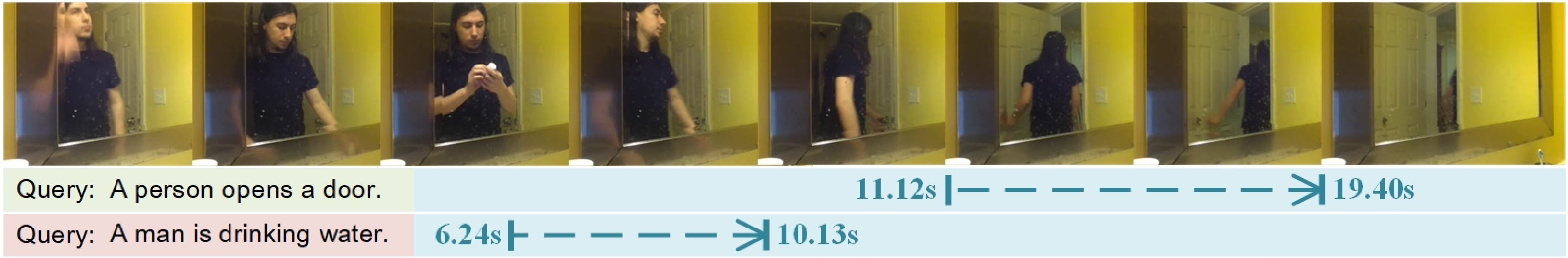
Temporal Sentence Grounding with Relevance Feedback in Videos

Jianfeng Dong[1], Xiaoman Peng[1], Daizong Liu[2], Xiaoye Qu [3],
Xun Yang[4], Cuizhu Bao[1], Meng Wang[5]

[1] Zhejiang Gongshang University
[2] Peking University
[3] Huazhong University of Science and Technology
[4] University of Science and Technology of China
[5] Hefei University of Technology

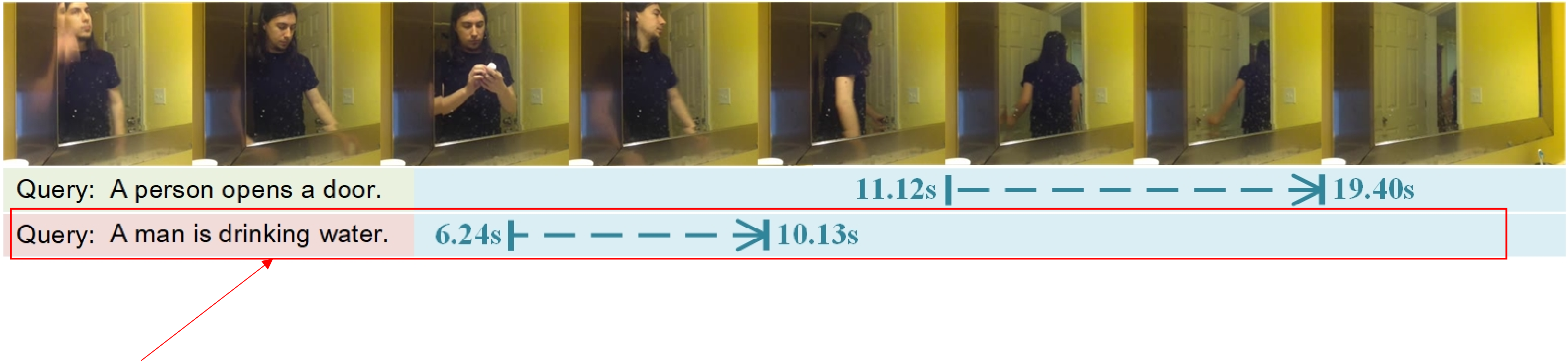
Temporal Sentence Grounding (TSG)

- **TSG** task aims to identify segments that are semantically relevant to a given query from a given long video, assuming that relevant **segments always exist in the given video.**


Query: A person opens a door.
11.12s
19.40s
Query: A man is drinking water.
6.24s
10.13s

Limitations of TSG

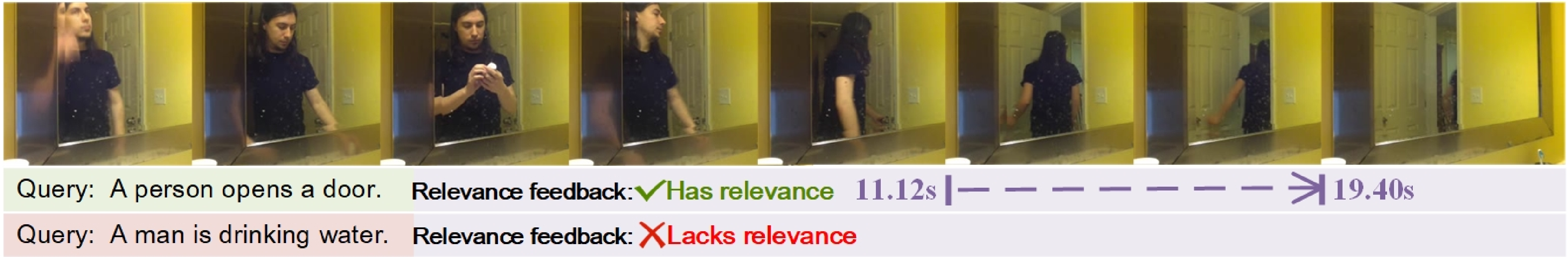
- Traditional TSG methods assume relevant segments always exist in videos. This assumption can lead to inaccuracies and poor performance in real-world scenarios.


Query: A person opens a door.
11.12s
19.40s
Query: A man is drinking water.
6.24s
10.13s

- **When there are no segments in the long video that are semantically relevant to the query text, it still predicts the start and end times of the segment and outputs the segment.**

Temporal Sentence Grounding with Relevance Feedback (TSG-RF)

- Temporal Sentence Grounding with Relevance Feedback (**TSG-RF**) accounts for the absence of relevant segments. **It provides definitive feedback on whether query-related content exists in the video.**



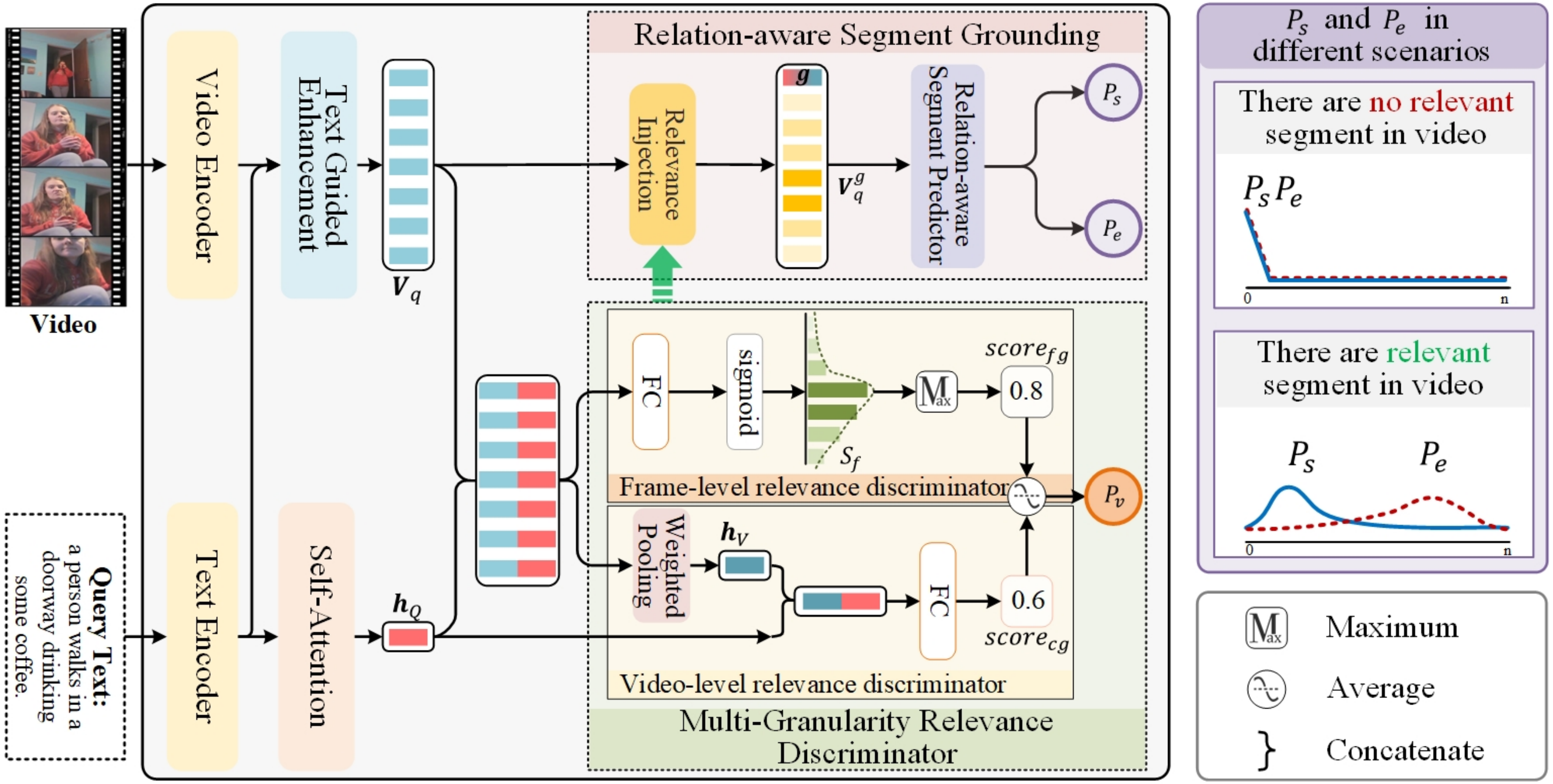
Query: A person opens a door. Relevance feedback: ✓ Has relevance

Query: A man is drinking water. Relevance feedback: ✗ Lacks relevance

Our Method

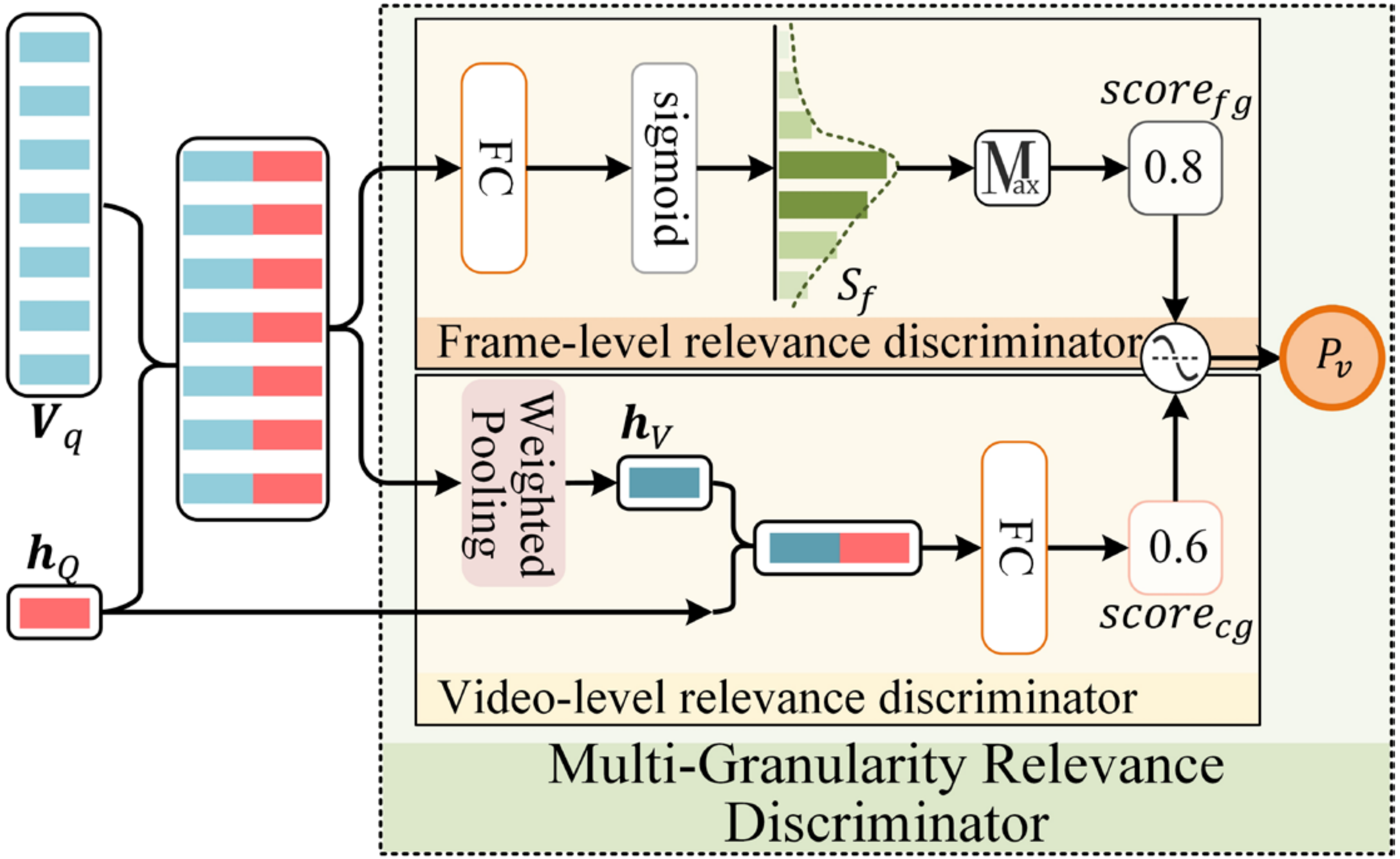
Relation-aware Temporal Sentence Grounding

- We propose the Relation-aware Temporal Sentence Grounding (RaTSG) network specifically designed for TSG-RF task.


Video
Query Text: a person walks in a doorway drinking some coffee.
Video Encoder
Text Encoder
Text Guided Enhancement
Self-Attention
V_q
h_Q
Relation-aware Segment Grounding
Relevance Injection
g
V_q^g
Relation-aware Segment Predictor
P_s
P_e
FC
sigmoid
S_f
Max
$score_{fg}$
0.8
Frame-level relevance discriminator
P_v
Weighted Pooling
h_V
FC
0.6
$score_{cg}$
Video-level relevance discriminator
Multi-Granularity Relevance Discriminator
P_s and P_e in different scenarios
There are no relevant segment in video
$P_s P_e$
0
n
There are relevant segment in video
P_s
P_e
0
n
Max Maximum
Average
} Concatenate

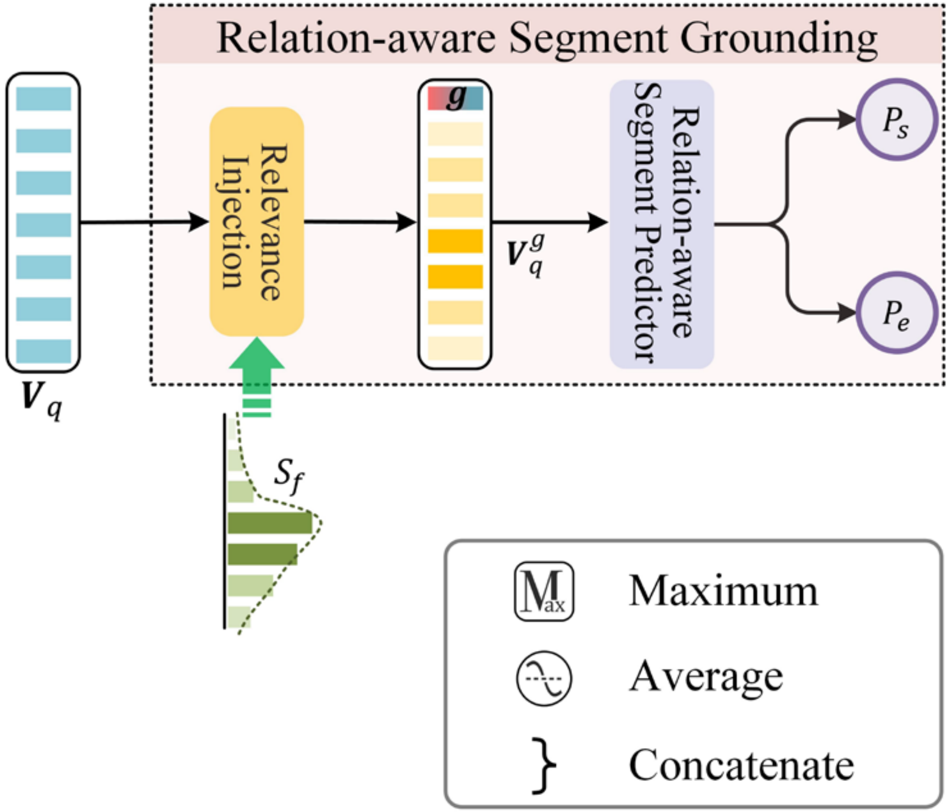
Multi-Granularity Relevance Discriminator

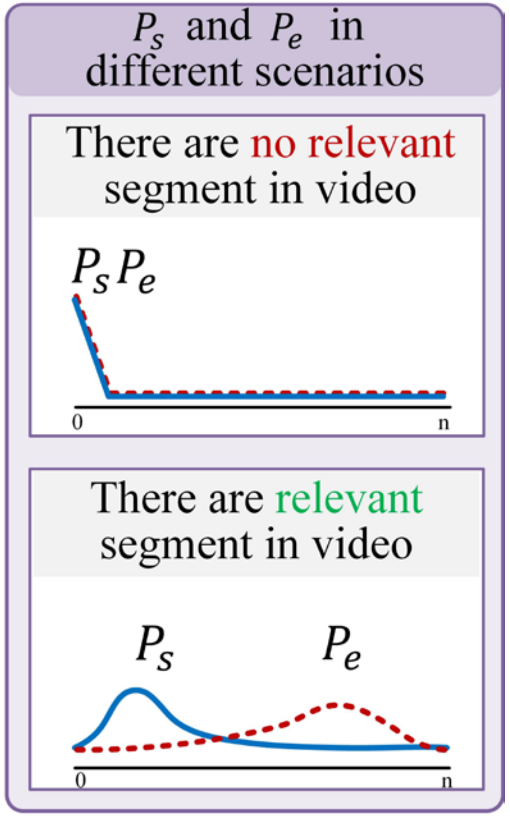
- Captures fine-grained and coarse-grained relevance between text and video. Determines whether query-related content exists at both frame and video levels.


V_q
h_Q
FC
sigmoid
S_f
Max
$score_{fg}$
0.8
Frame-level relevance discriminator
P_v
Weighted Pooling
h_V
FC
0.6
$score_{cg}$
Video-level relevance discriminator
Multi-Granularity Relevance Discriminator

Relation-Aware Segment Grounding

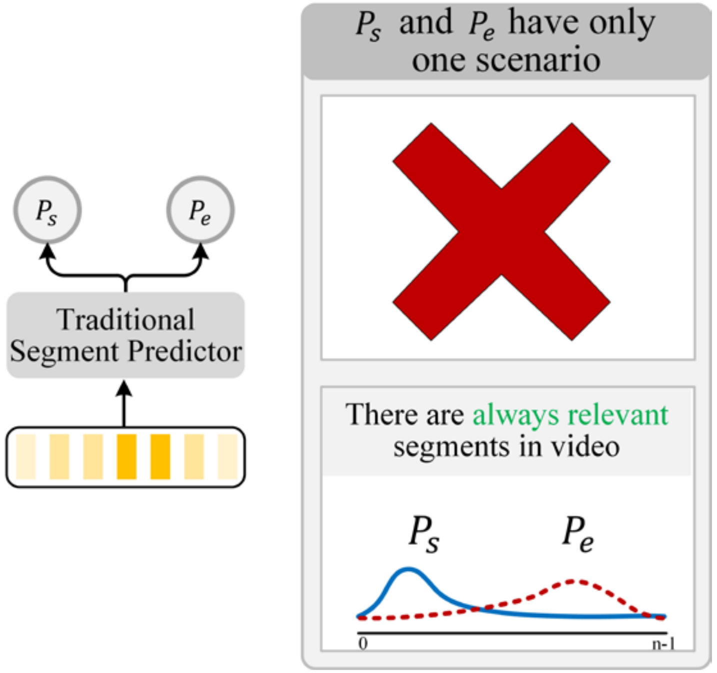
- Selectively predicts start and end boundaries based on relevance feedback. Adapts to the presence or absence of query-related segments.




P_s and P_e in different scenarios
There are no relevant segment in video
$P_s P_e$
0
n
There are relevant segment in video
P_s
P_e
0
n

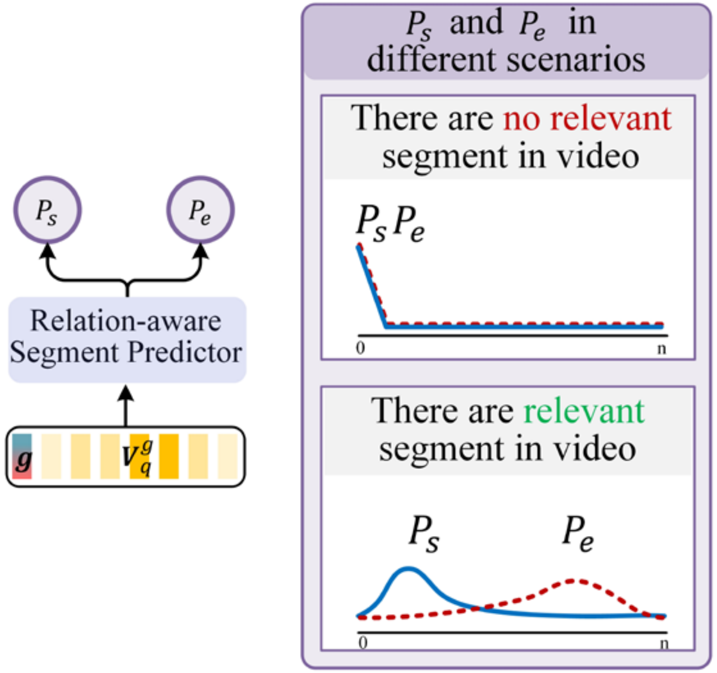
Relation-Aware Segment Predictor

- Compared to Traditional Segment Predictor, our Relation-Aware Segment Predictor uses a special token to encode the relationship between the query and video. The token helps the model focus on where each segment starts and ends, making it more accurate at identifying both relevant and irrelevant parts of the video.



P_s and P_e have only one scenario
P_s
P_e
Traditional Segment Predictor
There are always relevant segments in video
P_s
P_e
0
n-1

Traditional Segment Predictor



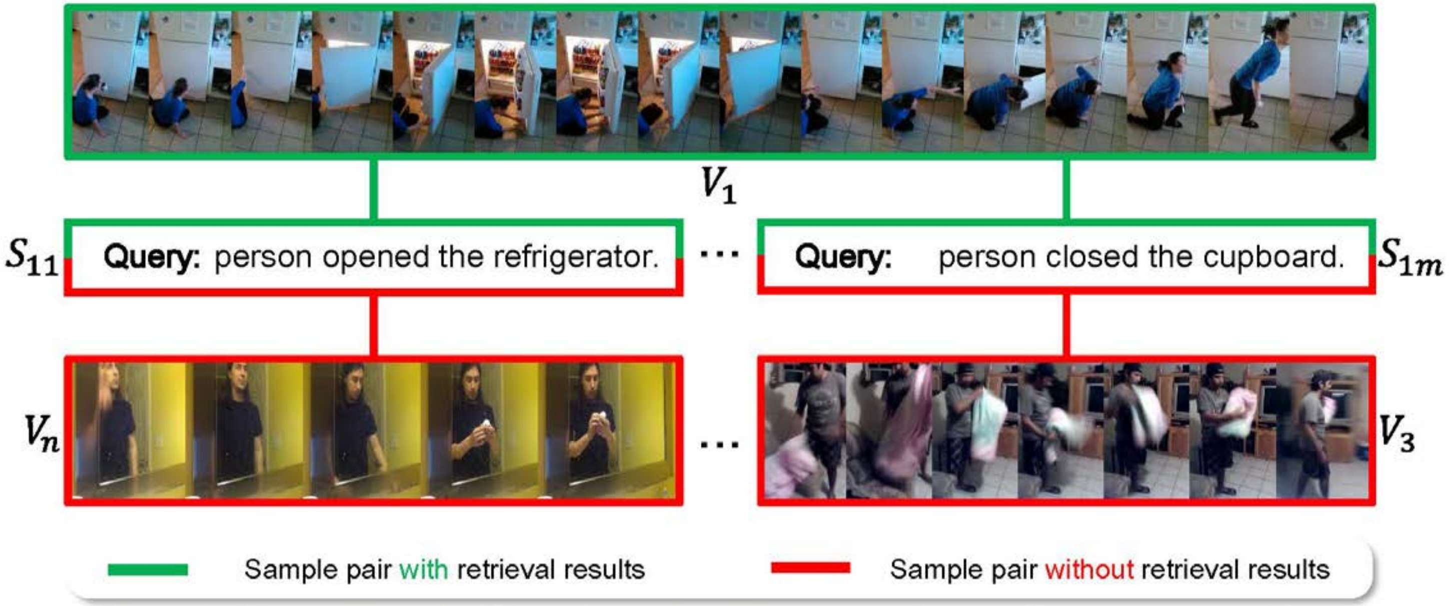
P_s and P_e in different scenarios
There are no relevant segment in video
$P_s P_e$
0
n
P_s
P_e
Relation-aware Segment Predictor
g
V_q^g
There are relevant segment in video
P_s
P_e
0
n

Our Relation-Aware Segment Predictor

Experiments

Datasets

Charades-RF and ActivityNet-RF: These are reconstructed versions of the Charades-STA and ActivityNet datasets, modified to include non-relevant samples, **obtained by pairing queries with randomly selected video that do not match the query.** This reconstruction enables the evaluation of the model's ability to handle queries where no relevant segment exists.


V_1
S_{11}
Query: person opened the refrigerator.
...
Query: person closed the cupboard.
S_{1m}
V_n
...
V_3
🟩 Sample pair with retrieval results
🟥 Sample pair without retrieval results

Evaluation Metrics

We use several key metrics to evaluate the performance of the RaTSG model:

1. **Accuracy:** Measures the correctness of the model's relevance feedback (i.e., determining if a relevant segment exists).
2. **Magic IoU:** A redefined mean IoU that accounts for samples with no grounding results, ensuring accurate overlap assessment.
3. **R{n}@{m}:** Indicates the percentage of queries where at least one segment in the top-n predictions has an IoU greater than m.

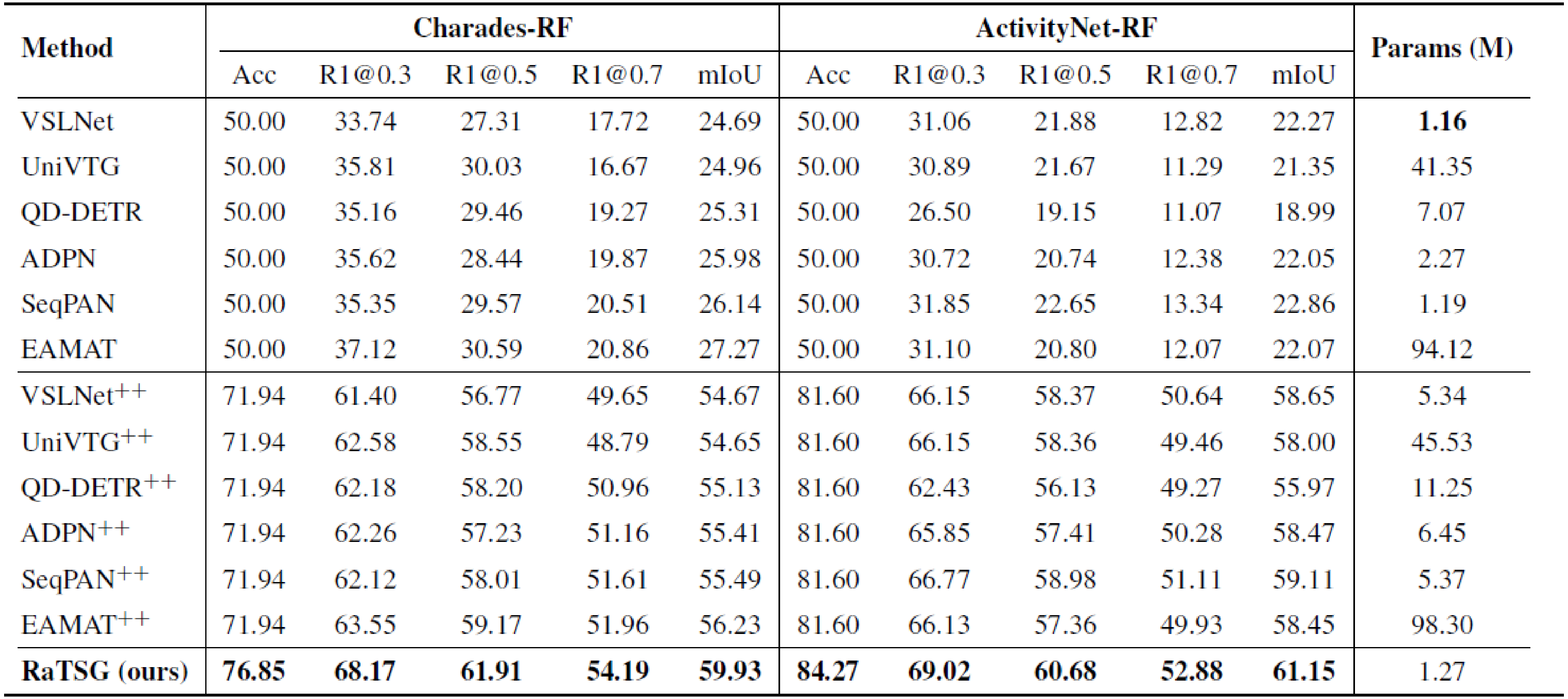
Baseline Models

There are no models are specifically designed for TSG-RF, so we adapted existing TSG models for this task.

- **Selected TSG Models:** We chose VSLNet, SeqPAN, EAMAT, ADPN, UniVTG, and QD-DETR, as they are recent, open-source TSG models.
- **Resulting Adapted Models:** By adding a relevance discriminator, we created enhanced versions: VSLNet++, SeqPAN++, EAMAT++, ADPN++, UniVTG++, and QD-DETR++.

Comparison with Baseline Methods

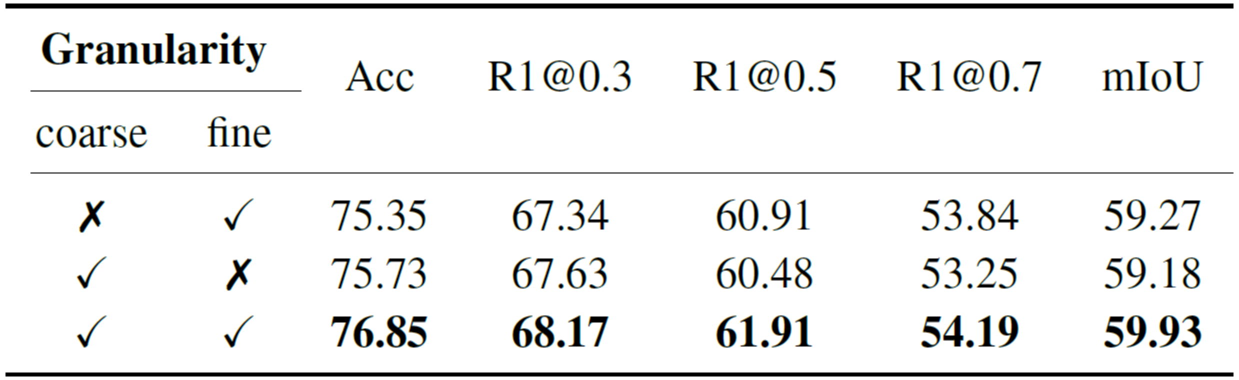
Our proposed RaTSG model consistently outperforms all Model++ baselines and previous state-of-the-art methods.



Method	Charades-RF					ActivityNet-RF					Params (M)
	Acc	R1@0.3	R1@0.5	R1@0.7	mIoU	Acc	R1@0.3	R1@0.5	R1@0.7	mIoU	
VSLNet	50.00	33.74	27.31	17.72	24.69	50.00	31.06	21.88	12.82	22.27	**1.16**
UniVTG	50.00	35.81	30.03	16.67	24.96	50.00	30.89	21.67	11.29	21.35	41.35
QD-DETR	50.00	35.16	29.46	19.27	25.31	50.00	26.50	19.15	11.07	18.99	7.07
ADPN	50.00	35.62	28.44	19.87	25.98	50.00	30.72	20.74	12.38	22.05	2.27
SeqPAN	50.00	35.35	29.57	20.51	26.14	50.00	31.85	22.65	13.34	22.86	1.19
EAMAT	50.00	37.12	30.59	20.86	27.27	50.00	31.10	20.80	12.07	22.07	94.12
VSLNet^{++}	71.94	61.40	56.77	49.65	54.67	81.60	66.15	58.37	50.64	58.65	5.34
UniVTG^{++}	71.94	62.58	58.55	48.79	54.65	81.60	66.15	58.36	49.46	58.00	45.53
QD-DETR^{++}	71.94	62.18	58.20	50.96	55.13	81.60	62.43	56.13	49.27	55.97	11.25
ADPN^{++}	71.94	62.26	57.23	51.16	55.41	81.60	65.85	57.41	50.28	58.47	6.45
SeqPAN^{++}	71.94	62.12	58.01	51.61	55.49	81.60	66.77	58.98	51.11	59.11	5.37
EAMAT^{++}	71.94	63.55	59.17	51.96	56.23	81.60	66.13	57.36	49.93	58.45	98.30
RaTSG (ours)	**76.85**	**68.17**	**61.91**	**54.19**	**59.93**	**84.27**	**69.02**	**60.68**	**52.88**	**61.15**	1.27

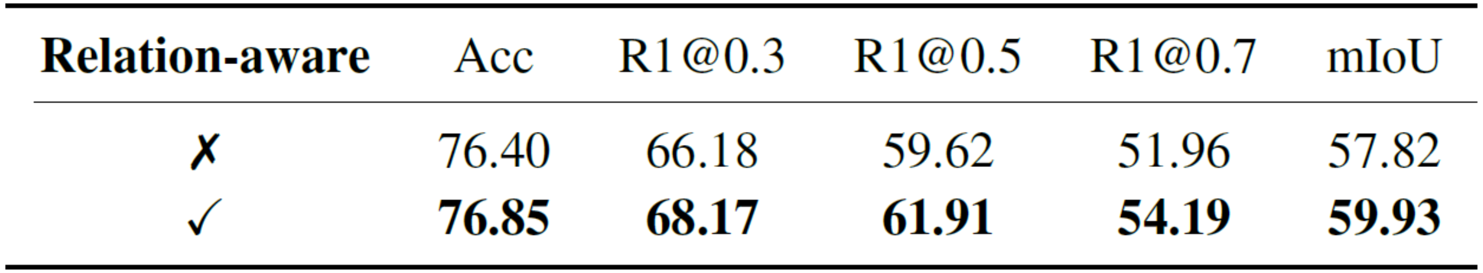
Ablation Studies

- The multi-granularity relevance discriminator outperforms models using only coarse or fine discrimination, effectively capturing partial relevance.



Granularity		Acc	R1@0.3	R1@0.5	R1@0.7	mIoU
coarse	fine					
✗	✓	75.35	67.34	60.91	53.84	59.27
✓	✗	75.73	67.63	60.48	53.25	59.18
✓	✓	**76.85**	**68.17**	**61.91**	**54.19**	**59.93**

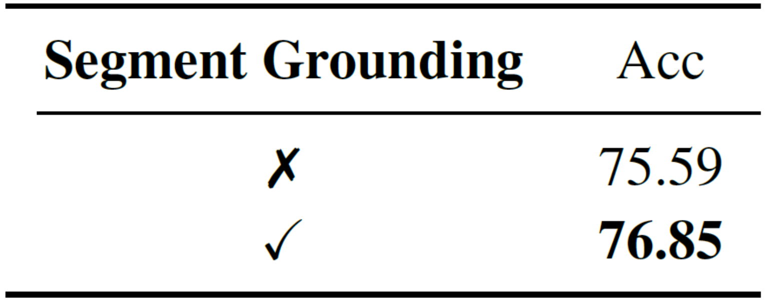
- The relation-aware segment grounding module outperforms models using random initialization, effectively leveraging prior relevance information.



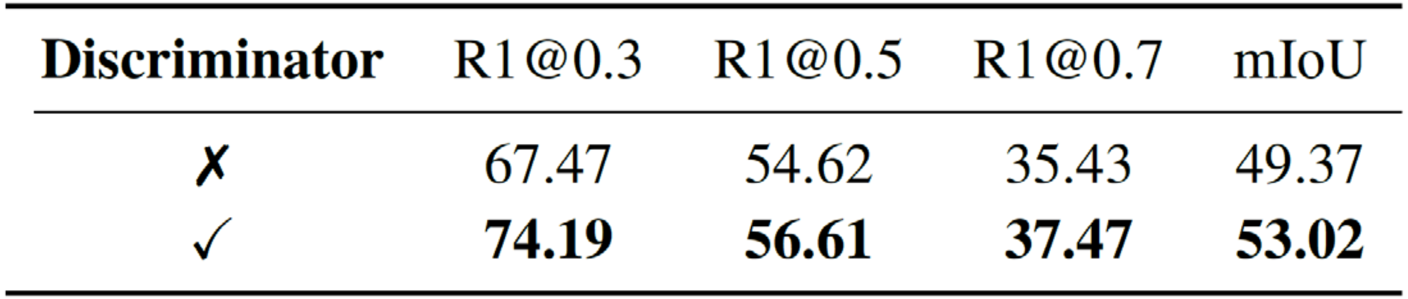
Relation-aware	Acc	R1@0.3	R1@0.5	R1@0.7	mIoU
✗	76.40	66.18	59.62	51.96	57.82
✓	**76.85**	**68.17**	**61.91**	**54.19**	**59.93**

Mutual Enhancement between Relevance Discrimination and Segment Grounding

- RaTSG model combines relevance discrimination and segment grounding. Removing either module reduces the performance of the other, showing that both components mutually enhance each other and validate our dual-branch framework.

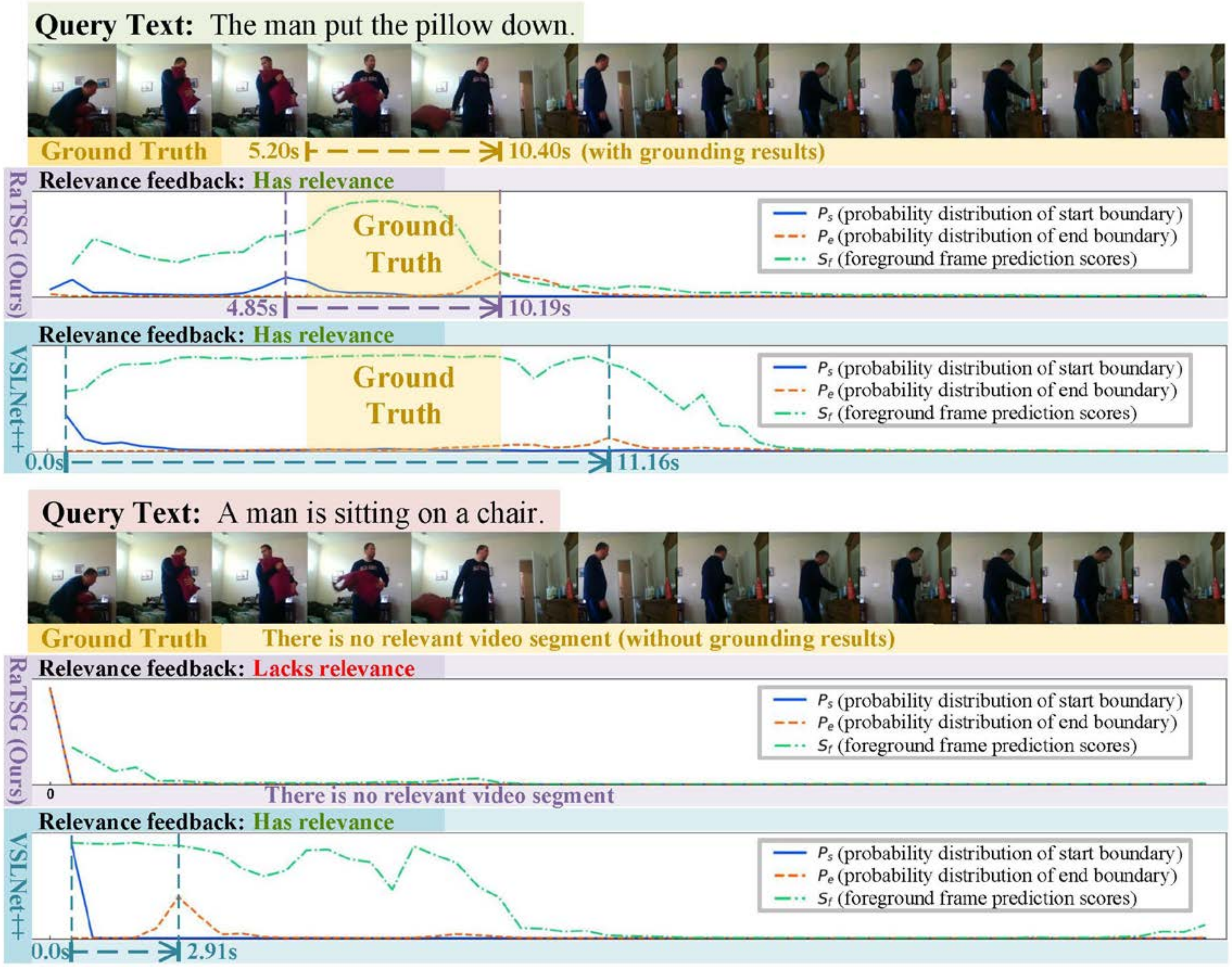


Segment Grounding	Acc
✗	75.59
✓	**76.85**



Discriminator	R1@0.3	R1@0.5	R1@0.7	mIoU
✗	67.47	54.62	35.43	49.37
✓	**74.19**	**56.61**	**37.47**	**53.02**

Analysis of Grounding Examples



- Our RaTSG effectively handles both relevant and non-relevant samples, providing more accurate localization and better distinction between foreground and background frames compared to VSLNet++.
- For samples without grounding results, RaTSG consistently gives correct feedback with low scores, while VSLNet++ often incorrectly predicts high scores, leading to errors.

Conclusions

- **TSG-RF Task:** We introduce a more realistic task, TSG-RF, which addresses the limitation of existing TSG methods that cannot handle cases without query-related segments.
- **Proposed Model (RaTSG):** Our model, RaTSG, integrates a multi-granularity relevance discriminator with segment grounding, effectively localizing relevant segments or providing clear feedback when none exist.
- **Dataset Contribution:** We contribute two new datasets specifically designed for the TSG-RF task.

Source code of paper: https://github.com/HuiGuanLab/RaTSG

If you have any questions, please free to contact us.

E-mail: dongjf24@gmail.com pengxiaoman1999@gmail.com