# Training Multimedia Event Extraction With Generated Images and Captions

Zilin Du
zilin003@e.ntu.edu.sg
Nanyang Technological University
Singapore

Yunxin Li
liyunxin987@163.com
Harbin Institute of Technology
Shenzhen, China

Xu Guo
xu008@e.ntu.edu.sg
Nanyang Technological University
Singapore

Yidan Sun
suny0053@e.ntu.edu.sg
Nanyang Technological University
Singapore

Boyang Li
boyang.li@ntu.edu.sg
Nanyang Technological University
Singapore

## ABSTRACT

Contemporary news reporting increasingly features multimedia content, motivating research on multimedia event extraction. However, the task lacks annotated multimodal training data and artificially generated training data suffer from the distribution shift from the real-world data. In this paper, we propose Cross-modality Augmented Multimedia Event Learning (CAMEL), which successfully utilizes artificially generated multimodal training data and achieves state-of-the-art performance. Conditioned on unimodal training data, we generate multimodal training data using off-the-shelf image generators like Stable Diffusion [45] and image captioners like BLIP [24]. In order to learn robust features that are effective across domains, we devise an iterative and gradual training strategy. Substantial experiments show that CAMEL surpasses state-of-the-art (SOTA) baselines on the $M^2E^2$ benchmark. On multimedia events in particular, we outperform the prior SOTA by 4.2% F1 on event mention identification and by 9.8% F1 on argument identification, which demonstrates that CAMEL learns synergistic representations from the two modalities.

## CCS CONCEPTS

• **Information systems** → **Multimedia information systems**; • **Computing methodologies** → **Structured outputs**.

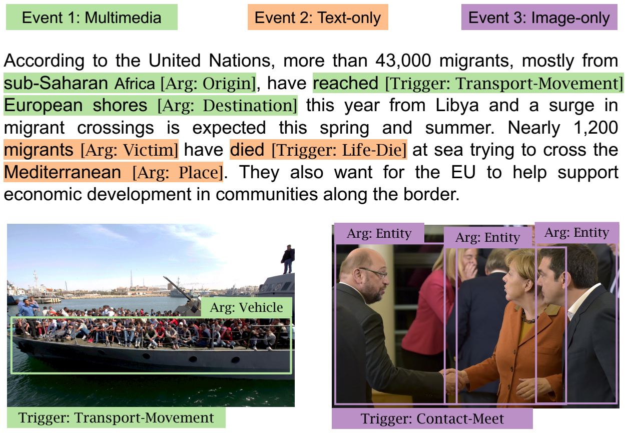
## KEYWORDS

Event Extraction; Multi-modal Learning; Data Augmentation; Cross-modality Generation

**ACM Reference Format:**
Zilin Du, Yunxin Li, Xu Guo, Yidan Sun, and Boyang Li. 2023. Training Multimedia Event Extraction With Generated Images and Captions. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612526

*MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.*

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612526



According to the United Nations, more than 43,000 migrants, mostly from sub-Saharan Africa [Arg: Origin], have reached [Trigger: Transport-Movement] European shores [Arg: Destination] this year from Libya and a surge in migrant crossings is expected this spring and summer. Nearly 1,200 migrants [Arg: Victim] have died [Trigger: Life-Die] at sea trying to cross the Mediterranean [Arg: Place]. They also want for the EU to help support economic development in communities along the border.

**Figure 1: Multimedia Event Extraction: three events are extracted from a multimedia news article. The multimedia event (Green) Transport-Movement is triggered by the word 'reached' and the image on the left. The textual event Life-Die (Orange) is triggered by the word 'died' only, and the visual event Contact-Meet (Purple) is solely triggered by the image on the right.**

## 1 INTRODUCTION

As a fundamental research topic in the domain of information extraction, event extraction aims to identify instances of events and their arguments from unstructured data [7, 11, 20, 44, 65]. An event refers to a specific incident that involves a change in state, which are marked by triggers such as verbs. The arguments of an event include the time and place of the event occurrence and its participants, such as the initiator, the recipient, and the instrument. Traditional research mostly focuses on a single modality, either language [8, 17, 27, 34, 58, 75] or visual data [3, 4, 39, 68].

As digital media quickly evolve, news reports today frequently present information with a combination of text and image, providing a more comprehensive view of events than text alone [9, 55]. This has spurred the emergence of the multimedia event extraction (MEE) task [26], which aims to jointly extract both textual and visual events from multimedia news articles. Figure 1 shows an MEE instance. Interestingly, not all events in a multimedia news article

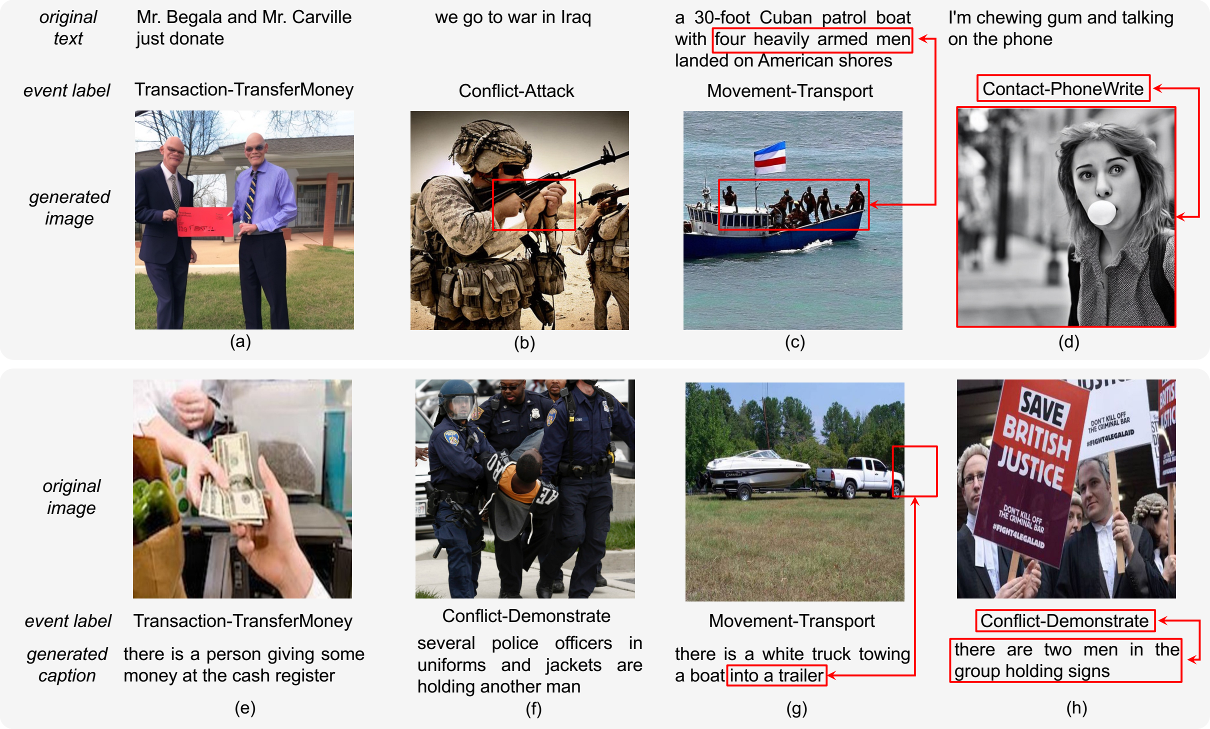

original text
Mr. Begala and Mr. Carville just donate
we go to war in Iraq
a 30-foot Cuban patrol boat with four heavily armed men landed on American shores
I'm chewing gum and talking on the phone
event label
Transaction-TransferMoney
Conflict-Attack
Movement-Transport
Contact-PhoneWrite
generated image
(a)
(b)
(c)
(d)
original image
event label
Transaction-TransferMoney
Conflict-Demonstrate
Movement-Transport
Conflict-Demonstrate
generated caption
there is a person giving some money at the cash register
several police officers in uniforms and jackets are holding another man
there is a white truck towing a boat into a trailer
there are two men in the group holding signs
(e)
(f)
(g)
(h)

**Figure 2: Examples of cross-modality augmented data. The red boxes indicate noise in the generated data, including inconsistency with the event label, hallucination, and unnatural image artifacts.**

are multimodal. For example, the event Transport-Movement is described by both the text and the image modalities, whereas the events Life-Die and Contact-Meet are contained respectively in text and image only.

A major challenge posed by the MEE problem is the lack of multimodal training data. The $M^2E^2$ dataset provided by [26] is only the test set. The labeled training datasets ACE2005 [52] and imSitu [68] contain event annotations in a single modality only. Despite recent progress [49, 74], transferring the knowledge learned from unimodal annotations to multimodal test data remains a difficult challenge.

After the explosive success of image generation networks such as DALL-E 2 [41] and Stable Diffusion [45], a natural thought is to perform cross-modality data augmentation in order to bridge the modality gaps of MEE. That is, conditioned on existing unimodal data, we can generate training data for the missing modality. After that, we use the resultant multimodal data to train a network. As the generative models capture world knowledge learned from observing correlative patterns among natural images and their textual descriptions, it is probable that such knowledge can be distilled and used to inform the task of event extraction.

However, a naive cross-modal data augmentation approach faces two obstacles. First, it is difficult to precisely control the generative models and produce data that are relevant to the event label and free of hallucination. In Figure 2 (d), the generated image depicts gum chewing but the event label is about talking on the phone. In Figure 2 (h), the generated caption describes the people as holding signs, whereas the label is demonstrating. In Figure 2 (g), the caption hallucinates a trailer that does not exist in the image. Second, in the case of image generation, existing models occasionally still generate images with significant deformation and unnatural artifacts. For example, the soldier in Figure 2 (b) is shown with three hands. For these reasons, the distribution of the generated data likely diverges from that of real-world data. In practice, we find that directly training on generated data results in performance degradation (Table 2).

To fully utilize the power of generative models to augment existing unimodal training data, we propose Cross-modality Augmented Multimedia Event Learning (CAMEL). After generating synthetic multimodal data, CAMEL applies an iterative and gradual training strategy that learns robust representations under noisy data and distribution shifts. We train the networks using text coupled with synthetic visual data and images coupled with synthetic textual data. The network is gradually frozen from the bottom up during training. Experimentally, we show that this training technique offers substantial benefits over naive data augmentation. In particular, on multimedia events, we outperform the previous best network, UniCL [29], by 4.1% F1 on event mention identification and 9.8% on argument role identification. In addition, the training strategy of CAMEL works robustly under different choices of image generation and captioning networks.

Our contributions can be summarized as follows:

- For multimodal event extraction, CAMEL utilizes synthetic data to fill in the missing modality in the unimodal ACE2005 and imSitu training datasets. To our best knowledge, this is the first work that successfully demonstrates the use of bidirectional cross-modality data augmentation (text-to-image and image-to-text) for multimodal learning. This results in superior data efficiency — with the unlabeled real-world multimodal VOA dataset [26] removed from training, we outperform previous work trained using VOA.
- We propose an incremental training strategy that handles artifacts, hallucination, and distribution shifts present in artificially generated multimodal data and avoids performance degradation caused by such noises.
- With CAMEL, we set a new state of the art on the $M^2E^2$ benchmark. On multimedia events in particular, we outperform the prior SOTA by 4.2% F1 on event mention identification and by 9.8% F1 on argument identification, which indicates that CAMEL learns synergistic representations from multimodal data.

# 2 RELATED WORK

## 2.1 Event Extraction

Event extraction [27] is a well-studied problem in information extraction. Many early works [8, 17, 30, 34, 58, 63, 75] focus on textual data and aim to identify event structures containing trigger words and arguments from unstructured text. Traditionally, textual event extraction is formulated as sequence labeling [12, 42, 61, 64]. More recent studies also formulate the problem as question answering [13, 59, 75]. Similarly, visual event extraction [5, 39, 46, 47, 60, 68], also referred to as situation recognition or visual semantic role labelling, aims to identify visual events and their participants. The earlier CRF-based methods [67, 68] jointly predict event type and the associated roles in one stage. [35] shows that identifying the action and the argument roles in two stages with separate networks to offer performance gains. More recent methods, such as GSRFormer [3] and SituFormer [60], adopt the two-stage approach.

Several studies investigate the use of multimodal data in unimodal event extraction. For example, [49] and [74] retrieve image relevant to the events, which can assist with disambiguation. [25] leverages image captions as distant supervision to interpret events in the associated images. Although they operate on multimodal data, these methods are aimed at events present in one modality.

Multimedia event extraction is proposed to extract events and arguments from multimedia documents [2, 26]. [26] tackles image-text documents while [2] focuses on video. WASE [26] uses weakly supervised learning to encode structured representations from textual and visual data into a shared embedding space. [29] introduces contrastive learning to bridge textual and visual modalities. Compare to these research, our method is the first to directly learn from synthetic multimodal training data, which are generated from labeled unimodal data.

## 2.2 Cross-modality Generation

### 2.2.1 *Cross-modality Generative Models.*

Text-to-image and image-to-text generative models are gaining traction. Text-to-image models [6, 22, 41, 43, 70, 71] are developed to produce high-quality images based on natural language descriptions. Early studies are based on GANs [22, 43, 70], auto-regression [41, 71], and VAEs [14, 51]. More recently, diffusion models have achieved impressive results [6, 72]. Some, like GLIDE [38] and Imagen [48], generate images at the pixel level directly; others, like DALL-E 2 [48] and Stable Diffusion [45], operate on a low-dimensional latent space. They have shown great promise in high-fidelity image generation. Meanwhile, remarkable improvement has also been achieved in the field of image-to-text generation (a.k.a. image captioning) with models like BLIP [24], GiT [53], and so on [21, 23, 28, 54].

### 2.2.2 *Cross-modality Generative Data Augmentation.*

Recent advances in generative models have propelled data augmentation research to a new level. On textual tasks, one approach is to generate additional textual training data [10, 16, 36, 37, 57, 62, 69]. Another, multimodal approach is to generate visual data to complement existing textual data [31, 33, 66, 76], which improves performance on textual tasks. For example, [31] generates visual data for machine translation. [76] uses generated images to guide text generation tasks, such as text completion, story generation, and concept-to-text generation. In addition, [33, 66] integrate synthetic images into language models to enhance the solution of plain language understanding tasks under low-resource settings. Unlike previous studies that address unimodal problems by synthesizing multimodal data, our work use the generated data to tackle multi-modal tasks. Doing so places a stringent requirement on the quality of generated data, as we need to train encoders in both modalities with the generated data. This necessitates overcoming the domain shifts between generated and real data. To the best of our knowledge, this is the first work to utilize bidirectional cross-modality data generation models for multimodal tasks.

# 3 TASK DEFINITION

Let $D = \langle M, S \rangle$ represent a multimedia document, which consists of a set of images $M = \{m_1, m_2, \ldots\}$ and a set of sentences $S = \{s_1, s_2, \ldots\}$. Each sentence $s$ consists of a sequence of words $[w_1, w_2, \ldots, w_L]$. The multimedia event extraction task contains the following two components.

**Event Mention Identification:** Given a multimedia document $D$, the first goal is to identify a set of event mentions from $D$. An event mention $e$ belongs to one of the predefined event types, $y_e$, and is grounded on a trigger word $w$ or a trigger image $m$ or both. A multimedia event contains both a trigger word $w$ and trigger image $m$, while a text-only or an image-only event only contains one type of trigger.

**Argument Role Identification:** The purpose of argument role identification is to extract, from the document $D$, all participants and attributes (i.e., arguments) of a given event $e$. For each event type, there is a predefined list of argument types. Each argument $a$ is classified into one argument type $y_a$ associated with the event type. The argument is grounded on a textual span $t$ in a sentence or one or more object bounding boxes in the image. The algorithm for argument role identification must also identify the position of the textual span $t$ and the bounding boxes.

If $e$ is a multimedia event, it must be grounded on both a textual trigger and a visual trigger. The arguments of multimedia events could contain both textual spans and visual objects. For example, in Figure 1, the multimedia event `Transport: Movement` is grounded on both the trigger word "reached" and the trigger image on the left. It also has two textual arguments and one visual argument.

# 4 APPROACH

The proposed approach, CAMEL, is trained with multimedia data that are artificially generated from unimodal data (Section 4.1). The cross-modality generative data augmentation approach can be thought of as distilling event-related knowledge from large generative models to the event identification network.

We show an overview of CAMEL in Figure 3. In a dual-encoder architecture, CAMEL first extracts features from the two modalities separately using unimodal encoders. To perform feature fusion and allow the network to pick relevant features among possibly noisy input, we design a modality-shared adapter module that perform cross-attention between the modalities. Further, to cope with possible distribution shifts and learn robust and generalizable features, we employ an iterative and gradual training strategy (Section 4.4). After these steps, we feed the resultant representation to domain-specific classifiers to identify the event mentions and arguments.

## 4.1 Cross-modality Generative Data Augmentation

A major obstacle for multimedia event extraction is the lack of multimodal training data. In the commonly used setup, first proposed by [26], the training data contains event annotations on text (ACE2005 [52]) and event annotations on images (imSitu [68]). The unlabeled VOA [26] dataset is often used as auxiliary training data;it contains parallel image-text data but no event annotations.

To tackle this problem, we utilize large text-to-image and image-to-text generative models to perform cross-modality generative data augmentation. Specifically, to augment the labeled textual data, we generate images using a text-to-image model. In addition, to augment labeled image data, we use an image-to-text model to generate image captions. This procedure yields labeled parallel image-text data. For most of our experiments, we use Stable Diffusion v2.1 [45] for image generation and BLIP [24] for captioning. However, CAMEL can be applied to a range of generative models with little loss in performance, as demonstrated in Section 5.3.

**Visual Data Augmentation.** We perform visual data augmentation on the labelled textual dataset, ACE2005, which consists of textual news reports. In order to extract textual spans that are relevant to the event, we utilize the existing annotations of event arguments and trigger words. For each event, we find the shortest continuous textual span that include all arguments and the trigger word, and use that as the textual input to image generation networks. We show one example of the extracted text span in the purple-lined box in Figure 3.

The image generation process is stochastic. Thus, we generate several images for each textual event in ACE2005 to cover different possible visual appearances and spatial arrangements. The number of images is a hyperparameter, which we set to four.

**Textual Data Augmentation.** To augment the visual dataset imSitu with the textual modality, we utilize the off-the-shelf image-to-text model to generate image captions. To generate diverse and detailed captions, we adopt nucleus sampling [15]. At each time step, the technique iteratively adds the most probable word to the candidate list until the total probability of the candidates exceeds a pre-defined probability. After that, the probabilities of candidates are normalized and one word is sampled accordingly. We generate one caption for each image in imSitu.

## 4.2 Model Architecture

**Feature Extractors.** CAMEL utilizes two pretrained Transformer encoders to extract unimodal features separately. Using the hidden states of the last network layer, the text encoder obtains a $d$-dimensional vector representation $h_i^{\text{text}}$ for each word $w_i$. Similarly, each patch of the image is encoded into a $d$-dimensional vector $h_i^{\text{img}}$. We denote the set of all text representations as $H_T$ and the set of all visual representations as $H_V$. We also prepend CLS tokens to the input of the two encoders. The corresponding encodings $h_{\text{CLS}}^{\text{text}}$ and $h_{\text{CLS}}^{\text{img}}$ can be understood as representing information from the entire sentence or image.

**Feature Fusion.** We devise a cross-attention network module, commonly used in Transformer networks, to fuse textual and visual features. The network consists of multi-head cross attention, layer normalization, and some linear layers. The detailed architecture is shown in Figure 4.

We refer to this module as the Adapter network. For simplicity, we denote the input to the Adapter as the query vector $q$, the key matrix $K$, and the value matrix $V$. The overall network is denoted as the function

$$g = \text{Adapter}(q, K, V). \tag{1}$$

We make repeated use of the same Adapter module with in the identification of event mentions and arguments, but change the $q$, $K$, and $V$ depending on the exact task. Most parameters are shared across tasks. However, parameters in the linear task-specific projection layer are specific to the four tasks (textual event mention, textual argument role, visual event mention, visual argument role).

The design of the Adapter module is motivated by the characteristics of multimedia documents, which usually do not explicit indicate the correspondence between images and the main text. When we try to identify a textual event and its arguments, we do not know which image is relevant to this event. The cross-attention mechanism allows the network to distinguish relevant images. Similarly, when extracting visual events, the network relies on the Adapter to select relevant portions of the text to facilitate its prediction.

**Textual Event Extraction.** The first sub-problem in textual event extraction is to identify the trigger word. This is word-level classification. The trigger word should be classified into the exact event type, whereas other words should be classified as non-triggers.

For the classification of the $i^{\text{th}}$ word, we first take its encoding from the textual encoder, $h_i^{\text{text}}$. After that, we feed $h_i^{\text{text}}$ to the

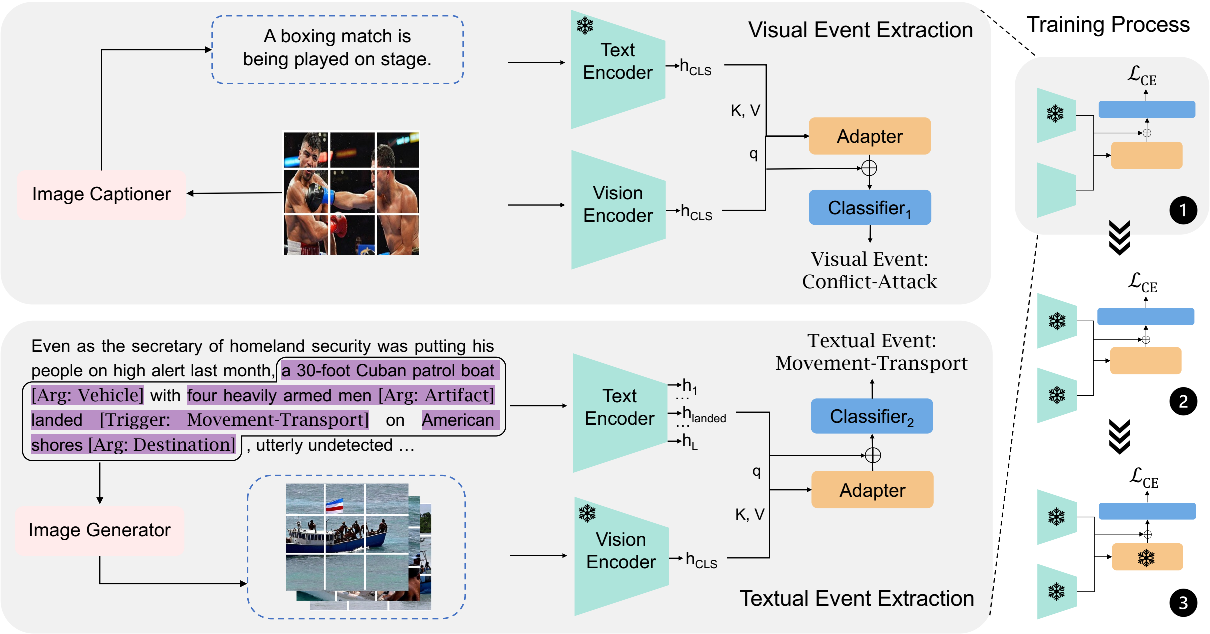

Visual Event Extraction
Training Process
A boxing match is being played on stage.
Image Captioner
Text Encoder
Vision Encoder
$h_{CLS}$
K, V
q
Adapter
Classifier$_1$
Visual Event: Conflict-Attack
$\mathcal{L}_{CE}$
1
2
3
Even as the secretary of homeland security was putting his people on high alert last month, a 30-foot Cuban patrol boat [Arg: Vehicle] with four heavily armed men [Arg: Artifact] landed [Trigger: Movement-Transport] on American shores [Arg: Destination], utterly undetected …
Image Generator
Textual Event: Movement-Transport
Classifier$_2$
$h_1$
$h_{landed}$
$h_L$
Textual Event Extraction

Figure 3: An overview of the CAMEL network architecture and its training strategy

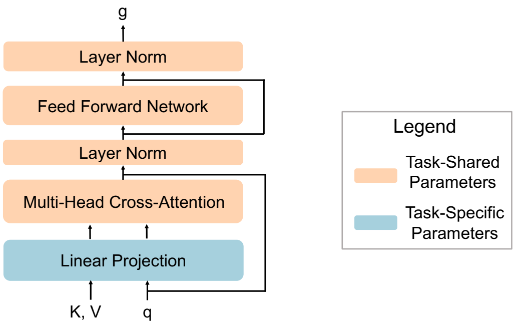


Figure 4: The architecture of the Adapter network.

Adapter network as the query vector. We use the CLS token encodings of all images in the entire multimedia document, denoted as $H^{\text{all-img}}$, as $K$ and $V$ in cross attention.

$$g_i^{\text{text}} = \text{Adapter}(h_i^{\text{text}}, H^{\text{all-img}}, H^{\text{all-img}}). \tag{2}$$

After that, we concatenate $h_i^{\text{text}}$ and $g_i^{\text{text}}$ and feed them through a linear classifier. The loss is cross-entropy.

The second sub-problem is the identification of event arguments. Following the convention in the literature [26, 29], we use the ground-truth list of entities for both training and inference. Each entity is a textual span that describes a person, an organization, a location and so on. We take the encoding of the first word in that entity as the entity feature $h^{\text{text-ent}}$, and feed it to the Adapter.

$$g_i^{\text{text-ent}} = \text{Adapter}(h^{\text{text-ent}}, H^{\text{all-img}}, H^{\text{all-img}}). \tag{3}$$

Similarly, we concatenate $h^{\text{text-ent}}$, $g_i^{\text{text-ent}}$, and the textual encoding of the trigger word, and feed them through a linear classifier, which classifies it into the argument classes. Though the types of valid arguments change depending on the event, here we do not exploit this fact for further performance improvement.

During training, if the ACE2005 sentence contains an event, we generate several positive images from the event text prompt (see Section 4.1). In addition, we also include some negative images generated for other events into $H^{\text{all-img}}$. This trains the network to distinguish between relevant images and irrelevant images. However, if the ACE2005 sentence does not contain any event, we would not be able to extract the event prompt using the method in Section 4.1. Instead, we randomly sample generated images from other text and use their encodings as $H^{\text{all-img}}$.

**Visual Event Extraction.** Similar to the textual modality, visual event extraction has two sub-problems, the classification of images into event types or non-events, and identification of objects as event arguments. For image event classification, we take the encoding of the image CLS token, $h_{\text{CLS}}^{\text{img}}$. Using the Adapter network again, we acquire an aggregated feature from the text modality, which we denote as $g^{\text{img}}$,

$$g^{\text{img}} = \text{Adapter}(h_{\text{CLS}}^{\text{img}}, H^{\text{all-text}}, H^{\text{all-text}}), \tag{4}$$

where the matrix $H^{\text{all-text}}$ contains the encoding vectors of the textual CLS token encodings of all sentences in the same batch. We feed the concatenation of $h_{\text{CLS}}^{\text{img}}$ and $g^{\text{img}}$ to a linear classifier.

For the second sub-problem, event argument identification, we first extract all objects in an image using an off-the-shelf object detector. For each object bounding box, we identify the three patches that contain its top-left corner, its center, and its bottom-right corner respectively. After that, we take the average of the three patch

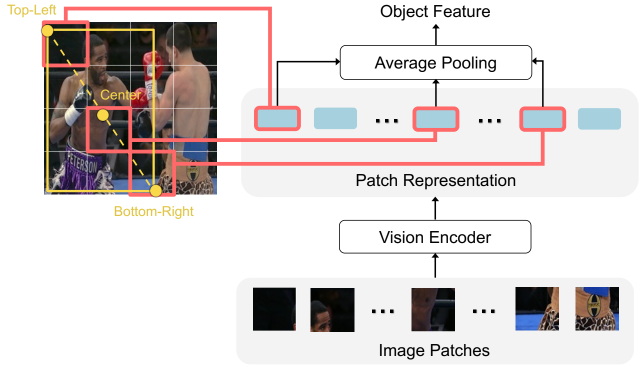


**Figure 5: Extracting features for objects in images.**

encodings, which we denote as $h^{\text{img-obj}}$. The object feature extraction process is illustrated in Figure 5. Once again, we apply feature fusion using the adapter network to obtain $g^{\text{img-obj}}$,

$$g^{\text{img-obj}} = \text{Adapter}(h^{\text{img-obj}}, H^{\text{all-text}}, H^{\text{all-text}}). \tag{5}$$

Finally, we concatenate three feature vectors, $h^{\text{img-obj}}$, $g^{\text{img-obj}}$, and $h^{\text{img}}$ into a single vector and feed it through a linear classifier.

### 4.3 Multimedia Event Extraction

For multimedia events, we need to resolve the coreference between text events and visual events. Given a multimedia document, we compute the similarity of each sentence-image pair. Following [26, 29], we treat a textual event and a visual event as the same event if and only if they have the same event type and the similarity between the sentence and image is greater than a threshold. We calculate the cosine similarity of the sentence-image pair using the CLIP model [40]. The multimedia event inherits all textual event arguments and visual event arguments as its own arguments.

### 4.4 Training Strategy

Robust representation learning is key to the success of cross-modality data augmentation and multimedia event extraction. As discussed in the introduction and shown in Figure 1, the automatically generated multimodal data often contain noise, such as inconsistency with the event label, hallucination, unnatural image artifacts, and so on. The discrepancy between the generated data distribution and the real-world data distribution may cause generalization difficulties. In addition, the M$^2$E$^2$ task itself poses a transfer learning problem because the training data, ACE2005 and imSitu, have different distributions from the test set. Hence, we need to learn robust feature representations that generalize well.

We propose an iterative and gradual training strategy, shown in the right column of Figure 3. We divide the training into the three stages. In the first stage, we first train on visual event mention, followed by textual event mention. The separation is a simple method to alleviate the well-known problem that different modalities learn at different speeds [56]. In the first stage, all network parameters are trained except the feature extractor corresponding to the generated synthetic data. For example, when training on real text data and generated image data, the text encoder is trainable but the image encoder is frozen. The rationale is to prevent the gigantic feature extractors (with hundreds of millions of parameters) from overfitting the low-level feature distributions of the augmented training data, which are likely idiosyncratic (e.g., soldiers with three hands) and not generalizable. However, we postulate that the high-level features extracted by the encoders are not heavily affected by shifts in lower-level feature distributions, so we train all the parameters after the encoders.

In the second stage, we again train the network on visual event mention identification, followed by textual event mention identification. Both encoders are frozen and only the Adapter and classifiers are trained. The design rationale is to allow visual classifiers to adapt to changes in the textual encoder in the first stage, and vice versa. In the third stage, we freeze all network parameters but finetune the visual event mention classifier using balanced event data. This technique is to mitigate the negative effects of imbalanced class proportions in the visual event mentions [19]. Finally, we separately finetune the visual encoder for visual event argument identification, and fintune the text encoder for textual event argument identification. This creates two models specialized for argument identification.

## 5 EXPERIMENTS

In this section, we extensively evaluate CAMEL by comparing against existing SOTA approaches, against ablated version of CAMEL, and against different choices for the image generators and image captioning networks.

### 5.1 Experimental Setting

**Datasets and Evaluation.** We evaluate on the M$^2$E$^2$ benchmark, a large-scale multimedia event extraction dataset that with the 8 types of events and 15 types of arguments. It contains 245 multimedia documents with 6,167 sentences and 1,014 images. There are 1,297 textual events and 391 visual events, among of which 192 textual event mentions and 203 visual event mentions are aligned into 309 multimedia events.

Since M$^2$E$^2$ does not provide training data, we follow the previous work [26, 29] to use the ACE2005 [52] and imSitu [68] (with the grounding information from [39]) for training. ACE2005 is a text dataset annotated with 33 event types, which contains the 8 specific types in M$^2$E$^2$. The image dataset imSitu is annotated with 504 activity verbs and 1,788 semantic roles. To utilize this dataset for 8-class classification, we follow [26] and map the 98 activity verbs to the 8 event types of M$^2$E$^2$. Following the previous works on event extraction [26, 30, 63], we use precision (P), recall (R), and F1 score (F1) as the default evaluation metrics.

**Baselines.** Following [29], we compare CAMEL with eight baselines for multimodal or unimodal event extraction.

Multimodal event extraction techniques can extract both textual events and visual events. WASE [26] first trains on different modalities independently and uses weakly supervised learning to align the two modalities. Two variations exist: WASE$_{\text{att}}$ locates the visual arguments using an attention heat map, whereas WASE$_{\text{obj}}$ leverages a object detection model. Flat$_{\text{att}}$ and Flat$_{\text{obj}}$ [26] are the simplified versions of $\text{WASE}_{\text{att}}$ and $\text{WASE}_{\text{obj}}$ respectively; they remove the graph convolution networks and concatenate features of different modalities for classification. UniCL [29] is the state-of-the-art on $\text{M}^2\text{E}^2$, which incorporates visual knowledge into textual event extraction but uses two separate modality-specific models for event extraction.

Unimodal event extraction methods only extract textual or visual events but not both. JMEE [30] is a state-of-the-art textual event extraction technique which utilizes an attention-based Graph Convolution Network. GAIL [73] is a reinforcement learning method for textual event extraction where rewards are estimated by a Generative Adversarial Network. VAD [74] augments textual documents with images retrieved from the Internet to improve textual event extraction. Clip-Event [25] utilizes the pretrained CLIP network to perform visual event extraction. WASE-T and WASE-V are the WASE model which trained on ACE2005 and imSitu only. The latter has two further varations $\text{WASE-V}_{\text{att}}$ and $\text{WASE-V}_{\text{obj}}$ [26].

**Hyperparameters.** During cross-modality data augmentation, for each event in ACE20005, we perform one-time generation of 4 images at 512×512 resolution with 100 denoising steps. In addition, we use nucleus (top-$p$) sampling [15] for image captioning with a probability threshold $p$ of 0.9. We generate one caption for each original image.

For fair comparisons with the SOTA baseline [29], we use the same 12-layer $\text{BERT}_{\text{Large}}$ as the text encoder, and the same 12-layer Transformer CLIP model [18] as the visual encoder with 16x16 patch size. To detect objects for visual argument roles, we leverage the pretrained YOLOv8 [50] as the object detector.

When training on visual event extraction, our batch size is set to 64 and learning rate to $10^{-4}$. For textual event extraction, the batch size is set to 10 and learning rate set to $10^{-4}$. We employ the AdamW optimizer [32] with $10^{-2}$ weight decay coefficient and the cosine learning rate schedule. In the first round of training, we train on the visual modality for 10 epochs and on the textual modality for 5 epochs. In the remainder of training, only one epoch is used for any modality. The maximum text input length is 200.

## 5.2 Main Results

Table 1 presents the performance of our proposed method CAMEL and several state-of-the-art baselines. The results show CAMEL significantly improves the event extraction performance over baseline methods. On textual events, we surpass UniCL by 1.7% F1 for event mention and 0.4% F1 for argument role. On visual events, we surpass UniCL by 0.9% F1 for event mention and 9.2% F1 for argument role. We speculate that the relatively small improvements for textual argument roles is that some textual arguments are pronouns (e.g., she) or proper noun (e.g., Saudi Arabia), which are not straightforward to visualize by the image generators.

Interestingly, the biggest performance boost appears on multimedia event extraction. We outperform the prior SOTA by 4.2% F1 on event mention identification and by 9.8% F1 on argument identification. This suggests CAMEL effectively learns synergistic representations from the two modalities.

**Table 1: Main results on event mention and argument role extraction for three types of events.**

| | | Event Mention | | | Argument Role | | |
|---|---|---|---|---|---|---|---|
| Event | Method | P | R | F1 | P | R | F1 |
| Textual | JMEE [30] | 42.5 | 58.2 | 48.7 | 22.9 | 28.3 | 25.3 |
| | GAIL [73] | 43.4 | 53.5 | 47.9 | 23.6 | 29.2 | 26.1 |
| | VAD [74] | 34.8 | 64.4 | 45.2 | 23.1 | 27.5 | 25.1 |
| | Flat [26] | 34.2 | 63.2 | 44.4 | 20.1 | 27.1 | 23.1 |
| | WASE-T [26] | 42.3 | 58.4 | 48.2 | 21.4 | 30.1 | 24.9 |
| | $\text{WASE}_{\text{att}}$ [26] | 37.6 | 66.8 | 48.1 | 27.5 | 33.2 | 30.1 |
| | $\text{WASE}_{\text{obj}}$ [26] | 42.8 | 61.9 | 50.6 | 23.5 | 30.3 | 26.4 |
| | UniCL [29] | 49.1 | 59.2 | 53.7 | 27.8 | 34.3 | 30.7 |
| | CAMEL (Ours) | 45.1 | 71.8 | **55.4** | 24.8 | 41.8 | **31.1** |
| Visual | Flat[26] | 27.1 | 57.3 | 36.7 | 4.3 | 8.9 | 5.8 |
| | $\text{WASE-V}_{\text{att}}$ [26] | 29.7 | 61.9 | 40.1 | 9.1 | 10.2 | 9.6 |
| | $\text{WASE-V}_{\text{obj}}$ [26] | 28.6 | 59.2 | 38.7 | 13.3 | 9.8 | 11.2 |
| | $\text{WASE}_{\text{att}}$ [26] | 32.3 | 63.4 | 42.8 | 9.7 | 11.1 | 10.3 |
| | $\text{WASE}_{\text{obj}}$ [26] | 43.1 | 59.2 | 49.9 | 14.5 | 10.1 | 11.9 |
| | CLIP-Event [25] | 41.3 | 72.8 | 52.7 | 21.1 | 13.1 | 17.1 |
| | UniCL [29] | 54.6 | 60.9 | 57.6 | 16.9 | 13.8 | 15.2 |
| | CAMEL (Ours) | 52.1 | 66.8 | **58.5** | 21.4 | 28.4 | **24.4** |
| Multi. | Flat [26] | 33.9 | 59.8 | 42.2 | 12.9 | 17.6 | 14.9 |
| | $\text{WASE}_{\text{att}}$ [26] | 38.2 | 67.1 | 49.1 | 18.6 | 21.6 | 19.9 |
| | $\text{WASE}_{\text{obj}}$ [26] | 43.0 | 62.1 | 50.8 | 19.5 | 18.9 | 19.2 |
| | UniCL [29] | 44.1 | 67.7 | 53.4 | 24.3 | 22.6 | 23.4 |
| | CAMEL (Ours) | 55.6 | 59.5 | **57.5** | 31.4 | 35.1 | **33.2** |

## 5.3 Ablation Study

In order to investigate the effects of different components in CAMEL, we create ablated systems by removing each of the components. First, we create two variations in the training strategy. In the **combined training** baseline, we merge the textual event task and the visual event task as one training set and train the model in one stage without freezing any model parameters. In the **one-round training** baseline, we separate the training of the textual event task and the visual event task. We freeze the visual encoder when training on real textual data and generated visual data, and freeze the textual encoder when training on real textual data and generated visual data. However, we only apply one stage of training and remove the two later stages.

Next, in the **w/o augmentation** baseline, we remove all generated multimodal training data and train the network on unimodal data alone. For example, in textual event mention identification, we train the textual encoder and the classifier; the Adapter is removed as well. Finally, the **w/o Adapter** ablation retains multimodal training data but removes the Adapter network. The cross-attention scores are computed as cosine similarity. For example, in text mention identification, we compute the cosine similarity between each word $h_i^{\text{text}}$ and the visual image encoding $h_{\text{CLS}}^{\text{img}}$. The similarities scores are normalized and used to compute a convex combination of image features, denoted as $g_i^{\text{text}}$. The concatenation of $h_i^{\text{text}}$ and $g_i^{\text{text}}$ is used for classification.

The results are shown in Table 2. The most interesting finding is that the w/o augmentation, unimodal baseline outperforms the simplistic combined training strategy by large margins (up to 12.9% F1 on multimedia event mentions). This clearly demonstrates the

Table 2: Ablation results of CAMEL on the $M^2E^2$ dataset.

|  | Textual Events |  |  |  |  |  | Visual Events |  |  |  |  |  | Multimedia Events |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | Event Mention |  |  | Argument Role |  |  | Event Mention |  |  | Argument Role |  |  | Event Mention |  |  | Argument Role |  |  |
| Method | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| CAMEL | 45.1 | 71.8 | **55.4** | 24.8 | 41.8 | **31.1** | 52.1 | 66.8 | **58.5** | 21.4 | 28.4 | **24.4** | 55.6 | 59.5 | **57.5** | 31.4 | 35.1 | **33.2** |
| combined training | 41.9 | 69.7 | 48.3 | 22.0 | 34.5 | 26.8 | 60.1 | 40.4 | 48.3 | 24.8 | 17.7 | 20.6 | 53.2 | 32.0 | 40.0 | 27.4 | 15.8 | 20.0 |
| one-round training | 45.1 | 70.6 | 55.0 | 22.4 | 40.6 | 30.6 | 66.5 | 36.6 | 47.2 | 24.1 | 13.5 | 17.3 | 55.9 | 33.6 | 42.0 | 31.5 | 19.2 | 23.8 |
| w/o augmentation | 40.0 | 73.2 | 51.7 | 25.7 | 30.5 | 27.9 | 48.9 | 62.9 | 55.0 | 19.3 | 25.9 | 22.1 | 51.5 | 54.4 | 52.9 | 31.6 | 26.4 | 28.8 |
| w/o adapter | 43.7 | 70.8 | 54.0 | 25.3 | 36.0 | 29.7 | 45.5 | 68.0 | 54.5 | 19.3 | 30.5 | 23.6 | 49.8 | 57.0 | 53.2 | 30.0 | 30.9 | 30.4 |

Table 3: Performance of CAMEL with different image captioners and image generators.

|  | Textual Events |  |  |  |  |  | Visual Events |  |  |  |  |  | Multimedia Events |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | Event Mention |  |  | Argument Role |  |  | Event Mention |  |  | Argument Role |  |  | Event Mention |  |  | Argument Role |  |  |
| Method | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| CAMEL | 45.1 | 71.8 | 55.4 | 24.8 | **41.8** | 31.1 | **52.1** | 66.8 | **58.5** | **21.4** | 28.4 | **24.4** | 55.6 | **59.5** | 57.5 | 31.4 | **35.1** | **33.2** |
| *Replacing the Image Captioner with ...* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| BLIPv2 [23] | 44.4 | 71.7 | 54.8 | 25.2 | 36.2 | 29.7 | 49.2 | 65.7 | 56.3 | 19.3 | 26.8 | 22.4 | 54.1 | 57.9 | 55.9 | 29.5 | 30.3 | 29.9 |
| GIT [53] | 44.0 | 71.9 | 54.6 | 25.8 | 38.5 | 30.9 | 49.9 | 65.7 | 56.7 | 19.6 | 28.2 | 23.1 | 53.5 | 57.0 | 55.2 | 29.8 | 31.9 | 30.8 |
| VIT-GPT2 [21] | 44.8 | 71.1 | 55.0 | **26.2** | 38.9 | **31.3** | 49.1 | 65.7 | 56.2 | 19.5 | 28.1 | 23.0 | 54.2 | 57.9 | 56.0 | **31.9** | 31.9 | 31.9 |
| OFA [54] | 44.8 | 71.3 | 55.0 | 26.1 | 36.3 | 30.4 | 49.8 | 65.7 | 56.7 | 19.7 | **28.5** | 23.3 | 54.6 | 57.9 | 56.2 | 30.4 | 30.8 | 30.6 |
| *Replacing the Image Generator with ...* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| SDv2 [45] | **45.4** | 72.0 | **55.7** | 25.0 | 40.4 | 30.9 | 49.6 | 66.0 | 56.6 | 18.7 | 26.2 | 21.8 | 54.7 | 58.9 | 56.7 | 30.1 | 31.8 | 30.9 |
| SDv1.5 [45] | **45.4** | 71.6 | 55.6 | 24.6 | 40.2 | 30.6 | 50.1 | **67.3** | 57.4 | 20.1 | **28.5** | 23.6 | 54.9 | 58.3 | 56.5 | 31.2 | 32.3 | 31.8 |
| Kandinsky [1] | 44.8 | **72.3** | 55.4 | 24.0 | 41.3 | 30.3 | 50.4 | 66.2 | 57.2 | 19.9 | 27.2 | 23.0 | **56.3** | **59.5** | **57.9** | 29.8 | 34.2 | 31.9 |
| UniCL [29] | 49.1 | 59.2 | 53.7 | 27.8 | 34.3 | 30.7 | 54.6 | 60.9 | 57.6 | 16.9 | 13.8 | 15.2 | 44.1 | 67.7 | 53.4 | 24.3 | 22.6 | 23.4 |

difficulties in training on generated multimodal data and the need for a carefully devised training strategy. Second, the one-round training strategy still falls behind the full-fledged CAMEL, showing the three-stage strategy to be effective. Additionally, the full-fledged CAMEL appears superior to unimodal training, surpassing by 4.6% and 4.4% on multimedia event mention and argument role extraction respectively. Finally, CAMEL outperforms the network without Adapter, indicating the advantage of the Adapter design.

### 5.4 Choice of Generative Models

We test if CAMEL can work with other large pretrained generative models. By default, CAMEL leverages Stable Diffusion v2.1 [45] as the image generator and BLIP [24] as image captioning model. In this experiment, we test out three different image generators, including Stable Diffusion v1.5 and v2 (SDv1.5 and SDv2) and the Kandinsky model [1]. For image captioning, we attempt BLIPv2 [23], GIT [53], OFA [54], and VIT-GPT2 [21].

Table 3 shows the results. We observe that, while the default settings works well, it often does not achieve the best F1 scores compared to other combinations. In addition, many model combinations outperform UniCL, the previous SOTA model. This demonstrates the generality of the CAMEL technique.

## 6 CONCLUSIONS

In this paper, we study the problem of multimedia event extraction and investigate the use of image generative networks and image captioning networks to complement existing unimodal training data. The automatically generated multimodal data often contain noise, such as inconsistency with the event label, hallucination, unnatural image artifacts, creating challenges for training. We propose a network, CAMEL, and a specialized training strategy to cope with augmented multimodal training data. CAMEL surpasses he prior SOTA by 4.2% F1 on event mention identification and by 9.8% F1 on argument identification. An ablation study shows that the design of network structure, the shared adapter, and the iterative training strategy in our method significantly improve performance. We also test the generality of the benefits of our approach to other cross-modality generative models.

## ACKNOWLEDGMENTS

This work has been supported by the Nanyang Associate Professorship and the National Research Foundation Fellowship (NRF-NRFF13-2021-0006), Singapore. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the funding agencies.

## REFERENCES

[1] Shakhmatov Arseniy, Razzhigaev Anton, Nikolich Aleksandr, Arkhipkin Vladimir, Pavlov Igor, Kuznetsov Andrey, and Dimitrov Denis. 2023. Kandinsky 2.1. https://github.com/ai-forever/Kandinsky-2.

[2] Brian Chen, Xudong Lin, Christopher Thomas, Manling Li, Shoya Yoshida, Lovish Chum, Heng Ji, and Shih-Fu Chang. 2021. Joint Multimedia Event Extraction from Video and Article. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 74–88.

[3] Zhi-Qi Cheng, Qi Dai, Siyao Li, Teruko Mitamura, and Alexander Hauptmann. 2022. GSRFormer: Grounded Situation Recognition Transformer with Alternate Semantic Attention Refinement. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3272–3281.

[4] Junhyeong Cho, Youngseok Yoon, and Suha Kwak. 2022. Collaborative transformers for grounded situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19659–19668.

[5] Junhyeong Cho, Youngseok Yoon, and Suha Kwak. 2022. Collaborative Transformers for Grounded Situation Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[7] 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*, Vol. 2. Lisbon, 837–840.

[8] Xinya Du, Alexander M Rush, and Claire Cardie. 2021. GRIT: Generative Role-filler Transformers for Document-level Event Entity Extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 634–644.

[9] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A survey on deep learning for multimodal data fusion. *Neural Computation* 32, 5 (2020), 829–864.

[10] Jiahui Gao, Renjie Pi, LIN Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023. Self-Guided Noise-Free Data Generation for Efficient Zero-Shot Learning. In *International Conference on Learning Representations*.

[11] Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics* 28, 3 (2002), 245–288.

[12] Tao Gui, Jiacheng Ye, Qi Zhang, Zhengyan Li, Zichu Fei, Yeyun Gong, and Xuan-Jing Huang. 2020. Uncertainty-Aware Label Refinement for Sequence Labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2316–2326.

[13] Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 643–653.

[14] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.

[15] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751* (2019).

[16] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. *arXiv preprint arXiv:2212.09689* (2022).

[17] 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*. 1890–1908.

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. PMLR, 4904–4916.

[19] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217* (2019).

[20] Paul R Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank.. In *LREC*. 1989–1993.

[21] Ankur Kumar. 2022. The Illustrated Image Captioning using transformers. *ankur3107.github.io* (2022). https://ankur3107.github.io/blogs/the-illustrated-image-captioning-using-transformers/

[22] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. 2019. Controllable text-to-image generation. *Advances in Neural Information Processing Systems* 32 (2019).

[23] 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* (2023).

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.

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

[26] 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*. 2557–2568.

[27] Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, et al. 2022. A Survey on Deep Learning Event Extraction: Approaches and Applications. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[28] Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, and Min Zhang. 2023. LMEye: An Interactive Perception Network for Large Language Models. *arXiv preprint arXiv:2305.03701* (2023).

[29] Jian Liu, Yufeng Chen, and Jinan Xu. 2022. Multimedia Event Extraction From News With a Unified Contrastive Learning Framework. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1945–1953.

[30] Xiao Liu, Zhunchen Luo, and He-Yan Huang. 2018. Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1247–1256.

[31] Quanyu Long, Mingxuan Wang, and Lei Li. 2021. Generative Imagination Elevates Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5738–5748.

[32] Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[33] Yujie Lu, Wanrong Zhu, Xin Wang, Miguel Eckstein, and William Yang Wang. 2022. Imagination-Augmented Natural Language Understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4392–4402.

[34] Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022. Prompt for Extraction? PAIE: Prompting Argument Interaction for Event Argument Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6759–6774.

[35] Arun Mallya and Svetlana Lazebnik. 2017. Recurrent models for situation recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 455–463.

[36] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *arXiv preprint arXiv:2202.04538* (2022).

[37] Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. 2022. Tuning Language Models as Training Data Generators for Augmentation-Enhanced Few-Shot Learning. *arXiv preprint arXiv:2211.03044* (2022).

[38] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*. PMLR, 16784–16804.

[39] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 314–332.

[40] 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*. PMLR, 8748–8763.

[41] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[42] Alan Ramponi, Rob van der Goot, Rosario Lombardo, and Barbara Plank. 2020. Biomedical event extraction as sequence labeling. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*. 5357–5367.

[43] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*. PMLR, 1060–1069.

[44] Ellen Riloff and Mark Schmelzenbach. 1998. An empirical approach to conceptual case frame acquisition. In *Sixth Workshop on Very Large Corpora*.

[45] 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*. 10684–10695.

[46] Arka Sadhu, Kan Chen, and Ram Nevatia. 2021. Video Question Answering with Phrases via Semantic Roles. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2460–2478.

[47] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5589–5600.

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep

language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

[49] Meihan Tong, Shuai Wang, Yixin Cao, Bin Xu, Juanzi Li, Lei Hou, and Tat-Seng Chua. 2020. Image enhanced event detection in news articles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9040–9047.

[50] Ultralytics. 2023. YOLOv8. https://github.com/ultralytics/ultralytics.

[51] Arash Vahdat and Jan Kautz. 2020. NVAE: A deep hierarchical variational autoencoder. *Advances in neural information processing systems* 33 (2020), 19667–19679.

[52] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia* 57 (2006), 45.

[53] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100* (2022).

[54] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*. PMLR, 23318–23340.

[55] Sitong Wang, Samia Menon, Tao Long, Keren Henderson, Dingzeyu Li, Kevin Crowston, Mark Hansen, Jeffrey V Nickerson, and Lydia B Chilton. 2023. ReelFramer: Co-creating News Reels on Social Media with Generative AI. *arXiv preprint arXiv:2304.09653* (2023).

[56] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What Makes Training Multi-Modal Classification Networks Hard?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[57] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-Instruct: Aligning Language Model with Self Generated Instructions. *arXiv preprint arXiv:2212.10560* (2022).

[58] Ziqi Wang, Xiaozhi Wang, Xu Han, Yankai Lin, Lei Hou, Zhiyuan Liu, Peng Li, Juanzi Li, and Jie Zhou. 2021. CLEVE: Contrastive Pre-training for Event Extraction. 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)*. 6283–6297.

[59] Kaiwen Wei, Xian Sun, Zequn Zhang, Jingyuan Zhang, Guo Zhi, and Li Jin. 2021. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction. 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)*. 4672–4682.

[60] Meng Wei, Long Chen, Wei Ji, Xiaoyu Yue, and Tat-Seng Chua. 2022. Rethinking the two-stage framework for grounded situation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2651–2658.

[61] Haoyang Wen, Yanru Qu, Heng Ji, Qiang Ning, Jiawei Han, Avirup Sil, Hanghang Tong, and Dan Roth. 2021. Event time extraction and propagation via graph attention networks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 62–73.

[62] Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4602–4625.

[63] Julia Wiedmann. 2017. Joint learning of structural and textual features for web scale event extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[64] Runxin Xu, Tianyu Liu, Lei Li, and Baobao Chang. 2021. Document-level Event Extraction via Heterogeneous Graph-based Interaction Model with a Tracker. 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)*. 3533–3546.

[65] Nianwen Xue and Martha Palmer. 2003. Annotating the propositions in the Penn Chinese Treebank. In *Proceedings of the second SIGHAN workshop on Chinese language processing*. 47–54.

[66] Yue Yang, Wenlin Yao, Hongming Zhang, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2022. Z-LaVI: Zero-Shot Language Solver Fueled by Visual Imagination. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 1186–1203. https://aclanthology.org/2022.emnlp-main.78

[67] Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, and Ali Farhadi. 2017. Commonly uncommon: Semantic sparsity in situation recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7196–7205.

[68] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5534–5542.

[69] Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922* (2022).

[70] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. 2021. Vector-quantized image modeling with improved VQGAN. *arXiv preprint arXiv:2110.04627* (2021).

[71] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* (2022).

[72] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. 2023. Text-to-image Diffusion Model in Generative AI: A Survey. *arXiv preprint arXiv:2303.07909* (2023).

[73] Tongtao Zhang and Heng Ji. 2018. Event extraction with generative adversarial imitation learning. *arXiv preprint arXiv:1804.07881* (2018).

[74] 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*. 270–278.

[75] Yang Zhou, Yubo Chen, Jun Zhao, Yin Wu, Jiexin Xu, and Jinlong Li. 2021. What the role is vs. what plays the role: Semi-supervised event argument extraction via dual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 14638–14646.

[76] Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022. Visualize Before You Write: Imagination-Guided Open-Ended Text Generation. *arXiv preprint arXiv:2210.03765* (2022).