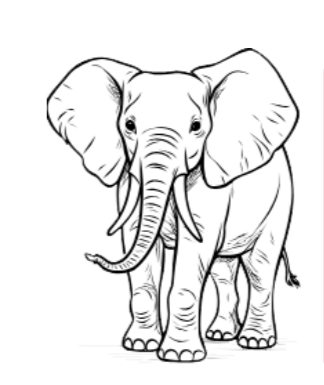


Recognizing Everything from All Modalities at Once: Grounded Multimodal Universal Information Extraction

Meishan Zhang¹ Hao Fei² Bin Wang¹ Shengqiong Wu¹
Yinxin Cao³ Fei Li⁴ Min Zhang¹

¹ Harbin Institute of Technology (Shenzhen)
³ Fudan University

² National University of Singapore
⁴ Wuhan University



ACL 2024
Bangkok, Thailand



Project: <https://haofei.vip/MUIE>

Paper: <https://arxiv.org/abs/2406.03701>

Highlight

First, to our knowledge, this is the first to propose a grounded Multimodal Universal Information Extraction (MUIE) setting, unifying all IE tasks across modalities, further with fine-grained multimodally grounded targets.

Second, we introduce an MLLM for the task, REAMO, excelling in MUIE prediction and achieving cross-modal grounding of static objects and dynamic events.

Third, we contribute a high-quality, diverse, and challenging dataset, setting an evaluation benchmark for follow-up grounded MUIE research.

Task Definition

Grounded MUIE

Text → NER

Please extract entity words from the text: Justin meet up with a fan in Beverly Hills, California yesterday.

(Justin, person)
(Beverly Hills, location)
(California, location)

Text + Image → NER + Image Segmentation

Please extract entity words in the text, and outline them in the image correspondingly: President Trump and Merkel address Korean Peninsula developments.

(Trump, person)
(Merkel, person)

Text + Audio → NER + Speech Segmentation

Please extract entity words in the text, and outline the audio correspondingly: Panasonic has reduced its workforce by nearly 10000 in both Japan and overseas companies.

(Panasonic, organization)
(Japan, location)

Text + Image + Audio → RE + Image & Speech Segmentation

Please extract relations between entities in the text, and outline them in given images and audios: Singer Mika enjoyed our Queen Shreya Ghoshal's spectacular Performances.

(Mika, peer, Shreya Ghoshal)

Text + Video → EE + Video Tracking

Please extract all possible events in the video, and track the argument mentions in the video: This boy is unpacking the plastic packaging of a new toy.

(Type: opening, Trigger: unpack, Agent: boy, Object: toy)

Video + Audio → EE + Video Tracking + Speech Segmentation

Please extract all events in the video, and track the arguments in the video, and outline them in audio: (Type: feeding, Trigger: feed, Agent: woman, Target: kangaroo)

MUIE (EE)

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.

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

MUIE (NER)

MUIE (RE)

Extract relations between entities in the text, and outline them in given images correspondingly. The output format should be "(subject, Relation, object)". Candidate relation labels: peer, award, part_of, locate_at, nationality, place_of_residence, member_of... Input Text: Rafael Nadal was interviewed by Bud Collins after his first French Open title in 2005.

The extracted relationships: (Rafael Nadal, award, French Open) (Rafael Nadal, interview_by, Bud Collins). The image is outlined as follows:

Experiment

Text+Image or standalone Image

Method	T+I Input						I Input							
	Twt17		MNRE		M ² E ²		PASCAL-C		VRD		imSitu			
	NER	I-Seg	RE	I-Seg	ET	EA	I-Seg	NER	I-Seg	RE	I-Seg	ET	EA	I-Seg
LLaVA+SEEM	23.0	45.8	15.4	51.8	22.8	13.5	48.3	17.8	26.1	10.4	36.9	19.5	8.2	29.8
InstructBLIP+SEEM	26.9	52.0	17.0	54.8	23.0	20.1	52.5	21.0	39.7	11.6	39.1	18.5	11.3	30.0
MiniGPT-v2	45.4	48.7	22.4	56.2	27.3	16.3	54.8	41.8	62.0	18.6	38.3	37.0	13.3	32.2
REAMO	47.4	53.5	24.6	56.9	30.2	25.6	60.1	43.0	64.6	26.0	43.9	41.5	16.3	39.6

Text+Audio or standalone Audio

Method	T+A Input				A Input			
	ACE05-Aud		ReTACRED		ACE05-Aud		ReTACRED	
	NER	A-Seg	RE	A-Seg	NER	A-Seg	RE	A-Seg
SpeechGPT	26.7	21.4	45.4	27.5	14.0	13.3	30.4	21.0
NEXT-GPT+SHAS	19.6	15.6	37.5	20.4	8.3	10.2	25.1	12.4
REAMO	28.5	24.3	46.8	29.1	17.4	16.7	33.4	25.1

Method	T+V (VidSitu-Txt)			V (VidSitu)		
	ET	ER	V-Trck	ET	ER	V-Trck
VideoChat+SEEM	28.8	18.5	28.1	14.3	9.2	20.9
Video-LLaVA+SEEM	31.0	22.4	31.4	18.6	8.8	20.6
REAMO	32.8	23.1	34.4	22.3	14.5	23.2

Text+Video or standalone Video

Complex modality-hybrid scenarios of MUIE

Recognize all entity words and categorize them by pre-defined labels in the given text, and outline them in the given image and audio correspondingly. The output format should be "(entity1, label1)(entity2, label2)". Candidate category labels: person, location, organization, country, title. Input Text: My cousin in Lebanon had a wedding and really invited Nancy Ajram to perform Lamaoooo. Please think step by step.

Based on the image, there is a woman in extravagant attire standing at the center of the stage, presumably the performer mentioned in the input text, named Nancy Ajram. She is performing "Lamaoooo". Additionally, according to the audio, the name mentioned is "Wendy," which is not explicitly mentioned in the text. Therefore, the extracted entity words based on the content from the image, text, and audio are: (Windy, person)(Nancy Ajram, person)(Lebanon, location)(Lamaoooo, title). The outlined image and audio are shown as follows: