Phrase-level Temporal Relationship Mining for Temporal Sentence Localization

Minghang Zheng, Sizhe Li, Qingchao Chen, Yuxin Peng, Yang Liu

Peking University,

Beijing Institute for General Artificial Intelligence

Task: Temporal Sentence Localization

Inputs: Video + Sentence query
Outputs: Target video clip (start and end timestamps)

Video:

Query:

A man puts on gloves and then clean the snow

Output

5.5s 15.1s

Phrase-level Query:

Query:

Puts on — 5.5s **Output** 9.8s

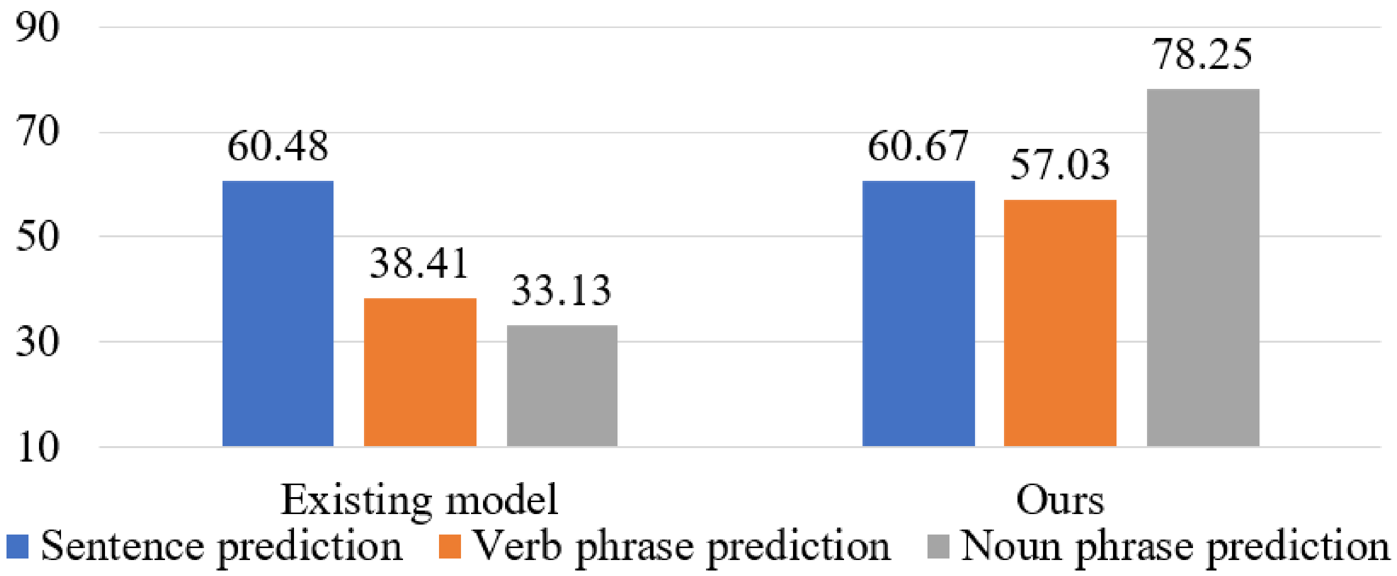
Gloves — 5.5s **Output** 18.3s

Motivation

Observations: Existing work can not deal with the **phrase-level** query well

Problems:

- Insufficient understandings of relationship between **simple visual and language concepts**
- Questioned model **interpretability** and **robustness**

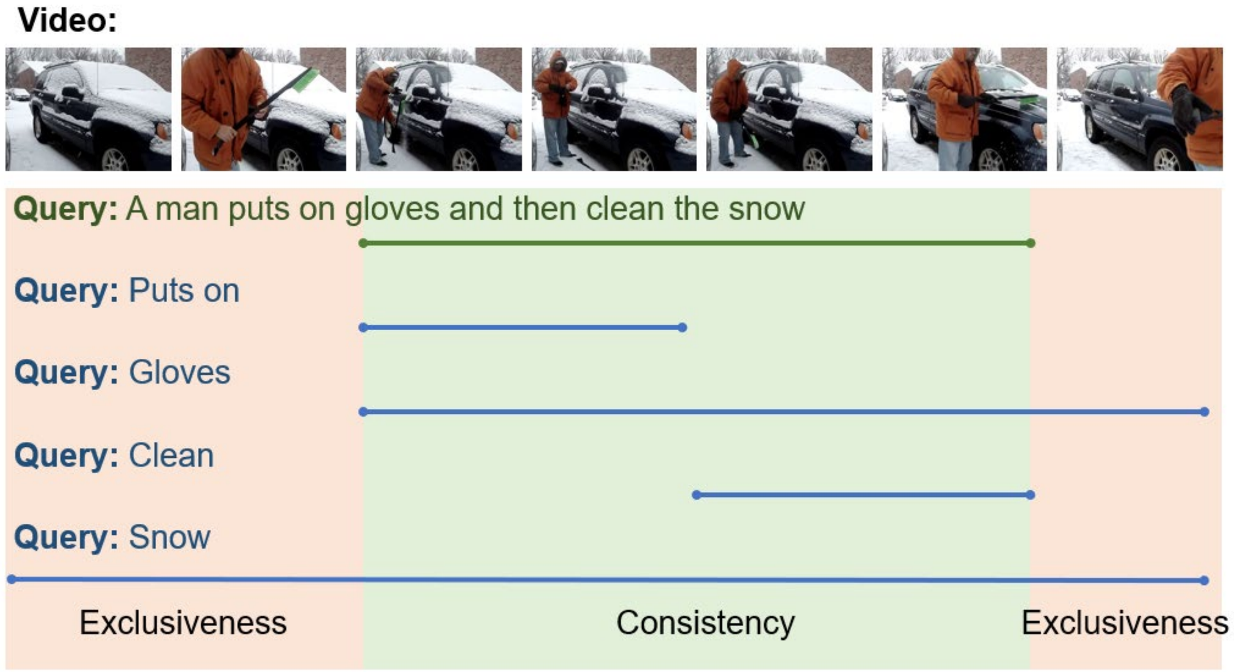


Motivation

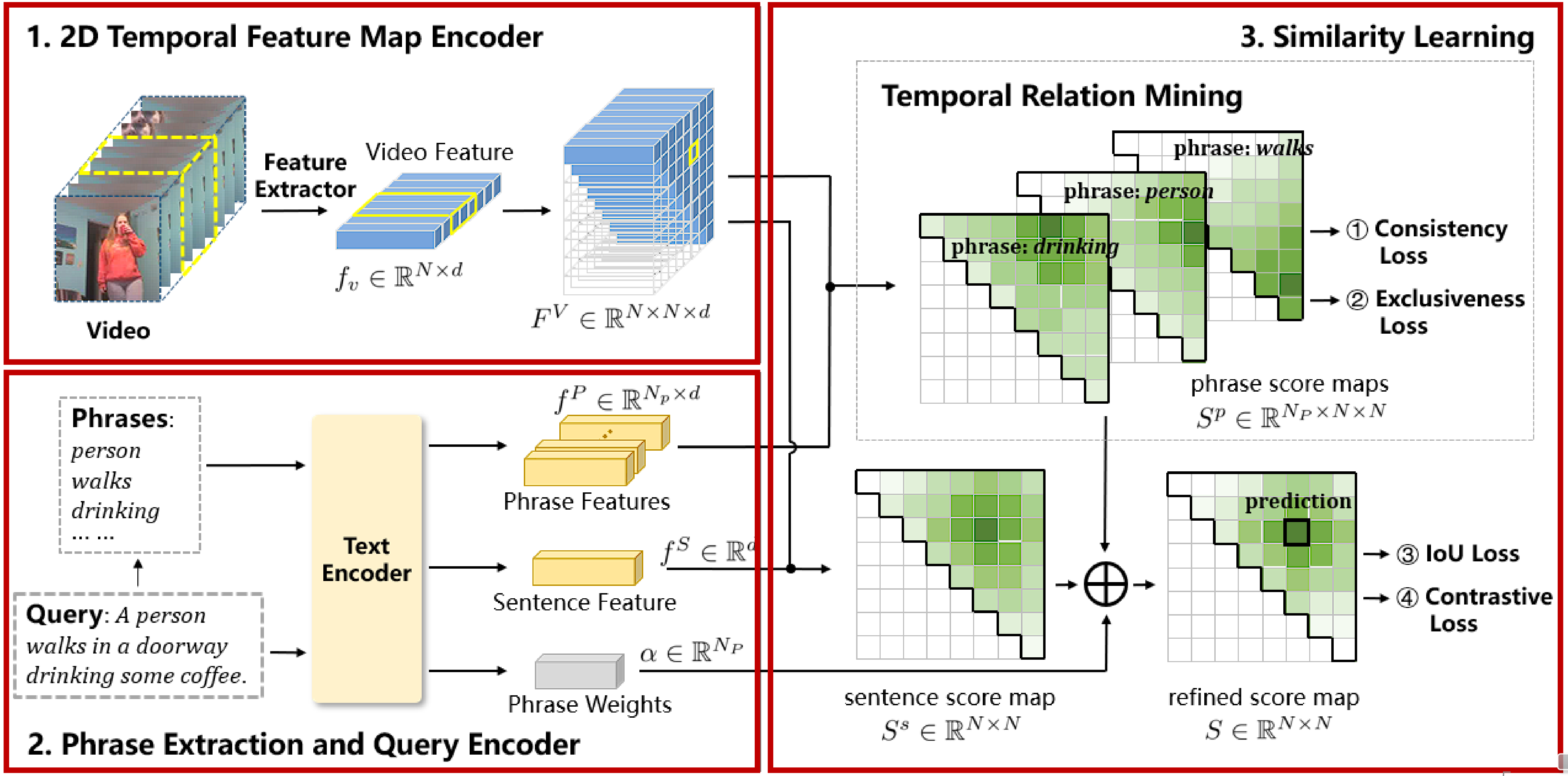
Difficulty: No phrase-level annotation

Solution: Phrase-level Temporal Relationship Mining (TRM)

- Consider **phrase-level prediction**
- Mining **temporal relationship** between phrase and sentence level prediction
- Two principles: **Consistency** & **Exclusiveness**



Overall Framework



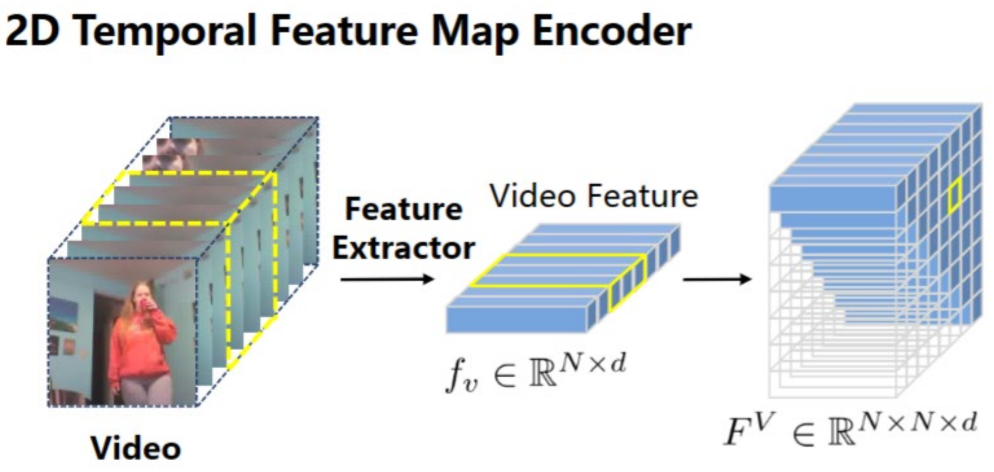
1. 2D Temporal Feature Map Encoder

Aims: Extract video features and generating 2D proposal feature map

Visual Feature Extraction

- Visual encoder: C3D[1] or VGG[2]

- Generate 2D feature map by Conv: F_{ij}^{V}: video candidate starting from the i-th clip and ending with the j-th clip



[1]Tran, et al. Learning Spatiotemporal Features with 3D Convolutional Networks. ICCV, 2015.
[2]Simonyan, et al. Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR, 2015

2. Phrase Extraction and Query Encoder

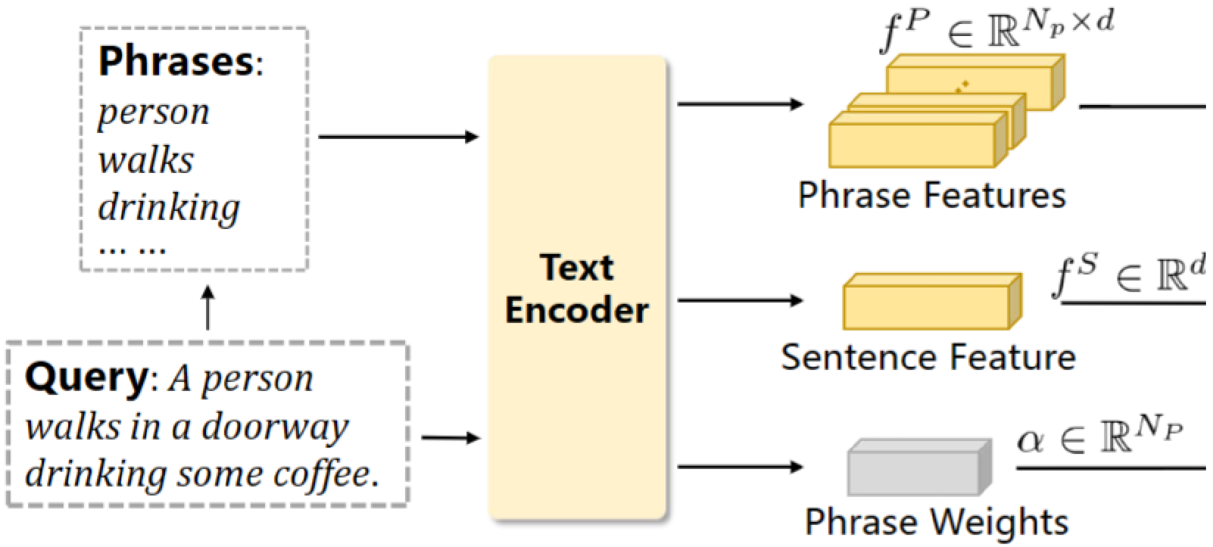
Aims: Extract fine-grained phrases and extract text features

Phrase Extraction

- From pretrained SRLBERT[1]
  N_p phrases: $[p_1, p_2, \ldots, p_{N_p}]$

Query Encoder

- Text encoder: DistilBERT[2]
  sentence features: $f^S \in \mathbb{R}^d$
  phrase features: $f^P \in \mathbb{R}^{N_p \times d}$
- Predict phrase weights by Attention
  $\alpha \in \mathbb{R}^{N_p}$: importance of each phrase



Phrase Extraction and Query Encoder

[1]Shi, et al. Simple bert models for relation extraction and semantic role labeling. arXiv, 2019.
[2]Sanh, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv, 2019

3.1 Similarity Learning

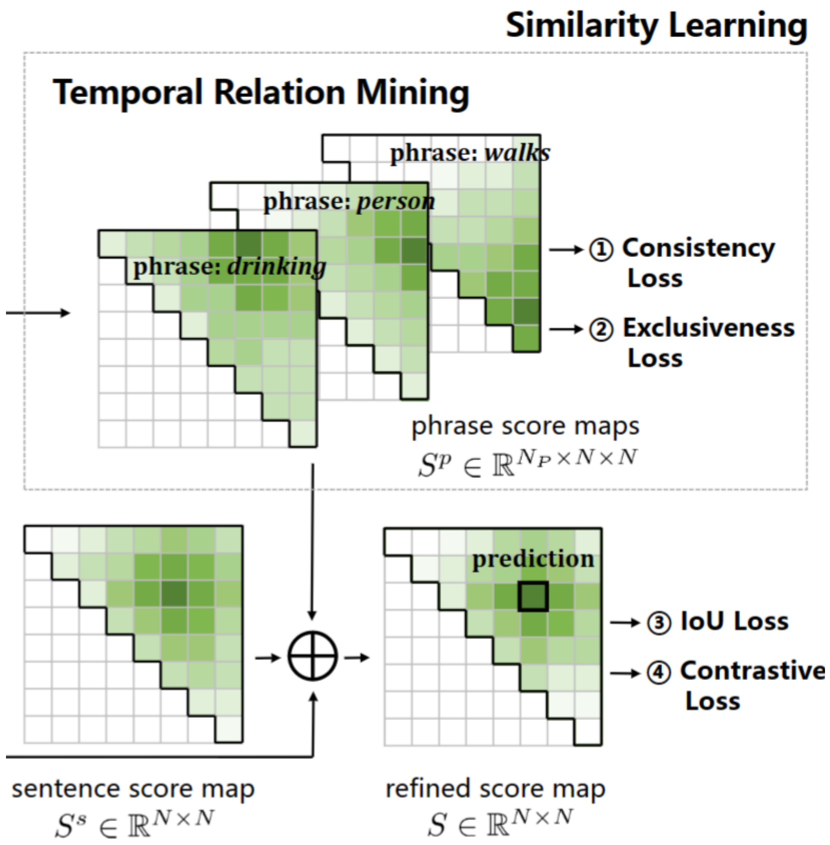
Aims: Learn semantic relevance of sentence/phrase with each proposal

Score Map Generation

- Calculate cosine similarity
  Sentence score map: $S^s = F^{VT} f^s$
  Phrase Score map: $S_i^p = F^{VT} f_i^p$

Temporal Relation Mining

- Improve the quality of phrase score map
- Consistency & Exclusiveness

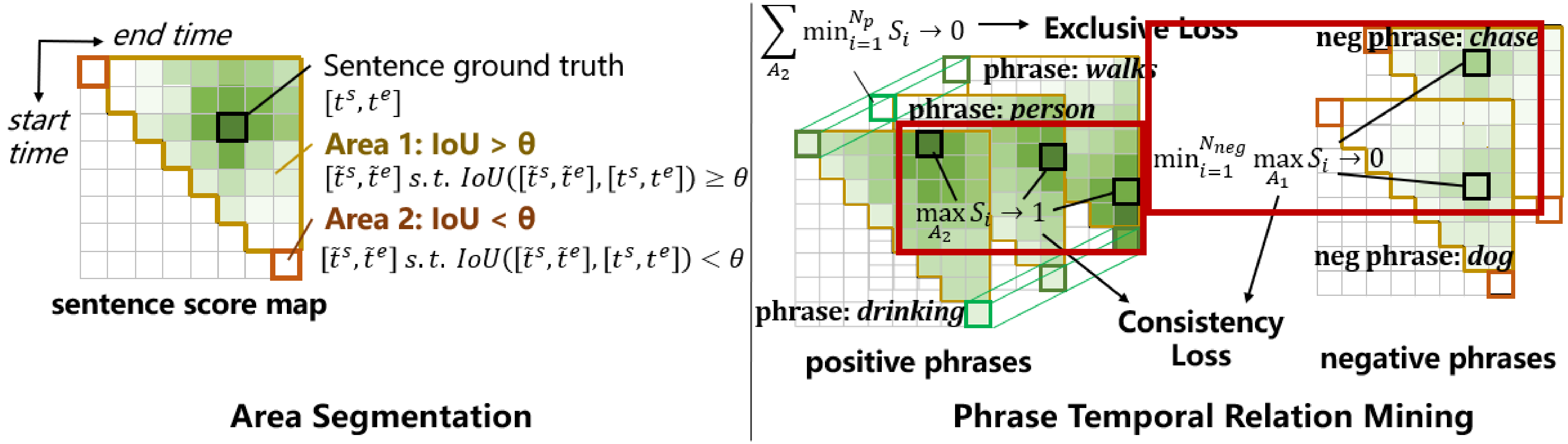


3.2 Temporal Relation Mining

Consistency

- **Paired sentence-video:** phrase-level prediction should **share** a period with the annotated sentence-level ground truth.
- **Unpaired sentence-video:** **at least one** phrase-level prediction does **not share** a period with the annotated ground truth.

$$\mathcal{L}_{con} = \max_{i=1}^{N_p}\left(L_f\left(\max_{(s,t)\in A_1} S_i^p[s,t],1\right)\right) + \min_{i=1}^{N_p}\left(L_f\left(\max_{(s,t)\in A_1} \hat{S}_i^p[s,t],0\right)\right)$$

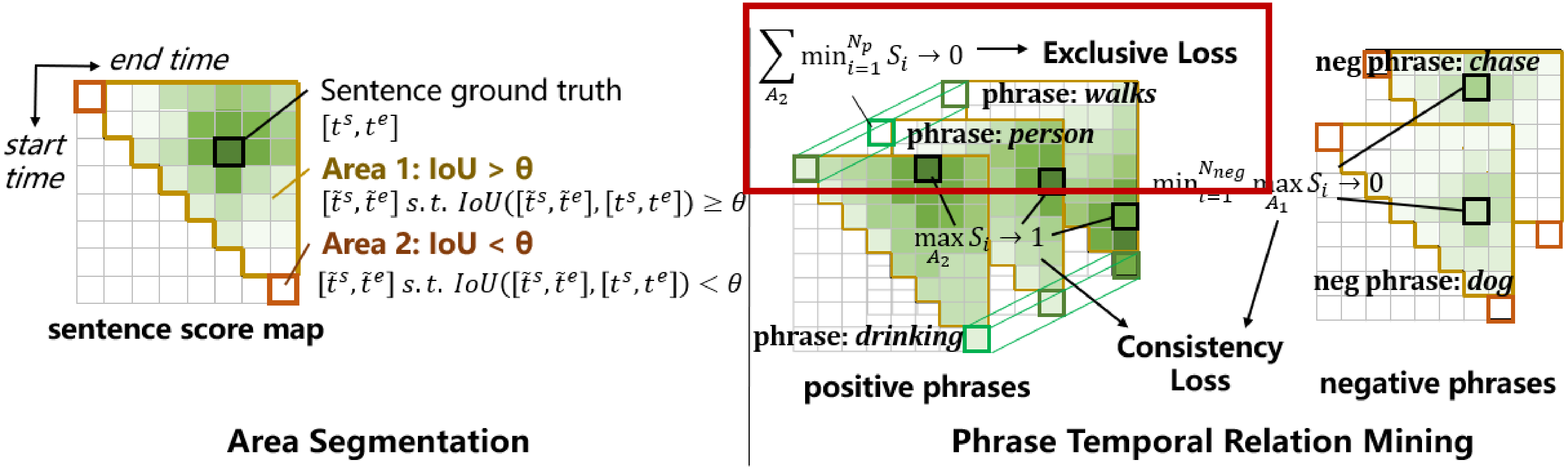


3.3 Temporal Relation Mining

Exclusiveness

- Each frame **outside** the ground truth is **not contained** in **at least one** phrase-level prediction

$$\mathcal{L}_{ex} = \frac{1}{|A_2|} \sum_{(s,t)\in A_2} L_f(\min_{i=1}^{N_p}(S_i^p[s,t]), 0)$$



3.4 Similarity Learning

Aims: Learn semantic relevance of sentence/phrase with each proposal

Sentence Score Map Refinement

- phrase-level score maps provide fine-grained information for sentence

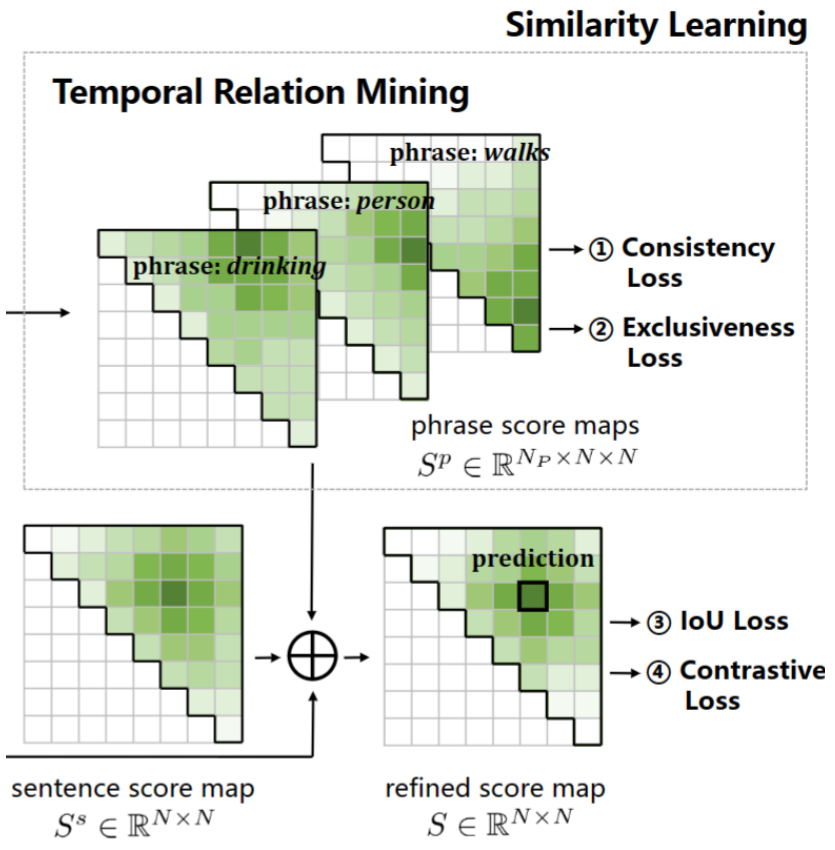
$$S = S^s + \sum \alpha_i S_i^p$$

$$\mathcal{L}_{iou} = -\frac{1}{C}\sum_{i=1}^{C}(y_i \log S_i + (1-y_i)\log(1-S_i)),$$

Sentence-level Contrastive Learning

- Use contrastive learning to provide more supervised signals

$$\mathcal{L}_{cont} = -(\sum_{s\in\mathbb{S}} \log p(v_s|s) + \sum_{v\in\mathbb{V}} \log p(s_v|v))$$



Experiments

Datasets

- ActivityNet Captions
- Charades-STA

Metrics

- Recall, IoU=m
- mIoU

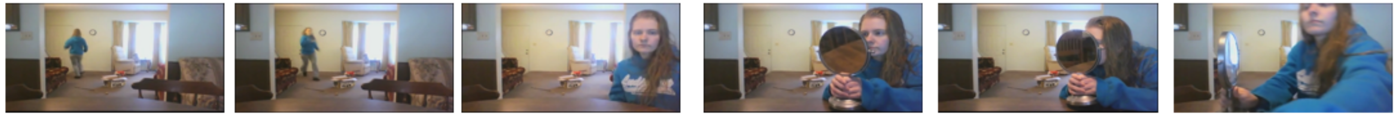
Evaluation for phrase

- Verb annotation from Temporal Action Localization task
- Action names as phrase queries

Query: A man in a red tank top is crossing the monkey bars.

(a) ActivityNet Captions

Query: Person runs to a table.



(b) Charades-STA

Charades-STA

Results

- Best performance on **phrase prediction**
- Phrase information can help **sentence prediction**

Method	feature	sentence prediction IoU=0.3	sentence prediction IoU=0.5	sentence prediction IoU=0.7	sentence prediction mIoU	phrase prediction IoU=0.3	phrase prediction IoU=0.5	phrase prediction IoU=0.7	phrase prediction mIoU
SAP (Chen and Jiang 2019)	VGG	—	27.42	13.36	—				
MAN (Zhang et al. 2019)	VGG	—	41.24	20.54	—				
LGI (Mun, Cho, and Han 2020)	VGG	57.20	40.70	20.13	38.75				
2D-TAN (Zhang et al. 2020b)	VGG	57.31	42.8	23.25	—	45.15	23.22	10.14	—
FVMR (Gao and Xu 2021)	VGG	—	42.36	24.14	—				
DRN (Zeng et al. 2020)	VGG	—	42.90	23.68	—				
SSCS (Ding et al. 2021)	VGG	—	43.15	25.54	—				
CBLN (Liu et al. 2021)	VGG	—	43.67	24.44	—				
CPN (Zhao et al. 2021)	VGG	**64.41**	46.08	25.06	**43.90**				
MMN (Wang et al. 2021b)	VGG	60.48	47.45	27.15	—	38.41	22.19	10.1	—
PLPNet (Li et al. 2022b)	VGG	57.82	41.88	20.56	39.12	46.24	22.94	7.69	28.46
TRM (ours)	VGG	60.67	**47.77**	**28.01**	42.77	**57.03**	**33.69**	**11.86**	**35.82**

*Experiments on ActivityNet Captions dataset can be found in the paper

Compositional Generalization

ActivityNet-CG[1] dataset

- **Novel-Composition:** unseen combination of seen phrases
- **Novel-Word:** unseen word

Results

- Best performance on **all test splits**
- Better **generalization**

	Method	Test-Trivial IoU=0.5	Test-Trivial IoU=0.7	Test-Trivial mIoU	Novel-Composition IoU=0.5	Novel-Composition IoU=0.7	Novel-Composition mIoU	Novel-Word IoU=0.5	Novel-Word IoU=0.7	Novel-Word mIoU
Weakly-supervised	WSLL (Duan et al. 2018)	11.03	4.14	15.07	2.89	0.76	7.65	3.09	1.13	7.10
RL-based	TSP-PRL (Wu et al. 2020)	34.27	18.80	37.05	14.74	1.43	12.61	18.05	3.15	14.34
Proposal-free	LGI (Mun, Cho, and Han 2020)	43.56	23.29	41.37	23.21	9.02	27.86	23.10	9.03	26.95
	VLSNet (Zhang et al. 2020a)	39.27	23.12	42.51	20.21	9.18	29.07	21.68	9.94	29.58
	VISA (Li et al. 2022a)	47.13	29.64	44.02	31.51	16.73	**35.85**	30.14	15.90	35.13
Proposal-based	TMN (Liu et al. 2018)	16.82	7.01	17.13	8.74	4.39	10.08	9.93	5.12	11.38
	2D-TAN (Zhang et al. 2020b)	44.50	26.03	42.12	22.80	9.95	28.49	23.86	10.37	28.88
	TRM (Ours)	**55.22**	**35.06**	**51.85**	**33.80**	**16.86**	35.80	**35.49**	**17.68**	**37.50**

[1]Li, et al. Compositional Temporal Grounding with Structured Variational Cross-Graph Correspondence Learning. CVPR, 2022.

Ablations on Charades-STA

Results

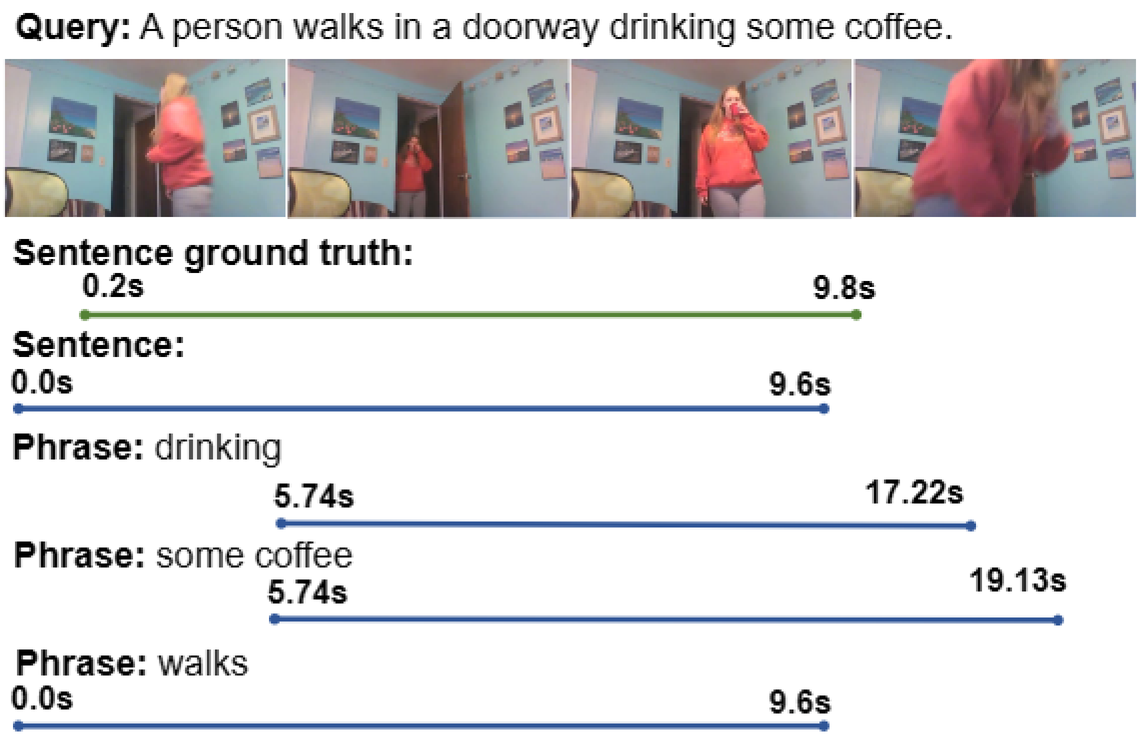
- Introducing phrase **without mining relationship** has **limited improvement**
- **Consistency** loss can greatly improve the performance
- Training with **only** exclusiveness loss has a **negative** impact
- **Consistency** loss and **exclusiveness** loss **together** can further improve the performance of **both sentences and phrase**

Method			Sentence prediction			Verb phrase prediction			Noun phrase prediction		
Phrase	Consistency	Exclusiveness	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
✗	✗	✗	60.48	47.45	27.15	38.41	22.19	10.01	33.13	8.17	3.15
✓	✗	✗	59.84	46.65	26.99	41.13	22.63	10.60	35.41	7.36	2.68
✓	✓	✗	60.22	46.56	27.31	56.69	30.85	10.85	71.12	51.67	8.57
✓	✗	✓	60.13	45.89	27.80	38.90	22.11	10.46	36.88	8.63	3.01
✓	✓	✓	**60.67**	**47.77**	**28.01**	**57.03**	**33.69**	**11.86**	**78.25**	**57.10**	**10.17**

Qualitative Results on Charades-STA

Observation

- Understands phrases: ‘drinking’, ‘some coffee’, and ‘walk
- Satisfy constraints of consistency and exclusiveness.



Conclusion

AAAI
Association for the Advancement of Artificial Intelligence

Phrase-level Temporal Relationship Mining (TRM)

- Consider **phrase-level prediction** in training **without** phrase-level annotation
- Propose the **consistency** and **exclusiveness** constraints

- Performance improved on **both** phrase and sentence prediction
- Better **interpretability**, and **generalization** performance

Thank you!