PIC
Person In Context

南京大学 NANJING UNIVERSITY

MAGUS
MediA recoGnition and UnderStanding

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

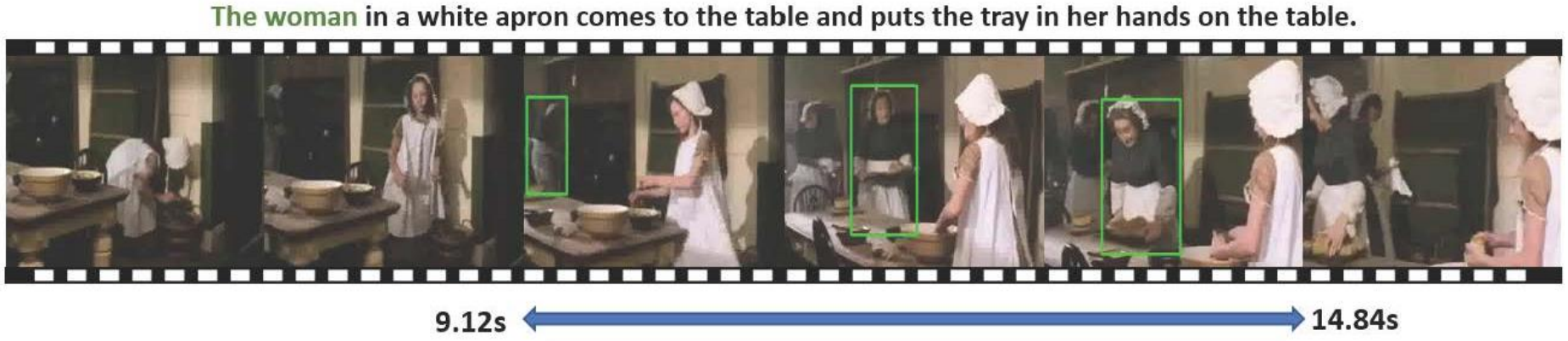
Fan Yu [1], Zhixiang Zhao [1], Yuchen Wang [1], Yi Xu [2], Tongwei Ren [1,*], Gangshan Wu [1]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

[2] School of Computer Science and Technology, Soochow University, Suzhou, China

Introduction

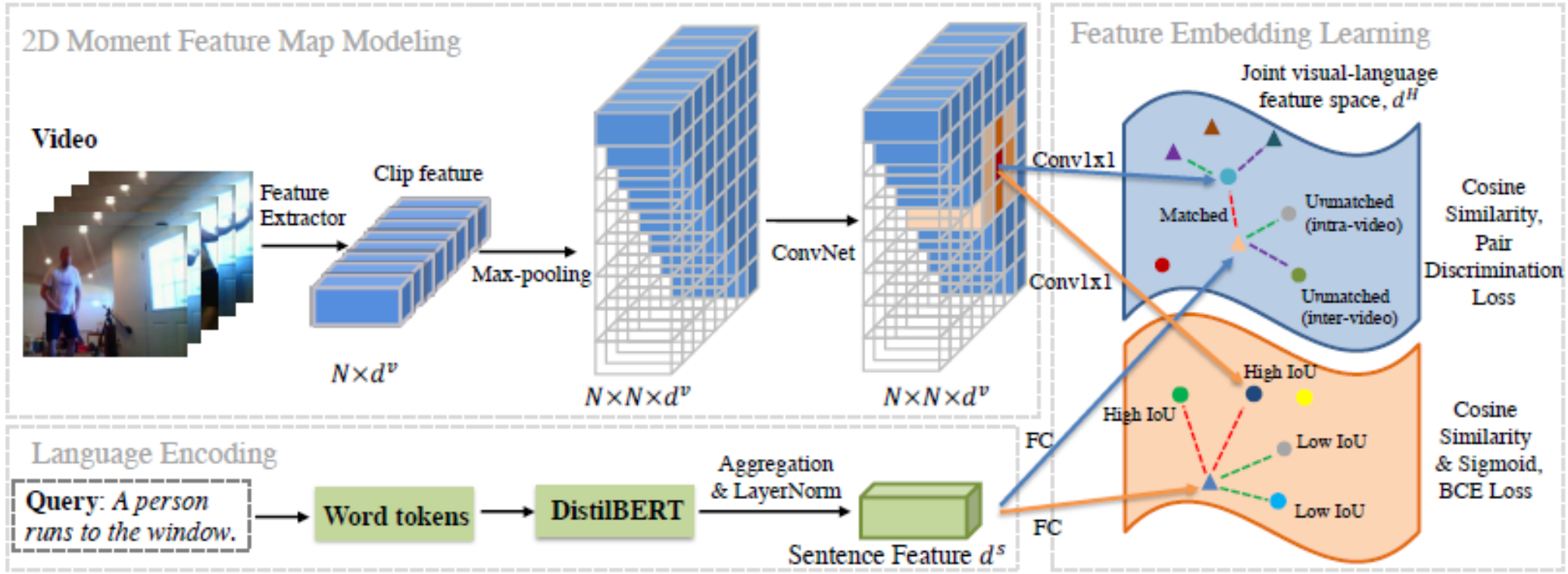
- **Human-centric Spatio-Temporal Video Grounding:** locate the spatio-temporal tube of the target object related to the description
  - Start time and end time
  - Person trajectory
- Combination of Mutual Matching Network and TubeDETR
  - **Temporal:** Mutual Matching Network
  - **Spatio:** TubeDETR



(c) Human centric **spatio-temporal** referring (ours)

Preliminary

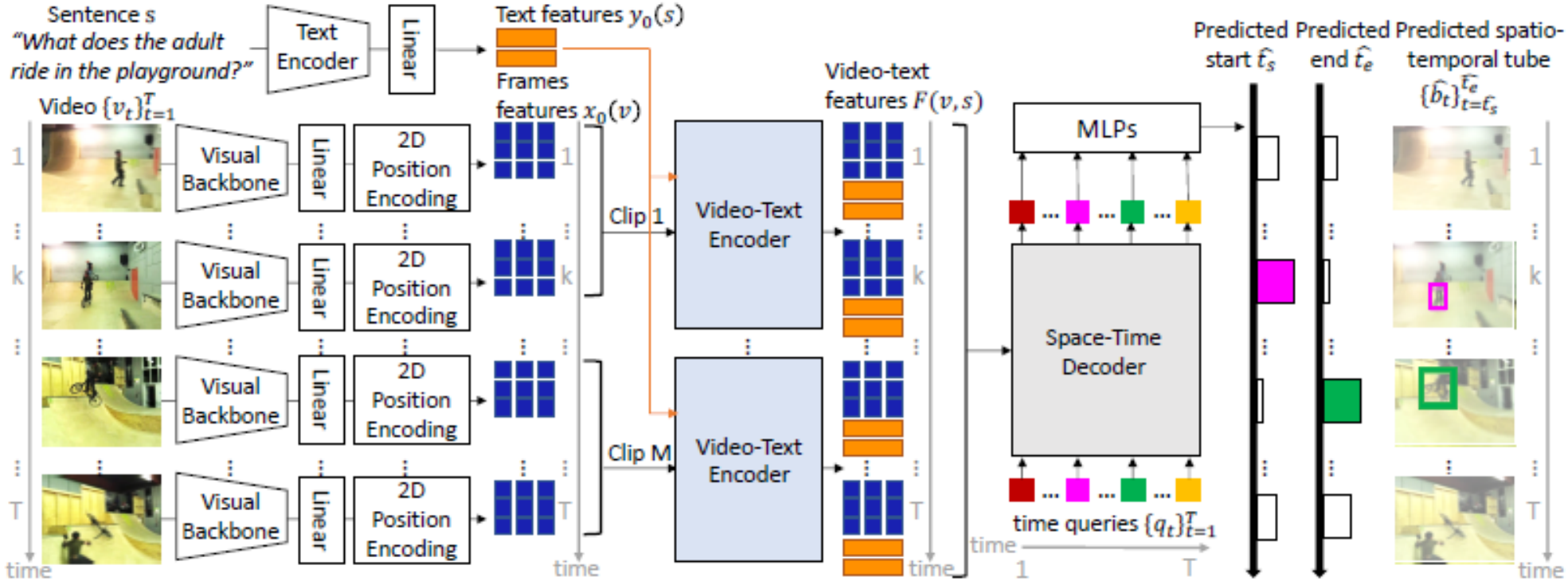
- **Negative Sample Matters:** cross-modal mutual matching in the metric-learning prospective
  - **First stage:** detect persons, generate tube candidates and extract features
  - **Second stage:** match textual description with tube candidates and trim the target tube



Wang et al. Negative Sample Matters: A Renaissance of Metric Learning for Temporal Grounding. AAAI, 2022.

Preliminary

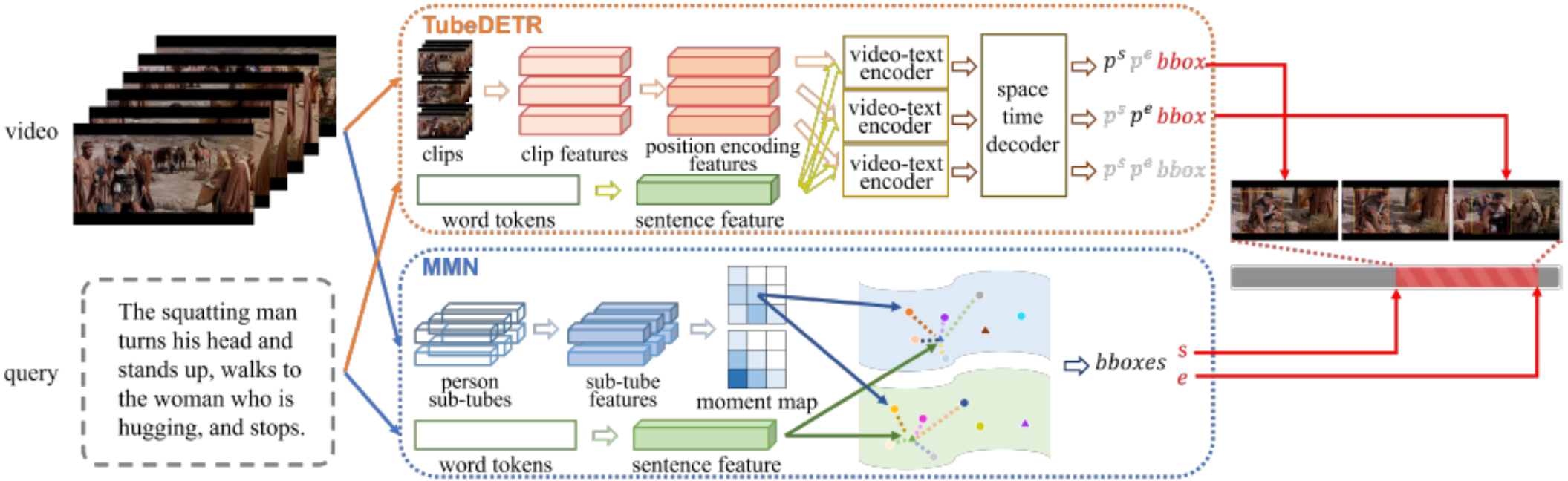
- **TubeDETR:** a unified framework with the encoder-decoder architecture
  - **Video-Text Encoder:** combine visual features with textual feature
  - **Space-Time Decoder:** take the time-sequentially combined features as input and predict the probability of starting and ending along with the tube for each clip



Yang et al. TubeDETR: Spatio-Temporal Video Grounding with Transformers. CVPR, 2022.

Solution

- **Combination of MMN and TubeDETR**
  - **Temporal:** MMN
  - **Spatio:** TubeDETR

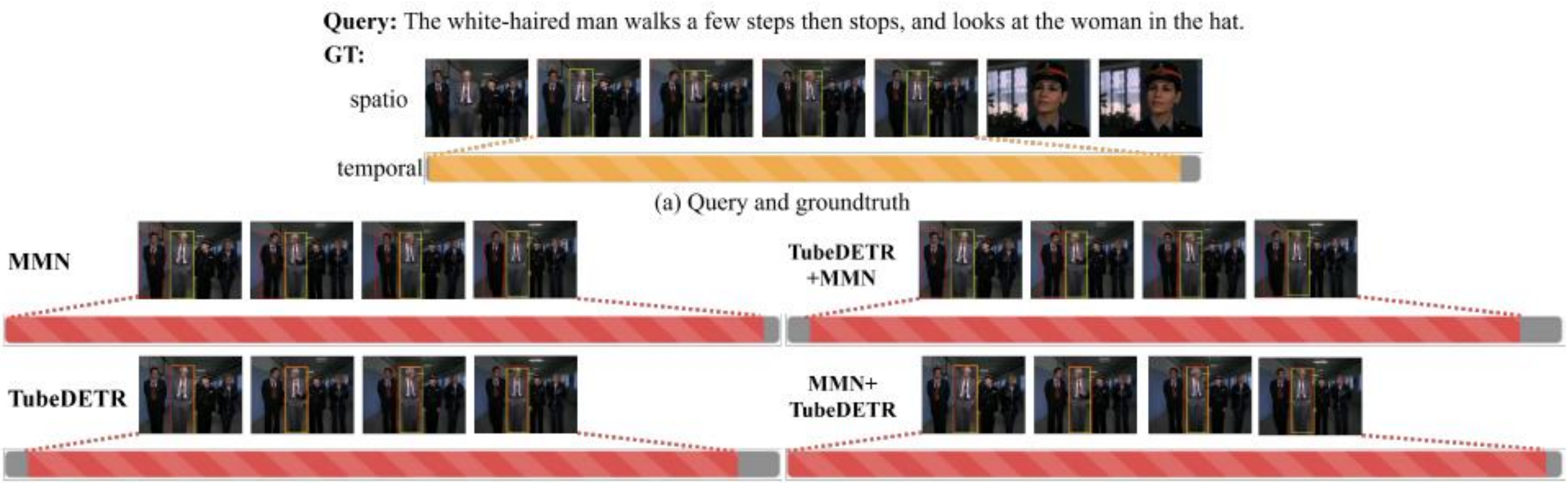


Experiments

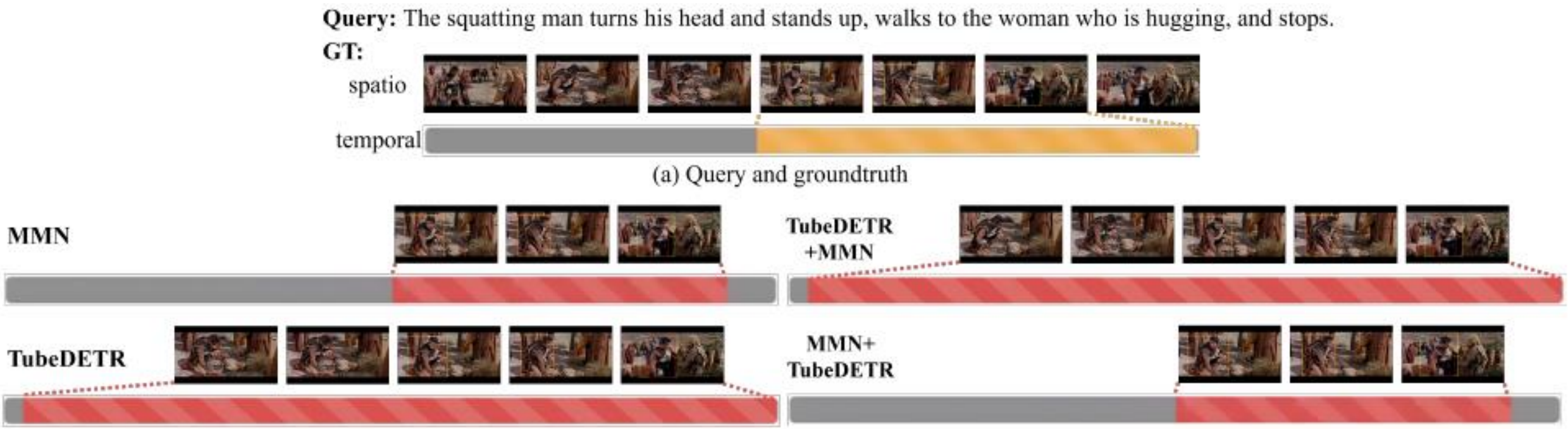
- TubeDETR achieves better performance in vIoU
- MMN achieves better performance in tIoU
- “TubeDETR+MMN”: temporal localization of TubeDETR and spatio localization of MMN
  - all metrics are worse than those of both MMN and TubeDETR
- “MMN+TubeDETR”: temporal localization of MMN and spatio localization of TubeDETR
  - achieves the best performance in all metrics

Methods	mvIoU	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**

Experiments



Experiments



THANK YOU

Fan Yu: yf@smail.nju.edu.cn
Tongwei Ren: rentw@nju.edu.cn

南京大学
NANJING UNIVERSITY

MAGUS
MediA recoGnition
and UnderStanding