MEEL: Multi-Modal Event Evolution Learning

Zhengwei Tao[1,2] **Zhi Jin**[1,2*] **Junqiang Huang**[3] **Xiancai Chen**[1,2]
Xiaoying Bai[4*] **Yifan Zhang**[1,2] **Chongyang Tao**[5]
[1]Key Laboratory of High Confidence Software Technologies (PKU), MOE, China
[2]School of Computer Science, Peking University [3]Peking University
[4]Advanced Institute of Big Data [5]SKLSDE Lab, Beihang University
`{tttzw, xiancaich, yifanzhang}@stu.pku.edu.cn`, `baixy@aibd.ac.cn`
`zhijin@pku.edu.cn`, `chongyang@buaa.edu.cn`

Abstract

Multi-modal Event Reasoning (MMER) endeavors to endow machines with the ability to comprehend intricate event relations across diverse data modalities. MMER is fundamental and underlies a wide broad of applications. Despite extensive instruction fine-tuning, current multi-modal large language models still fall short in such ability. The disparity stems from that existing models are insufficient to capture underlying principles governing event evolution in various scenarios. In this paper, we introduce Multi-Modal Event Evolution Learning (MEEL) to enable the model to grasp the event evolution mechanism yielding advanced MMER ability. Specifically, we commence with the design of event diversification to gather seed events from a rich spectrum of scenarios. Subsequently, we employ ChatGPT to generate evolving graphs for these seed events. We propose an instruction encapsulation process that formulates the evolving graphs into instruction-tuning data, aligning the comprehension of event reasoning to humans. Finally, we observe that models trained in this way are still struggling to fully comprehend event evolution. In such a case, we propose the guiding discrimination strategy, in which models are trained to discriminate the improper evolution direction. We collect and curate a benchmark M-EV^2 for MMER. Extensive experiments on M-EV^2 validate the effectiveness of our approach, showcasing competitive performance in open-source multi-modal LLMs. Code and Dataset are available on https://github.com/TZWwww/MEEL.

1 Introduction

Events are instances or occurrences that are the fundamental semantic units. Events are not independent, and they are usually interconnected by the following relations: causality, temporality, and intention. Multi-modal Event Reasoning (MMER) is to comprehend these events and their relations in both visual and textual modalities, and finally pave a path to better understanding the true world. MMER is expected to serve as the underpinning for various multi-modal applications, including visual storytelling (Huang et al., 2016), visual event prediction (Huang et al., 2021), event-related VQA (Park et al., 2020), MM knowledge graph construction (Ma et al., 2022), and video generation (Li et al., 2018; Liu et al., 2024). Such intricate tasks require an understanding of the event evolution mechanism across diverse scenarios.

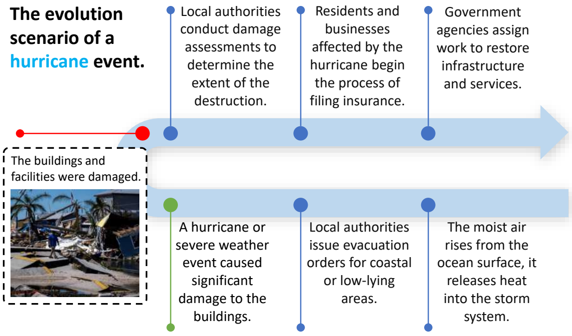


Figure 1: Part of the event evolution of a hurricane scenario. The queried event is in red. MEEL endows the model with the knowledge of all events in the scenario evolution. Current methods only train the model of few clips of event reasoning of the green one.

With the deepening of research on multi-modal instruction tuning, Multi-modal large language models (MLLM) have been able to handle various multi-modal tasks effectively (Liu et al., 2023; Zhu et al., 2023; Chen et al., 2023; Dai et al., 2023; Li et al., 2023b). These models master some abilities of MM event reasoning implicitly during training in diversified sorts of tasks. Among all the task categories, the perception tasks such as referring expression comprehension, referring expression generation, and grounded image captioning (Mao et al., 2016; Kazemzadeh et al., 2014; Peng et al., 2023) enable the model to comprehend the entities of the events in the image and text. The cognitive tasks, namely image caption and VQA (Lin et al.,

---

[*]*Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2024, pages 8912–8925
August 11-16, 2024

2014; Goyal et al., 2017), endow the model with the semantic understanding capability of events. However, the models trained by these tasks are unable to perceive event evolution because of the *static* nature of all modality inputs. Existing visual instruction-tuning methods only consist of questions for few clips of the entire event scenario. As shown in Figure 1, current methods only model the queried events with the green event and ignore the rest of the scenario. They lack a vision of a broad spectrum of other events in the evolving context. Such contextual absence impedes models from learning abundant evolution knowledge resulting in poor performances in MMER.

To address this issue, we propose Multi-Modal Event Evolution Learning (MEEL) for endowing the model to understand the event evolution to enhance the ability of MMER, leading to improved performances on downstream tasks. Specifically, we first design the event scenario diversification to acquire various events from abundant scenarios. Then, we employ ChatGPT to generate the evolving graphs of these seed events. The aim is to use these graphs to train the model to understand the rich knowledge of the evolution of events. To accomplish this goal, we propose the instruction encapsulation process to adapt the evolving graphs into instruction-tuning data to train the model. In this way, the training allows the model to comprehend more event evolutional knowledge of the scenario leading to better performance of MMER. However, allowing the model to learn only the evolving graphs is insufficient. Without acknowledging the incorrect evolving events, the model would improperly forward the process, resulting in the hallucination of event reasoning. To mitigate, we perform the guiding discrimination. The model requires judging the incorrect evolution. We design various negative mining strategies to harvest incorrect events. Then, we train the model to classify the right event. We also adapt the guiding discrimination into instruction tuning. After obtaining all the data, we finetune the LLaVA (Liu et al., 2023) model after its stage-1 pre-taining with LoRA (Hu et al., 2021) to get our model.

To validate the effectiveness of MEEL, we curate a benchmark M-EV^2 for **M**ulti-modal **EV**aluation of **EV**ent reasoning. M-EV^2 is collected or curated from nine existing datasets covering visual storytelling (Huang et al., 2016), visual event prediction (Huang et al., 2021), and event-related VQA (Yeo et al., 2018; Zhang et al., 2021a). Overall, M-EV^2 is a challenging task demanding the model to be capable of reasoning for diverse inter-event relations, like causality, temporality, and intent. It consists of two reasoning paradigms: close and open reasoning. We conducted extensive experiments on M-EV^2 and compare MEEL against some strong MLLM baselines. Our results demonstrate that MEEL does enhance the MMER ability of the model yielding significant improvements in downstream tasks. We conclude our contributions as:

- We propose the Multi-Modal Event Evolution Learning (MEEL). It aims to train the model to comprehend the intricate event evolution of diversified scenarios. Our method may shed light on other MM event reasoning research.
- We further design the Guiding Discrimination to guide the evolution and mitigate the hallucinations of MMER.
- We collect and curate the M-EV^2 benchmark for MMER. M-EV^2 covers diversified inter-event relations. We conduct extensive experiments on M-EV^2 to test the effectiveness of our model. We achieve competitive performance among open-source MLLMs.

2 Multi-Modal Event Evolution Learning

We strive to enhance a multi-modal large language model's capability in multi-modal event reasoning (MMER) to boost performance on downstream tasks. The key is to enable the model to comprehend event evolution. As shown in Figure 1, current multi-modal SFT data only model the target events with the green event and ignore the rest of the scenario. They lack a vision of a broad spectrum of other events in the evolving context. Such contextual absence impedes models from learning abundant evolution knowledge resulting in poor performances in MMER. The intuitive motivation is to endow the model with the knowledge of all events in the whole scenario.

To do that, we propose Multi-Modal Event Evolution Learning (MEEL). We leverage ChatGPT to obtain the evolution graph via our Event Graph Evolution mechanism, which starts from diversified seed events. The evolving graphs contain the entire event semantics of a whole scenario. Then we transform the evolving graphs into instruction-tuning data to train our model. Note that instruction tuning is one of the feasible ways to learn the knowledge of event-evolving graphs. One can also leverage

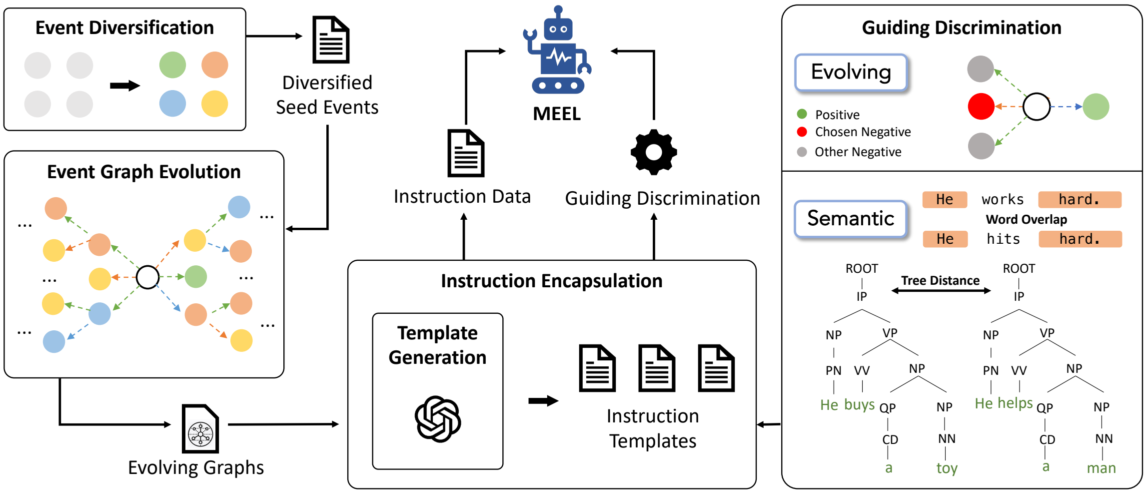


Figure 2: Overview of MEEL. We first implement the Event Diversification to harvest seed events. Then we perform the Event Graph Evolution to obtain the evolving graphs. We adapt the evolving graphs into instruction-tuning data through our Instruction Encapsulation. The Guiding Discrimination aims to improve the evolution learning with our two negative event mining strategies.

other methods such as in-context learning based on our data. We find that only when trained on the instruction-tuning data does the model turn to generate hallucinations. Therefore, we further add Guiding Discrimination loss to require the model to distinguish the correct events.

This section is organized as follows: Section 2.1 details the MMER task. The main purpose of MEEL is to enhance the comprehension of event evolution. We initiate with an event diversification step to generate a diverse mix of seed events of various scenarios (Section 2.2). Then we construct the event-evolving graphs through a novel method named event graph evolution (Section 2.3). Our next objective is to leverage these event-evolving graphs for model instruction tuning training (Section 2.4). Finally, we incorporate a guiding discrimination training strategy to refine evolution pathways and reduce reasoning errors (Section 2.5). MEEL's comprehensive framework is graphically represented in Figure 2.

2.1 Multi-Modal Event Reasoning

Multi-Modal Event Reasoning (MMER) involves deducing events based on certain inter-event relations across different modalities. Specifically, events as semantic units can be characterized by text, but their semantics are often more richly conveyed through associated images (Zhang et al., 2021b). The pursuit of MMER is to harness these multi-modal inputs to establish various relationships between events (temporal, causal, intentional, etc.), facilitating sophisticated reasoning processes (Tao et al., 2023b,a; Han et al., 2021). This reasoning underlies a spectrum of downstream tasks (Huang et al., 2016; Park et al., 2020).

We elaborate on the MMER formulation, wherein an event is expressed by a textual sentence $\mathcal{E}$ and represented by an image $\mathcal{I}$. Text provides argument structure, such as subject, verb, and object (Doddington et al., 2004), while images contextualize the event with environmental and situational details (Yang et al., 2023; Zellers et al., 2021). MMER can be modeled as inferring a target event $\mathcal{E}^t$ based on a given relation $\mathcal{R}$:

$$\mathcal{E}^t = \mathrm{M}\left(\mathcal{E}, \mathcal{I}, \mathcal{R}\right), \quad \mathcal{R} \in \mathbb{S}^{\mathcal{R}}. \tag{1}$$

Here, M denotes the model and $\mathbb{S}^{\mathcal{R}}$ represents the set of possible inter-event relations. For example, in Figure 1, $\mathcal{E}$ is the red event, $\mathcal{I}$ is the image, the queried relation $\mathcal{R}$ is "cause", the answer $\mathcal{E}^t$ is the green event. Therefore, the entire data is:

Question: *Given the image, what is the cause of "The buildings and facilities were damaged."*.

Answer: *A hurricane or severe weather event caused significant damage to the buildings.*

This question can not be answered only based on the $\mathcal{E}$ since there can be many reasons for building damage. Seeing the image, we can reason the damage could be caused by a hurricane. Models require analysis of both $\mathcal{E}$ and $\mathcal{I}$ to get the answer.

2.2 Event Diversification

Event diversification aims to curate a varied collection of seed events, encompassing multiple types and scenarios for ensuring evolutionary learning. We initiate this process with a corpus of text and

(a) Evolving Prompt

Consider the context that {caption}.Given a mentioned event: {event}, list events that each has a certain relation to the mentioned event. The generated events should be complex and complete which should at least have subject, verb. Leverage richer common sense knowledge to generate events. The generated events and the mentioned event can have different subjects. \nExamples: {examples} \nOutput:

(b) Instruction Generation & Generated Template

Give me 100 instructions. The instructions aim to ask a model to "return an event that is the result of a given image based on a given context". The generated instructions should be as rich as possible in syntax, semantics, and form, covering various task difficulties. Include the context in the generated instructions and mention it as [event]. Don't generate double quotation marks.

Considering the {event}, describe an event that unfolds as a consequence of the depicted image.

(c) Instruction Generation & Generated Template (MC)

Give me 50 instructions that aim to choose the most possible consequence of a given image from given choices based on a given context. The generated instructions should be as rich as possible in syntax, semantics, and form, covering various task difficulties. Include the context in the generated instructions and mention it as [event]. Don't generate double quotation marks.

Draw upon the {event} to interpret the image, then pick the consequence that emerges as the most plausible.

Figure 3: (a) Evolving prompt. The sentence in brown only exists if $\mathcal{E}$ is the seed event. In such a case, we add the caption of $\mathcal{I}$. (b) Instruction templates generation of `Result` relation and one example of generated template. (c) Multiple-choice Instruction templates generation of `Result` relation and one example of generated template. {caption} is the placeholder for the image caption. {event} and {examples} are for the event $\mathcal{E}$ and in-context examples.

image pairs $\{(\mathcal{E}_i, \mathcal{I}_i)\}$, where each pair jointly represents an event. We next extract the trigger words to represent the events. Trigger words are typically verbs that explicitly signify the event's occurrence (Doddington et al., 2004). We employ the Spacy tool[1] to identify the primary verb $\mathcal{V}_{\mathcal{E}_i}$ within each text $\mathcal{E}_i$ as the trigger.

Observing a long-tail distribution in trigger frequency, we only include K events per trigger to establish a balanced seed event set, denoted as $\mathbb{S}^{\mathcal{E}} = \{(\mathcal{E}_i, \mathcal{I}_i)\}$. The outcome of this event diversification step is more diversified event types and scenarios, thereby broadening our model's generalization capabilities and strengthening its understanding of varied contexts.

2.3 Event Graph Evolution

For the goal of enhancing the comprehension of event evolution, we utilize the seed events $\mathbb{S}^{\mathcal{E}}$ to construct event-evolving graphs through our designed event graph evolution methodology. Building on insights from prior work where LLMs like ChatGPT[2] have demonstrated proficiency in generating coherent event narratives (Gunjal and Durrett, 2023; Li et al., 2023e), we apply a breadth-first search (BFS) strategy using the ChatGPT to expand each seed event $(\mathcal{E}, \mathcal{I}) \in \mathbb{S}^{\mathcal{E}}$ both forward and backward in event happening time. We show the process of either direction in Algorithm 1.

Algorithm 1: Event Graph Evolution algorithm.

```
Input  : Seed event $\mathcal{E}$ and the caption $\mathcal{C}$, evolving
         relations $\mathbb{R}^E$, evolving steps $L$.
Output : Event-evolving graph $\mathbb{G}$.
$\mathbb{G}$.AddNode($\mathcal{E}$), $\tilde{\mathbb{E}} = [\mathcal{E}]$
for $i \leftarrow 1$ to $L$ do
    $\mathbb{N} = [\,]$
    for $\mathcal{E}_j$ in $\tilde{\mathbb{E}}$ do
        if $i == 1$ then
            $\{(\mathcal{E}_k, \mathcal{R}_k)\} =$
              Evolve($\mathcal{E}_j, \mathcal{C}$, SampleRel($\mathbb{R}^E, 2$))
        else
            $\{(\mathcal{E}_k, \mathcal{R}_k)\} =$
              Evolve($\mathcal{E}_j$, SampleRel($\mathbb{R}^E, 2$))
        end if
        for $\mathcal{E}_k, \mathcal{R}_k$ in SampleEvent($\{(\mathcal{E}_k, \mathcal{R}_k)\}, 2$)
          do
            $\mathbb{G}$.AddNode($\mathcal{E}_k$)
            $\mathbb{G}$.AddEdge($\mathcal{E}_j, \mathcal{R}_k, \mathcal{E}_k$)
            $\mathbb{N}$.Append($\mathcal{E}_k$)
        end for
    end for
    $\tilde{\mathbb{E}} = \mathbb{N}$
end for
return $\mathbb{G}$
```

We introduce the process of forward evolution. Starting from the seed event $\mathcal{E}$, we consider forward-oriented relations such as $\mathbb{R}^E = \{\texttt{Result}, \texttt{After}, \texttt{HasIntention}\}$[3]. For each iteration, we invoke the ChatGPT to produce events consistent with sampled relations from $\mathbb{R}^E$, as described in Equation 1. In the beginning, due to the bias of relying solely on textual events, we incorporate visual information of the seed event. Specifically, while evolving a seed event, we add its image caption to provide contextual details, promoting more accurate evolution. When evolving the intermediate events, we only use just their text. The prompt template for this evolution process is depicted in Figure 3(a). After L iterations, we acquire an event-evolving graph $\mathcal{G}$.

---

[1] https://spacy.io/

[2] https://openai.com/

[3] Relations are directed from the generated to the queried event, for instance, generating the `Result` for a given event. `HasIntention` implies the head event is intended by subjects in the tail event.

Besides, we also consider backward evolution. Our motivation for that is intuitive. We want the model to cognize event evolution in an complete timeline including both directions. Since we always start from an intermediate event in the timeline, we need to perform both forward and backward evolution. To do that, we consider evolving relations $\mathbb{R}^E = \{\texttt{Cause}, \texttt{Before}, \texttt{IsIntention}\}$ and remains the other steps the same.

After the both sides evolution, we denote the outputs as the event-evolving graph $\mathcal{G}$ which entails the rich evolution mechanism of the event scenario.

2.4 Instruction Encapsulation

To endow the knowledge of the evolving graphs $\mathcal{G}$ for model training, we turn to multi-modal instruction-tuning, a technique with proven efficiency in adapting models to human-like comprehension (Zhu et al., 2023; Sun et al., 2023; Li et al., 2023a; Liu et al., 2023; Li et al., 2023b; Dai et al., 2023). Our approach involves transforming the components of $\mathcal{G}$, represented as $\mathcal{G} = (\mathbb{V}, \mathbb{W})$ with nodes $\mathbb{V}$ and edges $\mathbb{W}$, into instruction-tuning data.

For each node $\mathcal{E}_i \in \mathbb{V}$, we aim to create a datum comprising the seed event $\mathcal{E}^s$, its associated image $\mathcal{I}$, the relation $\mathcal{R}_i$, and the event $\mathcal{E}_i$ [4]. However, directly inferring $\mathcal{R}_i$ between nodes $\mathcal{E}^s$ and $\mathcal{E}_i$ is not straightforward if they are non-adjacent. We address this by introducing induction rules that leverage the properties of inter-event relations, as detailed in Table 1. For example, in an evolving graph $\mathcal{G}$, there exists a path from the seed event $\mathcal{E}^s$ and another event $\mathcal{E}_2$: $\mathcal{E}^s\Rightarrow$[After]$\Rightarrow\mathcal{E}_1\Rightarrow$[Result]$\Rightarrow\mathcal{E}_2$. According to rule 1 in Table 1: (After)$\star$(Result)+(After)$\star$ infers Result, where $\star$ denotes there exists zero or more, + means there is at least one. We induce $\mathcal{E}^s\Rightarrow$[Result]$\Rightarrow\mathcal{E}_2$. By applying these rules, we derive the indirect relation $\mathcal{R}_i$.

Then we embed all the data with our instruction-tuning templates to form an instruction-tuning dataset. To avoid the laborious task of creating manual templates, we employ ChatGPT to generate diverse question templates for each relation type. With 100 templates from ChatGPT, the templates aim to reason about the tail event based on the provided visual and/or textual events in accordance with Equation 1. Considering the possible absence of textual input, we generate two variations for each of the $|\mathbb{S}^{\mathcal{R}}|$ relations: one with textual input and one without.

RULE	INDUCTION
`(After)⋆(Result)+(After)⋆`	`Result`
`(After)⋆(HasIntention)+( After)⋆`	`HasIntention`
`(After)+`	`After`
`(Before)⋆(Cause)+(Before)`	`Cause`
`(Before)⋆(IsIntention)+(Before)`	`IsIntention`
`(Before)+`	`Before`

Table 1: Relation induction rules. $\star$ denotes there exists zero or more. + means there is at least one.

GRAPH	NODE	TRAINSET	AVG INPUT TOKEN
3600	38.36	14,290	104.17

Table 2: Trainset statistics. GRAPH is the number of generated graphs. NODE stands for the average nodes in a generated graph. TRAINSET is the number of generated data. AVG INPUT TOKEN is the average number of tokens of the input instruction.

For any given data $(\mathcal{E}^s, \mathcal{I}, \mathcal{R}_i, \mathcal{E}_i)$, we randomly determine whether to include textual event information. We then match a suitable template to the relation type $\mathcal{R}_i$ and encapsulate all the items into our instruction-tuning dataset. An example of an encapsulated datum is illustrated in Figure 3(b).

2.5 Guiding Discrimination

To ensure accuracy during event graph evolution and guide the model away from generating erroneous events, we introduce a guiding discrimination training paradigm. This mechanism is pivotal in preventing the evolution process from producing hallucinations which is similar to DPO (Rafailov et al., 2023). In this paradigm, we task the model with identifying the correct event amongst a set of carefully selected negative events.

$$\mathcal{E}^t = \mathrm{M}\left(\mathcal{E}, \mathcal{I}, \mathcal{R}, \mathbb{D}\right), \quad \mathcal{R} \in \mathbb{S}^{\mathcal{R}}, \tag{2}$$

where $\mathbb{D}$ is the candidate set consisting of the correct event $\mathcal{E}^t$ and a few negative events.

The discrimination training is challenging to perform due to the sourcing of these negative events. For which we formulate two negative event acquisition strategies:

Semantic: This strategy requires model to discriminate the semantic similar events. To forge semantically similar negative events, we first compile a pool of all events of the generated graphs. For any positive event $\mathcal{E}$, utilizing Spacy for dependency parsing, we compute the tree edit distance and the word overlap rate between $\mathcal{E}$ and each event

---

[4] We also tried to keep the intermediate nodes between $\mathcal{E}^s$ and $\mathcal{E}^i$ into the training data but found poorer performances.

in this pool[5]. Filtering by the preset thresholds for these metrics, we select the top two events that are close to $\mathcal{E}$. This method sharpens the model's ability to distinguish between events with closely related linguistic structures.

Evolving: This strategy enhances the model's grasp on the directionality of event evolution. Leveraging the bidirectional nature of our event generation, namely forward and backward directions, we select two negative events from the opposite direction of the positive event's evolution. These negatives are particularly challenging as they maintain shared arguments within the same scenario but differ in their logical sequence. This practice further refines the model's reasoning skills for establishing the correct evolution path.

From the total four negative events generated through these strategies, we randomly select two of them. These, alongside the correct event, are then encapsulated into a multiple-choice format. We also create diverse multiple-choice question templates for each relation type via ChatGPT. An example of such a generation prompt and a corresponding template is presented in Figure 3(c).

2.6 Training

After acquiring both MMER and guiding discrimination dataset, we finetune the backbone by combining the MMER loss $\mathcal{L}^R$ (from Eq.1) and the guiding discrimination loss $\mathcal{L}^D$ (from Eq.2):

$$\begin{aligned}\mathcal{L}^R &= -\sum_{(\mathcal{E},\mathcal{I},\mathcal{R})} \log P(\mathcal{E}^t|\mathcal{E},\mathcal{I},\mathcal{R}),\\ \mathcal{L}^D &= -\sum_{(\mathcal{E},\mathcal{I},\mathcal{R},\mathbb{D})} \log P(\mathcal{E}^t|\mathcal{E},\mathcal{I},\mathcal{R},\mathbb{D}),\\ \mathcal{L} &= \mathcal{L}^R + \mathcal{L}^D\end{aligned} \tag{3}$$

3 Experiments

3.1 Construction of M-EV^2

To comprehensively evaluate the models' abilities of MMER on diversified inter-event relations, we collect and curate a benchmark M-EV^2. It incorporates nine test sets covering event-related visual question answering (`VCOPA`, `VisCa`, `VisualComet`), visual event prediction (`IgSEG`), and storytelling (`VIST`). M-EV^2 evaluates event relations of causality, temporality, and intent. Besides, M-EV^2 also covers two reasoning paradigms that are multiple-choice close reasoning tasks (CLOSE) and open reasoning without candidates (OPEN). We show the statistics of M-EV^2 in Table 7. We elaborate on the curation process as follows.

`VCOPA` This is the task of commonsense VQA (Yeo et al., 2018). Given an image $\mathcal{I}$ and two candidate options, the task is to select a more plausible cause or effect option. We also transform this dataset into an open reasoning task. We denote the original multiple-choice task as `VCOPA-C` and the transformed task as `VCOPA-O`.

`VisCa` This is a dataset of learning contextual causality from the visual and textual signals (Zhang et al., 2021a). The original task is formulated as that given two images as the context and two textual sentence descriptions, models need to determine if the former sentence causes the latter one. We transform it into our VQA task. We keep the image and first sentence and regard the second sentence as the label to generate. We retrieve one negative sentence by the ground truth and form it as a multiple-choice task. We also adapt the multiple-choice task into an open reasoning similar to `VCOPA-O`. We denote these two tasks as `VisCa-C` and `VisCa-O`.

`VisualComet` This is an open commonsense VQA task which is to answer situations before or after (Park et al., 2020). We also retrieve a negative answer to formulate it into a multiple-choice task. We denote these two tasks as `VC-O` and `VC-C`.

`IgSEG` This dataset aims to predict future events based on what has happened (Huang et al., 2021). Specifically, given a sequence of sentences in sequential order and the image of what will happen next, the models need to generate a sentence for this image. In addition, we also retrieve one negative event and form it as a multiple-choice task. We denote these two tasks as `IgSEG-O` and `IgSEG-C`.

`VIST` It's the storytelling task which is to generate the next story given the previous story in sentences and an image (Huang et al., 2016).

3.2 Baselines

We compare baselines as LLaVA-Lora (Hu et al., 2021), InstructBLIP (Dai et al., 2023), Otter (Awadalla et al., 2023), MiniGPT-4 (Zhu et al., 2023), MiniGPT-4-v2 (Chen et al., 2023). We show more details in Appendix B.

3.3 Implementation Settings

We use InstructBLIP (Dai et al., 2023) to generate the image captions for event graph evolution. We set the evolution steps as 3 and constructed 14,290 instruction-tuning data. Comprehensive statistics of the dataset are detailed in Table 2.

---

[5] https://github.com/timtadh/zhang-shasha

♣	`VCOPA-C`	`VisCa-C`	`VC-C`	`VCOPA-O`	`VisCa-O`	`VC-O`
	VQA					
InstructBLIP (Dai et al., 2023)	63.33	64.78	51.25	7.57 / 2.31 / 9.32	7.56 / 1.01 / 14.87	12.30 / **4.84** / 13.72
Otter (Li et al., 2023b)	57.27	55.97	45.10	11.78 / 1.35 / 17.12	10.29 / 0.51 / 10.51	7.96 / 3.18 / 9.13
LLaVA-Lora (Liu et al., 2023)	46.06	45.28	45.60	7.66 / 1.44 / 0.64	7.06 / 0.67 / 5.66	7.57 / 2.31 / 3.32
MiniGPT-4 (Zhu et al., 2023)	56.67	47.80	51.40	9.78 / 2.44 / 7.05	7.87 / 1.55 / 10.30	6.92 / 1.78 / 0.42
MiniGPT-4-v2 (Chen et al., 2023)	49.70	52.83	54.60	8.90 / 2.13 / 2.09	8.89 / 1.21 / 8.55	7.54 / 3.03 / 5.06
MEEL (Ours)	**66.06**	**72.33**	**68.10**	**19.18 / 2.92 / 26.02**	**19.16 / 3.40 / 29.58**	**16.28** / 3.99 / **22.93**

Table 3: Main results of VQA tasks. The bold number represents the highest score.

♣	`IgSEG-C`	`IgSEG-O`	`VIST`
	Prediction		StoryTelling
InstructBLIP	55.10	8.13/**2.63**/15.91	6.71/1.22/11.31
Otter	53.20	7.57/1.35/4.34	7.63/1.20/10.51
LLaVA-Lora	46.40	9.03/1.50/4.46	9.09/**3.03**/5.53
MiniGPT-4	49.90	8.72/1.54/3.24	8.66/1.67/9.64
MiniGPT-4-v2	51.30	8.69/1.45/3.73	8.95/1.68/10.44
MEEL (Ours)	**66.50**	**14.00**/1.41/**19.41**	**14.38**/1.44/**25.60**

Table 4: Main results of visual event prediction and storytelling. The bold numbers represent the best score.

For our model, we use LLaVA-v1.3 after the first pre-training stage as our backbone (Liu et al., 2023) and train with Lora (Hu et al., 2021). LLaVA-Lora-v1.3-7B is the most comparable baseline to our method since the only difference is the visual instruction-tuning dataset. We use deepspeed[6], zero-2 without CPU offloading. We set the batch size to 16 on 4×V100 GPUs.

In pilot experiments, we conducted tests with multiple input prompts for each model in order to identify the most effective prompts for evaluation. Despite variations in prompts, we observed only minimal fluctuations in the results. To ensure consistency and mitigate the other influences, we maintained uniformity by using the same prompt for all models performing a task. Detailed prompts can be located in the Appendix E. For the multiple-choice tasks, we transformed them into multiple-choice questions and instructed the model to respond with the corresponding label of choice. For Close tasks, we design an answer decoding strategy and show in Algorithm 2. We find this strategy can handle almost all situations. For Close tasks, we employ accuracy as the metric. For Open tasks we utilize BLEU-1/2 (Papineni et al., 2002) and BERT-SCORE (Zhang et al., 2019) as measures.

3.4 Main Results

We test our model on M-EV^2 benchmark. We show the VQA results in Table 3, visual event prediction and visual storytelling in Table 4. We calculate the various kinds of average scores in Table 5.

♣	VQA	Pred	Story	Open	Close	All
InstructBLIP	33.01	35.50	11.31	12.53	54.11	25.16
Otter	28.40	28.77	10.51	9.66	49.06	21.64
LLaVA-Lora	23.92	25.43	5.53	4.64	45.85	17.17
MiniGPT-4-v2	26.49	26.57	9.64	6.44	51.30	20.08
MiniGPT-4	28.86	27.51	10.44	7.84	53.11	21.60
MEEL (Ours)	**45.53**	**37.95**	**25.60**	**23.06**	**67.64**	**36.61**

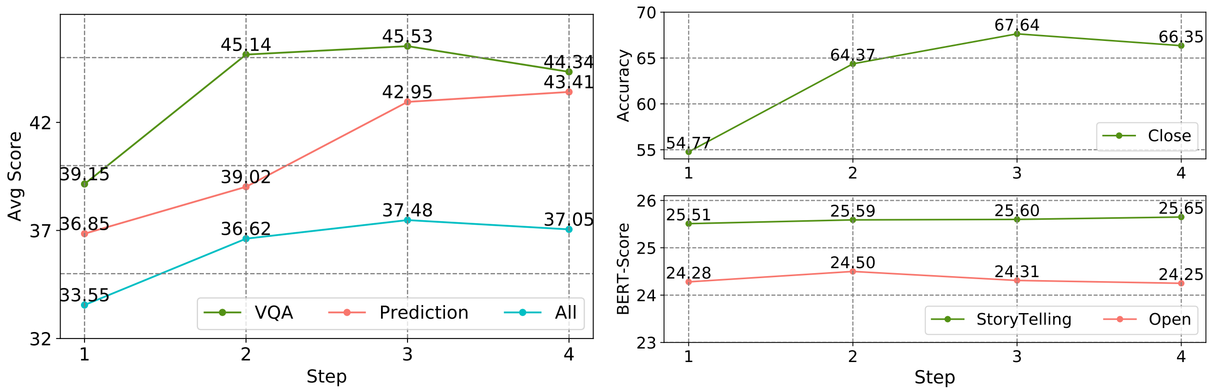
Table 5: Various kinds of average results. The bold numbers represent the best score. Pred stands for visual event prediction. Story is visual story telling. Close and Open are close and open reasoning tasks respectively. All is the average performance on all test set.

MEEL can effectively enhance performances of VQA. MEEL achieves the highest scores on three Close VQA, namely `VCOPA-C`, `VisCa-C`, and `VC-C` in Tabel 3. The results indicate MEEL can distinguish the right events since the improvements from event graph evolution with guiding discrimination. For the three Open VQA datasets, among all metrics, BERT-SCORE can mostly evaluate the answering quality. We find MEEL outperforms all other baselines to a large extent. These results demonstrate the effectiveness of our method on Open VQA. We also notice the BLEU-1/2 of MEEL is higher than almost all models. Since BLEU-1/2 measures lexical similarity, MEEL can generate more well-formed events as the ground truth. In all, our method improves the MMER.

MEEL outperforms baselines in visual event prediction. MEEL performs the best among all baselines in Table 4. The results demonstrate our training method enables the model to capture correct temporal relations leading to more precise prediction for the future. Compared to VQA tasks, We find all models perform worse in visual event prediction, indicating it needs more knowledge and reasoning ability to complete this task. In Open visual prediction, MEEL also achieves the highest scores in BERT-SCORE. This shows our model can forecast semantic similar events. However, we find MEEL performs slightly lower in BLEU-2

---

[6] https://www.deepspeed.ai/



(a) Average scores on VQA, PREDICTION, and all results. (b) Average scores on STORYTELLING, CLOSE, OPEN tasks.

Figure 4: Analysis of steps of event graph evolution.

♣	VCOPA-O	VisCa-O	VC-O	IGSEG-O	VIST
MEEL w.o. D	19.63	21.78	21.79	18.83	24.67
MEEL	26.02	29.58	22.93	19.41	25.60

Table 6: Ablation study. MEEL w.o. D is our method without guiding discrimination.

on `IgSEG-O`. Since BLEU calculates the 2-gram lexical similarity, this may indicate MEEL can predict more diversified events with correct semantics rather than words merely in the context.

MEEL can generate advanced story. In Table 4, we find MEEL can excel all baselines in VIST. The results show MEEL can tell better stories by capturing more scenario knowledge and comprehending the inter-event relations. The event graph evolution affects the training of the model to acknowledge enriched event information rather than merely shallow step reasoning.

In all, MEEL can significantly improve the performance of the downstream tasks attributed to boosted capabilities of MMER. In Table 5, MEEL excels all baselines on the average score of all datasets demonstrating the effectiveness of our method. Our event graph evolution process stimulates the contextual understanding of events. The guiding discrimination further mitigates the hallucinations of event reasoning yielding better performances.

Among all relation types, the improvements of VQA and STORYTELLING are larger than PREDICTION. It indicates our method benefits more for these tasks. PREDICTION is the hardest to learn attributed to its demand for pertaining for more abundant knowledge of events.

3.5 Analysis

Evolution steps. We conduct experiments on different evolution steps to verify the effectiveness of event graph evolution. We tested steps 1-4 respectively and calculated various average scores. We show the results in Figure 4.

As the average of all results, the performance of MEEL increases from steps 1 to 3 in Figure 4 (a). This is consistent with our motivation that the event graph evolution enables the model to learn the rich knowledge of event evolution. Then, the model can complete MMER better.

We find the performances drop when the step is too large, namely larger than four. This may be attributed to the semantic drift of the event graph evolution. ChatGPT would generate less relevant content compared to the seed event if it evolves further. We find that the drop is most obvious in VQA, which may be probably due to VQA being the most strict relation among all event interrelations.

We find MEEL can achieve a high score for STORYTELLING when the evolution step is only one in Table 4 (b). MEEL is 25.51 BERT-Score while InstructBLIP is 11.31. As the number of steps increases, MEEL maintains a high score. This indicates that MEEL completes the STORYTELLING even on few evolution steps.

Effect of guiding discrimination. We ablate guiding discrimination and show the results in Table 6. We find that all performances drop if MEEL trains without guiding discrimination. It indicates that discrimination can guide the evolution and mitigate hallucinations.

Examples of event graph evolution. We showcase two examples of event graph evolution in Figure 5. We find our evolving graphs can sufficiently contain information and knowledge of event scenarios. With the aid of event-evolving graphs, MEEL learns more abundant event knowledge and relation inter-connections.

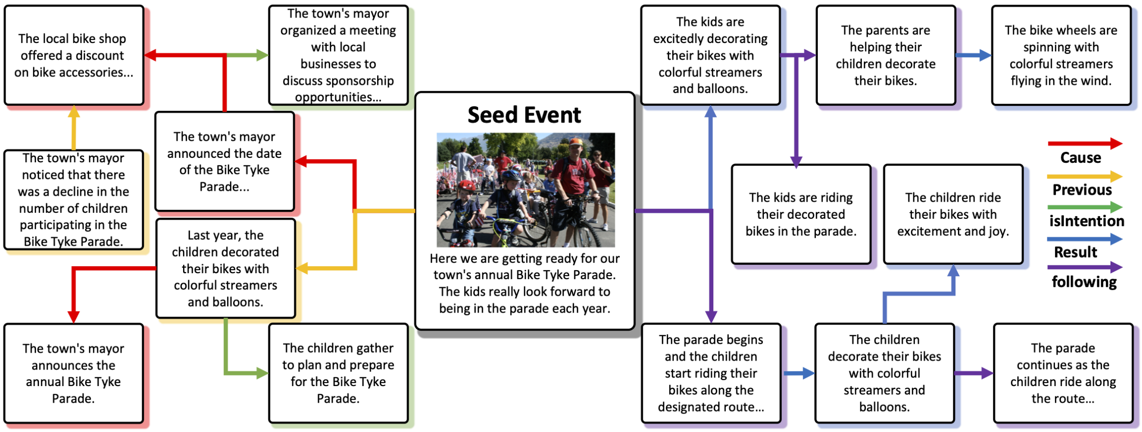


Figure 5: An example of an event-evolving graph. The event pointed to by the head cut is a tail event generated that satisfies the color relationship of the head cut.

4 Relation Works

Multi-Modal Event Relational Reasoning As one of the relation types, causality reasoning is crucial for exploring the cause and effect of events (Yeo et al., 2018; Zhang et al., 2021a; Chadha and Jain, 2021; Ignat et al., 2021). Apart from causality, event temporal reasoning forms a basic ability (Zellers et al., 2019; Park et al., 2020; Zellers et al., 2021). Event intentional reasoning uncovers the intentions of the subjects of the events (Park et al., 2020; Li et al., 2023c). Besides, there exists research on other relation types as well (Kim et al., 2022; Hessel et al., 2022). Multi-modal event relational reasoning constitutes a foundational capability for a range of downstream tasks in the realm of multi-modal reasoning. Our research endeavors to further enhance this crucial skill.

Multi-Modal Instruction tuning With the significant success of instruction tuning (Ouyang et al., 2022; Xu et al., 2023, 2024), current research has extended its capability to multi-modality. MM instruction tuning trains the model the follow instructions for questions about the images. Compared to textual instruction tuning, harvesting MM data with instructions is tougher. Zhu et al. (2023) trains MiniGPT-4 by further aligning pretrained EVA-CLIP (Fang et al., 2023) and Vicuna (Chiang et al., 2023). Liu et al. (2023) generate visual instruction data by requiring ChatGPT/GPT-4 with the given image and its caption. Dai et al. (2023) adapt human-labeled dataset into instruction data with pre-made templates. Li et al. (2023a) construct in-context learning data with instructions and use this dataset to train an MM LLM. These methods merely model shallow event evolving situations leading to poor ability of MMER.

Script Induction Script induction is to induce or generate chains or graphs of events representing the evolving mechanism. Du et al. (2022) induces 11 scripts of newsworthy scenarios from documents. Gunjal and Durrett (2023) attempt to generate event chains by querying large language models. Zhang et al. (2023) constructs scripts by designing interactions between humans and LLM. Li et al. (2023e) create event graphs in a pipeline operation with generation, ordering, and verification. In our work, we are the first to utilize the ability of script induction from ChatGPT to construct our MM event-oriented instruction-tuning data. We expect our work may shed light on other event-oriented approaches.

5 Conclusion

We propose the Multi-Modal Event Evolution Learning for MMER. We design the event graph evolution process based on the diversified seed events. We then encapsulate the evolving graphs into instruction-tuning data. We introduce the guiding discrimination training paradigm to further improve the learning of evolution. We conduct experiments on our collected and curated M-EV^2 benchmark for MMER. Results show the effectiveness of MEEL and it achieves competitive performance among open-source baselines.

Limitations

Our method is limited to MMER of a single image. However, a more complex MMER may contain several images to express a scenario. We leave the construction of methods and benchmarks of this complex MMER to future work.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grant No. 62192731.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Aman Chadha and Vinija Jain. 2021. ireason: Multimodal commonsense reasoning using videos and natural language with interpretability. *arXiv preprint arXiv:2107.10300*.

Jun Chen, Deyao Zhu1 Xiaoqian Shen1 Xiang Li, Zechun Liu2 Pengchuan Zhang, Raghuraman Krishnamoorthi2 Vikas Chandra2 Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: Large language model as a unified interface for vision-language multitask learning. *arXiv preprint arXiv:2310.09478*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Xinya Du, Zixuan Zhang, Sha Li, Pengfei Yu, Hongwei Wang, Tuan Lai, Xudong Lin, Ziqi Wang, Iris Liu, Ben Zhou, et al. 2022. Resin-11: Schema-guided event prediction for 11 newsworthy scenarios. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 54–63.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Anisha Gunjal and Greg Durrett. 2023. Drafting event schemas using language models. *arXiv preprint arXiv:2305.14847*.

Rujun Han, I-Hung Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, and Nanyun Peng. 2021. Ester: A machine reading comprehension dataset for reasoning about event semantic relations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7543–7559.

Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *European Conference on Computer Vision*, pages 558–575. Springer.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Qingbao Huang, Chuan Huang, Linzhang Mo, Jielong Wei, Yi Cai, Ho-fung Leung, and Qing Li. 2021. Igseg: Image-guided story ending generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3114–3123.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Oana Ignat, Santiago Castro, Hanwen Miao, Weiji Li, and Rada Mihalcea. 2021. Whyact: Identifying action reasons in lifestyle vlogs. *arXiv preprint arXiv:2109.02747*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Hyounghun Kim, Abhay Zala, and Mohit Bansal. 2022. Cosim: Commonsense reasoning for counterfactual scene imagination. *arXiv preprint arXiv:2207.03961*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023b. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. 2023c. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11963–11974.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023d. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Sha Li, Ruining Zhao, Manling Li, Heng Ji, Chris Callison-Burch, and Jiawei Han. 2023e. Opendomain hierarchical event schema induction by incremental prompting and verification. In *Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023)*.

Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. 2018. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*.

Yubo Ma, Zehao Wang, Mukai Li, Yixin Cao, Meiqi Chen, Xinze Li, Wenqi Sun, Kunquan Deng, Kun Wang, Aixin Sun, et al. 2022. Mmekg: multi-modal event knowledge graph towards universal representation across modalities. Association for Computational Linguistics.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 508–524. Springer.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

Zhengwei Tao, Zhi Jin, Xiaoying Bai, Haiyan Zhao, Yanlin Feng, Jia Li, and Wenpeng Hu. 2023a. Eveval: A comprehensive evaluation of event semantics for large language models. *arXiv preprint arXiv:2305.15268*.

Zhengwei Tao, Zhi Jin, Haiyan Zhao, Chengfeng Dou, Yongqiang Zhao, Tao Shen, and Chongyang Tao. 2023b. Unievent: Unified generative model with multi-dimensional prefix for zero-shot event-relational reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7088–7102.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Dexuan Xu, Yanyuan Chen, Jiayu Zhang, Yiwei Lou, Hanpin Wang, Jing He, and Yu Huang. 2023. Radiology report generation via structured knowledge-enhanced multi-modal attention and contrastive learning. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2320–2325. IEEE.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

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*.

Jinyoung Yeo, Gyeongbok Lee, Gengyu Wang, Seungtaek Choi, Hyunsouk Cho, Reinald Kim Amplayo, and Seung-won Hwang. 2018. Visual choice of plausible alternatives: An evaluation of image-based commonsense causal reasoning. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651.

Hongming Zhang, Yintong Huo, Xinran Zhao, Yangqiu Song, and Dan Roth. 2021a. Learning contextual causality between daily events from time-consecutive images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1752–1755.

Linhai Zhang, Deyu Zhou, Yulan He, and Zeng Yang. 2021b. Merl: Multimodal event representation learning in heterogeneous embedding spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14420–14427.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang, Isaac Tham, Zhaoyi Hou, Jiaxuan Ren, Liyang Zhou, Hainiu Xu, Li Zhang, Lara J Martin, Rotem Dror, Sha Li, et al. 2023. Human-in-the-loop schema induction. *arXiv preprint arXiv:2302.13048*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Appendix

A Baselines

We show statistics of M-EV^2 in Table 7.

B Baselines

LLaVA-Lora This is a MLLM trained on visual instruction-tuning. It's based on the visual encoder ViT-L/14-336px (Radford et al., 2021) and the textual chat LLM vicuna-v1.3-7b (Chiang et al., 2023). In the first pre-train stage, it is trained with image-text pairs. In the second stage, it is fine-tuned by LLM-generated instruction-tuning data with LoRA (Hu et al., 2021).

InstructBLIP It uses BLIP-2 (Li et al., 2023d) framework as its foundation, InstructBLIP strategically restructures 26 pre-trained public datasets, including image captioning and VQA, into a format conducive to instruction tuning (Dai et al., 2023).

Otter This model combines multi-modal in-context learning with multi-modal instruction tuning, building upon the foundation of Open-Flamingo (Awadalla et al., 2023). This involves updating the perceiver module and relevant components of the LLM throughout the training process. The instructional data is sourced from reputable datasets including VQAv2 (Antol et al., 2015), GQA (Hudson and Manning, 2019), LLaVA, as well as a proprietary video dataset not available to the public.

MiniGPT-4 This model conducts visual instruction tuning on the pre-trained BLIP-2 (Li et al., 2023d), specifically focusing on updating the linear layer (Zhu et al., 2023). The instructions primarily draw from the domain of image captioning tasks.

MiniGPT-4-v2 This model performs as a unified interface to complete various tasks such as VQA, visual grounding, and image caption (Chen et al., 2023). Different from MiniGPT-4, it adds task identifiers into the prompt to guide the task completion. The backbone of MiniGPT-4-v2 is LLama-2 (Touvron et al., 2023).

C Decoding Protocol

We show our decoding protocol for extracting answers of CLOSE tasks as in Algorithm 2.

D Effect of event diversification.

We compute the event verb distribution. We show two verb distributions with or without event diversification. The results are in Figure 6. We find the distribution is significantly diversified after the event diversification process. It enables MEEL to be trained in various event scenarios and domains.

Algorithm 2: CLOSE answer decoding.

```
Input  : Prediction $\mathcal{P}$, candidate set $\mathbb{D}$.
Output : Answer $\mathcal{A}$.
pattern =
  "the(?: correct)? (?:option|answer) is[\ s:]+([A-H])"
if $\mathcal{P}$.startsWithAlphabet() then
    $\mathcal{A}$ = starts_alphabet
else if re.match(pattern, $\mathcal{P}$) then
    $\mathcal{A}$ = re.extract($\mathcal{P}$, patten)
else
    $\mathcal{A}$=$\operatorname{argmax}_{c \in \mathbb{D}}$(WordOverlap($c$, $\mathcal{P}$))
return $\mathcal{A}$
```

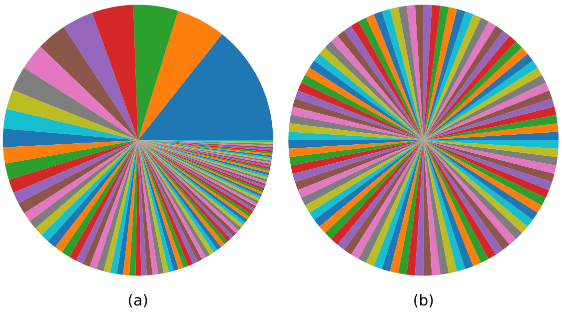

(a) (b)

Figure 6: Distribution of verbs before and after event diversification. Each part of the pie chart is the proportion of a verb. We present the 100 most frequent verbs with and without event diversification. (a) w.o. event diversification. (b) w.t. event diversification.

E Inference Prompts

We show inference prompts of all test set in Figure 7. We test various prompts in our pilot experiments and choose the prompts shown in Figure 7 which perform the best among others. We use the same prompts for all models.

	VCOPA-O	VCOPA-C	VisCa-O	VisCa-C	VC-O	VC-C	IgSEG-O	IgSEG-C	VIST
Number of tasks	330	330	282	159	2,000	2,000	1,000	1,000	4,379
Number of images	330	330	128	191	1,735	1,627	739	465	1,677
Relation types	C	C	C	C	T,I	T,I	T	T	T

Table 7: Statistics of M-EV^2. C, T, and I stand for Causal, Temporal, and Intentional inter-event relations. The number of tasks is not equal to the number of images resulting from duplicated images in these tasks.

VisualComet

temporal-open

Before:
From the picture, what happened before "{event}"?

After:
Form the picture, what happened after "{event}"?

temporal-close

Before:
Answer the question by returning A or B.

Question: From the picture, what happened before "{event}"?
Choices: {cs}
The answer is

After:
Answer the question by returning A or B.

Question: From the picture, what happened after "{event}"?
Choices: {cs}
The answer is

intentional-open

Input:
In the picture, {event}, what is the intent?

intentional-close

Input:
Answer the question by returning one of the choice from given Choices.

Question: What is the intent of the subject in "{event}"?
Choices: {cs}
The answer is

IgSEG

open

Before:
From the picture, what happened before "{event}"?

After:
Form the picture, what happened after "{event}"?

close

Before:
Answer the question by returning A or B.

Question: From the picture, what happened before "{event}"?
Choices: {cs}
The answer is

After:
Answer the question by returning A or B.

Question: From the picture, what happened after "{event}"?
Choices: {cs}
The answer is

VIST

open

Before:
From the picture, what happened before "{event}"?

After:
Form the picture, what happened after "{event}"?

VCOPA

open

Cause:
What caused "{event}"?

Effect:
What is the result of "{event}"?

close

Cause:
Answer the question by returning A or B.

Question: What is causes "{event}"?
Choices: {cs}
The answer is

Effect:
Answer the question by returning A or B.

Question: What is the result of "{event}"?
Choices: {cs}
The answer is

VisCa

open

Cause:
What caused "{event}"?

Effect:
What is the result of "{event}"?

close

Cause:
Answer the question by returning A or B.

Question: What is causes "{event}"?
Choices: {cs}
The answer is

Effect:
Answer the question by returning A or B.

Question: What is the result of "{event}"?
Choices: {cs}
The answer is

Figure 7: Inference prompts of all test set.