# Human-centric Spatio-Temporal Video Grounding via the Combination of Mutual Matching Network and TubeDETR

Fan Yu[1], Zhixiang Zhao[1], Yuchen Wang[1], Yi Xu[2], Tongwei Ren[1,*], Gangshan Wu[1]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2]School of Computer Science and Technology, Soochow University, Suzhou, China
{yf,zhaozx,181250147}@smail.nju.edu.cn,yxu9910@gmail.com,{rentw,gswu}@nju.edu.cn

## ABSTRACT

In this technical report, we represent our solution for the Human-centric Spatio-Temporal Video Grounding (HC-STVG) track of the 4th Person in Context (PIC) workshop and challenge. Our solution is built on the basis of TubeDETR and Mutual Matching Network (MMN). Specifically, TubeDETR exploits a video-text encoder and a space-time decoder to predict the starting time, the ending time and the tube of the target person. MMN detects persons in images, links them as tubes, extracts features of person tubes and the text description, and predicts the similarities between them to choose the most likely person tube as the grounding result. Our solution finally finetunes the results by combining the spatio localization of MMN and the temporal localization of TubeDETR. In the HC-STVG track of the 4th PIC challenge, our solution achieves the third place.

## 1 INTRODUCTION

Human-centric Spatio-Temporal Video Grounding (HC-STVG) task [10] is one of the three tracks in the 4th Person in Context (PIC) workshop and challenge. HC-STVG is a further exploration of visual grounding, which aims to locate the object of a given query with its bounding box [3, 14]. Video grounding requires to localize the starting and ending time of the given video according to a query [2, 15]. Given a sentence depicting an object, spatio-temporal video grounding (STVG) [11, 16] extracts the spatio-temporal tube of the object. The query of an input video in HC-STVG is a sentence describing a person in terms of the appearance, the action and the interaction with the environment. Similar to STVG, HC-STVG needs to localize the target person, *i.e.*, the starting and ending time with the bounding boxes of the target person during the video clip.

The first proposed method for HC-STVG is STGVT [10], which detects region proposals in frames, links the bounding boxes in consecutive frames to form spatio-temporal tube proposals and then uses a visual Transformer combining features extracted from videos and textual descriptions to match and trim the tubes with the given textual description. Su *et al.* [7] propose a unified STVG framework named STVGBert, which also exploits the Transformer model to encode visual and textual features but does not require to generate tube proposals in the begining. In the 2021 PIC challenge, three more solutions were proposed for HC-STVG. Tan *et al.* [8] propose to first localize the temporal segment with the Augmented 2D-TAN model and then predict the spatial location of the target person in each frame. Yu *et al.* [1] propose to extract human information from the query text, *i.e.,* gender, clothing color and clothing type, generate human tubes from the corresponding video, and finally exploit a Transformer to encode visual and textual features to perform tube-description matching and tube trimming. Wang *et al.* [12] introduces metric learning [17] on the basis of visual features extraction from linked human tubes and textual features extraction from the given query. Moreover, TubeDETR [13] is proposed as a unified framework for HC-STVG, which uses video-text encoders to combine visual and textual features and predicts starting time, ending time and the spatio-temporal tube with a space-time decoder.

Our solution is built on the basis of TubeDETR [13] and MMN [12]. We obverse that TubeDETR achieves desired results of spatio localization and MMN has better performance of temporal localization. Thus, we keep the temporal results of MMN and replace its spatio results with TubeDETR's.

## 2 DATASET

The first dataset for the HC-STVG task is *HC-STVG*, where each video is of 20 seconds and is labeled with a sentence describing a person and the corresponding spatio-temporal localization. The spatio-temporal localization in *HC-STVG* is represented by the staring frame, the ending frame and the bounding boxes during the segment. *HC-STVG* dataset has been updated to the third version. Compared with *HC-STVG* 1.0, data in *HC-STVG* 2.0 are expanded and the labels are cleaned. In *HC-STVG* 2.1, noisy data are further manually re-annotated and some videos are moved from the test set to the validation set. The difference among the three versions of data composition is shown in Table 1.

**Table 1: Number of video clips in different versions of *HC-STVG*.**

| version | 1.0 | 2.0 | 2.1 |
|---|---|---|---|
| training set | 4,500 | 10,131 | 10,131 |
| validation set | - | 2,000 | 3,482 |
| test set | 1,160 | 4,413 | 2,913 |

*Corresponding author.

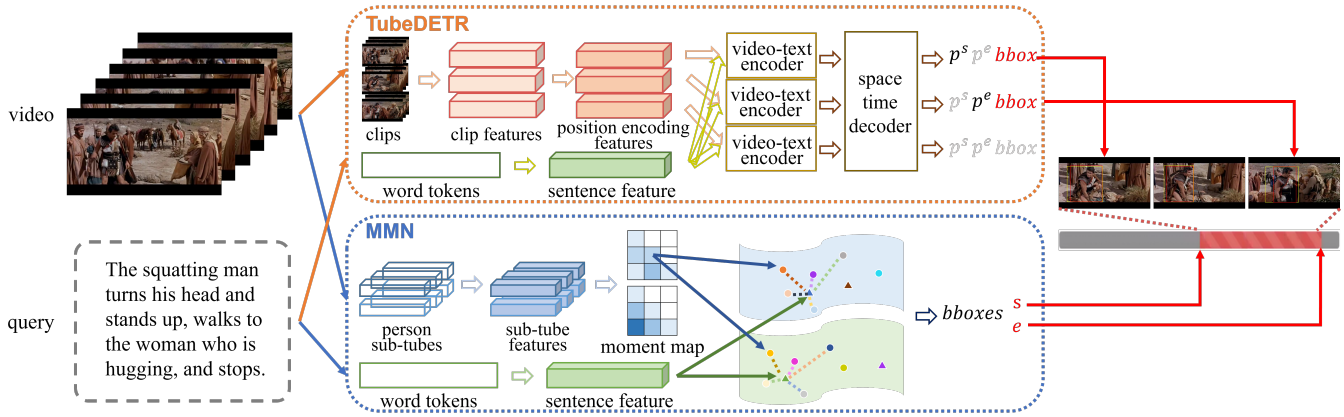


**Figure 1: Illustration of our solution. $s$ and $e$ represent starting time and ending time respectively, $p_s$ and $p_e$ are probabilities of starting time and ending time respectively, and $bbox$ represents bounding box.**

## 3 SOLUTION

As illustrated in Figure 1, our solution combines the temporal localization result of MMN and the spatio localization result of TubeDETR.

**MMN.** MMN performs cross-modal mutual matching in the metric-learning prospective. The framework of MMN contains two stages: the first stage aims to extract features and the second stage matches textual description with tube candidates and trims the target tube. MMN detects humans in frames with Faster R-CNN [6] and links the human bounding boxes following ACT [4] to generate tube candidates. Each candidate tube is split into 16 clips and each tube clip is considered as a unit. The visual feature of each unit is generated by CSN [9] and a 2D moment map is constructed for each tube candidate to predict the IoU score of a candidate sub-tube for the groundtruth tube with the max-pooled visual features. To predict the contrastive score, visual feature and textual feature are both used with metric learning. The final predicted tube is the one containing a moment with the maximum value of the multiplied IoU score and contrastive score, as well as the corresponding starting time and ending time. During training, the total loss is the summary of iou loss, video loss and sentence loss:

$$\mathcal{L}_M = \mathcal{L}_M^{iou} + \lambda(\mathcal{L}_M^{vid} + \mathcal{L}_M^{sen}), \tag{1}$$

$$\mathcal{L}_M^{iou} = -\frac{1}{C}\sum_{i=1}^{C}\left(y_{v_i}\log p_{v_i}^{iou} + (1-y_{v_i})\log(1-p_{v_i}^{iou})\right), \tag{2}$$

$$\mathcal{L}_M^{vid} = -\sum_{i=1}^{N}\log p(i_v|s_i), \tag{3}$$

$$\mathcal{L}_M^{sen} = -\sum_{i=1}^{N}\log p(i_s|v_i), \tag{4}$$

where $\lambda$ is the weight parameter, $C$ is the total number of valid moment candidates, $N$ is the total number of moment-sentence pairs for training, $i_v$ and $i_s$ denote the instance-level classes of the $i^{th}$ moment and the $i^{th}$ sentence respectively, $p_{v_i}^{iou}$ and $y_{v_i}$ denote the predicted and groundtruth iou of the $i^{th}$ moment respectively, and $v_i$ and $s_i$ refer to the $i^{th}$ moment and the $i^{th}$ sentence respectively.

**TubeDETR.** Different from MMN, TubeDETR is a unified framework with the encoder-decoder architecture. The input video is segmented into 20 clips, and the duration of each clips is 1 second. Visual features extracted from video clips are combined with the textual feature extracted from the corresponding query in video-text encoders. A space-time decoder then takes the time-sequentially combined features as input and predicts the probability of starting and ending along with the tube for each clip. During training, the total loss is the summary of bounding box loss, iou loss, Kullback-Leibler divergence loss and guided attention loss:

$$\mathcal{L}_{TD}^{sum} = \alpha\mathcal{L}_{TD}^{box} + \beta\mathcal{L}_{TD}^{Giou} + \gamma\mathcal{L}_{TD}^{KL} + \theta\mathcal{L}_{TD}^{att}, \tag{5}$$

$$\mathcal{L}_{TD}^{box} = \frac{1}{|B|}\sum_{b\in B} L1(b,\hat{b}), \tag{6}$$

$$\mathcal{L}_{TD}^{Giou} = \frac{1}{|B|}\sum_{b\in B}(1 - \frac{I}{U} + \frac{A^c - U}{A^c}), \tag{7}$$

$$\mathcal{L}_{TD}^{KL} = D_{KL}(\hat{\tau}^s\|\tau^s) + D_{KL}(\hat{\tau}^e\|\tau^e), \tag{8}$$

$$\mathcal{L}_{TD}^{att} = -\sum_{i=1}^{n}(1-\delta_{\tau^s\le i\le\tau^e})\log(1-a_i), \tag{9}$$

where $\alpha$, $\beta$, $\gamma$, $\theta$ are weight parameters, $B$ is the set of groundtruth bounding boxes, $\hat{b}$ is the predicted bounding box associated with a groundtruth bounding box element $b$, $L1$ represents L1 loss, $I$ and $U$ is the intersection and union area of the predicted bounding box and the groundtruth bounding box respectively, $A^c$ represents the area of the smallest enclosing box, $D_{KL}$ is the Kullback-Leibler divergence, $\hat{\tau}^s$ and $\hat{\tau}^e$ refer to the probabilities of the start and end of the output video tube respectively, $\tau^s$ and $\tau^e$ refer to the target start and end distribution respectively, $\delta$ is the Kronecker delta and $a_i$ is the $i^{th}$ column in the attention matrix $A$. In our solution, we use the MDETR [5] as the pretrained model, which assists the TubeDETR to achieve the best performance.

**Finetuning.** The bounding box results of TubeDETR is directly predicted by the space-time decoder together with the starting time and ending time and the network for jointly spatio-temporal prediction is trained on the HC-STVG dataset. However, the person tubes and the corresponding features in MMN are generated with pre-trained models. Thus, the spatio localization of TubeDETR is more accurate than that of MMN. The temporal location results of MMN are predicted with a starting-ending moment 2D matrix while the starting time and ending time are predicted in TubeDETR independently, Thus, MMN can achieve better performance in temporal localization. For these reasons, we keep the temporal results of MMN and replace the spatio results with TubeDETR's.

# 4 EXPERIMENTS

## 4.1 Metrics

To evaluate the performance of solutions for HC-STVG, three types of metrics are used.

**tIoU.** tIoU is used to evaluate the performance of temporal localization:

$$tIoU = \frac{|S_i|}{|S_u|}, \tag{10}$$

where $S_i$ is the set of frames in the intersection of predicted and ground truth tube, $S_u$ is the set of frames in the union of predicted and ground truth tube.

**vIoU.** vIoU evaluates both temporal localization and spatio trajectory:

$$vIoU = \frac{1}{|S_u|} \sum_{t \in S_i} IoU(Box^t, Box^{t'}), \tag{11}$$

where $Box^t$ and $Box^{t'}$ are the predicted bounding box and ground truth bounding box of frame $t$.

**vIoU@R.** vIoU@R represents the percentage of samples whose vIoU is larger than R, and vIoU@0.3 and vIoU@0.5 are used in this report.

**Table 2: Comparison results on the *HC-STVG* 2.1 validate set. It is worth noting that the final result of ours in leaderboard of HC-STVG 2022 is the result on the test set of MMN (corresponding to the first line).**

| Methods | vIoU | tIoU | vIoU@0.3 | vIoU@0.5 |
|---|---|---|---|---|
| MMN | 0.280 | **0.503** | 0.449 | 0.227 |
| TubeDETR | 0.285 | 0.445 | 0.426 | 0.192 |
| TubeDETR+MMN | 0.255 | 0.445 | 0.375 | 0.154 |
| MMN+TubeDETR | **0.313** | **0.503** | **0.501** | **0.252** |

## 4.2 Quantitative Analysis

We compare the results of MMN and Tube along with the finetuned results in Table 2. Compared with MMN, TubeDETR achieves better performance in vIoU but has worse performance in tIoU. "TubeDETR+MMN" represents the method that uses the temporal localization of TubeDETR and the spatio localization of MMN, all metrics of which are worse than those of both MMN and TubeDETR. However, "MMN+TubeDETR", which represents the method that uses the temporal result of MMN and replaces its spatio result with TubeDETR's, has the best performance in all metrics. These experimental data validate the effectiveness of our solution, which combines the temporal localization of MMN and the spatio localization of TubeDETR.

## 4.3 Qualitative Analysis

Two visualization examples (Figure 2 and Figure 3) show the performance difference between the solutions in Table 2. As shown in Figure 2, MMN has accurate temporal localization but detects the wrong person, TubeDETR has accurate spatio localization but its prediction of temporal localization is undesired. "TubeDETR+MMN" still detects the wrong person since it keeps the spatio result of MMN, while "MMN+TubeDETR" can detect the right person on the

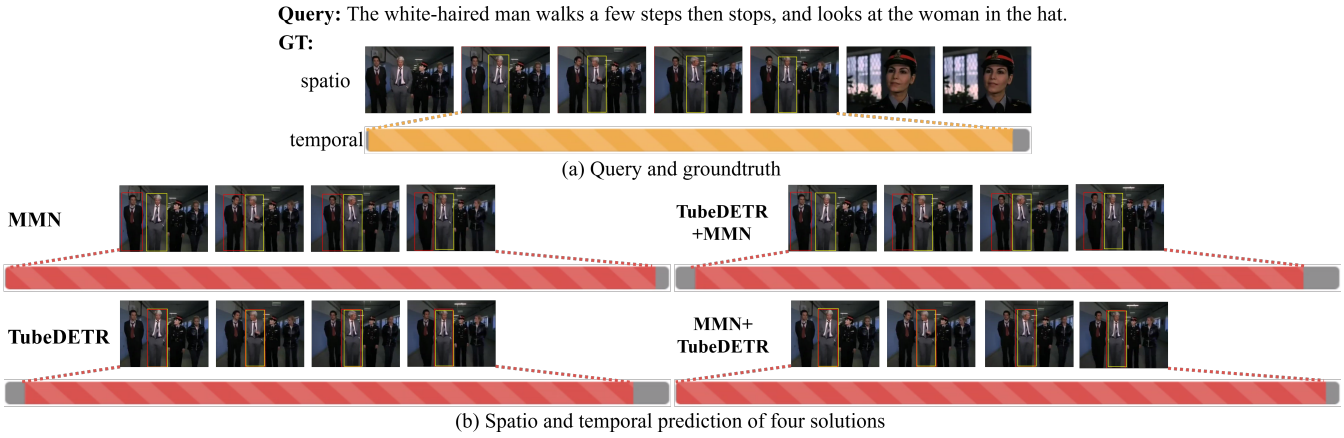


**Figure 2: An example result of MMN, TubeDETR and MMN+TubeDETR. Spatio and temporal annotations in groundtruth are in yellow, and those in prediction are in red.**

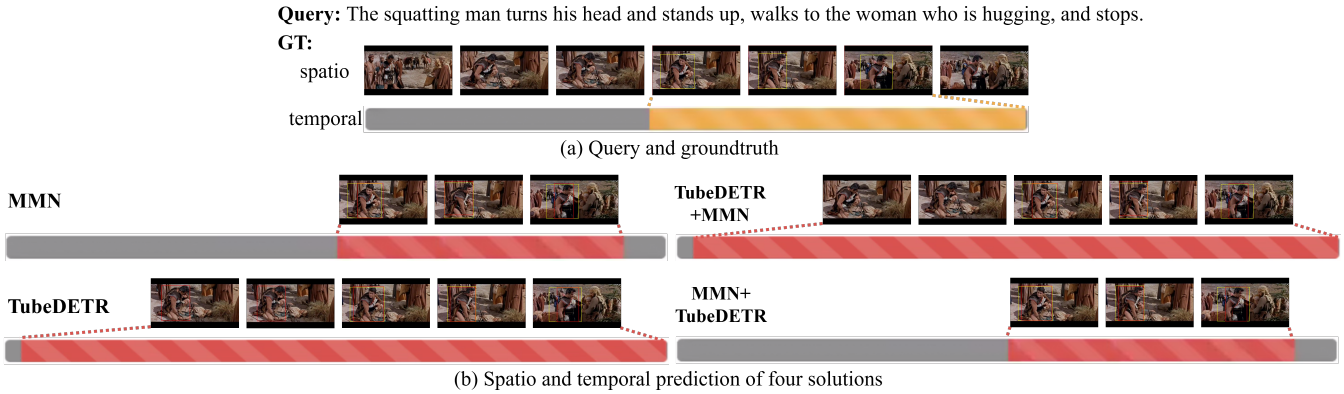


**Figure 3: Another example result of MMN, TubeDETR and MMN+TubeDETR. Spatio and temporal annotations in groundtruth are in yellow, and those in prediction are in red.**

basis of the accurate temporal localization. Figure 3 is another example, where MMN has better performance in temporal localization and TubeDETR almost keeps the whole video duration as the target time, but TubeDETR is more accurate in bounding box detection than MMN. Since "TubeDETR+MMN" uses the temporal result of TubeDETR and the spatio result of MMN, spatio localization is missing in almost half of its target time. "MMN+TubeDETR" keeps the accurate temporal localization of MMN and also uses the better spatio localization of TubeDETR, thereby achieving can achieve good performance in both spatio and temporal evaluation. These examples shows that combining the temporal prediction of MMN and the spatio prediction of TubeDETR is more effective.

## 5 CONCLUSIONS

In this report, we represented our solution for the HC-STVG track in PIC 2022 challenge. Our solution is built on the basis of the MMN and TubeDETR method, keeping the temporal localization result of MMN and the spatio localization result of TubeDETR. Experiments are conducted on the *HC-STVG* 2.1 dataset and validated the effectiveness of our solution.

## ACKNOWLEDGEMENT

This work is supported by National Science Foundation of China (62072232), Natural Science Foundation of Jiangsu Province (BK20191248) and Collaborative Innovation Center of Novel Software Technology and Industrialization.

## REFERENCES

[1] YiYu XinyingWang WeiHu XunLuo ChengLi. [n. d.]. 2rd Place Solutions in the HC-STVG track of Person in Context Challenge 2021. ([n. d.]).
[2] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *IEEE International Conference on Computer Vision*. 5267–5275.
[3] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4555–4564.
[4] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. 2017. Action tubelet detector for spatio-temporal action localization. In *IEEE International Conference on Computer Vision*. 4405–4413.
[5] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. MDETR-modulated detection for end-to-end multi-modal understanding. In *IEEE International Conference on Computer Vision*. 1780–1790.
[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems* 28 (2015).
[7] Rui Su, Qian Yu, and Dong Xu. 2021. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *IEEE International Conference on Computer Vision*. 1533–1542.
[8] Chaolei Tan, Zihang Lin, Jian-Fang Hu, Xiang Li, and Wei-Shi Zheng. 2021. Augmented 2d-tan: A two-stage approach for human-centric spatio-temporal video grounding. *arXiv preprint arXiv:2106.10634* (2021).
[9] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).
[10] 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).
[11] Gongmian Wang, Xing Xu, Fumin Shen, Huimin Lu, Yanli Ji, and Heng Tao Shen. 2022. Cross-modal dynamic networks for video moment retrieval with text query. *IEEE Transactions on Multimedia* 24 (2022), 1221–1232.
[12] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. 2022. Negative sample matters: A renaissance of metric learning for temporal grounding. *AAAI Conference on Artificial Intelligence*.
[13] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. TubeDETR: Spatio-Temporal Video Grounding with Transformers. *IEEE Conference on Computer Vision and Pattern Recognition*.
[14] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7282–7290.
[15] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. 2020. Dense regression network for video grounding. In *IEEE Conference on Computer Vision and Pattern Recognition*. 10287–10296.
[16] 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 *IEEE Conference on Computer Vision and Pattern Recognition*. 10668–10677.
[17] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. 2021. Weakly supervised contrastive learning. In *IEEE International Conference on Computer Vision*. 10042–10051.