

# Multimodal Event Transformer for Image-guided Story Ending Generation

Yucheng Zhou, Guodong Long

Australian AI Institute, School of Computer Science, FEIT, University of Technology Sydney  
yucheng.zhou-1@student.uts.edu.au, guodong.long@uts.edu.au

## Abstract

Image-guided story ending generation (IgSEG) is to generate a story ending based on given story plots and ending image. Existing methods focus on cross-modal feature fusion but overlook reasoning and mining implicit information from story plots and ending image. To tackle this drawback, we propose a multimodal event transformer, an event-based reasoning framework for IgSEG. Specifically, we construct visual and semantic event graphs from story plots and ending image, and leverage event-based reasoning to reason and mine implicit information in a single modality. Next, we connect visual and semantic event graphs and utilize cross-modal fusion to integrate different-modality features. In addition, we propose a multimodal injector to adaptive pass essential information to decoder. Besides, we present an incoherence detection to enhance the understanding context of a story plot and the robustness of graph modeling for our model. Experimental results show that our method achieves state-of-the-art performance for the image-guided story ending generation.

## 1 Introduction

Story ending generation (Guan et al., 2019) aims to generate a reasonable ending for a given story plot. It requires deep models to integrate powerful language understanding capability, which is crucial for artificial intelligence. Many efforts (Wang and Wan, 2019; Guan et al., 2019; Yao et al., 2019; Guan et al., 2020) have been proposed and achieved promising results since neural models designed for comprehending natural language allow them to understand story plots and reason reasonable story endings. With the advance of automatic story generation, it has attracted outstanding attention in multimodality research (Jung et al., 2020; Yu et al., 2021; Chen et al., 2021).

However, since story plots and story ending usually correspond to different content, the context

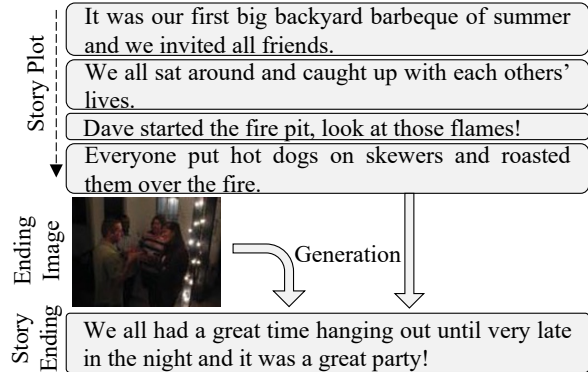


Figure 1: Given a multi-sentence story plot and an ending image, the image-guided story ending generation aims to generate a story ending related to the image.

with information bottleneck is not enough to deduce an informative story ending, i.e., generated endings tend to be inane and generic. To address this issue, Huang et al. (2021) propose an image-guided story ending generation (IgSEG) task that combines story plots and ending image to generate a coherent, specific and informative story ending. IgSEG demands not only introducing information from the ending image to story plots for story ending generation but also reasoning and mining implicit information from story plots and ending image, respectively. As shown in Figure 1, for story plots, “party” can be inferred from “big backyard barbeque” and “invited all friends”, and “all friends”, “all sat around” and “caught up with” can deduce “had a great time”. For the ending image, “dim indoor” and “bright lights” can infer “very late in the night”.

Existing methods (Huang et al., 2021; Xue et al., 2022) focus on cross-modal feature fusion but overlook reasoning and mining implicit information from story plots and ending images. Nonetheless, to effectively conduct cross-modal feature fusion, it is necessary to reason and mine more implicit information from single-modality data. An event is a fine-grained semantic unit, which refers to a text

span composed of a predicate and its arguments (Zhang et al., 2020). Recently, event-centric reasoning displays excellent capability for context understanding and subsequent event prediction (Zhou et al., 2022b). In this work, we propose a multi-modal event transformer (MET) to mine implicit information to improve cross-modal fusion. For story plots, we leverage semantic role labeling (SRL) parser (He et al., 2017) to extract events from story plots and then construct them into a semantic event graph. For an ending image, we utilize scene graph parser (Zellers et al., 2018) to capture visual concepts and their relation to construct visual event graphs. Since edges contain relationships between nodes in visual and semantic event graphs, we employ relational graph convolutional networks (RGCN) (Schlichtkrull et al., 2018) to encode event graphs to infer implicit information.

For cross-modal feature fusion, most recent works (Huang et al., 2021; Xue et al., 2022) adopt attention-based neural network models to implicitly integrate multi-modal features. However, due to the complexity of cross-modal features and the existence of dependency between single-modal features, it is often difficult for these models to complement cross-modal features. To tackle the issue, we propose cross-modal fusion to integrate different-modality features. Specifically, we merge visual and semantic event graphs and use RGCN to fuse cross-modal features for feature complement.

Moreover, since features from different modalities suffer from domain inconsistency, previous methods (Huang et al., 2021; Xue et al., 2022) directly concatenate them and pass them to the decoder, which is not a crafted manner. To appropriately combine features from different modalities, we design a multimodal injector to integrate relevant features into the decoder. In addition, we propose an incoherence detection to enhance the context understanding for a story plot and the robustness of graph modeling for our model.

In experiments, we conduct extensive evaluations on two datasets (i.e., VIST-E (Huang et al., 2021) and LSMDC-E (Xue et al., 2022)). Experimental results show that our method outperforms strong competitors and achieves state-of-the-art performance. In addition, we conduct further analysis to demonstrate the effectiveness of our method. Lastly, we compare the performance of our method and other methods through human evaluation.

## 2 Related Work

### 2.1 Story Ending Generation

Story ending generation aims to generate a story ending for given story plots, and it is one of the important tasks in natural language generation. Many efforts have been invested in story ending generation (Wang and Wan, 2019; Guan et al., 2019; Yao et al., 2019; Guan et al., 2020). To make the generated story ending more reasonable, Guan et al. (2019) propose a model encapsulating a multi-source attention mechanism, which can utilize context clues and understand commonsense knowledge. To ensure the coherence in generated story endings, Wang and Wan (2019) propose a transformer-based conditional autoencoder, which can capture contextual clues in story plot. To improve long-range coherence in generated stories, Guan et al. (2020) pre-train model on external commonsense knowledge bases for the story ending generation. Zhou et al. (2022b) propose a correlation-aware context-to-event pre-trained transformer, which applies to a wide range of event-centric reasoning and generation scenarios, including story ending generation. Beyond the limit of single-modal information, Huang et al. (2021) introduce visual information to enrich the generation of story endings with more coherent, specific, and informative. To improve cross-modal feature fusion, Xue et al. (2022) propose a multimodal memory transformer, which fuses contextual and visual information to capture the multimodal dependency effectively.

### 2.2 Visual Storytelling

Visual storytelling task is proposed by Huang et al. (2016), which aims to generate a story based on a given image stream. Wang et al. (2018) present an adversarial reward learning framework to learn an implicit reward function from human demonstrations. To inject imaginary concepts that do not appear in the images, some works (Yang et al., 2019; Chen et al., 2021; Xu et al., 2021) propose building scene graphs and injecting external knowledge into model to reason the relationship between visual concepts. Qi et al. (2021) propose a latent memory-augmented graph transformer to exploit the semantic relationships among image regions and attentively aggregate critical visual features based on the parsed scene graphs.

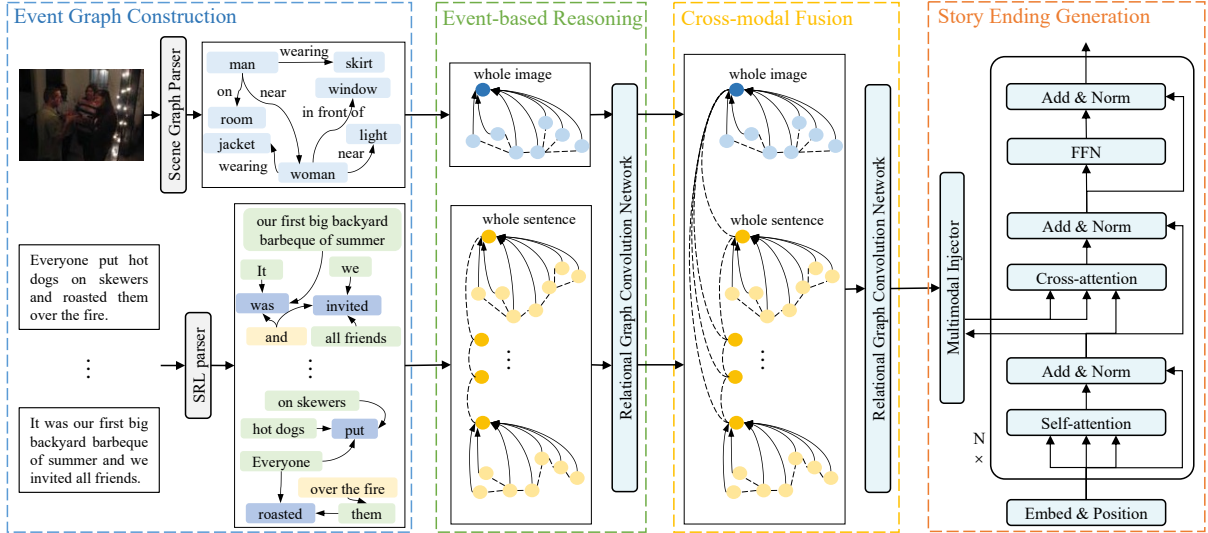


Figure 2: An overview of our model. Grey rounded rectangles denote fixed model. Blue rounded rectangles denote parameters that will be optimized.

### 2.3 Event-centric Reasoning

Events always play an essential role in a story because a story is composed of multiple events and implies the relationship between the events. An event is a text span composed of a predicate and its arguments (Zhang et al., 2020). Multiple events include relations between events that conform to human commonsense (Zhou et al., 2022a). Some works use plot events for story generation, which is generating a prompt and then transforming it into a text (Ammanabrolu et al., 2020; Fan et al., 2019). To generate a more coherent and specific ending, understanding events in story plots and their relationship can obtain informative context, which is a crucial step for story ending generation.

## 3 Method

This section will elaborate on our method for image-guided story ending generation, including event graph construction, event-based reasoning, cross-modal fusion, multimodal injector and story ending generation. The details of our method are shown in Figure 2. Lastly, details about objectives and training are elaborated.

### 3.1 Event Graph Construction

**Semantic Event Graph.** The story plot contains multiple events which are correlated with each other. The definition of an event is a text span composed of a predicate and its arguments (Zhang et al., 2020). The event-centric reasoning shows excellent capability for context understanding and

subsequent event prediction (Zhou et al., 2022b). To effectively reason and mine more implicit information from story plots, we use semantic role labeling (SRL) to parse the story and extract events from parsing results, as shown in Figure 2. Specifically, Given story plots  $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4\}$ , we construct semantic event graphs  $\mathcal{G}_i^s = (\mathcal{V}_i^s, \mathcal{E}_i^s)$  by SRL.  $\mathcal{E}_i^s$  consists of two vectors, one for the positive direction and one for the opposite direction, and  $\mathcal{V}_i^s = \{s_0^i, s_1^i, s_2^i, \dots, s_n^i\}$ . To obtain features of each node, we use a pre-trained transformer encoder to obtain token representations in sentence  $\mathcal{S}_i$ .

$$\mathbf{T}_i = \text{Trans-Enc}(\mathcal{S}_i), \mathbf{T}_i \in \{t_i^1, t_i^2, \dots, t_i^g\} \quad (1)$$

where  $t_i^g$  denotes token representation, and  $g$  is length of sentence  $\mathcal{S}_i$ . Next, we conduct a mean pooling operation for tokens presentations based on SRL parsing result  $\hat{\mathcal{S}}_i$  to get presentation  $\hat{s}_j^i$  for each node. In addition, we take pooling for all token presentations of sentence  $\mathcal{S}_i$  to obtain a presentation of sentence node  $\hat{s}_0^i$ . Each node  $\hat{s}_j^i$  in sentence  $\mathcal{S}_i$  is connected to the sentence node. To preserve the relationship between sequences, we connect sentence nodes in the order of the sequence.

**Visual Event Graph.** For ending images, previous works (Huang et al., 2021; Xue et al., 2022) use pre-trained convolutional neural networks (CNN) to extract feature maps directly. We construct visual event graphs to reason and mine more implicit information from ending images. Scene

graphs have been used for many tasks to produce structured graph representations of visual scenes (Zellers et al., 2018). Inspired by the success of these tasks, we parse the ending image  $I$  to a scene graph via the scene graph parser. A scene graph can be denoted as a tuple  $\mathcal{G}^I = \{\mathcal{V}^I, \mathcal{E}^I\}$ , where  $\mathcal{V}^I = \{v_0, v_1, v_2, \dots, v_k\}$  is a set of  $k$  detected objects.  $v_0$  denotes a representation of the whole image, and other  $v_i$  is a region representation of detected object.  $\mathcal{E}^I = \{e_1, e_2, \dots, e_m\}$  is a set of directed edges and each edge  $e_i$  refers to a triplet  $(v_i, r_{i,j}, v_j)$ , which includes two directional edges from  $v_i$  to  $r_{i,j}$  and from  $r_{i,j}$  to  $v_j$ . Specifically, the construction of the scene graph can be divided into two parts: one is object detection, and the other is visual relation detection.

For object detection, we leverage a well-trained object detector, Faster-RCNN (Ren et al., 2017) with a ResNet-152 (He et al., 2016) backbone, to classify and encode objects in the ending image  $I$ . The outputs of detector include a set of region representations  $\mathcal{V}^I = \{v_1, v_2, \dots, v_k\}$  and object categories  $\mathcal{O} = \{o_1, o_2, \dots, o_k\}$ . For visual relation detection, we leverage MOTIFS (Zellers et al., 2018) as our relation detector to classify the relationship between objects. We train the relation detector on Visual Genome dataset (Krishna et al., 2017). The output of relation detector is a set of relation  $\mathcal{E}^I = \{e_1, e_2, \dots, e_m\}$ , where  $e_i$  refers to a triplet  $(v_i, r_{i,j}, v_j)$ . Lastly, we obtain the scene graph  $\mathcal{G}^I = \{\mathcal{V}^I, \mathcal{E}^I\}$  of ending image by combining the results of object detection and relationship detection.

### 3.2 Event-based Reasoning

We perform graph-structure reasoning over semantic and visual event graphs to effectively reason and mine more implicit information from story plots and ending images. Since event graphs have multiple relations between nodes (e.g., relations between visual objects, relations between predicates and arguments, etc.), we select relational graph convolutional networks (RGCN), which can pass different messages along different relations. Specifically, for each layer  $l$  in  $L$ -layer RGCN, the node representation  $w_i^l$  is updated as follows:

$$w_i^{l+1} = \text{ReLU} \left( \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_r(i)} \frac{1}{|\mathcal{N}_r(i)|} \mathbf{W}_r \cdot w_j^l \right) \quad (2)$$

where  $\mathcal{R}$  denote a set of all edges types, and  $\mathcal{N}_r(i)$  is the neighborhood of node  $i$  under relation  $r$ .

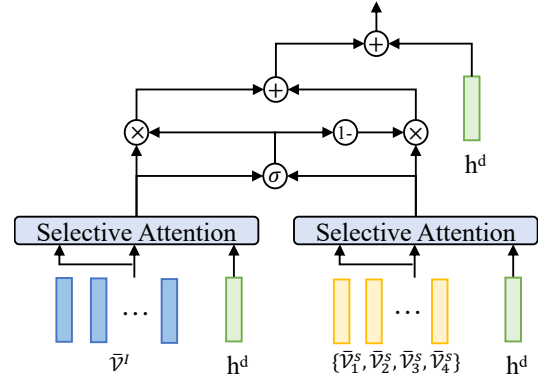


Figure 3: Details of the multimodal injector.

To reason and mine more implicit information in single-modality, we conduct event-based reasoning on semantic and visual event graphs, respectively.

### 3.3 Cross-modal Fusion

We propose cross-modal fusion for visual and semantic event graphs to integrate information from story plots and ending images. We adopt a layer normalization for node features to reduce the cross-modal gap between visual and semantic graphs. For cross-modal feature fusion, previous works (Huang et al., 2021; Xue et al., 2022) adopt attention-based neural network models to implicitly integrate multimodal features. However, these models neglect the dependency between single-modal features. Therefore, we maintain graph structure for visual and semantic features and connect nodes that represent whole image and sentences, as shown in Figure 2. Moreover, we utilize RGCN as Eq.2 to integrate cross-modal features in event graph, and outputs denote as  $\bar{V}_i^s = \{\bar{s}_0^i, \bar{s}_1^i, \bar{s}_2^i, \dots, \bar{s}_n^i\}$  and  $\bar{V}^I = \{\bar{v}_0, \bar{v}_1, \bar{v}_2, \dots, \bar{v}_k\}$ .

### 3.4 Multimodal Injector

To integrate different modal sources, we propose a multimodal injector, which adaptly extracts key information from different modal features and integrates them appropriately. As shown in Figure 3, inputs of multimodal injector include a hidden state  $h^d$  from the decoder, visual features  $\bar{V}^I$  and semantic features  $\bar{V}_i^s$ . Specifically, we first use selective attention for key information extraction, i.e.,

$$h_{attn}^u = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V, u \in \{I, S\} \quad (3)$$

where  $Q$  is  $h^d$  from decoder;  $K$  and  $V$  are visual features  $\bar{V}^I$  or semantic features  $\bar{V}_i^s$ ; and  $d_k$  is the

same as the dimension of  $\mathbf{h}^d$ . Then, the gate  $\lambda \in [0, 1]$  and the fused output are defined as:

$$\lambda = \sigma(\mathbf{U}\mathbf{h}_{attn}^I + \mathbf{V}\mathbf{h}_{attn}^S) \quad (4)$$

where  $\mathbf{U}$  and  $\mathbf{V}$  are trainable weights.  $\lambda$  controls how much visual information is attended.

$$\hat{\mathbf{h}}^d = \lambda \cdot \mathbf{h}_{attn}^I + (1 - \lambda) \cdot \mathbf{h}_{attn}^S + \mathbf{h}^d \quad (5)$$

where the fusion vector  $\hat{\mathbf{h}}^d$  is fed into the decoder.

### 3.5 Story Ending Generation

Recently, Transformer (Vaswani et al., 2017) shows its powerful ability to generate natural language (Radford et al., 2019). For story ending generation, we use a Transformer decoder as the decoder for our model. Specifically, the decoder input includes a segment of the generated story ending  $\bar{\mathbf{C}}$  and fusion vector  $\hat{\mathbf{h}}^d$  from the multimodal injector. The purpose of the decoder is to predict a probability distribution of the next word of the segment  $\bar{\mathbf{C}}$ , i.e.,

$$\mathbf{h}_i = \text{Trans-Dec}(\hat{\mathbf{h}}^d, \bar{\mathbf{C}}) \in \mathbb{R}^d$$

where  $\bar{\mathbf{C}} = [c_1, \dots, c_{i-1}]$  (6)

$$\mathbf{p}_i = \text{LM-Head}(\mathbf{h}_i) \in \mathbb{R}^V \quad (7)$$

where  $\mathbf{h}_i$  refers to the hidden representation in  $i$ -th step;  $V$  denotes token vocabulary and  $\mathbf{p}_i$  refers to a probability distribution over  $\mathcal{V}$ ;  $d$  in  $\hat{\mathbf{h}}^d$  denotes the current number of layer. Lastly, the story ending generation objective is defined as a maximum likelihood estimation. The loss function is defined as:

$$\mathcal{L}^{(gen)} = -\frac{1}{|N|} \sum_{i=1}^N \log \mathbf{p}_i(c_i), \quad (8)$$

where  $\mathbf{p}_i(c_i)$  denotes fetching the probability of the  $i$ -th step gold token  $c_i \in \mathbf{C}$  from  $\mathbf{p}_i$ .  $\mathbf{C}$  refers to the gold caption, and  $N$  is its length.

### 3.6 Incoherence Detection

To enhance the understanding context of a story plot and robustness of graph modeling for our model, we introduce a training objective: incoherence detection. We set a 10% probability to replace a whole sentence node in semantic event graph randomly. In the objective, the final step output  $\mathbf{h}_n$  of the decoder is passed into a MLP to classify whether each whole sentence node is changed, i.e.,

$$\mathbf{p}^{clf} = \sigma(\text{MLP}(\mathbf{h}_n)) \in \mathbb{R}^4 \quad (9)$$

where  $\sigma$  denotes a sigmoid function. The loss function is defined as:

$$\mathcal{L}^{(clf)} = -\frac{1}{4} \sum_{i=1}^4 y_i \cdot \log(\mathbf{p}_i^{clf}) + (1 - y_i) \cdot \log(1 - \mathbf{p}_i^{clf}) \quad (10)$$

### 3.7 Training

In model training, we set a trade-off parameter  $\alpha$  for two losses  $\mathcal{L}^{(gen)}$  and  $\mathcal{L}^{(clf)}$ . The total loss function of our model is definite as follows:

$$\mathcal{L} = \mathcal{L}^{(gen)} + \alpha \times \mathcal{L}^{(clf)} \quad (11)$$

## 4 Experiment

### 4.1 Dataset and Evaluation Metric

**VIST-Ending.** We compare our model and other state-of-the-art methods on the VIST-Ending (VIST-E) dataset (Huang et al., 2021). The dataset is built over VIST dataset (Huang et al., 2016). The VIST-E dataset comprises 39,920 samples for training, 4,963 samples for validation and 5,030 samples for testing. In experiments, we follow the data split in (Huang et al., 2021).

**LSMDC-Ending.** LSMDC-Ending (LSMDC-E) (Xue et al., 2022) contains 20,151 training samples, 1,477 validation samples and 2,005 test samples, which are collected from LSMDC 2021 (Rohrbach et al., 2017).

**Visual Genome.** We use the Visual Genome (VG) dataset to train a visual relationship detector. The dataset includes 108,077 images annotated with scene graphs, and we follow the setting in (Xu et al., 2017), which contains 150 object classes and 50 relation classes.

**Evaluation Metric.** As follow Xue et al. (2022), we utilize the same metrics to report evaluation results, and the evaluation code is open-source<sup>1</sup>. The evaluation metrics include: BLEU (Kingma and Ba, 2015), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), ROUGE-L (Lin, 2004) and Result Sum (rSUM) (Xue et al., 2022).

### 4.2 Implementation Details

For the scene graph, we limit the maximum number of objects to 10 and the maximum number of relationships to 20. The relational graph convolution network includes four relational graph convolution

<sup>1</sup><https://github.com/tylin/coco-caption>

Method	B@1	B@2	B@3	B@4	M	R-L	C	rSUM
Seq2Seq (Luong et al., 2015)	13.96	5.57	2.94	1.69	4.54	16.84	12.04	57.58
Transformer (Vaswani et al., 2017)	17.18	6.29	3.07	2.01	6.91	18.23	12.75	66.44
IE+MSA (Guan et al., 2019)	19.15	5.74	2.73	1.63	6.59	20.62	15.56	72.02
T-CVAE (Wang and Wan, 2019)	14.34	5.06	2.01	1.13	4.23	15.51	11.49	53.77
MG+Trans (Huang et al., 2021)	19.43	7.47	3.92	2.46	7.63	19.62	14.42	74.95
MG+CIA (Huang et al., 2021)	20.91	7.46	3.88	2.35	7.29	21.12	19.88	82.89
MGCL (Huang et al., 2021)	22.57	8.16	4.23	2.49	7.84	21.66	21.46	88.41
MMT (Xue et al., 2022)	22.87	8.68	4.38	2.61	15.55	23.61	25.41	103.11
MET (Ours)	<b>24.31</b>	<b>8.79</b>	<b>4.62</b>	<b>2.73</b>	<b>16.41</b>	<b>24.49</b>	<b>26.47</b>	<b>107.82</b>

Table 1: Comparison results on VIST-E test set. B@n, M, R-L, C and rSUM denote BLEU@n, METEOR, ROUGE-L, CIDEr and Result Sum, respectively.

Method	B@1	B@2	B@3	B@4	M	R-L	C	rSUM
Seq2Seq (Luong et al., 2015)	14.21	4.56	1.70	0.70	11.01	19.69	8.69	60.56
Transformer (Vaswani et al., 2017)	15.35	4.49	1.82	0.76	11.43	19.16	9.32	62.33
MGCL (Huang et al., 2021)	15.89	4.76	1.57	0.00	11.61	20.30	9.16	63.29
MMT (Xue et al., 2022)	18.52	5.99	2.51	1.13	12.87	20.99	12.41	74.42
MET (Ours)	<b>19.98</b>	<b>6.48</b>	<b>2.89</b>	<b>1.77</b>	<b>14.53</b>	<b>22.73</b>	<b>13.85</b>	<b>82.23</b>

Table 2: Comparison results on LSMDC-E test set.

layers, and the size of input and output sets of 768. For semantic event reasoning, we use a pre-trained BERT model (Devlin et al., 2019) as the language model. The layers and attention heads of the decoder are 12 and 8. The dimension of embedding vectors in the decoder is 768, and the dimension of hidden states is 768. The visual feature encoder is ResNet-152. For model training, we select the Adam optimizer (Kingma and Ba, 2015) to optimize the model with learning rate of  $2e-4$ . The maximum training epoch of our model is 25. The trade-off parameter  $\alpha$  in Eq.11 is 0.2. The batch size, weight decay and warm-up proportion are 128, 0.01 and 0.1. During inference, we use the beam search with a beam size of 3 to generate a story ending with maximum sentence length is 25. Our model is trained on one V100 GPU.

### 4.3 Baselines

We compare our model with following state-of-the-art baselines: (1) **Seq2Seq** is a stack RNN-based model (Luong et al., 2015) with attention mechanisms, and image features are directly concatenated. (2) **Transformer**, proposed by Vaswani et al. (2017), is an encoder-decoder model with self-attention mechanisms. (3) **IE+MSA** is a story ending generation model incorporating external

knowledge (Guan et al., 2019). (4) **T-CVAE** (Wang and Wan, 2019) is a conditional variational autoencoder based on transformer for missing story plots generation. (5) **MG+Trans** consists of multi-layer graph convolutional networks and a transformer decoder (Huang et al., 2021). (6) **MG+CIA** consists of multi-layer graph convolutional networks, top-down LSTM and one context-image attention unit in the decoder (Huang et al., 2021). (7) **MGCL** is an image-guided story ending generation model with multi-layer graph convolution networks and cascade-LSTM (Huang et al., 2021). (8) **MMT** is a multimodal memory transformer for image-guided story ending generation (Xue et al., 2022).

### 4.4 Main Results

The experimental results on VIST-E and LSMDC-E are shown in Table 1 and Table 2. From the tables, we can make two observations. Firstly, our model achieves state-of-the-art performance on the VIST-E and LSMDC-E datasets compared to other strong competitors. In addition, MG+CIA, MGCL, MMT and our model significantly and consistently outperform other models that directly concatenate visual features. It indicates that mining visual information is essential and can provide rich information to predict the ending. Moreover, our model

Method	B@1	B@2	B@3	B@4	M	R-L	C	rSUM
MET	<b>24.31</b>	<b>8.79</b>	<b>4.62</b>	<b>2.73</b>	<b>16.41</b>	<b>24.49</b>	<b>26.47</b>	<b>107.82</b>
w/o ID	23.84	8.70	4.51	2.56	15.91	24.10	25.86	105.48
w/o CMF	23.47	8.65	4.47	2.53	15.91	23.85	25.66	104.54
w/o MI	22.68	8.56	4.33	2.48	15.83	22.99	24.74	101.61
w/o VER	22.41	8.25	4.33	2.50	15.86	23.09	25.03	101.47
w/o SER	23.78	8.73	4.46	2.55	15.88	24.04	25.87	105.31
w/o CMF, MI	21.03	8.03	4.16	2.36	15.43	21.14	22.44	94.59

Table 3: Ablation study. “w/o ID” denotes removing the incoherence detection objective; “w/o CMF” denotes removing the cross-modal fusion; “w/o MI” denotes removing the multimodal injector; “w/o VER” denotes removing the event-based reasoning in visual event graph; “w/o SER” denotes removing the event-based reasoning in semantic event graph; “w/o CMF, MI” removing the cross-modal fusion and multimodal injector.

Method	B@1	B@2	B@4	M	R-L
Seq2Seq	14.27	4.27	1.05	6.02	16.32
Transformer	17.06	6.18	1.57	6.55	18.69
IE+MSA	20.11	6.62	1.68	6.87	<u>21.27</u>
T-CVAE	20.36	6.63	1.88	6.74	<u>20.98</u>
Plan&Write	20.92	5.88	1.44	7.10	20.17
KE-GPT2	<b>21.92</b>	<b>7.40</b>	1.90	<u>7.41</u>	20.58
MG+Trans	18.55	6.76	<u>2.33</u>	<u>7.31</u>	19.02
MGCL	20.27	6.26	1.81	6.91	21.01
MET	<u>21.88</u>	<u>7.28</u>	<b>2.36</b>	<b>7.41</b>	<b>21.32</b>

Table 4: Result of the SEG task on the VIST-E dataset (plain text). The bold / underline denotes the best and the second performance, respectively.

achieves better results than MG+CIA, MGCL and MMT, demonstrating that reasoning and mining implicit information from story plots and ending image is significant for image-guided story ending generation.

#### 4.5 Ablation Study

To verify the effectiveness of our method, we conduct an ablation study and show the results in Table 3. Firstly, the table shows that removing each component or objective decreases the model performance, which demonstrates our method’s effectiveness. In addition, we observe that removing cross-modal fusion and multimodal injector brings a great performance drop, which shows that cross-modal information mining and adaptive integration play a crucial role in story ending prediction.

#### 4.6 SEG Setting

To investigate the effectiveness of visual information mining in our method, we remove the image

from the VIST-E dataset and evaluate it on only plain text. The results are shown in Table 4. From the table, we observe that our model keeps competitive with Plan&Write (Yao et al., 2019) and KE-GPT2 (Guan et al., 2020) models designed especially for textual story generation. Moreover, our model outperforms MG+trans, which verifies the effectiveness of our incoherence detection and semantic event-based reasoning. Our model performs better when adding the image, as shown in Table 1. It demonstrates that mining implied visual information can help story ending generation.

#### 4.7 Analysis

##### 4.7.1 Impact of Event-based Reasoning

To investigate the effectiveness of event reasoning, we analyze its impact, and the results are shown in Table 5. From the table, we can observe that replacing semantic role labeling with dependency parsing leads to decreased performance. Moreover, replacing the visual event graph with whole image features (i.e., features extracted by pre-trained CNN) shows a performance drop. In addition, removing cross-modal fusion also shows a performance drop. These demonstrate the effectiveness of event-based reasoning for the image-guided story ending generation.

##### 4.7.2 Case Study

To extensively evaluate our method, we conduct a case study for our model and MGCL, and some random sampling examples are shown in Figure 4. For example, in the left case, we can observe that our model can reason that the man in the image is a soldier, while the result from MGCL is not significantly related to visual content. For example, in the right case, our model can generate the word "relax"

Method	B@1	B@2	B@3	B@4	M	R-L	C	rSUM
MET	<b>24.31</b>	<b>8.79</b>	<b>4.62</b>	<b>2.73</b>	<b>16.41</b>	<b>24.49</b>	<b>26.47</b>	<b>107.82</b>
w/ Dependence Parser	23.47	8.70	4.50	2.57	15.88	24.15	24.06	103.33
w/o Visual Event Graph	22.13	8.19	4.21	2.44	15.76	22.88	23.93	99.54
w/o CMF	23.47	8.65	4.47	2.53	15.91	23.85	25.66	104.54

Table 5: Impact of event reasoning. “w/ *Dependency Parser*” denotes replacing semantic role labeling with dependency parsing; “w/o *Visual Event Graph*” denotes removing the visual event graph and provides the whole image features as inputs; “w/o *CMF*” denotes removing the cross-modal fusion.



<p><b>Story Plot:</b> i went to the award ceremony yesterday . there were a lot of people there . everyone received an award for their effort . they had a great time .</p>	<p><b>Story Plot:</b> the day of our family vacation finally arrived. we made out way down to the lake after leaving our belongings in the lodge. there were a lot of other people out on the river. they really looked like they were having fun as well.</p>
<p><b>Ending Image:</b> </p>	<p><b>Generated Story Ending:</b> <b>MGCL:</b> we ended the day with a great time . <b>MET:</b> the soldiers were singing together at the end of the ceremony .</p>
<p><b>Ending Image:</b> </p>	<p><b>Generated Story Ending:</b> <b>MGCL:</b> at the end of the day , we ready to take a picture . <b>MET:</b> after we go home , we decided to take a relax in chair .</p>

Figure 4: Random sampling examples generated by MET and MGCL.



Figure 5: Interpretable visualization analysis of our method (better viewed in color).

based on the objects "human" and "chair". It shows that our model can mine the implicit information based on visual and semantic information.

#### 4.7.3 Interpretable Visualization Analysis

To investigate the effectiveness of the multimodal injector, we conduct an interpretable visualization analysis. The results are shown in Figure 5. The word with a blue underline denotes that the multimodal injector is assigned the greater probability in the node of visual event graph. Green corresponds to greater probability in the node of semantic event graph. The dotted boxes below represent the specific content of nodes. From the results, we can observe that nodes in visual and semantic event

Method	Gram.	Logic.	Rele.
MET	<b>3.49</b>	<b>3.37</b>	<b>2.94</b>
MGCL	3.36	3.15	2.66
MG+Trans	3.22	2.78	2.71

Table 6: Human evaluation.

graphs are able to deduce implicit information.

#### 4.7.4 Human Evaluation

To evaluate our method more comprehensively, we conducted a human evaluation to compare further the performance of our model and MGCL and MG+trans. As follow [Huang et al. \(2021\)](#), we considered three metrics for the story ending generated by models: Grammaticality (Gram.) ([Wang and Wan, 2019](#)) evaluates correctness, natural, and fluency of story endings; Logicity (Logic.) ([Wang and Wan, 2019](#)) evaluates reasonability and coherence of story endings; Relevance (Rele.) ([Yang et al., 2019](#)) measures how relevant between images and generated story endings. We randomly sampled 100 samples from the test set and display them to 3 recruited annotators. Thereby, each annotator worked on 300 items from 3 models. We show 3 annotators all outputs from all 3 models at once and shuffle the output-model correspondence to ensure that annotators do not know which model the output is predicted from. Following [Yang et al.](#)



(2019), we set a 5-grade marking system, where one is the worst grade, and five is the maximum. The results show that the performance of our model is significantly better than MGCL and MG+trans. That is, our model can generate higher-quality story endings.

## 5 Conclusion

In this work, we propose a multimodal event transformer, a framework for image-guided story ending generation. Our method includes event graph construction, event-based reasoning, cross-model fusion, multimodal injector and story ending generation. Different from previous work, our method not only focuses on cross-modal information fusion but also on reasoning and mining implicit information from single-modality data. In addition, we propose an incoherence detection to enhance the understanding context of a story plot and robustness of graph modeling for our model. In the experiments, results show that our method delivers state-of-the-art performance.

## Limitations

Although our proposed method can effectively reason and mine implicit information from story plots and ending image, it suffers from weaknesses in integrating cross-modal information. Specifically, our method connects visual and semantic event graphs by connecting whole image nodes and whole sentence nodes. It lacks fine-grained information to pass between semantic events to visual objects. In further work, we will study how to pass fine-grained information between visual and semantic event graphs.

## References

- Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J. Martin, and Mark O. Riedl. 2020. [Story realization: Expanding plot events into sentences](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7375–7382. AAAI Press.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: an automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.
- Hong Chen, Yifei Huang, Hiroya Takamura, and Hideki Nakayama. 2021. [Commonsense knowledge aware concept selection for diverse and informative visual storytelling](#). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 999–1008. AAAI Press.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Angela Fan, Mike Lewis, and Yann N. Dauphin. 2019. [Strategies for structuring story generation](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2650–2660. Association for Computational Linguistics.
- Jian Guan, Fei Huang, Minlie Huang, Zhihao Zhao, and Xiaoyan Zhu. 2020. [A knowledge-enhanced pretraining model for commonsense story generation](#). *Trans. Assoc. Comput. Linguistics*, 8:93–108.
- Jian Guan, Yansen Wang, and Minlie Huang. 2019. [Story ending generation with incremental encoding and commonsense knowledge](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6473–6480. AAAI Press.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep residual learning for image recognition](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.
- Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. [Deep semantic role labeling: What works and what’s next](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 473–483. Association for Computational Linguistics.

- Qingbao Huang, Chuan Huang, Linzhang Mo, Jielong Wei, Yi Cai, Ho-fung Leung, and Qing Li. 2021. [Igseg: Image-guided story ending generation](#). In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3114–3123. Association for Computational Linguistics.
- Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. [Visual storytelling](#). In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1233–1239. The Association for Computational Linguistics.
- Yunjae Jung, Dahun Kim, Sanghyun Woo, Kyungsu Kim, Sungjin Kim, and In So Kweon. 2020. [Hide-and-tell: Learning to bridge photo streams for visual storytelling](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11213–11220. AAAI Press.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. [Visual genome: Connecting language and vision using crowdsourced dense image annotations](#). *Int. J. Comput. Vis.*, 123(1):32–73.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. [Effective approaches to attention-based neural machine translation](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.
- Mengshi Qi, Jie Qin, Di Huang, Zhiqiang Shen, Yi Yang, and Jiebo Luo. 2021. [Latent memory-augmented graph transformer for visual storytelling](#). In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 4892–4901. ACM.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. [Faster R-CNN: towards real-time object detection with region proposal networks](#). *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149.
- Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Joseph Pal, Hugo Larochelle, Aaron C. Courville, and Bernt Schiele. 2017. [Movie description](#). *Int. J. Comput. Vis.*, 123(1):94–120.
- Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. [Modeling relational data with graph convolutional networks](#). In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. [Cider: Consensus-based image description evaluation](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.
- Tianming Wang and Xiaojun Wan. 2019. [T-CVAE: transformer-based conditioned variational autoencoder for story completion](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5233–5239. ijcai.org.
- Xin Wang, Wenhui Chen, Yuan-Fang Wang, and William Yang Wang. 2018. [No metrics are perfect: Adversarial reward learning for visual storytelling](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 899–909. Association for Computational Linguistics.
- Chunpu Xu, Min Yang, Chengming Li, Ying Shen, Xiang Ao, and Ruifeng Xu. 2021. [Imagine, reason and write: Visual storytelling with graph knowledge and relational reasoning](#). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3022–3029. AAAI Press.

- Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. 2017. [Scene graph generation by iterative message passing](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3097–3106. IEEE Computer Society.
- Dizhan Xue, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2022. [Mmt: Image-guided story ending generation with multimodal memory transformer](#). In *Proceedings of the 30th ACM International Conference on Multimedia, MM '22*, page 750–758, New York, NY, USA. Association for Computing Machinery.
- Pengcheng Yang, Fuli Luo, Peng Chen, Lei Li, Zhiyi Yin, Xiaodong He, and Xu Sun. 2019. [Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5356–5362. ijcai.org.
- Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. [Plan-and-write: Towards better automatic storytelling](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7378–7385. AAAI Press.
- Youngjae Yu, Jiwan Chung, Heeseung Yun, Jongseok Kim, and Gunhee Kim. 2021. [Transitional adaptation of pretrained models for visual storytelling](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12658–12668. Computer Vision Foundation / IEEE.
- Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. [Neural motifs: Scene graph parsing with global context](#). In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5831–5840. Computer Vision Foundation / IEEE Computer Society.
- Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. [ASER: A large-scale eventuality knowledge graph](#). In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2.
- Yucheng Zhou, Xiubo Geng, Tao Shen, Guodong Long, and Daxin Jiang. 2022a. [Eventbert: A pre-trained model for event correlation reasoning](#). In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 850–859. ACM.
- Yucheng Zhou, Tao Shen, Xiubo Geng, Guodong Long, and Daxin Jiang. 2022b. [ClarET: Pre-training a correlation-aware context-to-event transformer for event-centric generation and classification](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2559–2575, Dublin, Ireland. Association for Computational Linguistics.