Benchmarking and Improving LVLMs on Event Extraction from Multimedia Documents

Fuyu Xing*, **Zimu Wang***, **Wei Wang, Haiyang Zhang**†
School of Advanced Technology, Xi'an Jiaotong-Liverpool University
{Fuyu.Xing21,Zimu.Wang19}@student.xjtlu.edu.cn
{Wei.Wang03,Haiyang.Zhang}@xjtlu.edu.cn

Abstract

The proliferation of multimedia content necessitates the development of effective Multimedia Event Extraction (M²E²) systems. Though Large Vision-Language Models (LVLMs) have shown strong cross-modal capabilities, their utility in the M²E² task remains underexplored. In this paper, we present the first systematic evaluation of representative LVLMs, including DeepSeek-VL2 and the Qwen-VL series, on the M²E² dataset. Our evaluations cover text-only, image-only, and cross-media subtasks, assessed under both few-shot prompting and fine-tuning settings. Our key findings highlight the following valuable insights: (1) Few-shot LVLMs perform notably better on visual tasks but struggle significantly with textual tasks; (2) Fine-tuning LVLMs with LoRA substantially enhances model performance; and (3) LVLMs exhibit strong synergy when combining modalities, achieving superior performance in cross-modal settings. We further provide a detailed error analysis to reveal persistent challenges in areas such as semantic precision, localization, and cross-modal grounding, which remain critical obstacles for advancing M²E² capabilities.

1 Introduction

Event extraction (EE) is a fundamental task in NLP, aiming to identify structured events from unstructured data (Ahn, 2006; Peng et al., 2023b). It typically involves two subtasks: event detection (ED), which detects the event trigger words and classifies their event type, and event argument extraction (EAE), which extracts entities and labels their argument roles. Traditional EE methods have primarily focused on a single modality, either text (Wang et al., 2022, 2024; Zhang and Ji, 2021), image (Yatskar et al., 2016; Cho et al., 2022; Pratt et al., 2020), or video (Soomro et al., 2012; Ye et al., 2015; Caba Heilbron et al., 2015).

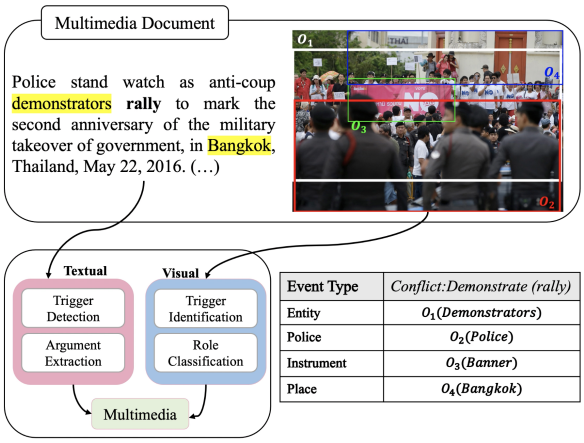

Figure 1: Illustration of the M²E² task, including an image and accompanying text, where event arguments may appear in either or both modalities.

In the era of digital media, content like news articles often contains both textual and visual modalities, and some arguments presented in images may not appear in the accompanying text. This has prompted the rise of Multimedia Event Extraction (M²E², Li et al., 2020), targeting to extract information from combined texts and images (see Figure 1). Prior works either rely on limited hand-crafted features or require expensive annotation efforts. Meanwhile, Large Vision-Language Models (LVLMs) have demonstrated strong cross-modal reasoning abilities (Gan et al., 2025; Zeng et al., 2025), but their utility for M²E² remains underexplored.

Motivated by this gap, we present the first systematic evaluation of LVLMs on the M²E² dataset. Evaluations are conducted across six subtasks for ED and EAE in text-only, image-only, and cross-modal settings, under both few-shot and fine-tuning paradigms. Our key findings reveal that: (1) Few-shot LVLMs demonstrate strong performance compared to specialized baselines in visual-only EE but lag behind established pre-trained language models (PLM) baselines textually. (2) Fine-tuning LVLMs dramatically improves their capabilities,

*Equal contribution.
†Corresponding author.

Proceedings of the 18th International Natural Language Generation Conference, pages 734–742
October 29–November 2, 2025.

enabling smaller models to surpass strong textual baselines and achieve state-of-the-art results. (3) Utilizing LVLMs on multimedia EE consistently yields the best results, showcasing the effective cross-modal synergy of LVLMs. (4) Despite performance improvements, persistent challenges exist across all modalities, including difficulties with semantic discrimination, localization, effective cross-modal grounding, and safety filter behaviors.

2 Related Work

Event Extraction has mainly been investigated in single modalities. Textual EE has progressed from token-level classification (Alan et al., 2020; Yang et al., 2024) to generative models (Peng et al., 2023a; Qi et al., 2024). On the other hand, visual EE has traditionally focused on coarse ED from images or videos (Gu et al., 2018; Wu et al., 2019), relying on predefined event types or bounding boxes (Yatskar et al., 2016; Silberer and Pinkal, 2018). Recent methods remove such constraints by integrating object detection and attention-based reasoning (Li et al., 2020). Emerging studies have also highlighted the improvements by integrating external visual knowledge or multimedia parameters on textual EE, showcasing the complementary strengths of different modalities (Ji and Grishman, 2008; Zhang et al., 2017; Zhang and Ji, 2018).

Multimedia Event Extraction combines textual and visual modalities to improve robustness, employing fusion strategies, such as feature-level concatenation (Horng et al., 2017; Chen et al., 2018), shared latent spaces (Khadanga et al., 2019), and decision-level aggregation. Recent frameworks such as WASE (Li et al., 2020), CLIP-Event (Li et al., 2022), and CAMEL (Du et al., 2023) incorporate weak supervision or inject structured event schemas to improve multimodal grounding. Instruction-tuned models like UMIE (Sun et al., 2024) and video-text paired learning (Cao et al., 2025) enhance event representation across modalities. However, despite the success of generative models in event understanding and reasoning tasks (Li et al., 2024, 2025), their potential in the critical M²E² remains underexplored. We pioneer this field by utilizing LVLMs to M²E², offering valuable analysis and insights for future research.

3 Problem Formulation

We focus on the textual, visual, and multimedia EE defined by Li et al. (2020). Each *Event* includes a trigger and one or more arguments. Formally: (1) *Event Trigger* (τ) is a verb or noun in the text indicating the occurrence of the event. (2) *Event Argument* (α) plays a role in the event, either a text entity, a temporal expression, a numerical value, or a detected object in the image. (3) *Argument Role* (ρ) is the semantic relationship linking an argument α to its event mention. Our two main tasks over a multimodal document are as follows:

Event Detection seeks to detect all event mentions $\{EM_1, \ldots, EM_n\}$ within the input context. Each EM_i is assigned a category $c_i \in C$ and is grounded in one or both modalities:

$$EM_i = (c_i, G_i),\ G_i \subseteq \{\tau_i, \upsilon_i\},$$

where τ_i is the trigger and υ_i is the corresponding image. Based on the content of the grounding set G_i, an event mention can be classified into one of the following types: **text-only** event mention where $G_i = \{\tau_i\}$, **image-only** event mention where $G_i = \{\upsilon_i\}$ and **multimodal** event mention where $G_i = \{\tau_i, \upsilon_i\}$, indicating that the same event is referred to both textually and visually.

Event Argument Extraction extracts a set of arguments $\{\alpha_{i1}, \ldots, \alpha_{ij}\}$ for each identified EM_i. Each argument α_{ij} is represented as:

$$\alpha_{ij} = (\rho_{ij}, H_{ij}),\ H_{ij} \subseteq \{e_{ij}, o_{ij}\},$$

where $\rho_{ij} \in R$ is the argument role and $H_{ij} \subseteq \{e_{ij}, o_{ij}\}$ indicates grounding in textual entity e_{ij} or visual object o_{ij} or both. Arguments referring to the same real-world object in text and image are merged into a single unified representation.

4 Evaluation Settings

4.1 Datasets and Evaluation Metrics

We conduct our experiments on M²E² (Li et al., 2020), a cross-media EE dataset of news articles, containing 1,105 text-only mentions, 188 image-only mentions, and 395 cross-media mentions. We utilize 70% for training and 30% for evaluation. All events are categorized into 8 fine-grained subtypes (e.g., `Movement.Transport`) with 18 argument roles. For auxiliary data, we leverage ACE 2005 (Walker et al., 2006), filtering it to match the eight event types in M²E². A conceptual overview of our evaluation pipeline is presented in Figure 2.

For all subtasks, we report the precision, recall, and F1-score under the following conditions: (1)

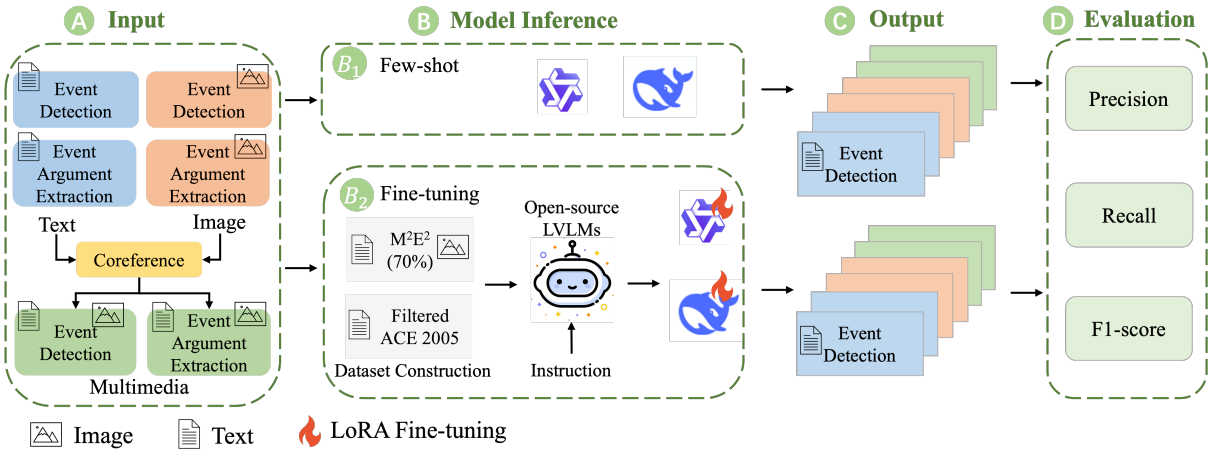

Figure 2: Evaluation pipeline under few-shot and fine-tuning paradigms. A multimedia document is processed, and the extracted events and arguments are compared against ground-truth annotations with strict matching.

Text-only ED, if both the trigger word and its event type are matched; (2) **Image-only ED**, if the image and event type are correct; (3) **Text-only EAE**, if both the argument and its role are matched; (4) **Image-only EAE**, if the argument and role match and the Intersection over Union (IoU) between the predicted and ground truth bounding box exceeds 0.5; (5) **Cross-media ED**, if the trigger (or image) and event type match exactly; and (6) **Cross-media EAE**, if the argument (or bounding box) and role are matched.

4.2 Prompt Design

Table 2 shows an example of the few-shot prompt used to guide the LVLMs. Each prompt consists of the following key components: a task-specific **Instruction**, four contrastive **Demonstrations** (both positive and negative) to enhance task understanding, the **Query** input (text and/or image), a set of candidate **Options** (e.g., event or role types), and a final **Answer** placeholder for the model's output. For the EAE subtask, the ground-truth **Event Type** is also supplied as context. In fine-tuning experiments, the structure of the prompt is similar, while the demonstrations are removed.

4.3 Experimental Setup

We conduct experiments across a range of open-source LVLMs: DeepSeek-VL2 (Wu et al., 2024), Qwen2-VL-7B (Bai et al., 2025), Qwen2.5-VL-7B, and Qwen2.5-VL-72B (Bai et al., 2025). Our evaluation includes a *few-shot* paradigm to assess out-of-the-box capabilities and a *fine-tuning* paradigm to measure performance after adaptation. During the fine-tuning process, we adapt Qwen2-VL-7B and Qwen2.5-VL-7B using LoRA (Hu et al., 2022). We train these models on a combined dataset created by merging the M²E² training split with filtered ACE 2005, separately for each sub-task. The fine-tuning is carried out with a learning rate of 1×10^{-4} and batch size of 3. We keep the inference temperature at 0 to ensure output stability. All experiments are conducted on 2 NVIDIA GeForce RTX 4090 graphics cards.

4.4 Baselines

We compare the performance of LVLMs against the following baselines across all modalities. These include unimodal methods such as the graph-based **JMEE** (Liu et al., 2018) for text and **Clip-Event** (Li et al., 2022) for vision. For multimedia models, we compare against diverse strategies, including those based on shared semantic embeddings (**WASE**, Li et al., 2020, **UniCL**, Liu et al., 2022), data synthesis (**CAMEL**, Du et al., 2023), instruction-tuning (**UMIE**, Sun et al., 2024), and the current state-of-the-art multi-task framework, **X-MTL** (Cao et al., 2025).

4.5 Experimental Results

Table 3 presents the evaluation results of LVLMs in comparison to baseline models, where X-TML, the state-of-the-art multi-task framework, is included. The full, unabridged table featuring all baseline models is provided in Appendix B. From the table, we have the following observations:

First, a notable performance disparity exists in the textual domain, where few-shot LVLMs fall significantly short compared to specialized models such as X-MTL. However, fine-tuning LVLMs effectively bridges this gap. For instance, Qwen2-VL-7B with LoRA yields a substantial 36.6% in-

Task	Metric	SOTA	Few-shot LVLMs				Fine-tuned (LoRA)	
		X-MTL	DS-VL	Qwen2-VL	Qwen2.5-VL	Qwen2.5-VL-72B	Qwen2-VL	Qwen2.5-VL
Textual Event Extraction								
Event Mention	P	49.7	30.5	13.3	70.0	80.0	46.8	41.5
	R	65.7	12.0	19.7	3.7	12.0	59.6	52.6
	F1	**56.6**	17.5	15.9	6.9	20.6	52.5	46.4
Argument Role	P	34.6	5.6	75.0	4.2	10.1	16.3	35.5
	R	37.6	2.2	2.6	1.4	7.5	24.1	50.3
	F1	36.0	3.1	5.1	2.1	8.6	19.5	**41.6**
Visual Event Extraction								
Event Mention	P	73.1	39.3	69.6	54.3	73.0	49.7	48.5
	R	70.3	76.4	62.1	47.3	70.6	74.4	72.5
	F1	71.7	51.9	65.6	50.5	**71.8**	59.6	58.1
Argument Role	P	33.2	0.7	3.3	23.7	62.8	23.5	10.8
	R	31.3	0.7	4.3	35.3	71.9	4.3	19.6
	F1	32.2	0.7	3.8	28.3	**67.0**	7.3	14.0
Multimedia Event Extraction								
Event Mention	P	78.3	41.3	65.4	62.9	84.0	41.7	56.6
	R	57.3	79.6	53.7	37.2	81.4	79.3	77.4
	F1	66.2	54.4	60.0	46.8	**82.5**	54.7	65.4
Argument Role	P	40.3	8.1	20.6	23.9	56.1	55.3	54.7
	R	42.6	5.8	21.9	14.9	71.9	70.8	80.1
	F1	41.4	6.8	21.2	18.4	63.0	62.1	**65.0**

Table 1: Main results on the M²E² benchmark. **Bold** indicates the best F1-score in each modality for each task. (DS-VL: DeepSeek-VL2, Qwen2-VL: Qwen2-VL-7B, Qwen2.5-VL: Qwen2.5-VL-7B)

crease in F1-score for ED and a 14.4% increase in EAE, elevating its performance to be on par with strong textual baselines. Notably, the fine-tuned Qwen2.5-VL-7B surpassed these baselines in EAE, highlighting that even smaller, fine-tuned LVLMs are capable of achieving state-of-the-art results on complex textual tasks.

Second, in the visual domain, the performance trend is reversed. Few-shot LVLMs, particularly Qwen2.5-VL-72B, exhibit exceptional out-of-the-box capabilities, surpassing all specialized visual baselines in both ED (achieving an F1-score of 71.8%) and EAE (with an F1-score of 67.0%). This indicates that their pre-training effectively captures the knowledge required for visual event understanding, though argument localization remains open challenges for models that have not been explicitly pre-trained on real-world coordinates.

Finally, LVLMs consistently achieved the highest overall performance in the multimedia setting, underscoring the complementary relationship between textual and visual modalities. Notably, even models that struggled with text-only ED show significant improvements when paired with the corresponding images, highlighting the critical role the visual context plays in enriching and disambiguating textual information while grounding events. Furthermore, the fine-tuned Qwen2.5-VL-7B further boosted performance, reaching an F1-score of 65.4% and 65.0% on ED and EAE, respectively. These results demonstrate that lightweight adaptation of a model with only 7B parameters can still match or surpass the few-shot performance of a 72B model, reinforcing the effectiveness of parameter-efficient fine-tuning approaches.

4.6 Error Analysis

Despite competitive quantitative results, a detailed analysis uncovers consistent shortcomings within each modality are categorized as follows:

Textual EE. In the textual setting, model errors primarily arise from insufficient *fine-grained understanding*. where models frequently confuse semantically similar event types (e.g., `Contact:Meet` and `Contact:Phone-Write`) and struggle with precise argument boundaries, leading to overextension of spans (e.g., predicting "*Iraqi government forces*" instead of "*forces*"). They also generate spurious arguments by misinterpreting context (e.g., labeling the country "*Uganda*" as a place argument for a specific attack) or failure to resolve coreference resolution (e.g., identifying pronoun "*he*" as an argument). Conversely, critical participants are often

Figure 3: Visual event detection error:
Ground Truth: `Conflict.Attack`,
Qwen2.5-VL-72B: `Movement.Transport`,
DeepSeek-VL2: `Conflict.Demonstrate`.

overlooked, such as omitting "*teenager*" or "*victim*" in event descriptions, indicating a potential over-reliance on explicit lexical cues.

Visual EE. In the visual setting, models closely resemble those observed in text. However, they still face difficulties in fine-grained event type classification, as exemplified in Figure 3, where an image of a vehicle incident labeled with `Conflict.Attack` may be misinterpreted by different models. Qwen2.5-VL-72B classifies it as `Movement.Transport`, while DeepSeek-VL2 predicts `Conflict.Demonstrate`. Moreover, models often identify visually prominent objects that are not actual event arguments (Figure 4), resulting in low precision under strict evaluation criteria. Except for the Qwen2.5-VL models benefit from real-world coordinate-aware pretraining (Bai et al., 2025), most variants fail to localize objects exactly.

Multimedia EE. Finally, in the multimedia setting, cross-modal grounding remains a challenge. Notably, errors arising from unimodal processing are not always resolved. For instance, fine-grained semantic confusion between event types or imprecise bounding box localization persist even when complementary information from the other modality is accessible. Furthermore, the internal safety filters of all LVLMs often hinder processing, refusing to process content related to sensitive topics like violence, which directly impacts recall for event types such as `Conflict:Attack` and `Life:Die`.

5 Conclusion

We present the first comprehensive evaluation of LVLMs for M²E², assessing their performance in both few-shot and fine-tuning settings. Our results reveal a critical interplay between modality and training strategy: Few-shot LVLMs excel on visual tasks but falter in the texts, a gap that is dramatically closed by fine-tuning. Ultimately, our work confirms that LVLMs excel in combining text and vision modalities, creating powerful cross-modal synergy that yields the best performance, highlighting the significant potential of adapted LVLMs for complex multimedia understanding. We further highlight the persistent challenges in our error analysis, including semantic precision, localization, and cross-modal grounding, which we will design more robust architectures to resolve in future research.

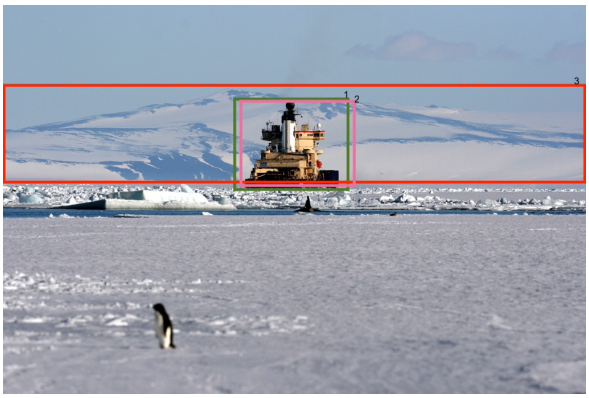

Figure 4: Visual event argument extraction error:
Ground Truth: **Green** Box,
Correct Prediction: **Pink** Box,
Additional Predicted: **Red** Box.

Limitations

We acknowledge the following limitations of this study that provide avenues for future work. First, due to the only dataset available, we only experimented in the news domain using the M²E² and ACE 2005 datasets. Second, while we evaluated several representative open-source LVLMs, we did not include closed-source models due to the high cost required, but it does not diminish the contribution and findings of this study.

Acknowledgements

We sincerely thank anonymous reviewers for their valuable comments. This research is supported by the Research Development Funding (RDF) (RDF-21-02-044) and Collaborative Research Project (RDS10120240248) at Xi'an Jiaotong-Liverpool University.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Rea-*

soning about Time and Events, pages 1–8.

Ramponi Alan, Rosario Lombardo, Plank Barbara, and 1 others. 2020. Biomedical event extraction as sequence labeling. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, pages 5357–5367.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Jianwei Cao, Yanli Hu, Zhen Tan, and Xiang Zhao. 2025. Cross-modal multi-task learning for multimedia event extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11454–11462.

David Chen, Naveed Afzal, Sunghwan Sohn, Elizabeth B Habermann, James M Naessens, David W Larson, and Hongfang Liu. 2018. Postoperative bleeding risk prediction for patients undergoing colorectal surgery. *Surgery*, 164(6):1209–1216.

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*, pages 19659–19668.

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*, pages 5504–5513.

Hong-Seng Gan, Muhammad Hanif Ramlee, Zimu Wang, and Akinobu Shimizu. 2025. A review on medical image segmentation: Datasets, technical models, challenges and solutions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(1):e1574.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, and 1 others. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056.

Steven Horng, David A Sontag, Yoni Halpern, Yacine Jernite, Nathan I Shapiro, and Larry A Nathanson. 2017. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PloS one*, 12(4):e0174708.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: Hlt*, pages 254–262.

Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. 2019. Using clinical notes with time series data for icu management. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6432–6437.

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*, pages 16420–16429.

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

Ruochen Li, Zimu Wang, and Xinya Du. 2025. Efficient document-level event relation extraction. In *Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025)*, pages 92–99.

Ruosen Li, Zimu Wang, Son Tran, Lei Xia, and Xinya Du. 2024. Meqa: A benchmark for multi-hop event-centric question answering with explanations. *Advances in Neural Information Processing Systems*, 37:126835–126862.

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*, pages 1945–1953.

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*, pages 1247–1256.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and 1 others. 2023a. When does in-context learning fall short and why? a study on specification-heavy tasks. *arXiv preprint arXiv:2311.08993*.

Hao Peng, Xiaozhi Wang, Feng Yao, Zimu Wang, Chuzhao Zhu, Kaisheng Zeng, Lei Hou, and Juanzi Li. 2023b. Omnievent: A comprehensive, fair, and easy-to-use toolkit for event understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 508–517.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *European Conference on Computer Vision*, pages 314–332. Springer.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2024. Adelie: Aligning large language models on information extraction. *arXiv preprint arXiv:2405.05008*.

Carina Silberer and Manfred Pinkal. 2018. Grounding semantic roles in images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2616–2626.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Lin Sun, Kai Zhang, Qingyuan Li, and Renze Lou. 2024. Umie: Unified multimodal information extraction with instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19062–19070.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *(No Title)*.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, and 1 others. 2022. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941.

Zimu Wang, Lei Xia, Wei Xjtlu, and Xinya Du. 2024. Document-level causal relation extraction with knowledge-guided binary question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16944–16955.

Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. 2019. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 284–293.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Jinshun Yang, Shuangxi Huang, and Mingfeng Huang. 2024. Mlee: Event extraction as multi-label classification task at token level. In *6th International Conference on Intelligence Science (ICIS)*, pages 55–65. Springer Nature Switzerland.

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.

Guangnan Ye, Yitong Li, Hongliang Xu, Dong Liu, and Shih-Fu Chang. 2015. Eventnet: A large scale structured concept library for complex event detection in video. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 471–480.

Xinyi Zeng, Zimu Wang, Haiyang Zhang, Yiming Luo, and Wei Wang. 2025. Skin disease classification with lvlms: An empirical study. In *2025 28th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 2304–2309. IEEE.

Tongtao Zhang and Heng Ji. 2018. Event extraction with generative adversarial imitation learning. *arXiv preprint arXiv:1804.07881*.

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

Zixuan Zhang and Heng Ji. 2021. Abstract meaning representation guided graph encoding and decoding for joint information extraction. In *Proc. The 2021 Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies (NAACL-HLT2021)*.

A Prompt

Component	Example for Text-only Event Detection (ED)
Instruction	`Please identify the trigger word and classify the events in the query into one or more appropriate categories based on the given choices. Then, output the trigger word, its position interval, and the option.`

Demonstrations	`– Example 1 (Positive) –`
	`Sentence: Two Chicago police officers take a man into custody during a protest march.`
	`Answer: custody; [8,9]; (D)`
	<*Other demonstrations (positive and negative) are provided in the actual prompt.*>

Query	`{sentence}`
Options	`(A) Movement.Transport (B) Conflict.Attack (C) Conflict.Demonstrate (D) Justice.ArrestJail (E) Contact.PhoneWrite (F) Contact.Meet (G) Life.Die (H) Transaction.TransferMoney`
Answer	<*placeholder*>

Table 2: Example of the few-shot prompt structure for the textual event detection (ED) subtask. The prompt includes a detailed instruction, contrastive demonstrations, and the final query for the model to complete. The structure is adapted for other subtasks, such as argument extraction.

B Full Experimental Results

	Model	Event Mention			Argument Role		
		P	R	F1	P	R	F1
Textual	JMEE	42.5	58.2	48.7	22.9	28.3	25.3
	GAIL	43.4	53.5	47.9	23.6	29.2	26.1
	WASE-T	42.3	58.4	48.2	21.4	30.1	24.9
	WASE$_{att}$	37.6	66.8	48.1	27.5	33.2	30.1
	WASE$_{obj}$	42.8	61.9	50.6	23.5	30.3	26.4
	UniCL	49.1	59.2	53.7	27.8	34.3	30.7
	CAMEL	45.1	71.8	55.4	24.8	41.8	31.1
	X-MTL	49.7	65.7	**56.6**	34.6	37.6	36.0
	Qwen2-VL-7B	13.3	19.7	15.9	75.0	2.6	5.1
	Qwen2.5-VL-7B	70.0	3.7	6.9	4.2	1.4	2.1
	DeepSeek-VL2-4.5B	30.5	12.0	17.5	5.6	2.2	3.1
	Qwen2.5-VL-72B	80.0	12.0	20.6	10.1	7.5	8.6
	Qwen2-VL-7B-LoRA	46.8	59.6	52.5	16.3	24.1	19.5
	Qwen2.5-VL-7B-LoRA	41.5	52.6	46.4	35.5	50.3	**41.6**
Visual	CLIP-Event	41.3	72.8	52.7	21.1	13.1	17.1
	WASE-V_{att}	29.7	61.9	40.1	9.1	10.2	9.6
	WASE-V_{obj}	28.6	59.2	38.7	13.3	9.8	11.2
	WASE$_{att}$	32.3	63.4	42.8	9.7	11.1	10.3
	WASE$_{obj}$	43.1	59.2	49.9	14.5	10.1	11.9
	UniCL	54.6	60.9	57.6	16.9	13.8	15.2
	CAMEL	52.1	66.8	58.5	21.4	28.4	24.4
	X-MTL	73.1	70.3	71.7	33.2	31.3	32.2
	Qwen2-VL-7B	69.6	62.1	65.6	3.3	4.3	3.8
	Qwen2.5-VL-7B	54.3	47.3	50.5	23.7	35.3	28.3
	DeepSeek-VL2-4.5B	39.3	76.4	51.9	0.7	0.7	0.7
	Qwen2.5-VL-72B	73.0	70.6	**71.8**	62.8	71.9	**67.0**
	Qwen2-VL-7B-LoRA	49.7	74.4	59.6	23.5	4.3	7.3
	Qwen2.5-VL-7B-LoRA	48.5	72.5	58.1	10.8	19.6	14.0
Multimedia	WASE$_{att}$	38.2	67.1	49.1	18.6	21.6	19.9
	WASE$_{obj}$	43.0	62.1	50.8	19.5	18.9	19.2
	UniCL	44.1	67.7	53.4	24.3	22.6	23.4
	CAMEL	55.6	59.5	57.5	31.4	35.1	33.2
	X-MTL	78.3	57.3	66.2	40.3	42.6	41.4
	Qwen2-VL-7B	65.4	53.7	60.0	20.6	21.9	21.2
	Qwen2.5-VL-7B	62.9	37.2	46.8	23.9	14.9	18.4
	DeepSeek-VL2-4.5B	41.3	79.6	54.4	8.1	5.8	6.8
	Qwen2.5-VL-72B	84.0	81.4	**82.5**	56.1	71.9	63.0
	Qwen2-VL-7B-LoRA	41.7	79.3	54.7	55.3	70.8	62.1
	Qwen2.5-VL-7B-LoRA	56.6	77.4	65.4	54.7	80.1	**65.0**

Table 3: Full results on the M²E² dataset, including all baseline models for comparison, which is the unabridged version of the condensed table presented in the main paper. **Bold** indicates the best result in each section.