multimodal event extraction 方向调研

匡栋栋

Case: “There was the free press in Qatar, Al Jazeera, but ***its’ offices in Kabul and Baghdad*** were ***bombed*** by ***Americans***.”

事件抽取相关概念

Typically, an event in a text is expressed by the following components:

- **Event mention**
- **Event trigger**
- **Event argument**
- **Argument role**

- **P R F1-score**

- Trigger Identification
- Trigger Type Classification
- Argument Identification
- Argument Role Classification

Cross-media Structured Common Space for Multimedia Event Extraction

ACL2020

任务1：MultiMedia Event Extraction $\mathbf{M^2E^2}$

贡献：structured representations and graph-based neural networks for multimedia

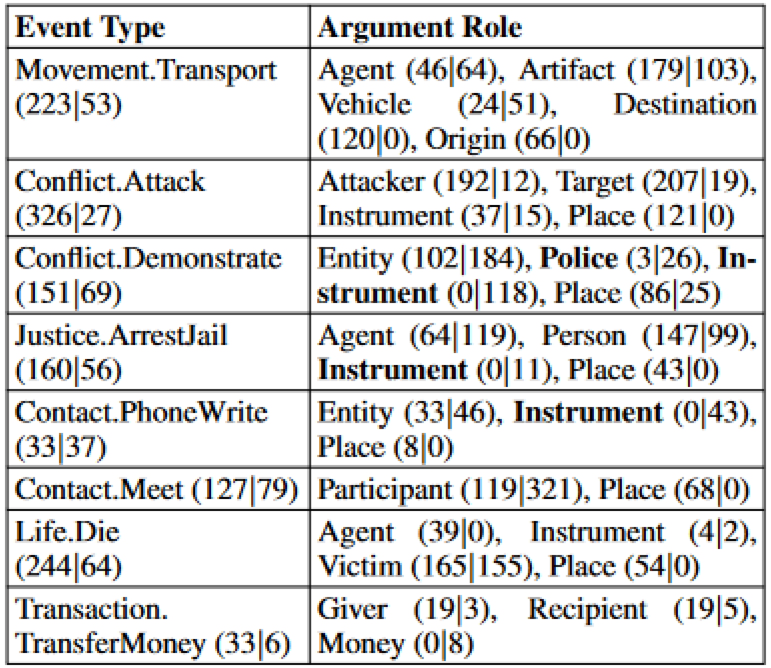
模型：**WASE**

MM documents

$$\left\{\begin{array}{l} \mathcal{M}=\{m_1,m_2,\ldots\} \\ S=\{s_1,s_2,\ldots\} \Rightarrow s=(w_1,w_2,\ldots) \\ \mathcal{T}=\{t_1,t_2,\ldots\} \end{array}\right.$$

$$\left\{\begin{array}{l} e=(y_e,\{w,m\}). \\ a=(y_a,\{t,o\}). \end{array}\right.$$

DataSet categories

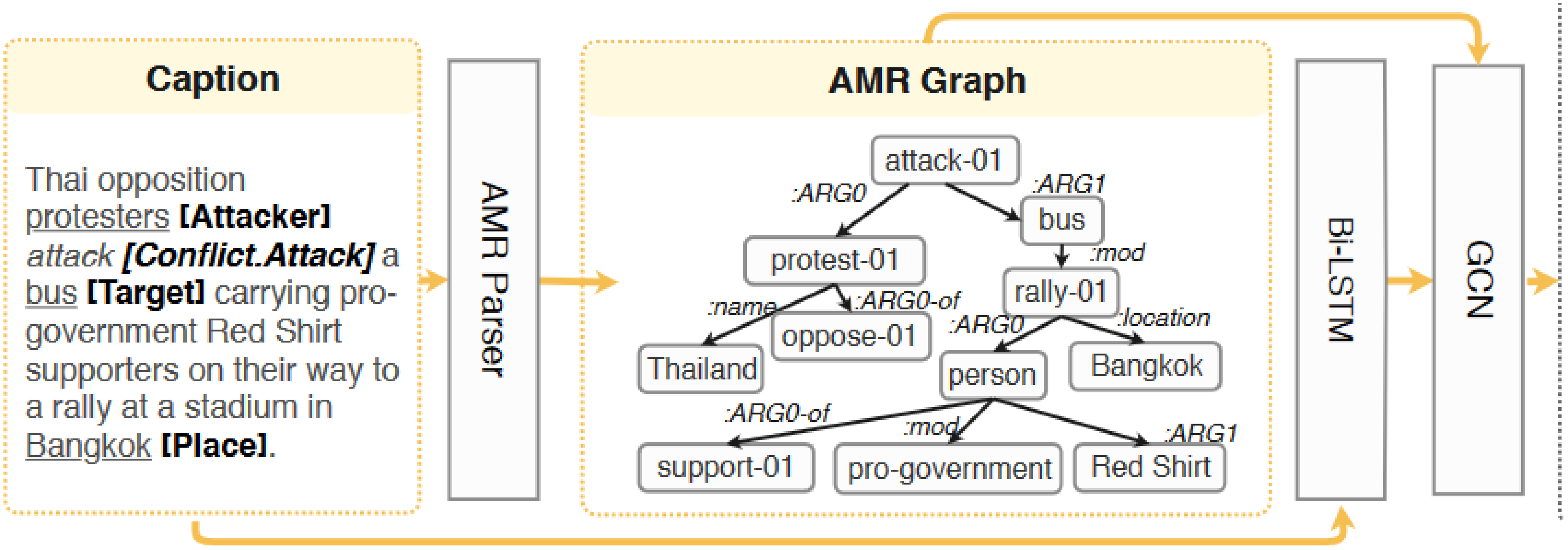


Event Type	**Argument Role**
Movement.Transport (223|53)	Agent (46|64), Artifact (179|103), Vehicle (24|51), Destination (120|0), Origin (66|0)
Conflict.Attack (326|27)	Attacker (192|12), Target (207|19), Instrument (37|15), Place (121|0)
Conflict.Demonstrate (151|69)	Entity (102|184), **Police** (3|26), **Instrument** (0|118), Place (86|25)
Justice.ArrestJail (160|56)	Agent (64|119), Person (147|99), **Instrument** (0|11), Place (43|0)
Contact.PhoneWrite (33|37)	Entity (33|46), **Instrument** (0|43), Place (8|0)
Contact.Meet (127|79)	Participant (119|321), Place (68|0)
Life.Die (244|64)	Agent (39|0), Instrument (4|2), Victim (165|155), Place (54|0)
Transaction. TransferMoney (33|6)	Giver (19|3), Recipient (19|5), Money (0|8)

对文本数据的结构化表征

1.将文本句子通过生成器产生AMR图（在预训练好的GloVe词嵌入、位置嵌入、实体类型嵌入基础上）
2.通过GCN网络来编码产生的图的上下文语义信息

$$\boldsymbol{w}_i^{(k+1)} = f\Big(\sum_{j\in\mathcal{N}(i)} g_{ij}^{(k)}\big(\boldsymbol{W}_{E(i,j)}\boldsymbol{w}_j^{(k)} + \boldsymbol{b}_{E(i,j)}^{(k)}\big)\Big)$$



Thai opposition protesters **[Attacker]** *attack* ***[Conflict.Attack]*** a bus **[Target]** carrying pro-government Red Shirt supporters on their way to a rally at a stadium in Bangkok **[Place]**.

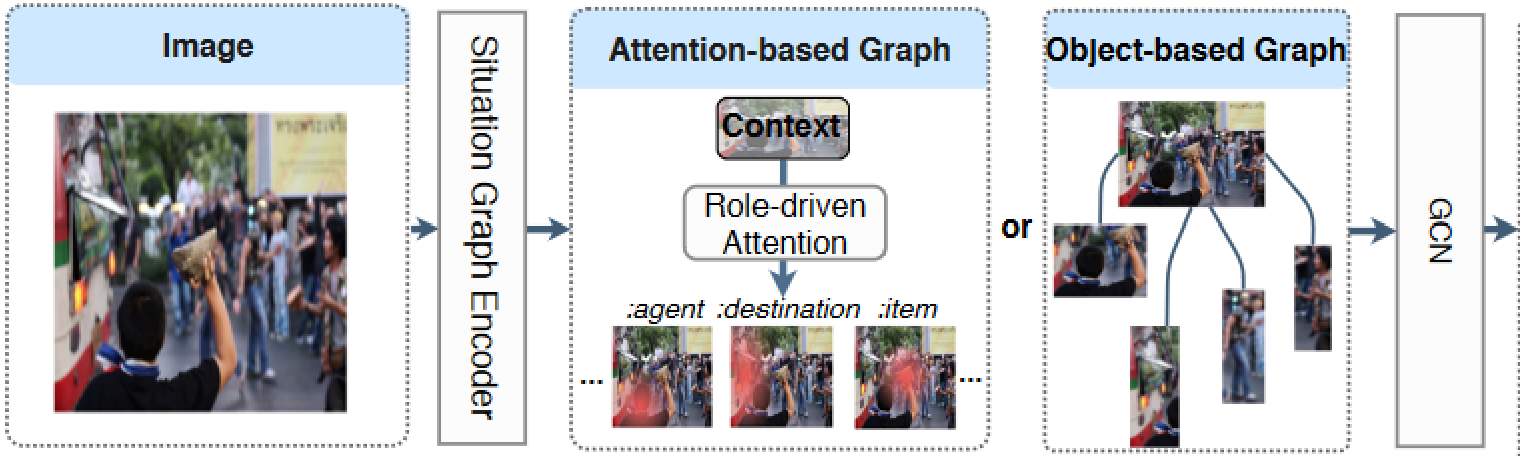
对图像数据的结构化表征

1.通过目标检测模型抽取目标信息和图片特征 via VGG-16
2.通过VGG-16 CNN抽取和所加上的MLP抽取出图片特征和目标检测的实体特征来在语义层面上转化为动作信息verb和实体信息noun
3.将verb和所提取出的noun信息建图
4.通过引入attention机制和利用VGG-16 CNN所抽取的特征图作为依据设置k v q对

$$\hat{\boldsymbol{m}} = \mathrm{MLP}_{\mathrm{m}}(\boldsymbol{m}) \ , \ \hat{\boldsymbol{o}}_i = \mathrm{MLP}_{\mathrm{o}}(\boldsymbol{o}_i).$$

$$P(v|m) = \frac{\exp(\hat{\boldsymbol{m}}\boldsymbol{v})}{\sum_{v'} \exp(\hat{\boldsymbol{m}}\boldsymbol{v}')},$$

$$P(n|o_i) = \frac{\exp(\hat{\boldsymbol{o}}_i\boldsymbol{n})}{\sum_{n'} \exp(\hat{\boldsymbol{o}}_i\boldsymbol{n}')},$$



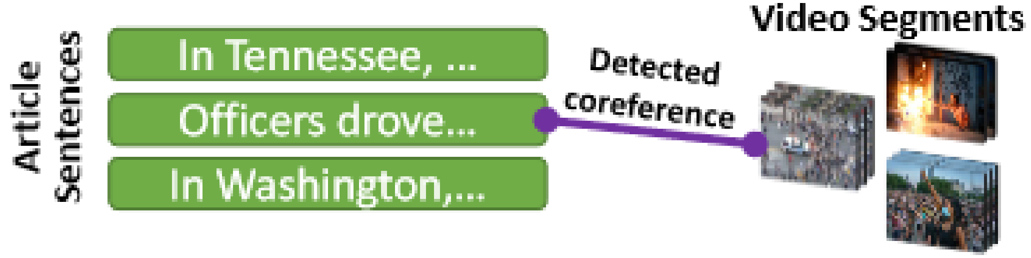
- 未解决的疑问
- GCN
- 对比学习
- CV任务 目标检测
- ....

Joint Multimedia Event Extraction from Video and Article --(EMNLP Findings)

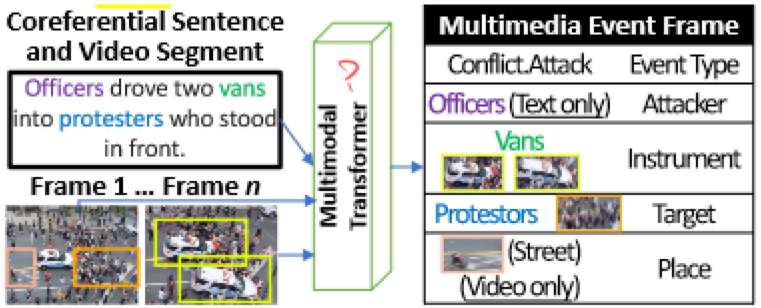
TASK:

$\mathrm{VM}^2\mathrm{E}^2$

Stage1:Multimodal Event Coreference Resolution



Stage 2: Joint Video and Text Event Extraction



Officers drove two vans into protesters who stood in front.

Multimedia Event Frame	
Conflict.Attack	Event Type
Officers (Text only)	Attacker
Vans	Instrument
Protestors	Target
(Street) (Video only)	Place

Method:Multimodal Event Coreference Resolution

思想：将多模态的特征投射到公共空间
学习方式： 自监督学习
产生标签方式：ASR-transcripts **(Miech et al., 2019)**

损失函数1：nce (Jozefowicz et al., 2016)

$$\max_{f,g} \sum_{i=1}^{n} \log \left(\frac{e^{f(x_i)^\top g(y_i)}}{e^{f(x_i)^\top g(y_i)} + \sum_{(x',y')\sim\mathcal{N}_i} e^{f(x')^\top g(y')}} \right)$$

(x, y)表示x一段句子和y一段视频片段f, g均表示投影函数

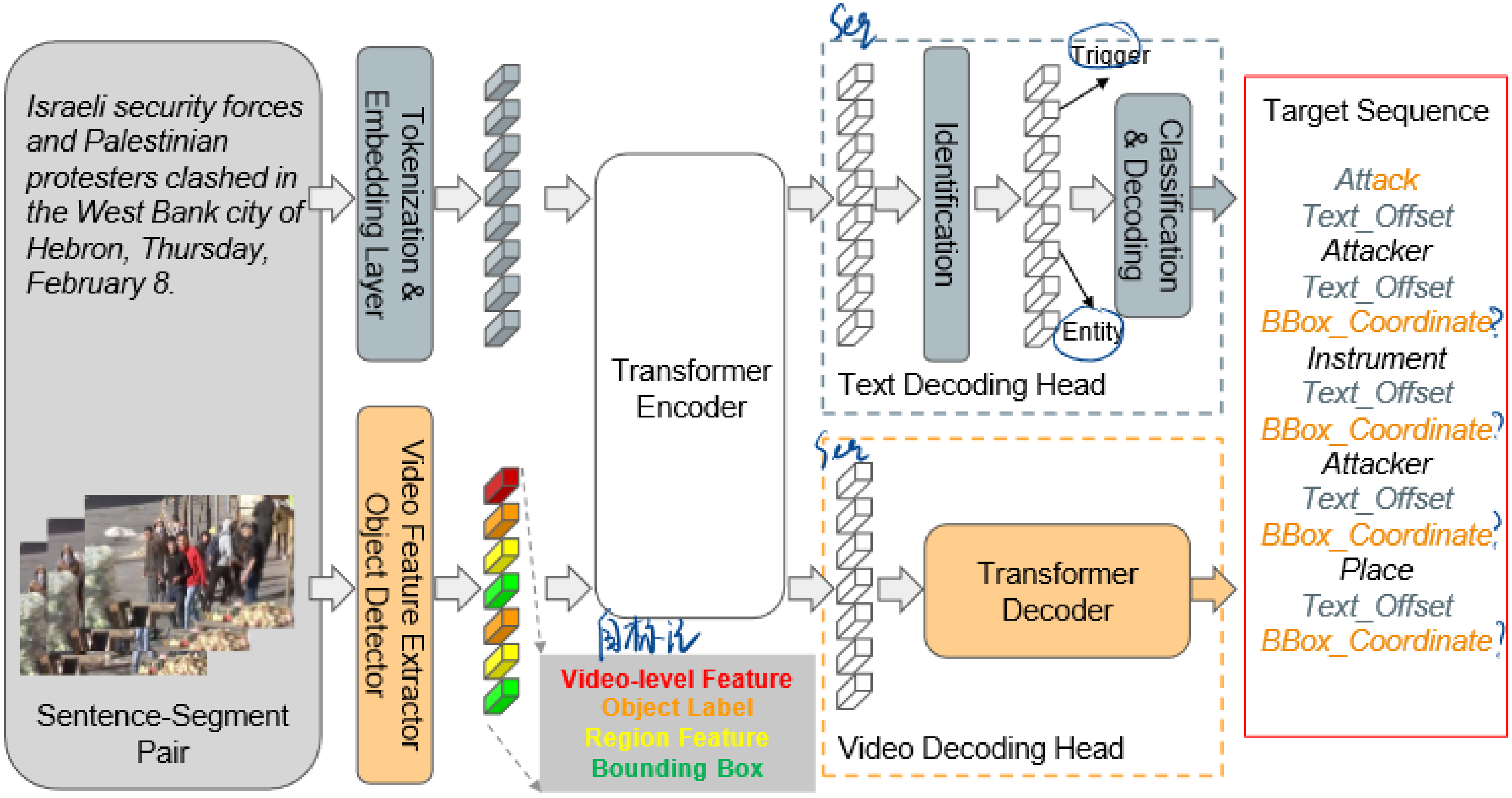
region information (arguments that participate in the event)

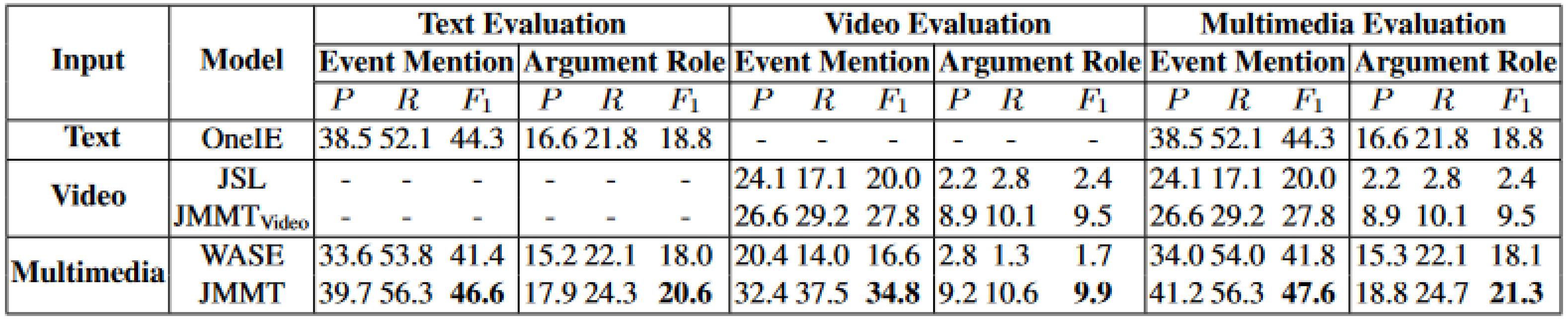
损失函数2：z表示视频片段中的区域信息

$$\max_{f,h} \sum_{i=1}^{n} \log \left(\frac{\sum_{(x,z)\in\mathcal{P}_i} e^{f(x)^\top h(z)}}{\sum_{(x,z)\in\mathcal{P}_i} e^{f(x)^\top h(z)} + \sum_{(x',z')\sim\mathcal{N}_i} e^{f(x')^\top h(z')}} \right)$$

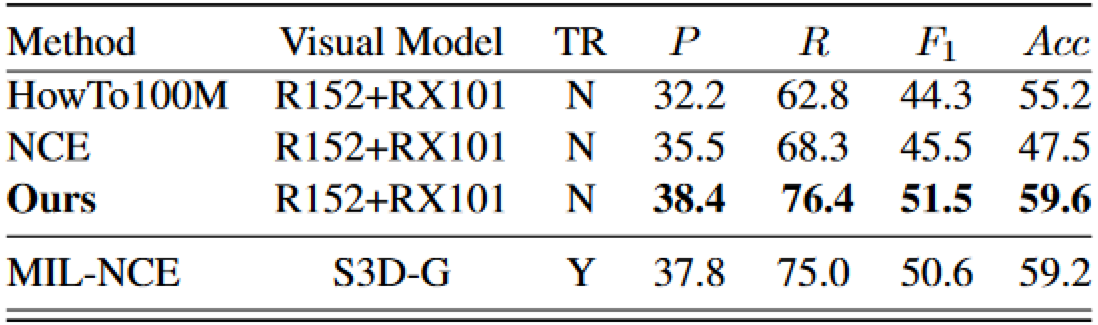
$$\mathcal{L}_{mmcoref} = \mathcal{L}_{NCE} + \mathcal{L}_{MILO} \ .$$

Joint Multimodal Event Extraction and Argument Role Labeling





Input	Model	Text Evaluation						Video Evaluation						Multimedia Evaluation					
		Event Mention			Argument Role			Event Mention			Argument Role			Event Mention			Argument Role		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Text	OneIE	38.5	52.1	44.3	16.6	21.8	18.8	-	-	-	-	-	-	38.5	52.1	44.3	16.6	21.8	18.8
Video	JSL	-	-	-	-	-	-	24.1	17.1	20.0	2.2	2.8	2.4	24.1	17.1	20.0	2.2	2.8	2.4
	$\text{JMMT}_{\text{Video}}$	-	-	-	-	-	-	26.6	29.2	27.8	8.9	10.1	9.5	26.6	29.2	27.8	8.9	10.1	9.5
Multimedia	WASE	33.6	53.8	41.4	15.2	22.1	18.0	20.4	14.0	16.6	2.8	1.3	1.7	34.0	54.0	41.8	15.3	22.1	18.1
	JMMT	39.7	56.3	**46.6**	17.9	24.3	**20.6**	32.4	37.5	**34.8**	9.2	10.6	**9.9**	41.2	56.3	**47.6**	18.8	24.7	**21.3**



Method	Visual Model	TR	P	R	F_1	Acc
HowTo100M	R152+RX101	N	32.2	62.8	44.3	55.2
NCE	R152+RX101	N	35.5	68.3	45.5	47.5
Ours	R152+RX101	N	**38.4**	**76.4**	**51.5**	**59.6**
MIL-NCE	S3D-G	Y	37.8	75.0	50.6	59.2