Two Heads are Better Than One: Hypergraph-Enhanced Graph Reasoning for Visual Event Ratiocination

Wenbo Zheng [1 2] **Lan Yan** [2 3] **Chao Gou** [4] **Fei-Yue Wang** [2]

Abstract

Even with a still image, humans can ratiocinate various visual cause-and-effect descriptions before, at present, and after, as well as beyond the given image. However, it is challenging for models to achieve such task–the visual event ratiocination, owing to the limitations of time and space. To this end, we propose a novel multi-modal model, **Hypergraph-Enhanced Graph Reasoning**. First it represents the contents from the same modality as a semantic graph and mines the intra-modality relationship, therefore breaking the limitations in the spatial domain. Then, we introduce the ***Graph Self-Attention Enhancement***. On the one hand, this enables semantic graph representations from different modalities to enhance each other and captures the inter-modality relationship along the line. On the other hand, it utilizes our built multi-modal hypergraphs in different moments to boost individual semantic graph representations, and breaks the limitations in the temporal domain. Our method illustrates the case of "two heads are better than one" in the sense that semantic graph representations with the help of the proposed enhancement mechanism are more robust than those without. Finally, we re-project these representations and leverage their outcomes to generate textual cause-and-effect descriptions. Experimental results show that our model achieves significantly higher performance in comparison with other state-of-the-arts.

[1]School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China [2]The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China [3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China [4]School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510275, China. Correspondence to: Fei-Yue Wang <feiyue.wang@ia.ac.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

1. Introduction

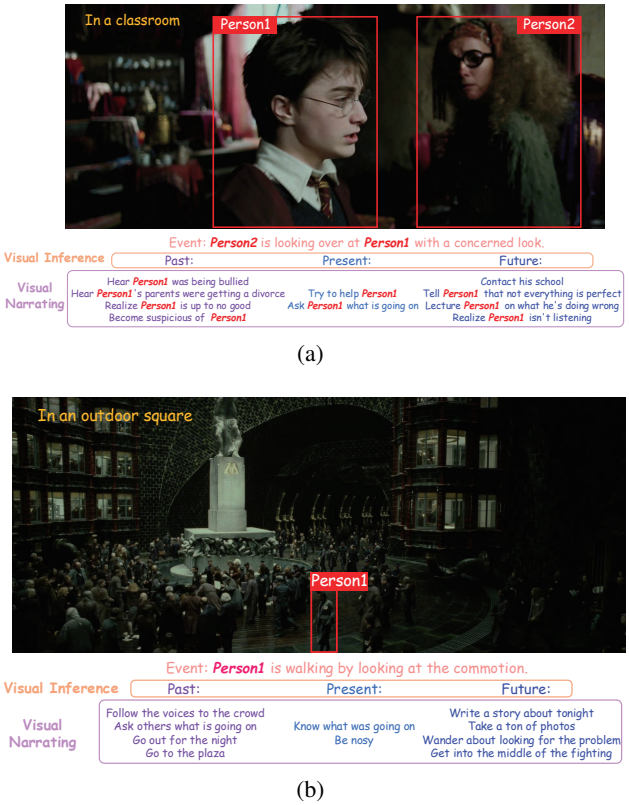

Figure 1. Illustration of The Visual Event Ratiocination. Given a person in the image, the model is required to reason about ① what needed to happen before, ② intents of the people at present, and ③ what will happen next. Our model has the excellent abilities of visual inference and visual narrating, detailed in Section 3.4 $\sim$ 3.5.

Visual Event Ratiocination is a novel challenging task about a combination of language and vision. Given an image and a description of the event in the image, it requires models to predict events that happen before/after and the present intents of the characters in the given image. In VisualCOMET benchmark dataset (Park et al., 2020), we show an example in Figure 1(a), given the image of the woman looking over at the man with a converned look in a classroom, the model can infer and generate/narrate three kinds of event ratiocination (a.k.a., event's cause-and-effect

descriptions): ① sometime in the past, she might unconsciously say many things to him, and when coming back to her senses, she might realize he looks panicked and become suspicious of him; ② sometime at present, she might ask he what is going on; ③ sometime in the future, she might realize he isn't listening. Besides, the event's cause-and-effect descriptions (e.g., 'ask what is going on') in the current moment of one image might be the same as these in the past time of one another image (shown in Figure 1(b)).

"A picture is worth a thousand words". The combination of two modalities (i.e., language and vision) is commonplace, so it is only natural to ask to what extent this combination may help machines understand the meaning. Some conventional tasks have been introduced for joint understanding from two modalities, e.g., visual question answering (Antol et al., 2015), referring expression reasoning (Liu et al., 2019). Different from these tasks, which only focus on visual recognition and inference about the current content of images, visual event ratiocination aims at reasoning the time-varying situation captured in the image, and pays attention to the model's abilities of both visual inference and visual narrating, as shown in Figure 1. Therefore, the study of this task has scientific significance: it opens the door of a significant leap from recognition-level understanding to cognitive-level reasoning.

From the above observation, it is obvious that there are three kinds of vital relationships here for generating event's cause-and-effect descriptions: ① relationships within the same modality, ② relationships between different modalities, and ③ spatio-temporal relationships among different images and their event ratiocination. Therefore, how to capture these three relationships from visual and semantic perspectives is essential for the task of visual event ratiocination.

Nevertheless, there has been little work on visual event ratiocination, while conventional visual-language tasks have been explored to a large extent. Park et al. (2020) proposed a dataset, VisualCOMET, which is the only one benchmark for this task at present. They also employ the Transformer (Vaswani et al., 2017) as a baseline model. Xing et al. (2021) propose the improved BART (Lewis et al., 2020) that incorporates textual information into the multi-modal model. Yet, these methods only look at conventional learning of visual and textual information while ignoring the link between modalities and among space-time. In short, current visual event ratiocination approaches have two main deficiencies:

① The existing models pay no attention to relationships from the same modality and different modalities.

② The current models ignore the spatio-temporal relationships from different samples with different moments.

To address the above two challenges, we propose a novel model, **Hypergraph-Enhanced Graph Reasoning**, to obtain a representation of the multi-modal contents in the task of visual event ratiocination. As shown in Figure 2, ① we construct semantic graphs for the same modality through translating intra-modality relationships from the spatial domain to the graph domain; furthermore, we propose an enhancement mechanism between these graphs to capture relationships between different modalities; ② we construct hypergraphs from different modalities with different moments to capture spatio-temporal relationships, and make these high-order semantic relations enhance the multi-modal graph representations by the proposed enhancement mechanism between graphs and the hypergraph.

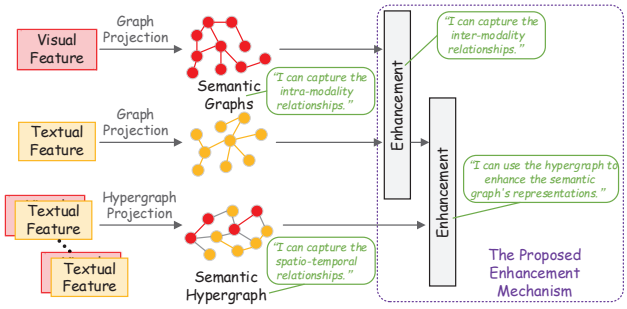

Figure 2. The Key Idea of Our Paper. The visual feature and the textual feature are projected into the semantic graphs that caputures the intra-modality relationships. The enhancement mechanism between these graphs can capture the inter-relationships. Similarly, different visual features and textual features are projected into the semantic hypergraph that caputures the spatio-temporal relationships. The enhancement mechanism between graphs and hypergraph can enhance the semantic graph's representations.

In summary, our main contributions are as follows:

☼ We propose a novel and effective hypergraph-enhanced graph reasoning model for visual event ratiocination. Our model captures intra- and inter- modality relationships as well as spatio-temporal relationships via learning in the graph domain. Experimental results show that our model has strong robustness and outperforms existing similar methods.

☼ We explore how to enhance two semantic graphs with each other as well as semantic graphs with hypergraphs, and propose a novel graph self-attention enhancement. The qualitative experiment shows that this mechanism is effective.

☼ Our hypergraph-enhanced graph reasoning model has the outstanding abilities of visual inference and visual narrating. The qualitative discussion reveals that our model achieves better performance than other state-of-the-art approaches on the evaluation of both visual inference and visual narrating.

2. Hypergraph-Enhanced Graph Reasoning

In this section, we present the hypergraph-enhanced graph reasoning model in detail, as shown in Figure 3. Specifi-

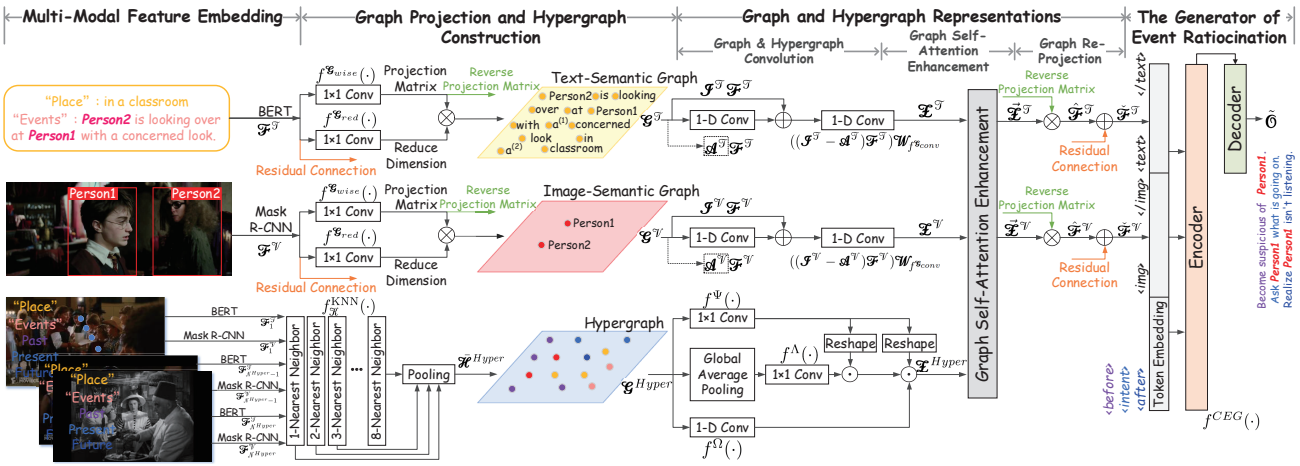

Figure 3. A Brief Illustration of the Proposed Framework, which can be concluded into four steps: *Step 1.* **Multi-Modal Feature Embedding**: multi-modal features are obtained with pre-trained models, consisting of visual feature $\mathcal{F}^{\mathcal{V}}$ and textual feature $\mathcal{F}^{\mathcal{T}}$; *Step 2.* **Graph Projection and Hypergraph Construction**: multi-modal features are projected into two semantic graphs $\mathcal{G}^{\mathcal{V}}$ and $\mathcal{G}^{\mathcal{T}}$, meanwhile these features are utilized to build the hypergraph $\mathcal{G}^{Hyper}$ via the $\mathcal{K}$-nearest neighbor function $f_{\mathcal{K}}^{\text{KNN}}$; *Step 3.* **Graph and Hypergraph Representations**: multi-modal semantic graphs perform with graph convolution to get the graph representations $\mathcal{E}^{\mathcal{V}}$ and $\mathcal{E}^{\mathcal{T}}$, meanwhile the semantic hypergraph performs with hypergraph convolution to get the hypergraph representations $\mathcal{E}^{Hyper}$, then perform the proposed ***Graph Self-Attention Enhancement*** shown in Figure 5 to obtain reinforced graph representations $\vec{\mathcal{E}}^{\mathcal{V}}$, $\vec{\mathcal{E}}^{\mathcal{T}}$, and finally these representations re-project back into original spatial feature domain, resulting in refined sample features $\tilde{\mathcal{F}}^{\mathcal{V}}$, $\tilde{\mathcal{F}}^{\mathcal{T}}$; *Step 4.* **The Generator of Event Ratiocination**: we employ the BART (Lewis et al., 2020) to build the generator $f^{CEG}(\cdot)$ of event ratiocination and obtain the event's cause-and-effect descriptions in an autoregressive manner based on graph re-projected representations (i.e., refined sample features).

cally, we utilize *Multi-Modal Feature Embedding* to obtain the features in the spatial domain and from two modalities, and then we build semantic graphs and hypergraph through *Graph Projection and Hypergraph Construction*. Furthermore, we construct *Graph and Hypergraph Representations*, including ***Graph Self-Attention Enhancement***, to capture the intra- & inter- modality relationships as well as the spatio-temporal relationship among past, present, and future samples. Finally, we employ *The Generator of Event Ratiocination* to generate the cause-and-effect descriptions.

2.1. Multi-Modal Feature Embedding

In this subsection, we formalize the way with pre-trained models to extract multi-modal features.

❶ Visual Features

Taking a given image as input, we detect the visual "person" using the Mask R-CNN (He et al., 2017), which extracts $\mathcal{N}^{\mathcal{V}}$ appearance features $\mathcal{V}^{a} = \{v_i^a\}_{i=1}^{\mathcal{N}^{\mathcal{V}}}$, and their corresponding bounding-box $\mathcal{V}^{b} = \{v_i^b\}_{i=1}^{\mathcal{N}^{\mathcal{V}}}$, where we encode the top-left position and the bottom-right position of the i-th bounding box using the 4-dimensional (-D) vector, i.e., $v_i^b = [x_i^{top}, y_i^{top}, x_i^{btm}, y_i^{btm}]$. To fuse image features, we calculate the visual features: $\mathcal{F}^{\mathcal{V}} = \{v_i\}_{i=1}^{\mathcal{N}^{\mathcal{V}}}$, $v_i \in \mathbb{R}^{d_{\mathcal{N}^{\mathcal{V}}}}$, where $v_i = w^a v_i^a + w^b v_i^b$, w^a and w^b are learn-able parameters, $d_{\mathcal{N}^{\mathcal{V}}}$ is the image feature dimension.

❷ Textual Features

Following the work of VisualCOMET (Park et al., 2020), a given image corresponding $\mathcal{N}^{\mathcal{T}}$-word textual descriptions, including two kind of information (place, and events), is fed into the pre-trained BERT model (Devlin et al., 2019) to obtain the textual feature $\mathcal{F}^{\mathcal{T}} = \{t_i\}_{i=1}^{\mathcal{N}^{\mathcal{T}}}$, where $t_i \in \mathbb{R}^{d_{\mathcal{N}^{\mathcal{T}}}}$ is the embedding of the i-th word, and $d_{\mathcal{N}^{\mathcal{T}}}$ is the dimension.

2.2. Graph Projection and Hypergraph Construction

In this subsection, in order to capture the intra-modality relationship from an individual modality, we build a multi-modal graph by the *Graph Projection*. Further, in order to have a picture of the whole situation that different images with descriptions of events at different moments, we construct a hyper-graph through the *Hypergraph Construction*.

❶ Graph Projection

As shown in Figure 3, given an image and its corresponding textual descriptions, we construct a multi-modal graph composed of two sub-graphs, i.e., image-semantic graph $\mathcal{G}^{\mathcal{V}}$, and text-semantic graph $\mathcal{G}^{\mathcal{T}}$ for representing the information in two modalities. For simplicity, we uniformly denote two semantic graphs as $\mathcal{G}^{tag}$ and original feature embedding $\mathcal{F}^{tag}$, $tag \in \{\mathcal{V}, \mathcal{T}\}$. We project the feature map $\mathcal{F}^{tag}$ from given training samples into the graph $\mathcal{G}^{tag} \in \mathbb{R}^{\mathcal{N}^{tag} \times d_{\mathcal{N}^{tag}}}$, where $\mathcal{N}^{tag}$ is the number of nodes

(i.e., the number of instances in each modality) and $d_{\mathcal{N}^{tag}}$ is the dimension of node features (i.e., the dimension of sample features). The entire projected $\mathcal{G}^{tag}$ can be built as a lightweight *fully-connected* graph. We use the project function $f^{\mathcal{G}_{proj}}(\cdot)$ that can be formulated as a linear combination with learn-able weights for acquiring the $\mathcal{G}^{tag}$:

$$\begin{aligned}\mathcal{G}^{tag} &= f^{\mathcal{G}_{proj}}(f^{\mathcal{G}_{red}}(\mathcal{F}^{tag}; \mathcal{W}_{f^{\mathcal{G}_{red}}})) \\ &= f^{\mathcal{G}_{wise}}(\mathcal{F}^{tag}; \mathcal{W}_{f^{\mathcal{G}_{wise}}}) \times f^{\mathcal{G}_{red}}(\mathcal{F}^{tag}; \mathcal{W}_{f^{\mathcal{G}_{red}}})\end{aligned} \tag{1}$$

where $f^{\mathcal{G}_{wise}}(\cdot)$ and $f^{\mathcal{G}_{red}}(\cdot)$ are two convolution layers (Chen et al., 2019b; Liang et al., 2018) for graph projection and feature dimension reduction, respectively. $\mathcal{W}_{f^{\mathcal{G}_{red}}}$ is weights of $f^{\mathcal{G}_{red}}(\cdot)$ and $\mathcal{W}_{f^{\mathcal{G}_{wise}}}$ is the weights of $f^{\mathcal{G}_{wise}}(\cdot)$.

❷ Hypergraph Construction

As shown in Figure 4, given $\mathcal{N}^{Hyper}$ images and their corresponding textual statements including events, places and three event ratiocination from the training dataset, we constructe a hyper-graph $\mathcal{G}^{Hyper}$ that can be represented by the incidence matrix $\mathcal{H}^{Hyper}$. For each hyper-graph vertex $\mathcal{F}_i^{Hyper} \in \{\mathcal{F}_i^{\mathcal{V}}\}_{i=1}^{\mathcal{N}^{Hyper}} \bigcup \{\mathcal{F}_i^{\mathcal{T}}\}_{i=1}^{\mathcal{N}^{Hyper}}$, where the visual feature $\mathcal{F}_i^{\mathcal{V}}$ and textual feature $\mathcal{F}_i^{\mathcal{T}}$ are both from i-th image. Note that $\mathcal{F}_i^{\mathcal{V}}$ is obtained by Mask-RCNN and image, and $\mathcal{F}_i^{\mathcal{T}}$ is obtained by BERT and textual statements. We find its $\mathcal{K}$ nearest neighbors (Chen et al., 2009), and then utilize each element $\mathcal{H}_{ij}(i, j = 1, \cdots, 2 \times \mathcal{N}^{Hyper})$ of the hyper-graph incidence matrix to connect these vertices:

$$\mathcal{H}_{ij} = \begin{cases} 1 & \mathcal{F}_j^{Hyper} \in f_{\mathcal{K}}^{\mathrm{KNN}}(\mathcal{F}_i^{Hyper}) \\ 0 & otherwise \end{cases} \tag{2}$$

where $f_{\mathcal{K}}^{\mathrm{KNN}}(\cdot)$ is the nearest neighbor function resulting in the neighborhood set containing the top-$\mathcal{K}$ neighbors.

With the above equation, it is obvious that $\mathcal{H}^{Hyper}$ is decided by $f_{\mathcal{K}}^{\mathrm{KNN}}(\cdot)$, which is dictated by the parameter $\mathcal{K}$. Therefore, according to the different values of $\mathcal{K}$, we can get different hyper-graph incidence matrices and thus different hyper-graphs. For the sake of writing, we denote the different incidence matrices as $\mathcal{H}_{\mathcal{K}}^{Hyper}$. In our model, we adopt the average as the final incidence matrix:

$$\bar{\mathcal{H}}^{Hyper} = \frac{1}{\mathcal{K}} \sum_{\hbar=1}^{\mathcal{K}} \mathcal{H}_{\hbar}^{Hyper} \tag{3}$$

2.3. Graph and Hypergraph Representations

In this subsection, in order to capture semantic relations, we update the node representation of semantic graphs by the *Graph Convolution*. Similarly, we update the hyper-graph representation through the *Hypergraph Convolution* to automatically capture high-order semantic relations. Since the images and their statements are from different modalities and different moments, we want to analyze the relationship among these samples to enrich the graph representation from the individual modalities. To this end, we propose the ***Graph Self-Attention Enhancement***, consisting of *Graph-Graph Attention Unit* and *Graph-HyperGraph Attention Unit* for the inter-modality relationship between two modalities as well as the spatio-temporal relationship among past, present, and future samples. Further, we employ the *Graph Re-Project* to transfer the reinforced graph representations into refined sample features.

❶ Graph Convolution

Based on the obtained graph $\mathcal{G}^{tag}$, we make use of *Graph Convolution* (Kipf & Welling, 2017) to further propagate information and aims at correlations between the feature of the relative nodes by learning edge weights. In particular, a single graph convolution with its parameter $\mathcal{W}_{f^{\mathcal{G}_{conv}}} \in \mathbb{R}^{d_{\mathcal{N}^{tag}} \times d_{\mathcal{N}^{tag}}}$ is defined as:

$$\begin{aligned}\mathcal{E}^{tag} &= f^{\mathcal{G}_{conv}}(\mathcal{A}^{tag}\mathcal{F}^{tag}\mathcal{W}_{f^{\mathcal{G}_{conv}}}) \\ &= ((\mathcal{I}^{tag} - \mathcal{A}^{tag})\mathcal{F}^{tag})\mathcal{W}_{f^{\mathcal{G}_{conv}}}\end{aligned} \tag{4}$$

where $\mathcal{A}^{tag}$ is the $\mathcal{N}^{tag} \times \mathcal{N}^{tag}$ adjacency matrix of graph $\mathcal{G}^{tag}$ for cross-nodes diffusion, $\mathcal{I}^{tag} \in \mathbb{R}^{\mathcal{N}^{tag} \times \mathcal{N}^{tag}}$ is the identity matrix. A Laplacian smoothing operator (Li et al., 2018) is performed to propagate the node features over the graph. Considering its own representation of each node, the adjacency matrix is added with self-connection. The graph convolution is implemented by two convolution layers along with channel-wise and node-wise directions as shown in Figure 3. The identity matrix $\mathcal{I}^{tag}$ is also a residual connection for every node. The adjacency matrix and its parameter $\mathcal{W}_{f^{\mathcal{G}_{conv}}}$ can be optimized by gradient descent.

❷ Hypergraph Convolution

To capture high-order semantic relations automatically, we use the *Hypergraph Convolution* (Feng et al., 2019) with the hyper-graph to propagate hypergraph information and update hypergraph embeddings, as illustrated in Figure 3:

$$\begin{aligned}\mathcal{E}^{Hyper} = f^{\Psi}(\bar{\mathcal{H}}^{Hyper}; \mathcal{W}_{f^{\Psi}}) \odot f^{\Lambda}(\bar{\mathcal{H}}^{Hyper}; \mathcal{W}_{f^{\Lambda}}) \\ \odot f^{\Psi}((\bar{\mathcal{H}}^{Hyper})^T; \mathcal{W}_{f^{\Psi}}) \odot f^{\Omega}(\bar{\mathcal{H}}^{Hyper}; \mathcal{W}_{f^{\Omega}})\end{aligned} \tag{5}$$

where $f^{\Psi}(\cdot)$ is the 1×1 convolution with its weights $\mathcal{W}_{f^{\Psi}}$ followed by a non-linear activation function (in our case ReLU function (Goodfellow et al., 2016)); $f^{\Lambda}(\cdot)$ with its weights $\mathcal{W}_{f^{\Lambda}}$ is channel-wise global average pooling (GAP) (Goodfellow et al., 2016) followed by a 1×1 convolution similar to (Hu et al., 2018), and it plays a role in a diagonal matrix, which helps in learning a better distance metric among the nodes for the incidence matrix $\bar{\mathcal{H}}^{Hyper}$; $f^{\Omega}(\cdot)$ with its weights $\mathcal{W}_{f^{\Omega}}$ is a single-layer convolution (Lin et al., 2014a) that is used to capture the global relationship of the features to develop better hyper-edges (Wu et al., 2020); $(\cdot)^T$ means the matrix transpose operation; $\odot$ means the matrix dot product calculation operation.

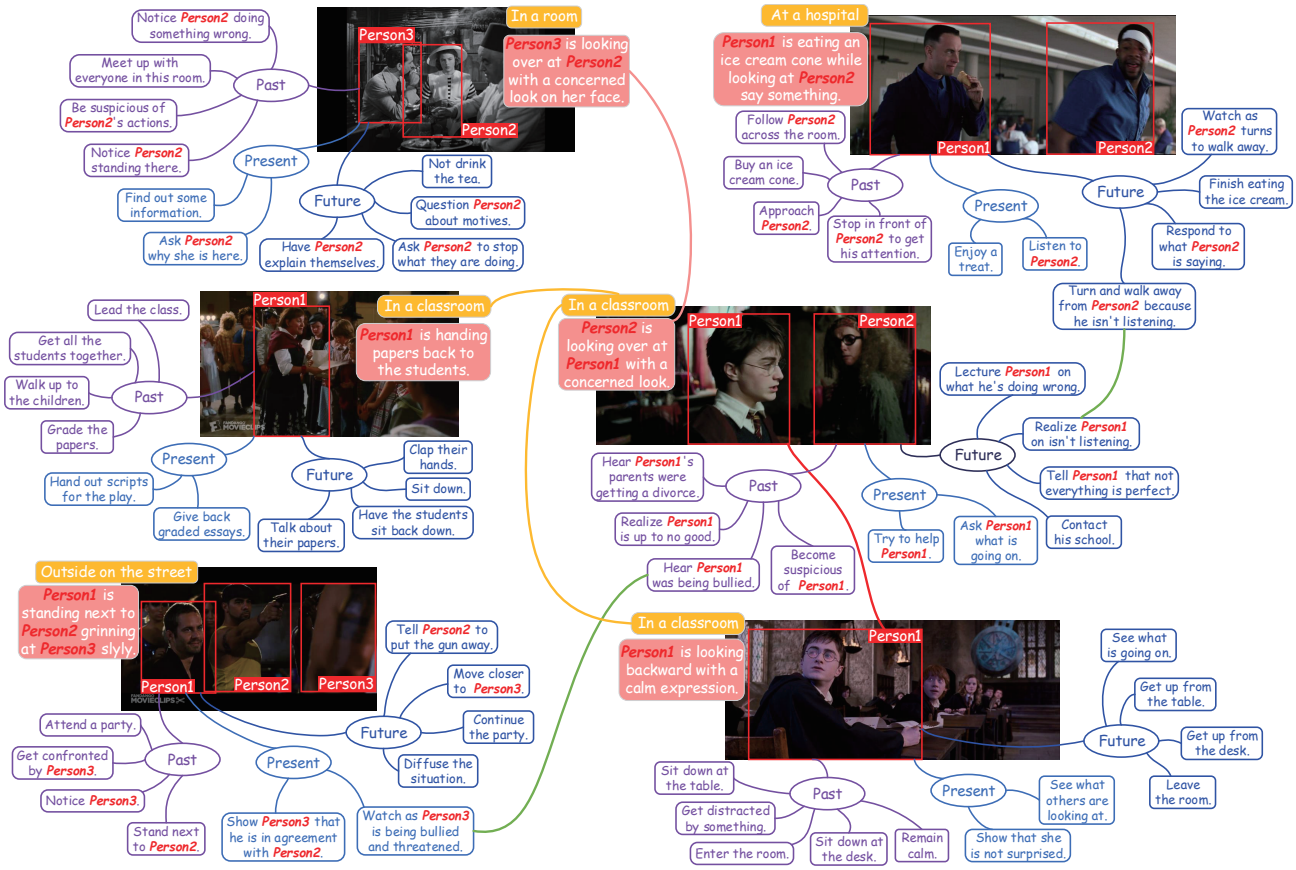

Figure 4. Snapshot of Our Built **HyperGraphs**. Images and their statements with different moments are connected by the same semantics. This way can capture spatio-temporal relationships.

❸ Graph Self-Attention Enhancement

Through the graph convolution process, we get two graph embeddings from two modalities: $\mathfrak{Z}^{\mathcal{V}}$, and $\mathfrak{Z}^{\mathcal{T}}$. Then, by the hyper-graph convolution process, we get the hyper-graph embedding from obtained hyper-graphs: $\mathfrak{Z}^{Hyper}$. We want to ① use each graph embedding of the two modalities to enhance the graph representation of each individually, ② use the hyper-graph embedding to enhance the graph representation of each modality. To this end, we propose the cascaded multi-modal structure of stacked attention layers, each of which contains our ***Graph Self-Attention Enhancement***, consisting of two kinds of self-attentions, as shown in Figure 5. At the last, we get final output $\vec{\mathfrak{Z}}^{\mathcal{V}}$ and $\vec{\mathfrak{Z}}^{\mathcal{T}}$.

We propose the graph self-attention enhancements that is an extension of multi-head attention consisting of some parallel heads, in which we replace original scaled dot product attention in classical multi-head attention (Vaswani et al., 2017) with the non-local attention block (Wang et al., 2018b; Zhu et al., 2019; Dong Zhang & Sun, 2020; Yin et al., 2020; Zhu et al., 2020) in each head. Our non-local attention block contains two kinds of attention units: ① *Graph-Graph Attention Unit* and ② *Graph-HyperGraph Attention Unit*.

① Graph-Graph Attention Unit

The graph-graph attention unit focuses on enriching the graph embedding of one modality with the graph embedding of the other modality. In particular, for len-th layer, suppose there are two sliced graph embeddings from two modalities as the input of our non-local attention block: $\mathfrak{Z}_{len}^{m_1}$ and $\mathfrak{Z}_{len}^{m_2}$, where $m_1, m_2 \in \{\mathcal{V}, \mathcal{T}\}$. The graph-graph attention unit can be represented as:

$$
\begin{aligned}
&\mathcal{Q}\boldsymbol{u}_{len}^{m_1} = \text{query}_{len}^{m_1}(\mathfrak{Z}_{len}^{m_1}), \\
&\mathcal{K}_{len}^{m_2} = \text{key}_{len}^{m_2}(\mathfrak{Z}_{len}^{m_2}), \mathcal{V}\boldsymbol{\ell}_{len}^{m_2} = \text{value}_{len}^{m_2}(\mathfrak{Z}_{len}^{m_2}); \\
&\mathcal{V}\boldsymbol{c}_{len}^{m_1 m_2} = \text{softmax}((\mathcal{Q}\boldsymbol{u}_{len}^{m_1})^T \times \mathcal{K}_{len}^{m_2}) \times (\mathcal{V}\boldsymbol{\ell}_{len}^{m_2})^T, \\
&\mathfrak{Z}_{len}^{m_1 m_2} = \text{cat}((\mathcal{V}\boldsymbol{c}_{len}^{m_1 m_2})^T, \mathcal{Q}\boldsymbol{u}_{len}^{m_1}; \mathcal{W}_{\text{cat}}).
\end{aligned} \tag{6}
$$

where $\text{query}_{len}^{m_1}(\cdot)$, $\text{key}_{len}^{m_2}(\cdot)$, and $\text{value}_{len}^{m_2}(\cdot)$ are three linear transformations; we use the softmax function (Goodfellow et al., 2016) to get the embeddings $\mathcal{V}\boldsymbol{c}_{len}^{m_1 m_2}$; by referring to the design of the non-local block (Wang et al., 2018b), $\text{cat}(\cdot)$ is implemented by a 1×1 convolution, with $\mathcal{W}_{\text{cat}}$ that acts as a weighting parameter to adjust the impor-

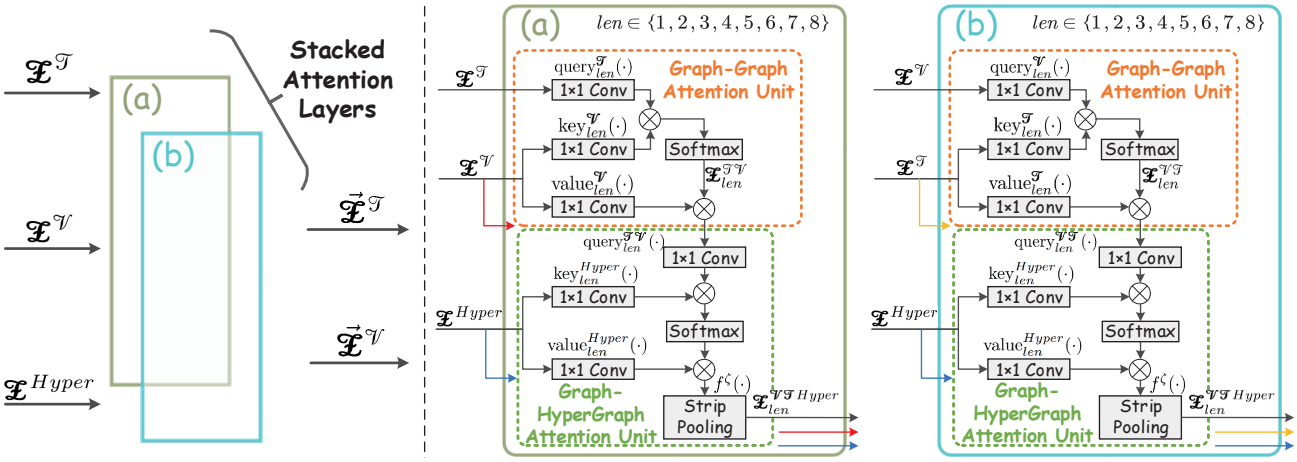

Figure 5. Illustration of ***Graph Self-Attention Enhancement***. It consists of two attention units: **Graph-Graph Attention Unit** and **Graph-HyperGraph Attention Unit**. There are two kinds of attention layers with the stack fashion. The graph-graph attention unit focuses on aggregating the graph embedding of one modality and the graph embedding of the other modality. Similarly, the graph-hypergraph attention unit aims at the aggregation of hypergraph embedding and the results of the former unit.

tance of non-local operation (Zhu et al., 2020; Srinivas et al., 2021). $\mathfrak{X}_{len}^{m_1m_2}$ is the fusion vector of $\mathfrak{X}_{len}^{m_1}$ and $\mathfrak{X}_{len}^{m_2}$.

② Graph-HyperGraph Attention Unit

Similarly, the graph-hypergraph attention unit aims at improving the graph embedding of each modality with hypergraph embedding from the obtained hyper-graph, under the condition of the previous step in the graph-graph attention unit. In particular, for len-th layer, we employ another non-local attention block to process the result $\mathfrak{X}_{len}^{m_1m_2}$ of in the graph-graph attention unit and the sliced hyper-graph embeddings $\mathfrak{X}_{len}^{Hyper}$ from obtained hyper-graphs, as follows:

$$\begin{aligned}\mathcal{V}\text{-}\boldsymbol{c}_{len}^{m_1m_2Hyper} &= \text{softmax}((\mathcal{Q}\boldsymbol{u}_{len}^{m_1m_2})^T \times \mathcal{K}_{len}^{Hyper}) \\ &\times (\mathcal{V}\boldsymbol{t}_{len}^{Hyper})^T; \\ \mathfrak{X}_{len}^{m_1m_2Hyper} &= f^{\zeta}(\text{cat}((\mathcal{V}\text{-}\boldsymbol{c}_{len}^{m_1m_2Hyper})^T, \\ &\mathcal{Q}\boldsymbol{u}_{len}^{m_1m_2}; \mathcal{W}_{\text{cat}})).\end{aligned} \tag{7}$$

where $\mathcal{Q}\boldsymbol{u}_{len}^{m_1m_2}$ is from $\mathfrak{X}_{len}^{m_1m_2}$; $\mathcal{K}_{len}^{Hyper}$ and $\mathcal{V}\boldsymbol{t}_{len}^{Hyper}$ is from $\mathfrak{X}_{len}^{Hyper}$. For rich contextual information, we use the strip pooling $f^{\zeta}(\cdot)$ (Hou et al., 2020) to enhance models.

At the last layer, we use the proposed graph self-attention enhancements for the vectors $\mathfrak{X}^{\mathcal{TV}Hyper}$ and $\mathfrak{X}^{\mathcal{VT}Hyper}$. It can keep the inputted size of the query in the graph-graph attention unit consistent with the final output of graph self-attention enhancements for the next in graph re-projection. Therefore, we re-denote these vectors as $\vec{\mathfrak{X}}^{\mathcal{T}}$ and $\vec{\mathfrak{X}}^{\mathcal{V}}$.

❹ Graph Re-Projection

We use the results of $f^{\mathcal{G}_{wise}}(\cdot)$ (i.e., the projection matrix from $\mathcal{F}^{\mathcal{V}}$ and $\mathcal{F}^{\mathcal{T}}$, correspondingly), which is obtained in *previous* graph projection (mentioned in Sec.2.2), to achieve the *Graph Re-Projection*. Similar to the work of Chen et al. (2019b), $f^{\mathcal{G}_{wise}}(\cdot)$ can re-project the final fusion vector $\vec{\mathfrak{X}}^{\mathcal{V}}$ and $\vec{\mathfrak{X}}^{\mathcal{T}}$ into original spatial feature domain, leading to new feature maps $\hat{\mathcal{F}}^{\mathcal{V}}$ and $\hat{\mathcal{F}}^{\mathcal{T}}$. In the end, these new feature maps are added with a residual connection of original features as final refined features $\check{\mathcal{F}}^{\mathcal{V}}$ and $\check{\mathcal{F}}^{\mathcal{T}}$.

2.4. The Generator of Event Ratiocination

We develop a cause-and-effect generator (CEG) as shown in Figure 3, based on Lewis et al. (2020)'s work and with the refined features $\check{\mathcal{F}}^{\mathcal{V}}$ and $\check{\mathcal{F}}^{\mathcal{T}}$ in the sequence-to-sequence task (Bao et al., 2020) to transfer into the predicted output $\tilde{\mathcal{O}}$. Formally, CEG $f^{CEG}(\cdot)$ with its parameter $\mathcal{W}_{fCEG}$ is:

$$\tilde{\mathcal{O}} = f^{CEG}(\check{\mathcal{F}}^{\mathcal{V}}, \check{\mathcal{F}}^{\mathcal{T}}; \mathcal{W}_{fCEG}) \tag{8}$$

❶ Encoder

Following the BART (Lewis et al., 2020) and its variant (Xing et al., 2021), the encoder of CEG is based on a multi-layer bidirectional Transformer (Dai et al., 2019), as shown in Figure 3. We use $< before >$, $< after >$, or $< intent >$ as the starting special token. To inform the model of different modalities of input, we add two sets of special tokens: for images, we use $< img >$ and $< /img >$ to indicate the start and the end of refined visual feature $\check{\mathcal{F}}^{\mathcal{V}}$, respectively. To inform the model textual inputs, we use $< text >$ and $< /text >$ for refined linguistic feature $\check{\mathcal{F}}^{\mathcal{T}}$.

❷ Decoder

The decoder of our model is also a multi-layer Transformer, similar to Wang et al. (2019b). Different from the encoder,

which is bidirectional, the decoder is unidirectional as it is supposed to be autoregressive when generating texts. The decoder does not take as inputs the visual embeddings.

❸ Pre-Training

We pre-train a CEG model with 12 layers in each of the encoder and decoder, and a hidden size of 1024. Following RoBERTa (Liu et al., 2019), we use a batch size of 8000, and train the model for $500,000$ steps. Documents are tokenized with the same byte-pair encoding as GPT-2 (Radford et al., 2019). To pretrain our model, we use three image-text datasets: Conceptual Captions Dataset (Sharma et al., 2018), Im2Text Dataset (Ordonez et al., 2011) and Microsoft COCO Dataset (Lin et al., 2014b). We pre-train the encoder in two steps, in both cases back-propagating the cross-entropy loss (Goodfellow et al., 2016; Weiss et al., 2015) from the output of the CEG model. In the first step, we freeze the parameter $\boldsymbol{\mathcal{W}}_{fCEG}$ and only update the randomly initialized encoder, and the self-attention input projection matrix of encoder first layer. In the second step, we pre-train all CEG model parameters for a small number of iterations.

3. Experiments and Results

In this section, we experimentally evaluate the proposed model on the benchmark datasets and compare its performance with other state-of-the-arts.

3.1. Benchmark Dataset Description

VisualCOMET dataset (Park et al., 2020) consists of over 1.4 million textual statements of visual event ratiocination carefully annotated over a diverse set of $59,000$ images, each paired with short video summaries of before and after.

3.2. Experimental Setup

In this subsection, we outline the used evaluation metrics and implementation details.

⌛ Evaluation Metrics

VisualCOMET dataset includes 1174K training examples and 146K validation examples. Some examples in the dataset share the same images or events, but with different ratiocination for events before/after or intents at present. Following Park et al. (2020), we report three metrics: BLEU-2 (Papineni et al., 2002), METEOR (Denkowski & Lavie, 2014), and CIDEr (Vedantam et al., 2015). Following the work of Xing et al. (2021), we report our model performance on the validation set as the test set is not available yet.

⌛ Implementation Details

In our **training** process, Adam optimizer (Kingma & Ba, 2015) is used with momentum parameters setting β_1 and β_2 to 0.9 and 0.999. The learning rate is initially set to 0.0001. The training batch size is set to 64. For our built hypergraph, we set the $\mathcal{K}$ to 8. For our graph self-attention enhancements, the number of head in multi-head attention is 8. We set the number of the stacked graph-graph attention unit and graph-hypergraph attention unit to 2.

3.3. Comparison with State-of-The-Arts

We compare the state-of-the-art methods with our model on the VisualCOMET benchmark.

Effect of Graph Reasoning. From Table 1, #1 is better than Vision&Lang Transformer and KM-BART. This suggests that the generator of event ratiocination is effective. Compared to #1, ours and other variants (i.e., #2∼#5) have better performance. Graph reasoning can help the model capture the intra-modality relationship, and the model with graph reasoning is better than without this process. *This implies that the design of graph reasoning is effective.*

Effect of Graph-Graph Attention Unit. The graph-graph attention unit strives to reinforce the graph embedding of one modality with the graph embedding of the other modality. In essence, the graph embedding as the query's input exploits this unit with the information of another modality to enhance itself. In this way, the model can capture the inter-modality relationship. From Table 1, it is clear that the method (i.e., #5) with the graph-graph attention unit is better than without this unit (i.e., #2∼#4) . *It shows the proposed graph-graph attention unit effectively improves the task of visual event ratiocination.*

Table 1. Comparison Results on the VisualCOMET dataset. "VGE" stands for Visual Graph Embedding. "LGE" stands for Linguistic Graph Embedding. "G-G" stands for Graph-Graph Attention Unit. "G-HG" stands for Graph-HyperGraph Attention Unit. ✔ stands for the RESERVED component; ✖ means the REMOVED component; ⊗ means the component that REPLACES with original feature embedding. Best is pointed in **bold**.

Method					BLEU-2	METEOR	CIDEr
Vision&Lang Transformer (Park et al., 2020)					13.50	11.55	18.27
KM-BART (Xing et al., 2021)					23.47	15.02	39.76
Our Variants							
#	VGE	LGE	G-G	G-HG			
1	⊗	⊗	✖	✖	24.19	16.00	40.26
2	✔	⊗	✖	✖	25.42	18.22	41.36
3	⊗	✔	✖	✖	26.48	19.48	42.48
4	✔	✔	✖	✖	29.31	20.46	43.11
5	✔	✔	✔	✖	31.77	23.02	47.39
Ours					**32.97**	**23.99**	**49.19**

Effect of Graph-HyperGraph Attention Unit. Similarly, the graph-hypergraph attention unit is the extension of the graph-graph attention unit and integrates the graph embedding from two modalities with the hypergraph. In other words, the graph embedding from the hypergraph is employed with the graph-hypergraph attention to strengthen the

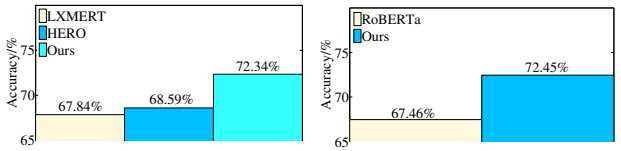

Figure 6. The Comparison Results of Visual Inference. (a) Results on The VIOLIN Dataset; (b) Results on The VLEP Dataset.

graph embedding from the two modalities. In this way, the model can capture the spatio-temporal relationship among past, present, and future samples. From Table 1, Ours is better than #5. From above, *our graph-hypergraph attention unit is effective for the task of visual event ratiocination.*

Effect of *Graph Self-Attention Enhancement.* Graph self-attention enhancement consists of two kinds of attention units: graph-graph attention unit and graph-hypergraph attention unit. Therefore, from the above analysis, *it is clear that our graph self-attention enhancement is effective.*

Effect of Our Approach. Our model can capture the intra- and inter- relationships from two modalities as well as the spatio-temporal relationship among past, present, and future samples. From Table 1, Ours is better than others. *It shows that our model is more robust than other state-of-the-art approaches on the VisualCOMET dataset.*

Summary. From these above analyses, our approach shows the case of "two heads are better than one" in the sense that semantic graph representations with the graph self-attention enhancement are more robust than those without.

3.4. Discussion on The Ability of Visual Inference

The visual inference is one of the key to the task of visual event ratiocination, and requires the model that devotes itself to the judgment of the temporal event's ratiocination for a given image. In order to further evaluate the performance of the proposed model, we perform two tasks: video-and-language inference (Liu et al., 2020) and video-and-language future event prediction (Lei et al., 2020).

❶ Video-and-Language Inference

Video-and-language inference aims at the joint understanding of video and text. Given a video clip with aligned subtitles as the premise, paired with a natural language hypothesis, a model needs to infer whether the hypothesis is entailed or contradicted by the given video clip. To fit this task, for our model, we replace the textual descriptions in our original task (i.e., visual event ratiocination) with subtitles and statements, split the given video clip into frames. We compare the ratiocination generated by our model with the candidates to obtain the final results. Specifically, we input the generated ratiocination and the candidate ones to the Bi-LSTM (Huang et al., 2015; Liu et al., 2020) to make the prediction. We use *accuracy* as evaluation metric.

Dataset. VIOLIN benchmark dataset (Liu et al., 2020) consists of $95,322$ video hypothesis pairs from $15,887$ video clips, spanning over 582 hours of video. These video clips contain rich content with event shifts, and interaction.

Performance of Different Methods. From Figure 6(a), Ours is better than LXMERT (Tan & Bansal, 2019; Liu et al., 2020), HERO (Li et al., 2020). *This suggests that our model has better performance than other state-of-the-art approaches for the task of video-and-language inference.*

❷ Video-and-Language Future Event Prediction

Video-and-language event prediction focuses on future event prediction from videos. In particular, given a video with dialogue, and two possible future events, the model is required to understand both visual and language semantics from this video, and choose the more likely event from two provided possible future events. Similarly, to fit this task, for our model, we split the given video clip into frames, and compare the ratiocination generated by our model with the candidates to obtain the final results by the Bi-LSTM (Huang et al., 2015). We use *accuracy* as evaluation metric.

Dataset. VLEP benchmark dataset (Lei et al., 2020) contains $28,726$ examples from $10,234$ short video clips. Each example consists of a short video clip with its dialogue and text summary, and two potential future event ratiocination.

Performance of Different Methods. From Figure 6(b), Ours is better than RoBERTa (Liu et al., 2019; Lei et al., 2020). As a results, *our model has state-of-the-art performance for the task of video-and-language event prediction.*

3.5. Discussion on The Ability of Visual Narrating

Visual narrating is another key for the task of visual event ratiocination and focuses on generating semantic descriptions from images or videos, e.g., video captioning (Shetty & Laaksonen, 2016) and visual storytelling (Huang et al., 2016). To further evaluate our model, we perform these two tasks in this subsection. To fit these two tasks, for our model, we split the given video clip into frames and fix the starting special token as $< intent >$. We use *BLEU-4* (Papineni et al., 2002), *METEOR*, and *CIDEr* as evaluation metrics.

❶ Video Captioning

The goal of video captioning is to generate a sentence to describe video content accurately. Here, we introduce the used dataset followed by the comparison analysis.

Dataset. MSVD (Chen & Dolan, 2011) contains $1,970$ video clips with multiple descriptions for each video clip. Following the work of Venugopalan et al. (2015), we use

Table 2. Comparison Results on The Task of Video Captioning. Best is pointed in **bold**.

Method	BLEU-4	METEOR	CIDEr	Method	BLEU-4	METEOR	CIDEr
S2VT (Venugopalan et al., 2015)	42.1	30.00	58.80	STAT (Tu et al., 2017)	51.1	32.7	67.5
MP-LSTM (Venugopalan et al., 2015)	50.40	32.50	71.00	M3 (Wang et al., 2018)	51.78	32.49	-
LSTM-E (Pan et al., 2016)	45.30	31.00	-	VRE (Shi et al., 2019)	51.7	34.3	86.7
p-RNN (Yu et al., 2016)	47.40	30.30	53.60	RecNet (Wang et al., 2018b)	52.3	34.1	80.3
Tempor-attention (Yao et al., 2015)	41.92	29.60	51.67	Xgating (Wang et al., 2019a)	52.5	34.1	88.7
Bi-GRU-RCN (Ballas et al., 2016)	48.42	31.70	65.38	MGSA (Chen & Jiang, 2019)	53.4	35.0	86.7
hLSTMat (Song et al., 2017)	48.50	31.90	-	PMI-CAP (Chen et al., 2020)	54.68	36.4	95.17
MAMRNN (Li et al., 2017)	41.40	32.20	53.90	OA-BTG (Zhang & Peng, 2019)	56.9	36.2	90.6
PickNet (Chen et al., 2018)	46.10	33.10	76.00	ELTI (Wei et al., 2020)	46.8	34.4	85.7
MCF (Wu & Han, 2018)	46.46	33.72	75.46	TTA (Tu et al., 2021)	51.8	35.5	87.7
RecNet (Wang et al., 2018a)	52.30	34.10	80.30	FCVC-CF-IA (Fang et al., 2019)	53.1	34.8	79.8
GRU-EVE (Aafaq et al., 2019)	45.60	33.70	74.20	S2VT+RL+DRPN (Xu et al., 2020)	49.2	32.6	86.4
Middle-out+self (Mehri & Sigal, 2018)	47.00	34.10	79.50	hLSTMat + DRPN (Xu et al., 2020)	57.3	34.2	78.3
TDConvED (Chen et al., 2019a)	48.30	32.90	72.30	Latest guiding DenseLSTM (Zhu & Jiang, 2019)	50.4	32.9	72.6
MemNet (Wu et al., 2020)	51.62	34.85	84.27	Ours	**57.67**	**38.38**	**97.13**

Table 3. Comparison Results on The Task of Visual Storytelling. Best is pointed in **bold**.

Method	BLEU-4	ROUGE	METEOR	CIDEr	Method	BLEU-4	ROUGE	METEOR	CIDEr
HSRL (Huang et al., 2019)	13.4	35.2	30.8	-	StoryAnchor (Zhang et al., 2020)	14.0	35.5	30.0	9.9
hattn-rank (Yu et al., 2017)	-	34.1	29.5	7.5	SGVST (Wang et al., 2020b)	14.7	29.9	35.8	9.8
CIDEr-RL (Wang et al., 2018a)	13.8	34.9	29.7	8.1	K-Storyteller (Yang et al., 2019)	12.8	29.9	35.2	12.1
GAN (Wang et al., 2018a)	14.0	35.0	29.5	9.0	TAVST (Wang et al., 2020a)	14.6	31.0	35.7	9.2
VSCMR (Li et al., 2019)	14.3	35.5	30.2	9.0	INet (Jung et al., 2020)	14.7	35.6	29.7	10.0
ARLE-IRL (Wang et al., 2018a)	14.1	35.0	29.5	9.4	SGEmb (Hong et al., 2020)	14.8	35.6	30.2	8.6
MemNet (Wu et al., 2020)	14.1	35.5	29.5	9.2	Ours	**16.7**	**37.4**	**37.8**	**14.1**

$1,200$ video clips for training, 100 video clips for validation, and 670 video clips for testing.

Performance of Different Methods. On the MSVD dataset, we compare ours with 29 state-of-the-arts. From Table 2, the proposed model significantly outperforms existing state-of-the-arts. From above, *our model is more robust than other state-of-the-arts on the task of video captioning.*

❷ Visual Storytelling

Visual storytelling requires the model to understand the event flow in these photos deeply. Here, we introduce the used dataset followed by the comparison analysis.

Dataset. The VIST dataset (Huang et al., 2016) is used for solving visual storytelling, which includes $10,117$ Flickr albums with $210,819$ unique images. After filtering the broken images, there are $40,098$ training, $4,988$ validation, and $5,050$ testing samples.

Performance of Different Methods. On the VIST dataset, we compare ours with 13 state-of-the-arts. From Table 3, the proposed model significantly outperforms existing state-of-the-arts. From above, *our model is more robust than other state-of-the-arts on the task of visual storytelling.*

Summary of Discussions

It is the key to solving the task of visual event ratiocination that visual inference and visual narrating. We utilize 4 tasks (i.e., video-and-language inference, video-and-language future event prediction, video captioning, and visual storytelling) to investigate these two key points. From Section 3.4 $\sim$ 3.5, *our model has the ability of not only visual inference but also visual narrating, and state-of-the-art performance.*

4. Conclusion and Future Work

In this paper, we present a novel multi-modal model **Hypergraph-Enhanced Graph Reasoning** for the task of visual event ratiocination. The model firstly represents the image with multi-modal contents as two semantic graphs, where each graph represents one modality. Then, the proposed ***Graph Self-Attention Enhancement*** in our model, rises up each graph representations with the help of our built hypergraph, followed by re-projecting back into the original spatial feature domain. Finally, we obtain the cause-and-effect descriptions with these finer representations of elements about the image. Experimental results show that our model achieves state-of-the-art performance. Moving forward, we will take rich structured information especially effective knowledge graphs as the guidance for our model.

Acknowledgements

We would like to thank all anonymous reviewers for their useful feedback. This work is supported in part by MOST and NNSF of China (2008AAA0101502, 61806198, U1811463), and Squirrel AI Learning.

References

Aafaq, N., Akhtar, N., Liu, W., Gilani, S. Z., and Mian, A. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Ballas, N., Yao, L., Pal, C., and Courville, A. C. Delving deeper into convolutional networks for learning video representations. In *ICLR (Poster)*, 2016. URL http://arxiv.org/abs/1511.06432.

Bao, H., Dong, L., Wei, F., Wang, W., Yang, N., Liu, X., Wang, Y., Gao, J., Piao, S., Zhou, M., and Hon, H.-W. UniLMv2: Pseudo-masked language models for unified language model pre-training. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 642–652, Virtual, 13–18 Jul 2020. PMLR.

Chen, D. and Dolan, W. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 190–200, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Chen, J., ren Fang, H., and Saad, Y. Fast approximate knn graph construction for high dimensional data via recursive lanczos bisection. *Journal of Machine Learning Research*, 10(69):1989–2012, 2009. URL http://jmlr.org/papers/v10/chen09b.html.

Chen, J., Pan, Y., Li, Y., Yao, T., Chao, H., and Mei, T. Temporal deformable convolutional encoder-decoder networks for video captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (01):8167–8174, Jul. 2019a. doi: 10.1609/aaai.v33i01.33018167. URL https://ojs.aaai.org/index.php/AAAI/article/view/4826.

Chen, S. and Jiang, Y.-G. Motion guided spatial attention for video captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 8191–8198, Jul. 2019. doi: 10.1609/aaai.v33i01.33018191. URL https://ojs.aaai.org/index.php/AAAI/article/view/4829.

Chen, S., Jiang, W., Liu, W., and Jiang, Y.-G. Learning modality interaction for temporal sentence localization and event captioning in videos. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M. (eds.), *Computer Vision – ECCV 2020*, pp. 333–351, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58548-8.

Chen, Y., Wang, S., Zhang, W., and Huang, Q. Less is more: Picking informative frames for video captioning. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.), *Computer Vision – ECCV 2018*, pp. 367–384, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01261-8.

Chen, Y., Rohrbach, M., Yan, Z., Shuicheng, Y., Feng, J., and Kalantidis, Y. Graph-based global reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019b.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL https://www.aclweb.org/anthology/P19-1285.

Denkowski, M. and Lavie, A. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3348. URL https://www.aclweb.org/anthology/W14-3348.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Dong Zhang, Hanwang Zhang, J. T. M. W. X. H. and Sun, Q. Feature pyramid transformer. In *European Conference on Computer Vision (ECCV)*, 2020.

Fang, K., Zhou, L., Jin, C., Zhang, Y., Weng, K., Zhang, T., and Fan, W. Fully convolutional video captioning with coarse-to-fine and inherited attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (01):8271–8278, Jul. 2019. doi: 10.1609/aaai.v33i01.33018271. URL https://ojs.aaai.org/index.php/AAAI/article/view/4839.

Feng, Y., You, H., Zhang, Z., Ji, R., and Gao, Y. Hypergraph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 3558–3565, Jul. 2019. doi: 10.1609/aaai.v33i01.33013558. URL https://ojs.aaai.org/index.php/AAAI/article/view/4235.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. The MIT Press, 2016. ISBN 0262035618.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017. doi: 10.1109/ICCV.2017.322.

Hong, X., Shetty, R., Sayeed, A., Mehra, K., Demberg, V., and Schiele, B. Diverse and relevant visual storytelling with scene graph embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 420–430, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.34. URL https://www.aclweb.org/anthology/2020.conll-1.34.

Hou, Q., Zhang, L., Cheng, M.-M., and Feng, J. Strip pooling: Rethinking spatial pooling for scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Huang, Q., Gan, Z., Celikyilmaz, A., Wu, D., Wang, J., and He, X. Hierarchically structured reinforcement learning for topically coherent visual story generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8465–8472, Jul. 2019. doi: 10.1609/aaai.v33i01.33018465. URL https://ojs.aaai.org/index.php/AAAI/article/view/4863.

Huang, T.-H. K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., Parikh, D., Vanderwende, L., Galley, M., and Mitchell, M. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1233–1239, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1147.

Huang, Z., Xu, W., and Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991, 2015.

Jung, Y., Kim, D., Woo, S., Kim, K., Kim, S., and Kweon, I. S. Hide-and-tell: Learning to bridge photo streams for visual storytelling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11213–11220, Apr. 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6780.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Lei, J., Yu, L., Berg, T., and Bansal, M. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8769–8784, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.706. URL https://www.aclweb.org/anthology/2020.emnlp-main.706.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://www.aclweb.org/anthology/2020.acl-main.703.

Li, J., Shi, H., Tang, S., Wu, F., and Zhuang, Y. Informative visual storytelling with cross-modal rules. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pp. 2314–2322, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368896. doi: 10.1145/3343031.3350918. URL https://doi.org/10.1145/3343031.3350918.

Li, L., Chen, Y.-C., Cheng, Y., Gan, Z., Yu, L., and Liu, J. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2046–2065, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.161. URL https://www.aclweb.org/anthology/2020.emnlp-main.161.

Li, Q., Han, Z., and Wu, X.-m. Deeper insights into graph convolutional networks for semi-supervised learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11604.

Li, X., Zhao, B., and Lu, X. Mam-rnn: Multi-level attention model based rnn for video captioning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 2208–2214, 2017. doi: 10.24963/ijcai.2017/307. URL https://doi.org/10.24963/ijcai.2017/307.

Liang, X., Hu, Z., Zhang, H., Lin, L., and Xing, E. P. Symbolic graph reasoning meets convolutions. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 1853–1863. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/cbb6a3b884f4f88b3a8e3d44c636cbd8-Paper.pdf.

Lin, M., Chen, Q., and Yan, S. Network in network. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014a. URL http://arxiv.org/abs/1312.4400.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014b. Springer International Publishing. ISBN 978-3-319-10602-1.

Liu, J., Chen, W., Cheng, Y., Gan, Z., Yu, L., Yang, Y., and Liu, J. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Liu, R., Liu, C., Bai, Y., and Yuille, A. L. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, art. arXiv:1907.11692, July 2019.

Mehri, S. and Sigal, L. Middle-out decoding. In *NeurIPS*, pp. 5523–5534, 2018. URL http://papers.nips.cc/paper/7796-middle-out-decoding.

Ordonez, V., Kulkarni, G., and Berg, T. L. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011.

Pan, Y., Mei, T., Yao, T., Li, H., and Rui, Y. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

Park, J. S., Bhagavatula, C., Mottaghi, R., Farhadi, A., and Choi, Y. Visualcomet: Reasoning about the dynamic context of a still image. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

Shetty, R. and Laaksonen, J. Frame- and segment-level features and candidate pool evaluation for video caption generation. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pp. 1073–1076, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450336031. doi: 10.1145/2964284.2984062. URL https://doi.org/10.1145/2964284.2984062.

Shi, X., Cai, J., Joty, S., and Gu, J. Watch it twice: Video captioning with a refocused video encoder. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pp. 818–826, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368896. doi: 10.1145/3343031.3351060. URL https://doi.org/10.1145/3343031.3351060.

Song, J., Gao, L., Guo, Z., Liu, W., Zhang, D., and Shen, H. T. Hierarchical lstm with adjusted temporal attention for video captioning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 2737–2743, 2017. doi: 10.24963/ijcai.2017/381. URL https://doi.org/10.24963/ijcai.2017/381.

Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P., and Vaswani, A. Bottleneck Transformers for Visual Recognition. *arXiv e-prints*, art. arXiv:2101.11605, January 2021.

Tan, H. and Bansal, M. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://www.aclweb.org/anthology/D19-1514.

Tu, Y., Zhang, X., Liu, B., and Yan, C. Video description with spatial-temporal attention. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, pp. 1014–1022, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349062. doi: 10.1145/3123266.3123354. URL https://doi.org/10.1145/3123266.3123354.

Tu, Y., Zhou, C., Guo, J., Gao, S., and Yu, Z. Enhancing the alignment between target words and corresponding frames for video captioning. *Pattern Recognition*, 111:107702, 2021. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2020.107702. URL http://www.sciencedirect.com/science/article/pii/S0031320320305057.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Curran Associates, Inc., 2017.

Vedantam, R., Zitnick, C. L., and Parikh, D. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., and Saenko, K. Sequence to sequence – video to text. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4534–4542, 2015. doi: 10.1109/ICCV.2015.515.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., and Saenko, K. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1494–1504, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1173. URL https://www.aclweb.org/anthology/N15-1173.

Wang, B., Ma, L., Zhang, W., and Liu, W. Reconstruction network for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018a.

Wang, B., Ma, L., Zhang, W., and Liu, W. Reconstruction network for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018b.

Wang, B., Ma, L., Zhang, W., Jiang, W., Wang, J., and Liu, W. Controllable video captioning with pos sequence guidance based on gated fusion network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019a.

Wang, J., Wang, W., Huang, Y., Wang, L., and Tan, T. M3: Multimodal memory modelling for video captioning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7512–7520, 2018. doi: 10.1109/CVPR.2018.00784.

Wang, R., Wei, Z., Cheng, Y., Li, P., Shan, H., Zhang, J., Zhang, Q., and Huang, X. Keep it consistent: Topic-aware storytelling from an image stream via iterative multi-agent communication. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 2250–2260, Barcelona, Spain (Online), December 2020a. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.204. URL https://www.aclweb.org/anthology/2020.coling-main.204.

Wang, R., Wei, Z., Li, P., Zhang, Q., and Huang, X. Storytelling from an image stream using scene graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9185–9192, Apr. 2020b. doi: 10.1609/aaai.v34i05.6455. URL https://ojs.aaai.org/index.php/AAAI/article/view/6455.

Wang, X., Chen, W., Wang, Y.-F., and Wang, W. Y. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 899–909, Melbourne, Australia, July 2018a. Association for Computational Linguistics. doi: 10.18653/v1/P18-1083. URL https://www.aclweb.org/anthology/P18-1083.

Wang, X., Girshick, R., Gupta, A., and He, K. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018b.

Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.-F., and Wang, W. Y. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019b.

Wei, R., Mi, L., Hu, Y., and Chen, Z. Exploiting the local temporal information for video captioning. *Journal of Visual Communication and Image Representation*, 67:102751, 2020. ISSN 1047-3203. doi: https://doi.org/10.1016/j.jvcir.2020.102751. URL http://www.sciencedirect.com/science/article/pii/S1047320320300018.

Weiss, S. M., Indurkhya, N., and Zhang, T. *Fundamentals of predictive text mining*. Springer, 2015.

Wu, A. and Han, Y. Multi-modal circulant fusion for video-to-language and backward. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 1029–1035. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/143. URL https://doi.org/10.24963/ijcai.2018/143.

Wu, A., Han, Y., Zhao, Z., and Yang, Y. Hierarchical memory decoder for visual narrating. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2020. doi: 10.1109/TCSVT.2020.3020877.

Wu, X., Chen, Q., Li, W., Xiao, Y., and Hu, B. Adahgnn: Adaptive hypergraph neural networks for multi-label image classification. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pp. 284–293, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3414046. URL https://doi.org/10.1145/3394171.3414046.

Xing, Y., Shi, Z., Meng, Z., Ma, Y., and Wattenhofer, R. KM-BART: Knowledge Enhanced Multimodal BART for Visual Commonsense Generation. *arXiv e-prints*, art. arXiv:2101.00419, January 2021.

Xu, W., Yu, J., Miao, Z., Wan, L., Tian, Y., and Ji, Q. Deep reinforcement polishing network for video captioning. *IEEE Transactions on Multimedia*, pp. 1–1, 2020. doi: 10.1109/TMM.2020.3002669.

Yang, P., Luo, F., Chen, P., Li, L., Yin, Z., He, X., and Sun, X. Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 5356–5362. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/744. URL https://doi.org/10.24963/ijcai.2019/744.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., and Courville, A. Describing videos by exploiting temporal structure. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pp. 4507–4515, USA, 2015. IEEE Computer Society. ISBN 9781467383912. doi: 10.1109/ICCV.2015.512. URL https://doi.org/10.1109/ICCV.2015.512.

Yin, M., Yao, Z., Cao, Y., Li, X., Zhang, Z., Lin, S., and Hu, H. Disentangled non-local neural networks. In *European Conference on Computer Vision (ECCV)*, 2020.

Yu, H., Wang, J., Huang, Z., Yang, Y., and Xu, W. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Yu, L., Bansal, M., and Berg, T. Hierarchically-attentive RNN for album summarization and storytelling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 966–971, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1101. URL https://www.aclweb.org/anthology/D17-1101.

Zhang, B., Hu, H., and Sha, F. Visual Storytelling via Predicting Anchor Word Embeddings in the Stories. *arXiv e-prints*, art. arXiv:2001.04541, January 2020.

Zhang, J. and Peng, Y. Object-aware aggregation with bidirectional temporal graph for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Zhu, F., Fang, C., and Ma, K. K. Pnen: Pyramid non-local enhanced networks. *IEEE Transactions on Image Processing*, 29:8831–8841, 2020. doi: 10.1109/TIP.2020.3019644.

Zhu, Y. and Jiang, S. Attention-based densely connected lstm for video captioning. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pp. 802–810, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368896. doi: 10.1145/3343031.3350932. URL https://doi.org/10.1145/3343031.3350932.

Zhu, Z., Xu, M., Bai, S., Huang, T., and Bai, X. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.