Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos

Yitian Yuan[1], Lin Ma[2], Jingwen Wang[2], Wei Liu[2], Wenwu Zhu[1]
[1]Tsinghua University; [2]Tencent AI Lab
yyt18@mails.tsinghua.edu.cn

website: https://github.com/yytzsy/SCDM

1 Temporal Sentence Grounding in Videos

Sentence query: She pins her hair followed by curling it along the sides.

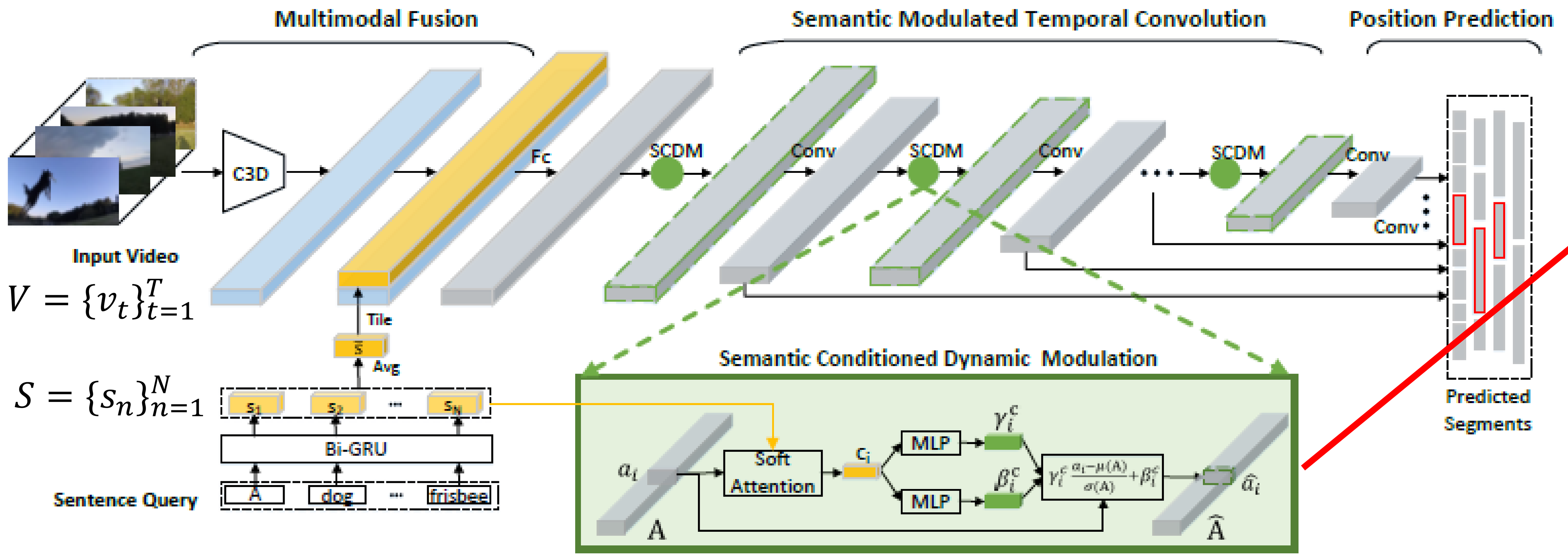
Given an untrimmed video and a natural sentence query, the task aims to identify *the start and end timestamps* of one specific video segment, which contains activities of interest semantically corresponding to the given sentence query.

2 Motivation: Leverage sentence semantics to correlate video contents over time

Previous methods mainly focus on semantically matching sentences and individual video segments or clips, while neglect *the important guiding role of sentences to help correlate and compose video contents over time.*

Sentence query: The woman takes the book across the room to read it on the sofa.

3 Proposed Model: Use sentence information to dynamically modulate feature normalization procedure in temporal convolution architecture



1. Multimodal Fusion

$$f_t = ReLU(W^f(v_t \| \bar{s}) + b^f)$$

2. Semantic Modulated Temporal Convolution

(1) Basic Temporal Convolution Network:

convolution kernel: $Conv(\theta_k, \theta_s, d_h)$

feature map: $A_k = \{a_{k,i}\}_{i=1}^{T_k}$

(2) Semantic Conditioned Dynamic Modulation

$$\beta_i^n = \text{softmax}(w^T \tanh(W^s s_n + W^a a_i + b))$$

$$c_i = \sum_{n=1}^{N} \beta_i^n s_n \qquad \gamma_i^c = \tanh(W^\gamma c_i + b^\gamma) \qquad \beta_i^c = \tanh(W^\beta c_i + b^\beta)$$

$$\hat{a}_i = \gamma_i^c \frac{a_i - \mu(A)}{\sigma(A)} + \beta_i^c$$

3. Position Prediction

convolution prediction: $(p^{over}, \Delta c, \Delta w)$

$$\varphi^c = \mu^c + \alpha^c \cdot \mu^w \cdot \Delta c$$

$$\varphi^w = \mu^w \cdot \exp(\alpha^w \cdot \Delta w)$$

Model training: *Training loss = center offset loss + width offset loss + overlap prediction loss*

5 Results & Illustration

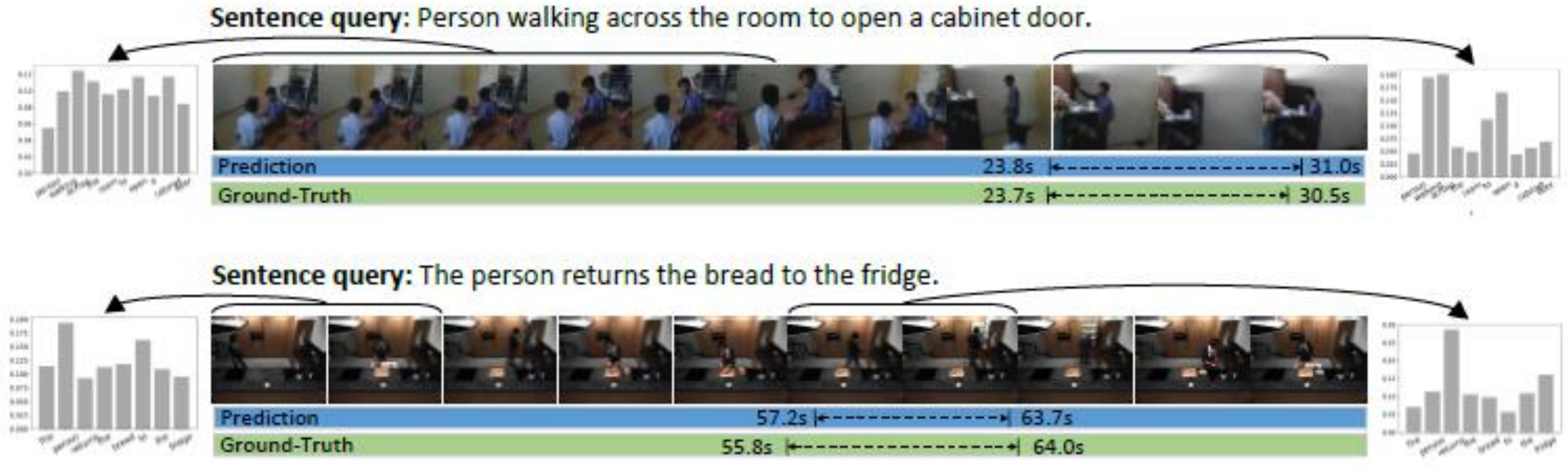
Table 1: Performance comparisons on the TACoS and Charades-STA datasets (%).

Method	TACoS				Charades-STA			
	R@1, IoU@0.3	R@1, IoU@0.5	R@5, IoU@0.3	R@5, IoU@0.5	R@1, IoU@0.5	R@1, IoU@0.7	R@5, IoU@0.5	R@5, IoU@0.7
CTRL [6]	18.32	13.30	36.69	25.42	23.63	8.89	58.92	29.52
MCF [23]	18.64	12.53	37.13	24.73	-	-	-	-
ACRN [14]	19.52	14.62	34.97	24.88	-	-	-	-
SAP [3]	-	18.24	-	28.11	27.42	13.36	66.37	38.15
ACL[7]	24.17	20.01	**42.15**	30.66	30.48	12.20	64.84	35.13
TGN [2]	21.77	18.90	39.06	31.02	-	-	-	-
Xu et al. [24]	-	-	-	-	35.60	15.80	79.40	45.40
MAN [26]	-	-	-	- -	46.53	22.72	**86.23**	53.72
Ours-SCDM	**26.11**	**21.17**	40.16	**32.18**	**54.44**	**33.43**	74.43	**58.08**

Table 2: Performance comparisons on the ActivityNet Captions 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
TGN [2]	45.51	28.47	-	57.32	43.33	-
Xu et al. [24]	45.30	27.70	13.60	75.70	59.20	38.30
Ours-SCDM	**54.80**	**36.75**	**19.86**	**77.29**	**64.99**	**41.53**

Qualitative examples



t-SNE projections of temporal feature maps yielded by models Ours-w/o-SCDM and Ours-SCDM