

multi-modal reasoning —> can make synthetic dataset

GenEARL: A Training-Free Generative Framework for Multimodal Event Argument Role Labeling

Hritik Bansal, Po-Nien Kung, P. Jeffrey Brantingham, Kai-Wei Chang, Nanyun Peng

University of California Los Angeles

hbansal@cs.ucla.edu, ponienkung@cs.ucla.edu

why not feature based event detection ?
Like how humans do. It would require a lot of data...

Abstract

Multimodal event argument role labeling (EARL), a task that assigns a role for each event participant (object) in an image is a complex challenge. It requires reasoning over the entire image, the depicted event, and the interactions between various objects participating in the event. Existing models heavily rely on high-quality event-annotated training data to understand the event semantics and structures, and they fail to generalize to new event types and domains. In this paper, we propose GenEARL, a training-free generative framework that harness the power of the modern generative models to understand event task descriptions given image contexts to perform the EARL task. Specifically, GenEARL comprises two stages of generative prompting with a frozen vision-language model (VLM) and a frozen large language model (LLM). First, a generative VLM learns the semantics of the event argument roles and generates event-centric object descriptions based on the image. Subsequently, a LLM is prompted with the generated object descriptions with a predefined template for EARL (i.e., assign an object with an event argument role). We show that GenEARL outperforms the contrastive pretraining (CLIP) baseline by 9.4% and 14.2% accuracy for zero-shot EARL on the M²E² and SWiG datasets, respectively. In addition, we outperform CLIP-Event by 22% precision on M²E² dataset. The framework also allows flexible adaptation and generalization to unseen domains.

1 Introduction

Multimodal event extraction (EE) aims at identifying the event types depicted in an image (e.g., ‘Arrest’), the event participates (objects), and event argument role labeling (EARL) for the objects (e.g., A person is the ‘Agent’ performing an ‘Arrest’ event). In many real-world applications such as analysis of news articles and social media, important event-related information is grounded in the



Figure 1: Overview of Multimodal Event Argument Role Labeling task. Given the image that depicts an ARREST event type, a list of possible event argument roles (AGENT, PERSON, INSTRUMENT), and three participant objects (bounding boxes) of color ‘red’, ‘blue’, and ‘yellow’. The task is to assign an event argument to each of the objects based on their role in the depicted event. Here, the ‘blue’ bounding box plays the role of the PERSON who gets arrested whereas the object in the ‘red’ and ‘yellow’ plays the role of the AGENT performing the arrest.

multimodal image-text data (Stephens et al., 1998; Blandfort et al., 2019). For example, a tweet about a protest may not mention whether the protest is violent, which can be readily discerned from the accompanying image (Giorgi et al., 2022). Hence, it becomes increasingly important to extract the multimodal events and relations, beyond text, as they provide a useful lens for analyzing social events.

In this work, our focus is on multimodal Event Argument Role Labeling (EARL) (illustrated in Figure 1) given its practical significance and the unique technical challenges it presents within the multimodal EE pipeline. For example, social scientists will benefit from understanding the DEMONSTRATORS, POLICE, and INSTRUMENT from a PROTEST scene. In the meanwhile, performing an accurate multimodal EARL requires fine-grained understanding of the entire image depicting the event along with comprehending the roles and interactions between various objects present in the image, which poses challenges for large-scale pre-trained vision-language models such as CLIP (Rad-

can we just say because you are using a relatively larger model we directly see the increase due to it over CLIP?

ford et al., 2019).

Existing methods (Li et al., 2020, 2022b) trained on event-annotated data, such as ACE (Doddington et al., 2004), imSitu (Yatskar et al., 2016), and VOA (Li et al., 2020) tend to overfit to the event types seen during training, and fail to generalize well to unseen event types and new domains (Parekh et al., 2023). Subsequent finetuning of these models will require the creation of new event-annotated data, which is expensive and hard to acquire.

To tackle these challenges, our major contribution is the proposal of **GenEARL**, a two-stage training-free generative framework for multimodal EARL. Our framework utilizes the ability of the instruction-following generative vision-language models (GVLM) and large language models (LLM) to effectively comprehend multimodal and text-only input prompts. First, we prompt the GVLM with visual features of the image and object, along with the target event and its associated argument role in a predefined template such that it generates an event-centric object description. This object description summarizes the role of the object in the context of the event depicted in the image. Since the generative VLM reasons over the task description in the input prompt, it does not require expensive event-annotated data for generating object role descriptions. In the next stage, we prompt instruction-following LLM with the generated object descriptions, image caption, along with the targeted event and its argument role labels in a predefined template to predict event argument role labels. We further show that in-context learning (Dong et al., 2022; Liu et al., 2023c) for each of the steps can improve EARL extraction.

Our second contribution is a comprehensive empirical study of the generalization ability of our framework to perform multimodal EARL on the M²E² and SWiG datasets, as used in Li et al. (2022b). We find that GenEARL outperforms the CLIP-L/14 (Radford et al., 2019) baseline by 9.4% and 14.2% accuracy in the zero-shot EARL on the M²E² (Li et al., 2020) and SWiG (Pratt et al., 2020) datasets, respectively (§5.1). In addition, we find that combining CLIP for event detection and GenEARL for EARL outperforms CLIP-Event (Lu et al., 2021) on M²E² by 22% on precision.

We further study the discrepancies between the training-free and supervised paradigms by finetuning a CLIP-B/32 on the training data of the SWiG dataset (§4.3), which enhances the CLIP’s understanding of event semantics and structures. We ob-

serve that our framework reduces the gap between the supervised and training-free paradigms from 44.5% to 16.1% on the SWiG dataset. In addition, our framework outperforms this finetuned CLIP model on the M²E² dataset by 3% accuracy with just three-shot in-context learning examples. This highlights that our generative paradigm is competitive with supervised methods without any training on event data.

2 Background: Multimodal EARL

In this section, we first describe the task of multimodal event argument role labeling (EARL). Given an input image \mathcal{I} that is depicting an event type E . The event type belongs to one of the possible events types $E \in \mathcal{E}$ under the predefined event ontology. Further, each event type E consists of a set of event arguments $\mathcal{A}_E = \{a_1, \dots, a_m\}$ that specifies the roles associated with it. In our datasets, each image contains a set of objects $\mathcal{O} = \{o_1, \dots, o_n\}$ where each object o_i is represented as a bounding box and participates in the depicted event.

too many requirements:
- must be from the set of events
- need event arguments
- need bounding box

The task of multimodal event argument role labeling (EARL) is to label every participating object (bounding box) o_i with an event argument role label a_j . The predicted argument role label for the participating object highlights its interactions with the other objects in the scene and its contribution to the understanding of the event scenarios. For example, Figure 1 the object in the ‘blue’ bounding box plays the role of the PERSON who gets arrested whereas the object in the ‘red’ and ‘yellow’ plays the role of the AGENT performing the arrest. Overall, EARL is a challenging task since it requires a model to reason over the fine-grained features of the image and the interactions between the participating objects. We highlight the assumptions of our work in Appendix §A.

3 The GenEARL Model

To tackle the challenging task of multimodal EARL, we propose GenEARL, a two-stage training-free generative framework that utilizes the unique capabilities of the modern generative models (Liu et al., 2023a; OpenAI, 2021) to understand event task descriptions and perform well on them. Specifically, the framework consists frozen (a) generative vision-language model (GVLM) and (b) large language model (LLM). Consider an input instance $\mathcal{Q}_i = \{\mathcal{I}, E, o_i, \mathcal{A}_E\}$ where we aim to label the event argument role for the object o_i in

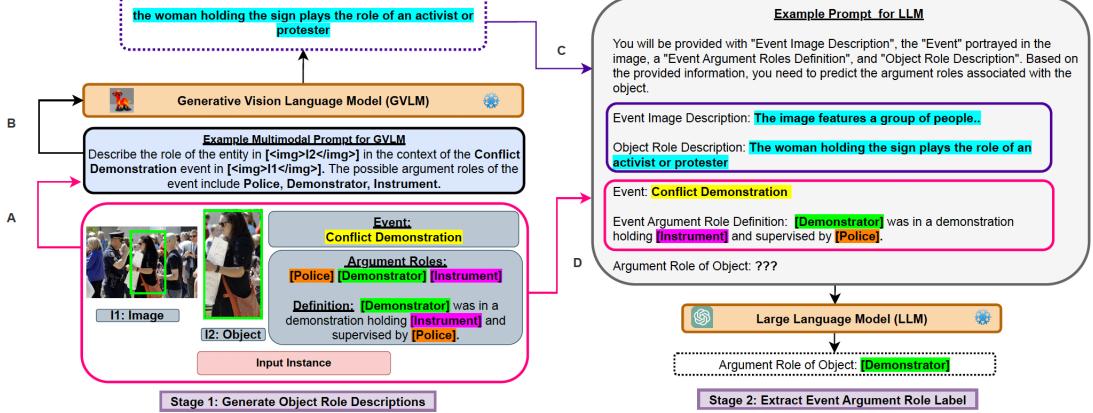


Figure 2: Overview of the GenEARL framework for multimodal event argument role labeling. It comprises two stages of generative prompting. In the first stage, a generative vision-language model like LLaVA is prompted with the multimodal input prompt to generate an event-centric object role description. In the second stage, we extract the argument role label based on the generated object description using a large language model like ChatGPT. **A:** The raw input instance data is converted into a multimodal prompt for GVLM. **B:** The input prompt is fed to the GVLM for object role description generation. **C:** The generated object role description is embedded in the template used to prompt the LLM. GVLM is also used to generate a caption for the image that depicts the event. **D:** The event details including the possible argument role labels are added in the LLM prompt.

the event E depicted in the image \mathcal{I} . \mathcal{A}_E comprises the possible event argument role labels for the event E along with their definition from the dataset guidelines.

We propose to decouple the task of multimodal EARL on the given input instance into two distinct sub-tasks. This approach allows each sub-task to be effectively addressed by the individual capabilities of the GVLM and LLM respectively. Specifically, we (a) generate an event-centric object description that captures the role played by the object o_i in the context of the event E in the image \mathcal{I} (§3.1) followed by (b) extracting the EARL from the generated object role descriptions (§3.2).

3.1 Generative Vision-Language Model

We consider a generative vision-language model (GVLM) \mathcal{G} that can reason over the text and visual input and generate an output text. Here, the GVLM is capable of capturing the fine-grained details in the event task description with image(s) in its context. We illustrate this stage in Figure 2 (left).

For the input instance Q_i we prompt the GVLM with a multimodal prompt $\mathcal{P} = \mathcal{T}(Q_i)$ where \mathcal{T} is a predefined template that converts the input query details into a multimodal prompt (Appendix Table 8). We highlight the example multimodal input prompt in Figure 2 (left). Subsequently, we generate the event-centric object role description d_{o_i} from the GLVM with the context \mathcal{P} .

3.2 Large Language Model

We illustrate this stage in Figure 2 (right). In the first stage, we generate an object role description for a given image \mathcal{I} and participant object o_i . Here, we use a large language model (LLM) \mathcal{L} since they are capable of generalizing to novel input queries to solve a task (Wei et al., 2021). We provide the LLM with a text-only prompt $\mathcal{C} = \mathcal{T}_{\ell}(d_{\mathcal{I}}, E, \mathcal{A}_E, d_{o_i})$. Here, \mathcal{T}_{ℓ} is a predefined template that converts the input information including the generated object role description d_{o_i} from the previous stage, event E , image description $d_{\mathcal{I}}$ ¹, and the possible argument roles \mathcal{A}_E to detailed task description for the LLM. We illustrate an example input prompt at the top of Figure 2 (right).

The template \mathcal{T}_{ℓ} is designed in a way that prompts the LLM to output the predicted event argument role \tilde{a}_i for the object o_i . Following this, we can evaluate if the predicted EARL matches with the ground-truth argument role label a_i . We show the example predicted EARL in the lower half of Figure 2 (right). In our work, we use ChatGPT (OpenAI, 2021), a state-of-the-art LLM that can adapt to new tasks and perform well on them.

¹ $d_{\mathcal{I}}$ is the image caption generated using the GVLM \mathcal{G} with the prompt ‘Describe this image in detail.’.

4 Experimental Setup

4.1 Dataset

We evaluate our approach on two datasets: M²E² (Li et al., 2020) and SWiG (Yatskar et al., 2016). In both these datasets, each image has an associated event type, a bounding box for the participating objects, and their associated event argument roles.

M²E² consists of 391 images with event mentions taken from the Voice of America (VOA) news website² where the event types and their argument roles are derived from the ACE ontology (Doddington et al., 2004). To maintain a good argument role density for evaluation, we filter the event types with less than three argument roles (Appendix Table 6). In total, we label 990 bounding boxes spanning 275 images in the M²E² dataset.

SWiG dataset consists of images that can belong to one of 504 action verbs. Following Li et al. (2022b), we consider the action verbs to be analogous to the event types. Like our M²E² setup, we filter the action verbs with less than three event argument roles (Appendix Table 7). In our experiments, we label 1600 bounding boxes spanning 600 test images from this dataset.

Following Li et al. (2022b), we do not consider ‘place’ as an event argument role since it cannot be grounded as a bounding box for both datasets. In the zero-shot setting, we evaluate the GenEARL’s capability to reason about the examples, without training on event data, by prompting the pretrained generative models with event information, its associated argument role labels, and the visual context from the image and the object. In the few-shot setting, we provide the k-shot examples for all the bounding boxes in k images sampled randomly from the dataset. For example, in case k = 1, we show the labels for the objects in one of the images from the dataset (Appendix Table 12).

4.2 Implementation Details

GenEARL is a two-stage framework that consists of a vision-language model and a large language model. In our experiments, we use LLaVA-7B (Liu et al., 2023a), a state-of-the-art generative vision-language model. It is finetuned with 150K multimodal instructions that enable it to understand a wide breadth of concepts and generate helpful and coherent outputs. Traditionally, LLaVa is used for generating outputs for a single image and text

²<https://www.voanews.com/>

inputs, however, we adapt it to reason over multiple images during inference i.e., an image depicting the event and a bounding box containing the object to be labeled. Due to this design, the multimodal input prompt can flexibly incorporate the input instance details to generate the event-centric object descriptions with any training on the event-annotated data. Existing models (Li et al., 2022b) cannot adapt to new event types or new domains without further finetuning on an event-annotated dataset.

For the second-stage of our framework where we predict the event argument role label based on the generated event-centric object description, we use ChatGPT (OpenAI, 2021) as our default LLM. It is a powerful LLM that can adapt to novel task descriptions and perform well on them. In addition, it is capable of learning about the target domains through few-shot examples in its context. In our work, we leverage this capability by prompting the LLM with a few-shot solved examples. Since prompting ChatGPT comes with cost considerations³, we label the event argument roles in batches where all the participating objects $o_i \in \mathcal{O}$ in an event depicting image \mathcal{I} form a single batch. We access this model through OpenAI’s API⁴ where it corresponds to ‘gpt-3.5-turbo’. In §5.2, we use Alpaca (Taori et al., 2023) by replacing ChatGPT in the second-stage for multimodal EARL.

4.3 Baselines

We compare the GenEARL framework against the open-source CLIP models (Radford et al., 2019). We evaluate our framework against zero-shot CLIP to understand the generalization capability of our model without training on any event-annotated data. Subsequently, we train a supervised CLIP model to assess the gap between our training-free paradigm and a supervised model that learns about the events from event annotated data.

Following Li et al. (2020), we compare against the large-scale representation learning vision-language models like CLIP-B/32 and CLIP-L/14 (Radford et al., 2019) since they can leverage any event knowledge acquired during its pretraining for multimodal EARL in a zero-shot setting.

Consider an event E depicted in an image, a set of argument role labels $\mathcal{A}_E = \{a_1, \dots, a_m\}$ for the event, and the object o_i that needs to be labeled. We calculate the similarity score $s(o_i, T_K(a_j))$ be-

³<https://openai.com/pricing>

⁴<https://openai.com/blog/openai-api>

multiple templates with event and role details

Paradigm	Method	M ² E ² Accuracy (%)	SWiG Accuracy (%)
Training Free	CLIP-B/32 (0-shot)	28.1	22.6
	CLIP-L/14 (0-shot)	30	24.7
	GenEARL (0-shot)	39.4	38.9
	GenEARL (1-shot)	44.1	50.1
	GenEARL (3-shot)	45.8	51
Supervised	CLIP-B/32 (+ SWiG)	41.9	67.1

Table 1: Multimodal EARL accuracy (%) of GenEARL (zero-shot to three-shot), pretrained CLIP (zero-shot), and supervised CLIP on the M²E² and SWiG dataset. Zero-shot GenEARL framework outperforms the zero-shot CLIP models on both datasets. The performance of our framework improves from zero-shot to three-shot example without any training of the underlying models on the event-annotated data. A supervised baseline is only used as a reference as it accesses event-annotated data. We present precision, recall, and F1 scores for the same in Appendix §C.

tween the object representation in the image embedding space and the text embedding of a set of text-based templates T_K containing information about the event and event argument role. Specifically, our template will be ‘An object playing a_j role in the E event.’ for every argument role label a_j in \mathcal{A}_E . The predicted argument role label \tilde{a}_i will be the one that maximizes the similarity score between the object representation and its corresponding text template i.e., $\tilde{a}_i = \text{argmax}_j(s(o_i, T_K(a_j))$.

To impart the knowledge of the event semantics and structures to the pretrained baseline, we find a CLIP-B/32 model on the 75K images of the SWiG dataset. We create a text-based prompt with the event and argument role label for every example, as describe above. Finally, the model is finetuned with the contrastive pretraining objective to bring the representations of the objects and their EARL together in the joint embedding space and pull apart the representations of the unmatched objects and event roles. We provide the finetuning details in Appendix §B.1. Given the lack of training data for the M²E² domain, we conduct an evaluation of CLIP (+SWiG) as a transfer learning setup.

5 Experimental Results

5.1 Main Results

We evaluate the accuracy of the GenEARL framework for multimodal EARL on the M²E² and SWiG datasets, in Table 1. We find that zero-shot GenEARL achieves a multimodal EARL accuracy of 39.4% and 38.9% on the M²E² and SWiG datasets, respectively, without any specialized training on the event-annotated data. We observe that zero-shot GenEARL framework outperforms the CLIP-L/14 baseline by 9.4% on M²E² and 14.2% on SWiG dataset. This highlights the strong generalization ability of the generative models to adapt

to multimodal EARL task descriptions and perform well on them.

Prior works have shown that the LLM are capable of improving their reasoning capabilities by learning from the few-shot examples in their context (Dong et al., 2022; Brown et al., 2020). Hence, we prompt the ChatGPT model in the GenEARL framework with one-shot and three-shot settings. We observe that even a single-shot example in the context greatly enhances the performance of the framework to 44.1% and 50.1% on the M²E² and SWiG datasets, respectively. In addition, we find that providing three-shot also slightly improves the EARL accuracy over the one-shot setting. Due to the context length limitations and cost considerations, while using the OpenAI API, we do not experiment with more than three-shot examples. Our results highlight the unique capability of our framework to improve its performance with a very small amount of event-annotated data.

We further aim to understand the performance improvements and generalization of the pretrained CLIP baseline when it is finetuned with the SWiG training data, which was originally acquired by expensive annotation from the human labelers. We observe that our framework reduces the gap between the supervised paradigm and training-free paradigm from 44.5% to 16.1% on the SWiG dataset. In addition, our framework outperforms the finetuned CLIP model on the M²E² dataset by 3% accuracy with just three-shot examples. Thus, our framework provides a flexible, training-free, and generalizable method for accurate multimodal EARL that can easily adapt to new event types. We provide a few qualitative examples for model predictions in Appendix Figure 5.

Comparison with CLIP-Event. We propose GenEARL as a multimodal EARL framework where

if feel 1-shot
works best to
understand the
output
formatting.

3-shot is
therefore not
giving more info
on the results

Method	M ² E ² Accuracy (%)	SWiG Accuracy (%)	Average (%)
GenEARL w/ ChatGPT (0-shot)	39.4	38.9	39.2
GenEARL w/ Alpaca-7B (0-shot)	50.4	36.5	43.5
GenEARL w/ ChatGPT (1-shot)	44.1	50.1	47.1
GenEARL w/ Alpaca-7B (1-shot)	54.6	37.5	46.1
LLaVA-7B (0-shot)	28.6	26	27.3

Table 2: Multimodal EARL accuracy (%) of the GenEARL framework with the choice of LLM (§5.2). The last row highlights the performance if we were to only use the GVLMLM for multimodal EARL (§5.4).

we assume access to the ground-truth event depicted in the image. This differs from prior work **CLIP-Event** (Li et al., 2022b) that performs event detection followed by EARL. Hence, we use a combination of the CLIP-B/32 for event type detection and GenEARL framework for EARL to directly compare with the reported CLIP-Event results.⁵

Method	M ² E ²		
	Precision	Recall	F1
CLIP-Event	21.1%	13.1%	17%
CLIP + GenEARL (0-shot)	42.4%	24.1%	31.4%
CLIP + GenEARL (3-shot)	43%	29%	34.6%

Table 3: Comparison between the combination of CLIP (for event detection) and GenEARL (for EARL) with CLIP-Event that performs end-to-end event detection followed by EARL.

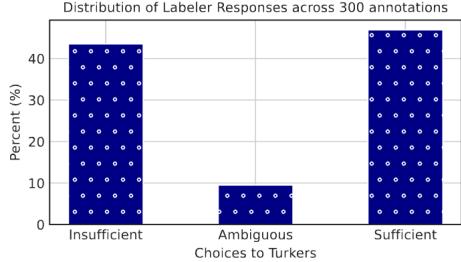
In Table 3, we find that the combination of CLIP and GenEARL framework outperforms CLIP-Event on precision and recall by 22%, 15%, 17% on precision, recall, and F1 respectively. With advancements in the event detection modules, the performance of the complete event detection and GenEARL pipeline will further improve.

5.2 The Impact of Large Language Models

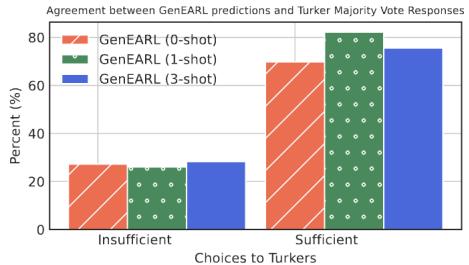
Here, we aim to study whether ChatGPT can be replaced by a relatively smaller instruction-following LLM like **Alpaca-7B** (Taori et al., 2023) in our GenEARL framework. We provide the details for predicting EARL using Alpaca in Appendix §E.

We compare the EARL accuracy with ChatGPT and Alpaca in the GenEARL framework under the zero-shot and one-shot settings in Table 2. We observe that Alpaca-7B outperforms ChatGPT in the zero-shot setting (43.5% vs 39.2%) while lagging behind ChatGPT in the one-shot setting (46.1% vs 47.1%) in the average EARL accuracy across the datasets.⁶ It suggests that ChatGPT is able to

improve more with the few-shot examples, while the base performance of Alpaca is already high. Specifically, we find that Alpaca-7B outperforms ChatGPT by a large margin on the M²E² dataset suggesting that it understands the events grounded in the news data better than ChatGPT. We observe that ChatGPT outperforms Alpaca-7B by a large margin on the SWiG dataset which suggests that ChatGPT is better at reasoning over the event types associated with the common action verbs. Such differences in the understanding of the event domains can be attributed to the difference in the base language models. InstructGPT (Ouyang et al., 2022) for ChatGPT and LLaMA (Touvron et al., 2023) for Alpaca, and their pretraining data.



(a) Distribution of labeler responses



(b) Agreement between GenEARL prediction and the quality of the object role descriptions assessed by the human labelers.

Figure 3: Human assessment of the quality of the generated object role descriptions from the LLaVA model.

⁵The code for CLIP-Event is publicly available, however, its pretraining dataset or pretrained checkpoint are unavailable.

⁶We do not perform a three-shot evaluation with Alpaca

due to the limitations of its context length and ability to capture long-range dependencies.

5.3 Human Assessment

We hypothesize that the GenEARL’s performance depends on the quality of the object role descriptions generated by the GVLM (§3.1). To verify this, we perform a human assessment for 100 instances of the generated object descriptions from the M²E² dataset. We ask the human labelers to choose whether the generated object role descriptions provide sufficient information for a human or LLM to correctly label the object’s event argument role. We conduct the experiment using 4 labelers from Amazon’s Mechanical Turk (MTurk) platform.⁷ Prior to the annotation, each labeler passed a qualification test, involving 10 sample instances, different from the actual 100 instances, to assess their understanding of the task. We provide the annotators with a detailed slide deck describing the task with a few solved examples. We attach a screenshot for one of the solved examples in Figure 14. We assign 3 labelers per instance to evaluate the annotator agreement and get majority votes on the description quality. The annotators were paid at a rate of 18\$ per hour⁸.

We find that the agreement between any two annotators averaged over 100 instances was 75% which is high given the subjective complexity involved in assessing the dense object role descriptions. In Figure 3a, we provide the distribution of all the labeler responses for 300 annotations (100 instances \times 3 labelers). We observe that the labelers find 47%, 10%, 43% of the object role descriptions sufficient, ambiguous, and insufficient, respectively, to correctly label the object’s argument role. This suggests that there is a scope for enhancements in the quality of the generated object descriptions with advancements in the GVLM design and prompting techniques. In Figure 3b, we assess the performance of the GenEARL framework on the instances that were marked containing ‘insufficient’ and ‘sufficient’ information to accurately predict the object role labels. We find that the results verify our hypothesis since the performance of the framework is $\sim 80\%$ and $\sim 20\%$ on the high and low-quality examples, respectively. The result highlights that innovations in the generative VLM will improve the performance of the framework.

5.4 Ablations

EARL via GVLM In our framework, we utilize LLaVA-7B for generating the event-centric object role descriptions that subsequently prompt the LLM for multimodal EARL prediction. Here, we study whether we can perform multimodal EARL without GenEARL’s two-stage pipeline. To answer this, we use LLaVA-7B to label the event argument roles for the participating objects given the visual features of the image and object, and the event and its associated argument roles. We provide more details for predicting EARL using LLaVA-7B in Appendix §F.

We present our findings regarding multimodal EARL with LLaVA-7B under the zero-shot setting, as shown in the last row of Table 2. Our analysis reveals that LLaVA-7B exhibits subpar performance when compared to the GenEARL with LLMs on both datasets. The underperformance of LLaVA-7B in the novel event argument role labeling task can potentially be attributed to its finetuning data, which lacked classification-oriented instructions and instead emphasized providing detailed image descriptions following the input instruction. Consequently, we leverage LLaVA’s capability to generate event-centric object descriptions when prompted, which can be effectively utilized by a robust LLM within the GenEARL framework. This LLM is likely to have encountered a greater number of classification-style instructions during its finetuning process which leads to its labeling capabilities.

Effect of the Conditioning Variables In the previous experiments, we show that the quality of the generated object role descriptions affects the predictions made by the LLM in the GenEARL framework. In this experiment, we aim to study the factors that affect this quality. Specifically, we compare the performance of our framework where object descriptions are generated (a) without the visual features of the event depicting image \mathcal{I} , and the event details including the event E and argument role label and their definitions \mathcal{A}_E , and (b) without the visual features of the event depicting image \mathcal{I} . In both scenarios, the visual features of the object o_i being labeled are present in the input prompt to the GVLM. The input template for these settings are present in Table 9 and Table 10.

We report the results in Table 4 by prompting ChatGPT under the one-shot setting. We find that the performance of the GenEARL framework with

⁷<https://www.mturk.com/>

⁸The total cost of human annotations was close to \$100.

GVLM prompting	M ² E ² (%)	SWiG (%)
w/ \mathcal{I} , (E, \mathcal{A}_E) [Ours]	44.1	51
wo/ \mathcal{I}	42.1	44
wo/ \mathcal{I} , (E, \mathcal{A}_E) [Object Caption]	39.9	43

Table 4: Multimodal EARL accuracy (%) on the M²E² and SWiG dataset. We compare three GVLM prompting strategies to understand the effect of various conditioning variables used in the multimodal prompt to GVLM for generating object descriptions. The experiment is performed in the one-shot setting.

the object descriptions generated without any context of the event image and event details i.e., object caption performs the worst on both datasets (Row 3). We observe that providing the event details improves the performance on both datasets (Row 2 and Row 3). Finally, we observe that the best performance is achieved when the GVLM is provided with the event image and event information in its context where the accuracy gains are higher for the SWiG dataset than the M²E² dataset. Our experiment reveals that we need to provide event-oriented information in the input prompt to generate event-centric object descriptions that will eventually contribute to improved multimodal EARL. Finally, we observe that our results are robust to the ordering perturbations in the predefined template to the LLM in Appendix §D, which makes our approach flexible.

6 Related Work

Event Extraction: Most of the prior work has focused on extracting events and their structures grounded in the text modality such as text documents (Ahn, 2006; Nguyen et al., 2016; Nguyen and Grishman, 2015; Nguyen and Nguyen, 2019; Lin et al., 2020; Yang and Mitchell, 2016; Paolini et al., 2021; Li et al., 2021). These works depend on training their models on the large-scale dataset with a complete event annotation (Doddington et al., 2004; Song et al., 2015). Previous studies have shown that these models do not generalize to unseen event types and new domains (Parekh et al., 2023). Other works in text-based event extraction have thus focused on training their models in a data-efficient manner (Lu et al., 2021; Hsu et al., 2022). In this work, we utilize modern generative models that are capable of understanding the event task descriptions with image contexts and performing well on them, thus keeping our approach training-free.

Multimodal Event Extraction: While traditional event extraction has focused on text modality, multimodal event extraction is popularized by

recent works (Yatskar et al., 2016; Li et al., 2020, 2022b). These works focus on training weakly supervised representation models such as WASE and CLIP-Event on event annotated data (Doddington et al., 2004; Yatskar et al., 2016; Li et al., 2020) to learn event semantics and structures. Similar to the previous works in text-based event extraction, these models will suffer from generalization issues with new event types. Additionally, there is no easy way to flexibly adapt them to new event types without further finetuning on expensive and hard-to-acquire event data. In this work, we show that GenEARL generalizes to the unseen events by reasoning over the visual features of the image and object, and the event and its argument roles in its context.

Generative Models: Recently, there has been a surge of large-scale pretrained language models (Brown et al., 2020; OpenAI, 2021; Touvron et al., 2023; Zhang et al., 2022; Chowdhery et al., 2022; Hoffmann et al., 2022; Lewis et al., 2019; Raffel et al., 2020) and multimodal generative models (Liu et al., 2023b; Li et al., 2022a, 2023; Dai et al., 2023; Awadalla et al., 2023; Alayrac et al., 2022; Zhu et al., 2023; Gong et al., 2023). These models have a remarkable ability to generalize to new tasks based on their descriptions in their context. Wang et al. (2022) use a combination of multimodal generative models and large language models to perform strong baselines on the few-shot video-language benchmarks. In our work, we leverage the flexibility and generalization capabilities of these models to perform well on multimodal EARL.

7 Conclusion

We propose GenEARL, a training-free framework for multimodal event argument role labeling. Our experiments reveal that GenEARL generalizes to various event types from the two datasets M²E² and SWiG. This generalization capability is desirable for event extraction models since the existing methods rely on event-annotated data, which is expensive and time-consuming to acquire. Our work relies on the input prompts that follow the predefined human-crafted templates. Future work can focus on the automation of such templates based on the event information. Existing text-only and multimodal generative models (Bansal et al., 2022; Pratt et al., 2020; Nadeem et al., 2020) have been shown to encode harmful social biases. Future work should aim to contextualize social biases in multimodal EARL, and focus on reducing them.

8 Acknowledgement

Hritik Bansal and Po-Nien Kung are supported in part by AFOSR MURI grant FA9550-22-1-0380. In addition, this work was partially supported by Defense Advanced Research Project Agency (DARPA) grant HR00112290103/HR0011260656, CISCO and ONR grant N00014-23-1-2780.

Limitations

GenEARL framework prompts modern vision-language generative models and large language models with event-specific task descriptions. These models have shown to encode harmful social biases and stereotypes that may be reflected in their output generations. For example, people with a specific color in an image might get labeled as ‘attackers’ or ‘offenders’ without understanding their true role in the depicted event. Future work should focus on the study of these generated object descriptions from GVLM and EARL predictions from the LLMs through human evaluation.

In our experiments, we perform the human assessment by employing mechanical turkers. All the labelers belonged to the US and thus will limit our assessment due to their own perceptual biases when presented with an event depicting an image and the participant objects. To obtain more reliable annotations, we can involve annotators from diverse regions.

Multimodal event extraction is a relatively new endeavor in comparison to traditional text-based event extraction methods. Thus, there are very few evaluation datasets to assess the multimodal models for their generalization capabilities to diverse domains and event types. It will be imperative to benchmark our approach on such diverse datasets for more comprehensive evaluation.

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, Mikołaj 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](#).
- Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. 2022. How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230*.
- Philipp Blandfort, Desmond U Patton, William R Frey, Svebor Karaman, Surabhi Bhargava, Fei-Tzin Lee, Siddharth Varia, Chris Kedzie, Michael B Gaskell, Rossano Schifanella, et al. 2019. Multimodal social media analysis for gang violence prevention. In *Proceedings of the International AAAI conference on web and social media*, volume 13, pages 114–124.
- 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.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.
- 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.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.
- Salvatore Giorgi, Sharath Chandra Guntuku, McKenzie Himelein-Wachowiak, Amy Kwarteng, Sy Hwang, Muhammad Rahman, and Brenda Curtis. 2022. Twitter data of the #blacklivesmatter movement and counter protests: 2013 to 2021.
- Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. [Multimodal-gpt: A vision and language model for dialogue with humans](#).

- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- 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.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- 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*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.
- Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022b. 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. *arXiv preprint arXiv:2005.02472*.
- Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. *arXiv preprint arXiv:2104.05919*.
- Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.
- Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023c. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. *arXiv preprint arXiv:2106.09232*.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.
- Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.
- Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 365–371, Beijing, China. Association for Computational Linguistics.
- 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.
- OpenAI. 2021. ChatGPT: Large-scale language model. Accessed: June 17, 2023.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *arXiv preprint arXiv:2101.05779*.
- Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023. Geneva: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. In *Proceedings of the Conference of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

- 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, pages 314–332. Springer.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ere: annotation of entities, relations, and events. In *Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98.
- Mitchell Stephens et al. 1998. *The Rise of the Image, the Fall of the Word*. Oxford University Press.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. 2022. Language models with image descriptors are strong few-shot video-language learners. *arXiv preprint arXiv:2205.10747*.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- Bishan Yang and Tom M. Mitchell. 2016. Joint extraction of events and entities within a document context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 289–299, San Diego, California. Association for Computational Linguistics.
- 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*, pages 5534–5542.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Devan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Multimodal EARL

Firstly, we assume that a given image has a single major event associated with it since there can be no event or multiple events depicted in the image. Secondly, we assume that every object (bounding box) participates in the event and hence has an associated event argument role label. However, we do incorporate the ability to predict ‘Other’ if the model does find a good match in the candidate argument role labels. Thirdly, we allow multiple objects in a given image to have the same event argument role label, as shown in Figure 1.

Existing methods (Li et al., 2020, 2022b) train the models on high-quality unimodal or multimodal data annotated with the events, entities, and their relations. For example, ACE is a text-based dataset, imSitu (Yatskar et al., 2016) is for actions (or events) in the images, and (Li et al., 2022b) utilizes VOA news as a source for multimedia events. In practice, collecting and annotating these datasets is time-consuming and expensive. Furthermore, the inability of the existing models to adapt to new event types and new domains without specialized training data makes them less generalizable.

B Setup-Baseline

B.1 Supervised Setting

We train our model with a batch size = 32, epochs = 5, AdamW optimizer (Loshchilov and Hutter, 2017) with a linear warmup of 100 steps followed by cosine annealing. We perform a hyperparameter search over three learning rates = $\{1e - 4, 1e - 5, 1e - 6\}$. We find that the model trained with $1e - 5$ performs the best on the test examples of the SWiG dataset during inference. Each finetuning experiment took 4 – 5 hours on a single A6000 48GB Nvidia GPU.

Given the lack of available training data for the M²E² domain, we conduct an evaluation of CLIP (+SWiG) as a transfer learning setup. Our findings reveal that the finetuned CLIP (+SWiG) model, utilizing a learning rate of $1e - 4$, demonstrates the most optimal performance on the M²E² dataset.

C Precision, Recall, and F1 Metrics

In Table 5, we report the precision, recall and F1 of GenEARL, Zero-shot CLIP baseline, and Supervised CLIP finetuned on the M²E² and SWiG dataset. We find that the trends (rankings) of the multimodal EARL methodologies are the same as

reported in Table 1. Specifically, we observe that GenEARL (3-shot) outperforms Supervised CLIP under the transfer learning settings on M²E². In addition, we find that the GenEARL (0-shot and 3-shot) outperform the zero-shot CLIP baseline on the SWiG dataset.

D Effect of ordering perturbation in the predefined template

In §3.2, we describe a predefined template \mathcal{T}_ℓ that converts the input variables (object role description, event type, image description, possible argument roles and their definitions) from the generative vision language model to detailed task description for LLM. In our experiments, we find that the LLM (ChatGPT) predictions are not sensitive to the order in which the input variables are presented to it in the predefined template. Specifically, we observe a slight change of $\pm 2\%$ for GenEARL (3-shot) on both datasets over different orderings of the input variables to the LLM template. This indicates that the predefined template is very flexible and can easily generalize to new datasets.

E Using Alpaca for EARL Prediction

We evaluate the probability scores from the Alpaca model \mathcal{M} over the possible event argument roles $a_j \in \mathcal{A}_E$ given by $p_{\mathcal{M}}(a_j|\mathcal{C})$ where $p_{\mathcal{M}}(\cdot|\mathcal{C})$ is the learned probability distribution of the Alpaca model, and \mathcal{C} is the context described in §3.2. The predicted EARL \tilde{a}_i for the object o_i is the one that maximizes the probability $\tilde{a}_i = \text{argmax}_j p_{\mathcal{M}}(a_j|\mathcal{C})$. We provide the EARL prediction template used for Alpaca experiments in the zero-shot settings in Table 13.

F Using LLaVA for EARL Prediction

To this end, we evaluate the probability scores from the LLaVA model \mathcal{G} over the possible event argument roles $a_j \in \mathcal{A}_E$ given by $p_{\mathcal{G}}(a_j|\mathcal{P})$ where $p_{\mathcal{G}}(\cdot|\mathcal{P})$ is the learned probability distribution of the LLaVA model, and \mathcal{P} is the context described in §3.1. The predicted EARL \tilde{a}_i for the object o_i is the one that maximizes the probability $\tilde{a}_i = \text{argmax}_j p_{\mathcal{G}}(a_j|\mathcal{P})$. We provide the EARL prediction template used for LLaVA in the zero-shot settings in Figure 4.

G Input Prompts for LLaVA

We provide the multimodal input templates used for prompting LLaVA in the GenEARL framework.

Method	M ² E ²			SWiG		
	Precision	Recall	F1	Precision	Recall	F1
Zero-shot CLIP-B/32	41.6%	28%	33.2%	22%	27.3%	24.4%
GenEARL (0-shot)	65.3%	38.4%	48.4%	36.5%	47%	41%
GenEARL (3-shot)	66.1%	44.7%	53.3%	47%	55%	50.7%
Supervised CLIP-B/32 (+SwiG)	66%	40.3%	50%	61.5%	72%	66%

Table 5: Precision, Recall, and F1 scores for Multimodal Event Argument Role Labeling on M²E² and SWiG datasets.

Input: Image \mathcal{I} , Object o , Event E , Event argument role labels (and definitions) \mathcal{A}
Prompt:
Image 1: $256 \times <\mathcal{I}$ visual feature tokens>
Image 2: $256 \times < o$ visual feature tokens>
What is the role of the entity in the "Image 2" in the context of the E event in the "Image 1"? The possible argument roles of the objects in the E event include \mathcal{A} . Choose only one of these options.

Figure 4: Multimodal input template used to prompt the GLVM for event argument role labeling. Here, the model is provided with the image, event and possible argument role labels (and definitions). We add ‘Other’ to the list of possible argument role labels in case it does not prefer any of the existing event argument role labels. We get 256 visual tokens for the image or the object by projecting the raw input into the vision embedding space using the visual processing module of the GVLM.

By default, we use the template presented in the Figure 8. We use the templates Figure 9 and Figure 10 for the experiments in §5.4. Since LLaVA is known to have a limited context token length of 2048 and like most transformer architectures (Tay et al., 2020) is susceptible to suffer from ignoring long-range dependencies, we create object role descriptions for one object o_i at a time.

H Event Types

To maintain the argument role density, we filter the event types with less than three argument roles from the M²E² dataset and the SWiG dataset. We provide the list of event types included in our multimodal EARL evaluation in Table 6 and Table 7.

I Qualitative Example

We provide three qualitative examples from the M²E² dataset in Figure 5.

J Human Assessment Screenshot

Here, we provide the screenshots for our human assessments. Figure 14 illustrates a solved example

shown to the labelers before the qualification test to make them understand the task. Figure 15 illustrates an unsolved example assigned to 3 labelers during the actual annotation.

K Can two-stage generative prompting perform multimodal event detection?

In this work, we present GenEARL, a combination of prompting a generative vision-language model followed by a large language model for multimodal EARL. Here, we aim to study whether the same principles can be adapted to perform multimodal event detection i.e., classifying the given image into one of the event types from the dataset.

To this end, we first generate an image caption from LLaVA-7B. Then, we prompt ChatGPT to predict the event type based on the generated image description and the list of possible event types (and their definitions). We perform this experiment in a zero-shot setting. Further, we compare our model against the zero-shot pretrained CLIP (Radford et al., 2019), and zero-shot and supervised CLIP-Event (Li et al., 2022b). We report the results in Table 8.

We find that the two-stage generative prompting framework outperforms all the other methods including the supervised baselines for multimodal event detection on the M²E² dataset. This suggests that generative prompting is a flexible and generalizable technique that can be used for event detection without access to any event-annotated training data. Surprisingly, we find that the two-stage framework performs poorly on the SWiG dataset. We perform a manual inspection of several examples where the predicted action does not match the ground action. Our investigation reveals that the two-stage generative prompting does provide reasonable predictions for the actions depicted in the images. For example, as shown in Figure 7, the ground-truth action annotated in the SWiG dataset is ‘dialing’. However, the two-stage setup

Event Types	Event Argument Role Definition
Life: Die	[Victim] died at place, or was killed by [Agent] using [Instrument]
Movement.Transport	[Agent] transported [Artifact or Person] in [Instrument] from [Origin] to [Destination]
Conflict.Attack	[Attacker] attacked or assaulted [Target] [Instrument] at place
Conflict.Demonstrate	[Demonstrator] was in a demonstration at place holding [Instrument] and supervised by [Police]
Justice.Arrest.Jail	[Agent] arrested or jailed [Person] using [Instrument] at place
Transaction.Transfer.Money	[Giver] gave [Money] to [Recipient]

Table 6: The list of six event types from the M²E² dataset included in our experiments. The argument role definitions are taken from (Li et al., 2020).

Image	Object	Event	Generated Image Description	Generated Object Description	Predicted Event Argument Role Label
		Movement:Transport Possible Args: Agent Artifact/Person Vehicle Origin Destination	The image shows a group of people wearing life jackets gathered together in the water. They appear to be having fun and enjoying themselves, possibly on a raft or a large inflatable floating device. There are at least 13 people in the water, with some of them closer to the surface while others are submerged. The group is spread out, with some individuals located closer to the left side of the image, others in the middle, and a few more towards the right side. The people are interacting and engaging with each other, creating a lively atmosphere in the water.	In the context of the Movement:Transport event, the inflatable raft appears to be a vehicle used for transportation. In 'Object', a man is inside the raft, either preparing to get on or already on the raft. The other people in the image might be either participants, event organizers, or bystanders observing the event. The inflatable raft's presence in the water suggests that it is part of the transport or transportation component, providing an alternative mode of travel for the participants, and contributing to the overall experience of the event.	Vehicle
		Transaction:Transfer-Money Possible Args: Giver Recipient Money	The image features three people standing close to each other, with one person in the center and the other two positioned on the left and right. They are holding a total of two large bills, and appear to be examining them or possibly exchanging them. The bills are of various sizes and orientations, suggesting that they might be from different parts of the world. The people seem to be focused on their task, likely discussing or comparing the bills' details.	In the context of the Transaction:Transfer-Money event, the entity in 'Object' is playing the role of the 'Recipient'. The person is receiving the ten-dollar bill from the other person, who is likely playing the role of the 'Giver'. This transfer of value is a typical scenario in everyday transactions and demonstrates how individuals exchange goods or services using physical currency.	Recipient
		Conflict:Demonstrate Possible Args: Demonstrator Instrument Police	The image features a group of people standing on a rooftop, with one person prominently holding a large flag, possibly a national or protest flag. The person holding the flag appears to be waving it, drawing attention from the others around them. In total, there are at least 13 people visible on the rooftop, gathered together and enjoying the event. The scene suggests a sense of unity and celebration among the group.	In the 'Object', a person is holding a large, written-in-Arabic flag, possibly representing a pro-Syrian or pro-Palestinian demonstration. The flag is a central symbol of the protest. The Conflict:Demonstrate event suggests that the entity is participating in the protest to show support for the cause. The flag serves as an important symbol for the demonstrators, emphasizing their unity and solidarity in advocating for their respective rights or opposing a conflict. The entity's role in the event is to raise awareness and express their stance on the issue by waving the flag and participating in the demonstration alongside other protesters.	Demonstrator

Figure 5: Qualitative Example for the predictions by the GenEARL framework in three-shot setting.

predicts ‘pressing’ and ‘navigating’ actions which are legitimate through visual inspection of Figure 7. Although we do not conduct a large-scale human evaluation of our approach, we attribute the low performance of the two-stage generative feedback on SWiG dataset to its restricted ground-truth action verb assignments (‘dialing’) despite the presence of other related actions in the list of event types (‘pressing’, ‘typing’). Although not shown here, we find that providing few-shot examples to ChatGPT does not improve the performance on the SWiG dataset.

‘tattooing’, ‘splashing’, ‘emerging’, ‘pasting’, ‘inflating’, ‘displaying’, ‘feeding’, ‘shredding’, ‘flicking’, ‘shaving’, ‘flinging’, ‘yanking’, ‘putting’, ‘decorating’, ‘mashing’, ‘baking’, ‘moistening’, ‘inserting’, ‘chopping’, ‘parachuting’, ‘giving’, ‘cooking’, ‘falling’, ‘ramming’, ‘pushing’, ‘wrapping’, ‘leaping’, ‘packing’, ‘smearing’, ‘whisking’, ‘sharpening’, ‘sealing’, ‘examining’, ‘raking’, ‘punching’, ‘whipping’, ‘watering’, ‘steering’, ‘distributing’, ‘spying’, ‘uncorking’, ‘squeezing’, ‘pawing’, ‘dripping’, ‘carrying’, ‘pinching’, ‘tying’, ‘arranging’, ‘tearing’, ‘blocking’, ‘practicing’, ‘potting’, ‘installing’, ‘dousing’, ‘making’, ‘unpacking’, ‘mowing’, ‘stacking’, ‘stinging’, ‘heaving’, ‘pulling’, ‘breaking’, ‘docking’, ‘destroing’, ‘drying’, ‘poking’, ‘tickling’, ‘tilting’, ‘hoisting’, ‘nailing’, ‘lighting’, ‘shelving’, ‘wagging’, ‘lifting’, ‘surfing’, ‘gluing’, ‘hauling’, ‘erasing’, ‘bathing’, ‘serving’, ‘drinking’, ‘soaking’, ‘writing’, ‘coaching’, ‘flipping’, ‘slicing’, ‘educating’, ‘bothering’, ‘sweeping’, ‘washing’, ‘spinning’, ‘dyeing’, ‘filling’, ‘shaking’, ‘signing’, ‘catching’, ‘scratching’, ‘shearing’, ‘ejecting’, ‘peeling’, ‘hanging’, ‘sliding’, ‘igniting’, ‘pumping’, ‘clenching’, ‘fastening’, ‘repairing’, ‘jumping’, ‘striking’, ‘manicuring’, ‘farming’, ‘photographing’, ‘mining’, ‘selling’, ‘bouncing’, ‘leaking’, ‘autographing’, ‘scraping’, ‘providing’, ‘tipping’, ‘patting’, ‘interrogating’, ‘emptying’, ‘hurling’, ‘scooping’, ‘sewing’, ‘massaging’, ‘attacking’, ‘signaling’, ‘dissecting’, ‘spanking’, ‘hugging’, ‘crashing’, ‘coloring’, ‘scrubbing’, ‘rinsing’, ‘fueling’, ‘begging’, ‘spraying’, ‘frying’, ‘stitching’, ‘spilling’, ‘covering’, ‘dipping’, ‘wheeling’, ‘constructing’, ‘burying’, ‘throwing’, ‘drenching’, ‘injecting’, ‘pricking’, ‘clearing’, ‘plummeting’, ‘wetting’, ‘weighing’, ‘deflecting’, ‘opening’, ‘lapping’, ‘vacuuming’, ‘kicking’, ‘picking’, ‘folding’, ‘operating’, ‘smashing’, ‘stroking’, ‘pouring’, ‘tuning’, ‘dusting’, ‘attaching’, ‘dropping’, ‘resting’, ‘rocking’, ‘paying’, ‘guarding’, ‘leaning’, ‘kissing’, ‘slapping’, ‘harvesting’, ‘piloting’, ‘welding’, ‘stuffing’, ‘rubbing’, ‘buttering’, ‘pruning’, ‘pinning’, ‘dragging’, ‘fetching’, ‘locking’, ‘shoveling’, ‘cramming’, ‘disciplining’, ‘planting’, ‘cleaning’, ‘spitting’, ‘climbing’, ‘caressing’, ‘flexing’, ‘drawing’, ‘assembling’, ‘lathering’, ‘mending’, ‘loading’, ‘combing’, ‘unloading’, ‘trimming’, ‘molding’, ‘checking’, ‘descending’, ‘aiming’, ‘performing’, ‘fishing’, ‘taping’, ‘immersing’, ‘ailing’, ‘sketching’, ‘grinding’, ‘filming’, ‘measuring’, ‘eating’, ‘buckling’, ‘submerging’, ‘helping’, ‘crafting’, ‘drumming’, ‘plunging’, ‘carting’, ‘curling’, ‘spreading’, ‘clipping’, ‘carving’, ‘hitting’, ‘retrieving’, ‘placing’, ‘launching’, ‘crushing’, ‘pitching’, ‘offering’, ‘strapping’, ‘adjusting’, ‘moisturizing’, ‘tasting’, ‘wiping’, ‘sprinkling’, ‘buying’, ‘applying’, ‘stapling’, ‘stirring’, ‘floating’, ‘shooting’, ‘microwaving’, ‘fixing’, ‘milking’, ‘brushing’, ‘vaulting’, ‘unlocking’, ‘extinguishing’, ‘building’, ‘painting’, ‘waxing’, ‘stripping’, ‘prying’, ‘crowning’

Table 7: The list of 262 event types from the SWiG dataset included in our experiments.

Method	M ² E ² Acc. (%)	SWiG Acc. (%)
Zero-shot-LLaVA-ChatGPT	74.3	11
Zero-shot-CLIP (B/32)	65.7	28.3
Zero-shot-CLIP-Event	70.8	31.4
Supervised-CLIP-Event	72.8	45.6

Table 8: Event Detection Accuracy (%) of generative prompting (LLaVA followed by ChatGPT), pretrained CLIP, and CLIP-Event (Li et al., 2022b) models on the M²E² and SWiG dataset.

You are a helpful AI assistant with extensive knowledge of situation recognition. Situation Recognition is the task of recognizing the activity happening in an image, the actors and objects involved in this activity, and the roles they play. Semantic roles describe how objects in the image participate in the activity described by the verb.

In this case, you will be provided with an "Image Description", and you must choose the set of all the actions portrayed in it. Please keep your answer concise and you can choose as many as actions from the 504 possible actions listed below:

[tattooing', 'raining', 'restraining', 'splashing', 'emerging', 'pasting', 'skiing', 'inflating', 'displaying', 'marching', 'feeding', 'hunting', 'snowing', 'shredding', 'chewing', 'slipping', 'teaching', 'flicking', 'shaving', 'ballooning', 'intermingling', 'flinging', 'sitting', 'yanking', 'putting', 'decorating', 'witing', 'mashing', 'baking', 'kneading', 'rearing', 'urinating', 'slouching', 'moistening', 'inserting', 'lecturing', 'bubbling', 'shushing', 'chopping', 'shrugging', 'parachuting', 'giving', 'cooking', 'waiting', 'saluting', 'falling', 'racing', 'ramming', 'pushing', 'wrapping', 'leaping', 'packing', 'foraging', 'smearing', 'mouming', 'stretching', 'ignoring', 'whisking', 'sharpener', 'jogging', 'browsing', 'sealing', 'examining', 'hiking', 'clapping', 'erupting', 'hunching', 'raking', 'punching', 'singing', 'standing', 'taxing', 'working', 'whipping', 'watering', 'steering', 'distributing', 'spying', 'uncorking', 'panhandling', 'weeping', 'squeezing', 'curlsyng', 'paving', 'spearng', 'whirling', 'dripping', 'carrying', 'pinching', 'flapping', 'tying', 'arranging', 'tearing', 'blocking', 'swooping', 'practicing', 'potting', 'installing', 'dousing', 'pressing', 'hoeing', 'making', 'unpacking', 'mowing', 'stacking', 'stinging', 'heaving', 'pulling', 'counting', 'breaking', 'dockng', 'scolding', 'destroying', 'dryng', 'poking', 'tickleng', 'bandaging', 'recuperating', 'butting', 'communicating', 'tilting', 'hoisting', 'nailing', 'lighting', 'shelving', 'wagging', 'surfing', 'gluing', 'twisting', 'hauling', 'complaining', 'crawling', 'punting', 'erasng', 'batshng', 'servng', 'telephoning', 'mimmg', 'congregating', 'smiling', 'drinking', 'scoring', 'pouting', 'yawning', 'soaking', 'glowing', 'writing', 'coaching', 'brawling', 'spoiling', 'poing', 'sucking', 'unveiling', 'flipping', 'slicing', 'educating', 'bothering', 'sweeping', 'washing', 'spinning', 'dyeing', 'filling', 'duckung', 'shaking', 'signing', 'braiding', 'stooping', 'celebrating', 'catching', 'juggung', 'preaching', 'scratching', 'shearing', 'ejecting', 'driving', 'puckering', 'peeling', 'hanging', 'bulldozing', 'slidng', 'igniting', 'turning', 'pumping', 'clenchng', 'ascendng', 'grimacing', 'fastening', 'plowing', 'repairng', 'trainng', 'recoverng', 'blossomng', 'cheering', 'jumpng', 'wading', 'leading', 'biting', 'grieving', 'sneezng', 'drooling', 'striking', 'knockng', 'encouraging', 'soaring', 'manicuring', 'squintng', 'boating', 'typng', 'lacing', 'farmng', 'photographng', 'speaking', 'minng', 'sallng', 'winkng', 'bouncing', 'leafng', 'cheerleading', 'autographng', 'chasing', 'boarding', 'nagging', 'scrapng', 'recording', 'providng', 'tipng', 'ruzzing', 'gatherng', 'reasuring', 'patting', 'talkng', 'interrogng', 'emptyng', 'decomposing', 'dancing', 'skiddng', 'flaming', 'subduing', 'hurlng', 'scopng', 'weeding', 'disembarkng', 'sewing', 'videotapng', 'collldng', 'socialzng', 'branchng', 'giggle', 'sprintng', 'massaging', 'votng', 'swarming', 'attckng', 'signllng', 'disectng', 'walkng', 'applaudng', 'grillng', 'spankng', 'hugng', 'friskng', 'gaspg', 'crashng', 'colorng', 'detainng', 'askng', 'ridng', 'clawng', 'scrubblng', 'rinsng', 'fordng', 'fuelng', 'sproutng', 'beggnng', 'frownng', 'sprayng', 'competing', 'fryng', 'imitatng', 'stitchng', 'sowng', 'spillng', 'coverng', 'rowng', 'dippng', 'lickng', 'callng', 'shellng', 'overflowng', 'wheelng', 'discussng', 'stormng', 'constructng', 'buryng', 'thowng', 'drenchnng', 'injectng', 'prickng', 'exterminatng', 'campng', 'tackling', 'clearng', 'intervewng', 'plummetng', 'wrinkln', 'wettng', 'readng', 'weighng', 'deflectng', 'openng', 'brewng', 'lapng', 'vacuumng', 'Kicking', 'phonng', 'pickng', 'starng', 'foldng', 'wrngng', 'operatng', 'bettng', 'amashng', 'packng', 'hppng', 'strokg', 'pourng', 'pedalng', 'tunng', 'dustng', 'chiseling', 'attachng', 'droppng', 'baptzng', 'instructng', 'restng', 'shopng', 'cryng', 'coughng', 'releasng', 'misbehavng', 'crestng', 'rockng', 'dialng', 'payng', 'guardng', 'queuing', 'grinnng', 'leanng', 'kissng', 'slappng', 'harvestng', 'bikng', 'pilotng', 'weldng', 'stuffng', 'tillng', 'archng', 'rubbg', 'sleepng', 'skippng', 'butterng', 'clng', 'swingng', 'prunng', 'pinng', 'dragng', 'barbecu', 'paradng', 'fetchng', 'lockng', 'shoveln', 'glarin', 'confrontng', 'crammg', 'disciplnn', 'plantng', 'runng', 'cleanng', 'spittng', 'climbng', 'peen', 'caressng', 'slitherng', 'flexng', 'studyn', 'drawng', 'handcuffng', 'assemblng', 'latherng', 'shiverng', 'mendng', 'distractng', 'circldng', 'prayng', 'kneeling', 'dining', 'loadng', 'flossng', 'exercisng', 'arrestng', 'shoutng', 'combing', 'unloadng', 'trimmng', 'unpluggng', 'moldng', 'checkng', 'descndng', 'sayng', 'aimng', 'pounce', 'waddln', 'performng', 'fishng', 'taping', 'taughng', 'immerng', 'gardenn', 'sillng', 'twirlng', 'dive', 'snifng', 'sketchng', 'grindng', 'filmg', 'rehabilitatng', 'measurng', 'eatng', 'buckllng', 'hitchhikng', 'bowng', 'tripplng', 'officiatng', 'submergng', 'helpng', 'craftng', 'drumpong', 'plungng', 'swimmng', 'cartng', 'stampndng', 'curlng', 'burnng', 'perspirng', 'spreadng', 'clipng', 'snuggng', 'carvg', 'tuggn', 'calmng', 'hitng', 'apprehendng', 'retrevng', 'camouflagn', 'placng', 'towng', 'prowlg', 'lanchng', 'crushng', 'pitchng', 'offerng', 'strappng', 'adjustng', 'landng', 'moisutnng', 'tastng', 'wipng', 'sprinkln', 'buyng', 'applyng', 'protestng', 'stumbldng', 'staplng', 'moppng', 'stirng', 'skatng', 'gnawng', 'floatng', 'shootng', 'commutng', 'microwavng', 'fixng', 'milkg', 'gamblng', 'brushng', 'vaulng', 'crouchn', 'embracng', 'whistln', 'waving', 'smellng', 'unlockng', 'extinguishng', 'building', 'rottng', 'paintng', 'admiring', 'waxng', 'stripng', 'pyng', 'crownng', 'buttonng]

If you do not find any action in the image description, then simply output "Other" rather than taking a random guess over the possible actions

Image Description: The image features a close-up of a person's hand holding a smartphone. The person is likely using the phone to interact with the device, possibly typing a message or navigating through an app. The focus of the image is on the individual's hand and the cell phone, capturing the interaction between the two.

Output: Typing, Pressing
Ground Truth: Dialing

Figure 6: Example prompt for predicting the event type using a two-stage generative framework using the SWiG dataset. We highlight the generated caption for image (Figure 7) in blue. The ChatGPT outputs are bolded in red while the ground truth action is bolded in pink.



Figure 7: An example image depicting 'dialing' action from the SWiG dataset.

Input: Image \mathcal{I} , Object o , Event E , Event argument role labels (and definitions) \mathcal{A}

Prompt:

Image 1: $256 \times < \mathcal{I}$ visual feature tokens>

Image 2: $256 \times < o$ visual feature tokens>

Describe the role of the entity in "Image 2" in the context of the E event in "Image 1". The possible argument roles of the objects performing E event include \mathcal{A} . Please be concise with your answer.

Figure 8: Multimodal input template used to prompt the GLVM in the GenEARL framework. Here, the model is provided with the image, event and possible argument role labels (and definitions). We get 256 visual tokens for the image or the object by projecting the raw input into the vision embedding space using the visual processing module of the GVLM.

Input: Object o , Event E , Event argument roles (and definitions) \mathcal{A}

Prompt:

Image: $256 \times < o$ visual feature tokens>

Describe the role of the entity in "Image" in the context of the E event. The possible argument roles of the objects performing E event include \mathcal{A} . Please be concise with your answer.

Figure 9: Multimodal input template used to prompt the GLVM in the GenEARL framework. Here, the model is provided with the event and possible argument role labels (and definitions). We get 256 visual tokens for the object by projecting the raw input into the vision embedding space using the visual processing module of the GVLM.

Input: Object o

Prompt:

Image: $256 \times < o$ visual feature tokens>

Describe the "Image" concisely.

Figure 10: Multimodal input template used to prompt the GLVM in the GenEARL framework. Here, the model is provided just the event participant object. We get 256 visual tokens for the object by projecting the raw input into the vision embedding space using the visual processing module of the GVLM.

Input: Image Caption \mathcal{I} , Generated object descriptions $\{g_1, g_2, \dots, g_k\}$ from LLaVA for the k participants objects in the image , Event E , Event argument role labels (and definitions) \mathcal{A}

Setting: Zero-shot

Prompt:

You are a helpful AI assistant with extensive knowledge of event argument extraction. Worldwide events are documented in raw text on various online platforms, and it is crucial to extract useful and concise information about them for downstream applications.

In this case, you will be provided with an “Event Image Description”, the “Event” portrayed in the image, a generic “Event Argument Roles Definition” that helps you to understand the argument roles that are grounded in different objects in the image, and the “Object Role” descriptions to describe the role of specific objects in the context of the "Event Image". Based on the provided information, you need to tell the argument roles associated with different objects in the image.

Remember that: 1) The number of possible event argument roles can sometimes be equal to, more, or less than the number of objects detected for the "Event Image".

2) Multiple objects may get identical event argument roles, but not always.

3) It is completely possible that some of the event argument roles are not grounded in any of the objects detected for the "Event Image".

Please keep your answer concise. You can choose to assign a "Other" argument role if you are not sure about the argument role for a particular object.

Event Image Description: \mathcal{I}

Event: E

Event Argument Role Definition: \mathcal{A}

Role of Object 1: g_1

.

.

Role of Object k: g_k

Argument Role of Object 1:

.

.

Argument Role of Object k:

Output:

Argument Role of Object 1: \hat{a}_1

.

.

Argument Role of Object k: \hat{a}_k

Figure 11: LLM template used to prompt the ChatGPT in the GenEARL framework. This template is utilized under the zero-shot settings. Due to cost considerations associated with prompting ChatGPT, we perform multimodal EARL in a batch of all the event participant objects in an image.

Solved Input: Image Caption \mathcal{I}_s , Generated object descriptions $\{g_{s,1}, g_{s,2}, \dots, g_{s,l}\}$ from LLaVA for the l participants objects in the image , Event E_s , Event argument role labels (and definitions) \mathcal{A}_s

Query Input: Image Caption \mathcal{I} , Generated object descriptions $\{g_1, g_2, \dots, g_k\}$ from LLaVA for the k participants objects in the image , Event E , Event argument role labels (and definitions) \mathcal{A}

Setting: One-shot

Prompt:

You are a helpful AI assistant with extensive knowledge of event argument extraction. Worldwide events are documented in raw text on various online platforms, and it is crucial to extract useful and concise information about them for downstream applications.

In this case, you will be provided with an “Event Image Description”, the “Event” portrayed in the image, a generic “Event Argument Roles Definition” that helps you to understand the argument roles that are grounded in different objects in the image, and the “Object Role” descriptions to describe the role of specific objects in the context of the "Event Image". Based on the provided information, you need to tell the argument roles associated with different objects in the image.

Remember that: 1) The number of possible event argument roles can sometimes be equal to, more, or less than the number of objects detected for the "Event Image".

2) Multiple objects may get identical event argument roles, but not always.

3) It is completely possible that some of the event argument roles are not grounded in any of the objects detected for the "Event Image".

Please keep your answer concise. You can choose to assign a "Other" argument role if you are not sure about the argument role for a particular object.

We will first show a single solved instance of the task, and then you will complete the task on a new query.

Solved Instance:

Event Image Description: \mathcal{I}_s

Event: E_s

Event Argument Role Description: \mathcal{A}_s

Role of Object 1: $g_{s,1}$

.

Role of Object l: $g_{s,l}$

Argument Role of Object 1: $a_{s,1}$

.

Argument Role of Object l: $a_{s,l}$

Query Instance:

Event Image Description: \mathcal{I}

Event: E

Event Argument Role Definition: \mathcal{A}

Role of Object 1: g_1

.

Role of Object k: g_k

Argument Role of Object 1:

.

Argument Role of Object k:

Output:

Argument Role of Object 1: \hat{a}_1

.

Argument Role of Object k: \hat{a}_k

Figure 12: LLM template used to prompt the ChatGPT in the GenEARL framework. This template is utilized under the one-shot settings.

Input: Image Caption \mathcal{I} , Generated object description g from LLaVA for the one of the participant objects in the image , Event E , Event argument role labels (and definitions) \mathcal{A}

Setting: Zero-shot

Prompt:

Given an “Event Image Description”, the “Event” portrayed in the image, a generic “Event Argument Roles Definition” that helps you to understand the argument roles that are grounded in different objects in the image, and the “Object Role” description to describe the role of the object in the context of the “Event Image”. Based on the provided information, you need to tell the argument roles associated with different objects in the image. You can choose to assign an “Other” argument role if you are not sure about the argument role for a particular object.

Event Image Description: \mathcal{I}

Event: E

Event Argument Role Definition: \mathcal{A}

Role of Object: g

Argument Role of Object:

Figure 13: LLM template used to prompt the Alpaca-7B. This template is utilized under the zero-shot settings. Due to the context length limitations, we label one participating object at a time.

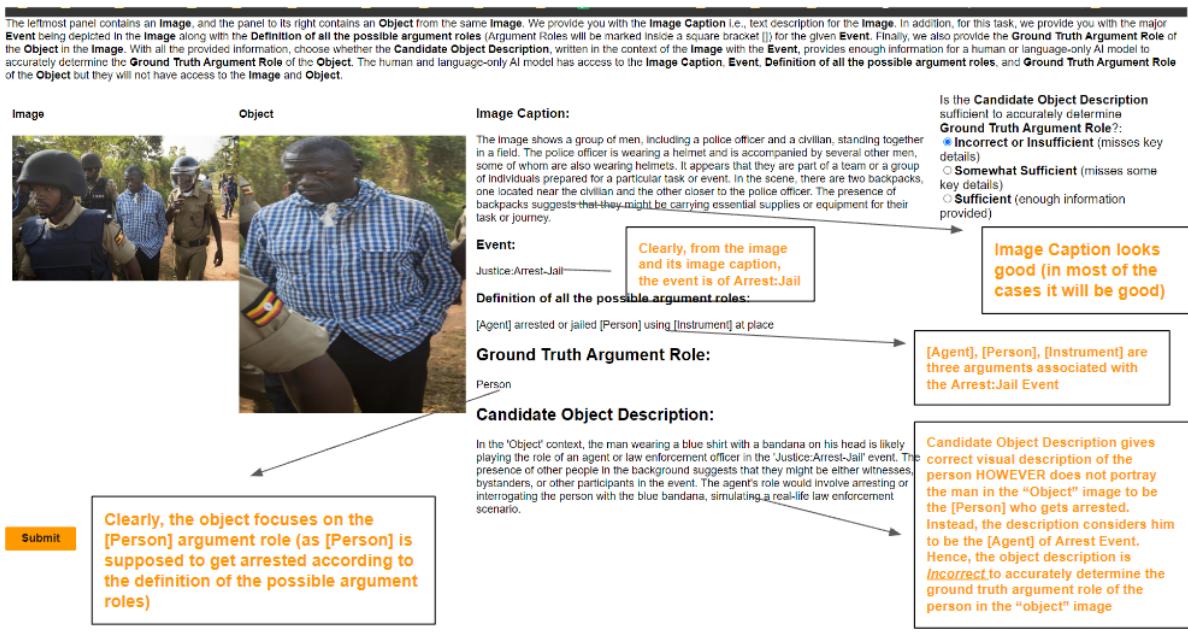


Figure 14: Screenshot of a solved example shown to the human labelers before they attempt the qualification test.

(WARNING: This HIT may contain adult content. Worker discretion is advised.) The leftmost panel contains an **Image**, and the panel to its right contains an **Object** from the same **Image**. We provide you with the **Image Caption** i.e., text description for the **Image**. In addition, for this task, we provide you with the major **Event** being depicted in the **Image** along with the **Definition of all the possible argument roles** (Argument Roles will be marked inside a square bracket []) for the given **Event**. Finally, we also provide the **Ground Truth Argument Role of the Object** in the **Image**. With all the provided information, choose whether the **Candidate Object Description**, written in the context of the **Image** with the **Event**, provides enough information for a human or language-only AI model to accurately determine the **Ground Truth Argument Role of the Object**. The human and language-only AI model has access to the **Image Caption**, **Event**, **Definition of all the possible argument roles**, and **Ground Truth Argument Role of the Object** but they will not have access to the **Image** and **Object**.

Image	Object	Image Caption:	Event:	Definition of all the possible argument roles:	Is the Candidate Object Description sufficient to accurately determine Ground Truth Argument Role?
		The image shows a car that has been involved in a severe accident, with the car's passenger side heavily damaged. The car is sitting off the side of the road, and the front end has been crushed in the collision. There are several people present at the scene, likely responding to or assessing the damage. One person can be seen near the middle of the scene, while two others are standing closer to the right side of the car. Additionally, there is a fire hydrant visible near the middle of the scene. The overall scene conveys the severity of the accident and the need for assistance or emergency services.	Conflict Attack	[Attacker] attacked or assaulted [Target] using Target [Instrument] at place	<input checked="" type="radio"/> Incorrect or Insufficient (misses key details) <input type="radio"/> Somewhat Sufficient (misses some key details) <input type="radio"/> Sufficient (enough information provided)
			Ground Truth Argument Role:	Target	
			Candidate Object Description:	In the context of the Conflict Attack event depicted in the image, the shattered car window and the broken car door are the direct results of an attack. The smashed window and door glass indicate that the car was intentionally damaged, possibly by a vandal or an aggressor. The shattered window and the debris scattered around the car can be seen as evidence of the conflict that occurred. The role of the entity in the 'Object' could potentially be a witness or a bystander observing the aftermath of the attack, or it could be the person responsible for the damage if they were involved in the conflict.	

Submit

Figure 15: Screenshot of an unsolved example shown to the human labelers during the actual annotation.