Cascaded Decoding and Multi-Stage Inference for Spatio-Temporal Video Grounding

Li Yang
Peixuan Wu
li.yang@nlpr.ia.ac.cn
wupeixuan2022@ia.ac.cn
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
School of Artificial Intelligence, University of Chinese Academy of Sciences

Chunfeng Yuan*
Bing Li
cfyuan@nlpr.ia.ac.cn
bli@nlpr.ia.ac.cn
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

Weiming Hu
wmhu@nlpr.ia.ac.cn
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
School of Artificial Intelligence, University of Chinese Academy of Sciences
CAS Center for Excellence in Brain Science and Intelligence Technology

ABSTRACT

Human-centric spatio-temporal video grounding (HC-STVG) is a challenging task that aims to localize the spatio-temporal tube of the target person in a video based on a natural language description. In this report, we present our approach for this challenging HC-STVG task. Specifically, based on the TubeDETR framework, we propose two cascaded decoders to decouple spatial and temporal grounding, which allows the model to capture respective favorable features for these two grounding subtasks. We also devise a multi-stage inference strategy to reason about the target in a coarse-to-fine manner and thereby produce more precise grounding results for the target. To further improve accuracy, we propose a model ensemble strategy that incorporates the results of models with better performance in spatial or temporal grounding. We validated the effectiveness of our proposed method on the HC-STVG 2.0 dataset and won second place in the HC-STVG track of the 4th Person in Context (PIC) workshop at ACM MM 2022.

CCS CONCEPTS

- **Information systems** → Multimedia and multimodal retrieval;
- **Computing methodologies** → Computer vision.

KEYWORDS

Spatio-temporal video grounding (STVG), Transformer, DETR

ACM Reference Format:
Li Yang, Peixuan Wu, Chunfeng Yuan, Bing Li, and Weiming Hu. 2022. Cascaded Decoding and Multi-Stage Inference for Spatio-Temporal Video Grounding. In *Proceedings of the 4th Person in Context Workshop and Challenge (PIC '22), October 14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3552455.3555814

1 INTRODUCTION

Spatio-temporal video grounding (STVG) aims to find the spatial location and temporal scope of the target object in a video given a natural language query, which is important for associating the linguistic expression with video understanding. On the basis of the STVG task, Tang et al. [17] introduce the human-centric spatio-temporal video grounding (HC-STVG) task, further focusing on humans as the targets in the video grounding process. This HC-STVG task has more practical applications in the real world, as humans are often the focus of video analysis and comprehension. To address this challenging problem, the HC-STVG challenge in the 4th Person in Context (PIC) workshop [18] is held in conjunction with ACM MM 2022.

In this report, we employ TubeDETR [22] as the basic framework to build our method for this HC-STVG challenge. Unlike the previous methods [17, 30] that rely on the pre-generated object proposals or tube proposals for the STVG task, TubeDETR [22] employs a transformer-based encoder-decoder structure to directly infer the spatio-temporal tube of the target object. The encode extracts the features of the input video frames and sentence, and then perform visual-textual feature fusion. Based on the features of two modalities, the decoder of TubeDETR establishes T time queries for T sampled frames, and conducts temporal self-attention and time-aligned cross-attention to jointly model the spatial and temporal information of the target. Evaluated on the challenging HC-STVG [17] and VidSTG [30] benchmarks, TubeDETR outperforms the previous methods by a significant margin.

Despite its success, TubeDETR utilizes a single decoder for both spatial and temporal grounding, which may cause conflicts and difficulties in learning since the two subtasks usually have different information to focus on. The fixed video frame sampling strategy (*e.g.* 100 frames for each video) may also be coarse for the target and could hamper the accuracy of the estimated bounding box sequences, especially for targets appearing for a short time in the videos. Thus, to address these issues, we propose several methodological improvements based on TubeDETR and build a stronger model for HC-STVG reasoning.

In our method, based on the features encoded by the video-text feature encoder of TubeDETR, we propose two cascaded decoders

*Corresponding author.

PIC '22, October 14, 2022, Lisboa, Portugal.

ACM ISBN 978-1-4503-9489-5/22/10…$15.00
https://doi.org/10.1145/3552455.3555814

for spatial and temporal video grounding, respectively. This decouples the reasoning process of the two grounding subtasks, allowing the model to focus on their respective related features for estimation. In addition, we devise a multi-stage inference strategy that enables the model to resample the key video frames focused on the target and produce more accurate grounding results in a coarse-to-fine manner. We also employ the model ensemble strategy to further improve the grounding accuracy. We conduct experiments on the dataset provided by the HC-STVG challenge and validate the effectiveness of our method.

2 RELATED WORK

2.1 Visual Grounding

Visual grounding aims to localize the target object in an image given a natural language expression, without the need for temporal grounding as in spatio-temporal video grounding. Existing methods usually extend a pre-trained object detector [5, 14, 15, 23] to address this task. Two-stage methods [7, 19, 27] first generate a set of object proposals, and then compare them to the language query to select the best match. One-stage methods [9, 25, 26] fuse visual and linguistic features and then generate dense detections with scores, where the top-ranked one is selected as the localization result. Some recent methods [3, 24] develop transformer-based models to directly identify the target object from the image, and achieve leading performance without performing ranking on the candidates.

2.2 Temporal Grounding

The goal of temporal grounding is to locate the relevant video moment corresponding to a given language query. Earlier approaches [1, 4, 10, 11] mainly use a sliding window based approach to generate multiple temporal candidates and then select the top-ranked one. TGN [2] proposes to exploit fine-grained frame-by-word interactions between the video and sentence to score the temporal candidates of multiple scales. Xu et al. [21] propose to fuse the query information with the fine-grained video clips to generate query-specific candidate segments. MAN [28] develops an iterative graph adjustment network to model the temporal relations of video moments for better moment alignment. 2D-TAN [29] and MMN [20] establish a two-dimensional temporal map to represent various video moments and model their temporal relationships.

2.3 Spatio-Temporal Video Grounding

Unlike visual grounding that identifies a target object in an image or temporal grounding that locates a specific video moment, spatio-temporal video grounding (STVG) seeks to pinpoint the spatio-temporal tube of the target object. This task is a combination of spatial and temporal localization for the target in the video, which requires a finer-grained understanding of the video information.

Zhang et al. [30] first introduce this STVG task and propose a spatio-temporal graph reasoning network (STGRN) to model the spatial and temporal relations of the detected object proposals for target tube retrieving. While STGRN does not need pre-generated tube proposals, it depends on a pre-trained object detector to generate a set of object proposals for each video frame. Another work STGVT [17] first forms spatio-temporal tube proposals by linking the detected object proposals in consecutive frames, and then employs a visual transformer to learn cross-modal features and perform tube-description matching. Recently, one-stage methods have also been developed without the use of pre-generated tube or object proposals. STVGBert [16] extends the pre-trained VilBERT model [13] to this task, and directly predicts the bounding box sequence as well as the start and end frames from the modeled cross-modal features. The recent one-stage method TubeDETR [22] proposes a space-time transformer decoder with time queries to perform spatial and temporal localization on the sampled video frames, outperforming prior methods by a large margin. In this report, we use TubeDETR as a strong basic framework and make methodological improvements to further enhance the performance on human-centric spatio-temporal video grounding (HC-STVG).

3 OUR METHOD

In Section 3.1, we first give an overview of our baseline framework, TubeDETR. Next, based on TubeDETR, we describe in detail our proposed methods, including spatio-temporal cascaded decoders (Section 3.2), multi-stage inference (Section 3.3), and model ensemble (Section 3.4).

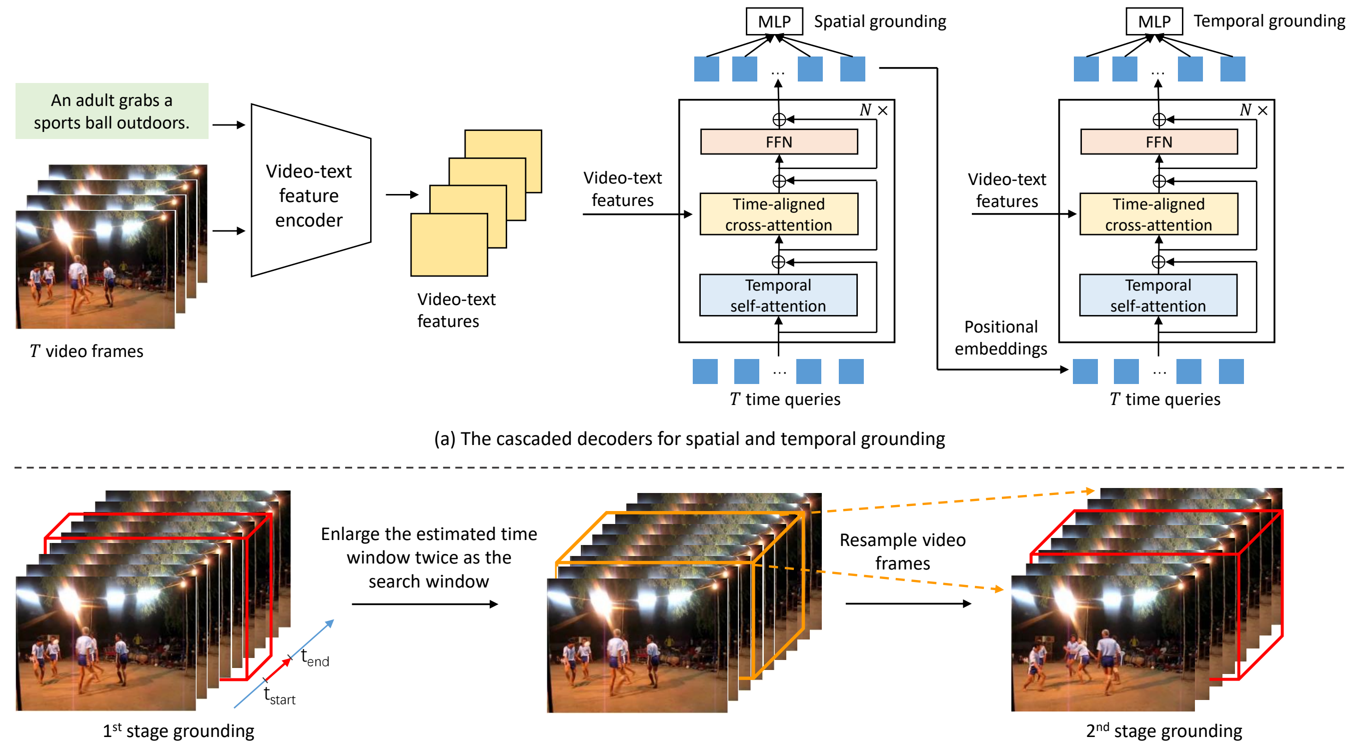
3.1 TubeDETR

We use TubeDETR [22] as our basic framework for human-centric spatio-temporal video grounding (HC-STVG). Given an untrimmed video and a description depicting the object, the HC-STVG task aims to localize the spatio-temporal tube of the target object, *i.e.* a sequence of bounding boxes and the corresponding temporal boundaries. To address this problem, TubeDETR adopts a transformer-based encoder-decoder architecture. In the two-stream video-text encoder, the video-language features are modeled on short clips of the k video frames (*e.g.* $k = 5$ frames per second) using a slow multi-modal branch and a fast vision-only branch, followed by slow-fast feature aggregation. The spatio-temporal decoder of TubeDETR establishes T time queries for T input video frames, and alternately performs temporal self-attention and time-aligned cross-attention to model the target within and across frames. Finally, two MLPs are applied to T time queries to predict the bounding boxes as well as the start and end times of the target on all input video frames.

3.2 Spatio-Temporal Cascaded Decoders

Spatio-temporal video grounding is challenging because it requires both temporal and spatial localization of the target in the video. These two localization subtasks naturally focus on different information (*e.g.* object features for spatial localization and temporal features for estimating the start and end times). Using a single decoder for these two subtasks may lead to suboptimal results. Thus we propose two cascaded decoders to perform the spatial and temporal grounding respectively.

As shown in Figure 1 (a), in the first decoder, we also establish T time queries for T input video frames, where each query is responsible for the spatial grounding of the target at the corresponding frame. To gather the target's features at each frame and ensure the consistency of the localized targets across frames, we employ the same decoder architecture as TubeDETR, applying temporal



(a) The cascaded decoders for spatial and temporal grounding

(b) Multi-stage inference

Figure 1: Our proposed methods for human-centric spatio-temporal video grounding. (a) Based on the encoded video-text features, we establish two cascaded decoders to decouple the spatial and temporal video grounding. (b) We devise the multi-stage inference strategy to resample the video frames focused on the target and thereby produce finer localization results.

self-attention and time-aligned cross-attention to update the time queries. The output queries are fed into an MLP to predict the bounding boxes for all frames.

In the second decoder, we also prepare T time queries and initialize their positional embeddings with the output queries of the first decoder. Since the first decoders' output queries already encode the information about the target object on all frames, we make use of such information to assist temporal grounding in the second decoder. The second decoder has the same architecture as the first decoder, but with unshared weights. We apply another MLP to the output queries of the second decoder to estimate the start and end probabilities for T input frames.

3.3 Multi-Stage Inference

TubeDETR samples a fixed number of T frames (*e.g.* $T = 100$) for all videos to perform spatio-temporal video grounding. However, this fixed sampling strategy may fail to capture information adequately or adaptively for targets with various temporal durations. For example, if a target appears for a short time in the video, only a few frames about the target will be sampled, which may make it difficult to recover the accurate positions of the target on all target-associated frames. To address this issue, we propose multi-stage inference to perform video grounding in a coarse-to-fine manner.

As shown in Figure 1 (b), in the first stage, the model infers the spatial locations and temporal boundaries of the target from T sampled frames of the entire video. With the estimated temporal boundaries, we get a coarse time window of the target and then enlarge it twice as the search window in the video. The video clips within this search window are generally more focused on the target than the entire video. Thus, in the next stage, we resample T video frames in this search window and input them into the model to perform video grounding again. Using these more densely sampled frames for the target, we are able to more accurately locate the bounding box sequence of the target from the video. In our implementation, we apply two-stage inference and directly replace the first stage's localization results with the second stage's estimations (on the corresponding frames), which further improves the grounding performance.

3.4 Model Ensemble

We also perform model ensemble to incorporate the results of models that perform better in spatial or temporal video grounding. Here, we empirically find that freezing the parameters of the CNN backbone during training leads to more accurate temporal localization results. This may be because temporal grounding relies more on the modeling of temporal features rather than learning finer visual features. We train two grounding models, one with a frozen

Table 1: Evaluation of the proposed methods on the test set of the HC-STVG challenge.

Method	m_vIoU	m_tIoU	vIoU@0.3	vIoU@0.5
TubeDETR (our implementation)	36.92	55.24	58.92	32.27
+ Cascaded decoders	37.45	56.31	60.15	33.40
+ Model emsemble	37.49	56.64	59.26	32.95
+ Multi-stage inference	37.71	56.64	59.81	33.47

backbone and the other without. During inference, we apply both models for video grounding and then perform fusion on their estimation results. Specifically, we adopt the temporal grounding results of the first model and select the spatial localization results of the second model within the estimated time window. The formed spatio-temporal grounding results further improve the accuracy by combining the advantages of both models.

4 EXPERIMENTS

4.1 Dataset

We evaluate our approach on the improved HC-STVG 2.0 dataset, which is provided by the HC-STVG challenge of the 4th Person in Context (PIC) workshop [18]. This new dataset has 16,544 video-sentence pairs, and it is split into three subsets with 10,131, 3,482, and 2,931 video-sentence pairs for training, validation, and testing, respectively. The new dataset has also been re-annotated compared to the previous HC-STVG 2.0 dataset.

4.2 Implementation Details

Following TubeDETR [22], we use ResNet-101 [6] and RoBERTa [12] as the visual backbone and text encoder, respectively. We initialize our model (including the feature encoder and decoder) with the pre-trained MDETR [8]. For each video, we sample a total of $T = 100$ frames as inputs (5 frames per second for videos of 20 seconds). For multi-stage inference, we also sample $T = 100$ frames in the second inference stage. To obtain the target bounding boxes on all video frames between the start and end frames, we perform linear interpolation to calculate the bounding boxes on the unsampled frames. During training, we combine the training and validation sets to form a trainval set for model training. All models are trained for 8 epochs with a batch size of 4. The other hyper-parameters are consistent with TubeDETR.

4.3 Evaluation Metrics

We follow [17, 22, 30] to use m_vIoU, the average of $vIoU$ tested on all videos, as the main metric to evaluate our model's performance on spatio-temporal grounding. For the grounding results of a single video, the calculation of $vIoU$ is defined as:

$$vIoU = \frac{1}{|S_u|} \sum_{t \in S_i} IoU\left(\hat{b}_t, b_t\right) \tag{1}$$

where S_u denotes the union of the frames contained in the predicted and ground-truth time segments, S_i refers to the intersection of the frames contained in the predicted and ground-truth time segments, $\hat{b}_t$ and b_t are the predicted bounding boxes and ground-truth bounding boxes at frame t, respectively. We also calculate $vIoU@R$, *i.e.* the proportion of samples with $vIoU > R$, for further evaluation. In order to evaluate the accuracy of temporal grounding, we also use m_tIoU, which is the average of temporal IoU ($tIoU$) between the predicted time segment and the ground truth.

4.4 Evaluation Results

As shown in Table 1, we gradually apply our proposed methods to the baseline to evaluate their effectiveness. In the first row, we retrain the TubeDETR baseline on the union of training and validation sets, which achieves 36.92% in m_vIoU. Based on this baseline, we employ the cascaded decoders to perform decoupled spatial and temporal grounding, which improves the m_vIoU to 37.45%, as shown in the second row of Table 1. It is worth noting that the m_tIoU metric is significantly improved by 1.07 percentage points, further demonstrating the efficacy of cascaded decoding. The model ensemble shows a slight improvement in m_vIoU, and increases the temporal localization accuracy (m_tIoU) by 0.33 points. Nevertheless, we expect that a more dedicated temporal grounding model [20] would improve the performance more significantly. Finally, we adopt the multi-stage inference strategy and further improve the m_vIoU to 37.71%. It is worth mentioning that we share the spatio-temporal decoders in both inference stages, but it may be better to apply separate decoders for these two stages, since they have different distributions of the target time windows. We leave this for future work.

5 CONCLUSION

In this report, we have developed several methodological improvements based on TubeDETR to achieve more accurate spatio-temporal video grounding. We first propose two cascaded decoders to decouple the reasoning process of the spatial and temporal grounding subtasks. Then, we propose the multi-stage inference strategy to re-sample the key frames focused on the target, which helps to produce finer localization results for the target in the video. We also employ the model ensemble strategy to further improve the grounding accuracy. The experiments on the dataset of the HC-STVG challenge validate the effectiveness of our proposed method.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (Grant No. 2020AAA0105802, 2020AAA 0105800), Beijing Natural Science Foundation (Grant No. JQ21017, 4224091), the Natural Science Foundation of China (Grant No. 619723 97, 62036011, 62192782, 61721004, 61906192), the Key Research Program of Frontier Sciences, CAS, Grant No. QYZDJ-SSW-JSC040, and the China Postdoctoral Science Foundation (Grant No. 2021M693402).

REFERENCES

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision.* 5803–5812.

[2] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing.* 162–171.

[3] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 1769–1779.

[4] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision.* 5267–5275.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision.* 2961–2969.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[7] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 1115–1124.

[8] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. MDETR-modulated detection for end-to-end multimodal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 1780–1790.

[9] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. 2020. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10880–10889.

[10] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *The 41st international ACM SIGIR conference on research & development in information retrieval.* 15–24.

[11] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia.* 843–851.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[14] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[15] S Ren, K He, R Girshick, and J Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2016), 1137–1149.

[16] Rui Su, Qian Yu, and Dong Xu. 2021. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 1533–1542.

[17] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. 2021. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology* (2021).

[18] The 4th Person in Context (PIC) Workshop 2022. Human-centric spatio-temporal video grounding. http://picdataset.com/challenge/task/hcvg/

[19] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1960–1968.

[20] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. 2022. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2613–2623.

[21] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9062–9069.

[22] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. TubeDETR: Spatio-Temporal Video Grounding with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 16442–16453.

[23] Li Yang, Yan Xu, Shaoru Wang, Chunfeng Yuan, Ziqi Zhang, Bing Li, and Weiming Hu. 2022. Pdnet: Towards better one-stage object detection with prediction decoupling. *IEEE Transactions on Image Processing* 31 (2022), 5121–5133.

[24] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. 2022. Improving Visual Grounding with Visual-Linguistic Verification and Iterative Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 9499–9508.

[25] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. 2020. Improving one-stage visual grounding by recursive sub-query construction. In *European Conference on Computer Vision.* Springer, 387–404.

[26] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 4683–4693.

[27] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 1307–1315.

[28] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1247–1257.

[29] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12870–12877.

[30] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. 2020. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10668–10677.