Cross-Modal Interaction Networks for Query-Based Moment Retrieval in Videos

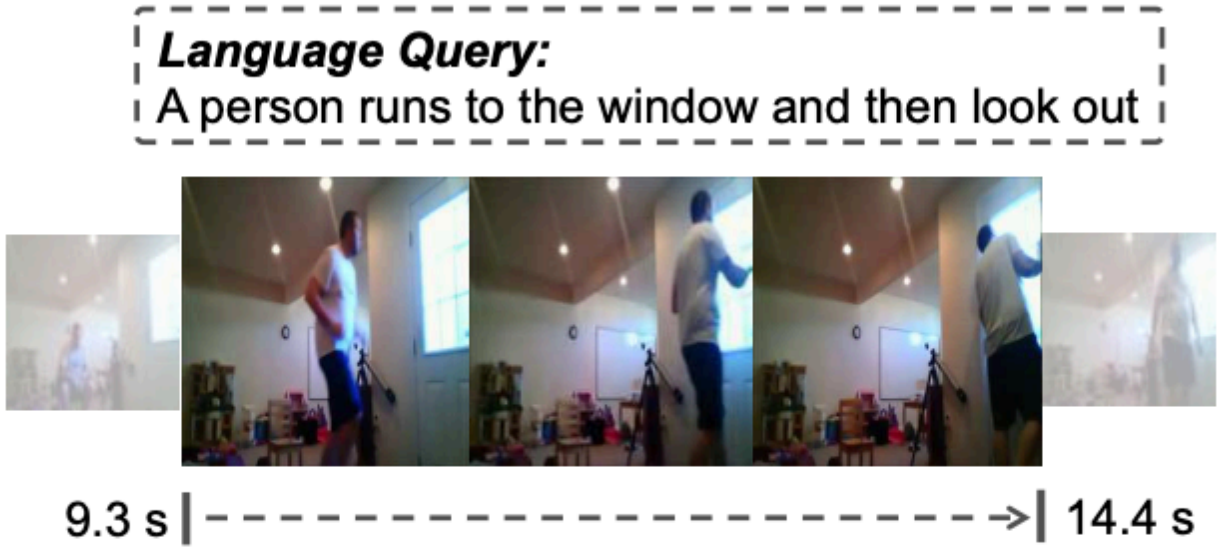
Zhu Zhang, Zhijie Lin, Zhou Zhao, Zhenxin Xiao

SIGIR 2019

Task

Query-Based Moment Retrieval

- 動画の一部分を説明した文章を受けて，それに対応する場面を動画内からローカライズするタスク



TALL

Related works

- MCN(Moment Context networks) (Hendricks, ICCV2017)
  動画特徴量(動画全体のglobal featureとタイムスタンプ内のlocal featureの合成)と言語特徴量の分布が最短距離となるタイムスタンプを推定

- CTRL(Cross-modal Temporal Regression Localizer) (Gao, ICCV2017)
  クリップ動画(複数の固定サイズ)と言語の特徴量をcross-modalに合成
  アライメントスコアを推定し，最大スコアのモーメントの回帰を行ってローカライズ

- ACRN(Attentive Cross-modal Retrieval Network) (Liu, SIGIR2018)
  attentionを組み込んで認識により有効な特徴量を強調
  各モーダル特徴量と融合特徴量を合成
  回帰モデルからアライメントスコアとオフセットを算出し，ローカライズ

- QSPN(Query-Guided Proposal Network) (Xu, AAAI2019)
  動画特徴量と言語特徴量をattentionを組み込んでearly fusionするマルチモデルからclassificationとregressionで学習
  クエリの再生成を補助タスク

Problem

query representation →RNN

- query文の文法的な文構造を捉え切れていない

video representation → CNN+RNN

- 動画のsemanticな関係性の情報を長期的に捉えることができない

Cross-modal Interaction → attention (fusion)

- 一般的なattentionが一層だけ設けられていているのみで十分な相互関係を見ることができない

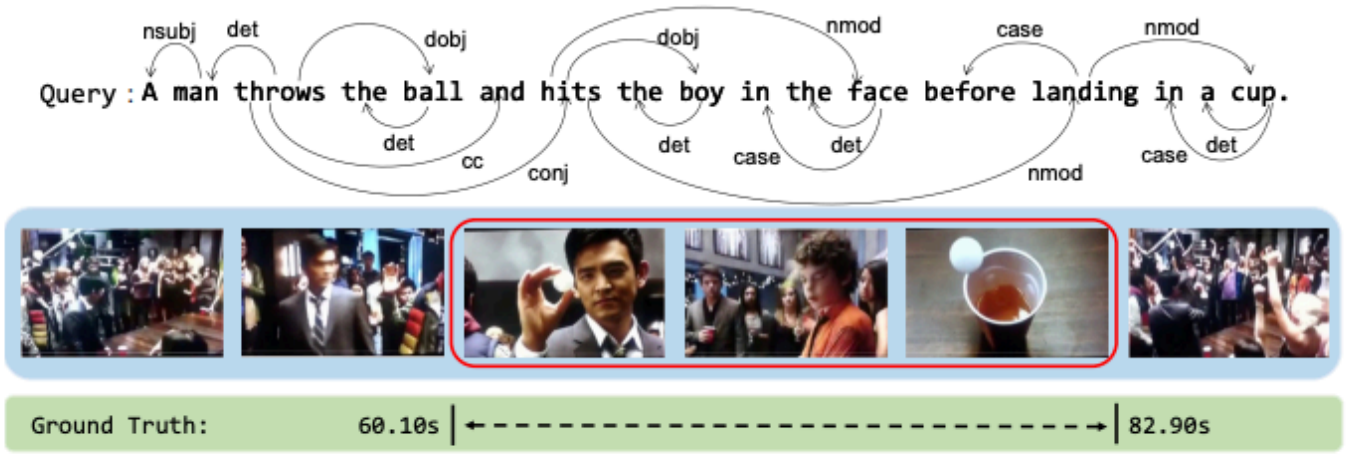


Figure 1: Query-Based Moment Retrieval in Video

Approach

query representation
→ GCN層によってquery文の文法的な構造を捉える

video representation
→ multi-head self-attention層によって動画の長期的でsemanticな関係性を捉える

Cross-modal Interaction → attention (fusion)
→ cross-modal interaction層の多層化によって効率的にモーダル間の統合的な情報を捉える

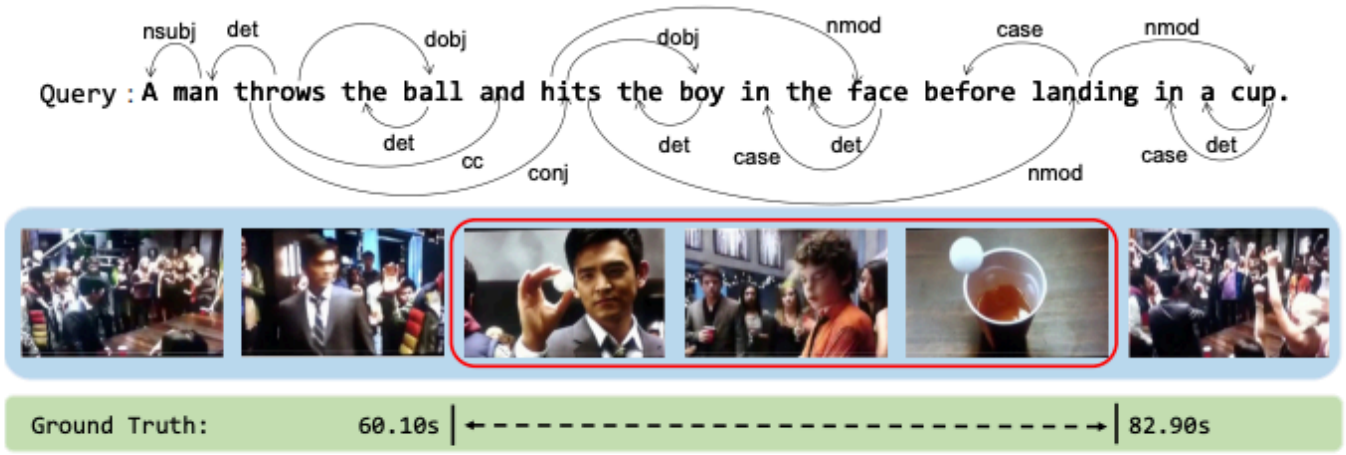
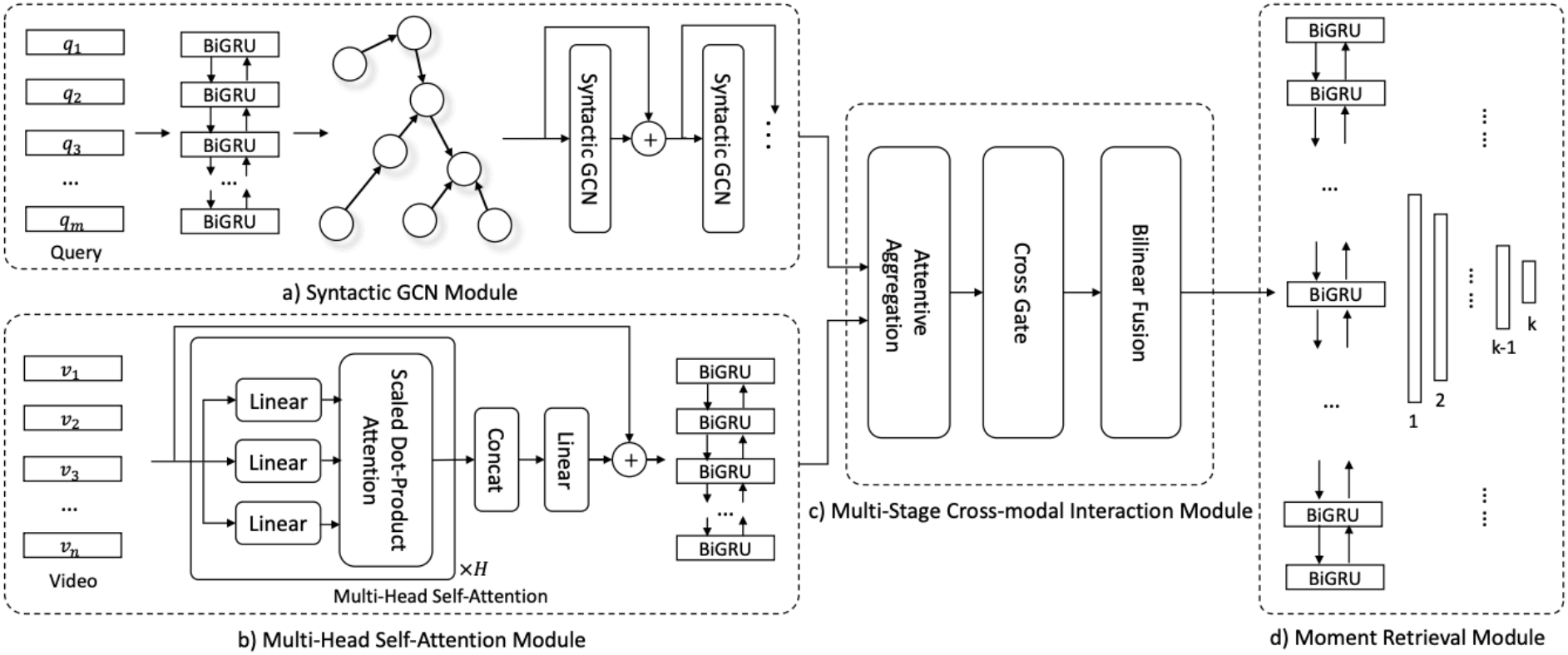


Figure 1: Query-Based Moment Retrieval in Video

Proposed Method



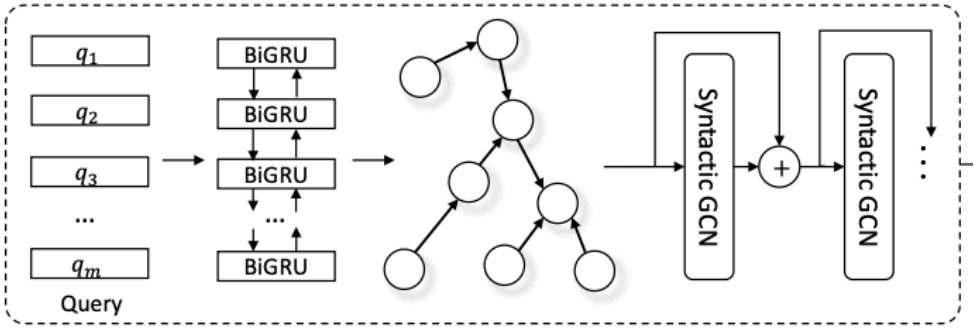
4モジュール構成

a. syntactic GCNモジュール
b. Multi-Head Self-Attentionモジュール
c. Multi-Stage Cross-Modal Interactionモジュール
d. Moment Retrievalモジュール

Proposed Method (Query)

文法的な文構造

入力: query
出力: syntactic-aware query representation



a) Syntactic GCN Module

- 文法的な構造関係の有向グラフを作成

$$\boldsymbol{g}_i^1 = \mathrm{ReLU}\left(\sum_{j\in\mathcal{N}(i)} \boldsymbol{W}^g_{dir(i,j)}\boldsymbol{h}_j^q + \boldsymbol{b}^g_{lab(i,j)}\right)$$

- syntactic GCNを介して文法的な文構造を考慮した特徴量

$$\boldsymbol{o}_i^1 = \boldsymbol{g}_i^1 + \boldsymbol{h}_i^q$$

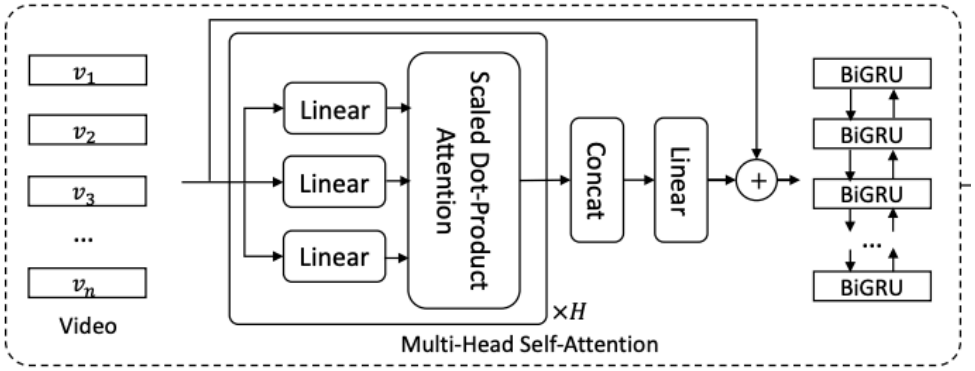
$$\begin{cases} \boldsymbol{g}^1 = \mathrm{synGCN}(\boldsymbol{h}^q), \boldsymbol{o}^1 = \boldsymbol{g}^1 + \boldsymbol{h}^q \\ \cdots \\ \boldsymbol{g}^l = \mathrm{synGCN}\left(\boldsymbol{o}^{l-1}\right), \boldsymbol{o}^l = \boldsymbol{g}^l + \boldsymbol{o}^{l-1} \end{cases}$$

$$\boldsymbol{o}^l = \left(\boldsymbol{o}_1^l, \boldsymbol{o}_2^l, \ldots, \boldsymbol{o}_m^l\right)$$

Proposed Method (Video)

動画の長期的でsemanticな関係性

入力: video
出力: video semantic representation



b) Multi-Head Self-Attention Module

- Multi-Head Self-Attentionによって近接フレーム間だけでなく，離れたフレーム間の関係性も考慮

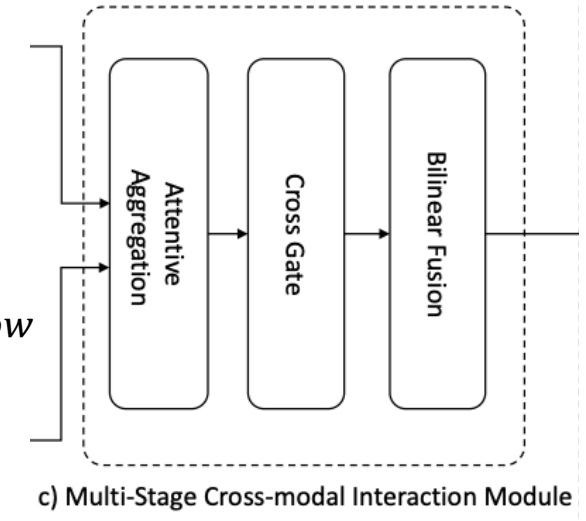
$$\text{Attention}(\overline{\boldsymbol{Q}}, \overline{\boldsymbol{K}}, \overline{\boldsymbol{V}}) = \text{Softmax}\left(\frac{\overline{\boldsymbol{Q}}^{\mathsf{T}}\overline{\boldsymbol{K}}}{\sqrt{d_k}}\right)\overline{\boldsymbol{V}}^{\mathsf{T}}$$

$$\text{MultiHead}(\overline{Q}, \overline{K}, \overline{V}) = \mathbf{W}^{\boldsymbol{O}}\text{Concat}(\text{head}_1, \ldots, \text{head}_\text{H})$$
$$\text{where head}_i = \text{Attention}\left(\boldsymbol{W}_i^Q\overline{\boldsymbol{Q}}, \boldsymbol{W}_i^K\overline{\boldsymbol{K}}, \boldsymbol{W}_i^V\overline{\boldsymbol{V}}\right)$$

$$\boldsymbol{V}^s = \text{MultiHead}(\boldsymbol{V}, \boldsymbol{V}, \boldsymbol{V}) + \boldsymbol{V}$$

Proposed Method (Cross-Modal)

入力: query and video semantic representation
出力: cross-modal representation



c) Multi-Stage Cross-modal Interaction Module

- Attentive Aggregation
  attentionベースで各単語jと各フレームiとの関係性M_{ij}^{row}
  フレーム毎でのqueryの集約表現$\boldsymbol{h}_i^s$

$$\boldsymbol{h}_i^s = \sum_{j=1}^{m} M_{ij}^{row} \boldsymbol{o}_j^l$$

- Cross Gate
  queryの集約表現$\boldsymbol{h}_i^s$とframe semantic表現$\boldsymbol{h}_i^v$との関係性から重要度が重み付け

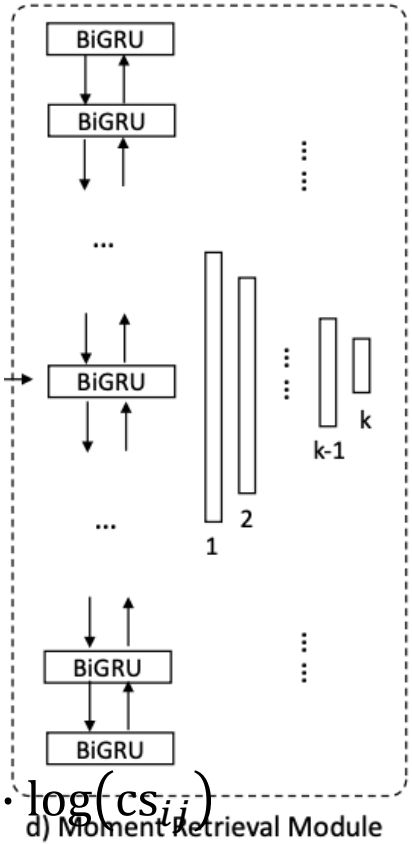
$$\widetilde{\boldsymbol{h}}_i^s = \boldsymbol{h}_i^s \odot \sigma(\boldsymbol{W}^v \boldsymbol{h}_i^v + \boldsymbol{b}^v)$$
$$\widetilde{\boldsymbol{h}}_i^v = \boldsymbol{h}_i^v \odot \sigma(\boldsymbol{W}^s \boldsymbol{h}_i^s + \boldsymbol{b}^s)$$

- Bilinear Fusion
  cross-modal表現$\boldsymbol{f} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)$を獲得

Proposed Method (Localization)

入力: cross-modal representation
出力: moment

- 複数サイズkの固定幅ウィンドウを候補として用意
- 各タイムステップiにおいてウィンドウをセット
- 同時にconfidence score $\boldsymbol{cs}_i$を算出
- これらのモーメントのオフセット(差分)$\widehat{\delta_s}, \widehat{\delta_e}$を推定



d) Moment Retrieval Module

Loss

- Alignment loss
  算出されるアライメントスコconfidence scoreに対して

$$L_{align} = -\frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}\left(1-\mathrm{IoU}_{ij}\right)\cdot\log\left(1-cs_{ij}\right)+\mathrm{IoU}_{ij}\cdot\log\left(\mathrm{cs}_{ij}\right)$$

- Regression loss
  オフセットに対して(δ_s, δ_e):GT, $\left(\widehat{\delta_s}, \widehat{\delta_e}\right)$:predicted

$$L_{reg} = -\frac{1}{N}\sum\left(\mathrm{L1}\left(\delta_s-\widehat{\delta_s}\right)+\mathrm{L1}\left(\delta_e-\widehat{\delta_e}\right)\right)$$

Results

Table 2: Performance Evaluation Results on the ActivityCaption Dataset ($n \in \{1, 5\}$ and $m \in \{0.3, 0.5, 0.7\}$).

Method	R@1 IoU=0.3	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.3	R@5 IoU=0.5	R@5 IoU=0.7
MCN	39.35	21.36	6.43	68.12	53.23	29.70
VSA-RNN	39.28	23.43	9.01	70.84	55.52	32.12
VSA-STV	41.71	24.01	8.92	71.05	56.62	34.52
CTRL	47.43	29.01	10.34	75.32	59.17	37.54
ACRN	49.70	31.67	11.25	76.50	60.34	38.57
QSPN	52.13	33.26	13.43	77.72	62.39	40.78
CMIN	**63.61**	**43.40**	**23.88**	**80.54**	**67.95**	**50.73**

Table 3: Performance Evaluation Results on the TACoS Dataset ($n \in \{1, 5\}$ and $m \in \{0.1, 0.3, 0.5\}$).

Method	R@1 IoU=0.1	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.1	R@5 IoU=0.3	R@5 IoU=0.5
MCN	3.11	1.64	1.25	3.11	2.03	1.25
VSA-RNN	8.84	10.77	4.78	19.05	13.90	9.10
VSA-STV	15.01	10.77	7.56	32.82	23.92	15.50
CTRL	24.32	18.32	13.30	48.73	36.69	25.42
ACRN	24.22	19.52	14.62	47.42	34.97	24.88
QSPN	25.31	20.15	15.23	53.21	36.72	25.30
CMIN	**32.48**	**24.64**	**18.05**	**62.13**	**38.46**	**27.02**

Results

Table 4: Performance Evaluation Results of Ablation Model on the ActivityCaption dataset.

Method	R@1 IoU=0.3	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.3	R@5 IoU=0.5	R@5 IoU=0.7
w/o. GCN	60.12	40.84	21.79	78.23	65.67	45.43
w/o. SA	61.22	41.56	22.36	79.43	66.91	48.12
w/o. CG	60.57	41.21	22.01	78.62	65.99	46.89
w/o. BF	61.32	41.89	22.12	79.27	66.21	47.92
full	**63.61**	**43.40**	**23.88**	**80.54**	**67.95**	**50.73**

Table 5: Performance Evaluation Results of Ablation Model on the TACoS dataset.

Method	R@1 IoU=0.1	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.1	R@5 IoU=0.3	R@5 IoU=0.5
w/o. GCN	30.54	23.22	10.03	57.69	37.12	26.16
w/o. SA	30.21	23.02	16.87	55.54	36.6	25.37
w/o. CG	31.96	23.59	17.47	61.87	38.11	26.79
w/o. BF	32.01	**24.79**	17.61	61.59	38.23	26.75
full	**32.48**	24.64	**18.05**	**62.13**	**38.46**	**27.02**

Results

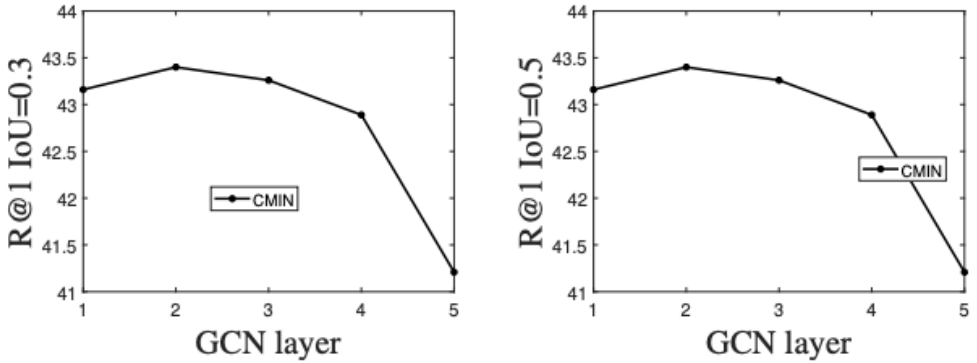


Figure 4: Effect of the Number of Stacked Syntactic GCN layers on the ActivityCaption Dataset.

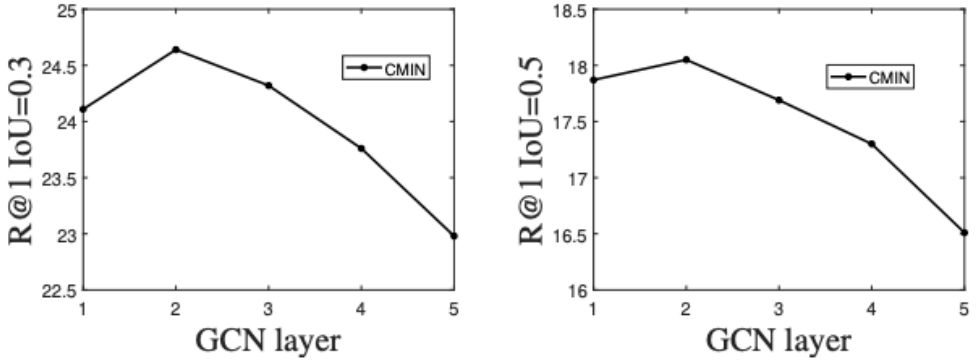


Figure 5: Effect of the Number of Stacked Syntactic GCN layers on the TACoS Dataset.

Syntactic GCN layerの層数による影響

Results

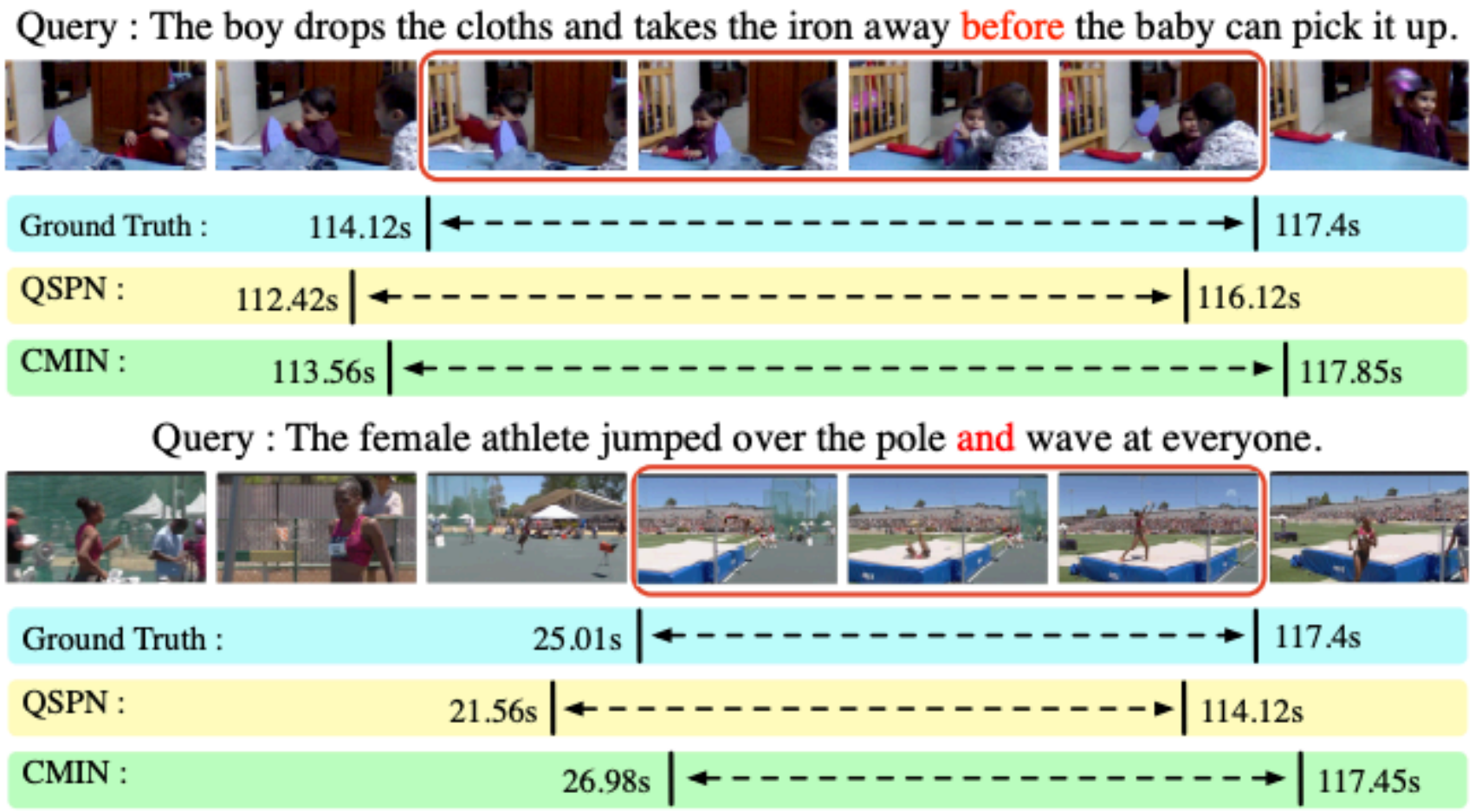


Figure 6: Examples on the ActivityCaption dataset.

Results

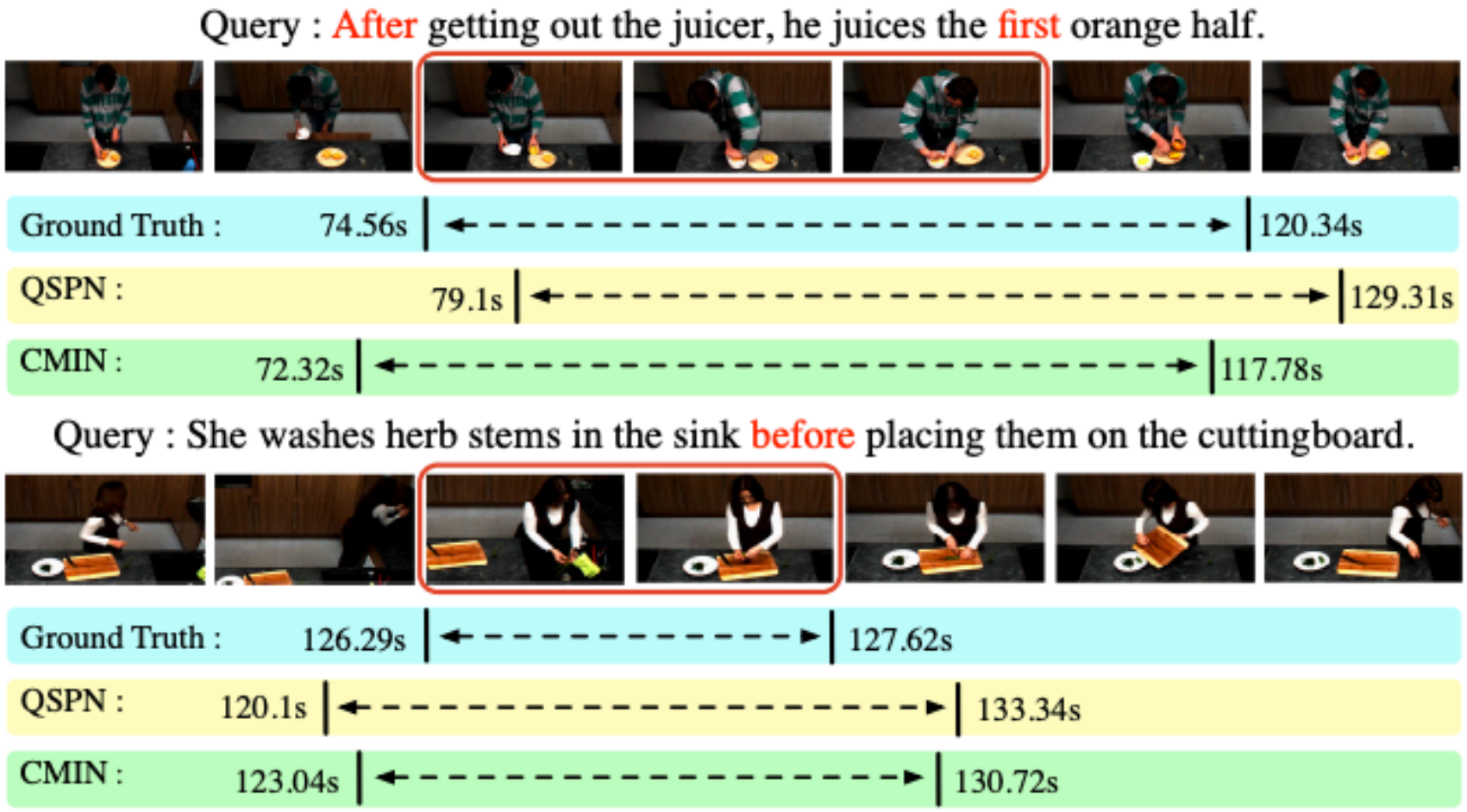


Figure 7: Examples on the TACoS dataset.

Results

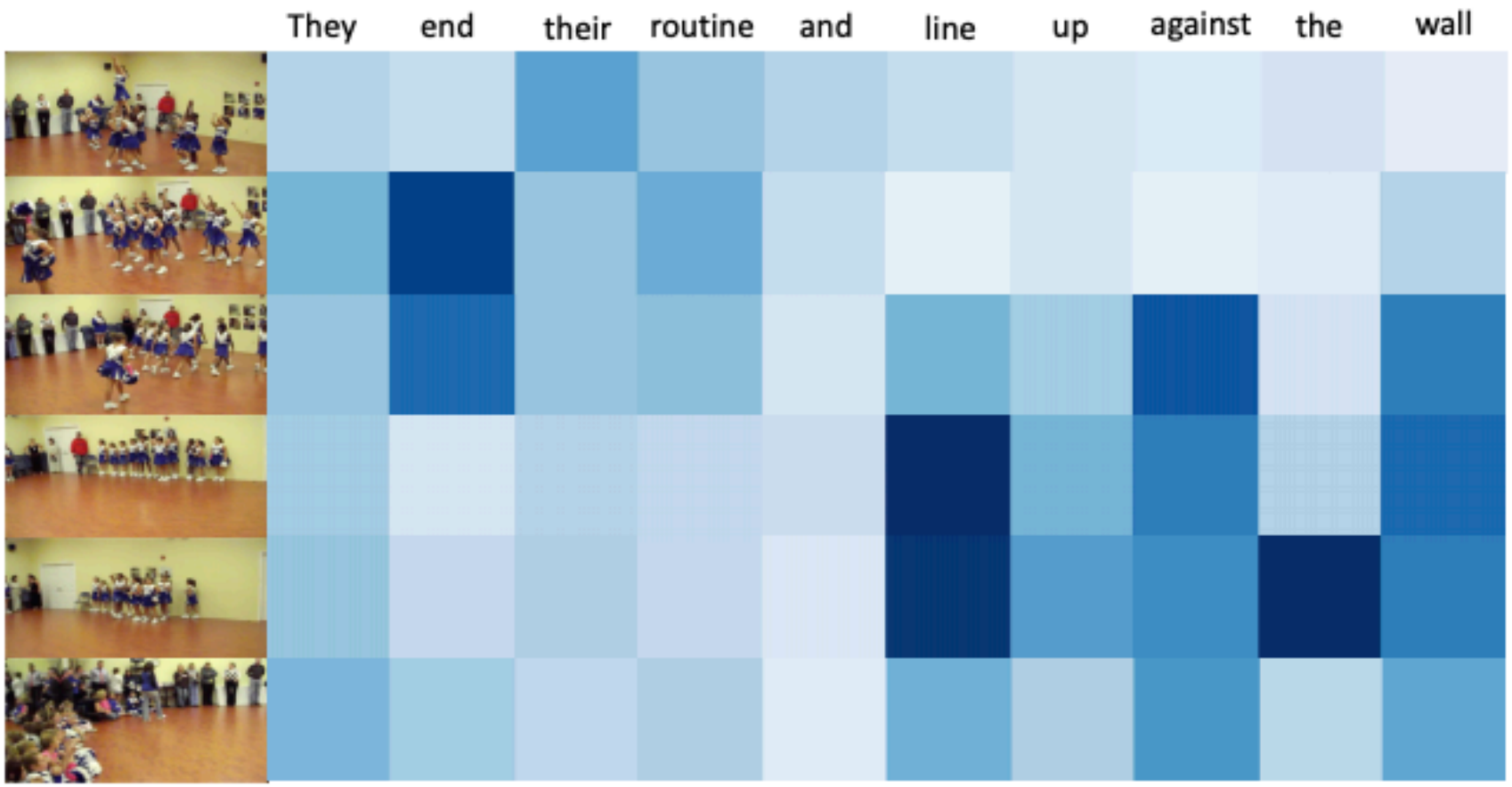


Figure 8: The Video-to-Query Attention Results in the Multi-Stage Cross-Modal Interaction Module