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

Meishan Zhang[1] Hao Fei[2] Bin Wang[1] Shengqiong Wu[1]
Yinxin Cao[3] Fei Li[4] Min Zhang[1]

[1] Harbin Institute of Technology (Shenzhen) [2] National University of Singapore
[3] Fudan University [4] Wuhan University

ACL 2024 Bangkok, Thailand

Project: https://haofei.vip/MUIE

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

Highlight

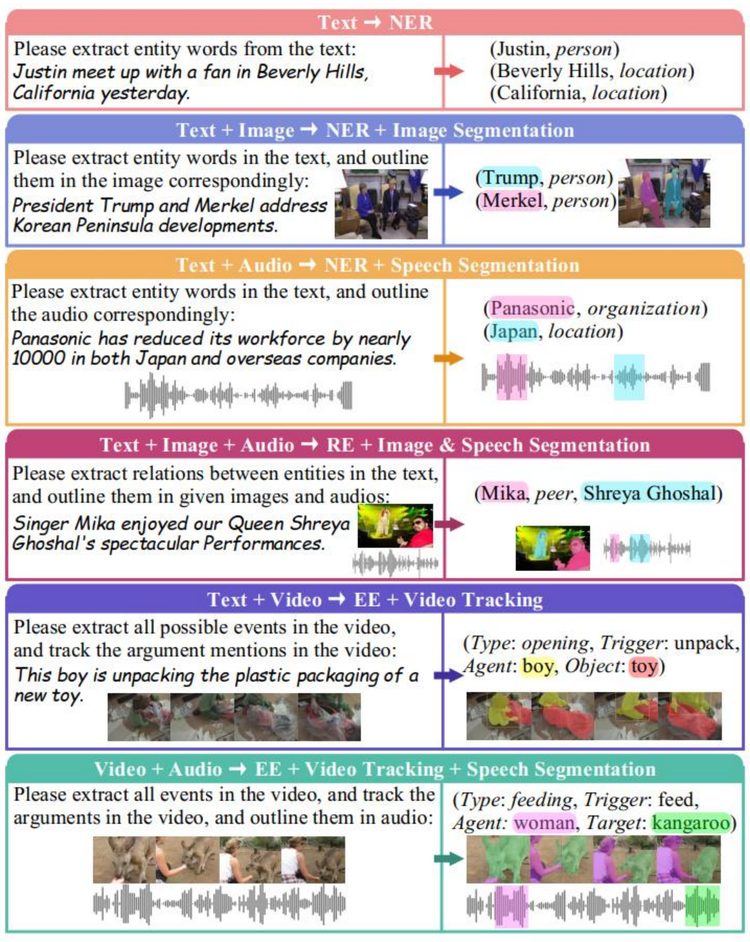
First, *to our knowledge, this is the first to propose a grounded Multimodal Universal Information Extraction (MUIE) setting, unifying all IE tasks across modalities, further with fine-grained multimodally grounded targets.*

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.*

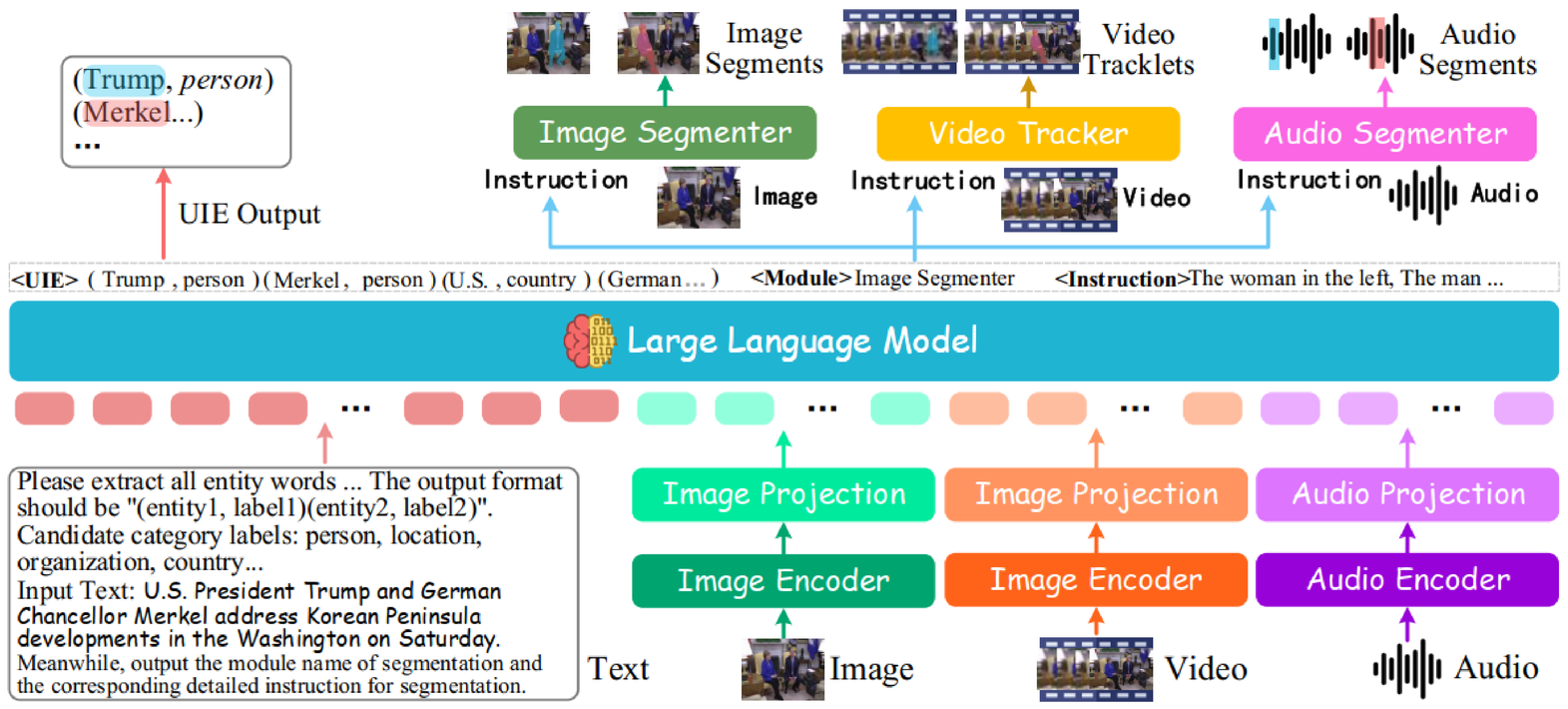
Task Definition

➢ **Grounded MUIE**



Model

➢ **REAMO:** ***R****ecognizing* ***E****verything from* ***A****ll* ***M****odalities at* ***O****nce*



Benchmark

➢ **Benchmark Evaluation Data for Grounded MUIE**

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)

Experiment

➢ Text+Image or standalone Image

Method	T+I Input							I Input						
	Twt17		MNRE		M^2E^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**

➢ Text+Video or standalone Video

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**

➢ Complex modality-hybrid scenarios of MUIE

Method	T+I+A (Twt17-Aud)			I+A (MNRE-Aud)			V+A (VidSitu-Aud)			
	NER	I-Seg	A-Seg	RE	I-Seg	A-Seg	ET	ER	V-Trck	A-Seg
Video-LLaMA+SEEM/+SHAS	-	-	-	-	-	-	12.0	4.8	12.7	8.4
NExT-GPT+SEEM/+SHAS	30.7	32.4	13.9	15.4	46.5	18.8	19.3	13.7	19.9	15.0
REAMO	**37.4**	**33.3**	**15.1**	**21.8**	**53.4**	**21.8**	**24.2**	**18.5**	**22.0**	**20.9**

➢ MUIE (EE)

➢ MUIE (RE)

➢ MUIE (NER)