning in videos. In *Proc. of ECCV*, 333–351. Springer.

Chen, S.; and Jiang, Y.-G. 2021. Motion guided region message passing for video captioning. In *Proc. of ICCV*, 1543–1552.

Chen, S.; Zhao, Y.; Jin, Q.; and Wu, Q. 2020c. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proc. of CVPR*, 10638–10647.

Deng, J.; Li, L.; Zhang, B.; Wang, S.; Zha, Z.; and Huang, Q. 2022. Syntax-guided hierarchical attention network for video captioning. *IEEE Trans. on CSVT*, 880–892.

Denkowski, M.; and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proc. of WMT*, 376–380.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. of CVPR*, 2625–2634.

Gao, L.; Guo, Z.; Zhang, H.; Xu, X.; and Shen, H. T. 2017. Video captioning with attention-based LSTM and semantic consistency. *IEEE Trans. on Multimedia*, 19(9): 2045–2055.

Guadarrama, S.; Krishnamoorthy, N.; Malkarnenkar, G.; Venugopalan, S.; Mooney, R.; Darrell, T.; and Saenko, K. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proc. of ICCV*, 2712–2719.

He, Y.; Liu, X.; Cheung, Y.-M.; Peng, S.-J.; Yi, J.; and Fan, W. 2021. Cross-Graph Attention Enhanced Multi-Modal Correlation Learning for Fine-Grained Image-Text Retrieval. In *Proc. of ACM*, 1865–1869.

Hou, J.; Wu, X.; Zhao, W.; Luo, J.; and Jia, Y. 2019. Joint syntax representation learning and visual cue translation for video captioning. In *Proc. of ICCV*, 8918–8927.

Jiang, Z.; Rahmani, H.; Angelov, P.; Black, S.; and Williams, B. M. 2022. Graph-Context Attention Networks for Size-Varied Deep Graph Matching. In *Proc. of CVPR*, 2343–2352.

Junchao, Z.; and Peng, Y. 2019. Object-aware aggregation with bidirectional temporal graph for video captioning. In *Proc. of CVPR*, 8327–8336.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv:1705.06950*, 1–22.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*, 1–15.

Kojima, A.; Tamura, T.; and Fukunaga, K. 2002. Natural language description of human activities from video images based on concept hierarchy of actions. *IJCV*, 50(2): 171–184.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2022a. Image-Text Embedding Learning via Visual and Textual Semantic Reasoning. *PAMI*, 1–14.

Li, L.; Gao, X.; Deng, J.; Tu, Y.; Zha, Z.-J.; and Huang, Q. 2022b. Long Short-Term Relation Transformer With Global Gating for Video Captioning. *Trans. on Image Processing*, 31: 2726–2738.

Li, Y.; Gu, C.; Dullien, T.; Vinyals, O.; and Kohli, P. 2019. Graph matching networks for learning the similarity of graph structured objects. In *Proc. of ICML*, 3835–3845.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lin, K.; Gan, Z.; and Wang, L. 2020. Semi-Supervised Learning for Video Captioning. In *Proc. of Conference on Empirical Methods in Natural Language Processing: Findings*, 1096–1106.

Liu, C.; Mao, Z.; Zhang, T.; Xie, H.; Wang, B.; and Zhang, Y. 2020. Graph structured network for image-text matching. In *Proc. of ICCV*, 10921–10930.

Liu, D.; Zhang, H.; Wu, F.; and Zha, Z.-J. 2019. Learning to assemble neural module tree networks for visual grounding. In *Proc. of ICCV*, 4673–4682.

Liu, S.; Ren, Z.; and Yuan, J. 2020. Sibnet: Sibling convolutional encoder for video captioning. *IEEE Trans. on PAMI*, 1–10.

Mahon, L.; Giunchiglia, E.; Li, B.; and Lukasiewicz, T. 2020. Knowledge graph extraction from videos. In *Proc. of ICMLA*, 25–32. IEEE.

Mocanu, B.; Tapu, R.; and Zaharia, T. 2019. Enhancing the accessibility of hearing impaired to video content through fully automatic dynamic captioning. In *Proc. of EHB*, 1–4. IEEE.

Monti, F.; Boscaini, D.; Masci, J.; Rodola, E.; Svoboda, J.; and Bronstein, M. M. 2017. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proc. of CVPR*, 5115–5124.

Nickel, M.; Murphy, K.; Tresp, V.; and Gabrilovich, E. 2015. A review of relational machine learning for knowledge graphs. *Proc. of the IEEE*, 104(1): 11–33.

Norcliffe-Brown, W.; Vafeias, S.; and Parisot, S. 2018. Learning conditioned graph structures for interpretable visual question answering. In *Proc. of NIPS*, 1–10.

Pan, B.; Cai, H.; Huang, D.-A.; Lee, K.-H.; Gaidon, A.; Adeli, E.; and Niebles, J. C. 2020. Spatio-Temporal Graph for Video Captioning with Knowledge Distillation. In *Proc. of CVPR*, 10870–10879.

Pan, P.; Xu, Z.; Yang, Y.; Wu, F.; and Zhuang, Y. 2016. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proc. of CVPR*, 1029–1038.

Pan, Y.; Yao, T.; Li, H.; and Mei, T. 2017. Video captioning with transferred semantic attributes. In *Proc. of CVPR*, 6504–6512.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*, 311–318.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP*, 1532–1543.

Qi, P.; Dozat, T.; Zhang, Y.; and Manning, C. D. 2019. Universal dependency parsing from scratch. *arXiv:1901.10457*, 1–10.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. of NIPS*, 1–14.

Rohrbach, M.; Qiu, W.; Titov, I.; Thater, S.; Pinkal, M.; and Schiele, B. 2013. Translating video content to natural language descriptions. In *Proc. of ICCV*, 433–440.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proc. of CVPR*, 815–823.

Shekhar, C. C.; et al. 2020. Domain-specific semantics guided approach to video captioning. In *Proc. of CVPR*, 1587–1596.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proc. of CVPR*, 4566–4575.

Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; and Saenko, K. 2015. Sequence to sequence-video to text. In *Proc. of ICCV*, 4534–4542.

Wang, B.; Ma, L.; Zhang, W.; Jiang, W.; Wang, J.; and Liu, W. 2019a. Controllable video captioning with pos sequence guidance based on gated fusion network. In *Proc. of ICCV*, 2641–2650.

Wang, B.; Ma, L.; Zhang, W.; and Liu, W. 2018. Reconstruction network for video captioning. In *Proc. of CVPR*, 7622–7631.

Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.-F.; and Wang, W. Y. 2019b. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proc. of ICCV*, 4581–4591.

Xiao, H.; and Shi, J. 2020. Video captioning with text-based dynamic attention and step-by-step learning. *Pattern Recognition*, 1–8.

Xie, Z.; Zhu, R.; Zhao, K.; Liu, J.; Zhou, G.; and Huang, X. 2020. A contextual alignment enhanced cross graph attention network for cross-lingual entity alignment. In *Proc. of ICCL*, 5918–5928.

Xiong, Y.; Huang, Q.; Guo, L.; Zhou, H.; Zhou, B.; and Lin, D. 2019. A graph-based framework to bridge movies and synopses. In *Proc. of CVPR*, 4592–4601.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proc. of CVPR*, 5288–5296.

Xu, N.; Liu, A.-A.; Wong, Y.; Zhang, Y.; Nie, W.; Su, Y.; and Kankanhalli, M. 2018. Dual-stream recurrent neural network for video captioning. *IEEE Trans. on CSVT*, 29(8): 2482–2493.

Yan, C.; Tu, Y.; Wang, X.; Zhang, Y.; Hao, X.; Zhang, Y.; and Dai, Q. 2019. Stat: spatial-temporal attention mechanism for video captioning. *IEEE Trans. on Multimedia*, 22(1): 229–241.

Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; and Courville, A. 2015. Describing videos by exploiting temporal structure. In *Proc. of CVPR*, 4507–4515.

Zanfir, A.; and Sminchisescu, C. 2018. Deep learning of graph matching. In *Proc of CVPR*, 2684–2693.

Zhang, W.; Wang, X. E.; Tang, S.; Shi, H.; Shi, H.; Xiao, J.; Zhuang, Y.; and Wang, W. Y. 2020a. Relational graph learning for grounded video description generation. In *Proc. of ACMM*, 3807–3828.

Zhang, Z.; Shi, Y.; Yuan, C.; Li, B.; Wang, P.; Hu, W.; and Zha, Z.-J. 2020b. Object Relational Graph with Teacher-Recommended Learning for Video Captioning. In *Proc. of CVPR*, 13278–13288.

Zhao, B.; Li, X.; and Lu, X. 2019. CAM-RNN: Co-attention model based RNN for video captioning. *IEEE Trans. on Image Processing*, 28(11): 5552–5565.

Zhao, W.; Wu, X.; and Luo, J. 2021. Multi-modal Dependency Tree for Video Captioning. *Advances in Neural Information Processing Systems*, 34: 1–12.

Zheng, Q.; Wang, C.; and Tao, D. 2020. Syntax-Aware Action Targeting for Video Captioning. In *Proc. of CVPR*, 13096–13105.

Zheng, Y.; Zhang, Y.; Feng, R.; Zhang, T.; and Fan, W. 2022. Stacked Multimodal Attention Network for Context-Aware Video Captioning. *IEEE Trans. on CSVT*, 32(1): 31–42.

Zhou, Y.; Wang, M.; Liu, D.; Hu, Z.; and Zhang, H. 2020. More grounded image captioning by distilling image-text matching model. In *Proc. of CVPR*, 4777–4786.