MEUR: A Benchmark for Evaluating Vision-Language Models on Multimodal Event Understanding and Reasoning

Zimu Wang[1,2], Yuqi Wang[3], Tong Chen[1,2], Changyu Zeng[4], Hongbin Na[5], Nijia Han[1,2], Fuyu Xing[6], Qi Chen[1], Qiufeng Wang[1], Anh Nguyen[2], Shuihua Wang[1], Ling Chen[5], Jionglong Su[1], Haiyang Zhang[1,†], Wei Wang[1,†]

[1]Xi'an Jiaotong-Liverpool University [2]University of Liverpool [3]Shanghai Jiao Tong University
[4]Eastern Institute of Technology [5]University of Technology Sydney [6]Carnegie Mellon University
Zimu.Wang19@student.xjtlu.edu.cn, {Haiyang.Zhang, Wei.Wang03}@xjtlu.edu.cn

Abstract

Event understanding and reasoning play critical roles in thoroughly evaluating the capabilities of Vision-Language Models (VLMs); however, existing Visual Question Answering (VQA) datasets predominantly focus on entity-centric questions, while event- or action-related questions are limited in scale and suffer from significant shortcut issues. We introduce MEUR, the first **M**ultimodal **E**vent **U**nderstanding and **R**easoning dataset consisting of 1,200 images and 4,217 questions, necessitating VLMs with a diverse range of multimodal understanding and reasoning capabilities to answer, ranging from basic event recognition to more complex tasks such as counting and comparison. To streamline the annotation process, we propose a novel semi-automated pipeline that combines advanced VLMs with human annotators, achieving high quality and efficiency. We conduct extensive experiments on state-of-the-art non-thinking and thinking VLMs to demonstrate their capabilities and limitations in multimodal event understanding and reasoning. Furthermore, we provide a detailed error analysis that points out promising directions for future research.

Keywords: Event Understanding, Event Reasoning, Vision-Language Models

1. Introduction

Events are defined as the instances or occurrences that constitute the basic semantic building units, encompassing the meanings of *Activities*, *Accomplishments*, *Achievements*, and *States* (Vendler, 1957). Effectively interpreting and analyzing events, along with their complex associations, is pivotal to understanding real-world dynamics and underlies applications such as event knowledge graph construction (Ma et al., 2022; Wei et al., 2024), future event prediction (Lin et al., 2022; Rong et al., 2025), and machine reading comprehension (Zhu et al., 2023; Ouyang et al., 2024).

Recent progress has witnessed an increasing emphasis on tackling issues in event understanding and reasoning, with several benchmarks have been introduced to evaluate tasks, including event detection (Wang et al., 2020; Yao et al., 2022), event argument extraction (Li et al., 2021; Wang et al., 2024), and event-centric reasoning, such as temporal/causal (O'Gorman et al., 2016; Wang et al., 2022) and multi-hop reasoning (Li et al., 2024b). However, current research has primarily focused on natural language texts, while multimodal event understanding and reasoning remain underexplored, largely due to the absence of dedicated benchmarks. While certain Visual Question Answering (VQA) datasets, such as VQA-E (Li et al., 2018) and GQA (Hudson and Manning, 2019), involve questions related to events or actions, they are limited in scale and diversity and suffer from significant shortcut problems. As exemplified in Figure 1, comprehending events and their associated information presents a substantially greater challenge than interpreting objects, attributes, and relation-

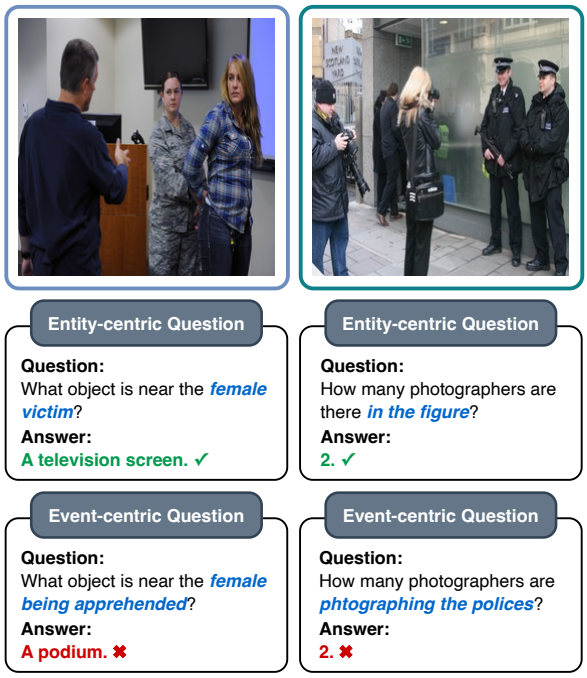


Figure 1: Examples of *entity-centric* and *event-centric* questions and their VLM responses.

[†]Corresponding authors.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1696–1709
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ships typically emphasized in current VQA datasets (Liu et al., 2024; Wang et al., 2024a). Such comprehension necessitates a more nuanced, compositional reasoning over both visual and non-visual cues. To effectively answer these questions, Vision-Language Models (VLMs) must capture the structure of events, including their related entities, roles, and interactions, posing substantial challenges and offering a robust benchmark for evaluating the reasoning capabilities of VLMs.

Motivated by the above phenomenon, we introduce the first **M**ultimodal **E**vent **U**nderstanding and **R**easoning (MEUR) dataset consisting of 1,200 images and 4,217 questions, necessitating VLMs to be equipped with a diverse range of multimodal understanding and reasoning capabilities to answer, including Event Recognition, Argument Identification, Argument Associate Identification, Counting and Comparison, and Unanswerable Questions. The first two tasks correspond to existing event understanding tasks (i.e., event detection and event argument extraction), the latter two address reasoning challenges, and the final one evaluates the hallucination of VLMs. To streamline the annotation process, we propose a novel semi-automated pipeline that repurposes the existing situation with groundings dataset, SwiG (Pratt et al., 2020), while aligning the event types and argument roles with the MEED dataset (Wang et al., 2021). Our approach to annotating event understanding questions follows a "VLM-then-Human" approach, where an advanced VLM, GPT-4o mini (Hurst et al., 2024), is used to generate initial "Event Recognition" and "Argument Identification" questions, which are then evaluated using GPT-4o under established criteria, followed by human verification to ensure accuracy. Human annotators then complete the remaining annotations, ensuring data quality and mitigating shortcut issues. Finally, the human-annotated questions undergo paraphrasing by GPT-4o mini to enhance linguistic diversity and variability.

We conduct comprehensive evaluations of state-of-the-art VLMs, including both non-thinking and thinking models, on the MEUR dataset. The results reveal the considerable difficulty posed by event understanding and reasoning tasks. Notably, the most advanced model, Seed 1.6 Vision, achieves only 40.65% overall accuracy. All models struggle with accurately identifying specific events and analyzing their associated participants, and thinking models, especially GPT-5 mini and Qwen3-VL, often produce overly specific answers that fail to handle unanswerable questions effectively. Meanwhile, non-thinking models exhibit substantial shortcomings in discrete reasoning tasks, with particularly weak performance in counting and comparison. To point out useful directions to guide future research, we conduct an in-depth error analysis, identifying key issues such as incorrect event recognition, inaccurate fine-grained argument identification and comprehension, and event-centric hallucinations.

The key contributions of this work can be summarized as follows:

- We introduce MEUR, the first dataset tailored for multimodal event understanding reasoning. It includes 1,200 images and 4,127 questions, grounded in 5 question strategies, necessitating a diverse range of multimodal understanding and reasoning capabilities to answer.
- We conduct extensive evaluations of state-of-the-art VLMs, including both non-thinking and thinking models, to analyze the capabilities and limitations of these models in multimodal event understanding and reasoning.
- We perform a thorough error analysis and identify key issues such as incorrect event recognition, inaccurate fine-grained argument identification and comprehension, and event-centric hallucinations, for future research in this field.

2. Related Work

Event Understanding and Reasoning. With the progressive advancement of text mining, numerous works have emerged on event understanding and reasoning. The predominant line centers on textual event understanding, commonly formulated as the event extraction task, decomposed into two subtasks: (1) *event detection*, which identifies lexical trigger words (i.e., keywords or phrases that evoke the events) and classifies their event types; and (2) *event argument extraction*, which identifies the arguments and their argument roles (Liu et al., 2020). Under this task, various datasets, such as ACE 2005 (Walker et al., 2006), TAC KBP (Ellis et al., 2014, 2015, 2016; Getman et al., 2017), MAVEN (Wang et al., 2020), and LEVEN (Yao et al., 2022), have been introduced. Another line of research targets inter-event semantics, modeling relations among event mentions, such as coreference (Pradhan et al., 2007; Cybulska and Vossen, 2014), temporal (Verhagen et al., 2009, 2010; UzZaman et al., 2013), causal (Dunietz et al., 2017), and subevent (Glavaš et al., 2014) relationships, and often combines multiple relation types (Mirza et al., 2014; O'Gorman et al., 2016; Caselli and Vossen, 2017; Wang et al., 2022) or advances to deeper reasoning, such as multi-hop reasoning (Li et al., 2024b), across event-argument structures. Many recent work extends the modalities of event extraction toward image (Pratt et al., 2020; Wang et al., 2021) and broader multimedia (Li et al., 2020), with methodologies using LLMs and LVLMs being proposed (Peng et al., 2023a,b; Wang et al., 2024b; Li

Strategy	Example 1	Example 2
Image		
Event Recognition	**Question:** What is the **soldier** doing to the **adult female**? **Answer:** **Apprehending**.	**Question:** What are the **people** doing to the **adult female** in gray? **Answer:** **Interviewing**.
Argument Identification	**Question:** Who is **apprehending** an **adult female** in room? **Answer:** A **soldier**.	**Question:** Who is being **interviewed** by **people**? **Answer:** An **adult female**.
Argument Associate Identification	**Question:** What object is near the female being **apprehended**? **Answer:** A television screen.	**Question:** What color of T-shirt is worn by the person being **interviewed**? **Answer:** Gray.
Counting and Comparison	(Not Applicable)	**Question:** How many people are conducting the **interview**? **Answer:** 3.
Unanswerable Questions	**Question:** Who is **apprehending** the adult male? **Answer:** Unanswerable.	(Not Applicable)

Table 1: Examples for the question strategies in the MEUR dataset. Event triggers and arguments are highlighted in **blue** and **green**, respectively. Some question strategies may not be applicable to some images, which are denoted as "(Not Applicable)".

et al., 2025a; Xing et al., 2025). Despite progress made, most existing work remains limited to recognizing events depicted while underexploring the crucial dimension of event-centric reasoning. We emphasize *reasoning* within event structures, introduce the first dataset, and delineate the unique challenges inherent to this problem.

Multimodal Understanding and Reasoning. A variety of studies have concentrated on multimodal understanding and reasoning, often framed as free-form, open-ended VQA tasks. In these tasks, natural language questions are constructed based on an image, and models are tasked with producing coherent natural language answers as outputs. Existing VQA datasets can be classified into four categories: object-centric (e.g., DAQUAR (Malinowski and Fritz, 2014) and Visual7W (Zhu et al., 2016)), relationship-centric (e.g., MovieQA (Tapaswi et al., 2016) and CLEVR (Johnson et al., 2017)), domain-specific (e.g., VQA-RAD (Lau et al., 2018) and VQA-Med (Abacha et al., 2020)), and knowledge-centric (e.g., KB-VQA (Wang et al., 2017) and E-VQA (Yang et al., 2023)) datasets. With the swift development of VLMs, researchers have also benchmarked their capabilities in multimodal reasoning, with a particular focus on mathematical reasoning (Kang et al., 2025a,b, 2026; Li et al., 2025b).

However, most existing datasets, particularly those designed for VQA, are primarily entity-centric, leaving event-centric datasets relatively underexplored. Additionally, the heavy reliance on automatically generated datasets (Sharma and Jalal, 2021) raises concerns about their suitability for robust reasoning tasks. While certain VQA datasets, such as VQA-E (Li et al., 2018) and GQA (Hudson and Manning, 2019), involve event- or action-related questions, these questions suffer from serious shortcut issues (see Section 3.2). In this paper, we build upon VQA-formed multimodal understanding and reasoning by proposing a novel semi-automated pipeline for understanding and reasoning questions, respectively, enhancing question diversity while mitigating shortcut problems in reasoning questions.

3. Dataset Construction

3.1. Multimodal Event Understanding and Reasoning Questions

Table 1 presents annotation examples for two images, each corresponding to the proposed question

strategies. To enhance the integration of both visual content and event structures, we propose five innovative strategies tailored for event understanding and reasoning questions, which we construct using a semi-automated process, the details of which are elaborated in Section 3.2.

Event Recognition: In this type of question, VLMs are evaluated on their ability to interpret the event occurring between the specified arguments, where the trigger words (i.e., verbs) annotated in the SWiG dataset (Pratt et al., 2020) are selected to represent the answers. For instance, in the first case from Table 1, models are expected to first identify the two event arguments, "*soldier*" and "*adult female*," within the image, and then determine the event connecting these arguments, as shown in the reasoning chain depicted in Figure 2(a).

Argument Identification: This type of question assesses the bridging capability of VLMs (Trivedi et al., 2022), where models are tasked with identifying the target argument based on the provided event (and its associated arguments). For the first example in Table 1, given the event trigger "*apprehended*" and an argument "*adult female*," VLMs should infer the target argument, the soldier, who apprehends the adult female, as illustrated in the reasoning chain in Figure 2(b).

Argument Associate Identification: Similar to "Argument Identification", this type of question also evaluates the bridging capability of VLMs while introducing an additional reasoning step involving the target argument. Inspired by entity-centric datasets (Liu et al., 2024), these questions prompt VLMs to deduce attributes (e.g., colors) and relations (e.g., next to) associated with the target arguments. As shown in Figure 2(b), VLMs are required to perform an extra reasoning step to identify the objects located near the soldier.

Counting and Comparison: This question type evaluates the discrete reasoning capability of VLMs (Dua et al., 2019). In scenarios where multiple entities are present, only a subset may possess a specific attribute towards the given event (e.g., *an interviewee*). Such questions pose unique challenges, requiring models to not only identify discrete entities but also associate them with the event while performing tasks such as counting and comparison. For the second example from Table 1, VLMs should first identify the people conducting the interview and then determine the number of interviewers involved in the event, which is 3.

Unanswerable Questions: The aforementioned question types are generally answerable through proper reasoning grounded in the image. However, not all propose a distinct question type intentionally designed to be unanswerable based on the image, which serves to evaluate the hallucination tendencies of VLMs, as shown in the example illustrated in Table 1. To create this type of questions, annotators are instructed to modify the event trigger by substituting it with another trigger from the same group of event types (e.g., `Conflict.Attack` → `Conflict.Demonstrate`) or to alter the argument by reassigning one argument to a different role.

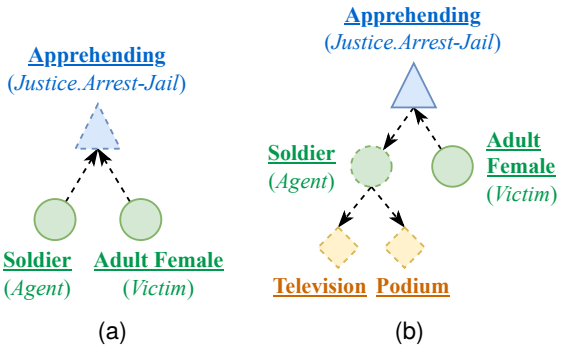


Figure 2: Reasoning chain examples for "Event Recognition," "Argument Identification," and "Argument Associate Identification" questions. The elements in the dash represent those not given in the question, and the dashed lines illustrate the reasoning chains to derive the answers.

3.2. Dataset Construction Pipeline

Figure 3 illustrates the overall pipeline for constructing the MEUR dataset, which is a semi-automatic framework that involves VLMs with human annotators to reduce annotation efforts to enhance efficiency. In this section, we introduce each of the steps in detail.

Data Sample Selection. We begin by selecting appropriate data samples from the Situation with Groundings (SwiG) dataset (Pratt et al., 2020) by referring to the event schema proposed by the MEED dataset (Wang et al., 2021), consisting of 66 event types and 81 argument roles, as illustrated in Figure 8. Each data sample includes an image I, three textual descriptions $d_i, i \in \{1, 2, 3\}$, an event trigger t corresponding to an event type $e \in \mathcal{E}$, and a list of arguments paired with their roles $\{(a_j, r_j)\}, r_j \in \mathcal{R}$. As described in Section 3.1, most types of questions necessitate comprehension across multiple arguments; therefore, we focus on data samples that contain at least two arguments, which support the creation of complex and informative questions. To assist human annotators' comprehension, we visualize the bounding boxes of the arguments in the image and label their roles within these boxes

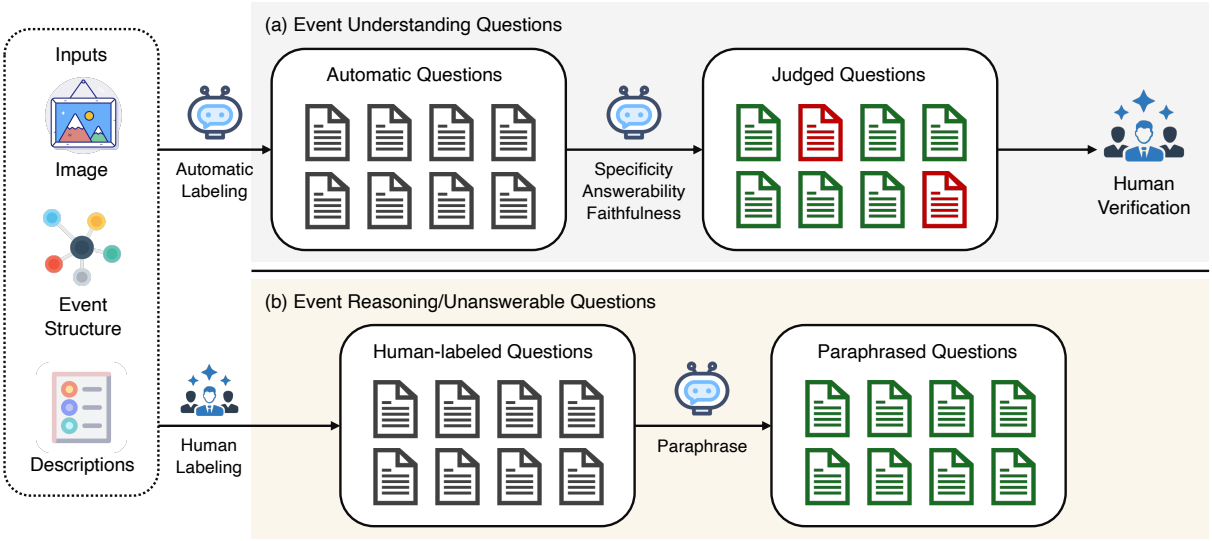


Figure 3: Overall collection pipeline of the MEUR dataset, with specific methods tailored for event understanding and reasoning questions, respectively.

Task Definition

You are a well-trained annotator for event understanding and reasoning questions. Given an image, its descriptions (Descriptions), an event trigger (Trigger), and a set of named arguments (Arguments), your goal is to create a specific, unique, and answerable Question-Answer (QA) pairs. [...]

Annotation Guidelines

1. *Clearly ask about the event, described by the Trigger, [...];*
2. *Be highly specific. If an argument cannot be clearly identified in the image by the given Argument role, [...];*
3. *Be unique and non-formulaic. Avoid using the same question template for every input. [...];*
4. *Be directly answerable. The question must be solely and unambiguously answerable by the given trigger word, [...].*

Rules

1. *Rule of Question Composition: [...];*
2. *Rule of Uniqueness and Answerability: [...];*
3. *Rule of Faithfulness: [...].*

In-context Examples

Figure 4: Prompt example for generating the "Event Recognition" questions, including a task definition, annotation guidelines, annotation rules, and two in-context examples.

using the OpenCV library[1].

Event Understanding Questions. LLMs and VLMs have shown great promise in understanding event structures (Peng et al., 2023a; Wang et al., 2024b) and performing event-centric annotations (Li et al., 2024a,b). Considering the structured nature of event understanding questions, such as recognizing the occurrence of events or identifying specific arguments, we adopt an advanced LLM, GPT-4o mini (Hurst et al., 2024), to automatically generate questions for "Event Recognition" and "Argument Identification," with the prompt detailed in Figure 4. It includes a task definition, annotation guides, rules, and two in-context examples, ensuring accuracy and instruction following through in-context learning (Brown et al., 2020). All relevant information from the SWiG and MEED datasets, including the image, descriptions, event triggers, and arguments, is incorporated to guide the generation outcomes. To evaluate the quality of the generated questions, as shown in Figure 3, we employ an "LLM-as-a-Judge" framework using GPT-4o to assess questions based on three metrics: *specificity*, *answerability*, and *faithfulness*. Any question failing at least one criterion undergoes human verification, and inaccurate questions are removed. After this process, a subset of 100 images is sampled for further validation, achieving a question accuracy rate exceeding 99%, thereby confirming the high quality of these two question strategies.

Event Reasoning Questions. The annotation of event reasoning questions is primarily conducted by human annotators, comprising undergraduate and postgraduate students from reputable universities. Annotators are provided with a comprehensive set of resources, including images, descriptions, event triggers, arguments, and the finalized event understanding questions, and their task is to craft high-quality event reasoning questions based on this information. Prior to annotation, all participants underwent thorough training, which covered basic knowledge of VQA and multimodal reasoning, as well as key concepts related to event understanding, such as event triggers, event arguments, and

[1] https://opencv.org/

Figure 5: Question examples with shortcut problem from the GQA dataset.

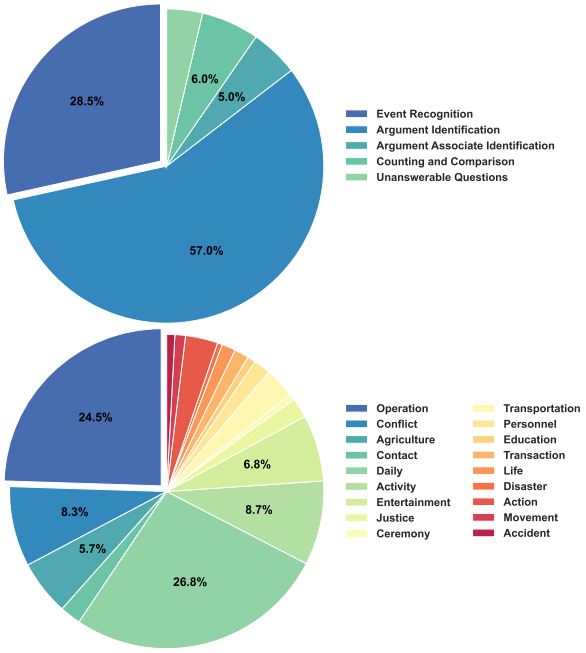


Figure 6: Distribution of question strategies (top) and event types (bottom) within the MEUR dataset.

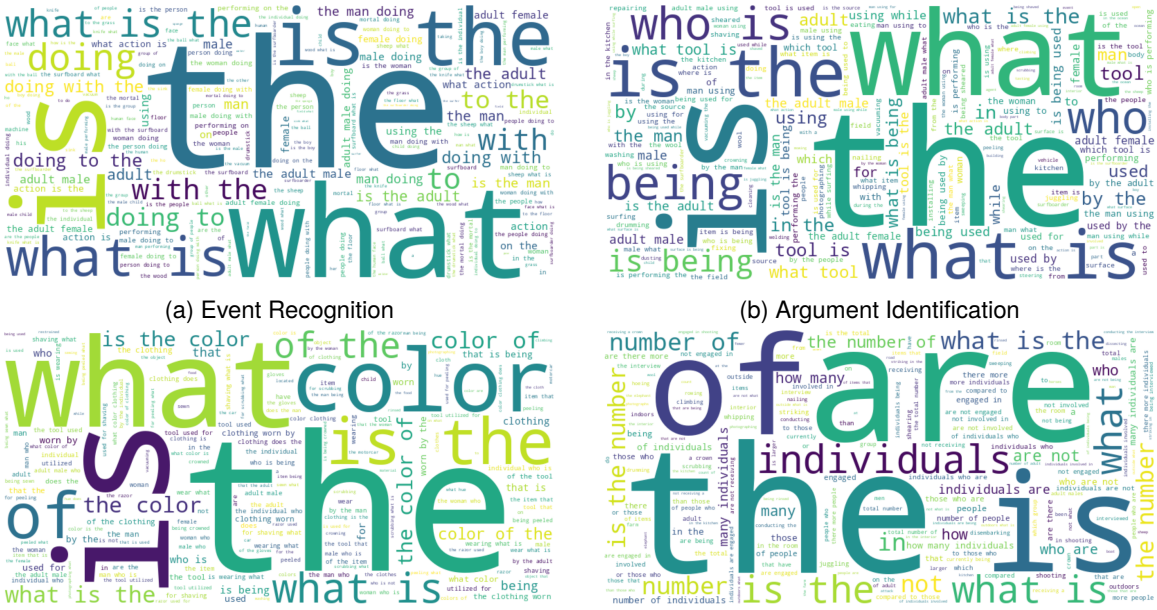
argument roles. Additionally, a detailed annotation guideline was provided to standardize the process and ensure consistent quality. Afterward, a qualification test was conducted, and the top two candidates demonstrating the highest accuracy and efficiency were selected. Each annotator was assigned 600 images and was also responsible for cross-checking the annotations of their counterpart. In cases of disagreement, a consensus was reached through discussions with the project lead. Finally, to further promote question diversity, the annotated questions were paraphrased using GPT-4o mini, ensuring a diverse and enriched dataset.

Desiderata on Shortcut Problems. Reasoning shortcuts present a critical challenge to the quality of multi-hop reasoning datasets (Trivedi et al., 2022; Li et al., 2024b), and this is similarly prevalent in existing VQA datasets, primarily due to their reliance on automatically generated data (Sharma and Jalal, 2021). Figure 5 illustrates two examples from the GQA dataset (Hudson and Manning, 2019), showcasing the image, the original question, and a modified version with redundant information removed. In the first example, although the image describes the person "*holding the umbrella*," this detail is unnecessary for reasoning, as only one person is present in the image. Consequently, despite many datasets involving event- or action-related questions, reasoning shortcuts render them unsuitable for comprehensive multimodal event reasoning. This highlights the crucial need for human annotation in developing reasoning questions. In our dataset curation process, we prioritize training human annotators to avoid introducing shortcut questions and deliberately checking each other's work to ensure such problems are fully addressed.

3.3. Dataset Characteristics

Question Distribution. The MEUR dataset consists of 1,200 images and 4,217 questions, which is tailored to evaluate the capabilities of VLMs on complex event understanding and reasoning tasks. Figure 6 illustrates the distribution of question strategies within the dataset, where "Event Recognition" and "Argument Identification" constitute the majority. Although "Argument Associate Identification" and "Counting and Comparison" account for a smaller proportion, it is important to highlight the distinct challenges involved in annotating these types of questions, particularly in formulating questions that avoid shortcut problems. Notably, the dataset's distribution aligns with established reasoning benchmarks, such as HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and MEQA (Li et al., 2024b). MEUR also features a wide range of event types, as illustrated in Figure 6. It includes 18 groups of event types, with each group containing multiple specific types (e.g., `Conflict.Attack` and `Conflict.Demonstrate` under the `Conflict` group). This diversity, both in questions and events, establishes MEUR as a pioneer benchmark for evaluating the event understanding and reasoning capabilities of VLMs.

Most Frequent n-grams. To analyze the semantic features of different question strategies, we present the most frequent n-grams within the questions, including unigrams, bigrams, and trigrams, in Figure 7. The analysis reveal that the most frequent n-grams align closely with their respective strategies. For instance, terms like "*actions*" and "*doing*" frequently appear in the "Event Recognition" strategy, whereas phrases such as "*how many*"



(a) Event Recognition

(b) Argument Identification

(c) Argument Associate Identification

(d) Counting and Comparison

Figure 7: The most frequent n-grams in the questions within different strategies.

and "*compared to*" are common in the "Counting and Comparison" strategy. This alignment not only demonstrates the consistency between the questions and their corresponding strategies but also highlights the high quality and reliability of the semi-automatic framework used to develop the dataset.

4. Experiments

4.1. Experimental Setup

Models and Setup. We conduct a comprehensive evaluation of state-of-the-art VLMs, including non-thinking and thinking models, on the full MEUR dataset. For non-thinking models, we experiment with GPT-4o (Hurst et al., 2024, 2024-11-20), GPT-4o mini, Grok 4 Fast[2] (Non Reasoning), DeepSeek-VL2 (Wu et al., 2024), and Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025). For thinking models, we evaluate GPT-5 mini[3], Gemini 2.5 Flash (Comanici et al., 2025), Grok 4 Fast (Reasoning), Seed 1.6 Vision, GLM-4.1V-Thinking (Hong et al., 2025), and Qwen3-VL-30B-A3B-Thinking. All hyperparameters are set to their default or recommended values during the inference process.

Evaluation Metrics. To ensure fair comparison, we adhere to standard metrics from VQA and multimodal reasoning (Kuang et al., 2025; Zeng et al., 2025) to evaluate model performance: (1) **Accuracy** (Acc.), including both overall accuracy and the per-question-strategy accuracy; (2) **Similarity** (Sim.) is computed as cosine similarity between model outputs and references in the BGE-Large (EN-v1.5, Xiao et al., 2024) embedding space.

4.2. Experimental Results

Table 2 presents the results of both non-thinking and thinking VLMs on the MEUR dataset. From the result, we observe the distinct challenges posed by these questions. Grok 4 Fast and Seed 1.6 Vision demonstrate the best performance in non-thinking and thinking models, respectively; however, their results remain unsatisfactory, with the top-performing model, Seed 1.6 Vision, achieving only 40.65% accuracy, highlighting substantial room for improvement in future model development.

Experimental results across different question strategies reveal that "Argument Associate Identification" achieves the highest performance, followed by "Event Recognition," while "Argument Identification" and "Counting and Comparison" show the weakest results. This indicates that VLMs excel at identifying apparent cues, such as recognizing basic attributes or recognizing the occurrence of events, but struggle with fine-grained argument understanding or discrete reasoning in counting and comparison. Furthermore, models exhibit varying performance when addressing unanswerable questions. Non-thinking models demonstrate relatively stable performance on such questions, whereas thinking models show significant variability. Interestingly, thinking models perform worse than non-thinking models on this question type, as they often produce overly confident and determined outputs.

[2] https://x.ai/news/grok-4-fast

[3] https://openai.com/index/gpt-5-system-card/

Model	#Param.	Overall		Per-class Accuracy				
		Acc.	Sim.	ER	AI	AAI	CC	UNA
Non-thinking VLMs								
GPT-4o	–	27.41	73.41	34.92	20.57	48.80	23.51	**52.90**
GPT-4o mini	–	29.00	75.46	**39.67**	22.65	37.32	27.49	36.13
Grok 4 Fast	–	**30.61**	**78.62**	35.92	**25.98**	49.28	38.25	23.87
DeepSeek-VL2	27B	20.18	70.31	17.42	16.15	38.28	37.85	50.97
Qwen3-VL (Inst.)	30B-A3B	28.01	74.51	26.42	23.69	**54.07**	**41.43**	50.32
Thinking VLMs								
GPT-5 mini	–	37.11	**82.39**	40.33	33.97	61.72	52.99	1.94
Gemini 2.5 Flash	–	34.43	79.29	42.50	27.89	46.89	49.40	32.26
Grok 4 Fast (Reason.)	–	33.58	79.10	41.08	27.27	53.11	41.83	**33.55**
Seed 1.6 Vision	–	**40.65**	81.46	**45.25**	**35.30**	**68.90**	52.99	29.68
GLM-4.1V-9B	9B	34.31	79.93	38.75	29.31	59.81	48.21	20.65
Qwen3-VL (Think.)	30B-A3B	36.64	80.93	44.33	30.43	65.07	**58.17**	0.00

Table 2: Experimental results of non-thinking and thinking models on the MEUR dataset. The best and the second-best values for each metric within each model type are highlighted in **bold** and underlined, respectively.

Example	Example 1	Example 2	Example 3
Image			
Question	What is the woman doing to the airplane?	How many individuals are hoeing in the field?	Which tool is used by the female child to wash the plate?
Ground Truth	**Disembarking.**	**2.**	**A sponge.**
Prediction	**Boarding.**	**3.**	**A washing machine.**

Table 3: Error types of VLMs on the MEUR dataset with examples. Ground truths and incorrect predictions are highlighted in **green** and **red**, respectively.

This observation aligns with prior findings (Chen et al., 2025), particularly for models like GPT-5 mini and Qwen3-VL, which frequently fail to correctly identify unanswerable questions.

4.3. Error Analysis

We conduct a detailed error analysis to highlight key challenges and suggest directions for future research, with examples organized in Table 3:

1. **Inaccurate Event Recognition:** VLMs occasionally fail to correctly identify events depicted in images, especially when those events fall within the same category or exhibit similar visual characteristics. For the first example in Table 3, the model detects an interaction between the woman and the plane but misclassifies the "*disembarking*" event as "*boarding*," both of which belong to the `Transportation` category.

2. **Inaccurate Fine-grained Argument Identification and Comprehension:** VLMs also exhibit difficulty in accurately identifying the arguments associated with events, often over-attributing participation by labeling all entities within the image as event participants. For the second example, although only two of the three individuals in the image are engaged in hoeing, the model incorrectly predicts that all three are participating in the event.

3. **Event-centric Hallucinations:** VLMs further suffer from hallucination issues when identifying events and their associated arguments,

often relying on commonsense knowledge instead of grounding their predictions in visual evidence, leading to responses that are inconsistent with the image content. For the third example, models answer "*washing machine*" as the tool used to wash a plate based on commonsense knowledge, while overlooking the sponge that is clearly being used in the image.

These challenges significantly hinder the ability of VLMs to perform accurate multimodal event understanding and reasoning. Addressing these limitations is critical for advancing the field and improving the robustness of VLMs in future research.

5. Conclusion

We introduce MEUR, the first dataset specifically designed to evaluate VLMs on the multimodal event understanding and reasoning task. MEUR includes 1,200 images and 4,127 questions, grounded in 5 question strategies, necessitating a diverse range of multimodal understanding and reasoning capabilities to answer. To streamline the annotation process, we propose a novel semi-automated pipeline that combines advanced VLMs with human annotators, achieving high quality and efficiency. We conduct extensive experiments on state-of-the-art non-thinking and thinking VLMs to demonstrate their capabilities and limitations in multimodal event understanding and reasoning, accompanied by an in-depth error analysis to point out useful directions for future research. In the future, we will extend the benchmark with more diverse yet underexplored reasoning questions and propose innovative methods to improve model capabilities in terms of multimodal event understanding and reasoning.

Limitations

Although MEUR pioneers the research by introducing the first dataset on multimodal understanding and reasoning, the high-quality dataset annotation leads to high annotation cost and difficulty, especially in ensuring reasoning questions that mitigate shortcut issues. How to further improve the automatic or semi-automatic data collection remains an open problem. Meanwhile, since the SWiG dataset (Pratt et al., 2020) is collected in English, MEUR is also limited to monolingual. We advocate for further research efforts to develop more diverse datasets with multilinguality.

Ethical Considerations

We discuss the following ethical considerations related to our MEUR dataset as follows: (1) **Intellectual Property.** The SWiG dataset (Pratt et al., 2020) is shared under the MIT License[4], and the MEED dataset (Wang et al., 2021) is published as an open resource[5], both of which are free for research use. We release our dataset under the MIT License on GitHub[6]. (2) **Annotators Treatments.** We hired student annotators and fairly pay them according to agreed salaries and workloads. (3) **Intended Use.** MEUR can be utilized to develop more persuasive models in the field of multimodal event understanding and reasoning. Researchers can also inherit our dataset design, especially the semi-automatic collection pipeline, to develop their own datasets. (4) **Controlling Potential Risks.** Since the documents of MEUR do not contain private information and the annotation process is not necessary to make many judgments about social risks, we believe MEUR does not introduce any additional risks. We manually verified some randomly sampled data to ensure the dataset did not contain risky issues.

Acknowledgments

We thank anonymous reviewers for their valuable feedback. This research is supported by the Postgraduate Research Scholarship (FOSA2212008), the Research Development Funds (RDF-21-01-069 and RDF-21-02-044), the Collaborative Research Project (RDS10120240248) at Xi'an Jiaotong-Liverpool University, and the Suzhou Industrial Park Interdisciplinary Innovation (Research) Platform for Affective Computing and Interactive Health (CXK2025101).

6. Bibliographical References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-vl technical report.

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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang,

[4] https://tlo.mit.edu/

[5] https://www.sigkg.cn/ccks2021/en/index.php/resource-track/

[6] https://github.com/zimuwangnlp/MEUR

Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card.

Xiaoqiang Kang, Zimu Wang, Xiaobo Jin, Wei Wang, Kaizhu Huang, and Qiufeng Wang. 2025a. Template-driven llm-paraphrased framework for tabular math word problem generation. *Proceedings of AAAI*, 39(23):24303–24311.

Xiaoqiang Kang, Zimu Wang, Xiaochen Zi, Xiaobo Jin, Kaizhu Huang, Fei Yin, and Qiufeng Wang. 2026. A benchmark and method for photographed table reasoning. *Pattern Recognition*, 178:113355.

Xiaoqiang Kang, Shengen Wu, Zimu Wang, Yilin Liu, Xiaobo Jin, Kaizhu Huang, Wei Wang, Yutao Yue, Xiaowei Huang, and Qiufeng Wang. 2025b. Can grpo boost complex multimodal table understanding?

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2025. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Comput. Surv.*, 57(8).

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 ACL*, pages 2557–2568.

Ruochen Li, Zimu Wang, and Xinya Du. 2025a. Efficient document-level event relation extraction. In *Proceedings of RepL4NLP*, pages 92–99.

Ruosen Li, Ziming Luo, and Xinya Du. 2025b. Fg-prm: Fine-grained hallucination detection and mitigation in language model mathematical reasoning.

Li Lin, Yixin Cao, Lifu Huang, Shu'Ang Li, Xuming Hu, Lijie Wen, and Jianmin Wang. 2022. What makes the story forward? inferring commonsense explanations as prompts for future event generation. In *Proceedings of SIGIR*, page 1098–1109.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models.

Kang Liu, Yubo Chen, Jian Liu, Xinyu Zuo, and Jun Zhao. 2020. Extracting events and their relations from texts: A survey on recent research progress and challenges. *AI Open*, 1:22–39.

Yubo Ma, Zehao Wang, Mukai Li, Yixin Cao, Meiqi Chen, Xinze Li, Wenqi Sun, Kunquan Deng, Kun Wang, Aixin Sun, and Jing Shao. 2022. MMEKG: Multi-modal event knowledge graph towards universal representation across modalities. In *Proceedings of ACL: System Demonstrations*, pages 231–239.

Siru Ouyang, Zhuosheng Zhang, and Hai Zhao. 2024. Fact-driven logical reasoning for machine reading comprehension. *Proceedings of AAAI*, 38(17):18851–18859.

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

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 EMNLP: System Demonstrations*, pages 508–517.

Huan Rong, Zhongfeng Chen, Zhenyu Lu, Xiao ke Xu, Kai Huang, and Victor S. Sheng. 2025. Pred-id: Future event prediction based on event type schema mining by graph induction and deduction. *Information Fusion*, 117:102819.

Himanshu Sharma and Anand Singh Jalal. 2021. A survey of methods, datasets and evaluation metrics for visual question answering. *Image and Vision Computing*, 116:104327.

Zeno Vendler. 1957. Verbs and times. *The Philosophical Review*, 66(2):143–160.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024a. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning.

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

Xiang Wei, Yufeng Chen, Ning Cheng, Xingyu Cui, Jinan Xu, and Wenjuan Han. 2024. CollabKG: A learnable human-machine-cooperative information extraction toolkit for (event) knowledge graph construction. In *Proceedings of LREC-COLING*, pages 3490–3506.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of SIGIR*, page 641–649.

Fuyu Xing, Zimu Wang, Wei Wang, and Haiyang Zhang. 2025. Benchmarking and improving lvlms on event extraction from multimedia documents.

Changyu Zeng, Yifan Wang, Zimu Wang, Wei Wang, Zhengni Yang, Muyi Bao, Jiming Xiao, Anh Nguyen, and Yutao Yue. 2025. Numina: A natural understanding benchmark for multi-dimensional intelligence and numerical reasoning abilities.

Jiazheng Zhu, Shaojuan Wu, Xiaowang Zhang, Yuexian Hou, and Zhiyong Feng. 2023. Causal intervention for mitigating name bias in machine reading comprehension. In *Findings of ACL*, pages 12837–12852.

7. Language Resource References

Asma Ben Abacha, Vivek Datla, Sadid A. Hasan, Dina Demner-Fushman, and Henning Müller. 2020. Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain. In *Conference and Labs of the Evaluation Forum*.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of EventStory*, pages 77–86.

Tong Chen, Zimu Wang, Yiyi Miao, Haoran Luo, Yuanfei Sun, Wei Wang, Zhengyong Jiang, Procheta Sen, and Jionglong Su. 2025. Medfact: A large-scale chinese dataset for evidence-based medical fact-checking of llm responses.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of LREC*, pages 4545–4552.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of NAACL-HLT*, pages 2368–2378.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of LAW*, pages 95–104.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M Strassel. 2015. Overview of linguistic resources for the TAC KBP 2015 evaluations: Methodologies and results. In *TAC*.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M Strassel. 2016. Overview of Linguistic Resources for the TAC KBP 2016 Evaluations: Methodologies and Results. In *TAC*.

Joe Ellis, Jeremy Getman, and Stephanie M Strassel. 2014. Overview of linguistic resources for the TAC KBP 2014 evaluations: Planning, execution, and results. In *TAC*.

Jeremy Getman, Joe Ellis, Zhiyi Song, Jennifer Tracey, and Stephanie Strassel. 2017. Overview of linguistic resources for the tac kbp 2017 evaluations: Methodologies and results. In *TAC*.

Goran Glavaš, Jan Šnajder, Marie-Francine Moens, and Parisa Kordjamshidi. 2014. HiEve: A corpus for extracting event hierarchies from news stories. In *Proceedings of LREC*, pages 3678–3683.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of COLING*, pages 6609–6625.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of CVPR*, pages 6693–6702.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of CVPR*, pages 1988–1997.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Chunyang Li, Hao Peng, Xiaozhi Wang, Yunjia Qi, Lei Hou, Bin Xu, and Juanzi Li. 2024a. MAVEN-FACT: A large-scale event factuality detection dataset. In *Findings of EMNLP*, pages 11140–11158.

Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of ECCV*, pages 570–586.

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

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of NAACL-HLT*, pages 894–908.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, volume 27.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of CAtoCL*, pages 10–19.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of CNS*, pages 47–56.

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. In *Proceedings of ICSC 2007*, pages 517–526.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *Proceedings of ECCV*, pages 314–332.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of CVPR*, pages 4631–4640.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of SemEval*, pages 1–9.

Marc Verhagen, Robert J. Gaizauskas, Frank Schilder, Mark Hepple, Jessica L. Moszkowicz, and James Pustejovsky. 2009. The tempeval challenge: identifying temporal relations in text. *Language Resources and Evaluation*, 43:161–179.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of SemEval*, pages 57–62.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. *Linguistic Data Consortium*, 57.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of IJCAI-17*, pages 1290–1296.

Shuo Wang, Qiushuo Zheng, Zherong Su, Chongning Na, and Guilin Qi. 2021. Meed: A multimodal event extraction dataset. In *Proceedings of CCKS*, pages 288–294.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of EMNLP*, pages 926–941.

Xiaozhi Wang, Hao Peng, Yong Guan, Kaisheng Zeng, Jianhui Chen, Lei Hou, Xu Han, Yankai Lin, Zhiyuan Liu, Ruobing Xie, Jie Zhou, and Juanzi Li. 2024. MAVEN-ARG: Completing the puzzle of all-in-one event understanding dataset with event argument annotation. In *Proceedings of ACL*, pages 4072–4091.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of EMNLP*, pages 1652–1671.

Zhenguo Yang, Jiale Xiang, Jiuxiang You, Qing Li, and Wenyin Liu. 2023. Event-oriented visual question answering: The e-vqa dataset and benchmark. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10210–10223.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of EMNLP*, pages 2369–2380.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. LEVEN: A large-scale Chinese legal event detection dataset. In *Findings of ACL*, pages 183–201.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of CVPR*, pages 4995–5004.

A. Event Schema

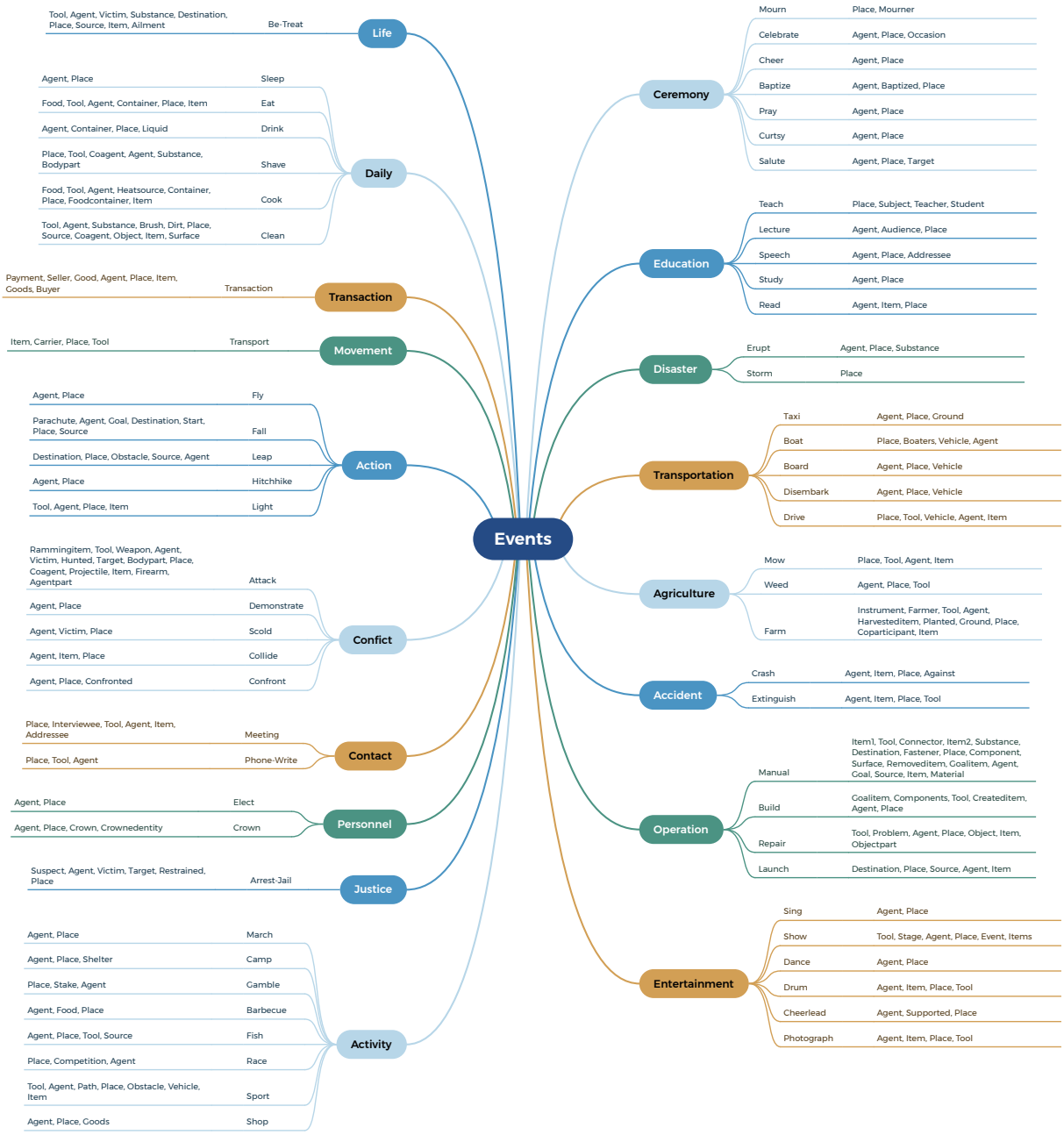


Figure 8: Event schema (i.e., event types and argument roles) of the MEED dataset.