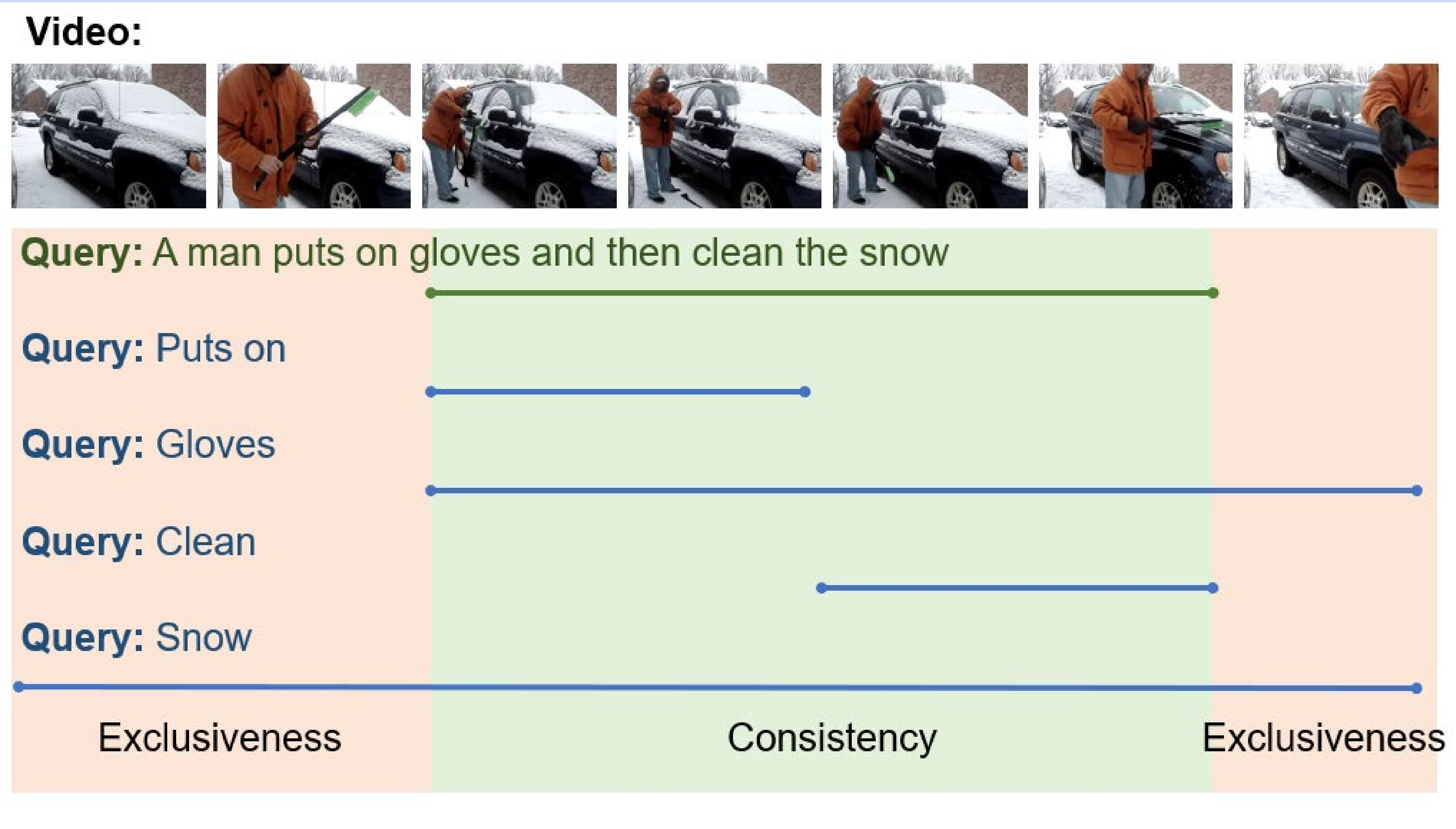
Phrase-level Temporal Relationship Mining for Temporal Sentence Localization

Minghang Zheng[1], Sizhe Li[1], Qingchao Chen[1], Yuxin Peng[1], Yang Liu[1,2]

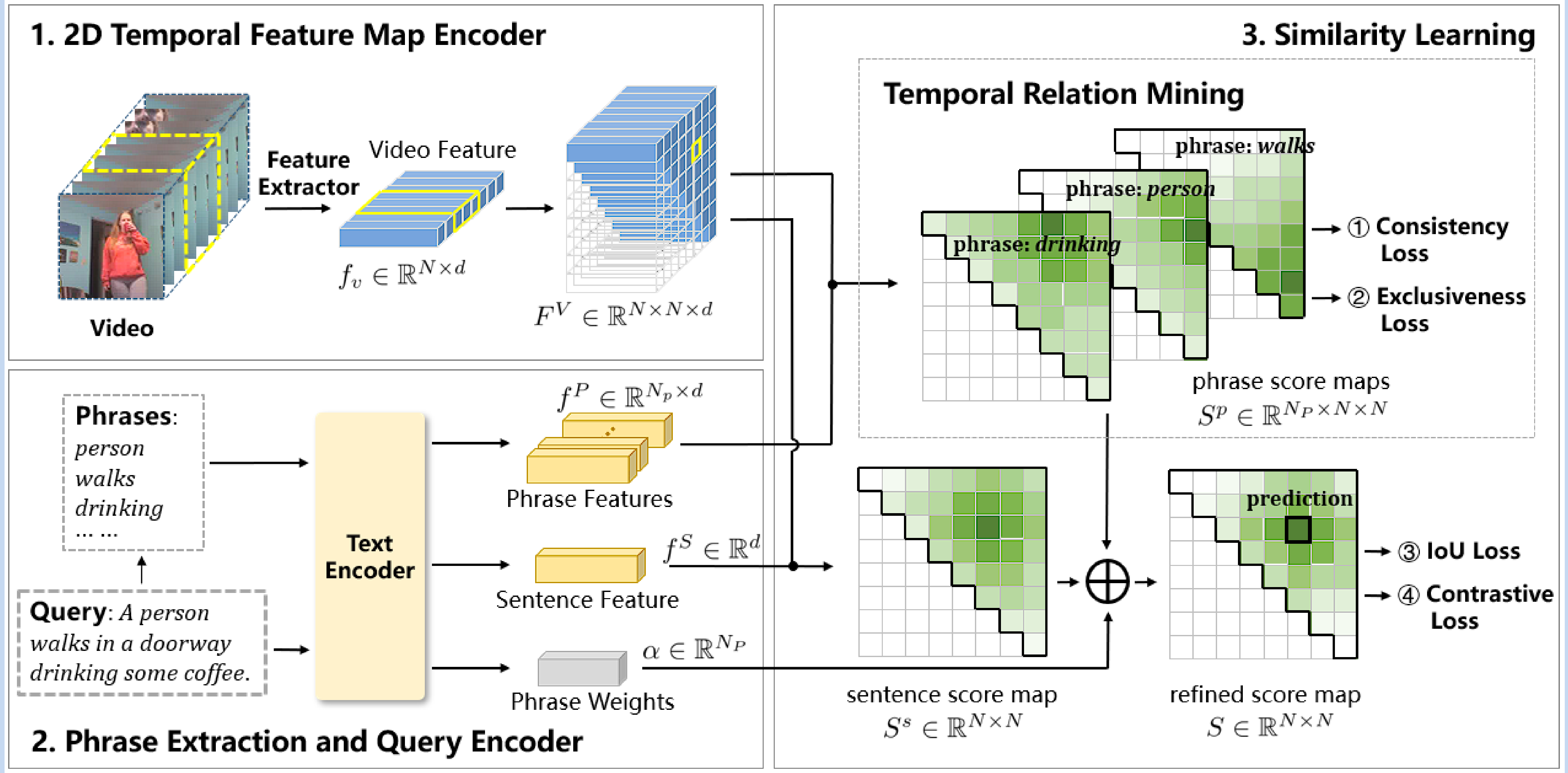
[1]Peking University [2]Beijing Institute for General Artificial Intelligence

Introduction



- **Task:** Temporal sentence grounding
  - **Inputs:** Video + Sentence query
  - **Outputs:** Target video clip (start and end timestamps)
- **Observations:** Existing work can not deal with the **phrase-level**
- **Problems:**
  - Insufficient understandings of relationship between **simple visual and language concepts**
  - Questioned model **interpretability** and **robustness**
- **Difficulty:** No phrase-level annotation
- **Solution:** Phrase-level Temporal Relationship Mining (TRM)
  - Consider **phrase-level** prediction
  - Mining **temporal relationship** between phrase and sentence
  - Two principles: **Consistency** & **Exclusiveness**

Method


1. 2D Temporal Feature Map Encoder
Video
Feature Extractor
Video Feature
$f_v \in \mathbb{R}^{N\times d}$
$F^V \in \mathbb{R}^{N\times N\times d}$
2. Phrase Extraction and Query Encoder
Phrases: person walks drinking … …
Query: A person walks in a doorway drinking some coffee.
Text Encoder
$f^P \in \mathbb{R}^{N_p\times d}$
Phrase Features
$f^S \in \mathbb{R}^d$
Sentence Feature
$\alpha \in \mathbb{R}^{N_P}$
Phrase Weights
3. Similarity Learning
Temporal Relation Mining
phrase: walks
phrase: person
phrase: drinking
① Consistency Loss
② Exclusiveness Loss
phrase score maps
$S^p \in \mathbb{R}^{N_P\times N\times N}$
sentence score map
$S^s \in \mathbb{R}^{N\times N}$
prediction
refined score map
$S \in \mathbb{R}^{N\times N}$
③ IoU Loss
④ Contrastive Loss

- **2D Temporal Feature Map Encoder**
  - Generate 2D visual feature map F_{ij}^V
- **Phrase Extraction and Query Encoder**
  - Extract phrase form pretrained SRLBERT
  - Encode sentence feature f^S and phrase feature f^P
- **Similarity Learning**
  - Calculate sentence score map $S^s = F^{VT} f^s$ and phrase score map $S_i^p = F^{VT} f_i^p$
  - **Consistency:** Phrase-level prediction should **share** a period with the annotated sentence-level ground truth
  - **Exclusiveness:** Each frame **outside** the ground truth is **not contained** in **at least** one phrase-level prediction
  - **Sentence Score Map Refinement:** phrase-level score maps provide fine-grained information for sentence

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

Experiments

- Charades-STA

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)		—	41.24	20.54	—				
LGI (Mun, Cho, and Han 2020)		57.20	40.70	20.13	38.75				
2D-TAN (Zhang et al. 2020b)		57.31	42.8	23.25	—	45.15	23.22	10.14	—
FVMR (Gao and Xu 2021)		—	42.36	24.14	—				
DRN (Zeng et al. 2020)		—	42.90	23.68	—				
SSCS (Ding et al. 2021)		—	43.15	25.54	—				
CBLN (Liu et al. 2021)		—	43.67	24.44	—				
CPN (Zhao et al. 2021)		**64.41**	46.08	25.06	**43.90**				
MMN (Wang et al. 2021b)		60.48	47.45	27.15	—	38.41	22.19	10.1	—
PLPNet (Li et al. 2022b)		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**

Ablation Study

- Compositional 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**

- Effectiveness of Proposed Modules

Method Phrase	Method Consistency	Method Exclusiveness	Sentence prediction IoU=0.3	Sentence prediction IoU=0.5	Sentence prediction IoU=0.7	Verb phrase prediction IoU=0.3	Verb phrase prediction IoU=0.5	Verb phrase prediction IoU=0.7
✗	✗	✗	60.48	47.45	27.15	38.41	22.19	10.01
✓	✗	✗	59.84	46.65	26.99	41.13	22.63	10.60
✓	✓	✗	60.22	46.56	27.31	56.69	30.85	10.85
✓	✗	✓	60.13	45.89	27.80	38.90	22.11	10.46
✓	✓	✓	**60.67**	**47.77**	**28.01**	**57.03**	**33.69**	**11.86**

Acknowledgements

This work was supported by the grants from the National Natural Science Foundation of China (61925201, 62132001, U21B2025), Zhejiang Lab (NO. 2022NB0AB05), the National Key R&D Program of China (2021YFF0901502) and CAAI-Huawei MindSpore Open Fund.

Code