LLaVA-MS-PIT: Multi-Modal Schema-Guided Progressive Instruction Tuning for Multi-Modal Event Extraction

Hui Zhang[1, 2, 3], **Po Hu**[1, 2, 3*], **Wei Emma Zhang**[4*]
[1]Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, China
[2]School of Computer Science, Central China Normal University, China
[3]National Language Resources Monitoring and Research Center for Network Media, Central China Normal University, China
[4]School of Computer and Mathematical Sciences, The University of Adelaide, Australia
feifei_cs_phd@mails.ccnu.edu.cn, phu@mail.ccnu.edu.cn, wei.e.zhang@adelaide.edu.au

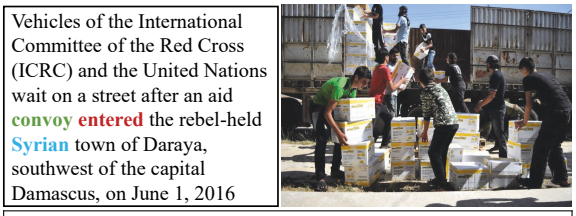
Abstract

The proliferation of multi-modal data on the internet has intensified the need for structured event understanding across textual and visual modalities. However, existing multi-modal event extraction models suffer from three major limitations: the absence of explicit event schema guidance, coarse-grained multi-modal alignment strategies, and reliance on heterogeneous, misaligned multi-modal training datasets. To address these issues, we propose LLaVA-MS-PIT, a Multi-modal Schema-Guided Progressive Instruction Tuning Framework that explicitly injects structured multi-modal event schema knowledge into the model before event extraction. Specifically, we introduce the textual event schema to establish the model's prior knowledge of event concepts and enhance its ability to reason about event structures, while the visual event schema is employed to bridge the representation gap between textual and visual modalities at the event level, enabling unified and semantically aligned event representations across modalities. Moreover, to alleviate data scarcity and modality misalignment inherent in current benchmarks, we construct imSitu-MEE, a high-quality multi-modal parallel dataset generated and annotated through schema-guided procedures. Extensive experiments demonstrate that LLaVA-MS-PIT achieves competitive performance on multi-modal event extraction benchmarks, underscoring the effectiveness and necessity of schema-guided progressive instruction tuning.

Code — https://github.com/zhanghuiecho/LLaVA-MS-PIT

Introduction

Event extraction, a fundamental task in information extraction and structured knowledge acquisition, aims to extract structured event information from unstructured input data. While substantial progress has been made in uni-modal (typically text-based) event extraction (Qi et al. 2024; Li et al. 2024; Ma et al. 2022), multi-modal event extraction (MEE), as shown in Figure 1, which aims to extract structured event information from both text and images remains relatively underexplored. MEE serves as a key enabler for various high-level applications, such as situational understanding and multimedia content analysis.



Vehicles of the International Committee of the Red Cross (ICRC) and the United Nations wait on a street after an aid **convoy** **entered** the rebel-held **Syrian** town of Daraya, southwest of the capital Damascus, on June 1, 2016

trigger: entered	**Event: Movement:Transport**	
Visual Arguments	Agent	**person**
	Artifact	**box**
Textual Arguments	Artifact	**convoy**
	Destination	**Syrian**

Figure 1: Overiew of multi-modal event extraction

Recent methods (Li et al. 2020; Liu, Chen, and Xu 2022; Du et al. 2023; Seeberger, Wagner, and Riedhammer 2024; Cao et al. 2025) have made notable progress, yet current methods still suffer from three major limitations: (1) the lack of explicit event schema guidance, which hinders models from internalizing the structural semantic constraints of events and prevents them from fully capturing intrinsic event structures and cross-modal consistency; (2) coarse-grained multi-modal alignment strategies that fail to achieve event-level semantic alignment across different modalities; and (3) reliance on heterogeneous and annotation-misaligned datasets, such as ACE 2005 (Walker et al. 2006), imSitu (Pratt et al. 2020), and VOA (Li et al. 2020), hampers unified event-level understanding across modalities, resulting in semantic discontinuity and training inefficiency. To address these challenges, we propose LLaVA-MS-PIT, a multi-modal schema-guided progressive instruction tuning framework. Our core insight is to inject multi-modal event schema knowledge into the model prior to event extraction, thereby enhancing its event perception and reasoning capabilities, and simultaneously achieving event-level cross-modal semantic alignment. Furthermore, we design a schema-guided strategy to construct a multi-modal event extraction dataset derived from imSitu, effectively mitigating data scarcity and cross-modal misalignment.

Existing research in text event extraction shows that event schemas provide the structural, semantic, and operational

*Corresponding author

knowledge necessary for both traditional sequence-labeling methods and modern LLM-based event extraction methods (Xu et al. 2024; Gui et al. 2024; Li et al. 2024). Motivated by these observations, we design multi-modal event schemas to enhance the performance of multi-modal event extraction. First, we construct Progressive Event Schema Decomposition (PESD) instructions based on textual event schemas, simulating human-like event cognition processes in text, ranging from event detection and type analysis to argument role assignment and structural reasoning. Second, we propose the Visual Event Schema (VES), a novel framework that bridges the semantic gap between abstract textual event types and concrete visual instances. VES decomposes images into event-consistent components, including core entities, attributes/actions, interactive relations, and scene context, thereby establishing the first isomorphic cross-modal event schema to enable fine-grained alignment between textual and visual modalities.

In addition, we design a two-stage progressive instruction tuning framework. During event schema-aware fine-tuning, we first fine-tune the model on textual event schema data to establish prior event knowledge and enhance event structural reasoning capability. Then further fine-tune the model with visual schema data to achieve event-level alignment and semantic consistency across modalities. During schema-guided event extraction fine-tuning, the model leverages event schema knowledge acquired in the first stage to perform structured multi-modal event extraction. Finally, to address data scarcity and semantic misalignment in existing training datasets, we employ a schema-guided strategy to construct a parallel multi-modal dataset based on imSitu. With the assistance of GPT-4o, we construct imSitu-MEE, a parallel multi-modal event extraction dataset aligned between imSitu and ACE 2005, ensuring semantic and structural consistency across modalities. Our progressive tuning framework and the constructed multi-modal parallel dataset provide a robust benchmark for training multi-modal event extraction models, experimental results demonstrate the superiority of our schema-guided framework.

In summary, our main contributions are as follows:

- We propose LLaVA-MS-PIT, a novel multi-modal schema-guided progressive instruction tuning framework, injecting event schema knowledge into the model prior to event extraction, enabling explicit event-level alignment and joint reasoning across modalities.
- We design structured multi-modal event schemas, effectively enhancing event perception, structural reasoning, and event-level cross-modal semantic alignment.
- We construct imSitu-MEE, a high-quality parallel multi-modal event extraction dataset, alleviating data scarcity and cross-modal misalignment, and providing a reliable benchmark for future research.
- Experiments demonstrate that our model outperforms existing methods and achieves competitive performance on multi-modal event extraction benchmarks.

Related Work

Sequence Labeling-based Methods

Structural Alignment. Multi-modal event extraction was first introduced in (Li et al. 2020), along with the M^2E^2 benchmark and the WASE framework. WASE encodes textual information via AMR graphs and visual content via Situation Graphs, aligning both in a shared semantic space for joint reasoning. This work establishes the first framework for multi-modal event extraction. CLIP-Event (Li et al. 2022) aligns textual and visual event structures, bridging the cross-modal semantic gap and boosting performance. Building on these works, MGIM (Liu et al. 2024b) proposes a coarse-to-fine alignment that sharpens argument localization and boosts multi-modal event extraction. These methods attempt to enhance MEE performance by aligning structural representations from text and images. However, their effectiveness is often constrained by the quality of structural modeling and the inherent complexity of multi-modal data.

Semantic Space Alignment. UniCL (Liu, Chen, and Xu 2022) learns a joint image-text-event embedding via contrastive learning, achieving strong zero-shot transfer. Most recently, X-MTL (Cao et al. 2025) addresses subtask conflicts and modality gaps by sharing parameters across four sub-tasks and refining them through pseudo-label distillation and adaptive loss weighting.

MMUTF (Seeberger, Wagner, and Riedhammer 2024) addresses multi-modal event argument extraction (MEAE) with a unified template-filling model that formulates event templates as natural-language prompts and extracts arguments by matching template queries with entity candidates.

Sequence labeling enables efficient and interpretable token-level prediction through an explicit label space; yet its reliance on a fixed schema constrains its adaptability to open-domain extension and cross-modal alignment.

Generative-based Methods

In recent years, generative event extraction approaches have also demonstrated significant potential. A clear trend has emerged toward integrating the three major sub-tasks of information extraction, into a unified generative framework, both text-only (Lu et al. 2022; Wang et al. 2023) and multi-modal information extraction. UMIE (Sun et al. 2024) unifies multi-modal NER, RE, and EE by instruction tuning on FLAN-T5 (Chung et al. 2024), achieving SoTA performance in multi-modal event detection. Generative-based methods use prompting or fine-tuning to flexibly handle many tasks. However, they often require expensive fine-tuning, have limited ability to follow instructions, and sometimes generate hallucinated content.

Data Synthesis-based Methods

Data scarcity and cross-modal misalignment remain major challenges in multi-modal event extraction. To address these issues, CAMEL (Du et al. 2023) employs bidirectional data augmentation to create parallel multi-modal data. However, generating missing modality data from uni-modal sources often introduces noise and distributional shifts.

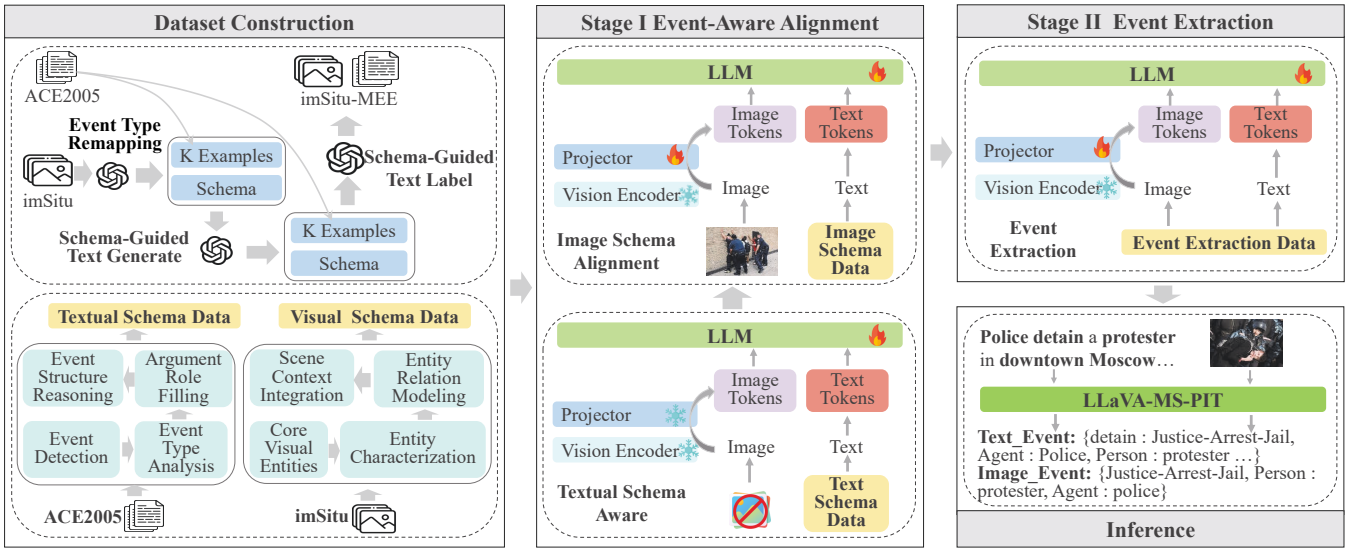


Figure 2: Framework of Multi-modal Schema-guided Progressive Instruction Tuning (LLaVA-MS-PIT)

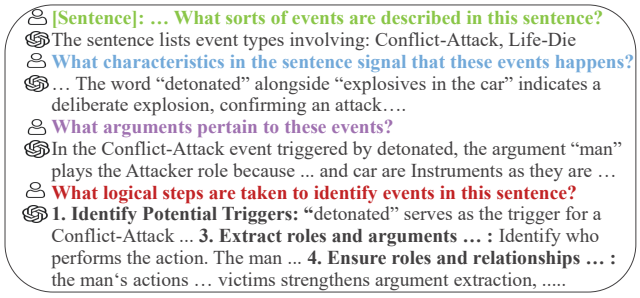


Figure 3: Example of Textual Event Schema

Methodology

In this section, we describe the proposed multi-modal schema-guided progressive instruction tuning framework.

Multi-modal Event Schema Construction

To mitigate the lack of explicit event schema guidance, we first establish structured textual event schema to capture the compositional nature of events in language, and additionally propose visual event schema tailored to represent entities and relations in the visual domain. Guided by these schema representations, we subsequently construct event schema datasets for both modalities, enabling effective and targeted injection of event schema knowledge.

Textual Event Schema Construction Textual event schema (TES) provides a formalized specification of the structure, semantics, and constraints of a particular type of event. It systematically defines the event type, argument roles with type constraints, and triggering conditions. However, traditional event extraction approaches rely on static, manually constructed schema or implicit schema induction, which makes it difficult for LLMs to effectively internalize these structured event patterns. These methods also lack robustness in complex or open-domain scenarios, limiting the model's capacity for deep event understanding and reasoning. To address these limitations and enable efficient and structured injection of event schema knowledge into LLMs, we propose a cognitively inspired paradigm. Specifically, we transform static textual event schema knowledge into dynamic, multi-turn progressive question-answer pairs to simulate the gradual human cognitive process of event understanding. This approach decomposes complex schema knowledge into learnable sub-tasks that can be incrementally internalized by the model. We design a set of progressive event dchema Decomposition (PESD) instructions that comprise four stages:

Event Detection: Guides the model to identify specific event types mentioned in text ("What events are mentioned?"), establishing initial associations between event types, triggers, and contextual information.

Event Type Analysis: Encourages the model to analyze the contextual semantics underlying event occurrence ("Why are these events present based on context?"), deepening its understanding of event triggers, semantic constraints, and contextual dependencies.

Event Argument Role Filling: Instructs the model to accurately locate and classify the semantic roles and corresponding arguments for each event ("What are the constituent semantic roles and arguments for each event?"), internalizing structural constraints and semantic norms for argument roles within specific event types.

Event Structure Reasoning: Requires the model to integrate prior information and explicitly infer the complete event structure and internal relationships ("What logical steps are taken to identify events in this sentence?"), achieving cognitive integration from discrete element identification to structured event generation.

This event schema decomposition-based multi-turn dialogue approach significantly enhances the model's under-

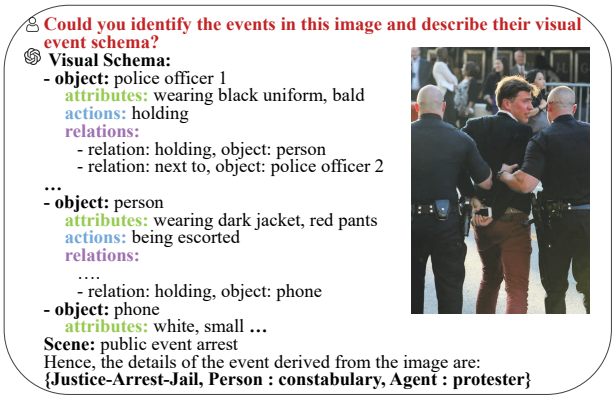


Figure 4: Example of Visual Event Schema

standing of event structure, semantic relationships between argument roles and events, and schema-based constraints, providing a solid semantic foundation for cross-modal alignment. An example is shown in Figure 3.

Visual Event Schema Construction Given the intrinsic semantic divergence between textual and visual modalities, where text encodes abstract event concepts while images convey concrete visual instances, directly applying textual event schema to visual data is fundamentally hindered by the representation gap. Meanwhile, existing multi-modal event extraction models primarily rely on coarse-grained multi-modal alignment strategies, limiting their ability to achieve fine-grained event-level correspondence. To address these constraints, we propose the novel concept of the visual event schema (VES). VES is designed to provide an isomorphic, structured intermediate representation for event instances within an image, thereby linking high-level abstract semantic frameworks to low-level perceptual instances.

The core of VES lies in decomposing visual content into key elements semantically associated with events:

Core Visual Entities: Identification of salient objects in the image that are highly relevant to the target event, corresponding to textual event arguments.

Entity Characterization: Description of the visual attributes of each core object and the key actions it performs within the event context, which correspond to textual triggers that signal event occurrences.

Entity Relation Modeling: Explicit modeling of the interactive relations between core objects in terms of spatial, action-based, or semantic interactions, reflecting implicit relational structure among event arguments on the textual side and forming the dynamic structure of the event.

Scene Context Integration: Recognition of the overall scene category presented in the image, providing essential background information for event understanding and corresponding to the event type in textual event schema.

Formally, we define the visual event schema as follows:

$$\text{VES} = \left(\mathcal{O},\ \{\mathcal{A}_{\text{attr}}(o_i), \mathcal{A}_{\text{action}}(o_i)\}_{o_i \in \mathcal{O}},\ \mathcal{R},\ \mathcal{C}\right) \quad (1)$$

where $\mathcal{O} = o_1, o_2, ..., o_n$ denotes the set of core objects relevant to the event. $\mathcal{A}_{\text{attr}}(o_i)$ represents the set of visual attributes for object o_i. $\mathcal{A}_{\text{action}}(o_i)$ denotes the set of actions performed by o_i in the event. $\mathcal{R} = \text{rel}_k(o_i, o_j), \text{rel}$ is the set of relations among objects. $\mathcal{C}$ denotes the overall scene category of the image.

As a semantic bridge, VES distills the fine-grained details of image instances into a structured representation that is semantically aligned with the textual modality. This unified framework establishes a solid foundation for cross-modal event understanding by reconciling abstraction with perceptual concreteness. An example is shown in Figure 4.

Phased Progressive Fine-tuning Framework

Based on the constructed multi-modal event schema data, we design a two-stage progressive fine-tuning framework (as illustrated in Figure 2), where the first stage focuses on schema knowledge injection, and the second stage performs schema-guided event extraction.

Stage I: Event Schema-Aware Fine-tuning The objective of this stage is to establish an event-aware cross-modal representation space within the model. It consists of two sequential sub-stages.

Textual Schema-Aware Tuning: The PESD instruction set is used to fine-tune the language model (LM), injecting structured textual event schema knowledge. In this sub-stage, the multi-modal projector and visual encoder are frozen, and only the LM parameters are efficiently fine-tuned using LoRA. The objective for this sub-stage is formally defined in Eq. (2) as follows:

$$\min_{\boldsymbol{\theta}_{\text{LLM}}} \mathcal{L}_{\text{text}} = -\sum_{(\mathbf{x}, \mathbf{s}^{\text{text}}, \mathbf{a}^{\text{text}}) \in \mathcal{D}_{\mathcal{T}}} \log p_{\boldsymbol{\theta}_{\text{LLM}}}\left(\mathbf{a}^{\text{text}} \mid \mathbf{x}, \mathbf{s}^{\text{text}}\right) \quad (2)$$

where $\mathbf{x}$ denotes the input text instances; $\mathbf{s}^{\text{text}}$ is the textual event schema data; $\mathbf{a}^{\text{text}}$ is the target annotations for the text instances; $\mathcal{D}_{\mathcal{T}}$ denotes the set of textual event schema data; $\boldsymbol{\theta}_{\text{LLM}}$ denotes the parameters of the language model; and $p_{\boldsymbol{\theta}_{\text{LLM}}}$ denotes the conditional probability distributions parameterized by the language model.

Visual Schema Alignment Tuning: After textual schema fine-tuning, we further fine-tune the model on visual event schema (VES) data to achieve event-level cross-modal alignment. During this sub-stage, only the multi-modal projection layer is updated, mapping visual features into the schema-guided event semantic space and bridging the gap between textual and visual representations. The objective for this sub-stage is formally defined in Eq. (3) as follows:

$$\min_{\boldsymbol{\theta}_{\text{LLM}}, \boldsymbol{\theta}_{\text{proj}}} \mathcal{L}_{\text{img}} = -\sum_{(\mathbf{I}, \mathbf{s}^{\text{img}}, \mathbf{a}^{\text{img}}) \in \mathcal{D}_{\mathcal{I}}} \log p_{\boldsymbol{\theta}_{\text{LLM}}, \boldsymbol{\theta}_{\text{proj}}}\left(\mathbf{a}^{\text{img}} \mid \mathbf{I}, \mathbf{s}^{\text{img}}\right) \quad (3)$$

where $\mathbf{I}$ denotes the input image instances; $\mathbf{s}^{\text{img}}$ denotes the visual event schema data; $\mathbf{a}^{\text{img}}$ denotes the target annotations for the image instances; $\mathcal{D}_{\mathcal{I}}$ denotes the set of visual event schema data; $\boldsymbol{\theta}_{\text{proj}}$ denotes the parameters of the multi-modal projector, $p_{\boldsymbol{\theta}_{\text{LLM}}, \boldsymbol{\theta}_{\text{proj}}}$ denotes the conditional probability distribution parameterized jointly by the language model and the multi-modal projector.

Upon completion of Stage I, the model establishes event awareness and achieves event-level cross-modal alignment,

providing a strong foundation for subsequent event extraction tasks.

Stage II: Schema-Guided Event Extraction Instruction Tuning Building on the event-aware representations established in Stage I, this stage endows the model with structured event parsing capability through task-driven instruction tuning. Based on the constructed multi-modal parallel dataset imSitu-MEE (as introduced in next subsection), we design end-to-end event extraction instructions to fine-tune the parameters of the language model and the multi-modal projector. During this stage, the event schema serves as an implicit constraint, guiding the model to focus on cross-modal cues relevant to event structures, thereby achieving precise identification of event elements and construction of event structures. The objective for this stage is formally defined in Eq. (4) as follows:

$$\min_{\boldsymbol{\theta}_{\text{LLM}},\, \boldsymbol{\theta}_{\text{proj}}} \mathcal{L} = - \sum_{(\mathbf{I},\, \mathbf{d},\, \mathbf{x},\, \mathbf{y}) \in \mathcal{D}} \log p_{\boldsymbol{\theta}_{\text{LLM}},\, \boldsymbol{\theta}_{\text{proj}}}(\mathbf{y} \mid \mathbf{I},\, \mathbf{d},\, \mathbf{x}) \tag{4}$$

where $\mathbf{d}$ represents the event extraction instruction; $\mathbf{y}$ denotes the target event extraction label; $\mathbf{x}$ denotes the input sentence; $\mathcal{D}$ is the overall set of training samples.

Event Schema-Guided Dataset Construction

Current MEE research is constrained by data scarcity (with only the M^2E^2 benchmark) and the inherent limitations of the traditional heterogeneous concatenation training approach, which combines VOA Caption alignment, ACE 2005 text, and imSitu images: (1) Semantic Gap: inconsistent annotation schemes (event/argument definitions) across different data sources; (2) Cross-modal Drift: text and images originate from different contexts, making it difficult to align event instances precisely with visual scenes; (3) Training Redundancy: separate training is required for uni-modal models, leading to inefficiency. To overcome these bottlenecks, we propose a schema-guided data reconstruction framework to construct a high-quality, modality-aligned parallel dataset. Notably, we adopt an image-to-text (Img2Txt) strategy instead of text-to-image (Txt2Img) generation, as our experiments show that models such as Stable Diffusion 3.5 (Esser et al. 2024) often generate complex event scenes with missing subjects or distorted details, thereby introducing additional noise.

imSitu-to-ACE Event Mapping & Data Cleaning The imSitu dataset (504 verbs) annotates visual scenes using $< verb, role, noun >$ triples. Previous studies (Li et al. 2020; Du et al. 2023) directly mapped verbs to ACE 2005 event types. However, we observed that a large proportion of images depict the action but fail to capture the full semantics of the corresponding ACE 2005 event (e.g., images depicting "burying" may not constitute a Die event). To address this, we design a two-stage validation strategy:

LLM Preliminary Filtering: GPT-4o is used to predict the most likely ACE 2005 event type for each image. If the prediction matches the original verb-to-ACE mapping (Li et al. 2020), the image is retained; otherwise, it is marked for further review.

Manual Review: Human experts review and confirm or revise the event type based on visual content and argument distribution.

After these two steps, approximately 53% of the images are discarded due to insufficient alignment with ACE 2005 events. Subsequently, based on the review results,we reconstruct a precise mapping table between the annotation scheme of the imSitu dataset and that of the ACE 2005 dataset, resulting in a high-quality image event subset named imSitu-Clean.

Schema-Guided Text Generation To generate parallel textual descriptions for imSitu-Clean images that are both stylistically consistent and semantically complete, we employ a schema-guided image-to-text generation strategy, in which the prompt incorporates the event schema and selected examples from ACE 2005 as few-shot demonstrations. The curated prompt is then fed into GPT-4o, which produces the corresponding textual description (imSitu-Clean-Text).

Automated Annotation We utilize a schema-aware annotation procedure to annotate imSitu-Clean-Text. Specifically, GPT-4o is prompted with the event schema and representative ACE 2005 examples to automatically label event types, triggers, and argument roles. Through this procedure, we construct imSitu-MEE, a structurally consistent and cross-modally aligned parallel dataset for multi-modal event extraction. imSitu-MEE effectively addresses the semantic gap, the modality drift, and the training redundancy, providing a reliable data foundation for the proposed LLaVA-MS-PIT framework and offering a high-quality training dataset for future MEE research.

Experiment

We rigorously evaluate our framework, benchmark it against recent state-of-the-art models, and analyze the impact of its modules and training strategies.

Experiment Setup

Datasets Training datasets used in our experiments include: (1) ACE 2005, a textual event extraction dataset; (2) imSitu-Clean, a refined visual event dataset derived from imSitu; and (3) imSitu-MEE, a multi-modal parallel dataset generated with explicit event schema guidance. For evaluation, we employ the widely adopted M^2E^2 benchmark (Li et al. 2020).

Evaluation Metrics Considering the generative nature of large language models in event extraction, we follow (Chen et al. 2024) and report text-based event extraction performance under two evaluation scenarios: (1) Strict Mode, requiring exact matches of both the event type and the trigger span; and (2) Loose Mode, requiring correct event types and arguments while relaxing trigger constraints, acknowledging LLMs' focus on holistic semantics. For visual event extraction, correctness is strictly defined as the accurate prediction of event type along with its corresponding arguments. Precision, recall, and F1 scores are consistently used as evaluation metrics.

	Textual Events						Visual Events						MM Events					
	Event Mention			Argument Role			Event Mention			Argument Role			Event Mention			Argument Role		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
$WASE_{obj}$	42.8	61.9	50.6	23.5	30.3	26.4	43.1	59.2	49.9	14.5	10.1	11.9	43	62.1	50.8	19.5	18.9	19.2
CLIP-Event	-	-	-	-	-	-	41.3	**72.8**	52.7	21.1	13.1	17.1	-	-	-	-	-	-
UNICL	49.1	59.2	53.7	27.8	34.3	30.7	54.6	60.9	57.6	16.9	13.8	15.2	44.1	67.7	53.4	24.3	22.6	23.4
MGIM	50.1	66.5	55.8	28.2	34.7	31.2	55.7	64.4	58.5	24.1	14.1	17.8	46.3	69.6	55.6	25.2	21.7	24.6
CAMEL	45.1	71.8	55.4	24.8	41.8	31.1	52.1	66.8	58.5	21.4	28.4	24.4	55.6	59.5	57.5	31.4	35.1	33.2
MMUTF	48.5	65.0	55.5	33.6	44.2	**38.2**	55.1	59.1	57.0	23.6	18.8	20.9	47.9	63.4	54.6	39.9	20.8	27.4
UMIE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	62.1	-	-	24.5
X-MTL	49.7	65.7	56.6	**34.6**	37.6	36.0	**73.1**	70.3	71.7	33.2	31.3	32.2	78.3	57.3	66.2	**40.3**	42.6	41.4
Our-S	53.9	71.9	61.6	23.9	44.5	31.1	72.3	72.3	72.3	57.3	54.3	55.7	72.5	90.0	80.3	33.8	57.1	42.5
Our-L	**60.9**	**81.3**	**69.6**	27.8	**51.9**	36.2	72.3	72.3	**72.3**	**57.3**	**54.3**	**55.7**	**79.4**	**97.4**	**87.5**	37.3	**63.0**	**46.8**

Table 1: Main results on textual, visual, and multi-modal event. **Our-S** and **Our-L** denote our results evaluated under the strict and loose mode, respectively. Best scores are in **bold**; second-best scores are underlined.

Baselines We compare our proposed framework with several representative and widely used approaches, including: $WASE_{obj}$ (Li et al. 2020), CLIP-Event (Li et al. 2022), UNICL (Liu, Chen, and Xu 2022), MGIM (Liu et al. 2024b), MMUTF (Seeberger, Wagner, and Riedhammer 2024), and X-MTL (Cao et al. 2025), UMIE (Sun et al. 2024) and CAMEL (Du et al. 2023).

Implementation Details We select LLaVA-1.5 (7B) (Liu et al. 2024a) as our foundational model. During schema-aware fine-tuning, the LM and the cross-modal projector are optimized with learning rates of 2e-4 and 2e-5, respectively, for 3 epochs with a batch size of 32. For schema-guided event extraction fine-tuning, we adopt the same learning rates and batch size as in schema-aware fine-tuning but reduce training to 1 epoch. All experiments are performed on NVIDIA A800 GPUs.

Main Results

As shown in Table 1, our experimental results demonstrate that LLaVA-MS-PIT achieves competitive results, particularly in event detection tasks. Specifically, our model surpasses the previous best-performing model, X-MTL, by substantial margins of 13.0, 0.6, and 21.3 F1 points on textual, visual, and multi-modal event detection tasks, respectively. Compared to CAMEL, which also employs data synthesis on uni-modal datasets, our approach achieves substantially superior performance, underscoring the effectiveness of our schema-guided multi-modal dataset construction strategy. In visual event argument extraction, our model also demonstrates substantial improvements, which we attribute primarily to our systematic filtering of the imSitu dataset and the refined remapping of argument roles. This process effectively eliminates noise from numerous irrelevant images, resulting in markedly improved visual event extraction performance.

However, for textual event argument extraction, our model exhibits pronounced advantages in recall but moderate performance in precision. A deeper analysis reveals that our model tends to predict a greater number of events and associated arguments, typically generating about 1.5 to 2 times more events and arguments than those present in the gold annotations. This tendency consequently leads to reduced precision.

Ablation Study

To further validate the effectiveness of our proposed training strategies and the explicit integration of multi-modal schema knowledge, we perform additional ablation studies under several comparative experimental conditions: (1) Origin Data Fine-tuning (ODT), where the model is directly fine-tuned on the original, unprocessed datasets; (2) Clean Data Fine-tuning (CDT), which employs the cleaned and aligned imSitu dataset for training; (3) Textual Event Schema Progressive Tuning (CDT-TES), leveraging only textual event schema data for progressive tuning; and (4) Visual Event Schema Progressive Tuning (CDT-VES), leveraging only visual event schema data for progressive tuning.

Table 2 presents the results of our ablation studies, validating the contributions of multi-modal event schema knowledge and progressive tuning to overall performance. Comparative analysis between ODT and CDT clearly demonstrates the effectiveness of our data cleaning methodology. Furthermore, comparing CDT with CDT-TES highlights the substantial contribution of textual event schema knowledge to multi-modal event extraction. Similarly, the comparison between CDT and CDT-VES underscores the critical role of visual event schema knowledge, confirming the importance of structured multimodal semantic guidance.

Further Analysis of Textual Event Schema Knowledge Injection

To more rigorously examine the effectiveness of textual event schema knowledge injection, we conducted further experiments on both the ACE 2005 and M^2E^2 datasets, focusing on the textual modality. While direct improvements on M^2E^2's text-based event extraction appear relatively modest in terms of the event argument extraction F1 scores, a

	Textual Events						Visual Events						MM Events					
	Event Mention			Argument Role			Event Mention			Argument Role			Event Mention			Argument Role		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ODT	57.9	80.1	67.3	24.8	48.2	32.8	60.7	60.7	60.7	69.7	22.7	34.2	-	-	-	-	-	-
CDT	59.3	79.3	67.9	27.7	51.1	36.0	72.3	72.3	72.3	54.3	50.0	51.9	71.3	95.6	81.8	32.1	63.0	42.5
CDT-VES	59.4	81.2	68.6	26.8	51.9	35.4	71.3	71.3	71.3	54.1	54.3	54.2	78.6	96.3	86.5	35.5	62.1	45.2
CDT-TES	60.4	80.8	69.1	27.5	**52.1**	36.0	**73.4**	**73.4**	**73.4**	55.3	51.3	53.2	75.6	97.4	85.1	34.6	62.4	44.5
Our	**60.9**	**81.3**	**69.6**	**27.8**	51.9	**36.2**	72.3	72.3	72.3	**57.3**	**54.3**	**55.7**	**79.4**	**97.4**	**87.5**	**37.3**	**63.0**	**46.8**

Table 2: Ablation experiment results on different Settings. Best scores are in **bold**.

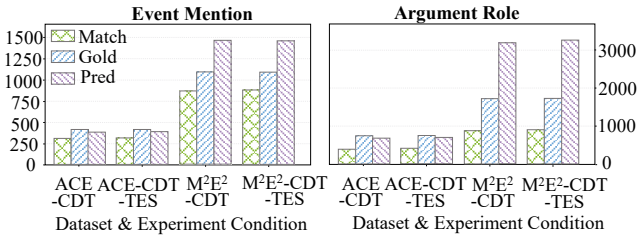


Figure 5: Match/Gold/Pred counts for Event Mention (left) and Argument Role (right) across ACE 2005 and M^2E^2 under two experimental conditions (CDT and CDT-TES). "Match" denotes the number of predictions that correctly match the gold annotations.

detailed investigation into the evaluation metrics and data characteristics reveals deeper insights into the value of textual event schema.

Dataset		CDT		CDT-TES	
		EM	AR	EM	AR
ACE 2005	P	81	58.1	**81.4**	**60.0**
	R	75.2	53.1	**76.2**	**56.2**
	F1	78	55.5	**78.7**	**58.0**
M^2E^2(text)	P	59.3	**27.7**	**60.4**	27.5
	R	79.3	51.1	**80.8**	**52.1**
	F1	67.9	**36.0**	**69.1**	**36.0**

Table 3: Performance comparison on ACE 2005 and M^2E^2(text) datasets under different settings, EM denotes Event Mention, and AR denotes Argument Role. Best scores are in **bold**.

As reported in Table 3, on ACE 2005, CDT-TES surpasses the CDT baseline in event argument extraction, improving F1 from 55.5 to 58.0, with gains in both precision and recall. This demonstrates that the progressive injection of textual event schema knowledge substantially enhances the model's capacity to identify and structure complex event-argument relations, benefiting from explicit schema-level priors that are otherwise lacking in traditional end-to-end fine-tuning methods.

However, our experiments on the M^2E^2 show an unexpected phenomenon. Although schema-guided tuning brings improvements in recall for argument extraction (from 51.1 to 52.1), the overall gain appears less pronounced than on ACE 2005. As shown in Figure 5, the predicted event and argument counts on M^2E^2 differ significantly from the gold annotations: for instance, the model predicts 1478 events versus 1105 gold, and 3266 arguments versus 1723 gold. This discrepancy is far less pronounced on ACE 2005, where predicted and gold counts are well aligned. Upon manual inspection of multiple M^2E^2 examples, we observed a non-negligible number of true event mentions and arguments that are present in the text but absent from the gold annotations, confirming our hypothesis that the gap arises from annotation incompleteness rather than model deficiency.

Crucially, this cross-dataset analysis underscores two important aspects. First, the schema-guided model exhibits a genuine ability to discover and extract plausible events and arguments even when the gold annotations are incomplete or noisy, demonstrating robust generalization and a strong event-centric inductive bias introduced by schema knowledge. Second, the significant performance boost observed on ACE 2005, which provides reliable event annotations, demonstrates that progressive fine-tuning with textual event schema data effectively enhances the model's event awareness and its ability to understand event structures.

Conclusion

In this paper, we propose a novel multi-modal event schema representation and incorporate it into a progressive fine-tuning framework, thereby systematically investigating the importance of event schema knowledge for multi-modal event extraction. Furthermore, we identify critical limitations of the widely used imSitu dataset for this task and perform rigorous data cleaning. Building on this, we develop a novel schema-guided strategy to construct a new parallel multi-modal dataset for event extraction, effectively alleviating the challenges of cross-modal misalignment and data scarcity in this domain. Experimental results show that our model achieves competitive performance on the M^2E^2 benchmark, confirming the effectiveness of our framework.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62476108) and the Project Funding of the Key Laboratory of Intelligent Sensing System and Security (Hubei University), Ministry of Education NO.KLISSS202504.

References

Cao, J.; Hu, Y.; Tan, Z.; and Zhao, X. 2025. Cross-modal Multi-task Learning for Multimedia Event Extraction. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 11454–11462. AAAI Press.

Chen, R.; Qin, C.; Jiang, W.; and Choi, D. 2024. Is a large language model a good annotator for event extraction? In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, AAAI'24/IAAI'24/EAAI'24. AAAI Press. ISBN 978-1-57735-887-9.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tai, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Castro-Ros, A.; Pellat, M.; Robinson, K.; Valter, D.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1).

Du, Z.; Li, Y.; Guo, X.; Sun, Y.; and Li, B. 2023. Training Multimedia Event Extraction With Generated Images and Captions. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 5504–5513. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Muller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; Podell, D.; Dockhorn, T.; English, Z.; Lacey, K.; Goodwin, A.; Marek, Y.; and Rombach, R. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. *ArXiv*, abs/2403.03206.

Gui, H.; Yuan, L.; Ye, H.; Zhang, N.; Sun, M.; Liang, L.; and Chen, H. 2024. IEPile: Unearthing Large Scale Schema-Conditioned Information Extraction Corpus. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 127–146. Bangkok, Thailand: Association for Computational Linguistics.

Li, M.; Xu, R.; Wang, S.; Zhou, L.; Lin, X.; Zhu, C.; Zeng, M.; Ji, H.; and Chang, S. 2022. CLIP-Event: Connecting Text and Images with Event Structures. *CoRR*, abs/2201.05078.

Li, M.; Zareian, A.; Zeng, Q.; Whitehead, S.; Lu, D.; Ji, H.; and Chang, S.-F. 2020. Cross-media Structured Common Space for Multimedia Event Extraction. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2557–2568. Online: Association for Computational Linguistics.

Li, Z.; Zeng, Y.; Zuo, Y.; Ren, W.; Liu, W.; Su, M.; Guo, Y.; Liu, Y.; Lixiang, L.; Hu, Z.; Bai, L.; Li, W.; Liu, Y.; Yang, P.; Jin, X.; Guo, J.; and Cheng, X. 2024. KnowCoder: Coding Structured Knowledge into LLMs for Universal Information Extraction. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8758–8779. Bangkok, Thailand: Association for Computational Linguistics.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved Baselines with Visual Instruction Tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26286–26296.

Liu, J.; Chen, Y.; and Xu, J. 2022. Multimedia Event Extraction From News With a Unified Contrastive Learning Framework. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 1945–1953. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Liu, Y.; Liu, F.; Jiao, L.; Bao, Q.; Sun, L.; Li, S.; Li, L.; and Liu, X. 2024b. Multi-Grained Gradual Inference Model for Multimedia Event Extraction. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10): 10507–10520.

Lu, Y.; Liu, Q.; Dai, D.; Xiao, X.; Lin, H.; Han, X.; Sun, L.; and Wu, H. 2022. Unified Structure Generation for Universal Information Extraction. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5755–5772. Dublin, Ireland: Association for Computational Linguistics.

Ma, Y.; Wang, Z.; Cao, Y.; Li, M.; Chen, M.; Wang, K.; and Shao, J. 2022. Prompt for Extraction? PAIE: Prompting Argument Interaction for Event Argument Extraction. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6759–6774. Dublin, Ireland: Association for Computational Linguistics.

Pratt, S.; Yatskar, M.; Weihs, L.; Farhadi, A.; and Kembhavi, A. 2020. Grounded Situation Recognition. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 314–332. Cham: Springer International Publishing.

Qi, Y.; Peng, H.; Wang, X.; Xu, B.; Hou, L.; and Li, J. 2024. ADELIE: Aligning Large Language Models on Information Extraction. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7371–7387. Miami, Florida, USA: Association for Computational Linguistics.

Seeberger, P.; Wagner, D.; and Riedhammer, K. 2024. MMUTF: Multimodal Multimedia Event Argument Extraction with Unified Template Filling. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 6539–6548. Miami, Florida, USA: Association for Computational Linguistics.

Sun, L.; Zhang, K.; Li, Q.; and Lou, R. 2024. UMIE: Unified Multimodal Information Extraction with Instruction Tuning. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 19062–19070. AAAI Press.

Walker, C.; Strassel, S.; Medero, J.; and Maeda, K. 2006. ACE 2005 Multilingual Training Corpus. *Progress of Theoretical Physics Supplement*, 110(110): 261–276.

Wang, X.; Zhou, W.; Zu, C.; Xia, H.; Chen, T.; Zhang, Y.; Zheng, R.; Ye, J.; Zhang, Q.; Gui, T.; Kang, J.; Yang, J.; Li, S.; and Du, C. 2023. InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction. *ArXiv*, abs/2304.08085.

Xu, Z.; Wang, P.; Ke, W.; Li, G.; Liu, J.; Ji, K.; Chen, X.; and Wu, C. 2024. Incorporating schema-aware description into document-level event extraction. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 6597–6605.