Recognizing Everything from All Modalities at Once: Grounded Multimodal Universal Information Extraction

Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li
National University of Singapore, Singapore Management University, Wuhan University

Abstract

In the field of information extraction (IE), tasks across a wide range of modalities and their combinations have been traditionally studied in isolation, leaving a gap in deeply recognizing and analyzing cross-modal information. To address this, this work for the first time introduces the concept of *grounded Multimodal Universal Information Extraction* (**MUIE**), providing a unified task framework to analyze any IE tasks over various modalities, along with their fine-grained groundings. To tackle MUIE, we tailor a novel multimodal large language model (MLLM), REAMO, capable of end-to-end extracting and grounding information from all modalities, i.e., 'recognizing everything from all modalities at once'. REAMO is updated via varied tuning strategies, equipping it with powerful capabilities for information recognition, fine-grained multimodal grounding, and in-depth reasoning. To address the absence of a suitable benchmark for evaluating grounded MUIE, we manually curate a high-quality, diverse, and challenging test set, which encompasses IE tasks across 9 common modality combinations. Meanwhile, we annotate the corresponding multimodal groundings, based on which we further formulate challenging cognitive questions&answers. The extensive comparison of REAMO with 11 existing MLLMs integrated into pipeline approaches demonstrates its significant advantages across all evaluation dimensions, establishing a strong benchmark for the follow-up research.

1 Introduction

IE is a pivotal topic, encompassing subtasks such as Named Entity Recognition (NER; Nadeau and Sekine, 2007), Relation Extraction (RE; Miwa and Bansal, 2016), and Event Extraction (EE; Ahn, 2006), which plays a crucial role in constructing domain-specific knowledge bases (Bosselut et al., 2019) and in facilitating deep semantic understanding of data (Satyapanich et al., 2020). In reality,

Figure 1: Examples of grounded multimodal universal information extraction.

a vast amount of information is conveyed through modalities beyond text. Consequently, research in IE has evolved from focusing solely on textual data to embracing various other modalities, leading to the development of multimodal IE (MIE; Liu et al., 2019), e.g., images, videos, and audio. Despite the growing research efforts dedicated to MIE, the exploration in this area remains insufficiently developed. We argue that several critical aspects must be fully considered for future MIE research trends.

Firstly, current studies primarily investigate MIE tasks within individual modalities (or certain modality combinations) (Sun et al., 2021; Chen et al., 2022b). With the existence of several modality categories and diverse definitions for different IE tasks, studying each modality separately to construct specialized MIE models would inevitably lead to resource wastage and inefficiency. In real-

world applications, there is a constant need for building unified systems with "one-for-all" robust generalizability for faster practical deployment. In light of the recent success of textual universal IE (UIE; Lu et al., 2022), MIE unification should also be promising. **Second**, the majority of existing studies (Zhang et al., 2017) exhibit a bias toward text-centric IE outputs, necessitating the decoding of detailed textual IE labels and inherently prioritizing text. While they often treat other modalities as auxiliaries and do not produce outputs for them, this practice does not align with reality, because all modalities can equally carry important information. For example, even infants who have not yet learned to speak can learn about entities from vision. Thus, each modality should be treated equally, and detect fine-grained information from all given modalities. **Last**, most current MIE research (Zheng et al., 2021) involving multiple modalities (e.g., Image&Text) tends to extract the modality-aligned part of the information under the assumption that different modalities associate with each other. However, in practical scenarios, the information carried by different modalities can be either shared (Li et al., 2022), or unrelated (Wu et al., 2023a). This suggests that information should be flexibly recognized from any modality sources.

In response to these challenges, this paper is dedicated to pioneering a novel task, grounded Multimodal Universal Information Extraction (**MUIE**). As illustrated in Fig. 1, MUIE aims to unify the modeling of various IE tasks (e.g., NER, RE, EE) with any (or combination) inputs across the most common modalities (e.g., text, audio, image, and video), and produce fine-grained multimodally grounded IE results. To solve MUIE, we consider taking advantage of the existing generative LLMs (OpenAI, 2022; Chung et al., 2022; Chiang et al., 2023) with in-context instructions (Dong et al., 2022). We develop an innovative multimodal LLM, REAMO, achieving "***R**ecognizing **E**verything from **A**ll **M**odalities at **O**nce*". REAMO not only outputs all possible textual IE labels but also identifies corresponding groundings across other modalities: 1) statically, by segmenting visual objects and audio speeches, and 2) dynamically, by tracking textual or vocal events in videos. REAMO is a fully end-to-end system, which outputs UIE label tokens as well as fine-grained groundings recurrently.

We then design a series of learning objectives to tune REAMO to endow it with robust MUIE and cross-modal grounding capabilities. First, we repurpose existing textual UIE annotation into instruction format, and use it to tune the backbone LLM for activating the UIE ability. Then, we perform both coarse-grained instance-level and fine-grained grounding-aware cross-modal alignment learning, enhancing the REAMO's capability in fine-grained multimodal semantic understanding. Furthermore, we instruction-tune REAMO on referring semantic segmentation tasks, to build its working behavior of recurrent generation. Finally, we strengthen its decision-making and more advanced cognitive level abilities with further instruction tuning, based on the reasoning-aware Chain-of-Thought (CoT; Wei et al., 2022) promoting.

In response to the absence of standard evaluation data for grounded MUIE, we further introduce a benchmark, where we annotate a high-quality test set of 3,000 instances covering NER, RE, and EE tasks under 9 common modality combinations. Besides, the data further advances by a) annotating both modality-shared/-specific content to simulate aligned and misaligned modality scenarios; and b) formulating cognition-level question-answer (QA) for each instance based on its grounding annotations. Extensive zero-shot experiments on these benchmarks demonstrate that REAMO shows much superior performance over existing 11 MLLMs with pipeline paradigm, with respect to IE tasks, multimodal grounding, QA abilities, and also stronger interpretability of predictions.

Overall, we make three key contributions:

- To our knowledge, this is the first to propose a grounded MUIE setting, unifying all IE tasks across modalities, further with fine-grained multimodally grounded targets.
- We introduce an MLLM for MUIE, REAMO, excelling in end-to-end MUIE prediction, and achieving cross-modal grounding of static objects and dynamic events.
- We contribute a high-quality, diverse, and challenging dataset, setting a benchmark for follow-up grounded MUIE research.

2 Related Works

IE (Lample et al., 2016; Wei et al., 2020) has long been a significant research direction, consistently attracting substantial interest and focus for decades. As the world contains various modalities of information, MIE has been consequently introduced, e.g., multimodal NER (Sun et al., 2021), multimodal RE (Chen et al., 2022b), and multimodal EE (Li et al., 2020). However, the majority of existing

well-established benchmarks for MIE still predominantly focus on texts, supplemented with images (Zhang et al., 2017; Wu et al., 2023a).

Historically, IE research has treated different tasks as separate studies for a long (Sun et al., 2021; Chen et al., 2022b). Recently, Lu et al. (2022) pioneer UIE, proposing to unify all IE tasks under a single generative model to produce all IE results, significantly reducing the maintenance cost for individual tasks. With the latest rapid development of LLMs, the latest advancements have utilized LLMs with in-context prompting for UIE, achieving promising zero-shot performance. Similarly, the swift progress of MLLMs should also ignite hope for MUIE. Yet research in MUIE remains under-explored.

To our best knowledge, the work most closely related to this paper is by Sun et al. (2023), who leverage existing MLLMs to unify various MIE tasks through a two-stage process of span extraction and classification in a multimodal QA format. Yet, we identify clear limitations in their approach that fall short of achieving comprehensive MIE unification. Firstly, the two-stage process simplifies the unification of IE, failing to support certain task settings, such as EE, which predicts both triggers and arguments, not just argument role classification. Our approach considers an end-to-end complete UIE prediction and proposes cross-modal grounding outputs. Secondly, while their work only considers text and image modalities, lacking comprehensiveness, we broadly cover the four most common modalities. Beyond addressing these limitations, we also introduce a novel MLLM tailored for grounded MUIE and contribute a new benchmark dataset for MUIE research. These efforts aim to pioneer the next stage of MUIE research.

3 Task Definition: Grounded Multimodal Universal Information Extraction

We now give a formal definition of grounded MUIE. Suppose the inputs are any of a text T, an image I, an audio A, a video V, or their combination.

NER task seeks to predict all possible textual labels of entities $\{E^{\text{ner}}\}$, with pre-defined entity types $C^{\text{ner}} \in \mathcal{C}^{\text{ner}}$ (e.g., person, location and organization), where each E may correspond to a span within T, or visual region within I, or a speech segment within A. We denote the visual grounding mask as M_{img} and the speech segment as M_{aud}.

RE task aims to first identify all possible entities $\{E^{\text{re}}\}$ following the NER step, and then determine a pre-defined relation label $R^{\text{RE}} \in \mathcal{R}^{\text{RE}}$ for two entities $< E_i^{\text{RE}}, E_j^{\text{RE}} >$ that should be paired. Also E^{RE} should correspond to T, I, or A, as in NER.

EE task detects all possible structured event records that consist of event trigger E^{et}, event type $C^{\text{et}} \in \mathcal{C}^{\text{et}}$, event argument E^{ea} and event argument role $C^{\text{er}} \in \mathcal{C}^{\text{er}}$. Here E^{et} and E^{ea} correspond to a continuous span within T or a speech segment within A. Also E^{ea} might refer to the visual region within I, or the temporal dynamic tracklet in video V (i.e., object tracking). We denote the video tracking mask as M_{vid}. $\mathcal{C}^{\text{et}}$ and $\mathcal{C}^{\text{er}}$ are pre-defined label sets.

Following Wang et al. (2023), we employ in-context learning (ICL) (Dong et al., 2022) to prompt LLMs for MUIE, with the specific task executed depending on the user's intention. In the bottom left of Fig. 2 we simply illustrate the ICL prompt. Appendix §A.3 presents prompts for different MUIE tasks in detail.

4 Our Proposed Methodology

4.1 REAMO MLLM Framework

Fig. 2 presents a schematic overview of REAMO MLLM, which consists of three main parts: multimodal encoder, LLM, and decoder for UIE prediction & multimodal grounding.

Multimodal Encoding. While REAMO takes four types of modality sources, except texts that are directly input to LLM, the image, audio, and video inputs should be encoded. Following Wu et al. (2023b), we leverage the high-performance ImageBind (Girdhar et al., 2023) as a unified multimodal encoder. Then, via a projection layer, different input representations are aligned into language-like embeddings that are understandable to the LLM.

LLM Reasoner. An LLM serves as the center unit of REAMO for content semantics understanding and reasoning. Specifically, we choose FlanT5 (Chung et al., 2022) due to its prominence in text understanding for NLP tasks. The choice is also made based on our pilot study, cf. §6.2.

MUIE Decoding with Grounding. Based on the input prompt, LLM will autoregressively produce textual responses, containing the UIE task results as required. To further generate fine-grained multimodal groundings, we consider integrating the existing high-performance SEEM model (Zou et al., 2023) for image segmentation and video tracking, and the SHAS model (Tsiamas et al., 2022) for audio segmentation. Technically, we

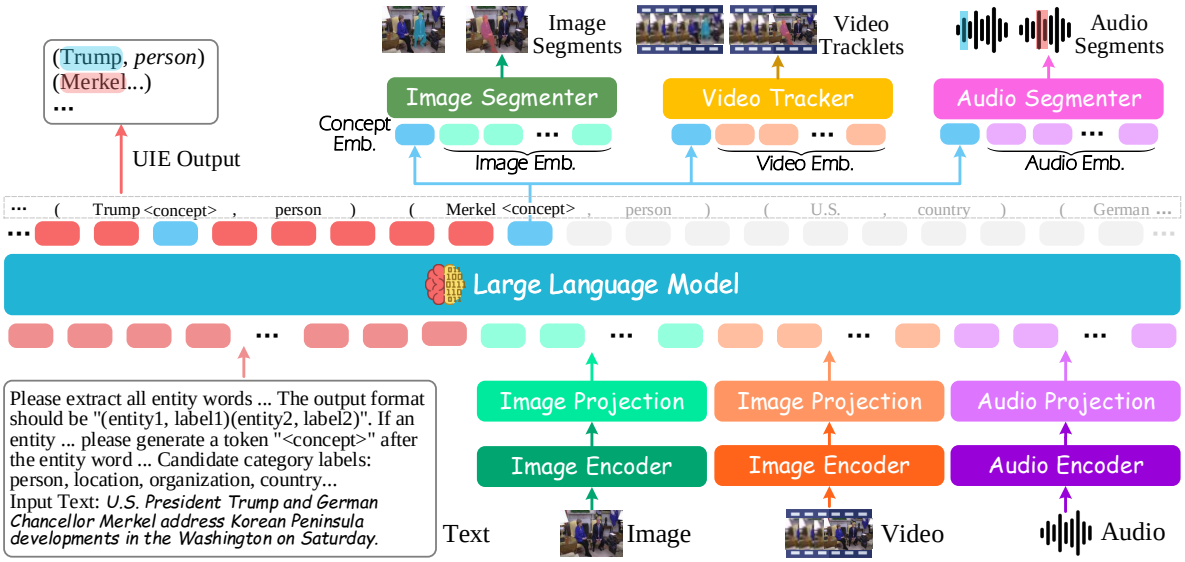

Figure 2: An overview of the proposed REAMO MLLM for MUIE.

design a special token after an entity/object token, i.e., '<concept>', as depicted in Fig. 2, which will be generated by LLM when it believes the entity should be grounded in other non-text modality(s). The embedding of the <concept> token $\boldsymbol{r}$ associated with each modality-specific embedding $\boldsymbol{H}_{img}, \boldsymbol{H}_{vid}, \boldsymbol{H}_{aud}$ will be passed to the corresponding modality segmenter(s), to activate them to generate grounding(s). We denote this step as:

$$M_{img} = \text{I-Segmenter}(\boldsymbol{r}, \boldsymbol{H}_{img}), \quad (1)$$

$$M_{vid} = \text{V-Trakcer}(\boldsymbol{r}, \boldsymbol{H}_{vid}), \quad (2)$$

$$M_{aud} = \text{A-Segmenter}(\boldsymbol{r}, \boldsymbol{H}_{aud}). \quad (3)$$

We extend more technical details of REAMO architecture in Appendix §A.

4.2 MUIE Fine-tuning for REAMO

With REAMO at hand, we now consider fine-tuning it through multiple objectives to enable REAMO with strong MUIE capability.

UIE Instruction Tuning. Our initial goal is to equip the system with the fundamental capability for UIE in the text modality. To achieve this, we consider tuning the backbone LLM specifically for UIE. Following the practices of Wang et al. (2023), we repurpose existing annotation data to form a set of instruction-tuning datasets for UIE. To avoid the huge cost of fully updating the LLM, we leverage the LoRA technique (Hu et al., 2022), achieving the goal by tuning only a small subset of parameters, without altering the overall LLM.

Multimodal Alignment Learning. REAMO integrates the ImageBind encoder to enable the LLM to comprehend basic multimodal signals. Following this, we proceed with multimodal alignment learning. We consider using the language-centric LLM as the core, requiring only the alignment of other modalities to text. We mainly utilize the vast array of available 'X-caption' pair data ('X' stands for image, audio, or video). We adopt an 'X-to-text' generation, where the input is 'X', and the LLM generates the corresponding caption. During this process, we fix the ImageBind and LLM while only updating the projection layer.

Fine-grained Cross-modal Grounding-aware Tuning. Above we merely enable REAMO with a coarse-grained multimodal understanding. Yet our goal is to attain a subtle modality comprehension. Thus, we further engage in fine-grained multimodal grounding. Our primary approach revolves around utilizing existing 'X-to-text' phrase grounding data, e.g., MS-COCO (Lin et al., 2014), where REAMO's input consists of textual phrases and grounded regional modality features, and then prompt LLM to determine their match.

Referring Semantic Segmentation Tuning. Next, we plan to cultivate the system's capability for autoregressive generation behavior. This step is crucial, especially since REAMO recurrently generates the entire MUIE results with groundings. Specifically, we consider the referring semantic segmentation tasks based on relevant datasets, e.g., MUSE (Ren et al., 2023) and RefCOCO (Kazemzadeh et al., 2014), which require the system to generate all textual tokens while simultaneously updating concept token embeddings with the whole decoders (i.e., SEEM and SHAS models). Through this learning objective, we can not only

Modality	NER	RE	EE
I	PASCAL-C	VRD	imSitu
V			VidSitu
A	ACE05-Aud	ReTACRED	
T+I	Twt17	MNRE	M^2E^2
T+V			VidSitu-Txt
T+A	ACE05-Aud	ReTACRED	
I+A		MNRE-Aud	
T+I+A	Twt17-Aud		
V+A			VidSitu-Aud

Table 1: Summary of the sources of our MUIE data.

strengthen the system's autoregressive generation capability but also jointly update the LLM core and various grounding decoders.

CoT-based Cognitive-aware Tuning. The above series of learning have already equipped REAMO with the capability for solid grounded MUIE. Yet we aim to go a step further, fully leveraging the LLM's powerful reasoning abilities. We primarily seek to achieve two goals: 1) to flexibly determine UIE label results based on the content of different input modalities, deciding which modalities the labels should be derived from; 2) to further link the grounded multimodal information and labels to the factual world, providing cognitive commonsense knowledge about the content being grounded. We leverage the CoT technique to prompt the LLM decision-maker for further deliberation and to provide the rationale behind its decisions.

Appendix §B provides more details about the various fine-tuning targets and datasets used.

5 A Benchmark for Grounded MUIE

To evaluate the performance of our grounded MUIE system, we develop a benchmark testing set. To begin with, we select 9 existing datasets from different modalities (or combinations thereof) for IE/MIE tasks. We then process these datasets, such as Text↔Speech, to create 6 new datasets under new multimodal (combination) scenarios. Before annotation, we carefully select 200 instances from their corresponding testing sets, ensuring each instance contained as much IE information as possible. Table 1 summarizes these datasets.

Firstly, we manually annotate the grounding information for MUIE across these 15 tasks and 10 modality combinations, including image segmentation, video tracking, and speech segmentation. This process results in a total of 3,000 high-quality grounded MUIE test instances. Next, we revisit the annotation information, specifically re-annotating instances from the combined modality datasets where cross-modal content is not fully aligned. This ensures that the dataset covers both modality-shared and modality-specific instances. Finally, based on the annotated multimodal grounding information and specific to the task at hand, we set a question for each instance. Answering this question correctly requires reliance on 1) precise grounding results and 2) relevant commonsense knowledge. We format it as multiple-choice QA, providing candidate answers for each question. Appendix §D extends more details about our datasets.

Method	Twt17	MNRE	M^2E^2		*Avg*
	NER	RE	ET	ER	
• Input with only Text only					
UIE (T5-L)	22.4	17.3	23.7	10.9	18.6
InstructUIE (FlanT5-XXL)	40.8	29.5	42.6	24.1	34.2
• Input with Text+Image					
MiniGPT4 (Vicuna 7B)	25.9	20.6	24.8	14.2	21.4
LLaVA (Vicuna 7B)	33.8	23.7	40.2	18.5	29.1
LLaVA (Vicuna 7B) w/o UIE-tune	20.5	18.2	15.9	7.4	15.5
InstructBLIP (Vicuna 7B)	27.3	22.5	30.4	15.0	23.8
InstructBLIP (FlanT5-XL)	39.4	27.3	39.3	23.5	32.4
InstructBLIP (FlanT5-XXL)	45.2	32.7	47.0	27.6	38.1
REAMO (Vicuna 7B)	29.0	21.4	28.9	13.0	23.1
REAMO (FlanT5-XXL)	**53.6**	**37.8**	**50.4**	**30.3**	**43.0**

Table 2: Preliminary experiment on text-image MIE tasks with LLMs in different types and sizes. Note that here we use three original whole-scale testing sets.

6 Experiments

6.1 Settings

We measure the system performance by following most practices of end-to-end UIE: F1 of entity span with type for NER, F1 of all subject&object entities and their relation label for RE. For EE, we consider event trigger (ET) F1 including both trigger and event type, and event argument F1 including both arguments with role types. W.r.t. the multimodal grounding, for both image and audio segmentation, we consider the mean Intersection over Union (mIoU); for video segmentation, we use the average Jaccard (J). For the QA, we use accuracy.

REAMO take the FlanT5-XXL (11B) as default LLM, as it shows to give the best performance. The encoder projection is a linear layer with a hidden size of 4,096. As no prior method is designed for grounded MUIE, we implement pipeline systems. Specifically, we first employ an MLLM to do UIE on multimodal input. And then we pass the raw multimodal source and the necessary UIE label to SEEM or SHAS model for fine-grained grounding of image, video and audio. We consider the existing 11 well-exposed MLLMs. For image-related UIE: MiniGPT4 (Zhu et al., 2023), InstructBLIP (Dai et al., 2023), LLaVA (Liu et al., 2023), and also

Method	T+I Input										I Input									
	Twt17			MNRE			M^2E^2				PASCAL-C			VRD			imSitu			
	NER	I-Seg	QA	RE	I-Seg	QA	ET	ER	I-Seg	QA	NER	I-Seg	QA	RE	I-Seg	QA	ET	ER	I-Seg	QA
• ***Pipeline***																				
LLaVA+SEEM	23.0	45.8	20.6	15.4	51.8	16.8	22.8	13.5	48.3	10.4	17.8	26.1	14.4	10.4	36.9	12.1	19.5	8.2	29.8	5.8
InstructBLIP+SEEM	45.4	48.7	27.0	22.4	56.2	24.3	43.0	20.1	52.5	16.7	41.8	32.0	23.6	18.6	38.3	17.8	37.0	13.3	32.2	11.2
• ***End-to-end***																				
MiniGPT-v2	26.9	52.0	35.7	17.0	54.8	30.5	24.3	16.3	54.8	21.0	21.0	39.7	28.8	11.6	39.1	24.0	18.5	11.3	30.0	15.3
MiniGPT-v2(+CoT)	29.4	57.8	40.4	20.8	59.5	36.8	26.0	19.2	53.0	24.8	24.4	44.8	36.3	14.0	44.9	32.9	22.8	12.7	34.8	20.6
REAMO	58.2	66.0	48.8	33.4	63.4	43.4	49.7	31.5	68.8	39.4	53.8	58.2	43.1	29.7	53.4	38.4	43.8	27.0	45.0	33.2
REAMO (+CoT)	**63.4**	**68.5**	**54.0**	**41.6**	**64.9**	**50.7**	**57.2**	**39.6**	**70.1**	**48.6**	**57.0**	**56.6**	**50.8**	**36.0**	**58.9**	**44.4**	**51.5**	**36.3**	**49.6**	**40.5**

Table 3: Zero-shot performance in the UIE scenario of text+image or standalone image input. I-Seg: grounding by image segmentation. Here InstructBLIP uses FlanT5-XXL. Performance of REAMO is with blue background.

Method	T+A Input						A Input					
	ACE05-Aud			ReTACRED			ACE05-Aud			ReTACRED		
	NER	A-Seg	QA	RE	A-Seg	QA	NER	A-Seg	QA	RE	A-Seg	QA
• ***Pipeline***												
NExT-GPT+SHAS	19.6	15.6	25.8	37.5	20.4	23.1	8.3	10.2	14.3	25.1	12.4	10.7
• ***End-to-end***												
SpeechGPT	26.7	21.4	32.4	45.4	27.5	29.5	14.0	13.3	22.4	30.4	21.0	19.5
SpeechGPT(+CoT)	28.4	23.5	35.8	48.7	30.0	36.4	15.4	14.8	30.5	30.8	23.4	24.0
REAMO	35.4	49.0	46.5	59.4	52.4	45.1	26.9	27.4	38.2	41.0	36.0	34.4
REAMO (+CoT)	**38.5**	**54.3**	**52.4**	**63.8**	**59.1**	**52.7**	**27.4**	**26.7**	**56.7**	**43.4**	**39.1**	**51.5**

Table 4: Zero-shot performance in the scenario of text+audio or standalone audio input.

Method	T+V (VidSitu-Txt)				V (VidSitu)			
	ET	ER	V-Trck	QA	ET	ER	V-Trck	QA
• ***Pipeline***								
VideoChat+SEEM	28.8	18.5	28.1	21.6	14.3	9.2	20.9	15.1
Vid-ChatGPT+SEEM	27.4	19.2	30.6	24.2	13.7	7.6	22.8	19.0
Video-LLaVA+SEEM	31.0	22.4	31.4	27.4	18.6	8.8	20.6	20.5
• ***End-to-end***								
REAMO	45.0	35.6	43.1	36.8	26.0	16.6	36.7	30.4
REAMO (+CoT)	**48.8**	**40.1**	**47.4**	**48.6**	**31.3**	**22.5**	**40.2**	**41.1**

Table 5: Zero-shot results in the scenario of text+video or standalone video input.

MiniGPT-v2 (Chen et al., 2023) that can output image segmentation end-to-end. For or video-related UIE: VideoChat (Li et al., 2023), Video-ChatGPT (Maaz et al., 2023), Video-LLaVA (Lin et al., 2023). SpeechGPT (Zhang et al., 2023a) for audio-related UIE. Video-LLaMA (Zhang et al., 2023b) supporting video+audio; PandaGPT (Su et al., 2023) and NExT-GPT (Wu et al., 2023b) supporting all four modalities. All these systems take the 7B LLM, unless otherwise specified. For fairness, all baselines are further tuned using the same UIE instruction-tuning data as ours. All system takes zero-shot inference, without tuning on in-house datasets. For comparison, we also show both the results of our REAMO with/without CoT prompting. More configurations are detailed in Appendix §C.

6.2 Pilot Study

Before conducting the formal experiments, we carry out a preliminary study to gather some basic observations. Table 2 shows the MIE results on several text-image tasks, where we also include the results with only text input by UIE (Lu et al., 2022) and InstructUIE (Wang et al., 2023). From the data, we can observe: 1) Incorporating multimodal input modeling adds supplementary information that benefits textual UIE. 2) Observing the results from the LLaVA group, it is evident that UIE tuning is crucial for MUIE performance. 3) The FlanT5 backbone demonstrates a stronger advantage for NLP-oriented IE tasks, compared to the Vicuna (Chiang et al., 2023) LLM backbone. 4) Larger models yield better results, leading us to select FlanT5-XXL as our backbone.

6.3 Zero-shot Results on Image-related MUIE

Table 3 presents the results of different models on our MUIE dataset under both text+image and pure image conditions. From the data, we observe: 1) Overall, end-to-end approaches outperform pipeline methods. 2) Although InstructBLIP is a pipeline model, it utilizes FlanT5-XXL, achieving relatively high results on pure IE tasks. However, its performance on subsequent multimodal grounding jobs does not match that of joint methods. 3) Significantly, end-to-end methods consistently perform better on QA tasks. This is because accurate grounding is essential for correct answers in QA. 4) Our system demonstrates a clear and consistent advantage overall. 5) Results equipped with CoT inference significantly surpass those with a straightforward prompting method.

Method	T+I+A (Twt17-Aud)				I+A (MNRE-Aud)				V+A (VidSitu-Aud)				
	NER	I-Seg	A-Seg	QA	RE	I-Seg	A-Seg	QA	ET	ER	V-Trck	A-Seg	QA
• ***Pipeline***													
Video-LLaMA+SEEM/+SHAS	-	-	-	-	-	-	-	-	12.0	4.8	12.7	8.4	17.0
PandaGPT+SEEM/+SHAS	27.3	25.3	10.1	32.7	10.2	41.2	13.0	21.4	20.4	11.8	17.5	12.4	24.8
NExT-GPT+SEEM/+SHAS	30.7	32.4	13.9	36.3	15.4	46.5	18.8	24.8	19.3	13.7	19.9	15.0	26.3
• ***End-to-end***													
REAMO	57.0	49.2	42.4	56.4	36.7	61.0	46.3	37.0	32.4	21.0	37.6	38.2	33.0
REAMO (+CoT)	**67.4**	**51.3**	**45.1**	**62.5**	**41.8**	**63.4**	**51.8**	**45.2**	**34.2**	**28.5**	**42.0**	**40.9**	**43.5**

Table 6: Zero-shot performance in more complex modality-hybrid scenarios of MUIE.

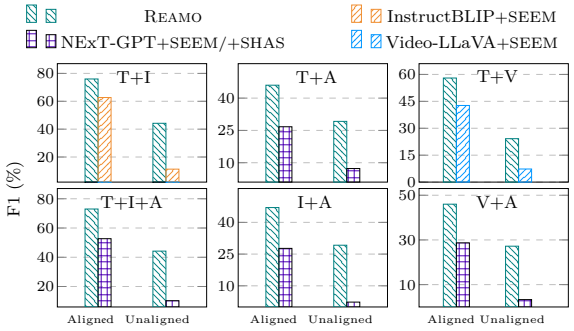

Figure 3: Performance gap between modality-shared (aligned) and modality-specific (unaligned) MUIE.

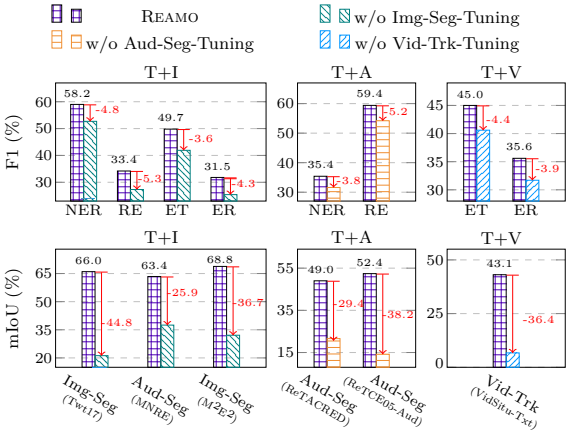

Figure 4: Removing grounding tuning of each modality.

6.4 Zero-shot Results on Audio-related MUIE

In Table 4, we present the performance of various models under text+audio and pure audio input, respectively. The trends observed are similar to those seen in the previous table: 1) End-to-end approaches demonstrate stronger performance compared to pipeline methods, effectively mitigating the issue of error propagation. 2) MLLMs that incorporate CoT exhibit significant improvements. 3) Our REAMO consistently outperforms others across all subtasks and scenarios.

6.5 Zero-shot Results on Video-related MUIE

In Table 5, we present the final set of results for EE task based on text+video and pure video input. The overall trend observed here again aligns with that of the previous tables, with our model achieving the strongest performance. Following, we can confirm

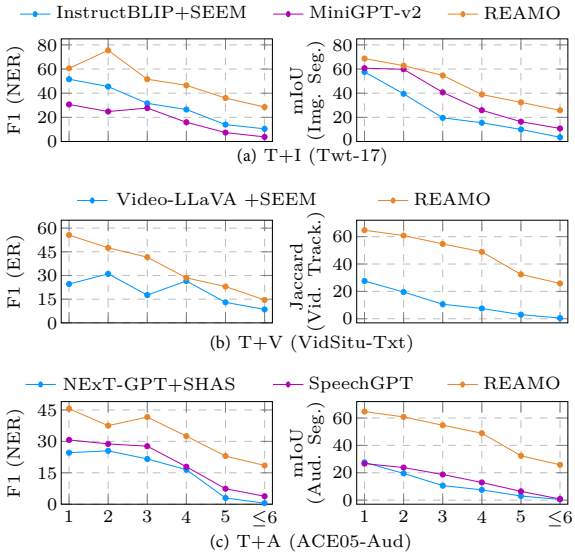

Figure 5: Impact of different speech

that for single modalities (such as image, audio, video), without the assistance of textual modality, the effectiveness of MUIE significantly diminishes.

6.6 Results on Modality-Compound MUIE

Finally, we present the performance of our model across a broader range of complex mixed-modality scenarios for MUIE. As shown in Table 6, our REAMO consistently delivers the best results. In comparison, all pipeline-based solutions that combine an LLM with an external grounding model exhibit significantly limited effectiveness. This discrepancy becomes even more pronounced in complex mixed-modality contexts.

6.7 Gap between Modality-Shared vs. Modality-Specific MUIE

Previously, we mentioned that in the MUIE context, different models may encounter both aligned shared information and misalignments between modalities. Now, we have categorized cases in the data as either Modality-Shared or Modality-Specific and compared the performance of various models under these two scenarios. As illustrated in Fig. 3, REAMO outperforms all baseline models in addressing challenges arising from misalignments between modalities.

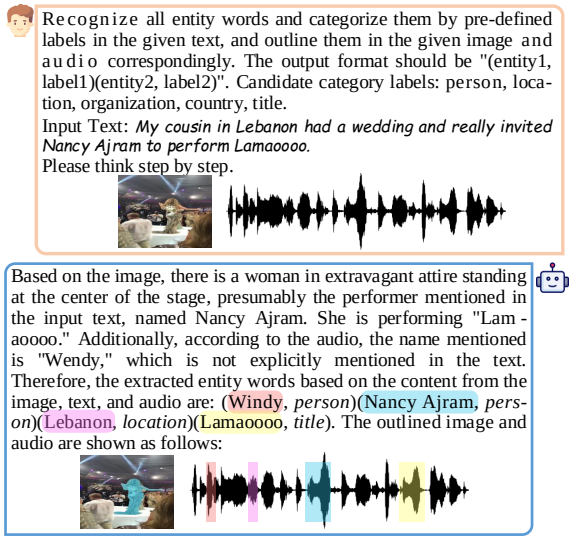

Figure 6: Qualitative result A on MUIE (NER) with modality-specific case via reasoning.

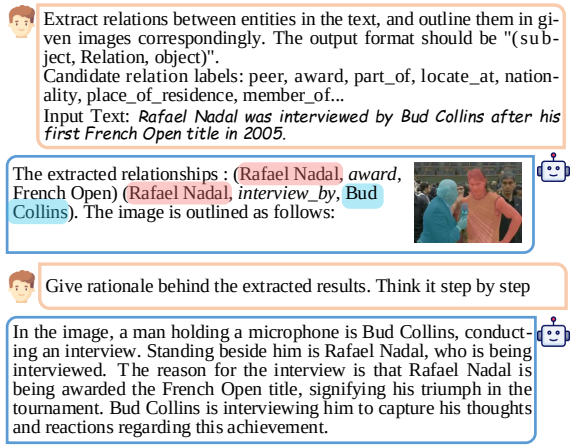

Figure 7: Qualitative result B on MUIE (RE) with grounding rationale via reasoning.

6.8 Impact of Grounding-aware Tuning

In this step, we study the impact of the entire REAMO system when all multimodal-related, grounding-aware tuning is removed (i.e., image/audio segmentation or video tracking). The results are plotted in Fig. 4. Overall, the absence of grounding tuning results in a minor loss to the MUIE extraction outcomes. However, unsurprisingly, it leads to a significant deterioration in the performance of fine-grained multimodal grounding analysis within the respective modalities.

6.9 Influence of Entity/Object Numbers

Fig. 5 further illustrates the impact of the number of entities (or objects) in an instance on the performance of the MUIE system. It is evident that our REAMO system maintains commendable performance in extracting an increasing number of objects across T+I/T+A/T+V scenarios, both in terms of MUIE results and fine-grained grounding,

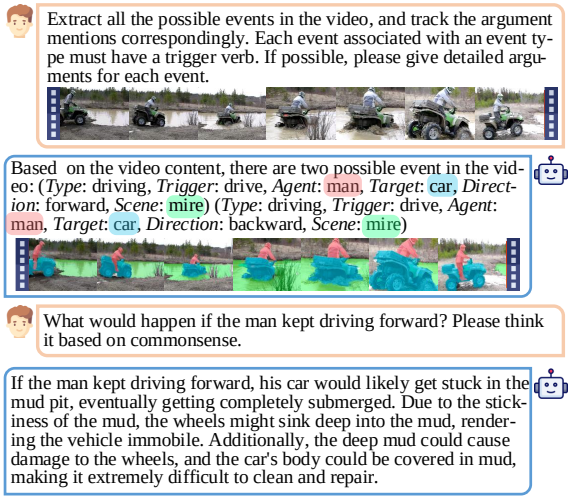

Figure 8: Qualitative result C on MUIE (EE) with commonsense-aware cognitive reasoning.

being clearly superior to baseline pipeline systems.

6.10 Case Study

Finally, we provide visualizations, through which we aim to offer a more intuitive and comprehensive demonstration of our MUIE system's enhanced capabilities, achieved via CoT-based reasoning. Fig. 6, 7, and 8 each display an example for NER, RE, and EE tasks, respectively. In case A, our system demonstrates how, through simple CoT prompting, it can engage in thought processes and flexibly determine MUIE labels from different modalities. In case B, the model is prompted to provide a rational basis for explainable decision-making. In case C, most impressively, it manages to link semantic information to factual commonsense understanding based on accurate fine-grained cross-modal grounding, and correctly answer the question.

7 Conclusion

This paper introduces a novel multimodal information extraction setting: grounded Multimodal Universal Information Extraction (MUIE). First, MUIE definition unifies all IE tasks across various modalities, including text, audio, images, and video, with fine-grained multimodally grounded targets. To solve MUIE, we devise REAMO, a novel MLLM that can extract and ground information from all modalities end-to-end. REAMO is tuned through various strategies to achieve proficiency in recognizing and grounding multimodal information. Further we introduce a high-quality, diverse, and challenging benchmark dataset for evaluating MUIE systems. Experimental results demonstrate REAMO's superior performance in extracting and grounding, and reasoning, setting a strong benchmark for the following grounded MUIE research.

8 Limitations

This work's limitations possibly stem from the following two main aspects.

8.1 Model Perspective

In this paper, we introduce an MLLM designed specifically for Grounded Multimodal Universal Information Extraction (MUIE).

First, our model encompasses four modalities: text, images, video, and audio. It undergoes specialized fine-tuning for these modalities to achieve robust functionality. Expanding to accommodate additional modalities in the future would require substantial training efforts. However, fortunately, the four modalities we currently focus on are among the most common worldwide, and it's unlikely that new ones will need to be added in the short term.

Moreover, although we have developed capabilities for modality-specific multimodal information extraction, our model still faces challenges in accurately predicting more complex implicit information, necessitating further improvements in extracting such content.

Thirdly, despite integrating CoT techniques to enhance the model's reasoning abilities, it still struggles with complex cognitive-level problems, often providing incorrect interpretations or reasoning. Future research should aim to strengthen this aspect by developing more sophisticated reasoning techniques. Finally, our model currently only considers tracking in video for EE tasks, as events intuitively represent dynamic content that aligns with video dynamics. However, for NER and RE tasks involving video, the system should also support tracking entities within videos. Our instruction tuning phase was limited to training on event data for video tracking, an area that future work could explore further.

8.2 Data Perspective

On the other hand, we introduced a new dataset for grounded MUIE evaluation.

Our dataset currently consists only of a test set, which supports evaluation but does not allow models to undergo in-house training and learning. Our next steps will include significantly expanding the annotation volume to add a training set.

Besides, while our dataset covers nine common modalities or combinations thereof, many more combinations exist, each of which should include the three IE tasks: NER, RE, and EE. Future efforts could look into expanding the annotation of information for more modality combinations.

9 Ethics Statement

The development and application of MUIE systems, as explored in this paper, raise several potential ethical considerations or risks that should be properly treated to ensure responsible research and deployment. These considerations span privacy, fairness, transparency, and the potential for misuse, among others. By incorporating ethical considerations into the research, development, and deployment processes, we can ensure that MUIE systems are developed responsibly and benefit society as a whole.

Privacy and Consent. Multimodal data, including text, images, video, and audio, can contain sensitive information about individuals, such as identifiable features, personal preferences, and private activities. It is crucial to ensure that data used for training and evaluating MUIE systems are collected and processed with explicit consent from the subjects involved, adhering to relevant data protection laws and guidelines.

Bias and Fairness. Machine learning models, including MLLMs, are susceptible to biases present in their training data. These biases can lead to unfair or discriminatory outcomes when models are applied in real-world scenarios. Efforts must be made to identify, mitigate, and monitor biases within MUIE systems to ensure equitable treatment of all individuals, regardless of their background, ethnicity, gender, or other characteristics.

Transparency and Accountability. The complex nature of MUIE systems, particularly those utilizing large language models and advanced algorithms for multimodal information extraction, can result in a lack of transparency regarding how decisions are made. Researchers and developers should strive for explainable AI, providing clear documentation and rationale for the model's predictions to foster trust among users and stakeholders.

Misuse Potential. The capabilities of MUIE systems to extract and synthesize information from multiple modalities also present opportunities for misuse, such as creating misleading content, conducting surveillance without consent, or other malicious applications. It is thus imperative to establish guidelines and safeguards against the misuse of

MUIE technologies, including legal and ethical frameworks that govern their use.

Environmental Impact. The training and operation of large-scale MLLMs are resource-intensive, contributing to significant energy consumption and carbon emissions. Researchers should consider the environmental impact of developing MUIE systems, and adopting more efficient computing techniques and models to minimize their ecological footprint.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8.

Richard Arratia and Stephen Desalvo. 2016. Probabilistic divide-and-conquer: A new exact simulation method, with integer partitions as an example. *Comb. Probab. Comput.*, 25(3):324–351.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the ICCV*, pages 1708–1718.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

Boli Chen, Guangwei Xu, Xiaobin Wang, Pengjun Xie, Meishan Zhang, and Fei Huang. 2022a. AISHELL-NER: named entity recognition from chinese speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 8352–8356.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022b. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1607–1618.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. 2023. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 9887–9897.

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

Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. 2020. TAO: A large-scale benchmark for tracking any object. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, pages 436–454.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. *CoRR*, abs/2305.05665.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the ICLR*.

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.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the NAACL*, pages 119–132.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the NAACL*, pages 260–270.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429.

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

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *CoRR*, abs/2311.10122.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *CoRR*, abs/2304.08485.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph convolution for multimodal information extraction from visually rich documents. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 32–39.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 852–869. Springer.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *CoRR*, abs/2306.05424.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

OpenAI. 2022. Introducing chatgpt.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. 2023. Pixellm: Pixel reasoning with large multimodal model. *arXiv preprint arXiv:2312.02228*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III*, pages 148–163.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual semantic role labeling for video understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. CASIE: extracting cybersecurity event information from text. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8749–8757.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the ACL*, pages 2556–2565.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *CoRR*, abs/2305.16355.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI conference on artificial intelligence*, 15, pages 13860–13868.

Yuxuan Sun, Kai Zhang, and Yu Su. 2023. Multimodal question answering for unified information extraction. *arXiv preprint arXiv:2310.03017*.

Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2019. A hierarchical framework for relation extraction with reinforcement learning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7072–7079. AAAI Press.

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022. SHAS: approaching optimal segmentation for end-to-end speech translation. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 106–110.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2011. Ace 2005 multilingual training corpus. *LDC2006T06, Philadelphia, Penn.: Linguistic Data Consortium*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

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 the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1652–1671. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 24824–24837.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the ACL*, pages 1476–1488.

Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. 2023a. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023b. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Tongtong Wu, Guitao Wang, Jinming Zhao, Zhaoran Liu, Guilin Qi, Yuan-Fang Li, and Gholamreza Haffari. 2022. Towards relation extraction from speech. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10751–10762.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9777–9786.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5534–5542.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *CoRR*, abs/2305.11000.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. *CoRR*, abs/2306.02858.

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

Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021. Multimodal relation extraction with efficient graph alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5298–5306.

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

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. 2023. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*.

A Model Specification

A.1 LLM Reasoning

The input text, combined with the input image, or video, or audio, is then processed by the LLM to generate interleaved response y in an autoregressive way:

$$y = \text{LLM}(\boldsymbol{H}_{txt}, \{\boldsymbol{H}_{img}|\boldsymbol{H}_{vid}|\boldsymbol{H}_{aud}\}), \quad (4)$$

where $\boldsymbol{H}_{txt}$ is the text embedding of the input text prompt, and $\{\cdot|\cdot\}$ is the two individual features or their combination. To help understand this process, consider an example of the text prompt x_{txt} and an input video x_{vid}:

> ▶ **Text Prompt:**
> Extract all the possible events in the video, and track the argument mentions correspondingly. Each event associated with an event type must have a trigger verb. If possible, please give detailed arguments for each event.
> ▶ **Video Prompt:**
> 

Then, the output y contains L tokens:

> Based on the video content, there are two possible events in the video: (*Type*: driving, *Trigger*: drive, *Agent*: man <concept>, *Target*: car <concept>, *Direction*: forward, *Scene*: mire <concept>) (*Type*: driving, Trigger: drive, *Agent*: man <concept>, *Target*: car <concept>, *Direction*: backward, *Scene*: mire <concept>

The corresponding hidden embeddings (i.e. the output of the last layer of LLM) of '<concept>' are represented as $\boldsymbol{r}$, which are inputs to the modality-specific grounding module alongside video features for mask generation.

A.2 Detailed Decoding with Grounding

Mask Definition. In our method, we ground the fine-grained segments within each modality instead of the bounding boxes during the decoding process. We denote the grounded image segment as $\hat{M}_{img} \in \mathbb{R}^{w_{img} \times h_{img}}$ which is a binary mask of each pixel, where w_{img} and h_{img} are the shape of the image. Similarly, the video tracklets consist of n frame segments $\hat{M}_{vid} = \{M^0_{vid}, \cdots, M^n_{vid}\}$, where $M^n_{vid} \in \mathbb{R}^{w_{vid} \times h_{vid}}$, and w_{img} and h_{img} are the shape of the each frame. The audio segment is the binary sequence $\hat{M}_{aud} \in \mathbb{R}^b$, where b represents the number of frames of each audio.

Image Segmenter. For an input image x_{img}, the image encoder I extracts visual features $\boldsymbol{I}_{img} = \text{I-Encoder}(x_{img})$, which is then transformed into the language space via an image linear projection layer $\boldsymbol{H}_{img} = \text{I-Projection}(\boldsymbol{I}_{img})$.

The design of Image Segmenter, inspired by the SEEM (Zou et al., 2023), employs cross-attention and self-attention layers to update all embeddings.

$$\begin{aligned} \boldsymbol{F}_{img} &= \text{Cross-Attn}(\boldsymbol{r}, \boldsymbol{H}^l_{img}), \\ \boldsymbol{H}^{l+1}_{img} &= \text{Self-Attn}(\boldsymbol{F}_{img}), \end{aligned} \quad (5)$$

where $\boldsymbol{H}^0_{img}$ is initialized with $\boldsymbol{H}_{img}$. After running L layers, we upsample the image embedding and an MLP maps the output token to a dynamic linear classifier, which then computes the mask foreground probability at each image location.

Video Tracker. For encoding a video, we sample n video frames $\{x^0_{vid}, \cdots, x^n_{vid}\}$. Then, similar to image encoding and projecting, we employ another video encoder and projection to connect the video features to the LLM:

$$\begin{aligned} \boldsymbol{V}_{vid} &= \text{V-Encoder}(\{x^0_{vid}, \cdots, (x^n_{vid}\}), \\ \boldsymbol{H}_{vid} &= \text{V-projection}(\boldsymbol{V}_{vid}). \end{aligned} \quad (6)$$

Then, during video tracking, we apply the same image segmenter on each video frame. Finally, we integrate the segments of video frames to obtain the video tracklets.

Audio Segmenter. Following Imagebind, we first convert a 2-second audio sample at 16kHz into spectrograms $\{x^0_{aud}, \cdots, x^m_{aud}\}$ using 128 mel-spectrogram bins. Then an audio encoder and projection are employed to extract the audio features and convert such features into the language space:

$$\begin{aligned} \boldsymbol{A}_{aud} &= \text{A-Encoder}(\{x^0_{aud}, \cdots, x^m_{aud}\}), \\ \boldsymbol{H}_{aud} &= \text{A-projection}(\boldsymbol{A}_{aud}). \end{aligned} \quad (7)$$

During the decoding, we leverage an existing speech segmentation model as the audio segmenter, named SHAS (Tsiamas et al., 2022), which consists of a Transformer encoder to extract the audio representation, followed by a linear sigmoid layer that maps them to a sequence of binary probabilities. After obtaining the probabilities for each frame of the audio, we use the probabilistic Divide-and-Conquer (pDAC) segmentation algorithm (Arratia and Desalvo, 2016) to obtain the final audio segments.

A.3 Prompt Settings

Here, we show the prompts for each subtask (i.e., NER, RE, EE) in MUIE, where *xxxx* represents the input text placeholder:

► Prompt for Named Entity Recognition:
Please recognize all entity words and categorize them by pre-defined labels in the given text, and outline them in the given image or video or audio correspondingly. The output format should be "(entity1, label1)(entity2, label2)". If an entity possibly has a counterpart in the given image or video or audio, please generate a token "<concept>" after the entity word, for subsequent cross-modal grounding.
Candidate category labels: *person, location, organization, country...*
Input Text: *xxxx*.
Input Image/Video/Audio: .

► Prompt for Relation Extraction:
Please extract all relations between named entities, and outline them in the given image or video or audio correspondingly. The output format should be "(subject entity, relation, object entity)". If an entity possibly has a counterpart in the given image or video or audio, please generate a token "<concept>" after the entity word, for subsequent cross-modal grounding.
Candidate relation labels: *peer, award, part_of, locate_at, nationality, place_of_residence, member_of...*
Input Text: *xxxx*.
Input Image/Video/Audio: .

► Prompt for Event Extraction:
Extract all the possible events in the video, and track the argument mentions correspondingly. Each event associated with an event type must have a trigger verb. If possible, please give detailed arguments for each event.
Candidate event types: *Marry, Attack, Injure, Beborn, Meet, Transport, Start-position...*,
Candidate event argument types: *Agent, Target, Direction, Time, Place, Instrument, Organization, Duration...*
Input Text: *xxxx*.
Input Image/Video/Audio: .

B Specification of MUIE Fine-tuning

In this part, we detail the overall process of tuning our REAMO, including the datasets used and how the prompting format. Essentially, all training utilizes the hyperparameters specified in Table 7. We use the FlanT5-XXL (11B)[1] in default. For all the training of each stage, we use 8 A100 (80G) GPUs. Each training cost different hours according to the datasets used.

Config.	Value
Multimodal Cncoder	ImageBind
LLM-bakcbone	FlanT5-XXL (11B)
Image Segmentor	SEEM
Video Tracker	SEEM
Audio Segmentor	SHAS
Encoder projector	MLP (4,096-d)
Size of Concept Emb.	4,096-d
Size of Image Emb.	4,096-d
Size of Video Emb.	4,096-d
Size of Audio Emb.	1,024-d
Optimizer	AdamW
Base learning rate	3.0e-4
Weight decay	2.0e-5
Optimizer momentum	β_1, β_2=0.9,0.95
Learning rate schedule	WarmipDecayLR
Warmup iterations	100

Table 7: Summary of hyperparams during training.

B.1 UIE Instruction Tuning

In this step, we train only the core LLM. To avoid the significant cost associated with fully updating the LLM, we employ the LoRA technique, which allows us to achieve our objectives by tuning only a small subset of parameters, thus leaving the overall architecture of the LLM unchanged. The datasets we utilize are sourced from existing IE datasets. We list these datasets in Table 8. These data are converted into an instruction format, following the practices outlined by Wang et al. (2023).

Task	Data Source	Amount
NER	OntoNotes 5.0[2]	76,714
	CoNLL2003[3]	20,744
RE	NYT (Riedel et al., 2010)	56,196
	NYT11 HRL (Takanobu et al., 2019)	62648
EE	MAVEN (Wang et al., 2020)	49,873

Table 8: Datasets used for UIE tuning.

B.2 Multimodal Alignment Learning

To accomplish the alignment, we adopt an 'X-to-text' generation task trained on the 'X-caption' pair ('X' stands for image, audio, or video) data from existing corpus and benchmarks, i.e., given the representation of an 'X', to prompt the frozen LLM to

[1] https://huggingface.co/google/flan-t5-xxl

generate the corresponding text description. Specifically, we utilize three types of 'X-caption' pair data, including: 1) 'Video-caption' pair dataset: Webvid-2M (Bain et al., 2021), a large-scale dataset of short videos with textual description sourced from stock footage sites, 2) 'Image-caption' pair dataset: CC3M (Sharma et al., 2018), contains over 3 million images accompanied by diverse styles of natural-language descriptions, and 3) 'Audio-caption' pair dataset: AudioCaps (Kim et al., 2019), an extensive dataset of approximately 46k audio clips paired with human-written textual descriptions collected via crowdsourcing.

B.3 Fine-grained Cross-modal Grounding-aware Tuning

For this step, our focus is on achieving fine-grained, concept-level cross-modal alignment, primarily aligning the other three modalities to the textual modality. Our primary method involves utilizing existing 'X-to-text' phrase grounding datasets. REAMO's input comprises textual phrases and grounded regional modality features, prompting the LLM to determine their match. For fine-grained image-to-text alignment, we consider the MS-COCO dataset (Lin et al., 2014). For video-to-text alignment, we turn to the TAO dataset (Dave et al., 2020). For entity-level audio-to-text alignment, we primarily utilize the dataset for the Speech NER task (Chen et al., 2022a). Since there is a lack of dataset in English speech, we thus use TTS tools to generate the CoNLL 2003 textual NER data into speech NER data.

B.4 Referring Semantic Segmentation Tuning

At this stage, we consider conducting a full-scale update of the entire REAMO model, including all subsequent grounding modules. The datasets used are as follows. For image segmentation, we consider two sets of data: MUSE (Ren et al., 2023) and RefCOCO (Kazemzadeh et al., 2014). In these datasets, each image's segmentation is accompanied by textual descriptions, which can serve as the textual generation targets for the model. For video tracking, the datasets we utilize include TAO (Dave et al., 2020) and SportsMOT (Cui et al., 2023). The textual descriptions in these datasets are constructed based on each object that needs to be tracked within the corresponding videos. For speech segmentation, we continue to use the speech NER data, CoNLL 2003, prompting the model to paraphrase sentences and decode audio segments.

B.5 CoT-based Cognitive-aware Tuning

This type of tuning mainly tries to achieve two goals.

First, we expect REAMO to flexibly determine UIE label results based on the content of different input modalities, deciding which modalities the labels should be derived from. Technically, we write the instructions by adding one additional question to the raw prompt to instruct MLLM to carefully think about the input data. Here, we can provide an example of NER:

> ▶ **Prompt for Reasoning Modal-specific MUIE:**
> Please recognize all entity words and categorize them by pre-defined labels in the given text, and outline them in the given image or video or audio correspondingly. The output format should be "(entity1, label1)(entity2, label2)". If an entity possibly has a counterpart in the given image or video or audio, please generate a token "<concept>" after the entity word, for subsequent cross-modal grounding.
> Candidate category labels: *person, location, organization, country...*
> *Please pay attention to identifying the content of each input modality and extract all entities from each modality, not just limiting entity recognition to the textual modality. This is because different modalities may contain unique, non-shared content, and each modality can contribute some modality-specific entities.*
> Please think step by step.
> **Input Text:** *xxxx*.
> **Input Image/Video/Audio:**    .

Second, we hope the system can further link the grounded multimodal information and labels to the factual world, providing cognitive commonsense knowledge about the content being grounded. We leverage the CoT technique to prompt the LLM decision-maker for further deliberation and to provide the rationale behind its decisions. For this, we devise a two-turn prompting template. The first step involves performing the standard Information Extraction (IE) task. The second step prompts the LLM to closely examine the input from various modalities and pay attention to the identified grounding information. Combining this with its commonsense knowledge, the model is then asked to reconsider: 1) whether the grounding results are reasonable, or 2) why certain outcomes are observed in the grounding. Again let's take an example of NER for the second step:

► **Prompt for Reasoning with Commonsense:**
Please carefully examine the information from all previously input modalities and, based on the identified grounding information combined with your commonsense knowledge, reconsider:
1) whether the grounding results are reasonable,
2) using your commonsense knowledge, why such results have occurred within the grounding.
Please think step by step.

C Extended Experimental Settings and Configurations

C.1 Baseline Specification

C.1.1 Baseline MLLMs

These MLLMs are adopted as our baselines.

Pure Textual LLM:

- **UIE**[4] (Lu et al., 2022), which takes T5 (Raffel et al., 2020) large (770 M) version as the core LLM.
- **InstructUIE**[5] (Wang et al., 2023), with FlanT5-XXL (11 B) as backbone LLM.

Image-aware MLLM:

- **MiniGPT4**[6] (Zhu et al., 2023), with Vicuna (7B) as backbone LLM.
- **InstructBLIP**[7] (Dai et al., 2023), with Vicuna (7B), FlanT5-XL (3 B), FlanT5-XXL (11 B) as backbone LLM, respectively. By default, we use the FlanT5-XXL (11 B).
- **LLaVA**[8] (Liu et al., 2023), with Vicuna (7B) as backbone LLM.
- **MiniGPT-v2**[9] (Chen et al., 2023), with Vicuna (7B) as backbone LLM. MiniGPT-v2 can output image segmentation end-to-end.

Video-aware MLLM:

- **VideoChat**[10] (Li et al., 2023), with Vicuna (7B) as backbone LLM.
- **Video-ChatGPT**[11] (Maaz et al., 2023), with Vicuna (7B) as backbone LLM.
- **Video-LLaVA**[12] (Lin et al., 2023), with Vicuna (7B) as backbone LLM.

Audio-aware MLLM:

- **SpeechGPT**[13] (Zhang et al., 2023a), with Vicuna (7B) as backbone LLM.

Full-modality-aware MLLM:

- **Video-LLaMA**[14] (Zhang et al., 2023b), with Vicuna (7B) as backbone LLM.
- **PandaGPT**[15] (Su et al., 2023), with Vicuna (7B) as backbone LLM.
- **NExT-GPT**[16] (Wu et al., 2023b), with Vicuna (7B) as backbone LLM.

C.1.2 Pipeline Implementation

Since there are currently no models specifically designed to support our grounded MUIE task, we have decided to implement some pipeline baselines for comparison. We mainly reproduce these models using their publicly available code and parameters. Here, we provide the detailed implementation of the pipeline baseline, which comprises two steps:

- **Step-1**: The multimodal inputs first are fed into LLM, and then the structured extraction results are obtained.
- **Step-2**: The output textual entities along with the original image/video/audio will be input into SEEM for fine-grained image grounding/video tracking, and into SHAS for speech segmentation. Note that we initially fine-tuned the SHAS on Speech NER task (Chen et al., 2022a) to ensure its capability to recognize named entities.

C.1.3 UIE Fine-tuning for Baselines

For baseline MLLMs, since they have not undergone IE learning, directly comparing our system with the baselines may seem unfair. Therefore, we propose to level the playing field by allowing all baseline models to undergo UIE pre-training on our UIE instruction tuning dataset. The learning process and data, prompts will be kept exactly the same as ours, as shown in Appendix §B.1. Of course, we do not consider UIE training for Pure Textual LLM baselines (such as UIE and InstructUIE) since they have already been specifically pre-trained on large-scale IE datasets.

C.1.4 Baseline Prompt Settings

This is an appendix.

[4] https://github.com/universal-ie/UIE
[5] https://github.com/BeyonderXX/InstructUIE
[6] https://github.com/Vision-CAIR/MiniGPT-4
[7] https://github.com/salesforce/LAVIS/tree/main/projects/instructblip
[8] https://github.com/haotian-liu/LLaVA
[9] https://github.com/Vision-CAIR/MiniGPT-4
[10] https://github.com/OpenGVLab/Ask-Anything
[11] https://github.com/mbzuai-oryx/Video-ChatGPT
[12] https://github.com/PKU-YuanGroup/Video-LLaVA
[13] https://github.com/0nutation/SpeechGPT
[14] https://github.com/DAMO-NLP-SG/Video-LLaMA
[15] https://github.com/yxuansu/PandaGPT
[16] https://github.com/NExT-GPT/NExT-GPT

C.2 Grounded MUIE Evaluation

Our grounded MUIE evaluation dataset involves predictions for three tasks, including UIE label prediction, multimodal grounding prediction, and cognitive QA task prediction. Here, we provide detailed evaluation metrics for these three subtasks.

C.2.1 UIE Evaluation Metrics

To evaluate textual UIE results of the model, we use `span-based offset Micro-F1` as the primary metric.

- For NER task, we follow a span-level evaluation setting, where the entity boundary and entity type must be correctly predicted.
- For RE task, a relation triple is correct if the model correctly predicts the boundaries of the subject entity, the object entity, and the entity relation.
- For EE task, we report two evaluation metrics:
  - Event Trigger: an event trigger is correct if the event type and the trigger word are correctly predicted.
  - Event Argument: an event argument is correct if its role type and event type match a reference argument mention.

C.2.2 Modality Grounding Evaluation Metrics

For the evaluation of the fine-grained modality grounding accuracy, the key idea is to measure the `mean Intersection over Union (mIoU)`.

Image Segmentation. Let us denote by $\hat{M}_{img} = \{M_g\}_{g=1}^{G}$ the ground truth set of G regions, and $M_{img} = \{M_k\}_{k=1}^{K}$ the set of K predictions. Inspired by prior work, if $K \neq G$, we employ padding with $\emptyset$ to equalize the sizes of both sets, resulting in a final size of $P = \max(G, K)$. Then, we find a bipartite matching between these two sets by searching for a permutation of P elements, $\sigma \in \mathcal{S}_P$, with the lowest cost:

$$\hat{\sigma} = \text{argmin}_{\sigma \in \mathcal{S}_P} \sum_{i}^{P} \mathcal{L}_{match}(\hat{M}_i, M_{\sigma(i)}), \quad (8)$$

where $\mathcal{L}_{match}(\hat{M}_i, M_{\sigma(i)})$ is a pairwise matching cost between ground truth Mi and a prediction with index $\sigma(i)$. We compute this optimal assignment efficiently with the Hungarian algorithm. We define $\mathcal{L}_{match}(\hat{M}_i, M_{\sigma(i)})$ as $\mathcal{L}_{bce}(\hat{M}_i, M_{\sigma(i)}) + \mathcal{L}_{dice}(\hat{M}_i, M_{\sigma(i)})$. The final IoU of each prediction is:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \quad (9)$$

Based on the IoU scores, we can calculate mIoU metric by referring image segmentation dataset.

Video Tracking. For videos, we compute the Jaccard Index (a.k.a, mIoU score) for each frame via the above calculations, and then average them.

Audio Segmentation. Similarly, the mIoU score for each audio segment is computed to evaluate the quality of speech segmentation results. We measure the 1D span of the extracted segments and the 1D span of gold segments.

C.2.3 QA Grounding Evaluation Metrics

In assessing the grounding performance of QA systems, we utilize the metric of `accuracy`, following all the existing QA datasets under examination with the multi-choice QA formats (Xiao et al., 2021).

D MUIE Data Specification

In this chapter, we provide a detailed account of the construction process and content of the grounded MUIE testing dataset that we have developed.

D.1 Data Source

The datasets we used are summarized in Table 9. We primarily utilize the following raw datasets:

- **PASCAL-C**[17]: is an object detection dataset for evaluating the robustness.
- **VRD**[18], Visual Relationship Dataset: is designed to assess the precision of detecting interactions among pairs of objects. Comprising 5,000 images, the dataset encompasses 100 object categories and 70 predicates. In total, the dataset contains 37,993 relationships with 6,672 relationship types and 24.25 predicates per object category.
- **imSitu**[19]: serves as a resource for facilitating situation recognition, a task concerned with generating a succinct depiction of the scenario portrayed in an image. This includes (1) the main activity, (2) the participating actors, objects, substances, and locations and most importantly (3) the roles these participants play in the activity.
- **ACE2005**[20]: is a benchmark dataset extensively used in the field of information extrac-

[17] https://github.com/bethgelab/robust-detection-benchmark

[18] https://cs.stanford.edu/people/ranjaykrishna/vrd/

[19] http://imsitu.org/

[20] http://projects.ldc.upenn.edu/ace/

Modality	Tasks		
	NER	RE	EE
I	PASCAL-C (Mottaghi et al., 2014)	VRD (Lu et al., 2016)	imSitu (Yatskar et al., 2016)
V			VidSitu (Sadhu et al., 2021)
A	ACE05-Aud (Walker et al., 2011)	ReTACRED (Wu et al., 2022)	
T+I	Twt17 (Lu et al., 2018)	MNRE (Zheng et al., 2021)	M^2E^2 (Li et al., 2020)
T+V			VidSitu-Txt (Sadhu et al., 2021)
T+A	ACE05-Aud (Walker et al., 2011)	ReTACRED (Wu et al., 2022)	
I+A		MNRE-Aud (Zheng et al., 2021)	
T+I+A	Twt17-Aud (Lu et al., 2018)		
V+A			VidSitu-Aud (Sadhu et al., 2021)

Table 9: Summary of the grounded MUIE test data we build in this work. Items in the light yellow background mean they are the data after preprocessing, i.e., via modality translation, where the colored postfix means the target modality.

tion and natural language processing. It comprises annotated news articles in English, covering a diverse range of topics and events, with annotations including named entities, relations, and events.

- **ReTACRED**[21]: is a revised version of TACRED for relation detection, containing over 91 thousand sentences spread across 40 relations.
- **VidSitu**[22]: is a large-scale dataset containing diverse 10-second videos from movies depicting complex situations (a collection of related events). Events in the video are richly annotated at 2-second intervals with verbs, semantic-roles, entity co-references, and event relations.
- **Twt17**[23], Twitter-17: is a publicly available Twitter dataset for named entity recognition. It encompasses 3,373 training tweets, 723 validation tweets, and 723 test tweets, with annotations covering four entity types, namely, *person*, *location*, *organization*, *miscellaneous*.
- **MNRE**[24], Multimodal Neural Relation Extraction: comprises 15,484 samples and 9,201 accompanying images across 23 distinct relation categories, partitioning into training, development, and testing subsets, consisting of 12,247, 1,624, and 1,614 samples respectively.
- **M^2E^2** [25]: is comprised of 245 multimedia news articles meticulously annotated with events and their corresponding arguments.

Pre-processing And Modality Translation Before we begin the annotation work for grounding, our first step is to enrich the types of modality combinations. This is essential because most MIE datasets currently focus on the combination of images and text. However, we aim to simulate a variety of common modality combinations that could occur in real-world scenarios. To achieve this, we plan to transform and preprocess existing datasets. Specifically, our approach involves cross-modal parallel translation to generate data in another modality. We intend to preprocess the following datasets:

- **VidSitu-Aud**: we start by captioning videos from the VidSitu dataset, then use the given video event annotations combined with the captions to have ChatGPT generate a coherent sentence, serving as the paired text for each piece of data.
- **VidSitu-Aud**: Based on VidSitu-Txt, we convert each sentence into speech using Text-To-Speech (TTS) tools. We employ state-of-the-art open-source TTS models: Bark[26] and Edge-TTS[27].
- **ACE-Aud**: we take original textual sentences from the ACE dataset and record speech using TTS technology.
- **MNRE-Aud**: we record speech for sentences from the MNRE using TTS.
- **Twt17-Aud**: Similarly, we record speech for sentences from the Twt17 dataset using TTS.

However, we emphasize that such parallel data generation can only produce modality-aligned content. To create diverse content, we plan to introduce randomness by adding noise to some instances. For example, we might alter parts of the original text before synthesizing the speech with TTS.

[21] https://github.com/gstoica27/Re-TACRED
[22] https://vidsitu.org/
[23] https://github.com/jefferyYu/UMT
[24] https://github.com/thecharm/MNREp
[25] https://github.com/limanling/m2e2
[26] https://github.com/suno-ai/bark
[27] https://github.com/rany2/edge-tts

Table 9 shows our original datasets and the new datasets under different modality combinations obtained through preprocessing.

Subsection Selection Criteria Considering the aim to only annotate a test set for the sake of conserving labor and reducing costs, we plan to select subsets from the test sets of various original datasets obtained. We aim to select 200 from each. Our selection criterion focuses on ensuring a high quantity of entities and objects in the content, and the final labels should cover a rich vocabulary. Given there are 15 combinations of modalities and tasks (including those augmented through our post-processing), we will have a total of 3,000 data entries.

D.2 Modality-specific MUIE Annotation

In addition to this, we also plan to manually annotate a portion of data where modalities are not aligned, simulating the scenario of modality-specific MUIE. For this purpose, we identify misaligned parts from the content of the original data and annotate them according to the original dataset's label set.

D.3 Grounding Annotation

In this part, we introduce the principles and methods for fine-grained annotation of data pertaining to the three non-textual modalities.

Concerning the annotation tool, we choose the open-source labeling tool **Label Studio**[28] for pixel-wise segmentation.

For each instance, we ask three well-trained people to give their annotations in parallel. When starting annotating, with certain modality of information, e.g., image, video or audio, by coreferring the gold IE labels, annotators should annotate the corresponding visual objects or speech segments of the counterparts. Besides, if identifying any misaligned parts from the content of any modalities, annotators should annotate it and add it to the UIE label of the instance. Once an instance is fully annotated, annotators should review the work for any missed objects or inaccuracies in segmentation/tracking, and make any necessary adjustments.

After annotating we will gather the annotations from three annotators to cross-examine the labels. Finally, Cohen's Kappa score of each instance will be calculated. For our case, for the image-related annotation, the average score reaches to 0.87; for the video-related annotation, the average score reaches to 0.75; for the image-related annotation, the average score reaches to 0.98. This indicates our annotation corpus has reached a high-level agreement.

[28] https://labelstud.io/

Modality	Tasks		
	NER-QA	RE-QA	EE-QA
I	85	76	74
V	-	-	73
A	83	81	-
T+I	90	82	85
T+V	-	-	78
T+A	86	87	-
I+A	-	87	-
T+I+A	93	-	-
V+A	-	-	76

Table 10: Human performance on the QA problems in each set.

D.4 QA Annotation

We believe that IE should not only involve extracting information but more utilizing the extracted information to aid in the deeper semantic understanding of the input modalities. For the annotation of the third part of our dataset, our main idea revolves around requiring the model to answer questions based on the identified and grounded entities, relations, and events.

Therefore, our setup mandates that the model must rely on the following two aspects to arrive at the correct answer: 1) accurate grounding results, and 2) the integration of necessary cognitive-level commonsense knowledge. This means our questions are designed to be challenging. Following previous relevant works, we format the questions as multiple-choice, providing several candidate answers with only one correct answer.

In Table 10, we present human performance as an upper-bound reference point for follow-up research.

E More experiment & analysis

E.1 Comparisons with GPT-4V

Here, we conducted a performance comparison between our model and GPT-4V. Given that GPT-4V can only handle image and text inputs, we solely evaluated information extraction in scenarios where inputs were either images or text. For videos, we adopted a sampling approach, treating multiple frames as individual images and inputting them into GPT-4V. Regarding output grounding,

we prompted GPT-4V to provide bounding boxes corresponding to identified entities. The specific prompt used for this purpose is outlined as follows:

► GPT-4V Prompt for NER:
Please recognize all entity words and categorize them by pre-defined labels in the given text, and outline them in the given images correspondingly. The output format should be "(entity1, label1)(entity2, label2)". If an entity possibly has a counterpart in the given images, please generate a bounding box with "entity1: [xmin, ymin, xmax, ymax]" to indicate the position of the entity.
Candidate category labels: *person, location, organization, country...*
Input Text: *xxxx*.
Input Image: .

► GPT-4V Prompt for RE:
Please extract all relations between named entities, and outline them in the given images correspondingly. The output format should be "(subject entity, relation, object entity)". If an entity possibly has a counterpart in the given images, please generate a bounding box with "entity1: [xmin, ymin, xmax, ymax]" to indicate the position of the subject/object entity.
Candidate relation labels: *peer, award, part_of, locate_at, nationality, place_of_residence, member_of...*
Input Text: *xxxx*.
Input Image: .

► GPT-4V Prompt for EE:
According to frame_1 <image>, frame_2 <image> ... frame_n <image>, extract all the possible events, and track the argument mentions correspondingly. When detailing the argument mentions' trajectories in your response, adhere to the output format <mention1> Frame_1:[xmin, ymin, xmax,ymax] ... Frame_n:[xmin, ymin, xmax,ymax]</mention1>
Candidate event types: *Marry, Attack, Injure, Beborn, Meet, Transport, Start-position...*,
Candidate event argument types: *Agent, Target, Direction, Time, Place, Instrument, Organization, Duration...*
Input Text: *xxxx*.
Input Image: .

For a fair comparison, we also convert our segments into bounding boxes and compare them with the results from GPT-4V, as depicted in Figure 9. We find that GPT-4V's capability is on par with that of ours regarding IE label prediction. However, when it comes to grounding, our model notably outperforms GPT-4V.

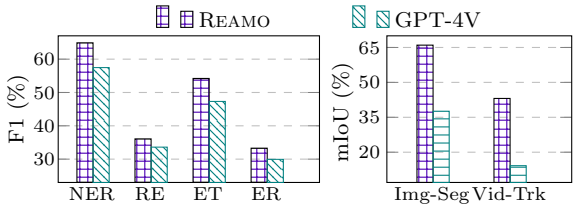

Figure 9: Comparison between REAMO and GPT-4V on image generation and

E.2 Error Analysis

Via our experiments, we here summarize several error types that our system will commit:

- **Repetition of Extracted Content**: When text and information from other modalities are not strictly consistent, our method may output different entity names, arguments, or relationships. However, upon integrating information from different modalities, they should correspond to the same entity names, arguments, or relationships.
- **Incomplete Information Extraction**: The outcomes of information extraction are incomplete, such as incomplete named entity recognition, failure to identify relations involving in-depth reasoning, or incomplete identification of event arguments.
- **Incorrect Grounding Match**: The entity or arguments do not match with the grounding results. For instance, when the text mentions 'Obama' and 'Trump' and the image depicts both individuals, the image object segmenter fails to ascertain who is 'Obama' and 'Trump', resulting in an erroneous grounding match.
- **Miss-grounding**: Our model may output entities or arguments without successfully grounding the corresponding regions in the respective image, video, or audio.
- **Over-grounding**: Our model may generate multiple '<concept>' trigger words and perform grounding in the image, video, or audio, yet no corresponding regions actually exist in the visual or auditory content.
- **Error Propagation**: Due to inaccuracies in the information extraction results or grounding segments, our model fails to answer the question correctly.
- **In-depth Knowledge Lacking**: When certain questions require deep commonsense knowledge and complex reasoning ability, our proposed model fails to incorrect answers.