

Appendices

I. DETAILED CONTENT OF PROMPTS IN SSGPF

A. M2E2 Event Schema

The prompt we designed in Event Type Schema Guided Prompting (ETSGP) to complete Multimedia Event Detection (MED) and multi-step prompts in Argument Role Schema Guided Prompting (ARSGP) to complete Multimedia Event Argument Extraction (MEAE) all contain information related to M2E2 Multimedia Event Extraction (MEE) event schema, thus we first introduce M2E2 MEE event schema. M2E2 MEE event schema is defined in [1], which is a subset of text-only event annotated dataset ACE2005's [2] event schema. M2E2 event schema contains 8 event types from ACE2005 event schema, each event type inherits corresponding candidate argument role labels defined in ACE2005 event schema. We list M2E2 event schema as follows. For each event type's definition, we follow previous work [3].

1) *Transport*: **Event type definition**: Activities involving the movement or transportation of people or goods from one place to another. **Candidate argument role labels**: Agent, Artifact, Vehicle, Destination, Origin.

2) *Attack*: **Event type definition**: Aggressive actions or assaults by one party against another. **Candidate argument role labels**: Instrument, Place, Attacker, Target.

3) *Demonstrate*: **Event type definition**: Public displays of disagreement or protest to express opinions or demands. **Candidate argument role labels**: Entity, Police, Instrument, Place.

4) *ArrestJail*: **Event type definition**: Incidents involving the arrest and subsequent detention in jail or custody of individuals. **Candidate argument role labels**: Agent, Person, Instrument, Place.

5) *PhoneWrite*: **Event type definition**: Interactions between individuals through phone calls or written communication. **Candidate argument role labels**: Entity, Instrument, Place.

6) *Meet*: **Event type definition**: Instances where individuals physically meet or come into contact with each other. **Candidate argument role labels**: Participant, Place.

7) *Die*: **Event type definition**: The life of a person ends. **Candidate argument role labels**: Agent, Instrument, Victim, Place.

8) *TransferMoney*: **Event type definition**: The exchange of money or financial resources between parties. **Candidate argument role labels**: Giver, Recipient, Money.

B. Prompt in ETSGP

At the initial step of SSGPF, we design ETSGP for MED. The detailed content of the final prompt Pr_{ED} inputting to MLLM is:

According to the image and the input text, extract all the possible events, each event associated with an event type must include a trigger word. Note that trigger word must only be noun or verb. Input text:[Input News Text]. Candidate event types: Transport, Attack, Demonstrate, Arrest-Jail, PhoneWrite, Meet, Die, TransferMoney, None. Please output the event type and trigger word separated with ';'. If multiple events are spotted, use <seg> to separate each event. The output format is: (event type;trigger word). If no event is detected, output 'None'. The definitions of these event types are: Transport:Activities involving the movement or transportation of people or goods from one place to another. Attack:Aggressive actions or assaults by one party against another. Demonstrate:Public displays of disagreement or protest to express opinions or demands. ArrestJail:Incidents involving the arrest and subsequent detention in jail or custody of individuals. PhoneWrite:Interactions between individuals through phone calls or written communication. Meet:Instances where individuals physically meet or come into contact with each other. Die:The life of a person ends. TransferMoney:The exchange of money or financial resources between parties.

The [Input News Text] denotes the place to insert the input news text's content.

C. Prompt in ARSGP

After ETSGP, we design ARSGP containing multi-step prompts for MEAE. Suppose MLLM outputs *Attack;fight* at ETSGP, we fetch candidate argument role labels of *Attack* event from M2E2 event schema. At each step of ARSGP, we focus on one argument role. Suppose that at this step we aim to extract arguments belonging to role label *Instrument*, the detailed content of the final prompt Pr_{AE} of this step inputting to MLLM is:

According to the image and the input text, extract the arguments with role 'Instrument' of the event 'Attack'. The trigger word in the text of this event is 'fight'. Argument of role 'Instrument' is the tools or weapons used in the demonstration. Input text:[Input News Text]. If no argument belongs to role 'Instrument', output 'None'. First output all arguments in text, then output all arguments in image. Use <seg> to separate text arguments and image arguments. If there are multiple arguments of this role in text or image, use ';' to separate different arguments. For each image argument, output the description of this argument in image. The output format is: (text argument a<seg>image argument b;image argument c).



sentence in ACE2005	matched image in SWiG	original annotations of the image in SWiG	gold answers in our constructed tuning set																					
<p>The Iraqis arrested looked bemused and plead innocence but with many militia here, pretending to surrender only to open fire on their captors later, first impressions can be deceptive and lethal.</p>		<table><tr><th colspan="3">activity verb: detaining</th></tr><tr><td>role: agent</td><td>name: soldier</td><td>bbox: region1</td></tr><tr><td>role: victim</td><td>name: man</td><td>bbox: region2</td></tr></table>	activity verb: detaining			role: agent	name: soldier	bbox: region1	role: victim	name: man	bbox: region2	<table><tr><th colspan="2">event type and trigger word: ArrestJail;arrested</th></tr><tr><td>role: Agent</td><td>None <seg> soldier, holding a gun</td></tr><tr><td>role: Person</td><td>The Iraqis <seg> man, being tied up by an israeli soldier</td></tr><tr><td>role: Instrument</td><td>None</td></tr><tr><td>role: Place</td><td>None</td></tr></table>	event type and trigger word: ArrestJail;arrested		role: Agent	None <seg> soldier, holding a gun	role: Person	The Iraqis <seg> man, being tied up by an israeli soldier	role: Instrument	None	role: Place	None		
			activity verb: detaining																					
role: agent	name: soldier	bbox: region1																						
role: victim	name: man	bbox: region2																						
event type and trigger word: ArrestJail;arrested																								
role: Agent	None <seg> soldier, holding a gun																							
role: Person	The Iraqis <seg> man, being tied up by an israeli soldier																							
role: Instrument	None																							
role: Place	None																							
<hr/>																								
<p>It was the first unit to cross the Euphrates River and then punch northward to within 60 miles of Baghdad.</p>		<table><tr><th colspan="3">activity verb: boating</th></tr><tr><td>role: vehicle</td><td>name: ferry</td><td>bbox: region1</td></tr><tr><td>role: boaters</td><td>name: person</td><td>bbox: region2</td></tr></table>	activity verb: boating			role: vehicle	name: ferry	bbox: region1	role: boaters	name: person	bbox: region2	<table><tr><th colspan="2">event type and trigger word: Transport;cross</th></tr><tr><td>role: Agent</td><td>None</td></tr><tr><td>role: Artifact</td><td>the first unit <seg> person is standing on the top of a boat</td></tr><tr><td>role: Vehicle</td><td>None <seg> a green and yellow boat with a white hull</td></tr><tr><td>role: Origin</td><td>Euphrates River <seg> None</td></tr><tr><td>role: Destination</td><td>within 60 miles of Baghdad <seg> None</td></tr></table>	event type and trigger word: Transport;cross		role: Agent	None	role: Artifact	the first unit <seg> person is standing on the top of a boat	role: Vehicle	None <seg> a green and yellow boat with a white hull	role: Origin	Euphrates River <seg> None	role: Destination	within 60 miles of Baghdad <seg> None
			activity verb: boating																					
role: vehicle	name: ferry	bbox: region1																						
role: boaters	name: person	bbox: region2																						
event type and trigger word: Transport;cross																								
role: Agent	None																							
role: Artifact	the first unit <seg> person is standing on the top of a boat																							
role: Vehicle	None <seg> a green and yellow boat with a white hull																							
role: Origin	Euphrates River <seg> None																							
role: Destination	within 60 miles of Baghdad <seg> None																							

Fig. 1. Examples of weakly-aligned image-text pairs matched by our crossmodal news retriever and annotations in the original SWiG dataset and our constructed dataset. Upper part: matched image with the given sentence considered as jointly describing an ArrestJail event. Lower part: matched image with the given sentence considered as jointly describing a Transport event.

II. MORE DETAILS OF THE DATASETS

A. Datasets Statistics

We use text-only event annotated dataset ACE2005 [2] and image-only event annotated dataset SWiG [4] to build our weakly-aligned multimodal event labeled instruction tuning set. ACE2005 contains 14,671 news-related sentences annotated with event types, trigger words, and argument roles. It includes 33 possible event types. SWiG has approximately 125,000 images, each depicting a more general event beyond news, and is labeled with an activity verb such as “detaining”, “boating”, and contains bounding box and role label for each argument in image contributing to the general event. The event schema definition of M2E2 MEE benchmark [1] is a subset of ACE2005 event schema, and for SWiG, previous work [1] established the mapping from SWiG event schema to MEE event schema, where 98 of 504 SWiG event types are aligned with 8 event types of M2E2. All argument role labels in these 98 event types in SWiG can be mapped to the corresponding event type’s candidate argument role labels in M2E2 event schema, which is also provided by [1]. Thus our constructed multimodal event labeled dataset collects all sentences from ACE2005 whose event type is included in the M2E2 schema. For each sentence, all images in SWiG corresponding to the same event type as the sentence will be selected as candidates. We match the most event-related image from all candidates with the sentence to form one sample in our constructed dataset. The statistics of the total count of multimodal events of MED, textual arguments and visual arguments of MEAE in our constructed dataset for instruction tuning and M2E2 for MEE evaluation are presented in Table I.

B. Dataset Visualization

We present actual examples of our constructed weakly-aligned multimodal event labeled instruction tuning set. As shown in Fig. 1, we present two examples of our dataset

TABLE I
STATISTICS OF MULTIMODAL EVENTS, TEXTUAL ARGUMENTS AND VISUAL ARGUMENTS IN OUR CONSTRUCTED INSTRUCTION TUNING SET AND M2E2 MEE EVALUATION DATASET

Dataset	MED subtask	MEAE subtask	
	Events	Textual Args	Visual Args
Our constructed	2,316	6,376	9,635
M2E2	309	465	995

construction. We describe the upper example in detail as an illustration. Given the text from ACE2005, we employ a crossmodal news retriever to find the most event-related image in SWiG. As shown in the upper example in Fig. 1, the original annotations of this image in SWiG contain an activity verb *detaining* and two arguments. Each argument contains annotations of its role, name, and bounding box region in image representing this argument. We match this image with the sentence to create gold answers. We consider the image and the sentence as multimodal inputs which jointly describe an *ArrestJail* event in M2E2 event schema, and the trigger word in text expressing such event is *arrested*.

For each argument role in current event type’s argument role schema, we create gold answers containing arguments from text and image belonging to this role. Gold answers of trigger word and arguments in text all come from original annotations of this sentence in ACE2005. As shown at “gold answers in our constructed tuning set” of the upper example in Fig. 1, for role label *Agent*, no argument in sentence belongs to this role, while *region1* (bounding box colored blue in image) has original role label *agent* mapped to M2E2 role label *Agent*, thus we crop the original image to retain only the bounding box region of *region1*, sending this image region to an image captioning model [5], asking the model to generate description about the region. We insert the corresponding name

of this region in original SWiG annotations into the image captioning model for a more accurate and detailed description. For *region1*, we create prompt “Describe the soldier in the image.” for the captioning model to generate description. The captioning model generates “soldier, holding a gun” as this image argument’s description. Then we concatenate all candidate arguments in text and image belonging to this role as the gold answer of this argument role following the output format requirements defined in ARSGP for MEAE. As shown in this example, the gold answer for role label *Agent* is *None* <seg> soldier, holding a gun, and the gold answer for role label *Person* is *The Iraqis* <seg> man, being tied up by an israeli soldier, as the sentence contains argument *The Iraqis* with this role label, and *region2* (bounding box colored green in image) with original role label *victim* is mapped to role label *Person* in M2E2 following the mapping strategy provided by [1], then the captioning model generates description “man, being tied up by an israeli soldier” of this region.

Based on the examples shown in Fig. 1, we can see that the crossmodal news retriever is indeed capable of finding image-text pairs that are relatively consistent and aligned towards the described event scenario. Therefore, the matched image-text pair can be considered as weakly-aligned to jointly describe a multimodal event, ensuring the effectiveness of the multimodal supervised signals in our constructed dataset.

III. IMPLEMENTATION DETAILS

The complete SSGPF contains two sub-modules, ETSGP and ARSGP. They share the same neural network architecture with different textual input and output. They both contain 320M trainable LoRA parameters, which is significantly fewer than the 7B (=7,000M) parameters of the Vicuna-v1.5-7B [6] LLM used in LLaVA-v1.5-7B [7] MLLM. ETSGP and ARSGP are trained with mixed precision. We use SEEM [8] of *seem_focalt_v0* version as the Visual Grounding (VG) model for final image argument locating. As SEEM output segmentation masks for the grounded region, we find the topmost, bottommost, leftmost, and rightmost boundaries of the grounded object’s segmentation masks to form the final bounding box of the argument. We use BLIP-2 [5] of *blip2-flan-t5-xl* version as the image captioning model to generate image arguments’ gold answers in our constructed tuning set. We use CLIP [9] of *clip-vit-base-patch16* version as the crossmodal news retriever and train it on VOA image-text corpora with 10 epochs, batch size 196, learning rate 1e-5, and AdamW optimizer. As some image links provided by [1] for VOA image-text corpora are broken, we finally collect 88,897 of 123,078 VOA image-text pairs for crossmodal news retriever training. All models are deployed on one 48GB NVIDIA RTX 6000 Ada GPU.

IV. MORE DETAILS OF MODELS IN ABLATION STUDIES

In our ablation studies, we design SSGPF_{JALL} to jointly complete MED and MEAE in a single response. We merge the prompts in original ETSGP and ARSGP of our complete SSGPF, and the single step prompt of SSGPF_{JALL} contains

all possible event types and all possible argument roles. To distinguish different argument roles of different event types, we define the argument role output sequence format of all event types and insert it into SSGPF_{JALL} prompt, asking MLLM to output predicted arguments following the sequence format. The argument role output sequence format is:

Transport: [Agent] [Artifact] [Vehicle] [Origin] [Destination]. *Attack*: [Attacker] [Target] [Instrument] [Place]. *Demonstrate*: [Entity] [Police] [Instrument] [Place]. *Arrest-Jail*: [Agent] [Person] [Instrument] [Place]. *PhoneWrite*: [Entity] [Instrument] [Place]. *Meet*: [Participant] [Place]. *Die*: [Agent] [Victim] [Instrument] [Place]. *TransferMoney*: [Giver] [Recipient] [Money].

Following the sequence format, SSGPF_{JALL} first outputs the event type, then outputs each argument role in the event type, and the extraction results of arguments are placed after each argument role with the same outputting format as one single step’s output in original ARSGP. As the complete SSGPF contains two sets of LoRA modules for ETSGP and ARSGP, SSGPF_{JALL} aims to complete MED and MEAE in one single step, thus it only retains one set of LoRA modules.

For SSGPF_{JMEAE} to complete MEAE in a single step, the initial step ETSGP in SSGPF_{JMEAE} for MED is the same as full SSGPF model, and as ETSGP and ARSGP are trained independently, SSGPF_{JMEAE} changes ARSGP prompting logic while maintaining ETSGP the same as full SSGPF, thus the experimental results on MED of SSGPF_{JMEAE} are the same as full SSGPF. SSGPF_{JMEAE} also contains two sets of LoRA modules. For ARSGP in SSGPF_{JMEAE}, it only retains a single step prompt asking MLLM to output all arguments of the predicted event type, thus we fetch the predicted event type outputted by ETSGP, then we only fetch this targeted event type’s argument role output sequence format and insert it into the prompt of ARSGP. For example, if ETSGP step outputs event type *Attack*, the argument role output sequence format in ARSGP prompt will be [Attacker] [Target] [Instrument] [Place]. The ARSGP in SSGPF_{JMEAE} outputs each argument role in the sequence format, and the extraction results of arguments are placed after each argument role, and the output format of the predicted arguments is the same as one single step’s output of the original ARSGP in our full SSGPF.

For “SSGPF-unimodal” in the ablation studies of the constructed dataset, we directly leverage two unimodal annotated training sets ACE2005 [2] and SWiG [4] for SSGPF training, following the weakly supervised unimodal training framework of previous MEE works. As our MLLM backbone requires both image and text inputs simultaneously, for training samples in text-only ACE2005, we set the visual input as a plain white image with no actual content, gold answers for MED contain event types and trigger words, gold answers for MEAE only contain arguments from text, the content in gold answers for the image argument (content after special token <seg>) section is set as “None”. For training samples in image-only SWiG, we first map the SWiG event schema to M2E2 schema following [1], then for each training sample in SWiG, we set the input news text directly as the activity verb of each

image in original SWiG annotations, and set the gold answer of trigger word for MED as this activity verb. Gold answers in MEAE only contain arguments from image, the content in gold answers for the argument in text (content before special token `<seg>`) section is set as “None”.

REFERENCES

- [1] Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang, “Cross-media structured common space for multimedia event extraction,” in *ACL*, 2020, pp. 2557–2568.
- [2] Walker Christopher, Strassel Stephanie, Medero Julie, and Maeda Kazuaki, “Ace 2005 multilingual training corpus,” in *Linguistic Data Consortium, Philadelphia*, 2006.
- [3] Yuxuan Sun, Kai Zhang, and Yu Su, “Multimodal question answering for unified information extraction,” *arXiv preprint arXiv:2310.03017*, 2023.
- [4] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi, “Grounded situation recognition,” in *ECCV*, 2020, pp. 314–332.
- [5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *ICML*, 2023, pp. 19730–19742.
- [6] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica, “Judging llm-as-a-judge with mt-bench and chatbot arena,” in *NeurIPS*, 2023, pp. 46595–46623.
- [7] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee, “Improved baselines with visual instruction tuning,” in *CVPR*, 2024, pp. 26296–26306.
- [8] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee, “Segment everything everywhere all at once,” in *NeurIPS*, 2023, pp. 19769–19782.
- [9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., “Learning transferable visual models from natural language supervision,” in *ICML*, 2021, pp. 8748–8763.