Three Stream Based Multi-level Event Contrastive Learning for Text-Video Event Extraction

Jiaqi Li[1,3*], Chuanyi Zhang[2*], Miaozeng Du[1,3], Dehai Min[1,3], Yongrui Chen[1,3], Guilin Qi[1,3†]

[1] School of Computer Science and Engineering, Southeast University, Nanjing, China

[2] College of Artificial Intelligence and Automation, Hohai University, Nanjing, China

[3] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China

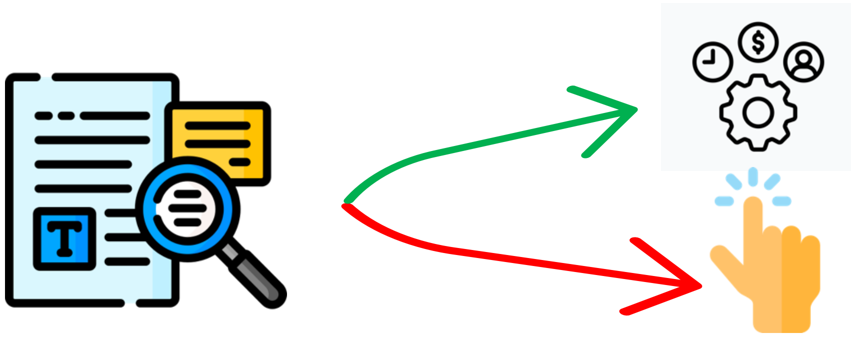
Background

- Event Extraction

Triggers

Arguments

··· and **arrested** [Arrest] six people [Arrest.Person] on charges of conspiracy to publish seditious publications [Arrest.Crime].



Arguments: Participants in the event

Triggers: A word or phrase that triggers an event

- ***Traditional event extraction only considers the information in the text, but lacks the rich event information in other modalities.***

Background

- Multimodal Event Extraction (text-video)
  - Identifying event information from the given text-video pairs
  - Input: Text x_i, video clip y_i
  - Output: Triggers x_i^t, arguments x_i^a

Triggers

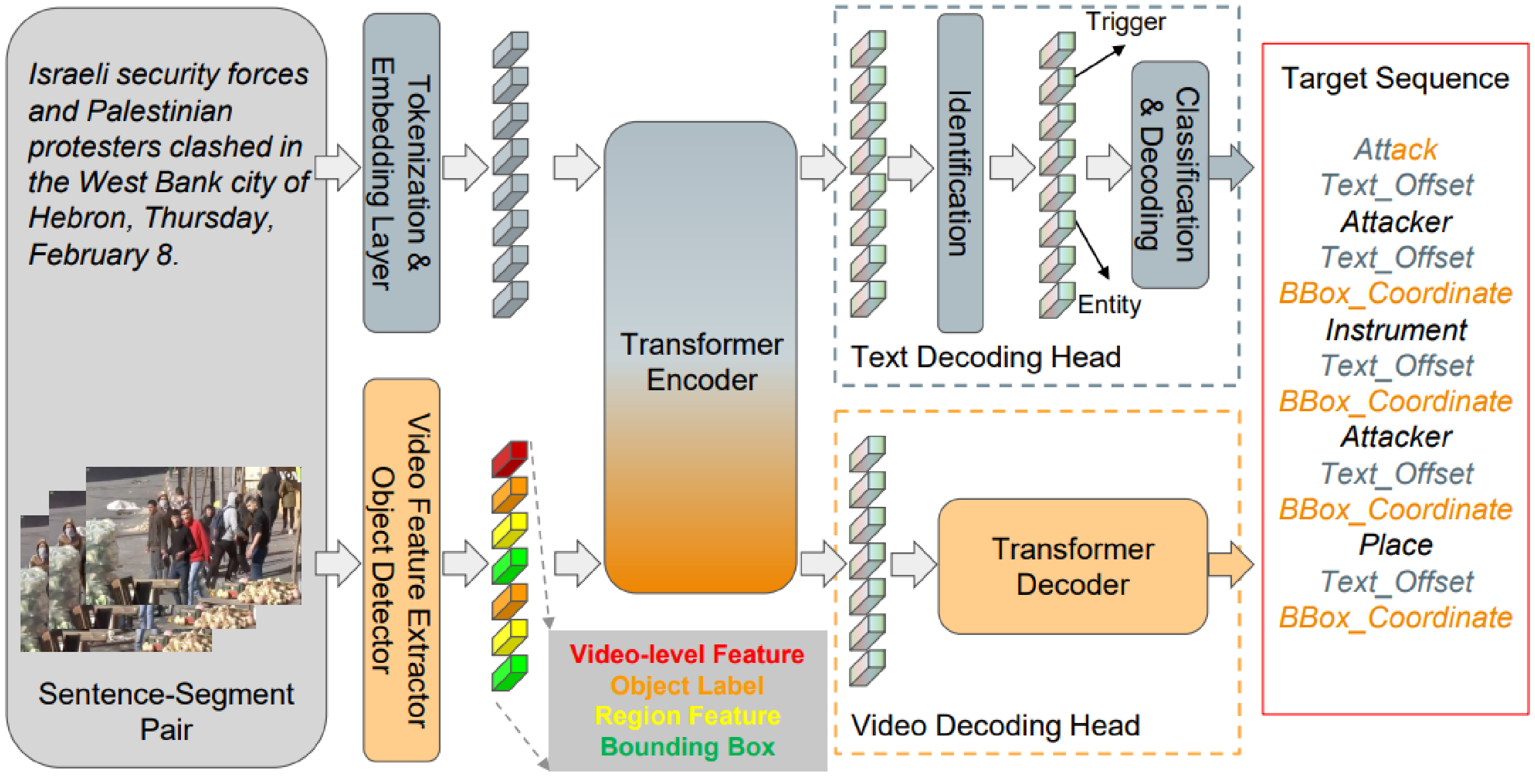
Arguments

··· and **arrested** [Arrest] six people [Arrest.Person] on charges of conspiracy to publish seditious publications [Arrest.Crime].

Existing Work

- Diverse format of features
  - Video-level feature
  - Object label
  - Region feature
  - Bounding box

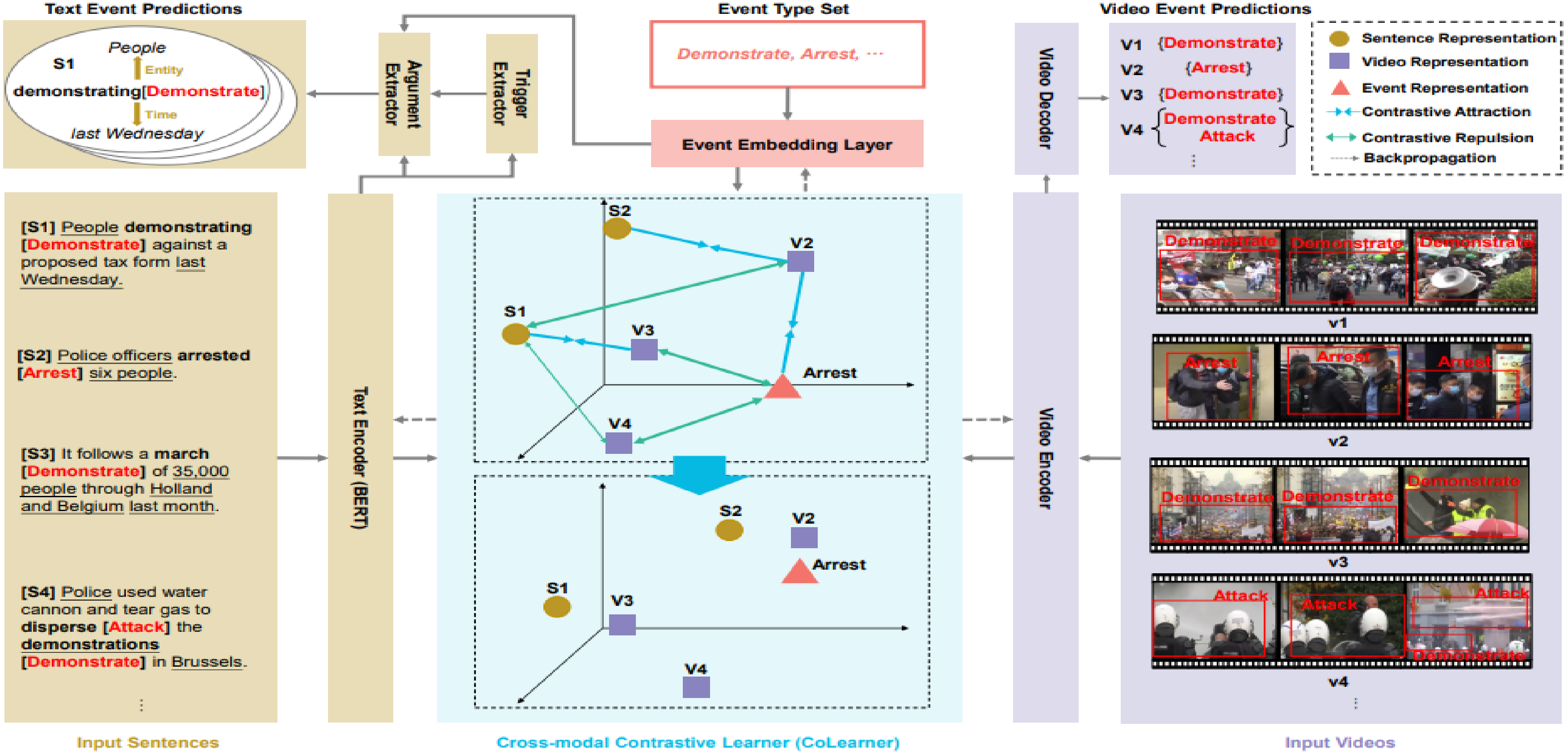
Encoder-Decoder Architecture



Joint Multimedia Event Extraction from Video and Article (EMNLP 2021)

Existing Work

- Contrastive learning for modality alignment
  - Contrast global video feature and text feature
  - Contrast global video feature and event type feature



Cross-modal Contrastive Learning for Event Extraction (DASFAA 2023)

Issues of existing work

- **Disregard motion representation**
  - Existing works ignore the rich motion features in videos

- **Global video features are hard to be directly aligned with event types**
  - Event types in videos refer to specific video clip
  - Abundant background noises in the appearance features of videos

Issues of existing work

… and **arrested** [Arrest] six people [Arrest.Person] on charges of conspiracy to publish seditious publications [Arrest.Crime].

Background noise

Motion representation

Motivation of our work

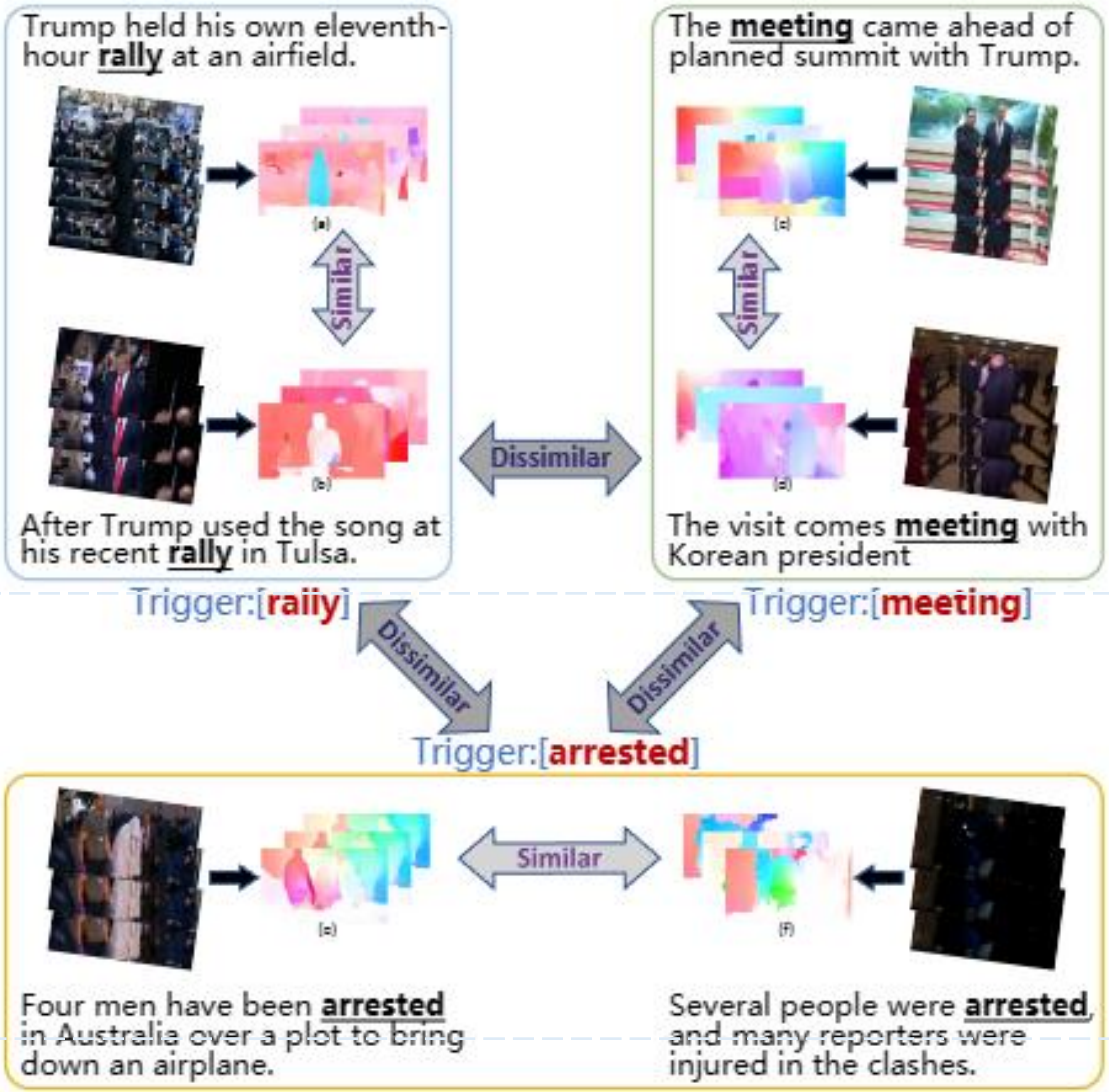
- ***Introducing optical flow***

a) Utilizing motion representations
b) Excluding background noises

Motivation of our work

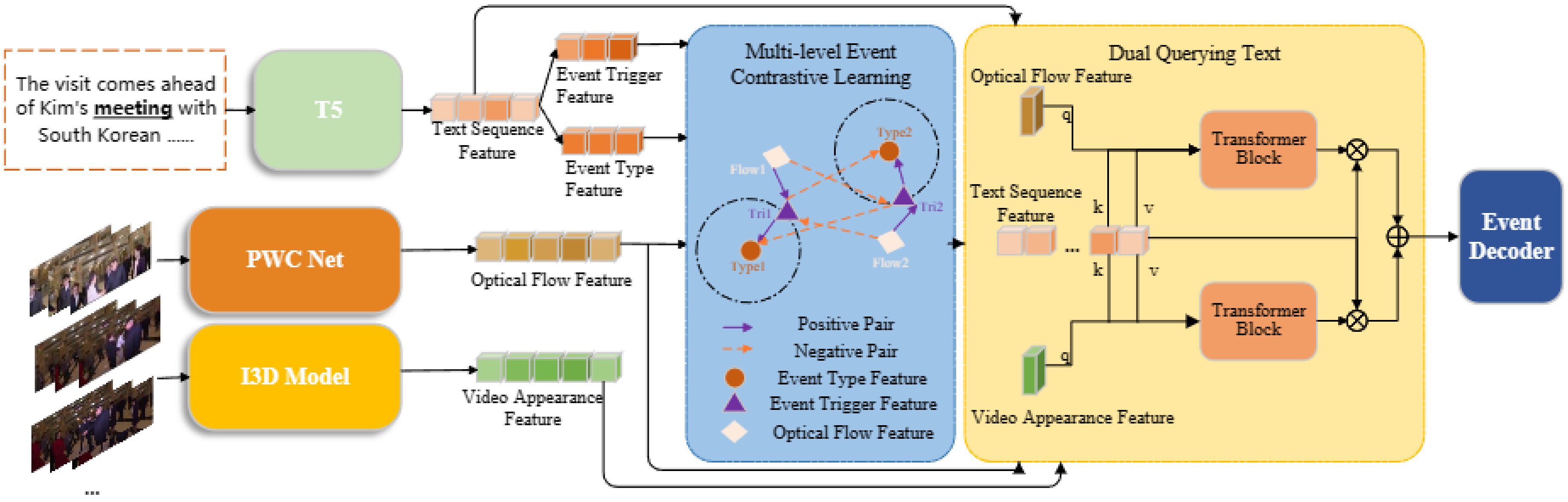
- ***Introducing optical flow***

We observe that the same event triggers correspond to similar motion trajectories



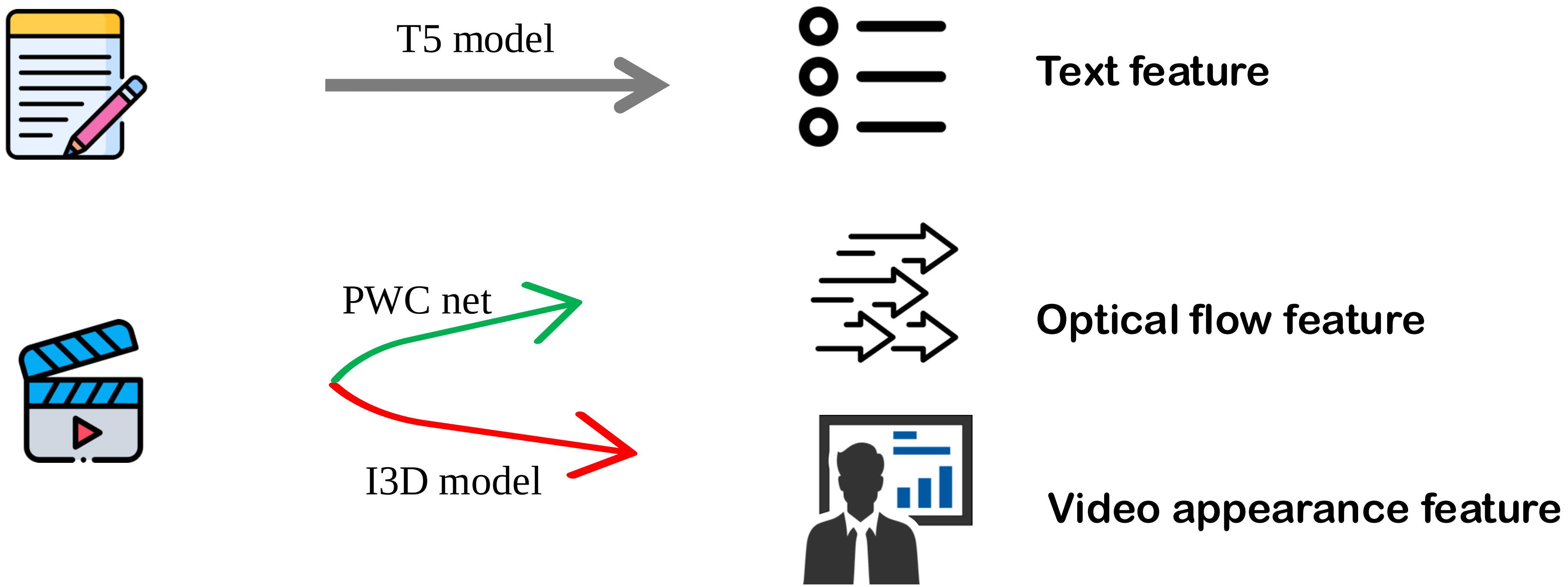
Our Work

a) Introduce optical flow to multimodal event extraction
b) Propose multi-level event contrastive learning to align the embedding space between optical flow and event trigger
c) Design dual querying text module to enhance the interaction among multiple modalities.



Base multimodal feature extractors

- A strong language model T5 as the text encoder;
- PWC net utilized as the optical flow feature extractor;
- A powerful video appearance feature encoder I3D model



Multi-level Event Contrastive Learning

Why event Contrastive learning

- The background noises in various videos make it hard to align global features and event features

- Event features could supply more specific information and be directly beneficial to event information extraction

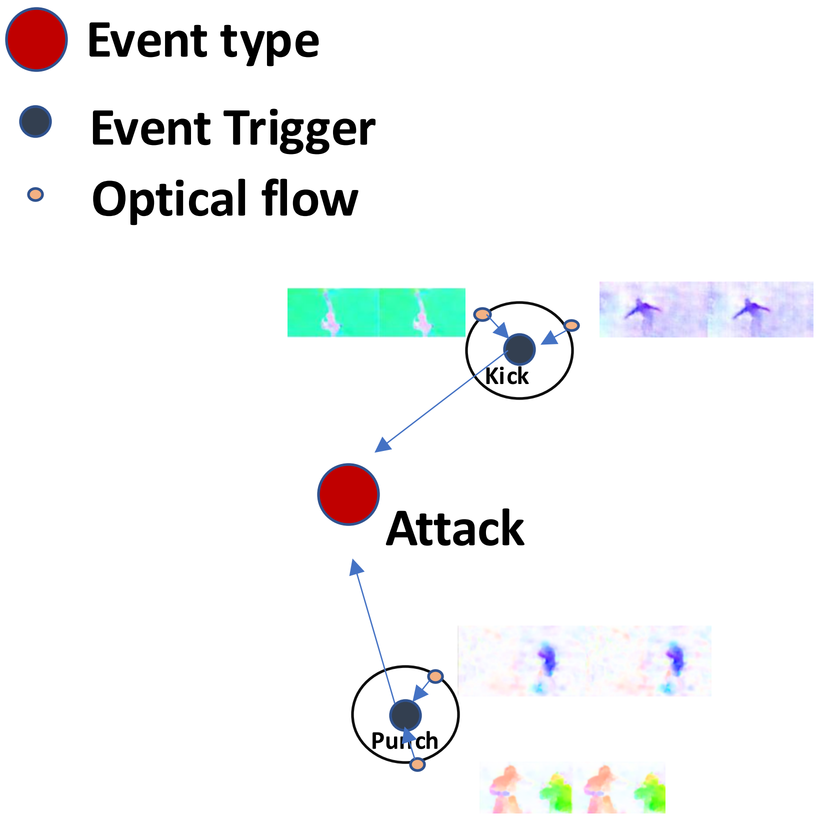
Multi-level Event Contrastive Learning

Why Multi-level

- Identical event triggers usually involve similar motion representations

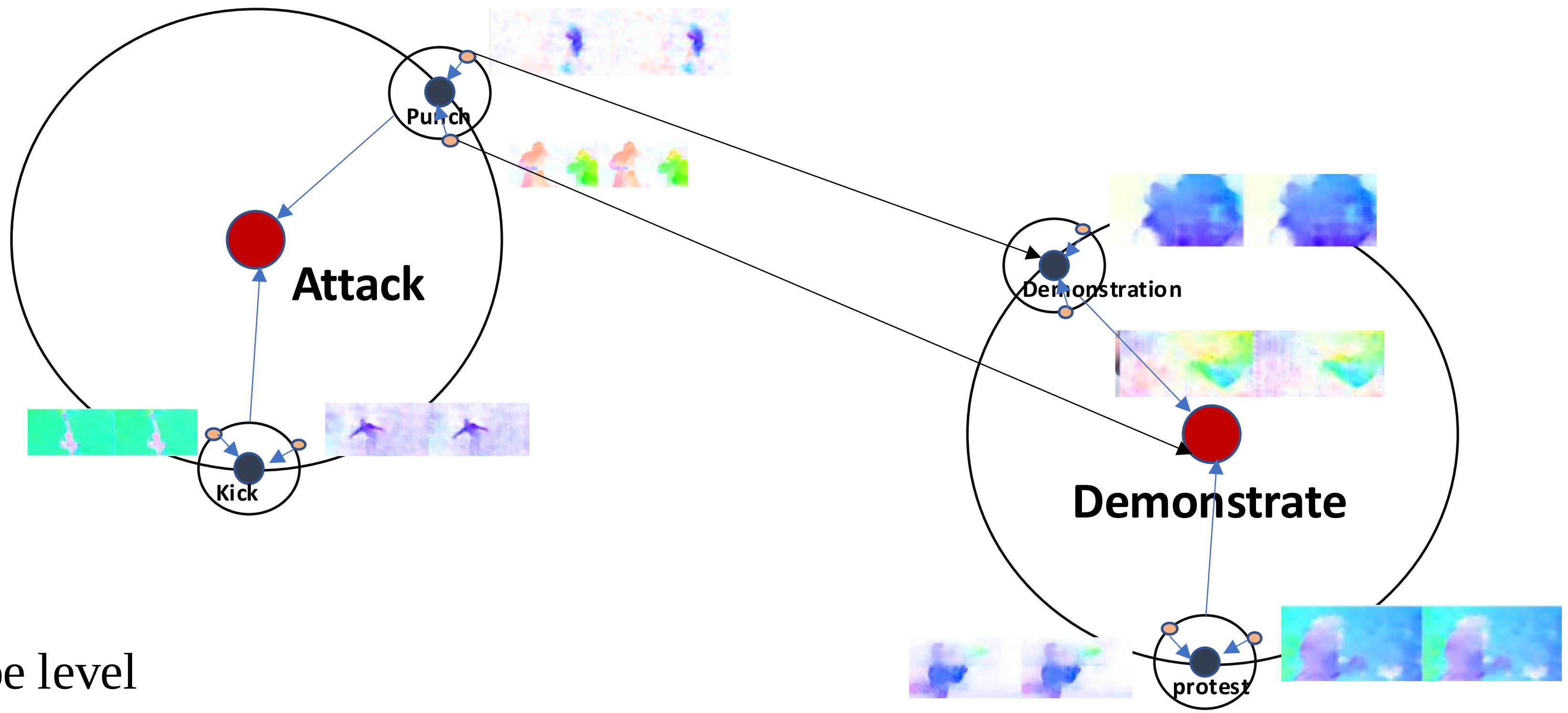
- In the event extraction, an event type is correlated to various triggers

Multi-level Event Contrastive Learning



- Attack → kick, punch;
- Kick → flow1, flow2, …;
- Punch → flow3,flow4, …

Multi-level Event Contrastive Learning



- Event type level

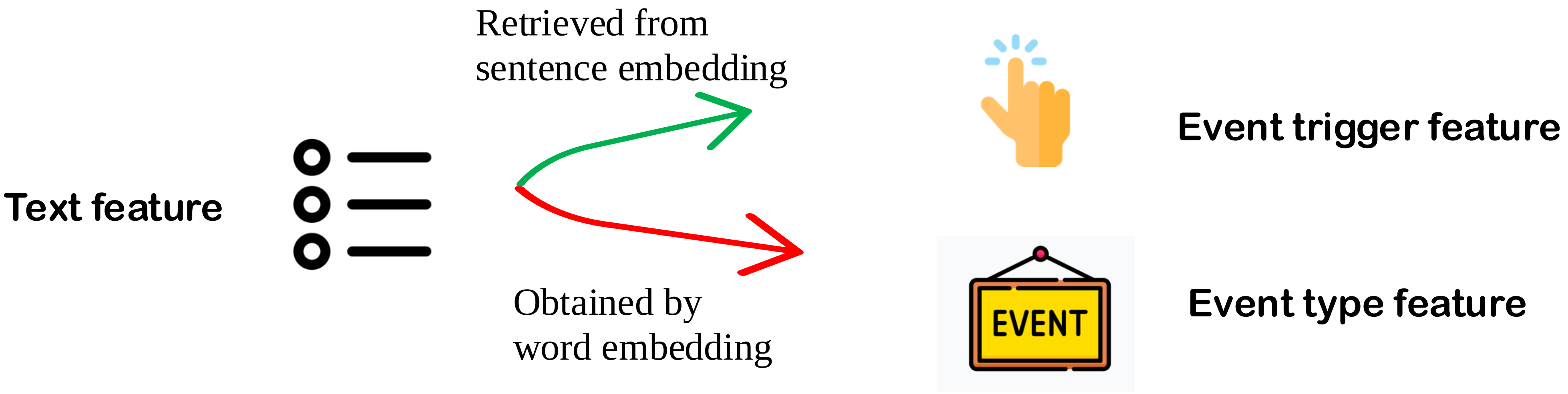
  Since an event type corresponds to various triggers, we use event types as the anchors for triggers

- Event trigger level

  Considering the same event triggers correspond to similar motion trajectories in videos, we regard the triggers as the anchors for optical flows

Multi-level Event Contrastive Learning

- Firstly we obtain the event type features and event trigger features from text features



Multi-level Event Contrastive Learning

- Then we set the positive and negative pairs in training process

  - Event type level
    - Positive pairs of each event type consist of all referring trigger words and the event type itself
    - The negative pairs comprise irrelevant trigger words and the event type itself

  - Event trigger level
    - Each trigger's positive pairs are composed of optical flow features that point to the trigger and the trigger itself
    - The negative pairs are made up of optical flow features that are unrelated to the trigger and the trigger itself

Multi-level Event Contrastive Learning

- Finally we define the loss function of contrastive learning.
    - Multi-level training loss
    - Supervised contrastive learning form

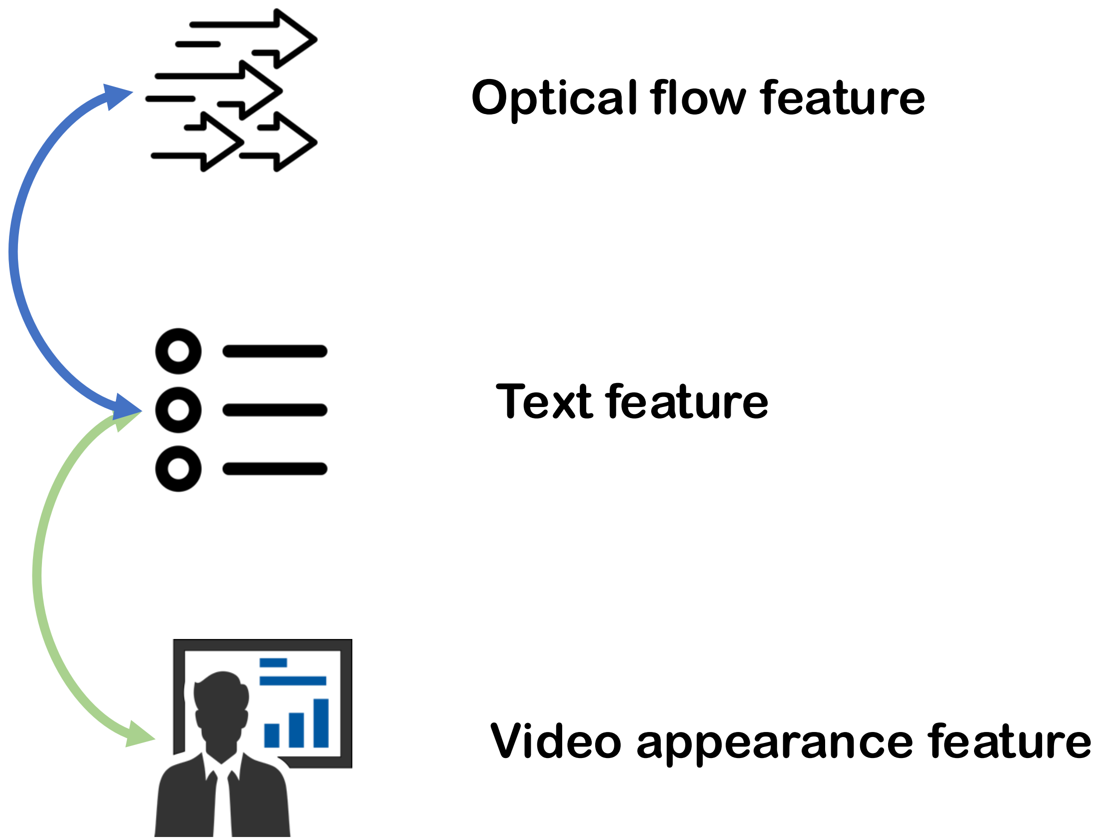
$$\mathcal{L}_{type} = -\sum_{i=1}^{B} log \frac{exp(x^i \cdot z^i/\tau)}{\sum_{z^l \in W_c \setminus z^i} exp(x^i \cdot z^l/\tau)},$$

$$\mathcal{L}_{trig} = -\sum_{i=1}^{B} log \frac{exp(z^i \cdot F_O^i/\tau)}{\sum_{F_O^u \in F_{O_c} \setminus F_O^i} exp(z^i \cdot F_O^u/\tau)},$$

- τ is the temperature parameter of supervised contrastive learning
- x^i, z^i,F_O^i are the features of event types, event triggers and optical flows respectively

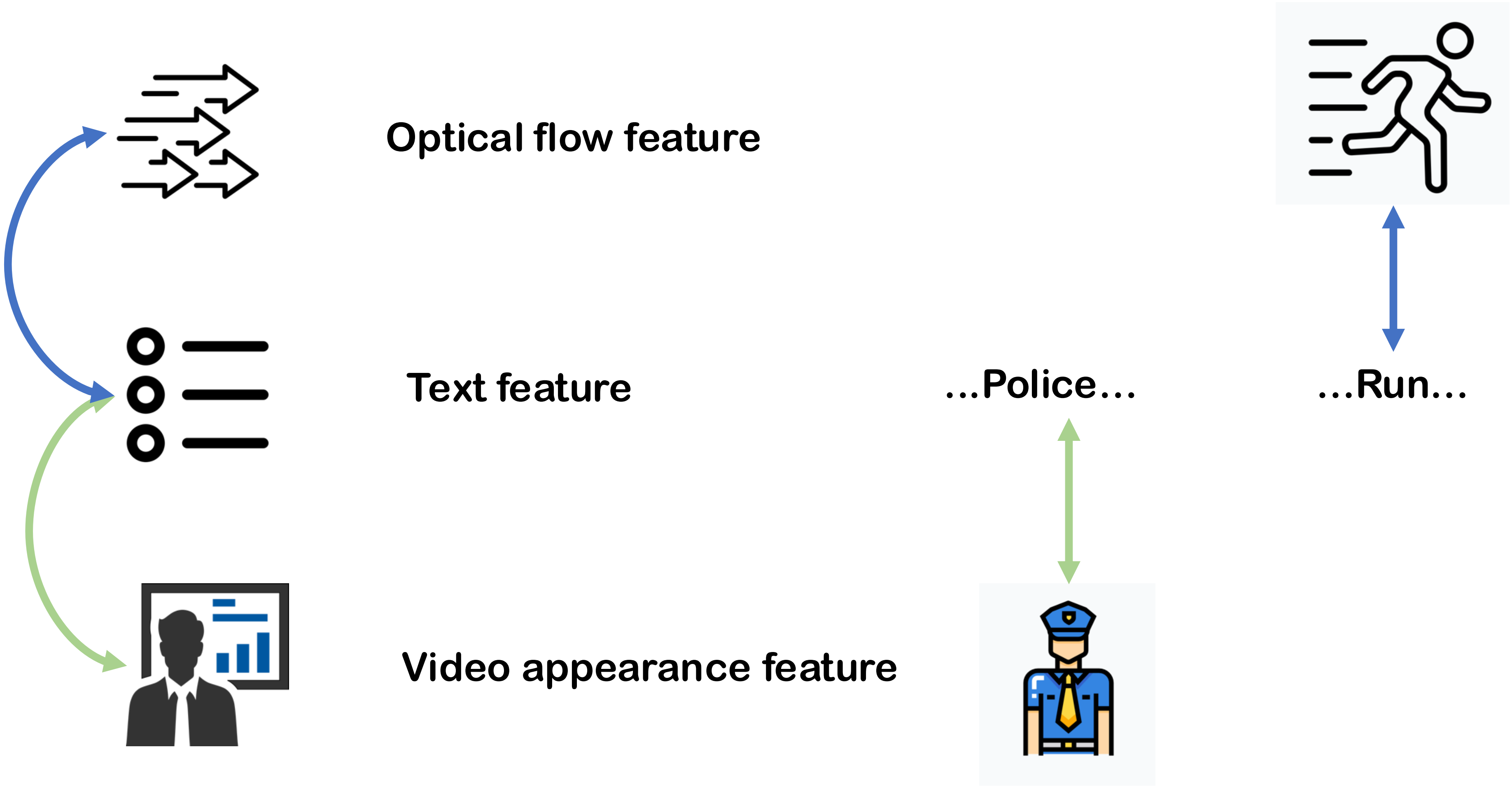
Dual Querying Text

- Enhance the interaction among three modalities
- Improve the explainability



Dual Querying Text

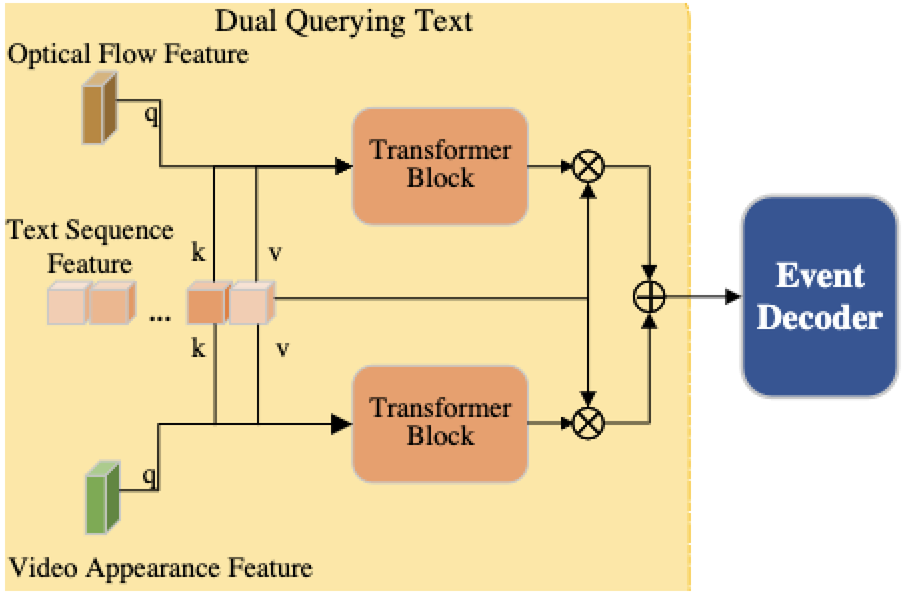
- Query each token in the text to find out which token reflects the optical/video most.



Dual Querying Text

- Employ two transformer architectures
- Set optical & video as the query input and text as the key & value input

$$A_v = softmax(\frac{F_V H_{q_1} H_{k_1}^\top F_T^\top}{\sqrt{dt}}) F_T H_{v_1},$$
$$A_o = softmax(\frac{F_O H_{q_2} H_{k_2}^\top F_T^\top}{\sqrt{dt}}) F_T H_{v_2},$$



Experimental Setup

- Datasets
  - TVEE
  - VM2E2
- Evaluation Metrics
  - Trigger: Precision, Recall, F1
  - Argument: Precision, Recall, F1
- The event schema is from ACE2005 benchmark that consists of 8 superior event types and 33 event types
- Contact, Speech, Disaster, Accident are added to the event schema because schema in ACE2005 could not cover all the event types in videos

Results

- Overall Results

Dataset	Input	Model	Text Evaluation						Video Evaluation			Multimodal Evaluation		
			Trigger			Argument								
			P	R	F1	P	R	F1	P	R	F1	P	R	F1
TVEE	Text	DEEPSTRUCT	76.4	75.2	75.8	53.1	48.9	50.9	-	-	-	76.4	75.2	75.8
		CoCoEE$_T$	76.0	76.6	76.3	62.9	44.2	51.9	-	-	-	76.0	76.6	76.3
		TSEE$_T$	75.7	77.2	76.4	63.3	45.0	52.6	-	-	-	75.7	77.2	76.4
	Video	JSL	-	-	-	-	-	-	48.2	51.6	49.8	48.2	51.6	49.8
		CoCoEE$_V$	-	-	-	-	-	-	49.1	60.7	54.3	49.1	60.7	54.3
		TSEE$_V$	-	-	-	-	-	-	48.7	62.1	54.6	48.7	62.1	54.6
	Multimodal	JMMT	74.3	80.2	77.1	50.1	**54.9**	52.3	55.4	57.0	56.2	87.2	88.6	87.9
		CoCoEE	80.7	76.4	78.5	65.6	45.4	53.6	56.4	57.4	56.9	92.9	92.9	92.9
		TSEE (ours)	**82.6**	**80.5**	**81.5**	**67.0**	49.3	**56.8**	**58.2**	**58.6**	**58.4**	**94.4**	**93.7**	**94.0**
VM2E2	Text	DEEPSTRUCT	44.7	43.1	43.9	19.8	13.2	15.9	-	-	-	44.7	43.1	43.9
		CoCoEE$_T$	41.5	45.6	43.5	20.5	15.3	17.5	-	-	-	41.5	45.6	43.5
		TSEE$_T$	45.2	41.8	43.4	21.2	17.1	18.9	-	-	-	45.2	41.8	43.4
	Video	JSL	-	-	-	-	-	-	21.2	18.6	19.8	21.2	18.6	19.8
		CoCoEE$_V$	-	-	-	-	-	-	27.3	31.2	29.1	27.3	31.2	29.1
		TSEE$_V$	-	-	-	-	-	-	26.5	30.7	28.4	26.5	30.7	28.4
	Multimodal	JMMT	39.7	**56.3**	46.6	17.9	24.3	20.6	32.4	37.5	34.8	76.1	69.5	72.7
		CoCoEE	47.3	47.7	47.5	**26.7**	18.5	21.8	33.2	37.2	35.1	78.2	75.6	76.9
		TSEE (ours)	**49.2**	53.5	**51.6**	24.5	**27.4**	**25.9**	**35.1**	**38.0**	**36.5**	**78.9**	**77.2**	**78.0**

- Multimodal methods perform better than unimodal methods;
- Our method outperforms all the baselines in terms of trigger evaluation on TVEE dataset

Results

- Ablation Study

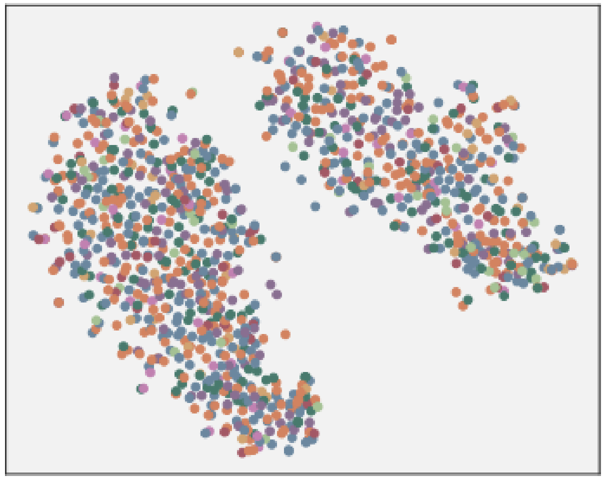
Dataset	Units			Trigger			Argument		
	O	H	D	P	R	F1	P	R	F1
TVEE				76.2	76.9	76.5	62.8	46.1	53.2
	✔			76.8	77.3	77.0	63.9	45.7	53.3
	✔	✔		80.5	79.2	79.8	64.5	47.3	54.6
	✔	✔	✔	**82.6**	**80.5**	**81.5**	**67.0**	**49.3**	**56.8**
VM2E2				42.3	45.9	44.0	21.3	16.6	18.7
	✔			44.0	47.2	45.5	20.8	18.1	19.4
	✔	✔		47.9	50.6	49.2	22.7	25.3	23.9
	✔	✔	✔	**49.2**	**53.5**	**51.6**	**24.5**	**27.4**	**25.9**

Table 2: Ablation study on three units in TSEE. ‘**O**’ represents OFF (Optical Flow Features). ‘**H**’ means MECL (Multi-level Event Contrastive Learning) module. ‘**D**’ denotes DQT (Dual Querying Text) module. ‘✔’ represents our framework is equipped with the unit.

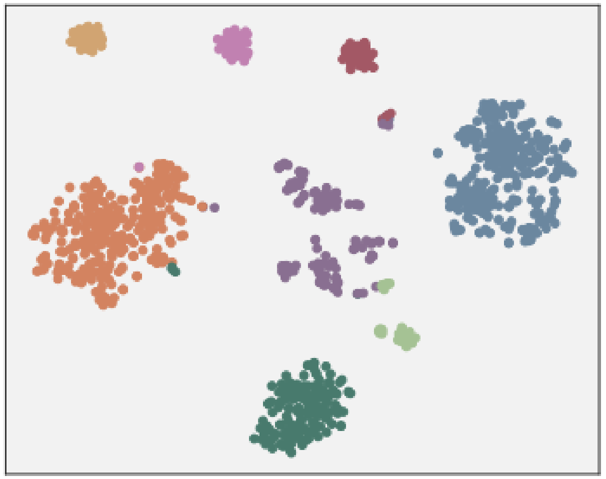
- We add the module one by one to our model
- Our model equipped with all modules performs best in terms of trigger and argument.

Results

- T-SNE visualization for MECL module



(a) w/o MECL

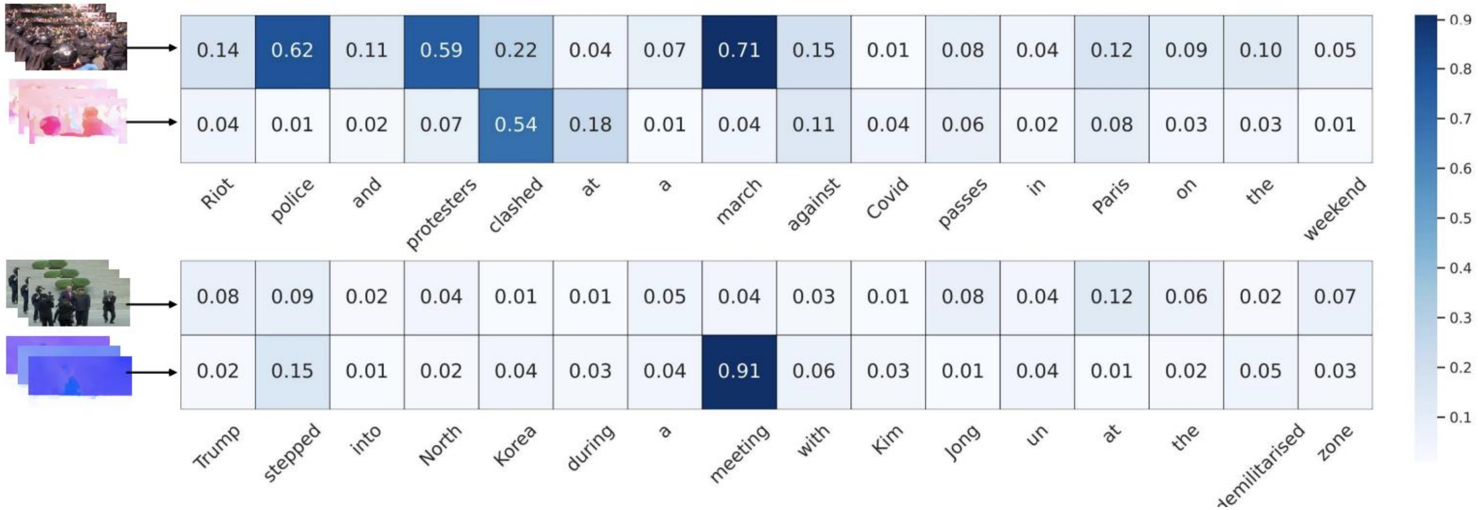


(b) w/ MECL

- The a picture removed MECL from our method.
- Each dot represents one optical flow and each color denotes a specific event trigger.

Results

- Case study on Dual Querying Text



- The first line is video and the second line is optical flow.
- we visualize the attention heatmaps based on the attention scores output by Dual Querying Text.

Conclusions

- We propose a novel framework called TSEE that leverages the motion representations in videos. To the best of our knowledge, we are the first to introduce optical flow features into text-video multimodal event extraction.

- Our proposed modules, MECL and DQT, significantly improve the model performance.

- The experimental results on two benchmark datasets demonstrate the superiority of our framework over the state-of-the-art methods.

Thanks for watching!