Iteratively Refining Transformer Decoder for Temporal Video Grounding

Ziyi Chen, Xiao Liang and Te Tao

Data science and Information Technology, Tsinghua Berkeley Shenzhen Institute

ABSTRACT

Temporal video grounding is a crucial task in vision-language learning, which aims to retrieve a segment from an untrimmed video semantically corresponding to a natural language query sentence. The main insight in this work is to effectively extract and aggregate multi-modal contextual features, followed by classifiers or regressors as the final moment localizer. In this project we propose a novel transformer-based model with language guidance to update the predicted segment in decoder layers, which can help correct the initial prediction step by step. Extensive experiments on the main-stream datasets have demonstrated the effectiveness of our proposed method.

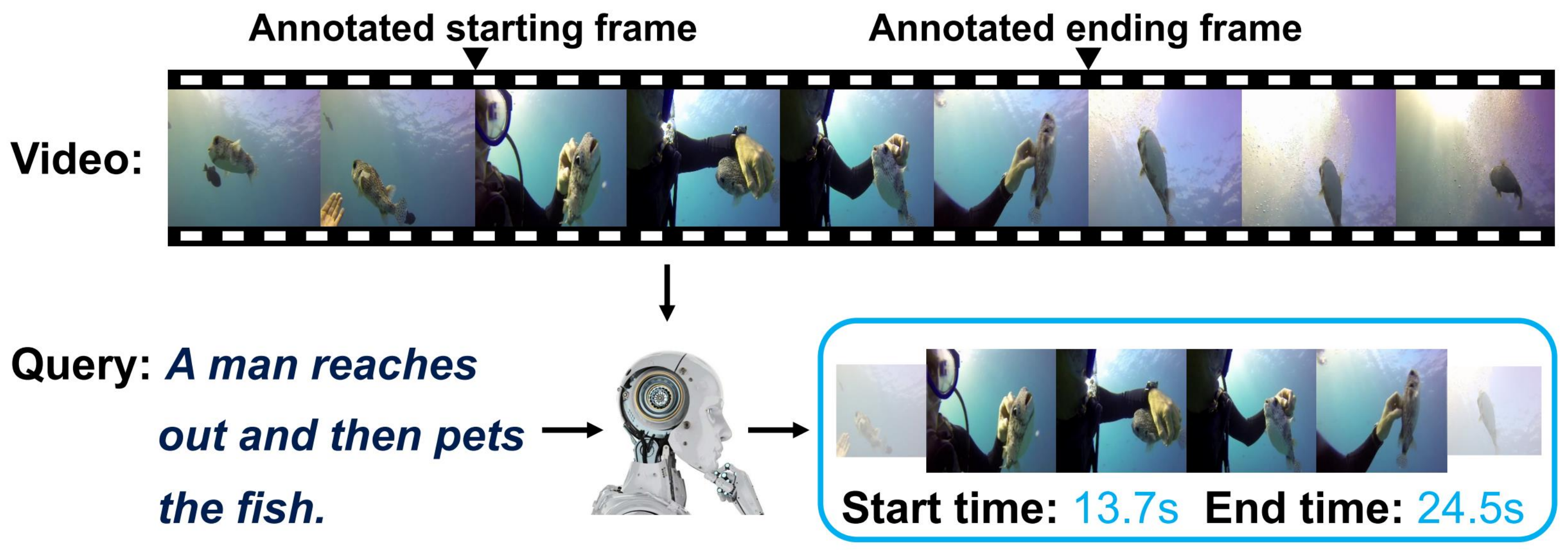
Annotated starting frame
Annotated ending frame
Video:
Query: *A man reaches out and then pets the fish.*
Start time: 13.7s End time: 24.5s

Fig1. An illustrative example of the TVG task. Given a video and a sentence query, the TVG task aims to identify the starting and ending time of the video segment relevant to the query.

INTRODUCTION

With the development of multimedia, video understanding is catching increasing attention in computer vision community. In order to accomplish that people can quickly search the desired segments from the whole video, researchers in both computer vision and natural language processing communities have proposed the Temporal Video Grounding (TVG) task. As shown in Fig 1, given an untrimmed video and a sentence query, this task aims to predict the specific video segment relevant to the activities described in the sentence. Existing TVG methods can be simply divided into two categories:

In this project, we proposed a transformer-based model that is more suitable for TVG tasks. Specifically, we first extract video features with a 3D backbone and sentence features with BERT model. Then we simply fuse the multi-modal features through the transformer encoder, after which we apply the query sentence to generate a series of continuously updated anchor segments that guide temporal localization in the transformer decoder.

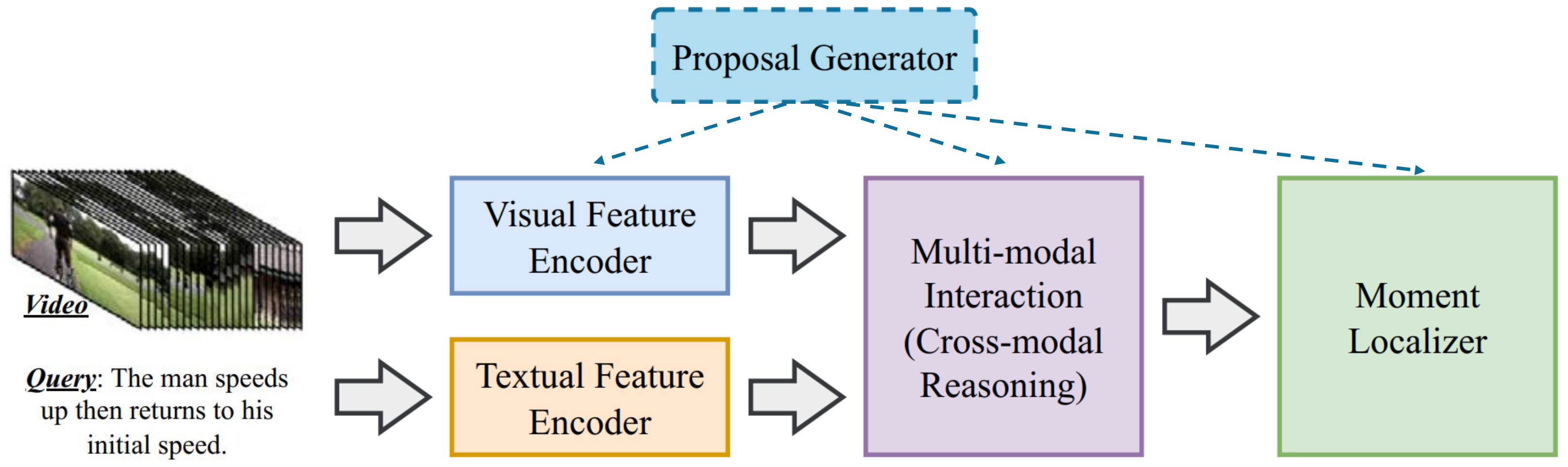
Proposal Generator
Video
Query: The man speeds up then returns to his initial speed.
Visual Feature Encoder
Textual Feature Encoder
Multi-modal Interaction (Cross-modal Reasoning)
Moment Localizer

Fig2. A general pipeline for temporal sentence grounding in videos[1].

METHOD

Our proposed baseline model consists of a multimodal encoder and a language-guided decoder. The output of the encoder part is the fused video and query features attended to each other, which can provide the aligned information for both modalities. After we got the multimodal features from the transformer encoder, we first split the multimodal sequence into video and query features. A cross-attention layer is applied to generate the segment representations and the local priors from the query-specific multi-modal representations.

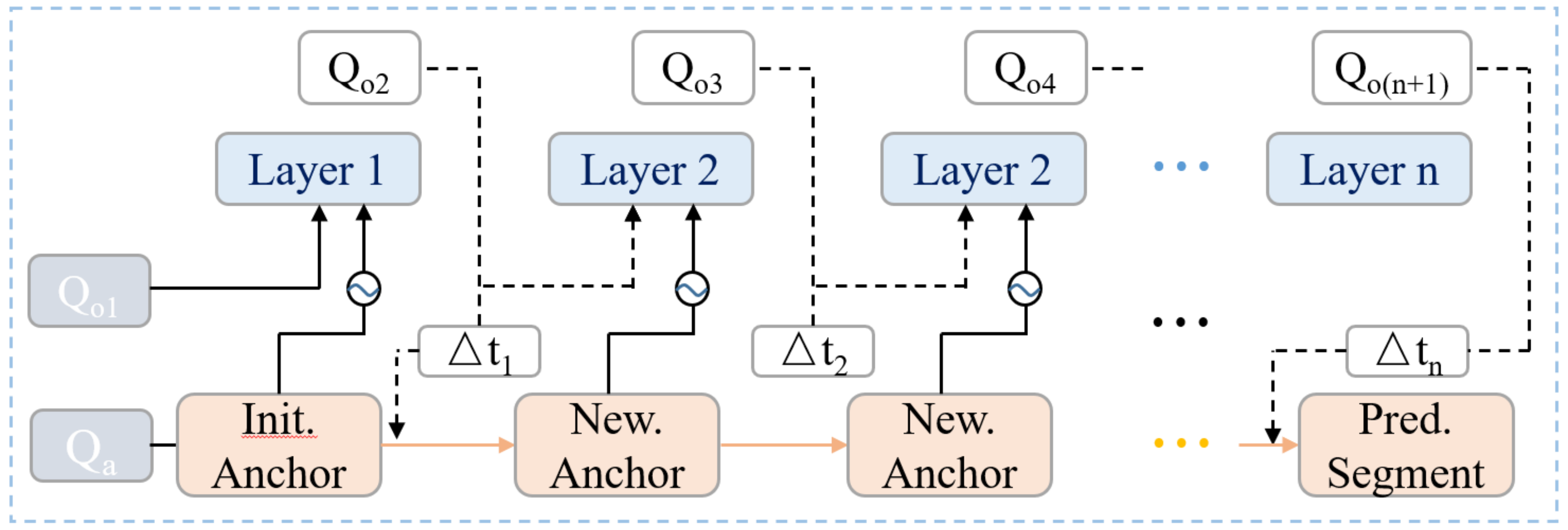

Fig3. Language-guided Iteratively Refining Decoder

To provide language guidance for segment localization, in the decoder layers we update the predicted segment layer by layer to refine the precision of localization. Specifically, we compute the deviation of the previous predicted segment from the actual segment with multimodal features and add it to the predicted segment in every decoder layer. Through the above refinement in the decoder, the predicted segment would be closer to the actual localization layer by layer. Finally, the last layer of the transformer decoder can directly output the final prediction for the described segment.

Loss: we use the sum of the Generalized IoU loss and the L1 loss for training our model, which is defined as:

$$\mathcal{L} = \lambda_0 \mathcal{L}_{giou}\,(\mathrm{S}, \tilde{\mathrm{S}}) + \lambda_1 \parallel \mathrm{S} - \tilde{\mathrm{S}} \parallel_1$$

where S is the ground truth and $\tilde{S}$ is our prediction, λ_0 and λ_1 are the balancing weights.

RESULTS

Method	Venue	Charades-STA		ActivityNet Captions	
		IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5
CTRL[2]	ICCV'17	39.86	23.63	36.64	26.60
ROLE[3]	MM'18	25.26	12.12	-	-
SLTA[4]	ICMR'19	38.96	22.81	39.04	29.24
CBP[5]	AAAI'20	54.70	**35.60**	54.30	35.76
Ours	-	**55.33**	35.46	**57.62**	**37.20**

Tab1. Results of supervised methods on TVG task.

CONCLUSIONS

We have presented a Transformer-based decoding method for TVG task, which exploits a series of language-guided anchor segments as temporal cues for guiding context pooling in a Transformer decoder. Extensive experiments on mainstream benchmarks demonstrate its competitive performance with respect to existing methods.

REFERENCES

[1] Zhang, H., Sun, A., Jing, W., & Zhou, J.T. (2022). Temporal Sentence Grounding in Videos: A Survey and Future Directions.

[2] J. Gao, C. Sun, Z. Yang, and R. Nevatia, “Tall: Temporal activity localization via language query,” in ICCV, 2017.

[3] M. Liu, X. Wang, L. Nie, Q. Tian, B. Chen, and T.-S. Chua, “Crossmodal moment localization in videos,” in ACM MM, 2018.

[4] B. Jiang, X. Huang, C. Yang, and J. Yuan, “Cross-modal video moment retrieval with spatial and language-temporal attention,” in ACM ICMR, 2019.

[5] S. Chen and Y.-G. Jiang, “Semantic proposal for activity localization in videos via sentence query,” in AAAI, vol. 33, 2020.