Theia: Weakly Supervised Multimodal Event Extraction from Incomplete Data

Farhad Moghimifar[1] and **Fatemeh Shiri**[1] and **Van Nguyen**[2]
Yuan-Fang Li[1] and **Gholamreza Haffari**[1]
[1] Faculty of Information Technology, Monash University, Australia
[2] Information Sciences Division, Defence Science and Technology Group, Australia
{first.lastname}@monash.edu
van.nguyen5@defence.gov.au

Abstract

Event extraction from multimodal documents is an important yet under-explored problem. One challenge faced by this task is the scarcity of paired image-text datasets, making it difficult to fully exploit the strong representation power of multimodal language models. In this paper, we present Theia, an end-to-end multimodal event extraction framework that can be trained on incomplete data. Specifically, we couple a generation-based event extraction model with a customised image synthesizer that can generate images from text. Our model leverages capabilities of pre-trained vision-language models and can be trained on incomplete (i.e. text-only) data. Experimental results on existing multimodal datasets demonstrate the effectiveness of our approach for both synthesising missing data and extracting events over state-of-the-art approaches.

1 Introduction

Event extraction is an important task in natural language processing that aims to identify and extract structured information about events and their arguments from text. Despite the challenging nature of this task, being rooted in the ambiguity, complexity and diversity of natural language, recent years have seen significant improvements from end-to-end deep learning models (Wadden et al., 2019; Du and Cardie, 2020; Hsu et al., 2022) over the traditional rule-based approaches (Valenzuela-Escárcega et al., 2015; Bui et al., 2013).

There is a rapid increase of *multimodal* documents online, propelled by the high prevalence of camera-enabled devices. It has been shown that other modalities supplement the information that is available in text (Li et al., 2020). However, event extraction from multimodal information is an under-explored area, as existing methods were primarily developed for textual information.

Some recent works studied this question through the use of the visual modality (Li et al., 2022a, 2020). These approaches, however, suffer from two shortcomings. Firstly, existing models formulate event extraction as a classification problem, in which trigger words and entity recognition modules are used as features. This limits their performance in capturing high-level complex event structures. Secondly, these methods require a complete set of text-(corresponding)image pairs at the time of training, which limits their performance when models face *missing* data, where images are unavailable for all or a portion of the training data.

To overcome these shortcomings, we propose an end-to-end sequence-generation-based multimodal event extraction model. Our proposed approach incorporates an image synthesizer model to handle the missing images during training. The image synthesizer, conditioned on the given textual information, generates a visual representation that helps to train the encoder-decoder structure of our event extraction model, and customises the generated images by reducing the domain shift between the original domain of the pre-trained models and the target domain (Figure 1). Experiments on the task of multimodal event extraction confirm the strong superiority of our model, by a margin of 7%, over state-of-the-art models. Furthermore, empirical results using the images synthesise show the capability of our model to portray domain-dependant visual context.

2 Related Work

The task of event extraction has been widely studied (Ahn, 2006; Hogenboom et al., 2011), where initially rule-based classification approaches were developed to address this task (Bui et al., 2013; Ritter et al., 2012). With the advances in deep learning, models based on neural networks have also been developed for this task (Nguyen and Nguyen, 2019; Zhang et al., 2019). More recently, several studies leverage the strong representation and reasoning capabilities of pre-trained language models (Li et al.,

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 139–145
November 1–4, 2023.

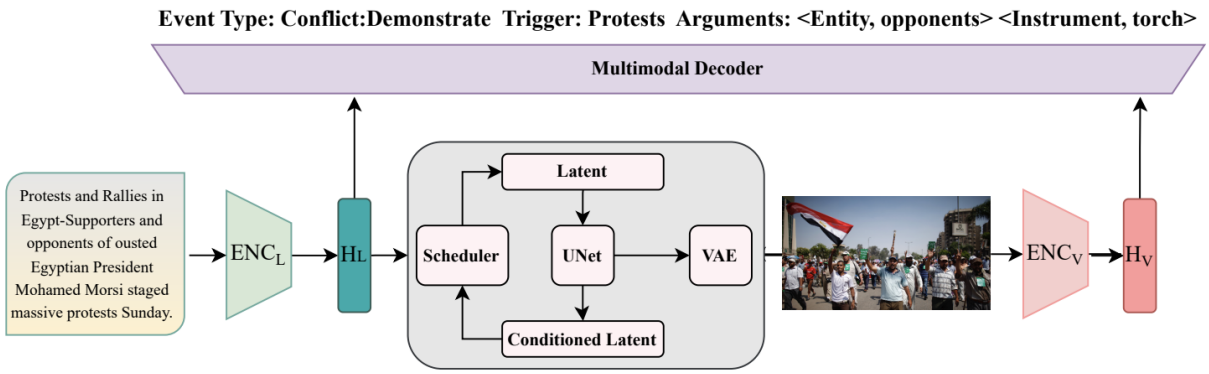

Figure 1: An overview of our proposed model. The textual encoder ($\mathbf{ENC}_L$) takes the text as input and feeds the encoded version to the image synthesizer, as well as the decoder. The visual encoder ($\mathbf{ENC}_V$) takes the synthesised image as input and feeds the encoded image into the multimodal transformer-based decoder, and the structured event information is generated as a sequence.

2021; Wu et al., 2022; Hsu et al., 2022).

While these studies achieve excellent performance on text-only benchmarks, they fail to account for other modalities. By proposing a new dataset and a multimodal event extraction model, Li et al. (2020); Zhang et al. (2017); Moghimifar et al. (2023) showed that including the visual context results in better extraction performance. Li et al. (2022a) proposed an approach based on pre-trained vision-language models (Radford et al., 2021) for addressing multimodal event extraction. However, these approaches cast event extraction as a classification problem, whereby their model is limited to a fixed schema. Furthermore, these models require a complete set of text-image pairs for training, which hinders their real-world practicality where it is highly likely that a part of the visual data is missing. Unlike these approaches, we propose a sequence generation model that is capable of synthesising the visual context to alleviate the problem of *missing* data.

3 Multimodal Event Extraction

Problem Formulation. Given a corpus of sequences of tokens $\mathbb{X} = \{x_1, \ldots, x_N\}$ of length N, where each $x_i = (w_i^1, w_i^2, \ldots)$ represents the i-th sentence of the corpus and w_i^m is a token from the vocabulary $\mathcal{W}$. In addition, there is the corresponding visual context $\mathbb{V} = \{v_1, \ldots, v_M\}$ of length M, where $M < N$, and for each $v \in \mathcal{V}$, it is paired with one sentence $x \in \mathcal{X}$. Our goal is to extract event mentions, including the event type $e \in \mathcal{E}$, and roles $r \in \mathcal{R}$ together with their arguments $a \in \mathcal{W}$.

Multimodal Event Generation. We formulate the task of multimodal event extraction as a sequence generation task, where our proposed model outputs a linearised representation of event mentions in a given sentence-image pair (x_i, v_i). Each event then is represented in the form of $y_i = \langle t_i, e_i, \langle a_i, r_i \rangle_1, \langle a_i, r_i \rangle_2, \ldots \rangle$, where $t_i \in W$ refers to a trigger word, $e_i \in \mathcal{E}$ indicates the event type, and $a_i \in W$ and $r_i \in \mathcal{R}$ represents an argument token and role, respectively. Thus, given an input pair (x, v), the goal of our proposed model is to generate sequence y of length T as follows:

$$p_{\theta_{\mathbf{SEQ}}}(y|x,v) = \prod_{t=1}^{T} p_{\theta_{\mathbf{SEQ}}}(y_t|y_{<t}, x, v) \quad (1)$$
$$= \prod_{t=1}^{T} \mathbf{DEC}_{\theta_{\mathbf{DEC}}}(y_t|y_{<t}, \mathbf{ENC}_{\theta_{\mathbf{ENC}}}(x,v)),$$

where $\mathbf{ENC}_{\theta_{\mathbf{ENC}}} = (\mathbf{ENC}_L, \mathbf{ENC}_V)$ and $\mathbf{DEC}_{\theta_{\mathbf{DEC}}}$ refers to an encoder and a decoder structure, respectively, and $\theta_{\mathbf{SEQ}} := \{\theta_{\mathbf{ENC}}, \theta_{\mathbf{DEC}}\}$ denotes the parameters of our sequence generation model.

Multimodal Architecture. The encoder $\mathbf{ENC}(\cdot)$ consists of a language encoder ($\mathbf{ENC}_L$) and a visual encoder ($\mathbf{ENC}_V$) to compute hidden representations of the textual ($\boldsymbol{H}_L$) and visual ($\boldsymbol{H}_v$) inputs separately. Hence, the hidden representation ($\boldsymbol{H}$) of the input (x, v) is formulated as:

$$\boldsymbol{H} = \mathbf{ENC}(x,v) = [\mathbf{ENC}_L(x); \mathbf{ENC}_V(v)] \quad (2)$$

The $\mathbf{DEC}(\cdot)$ is a multi-layer Transformer-based decoder, where each layer is a Transformer block.

At step t, the decoder generates the t-th token of sequence y_i and the hidden state $\boldsymbol{h}_t$ as follows:

$$y_t, \boldsymbol{h}_t = \textbf{DEC}(y_{t-1}; \boldsymbol{H}_{y<t}, \boldsymbol{H}), \quad (3)$$

where $\boldsymbol{H}_{y<t} \in \mathbb{R}^{(t-1)\times d}$ denotes the past hidden state used for self-attention during decoding. Given the *complete* data (x, v, y), the **ENC-DEC** architecture can be trained by minimising the loss:

$$\mathcal{L}_{seq} = \mathbb{E}_{x,v,y}[-\log p(y|x, v; \theta_{seq})] \quad (4)$$

4 Training with Incomplete Data

During training, when the visual context of an input x_i is unavailable, we leverage an *image synthesizer* that produces visual representation z_i based on x_i. We then adapt the image synthesizer based on the complete and incomplete event extraction data.

Our proposed method is based on Denoising Diffusion Probabilistic Models (Ho et al., 2020). This generative model consists of a pre-trained autoencoder that maps images to a spatial latent code, a corresponding decoder that learns to map the latent representation back to the image, and a diffusion model that is conditioned on the textual input (x_i). Inspired by Ruiz et al. (2022), during training, we use the textual input of complete pairs of $(x^{\text{com}}, v^{\text{com}})$ to condition the model to regenerate v^{com}. Furthermore, we use the textual input of incomplete data points x^{inc} to synthesise visual context z. We then resort to the following reconstruction-based loss function to train the image synthesiser on the synthesised image z,

$$\begin{aligned}\mathcal{L}_{syn} = \mathbb{E}_{x^{\text{com}},x^{\text{inc}},v^{\text{com}},z,\epsilon,\epsilon',t,t'} \quad (5)\\ [\omega_t||\hat{V}_{\theta_{syn}}(\alpha_t v + \sigma_t\epsilon, x^{com}) - v^{com}||_2^2 +\\ \lambda_{syn}\omega_{t'}||\hat{V}_{\theta_{syn}}(\alpha_{t'} z + \sigma_{t'}\epsilon', x^{inc}) - z||_2^2],\end{aligned}$$

where $\hat{V}_{\theta_{syn}}$ is the image synthesizer function, $\epsilon \sim \mathcal{N}(0, I)$ is a noise term, and $\omega_t, \alpha_t, \sigma_t$ and $\omega_{t'}, \alpha_{t'}, \sigma_{t'}$ are the terms that control the noise schedule and sample quality for complete and incomplete data, respectively, where t (and t') $\sim \mathcal{U}([0, 1])$. λ_{syn} controls the trade-off between the images synthesizer's capability to regenerate the v^{com} conditioned on x^{com} and synthesising images conditioned on x^{inc}. The only trainable parameters of the image synthesier are those of the textual encoder, and we keep the other parameters frozen.

We train our models in an end-to-end manner, where the overall optimisation objective is defined as a weighted sum of the sequence generation loss and the image synthesizer loss:

$$\mathcal{L} = \mathcal{L}_{seq} + \lambda\mathcal{L}_{syn}, \quad (6)$$

where the hyperparameter λ controls the trade-off between extracting textual and visual semantic information for sequence generation and synthesising the visual features.

To train our model, we apply online hard EM (Neal and Hinton, 1999) by interleaving the following steps in an iterative manner:

- **E-step**: generate images z for data points x^{inc} using the current parameters of our proposed image synthesizer θ_{syn}.
- **M-step**: the model parameters (θ_{seq} and θ_{syn}) are the updated by minimising the loss Eq. 6.

5 Experiments

In this section, we report the performance of our model on the task of multimodal event extraction in comparison to the current state-of-the-art models.

5.1 Experimental Setup

Evaluation Metrics. Following previous work on multimodal event extraction (Li et al., 2020), we report the results of the macro-averaged F1 score (F1), precision (P) and recall (R). Since we have formulated this as a sequence generation task, we also report the BLEU score.

Baselines. We compare our model, Theia (VL), against Valhalla (Li et al., 2022b), RMMT (Wu et al., 2021), and Gated Fusion (Wu et al., 2021)[1]. Furthermore, we report the performance of GPT3.5 (Brown et al., 2020) and Flamingo (Alayrac et al., 2022), as two in-context learning approaches. In order to use visual information, we use BLIP-2 (Li et al., 2023) to extract *scene description* from images and then we feed them to the context of the model (GPT3.5/SC). In addition, we report the performance of our model in two ablation settings, firstly, when the visual context is fully disregarded in training (Theia (L)), and secondly, instead of the image synthesizer, we use an image retrieval model to fill in the missing data with the most relevant images from the full set of images in the dataset (Theia (R)).

Dataset. **M2E2** (Li et al., 2020) is a multimodal news event extraction dataset that expands upon ACE (Doddington et al., 2004). It consists of 6,167 sentences, with 1,297 event mentions and 1,965

[1]The results of Li et al. (2022a, 2020) are not reproducible, due to incomplete source code/data.

		Event			Argument				
Model	**BLEU**	**F1**	**P**	**R**	**F1**	**P**	**R**	Size	Method
Gated Fusion (Wu et al., 2021)	13.34	21.45	21.38	22.50	12.24	12.36	13.59	21.7M	SL
RMMT (Wu et al., 2021)	10.81	26.41	36.12	23.96	11.67	12.90	12.37	127.9M	SL
Valhalla (Li et al., 2022b)	15.23	19.19	17.53	24.62	10.83	10.51	15.10	49.1M	SL
OpenFlamingo (Awadalla et al., 2023)	-	7.7	13.49	14.17	5.42	7.68	8.27	9B	ICL
GPT3.5 (Brown et al., 2020)	-	19.56	17.78	31.31	12.11	10.77	**21.62**	-	ICL
GPT3.5/SC (Brown et al., 2020)	-	11.49	12.62	17.77	6.95	7.9	8.62	-	ICL
Theia (VL)	**18.48**	**52.48**	**51.84**	**53.05**	**16.44**	**16.31**	18.39	1.4B	SL
Theia (L)	17.83	46.75	47.31	48.03	15.21	15.70	17.47	1.4B	SL
Theia (R)	17.74	48.38	48.12	50.75	13.56	15.00	14.93	1.4B	SL

Table 1: Experimental results of multimodal event extraction and ablation studies, on M2E2 dataset. The column "Method" refers to the learning method, where ICL and SL indicate in-context and supervised learning.

argument roles. We split M2E2 into training, development, and test sets with a ratio of 60:20:20. We consider a *missing* data setting, where the training session has only 10 complete data points with text-image pairs, and the rest of the data lacks the visual modality. During inference, complete data points of text-image pairs are provided to the model.

Results. Table 1 summarises the performance of the models on M2E2, where at the training stage only 10 data points include both modalities. On the BLEU score, our model outperforms the second-best model by more than 3 absolute points. This indicates a higher quality of text generation, which results in better extraction of event information. This superiority is also backed up by the F1, precision and recall scores associated with both event and argument extractions. On F1 score, our model achieves a substantial 32% improvement on event trigger detection and a 5% improvement over the second-best model on argument extraction. The comparatively lower improvements of argument role extraction, when compared to event extraction, suggest the difficulty and complexity of argument representations. This is explained by the diversity of the mentions and the lower frequency of the mentions in the training data. Furthermore, we believe that this be caused by the degradation in the quality of longer generated sequences.

Compared to our full model, the poorer performance when the visual context is discarded (Theia (L)) suggests that using the visual context indeed improves the extraction of events. Moreover, the lower performance of Theia (R) indicates the positive effect of our image synthesizer in providing more context-related features.

To evaluate the effect of missing visual data, we reconfigure the training settings from 10 complete data points to 5, 50, 100, and the full set of

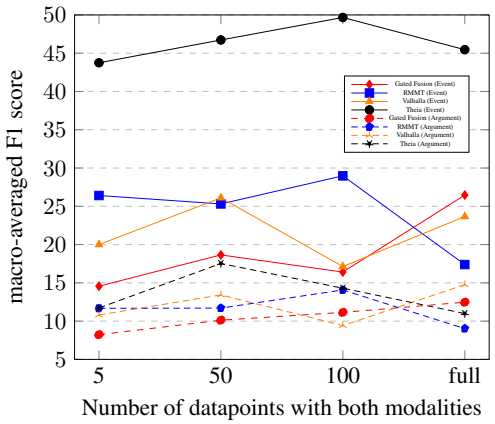

Figure 2: Performance of the models on M2E2, with different numbers of data point with both modalities during the training time.

data points, respectively. Figure 2 shows the results of this study. The x-axis shows the number of data points that, during training, include both textual and visual inputs, and the y-axis shows the F1 score. We see that on event trigger detection, our model outperforms all of the baselines in all settings. This observation suggests that the images generated by our image synthesizer model can capture more contextual information related to the textual input, hence resulting in improved performance. For argument extraction, our model performs the best in all settings except *full*, where Valhalla and Gated Fusion achieve better performance. The drop in performance of our model from 100 to full setting, suggests that the images associated with textual inputs (on full setting) carry less relevant information, and on the 100 setting when missing images are synthesised, the auxiliary images can provide better contextual information.

6 Conclusion

In this work, we address the task of *multimodal* event extraction, when the model faces *missing*

vision modality in the training data. We propose an end-to-end sequence generation model, which leverages a customised image synthesizer model to provide visual context to the model. Empirical results show that our proposed approach effectively exploits the incomplete training data, and outperforms state-of-the-art techniques in this task.

7 Limitations

This paper discusses multimodal event extraction where textual input is represented at sentence level, therefore our model is limited by the length of the inputs and performs poorly on document-level event extraction. While our model showed a strong ability in addressing this task, real-world scenarios are more challenging due to distributional shift in data and noisy environments. While our model has proven to generate context-dependent images, its ability in generating facial features is limited, hence generated images of people are unrecognisable. In addition, the core part of our model, which is based on large-scale vision-language pre-trained models, can limit the deployment in situations with limited resources. Besides, all the reported results are by fixing a random seed and running all experiments once.

8 Ethics Statement

This project aimed at customising existing pre-trained image generation models (Stable Diffusion) towards the domain of news. While our image synthesizer might be biased toward specific attributes of the domain, it enables our model to extract better events from the text. Similar to existing generative models, the generated content might be used to mislead or manipulate. Therefore, our model inherits similar potential risks from this family of generative models.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198.

Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Quoc-Chinh Bui, David Campos, Erik van Mulligen, and Jan Kors. 2013. A fast rule-based approach for biomedical event extraction. In *proceedings of the BioNLP shared task 2013 workshop*, pages 104–108.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska De Jong. 2011. An overview of event extraction from text. *DeRiVE@ ISWC*, pages 48–57.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. Degree: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. san diego. In *Third Annual International Conference on Learning Representations*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022a. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908.

Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Richard Chen, Rogerio S Feris, David Cox, and Nuno Vasconcelos. 2022b. Valhalla: Visual hallucination for machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5216–5226.

Farhad Moghimifar, Fatemeh Shiri, Reza Haffari, Yuan-Fang Li, and Van Nguyen. 2023. Few-shot domain-adaptative visually-fused event detection from text. In *2023 26th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE.

Radford Neal and Geoffrey Hinton. 1999. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. MIT Press.

Trung Minh Nguyen and Thien Huu Nguyen. 2019. One for all: Neural joint modeling of entities and events. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6851–6858.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alan Ritter, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*.

Marco A Valenzuela-Escárcega, Gus Hahn-Powell, Mihai Surdeanu, and Thomas Hicks. 2015. A domain-independent rule-based framework for event extraction. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 127–132.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. 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)*, pages 5784–5789.

Tongtong Wu, Fatemeh Shiri, Jingqi Kang, Guilin Qi, Gholamreza Haffari, and Yuan-Fang Li. 2022. Kcgee: Knowledge-based conditioning for generative event extraction.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166.

Tongtao Zhang, Spencer Whitehead, Hanwang Zhang, Hongzhi Li, Joseph Ellis, Lifu Huang, Wei Liu, Heng Ji, and Shih-Fu Chang. 2017. Improving event extraction via multimodal integration. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 270–278.

Yunyan Zhang, Guangluan Xu, Yang Wang, Xiao Liang, Lei Wang, and Tinglei Huang. 2019. Empower event detection with bi-directional neural language model. *Knowledge-Based Systems*, 167:87–97.

A Examples of Images by Theia

In this section, we present two samples of images generated by our model, Theia. We randomly selected two data points from the test set of M2E2. We then used the textual information to condition our image synthesizer into generating the visual representation of the text (Figure 3). As it can be seen, in the first example, *churches* is properly captured in the image generated by our model, Theia.

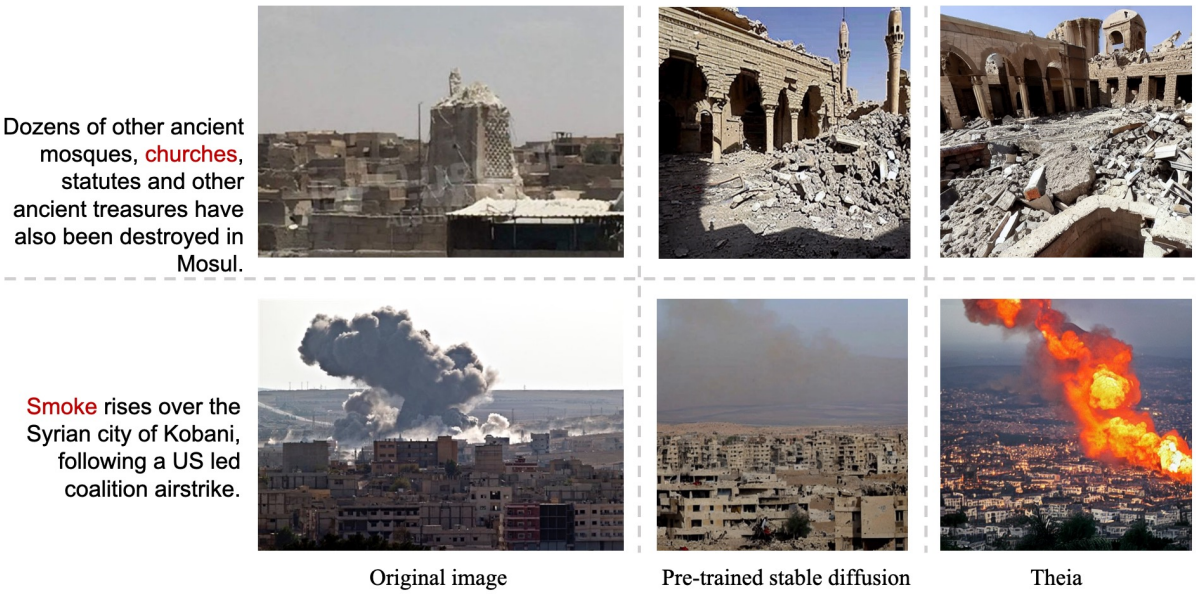

Figure 3: The images generated by our model for given textual inputs, compared to the original images from the news article, and the image generated by Stable Diffusion model (Rombach et al., 2022).

Similarly, in the second example, while stable diffusion fails to portray the *smoke*, our model represents this with fire/smoke, similar to the original image.

B Data Preparation

In order to develop our model, and run the experiments for all of the baselines as well, we convert the structured events into a sequence of tokens. To this end, a structured event represented in form of $y_i = \langle t_i, e_i, \langle a_i, r_i \rangle_1, \langle a_i, r_i \rangle_2, \dots \rangle$ (Section 3), is converted to the following span of tokens:
<TRIGGER> t_i </TRIGGER> <EVENT> e_i </EVENT> <ARG> a_i </ARG> <ROLE> r_i </ROLE>

where the tokens between <·> are special tokens added to the vocabulary set $\mathcal{W}$, and the embedding space of our encoder is then adjusted to the new length of vocab.

C Experimental Details

To train our model, we use CLIP ViT-L/14 (Radford et al., 2021) as a text and vision encoder. We also use the parameters of the pre-trained stable diffusion model (Rombach et al., 2022) to initialise the image synthesizer. This model is licensed with a CreativeML OpenRAIL++ license and free to use. We use an embedding size of 768, and a dropout of 0.3 over the textual encoder. We use Transformer-Base decoder in the sequence generator. Our model include more than 1.4B parameters, which belongs to both image synthesise and sequence generator. Optimisation is handled by Adam (Kingma and Ba, 2015) with a square root learning rate schedule and warm-up steps. We set the λ to 0.01 and use beam search with a beam size of 5 during inference. We ran our model on a machine with one NVIDIA A100 gpu. For the baselines, we follow the implementation details explained in their papers, and each model is trained with the same data preparation as our model. The best performing checkpoint on the development set then was used for testing. Since we kept the random seed constant throughout the experiments, we report the single-run results of each model. For running the experiments of GPT3.5, we prepare a prompt including task definition, definition of each event type and argument type and one example per event type, in addition to scene description, and ask the model to generate event information according to the instruction. We use model engine `gpt-3.5-turbo`. BLIP-2 [2] is used to extract description of the images, as a part of the prompt. The experiments related to OpenFlamingo [3] are conducted by using the open source implementation of Flamingo (Alayrac et al., 2022), where a prompt similar to prompt used in GPT3.5 experiments is used to instruct the model to generate information about events given a text and the corresponding image.

[2] https://github.com/salesforce/LAVIS/tree/main/projects/blip2
[3] https://github.com/mlfoundations/open_flamingo