Generating Structured Pseudo Labels for Noise-resistant Zero-shot Video Sentence Localization

Minghang Zheng[1], **Shaogang Gong**[2], **Hailin Jin**[3], **Yuxin Peng**[1, 4], **and Yang Liu**[1,5*]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Queen Mary University of London, [3]Adobe Research
[4] National Key Laboratory for Multimedia Information Processing, Peking University
[5] National Key Laboratory of General Artificial Intelligence, BIGAI
{minghang, pengyuxin, yangliu}@pku.edu.cn
s.gong@qmul.ac.uk, hljin@adobe.com

Task

Zero-shot Temporal Sentence Localization

- **Inputs:** Video + Sentence query
- **Outputs:** Target video clip (start and end timestamps)
- **Zero-shot Setting:** No manual annotation required

Video:

Query:

A man puts on gloves and then clean the snow

Output

5.5s — 15.1s

} **Unavailable**

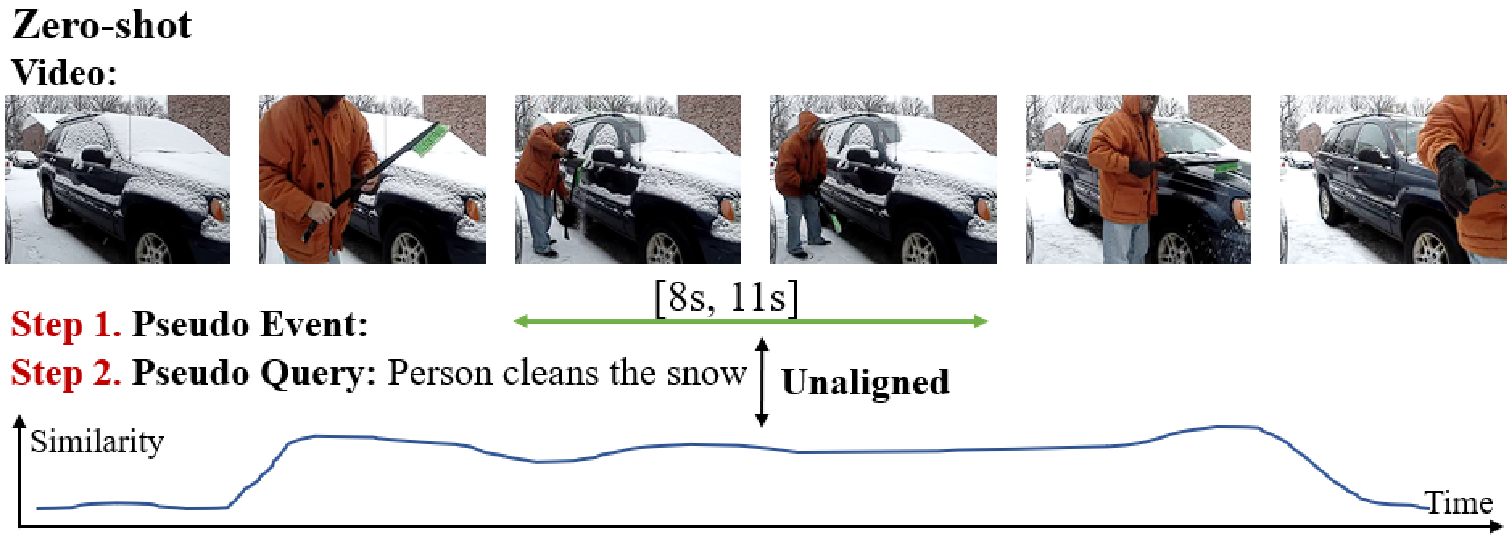
Motivation

Existing zero-shot methods:

- Generating **pseudo-events** and **pseudo-queries**
- **Training** with pseudo-event and pseudo-query

Drawbacks:

- 1. Pseudo queries are **too simple**
- 2. **Unalignment** between pseudo-events and pseudo-queries
- 3. Ignoring the **noise** in the pseudo labels

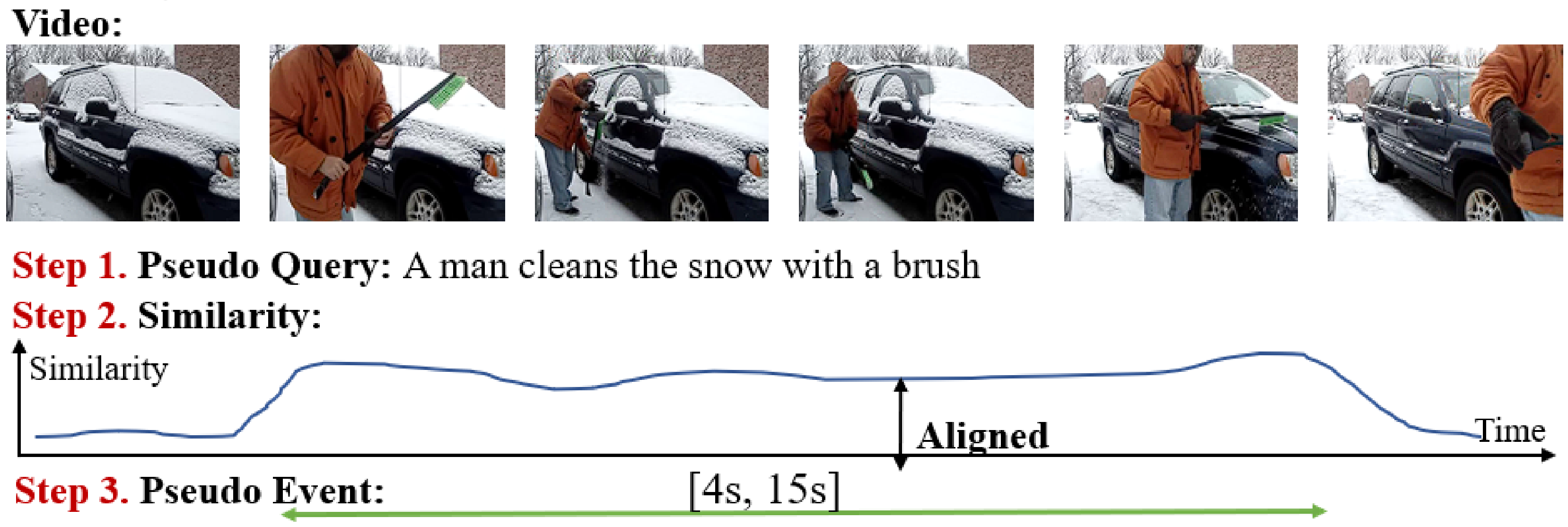


(c) Existing pipeline

Method

Our Structured Pseudo-Label (SPL) generation:

- Generate **free-form** pseudo-queries using image description models
- Generate pseudo-events based on the **event temporal structure**
  - the video **inside** the event has a **high** correlation with the query
  - the video **outside** the event has a **low** correlation with the query
- **Reduce noise** during training
  - Sample **re-weight** and label **refinement**



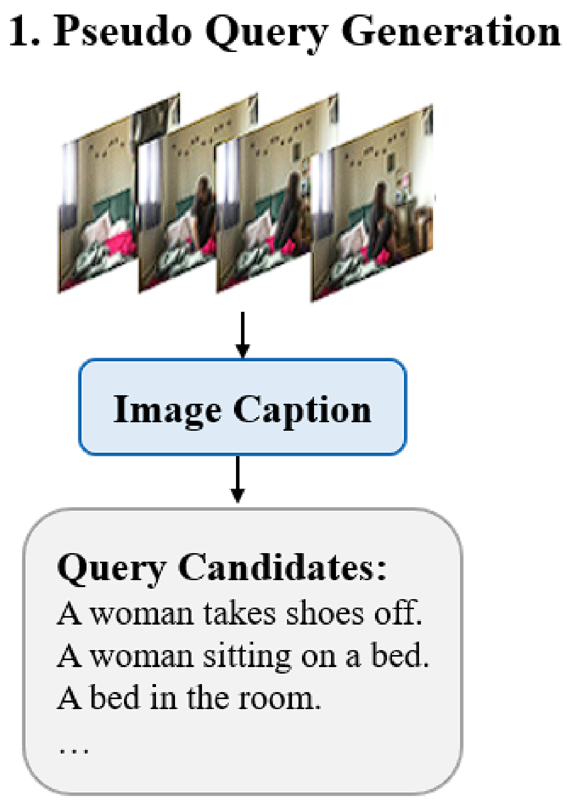
(d) Ours pipeline

1. Pseudo Query Generation

Aims: Generate **free-form** pseudo-queries

- Densely sample video frames
- Generate pseudo queries from video frames using pretrained BLIP[1] model

[1]Li, Junnan, et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." International Conference on Machine Learning. ICML, 2022.



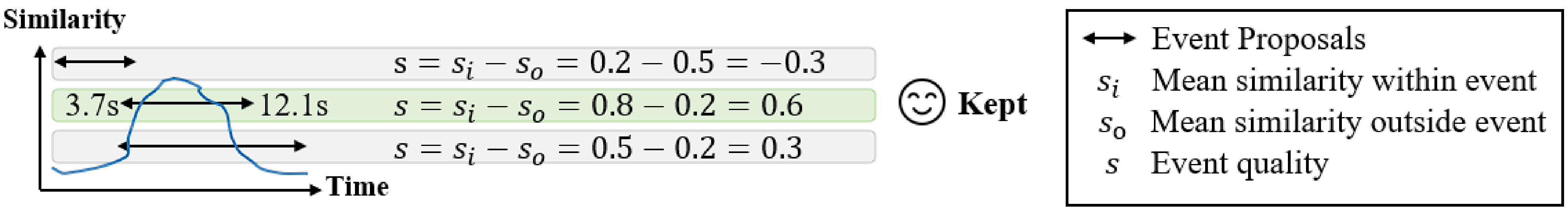
2. Pseudo Event Generation

Aims: Generate pseudo-events based on **event temporal structure**

- Videos **within** events have **high relevance** to queries
- Videos **outside** events have **low relevance** to queries

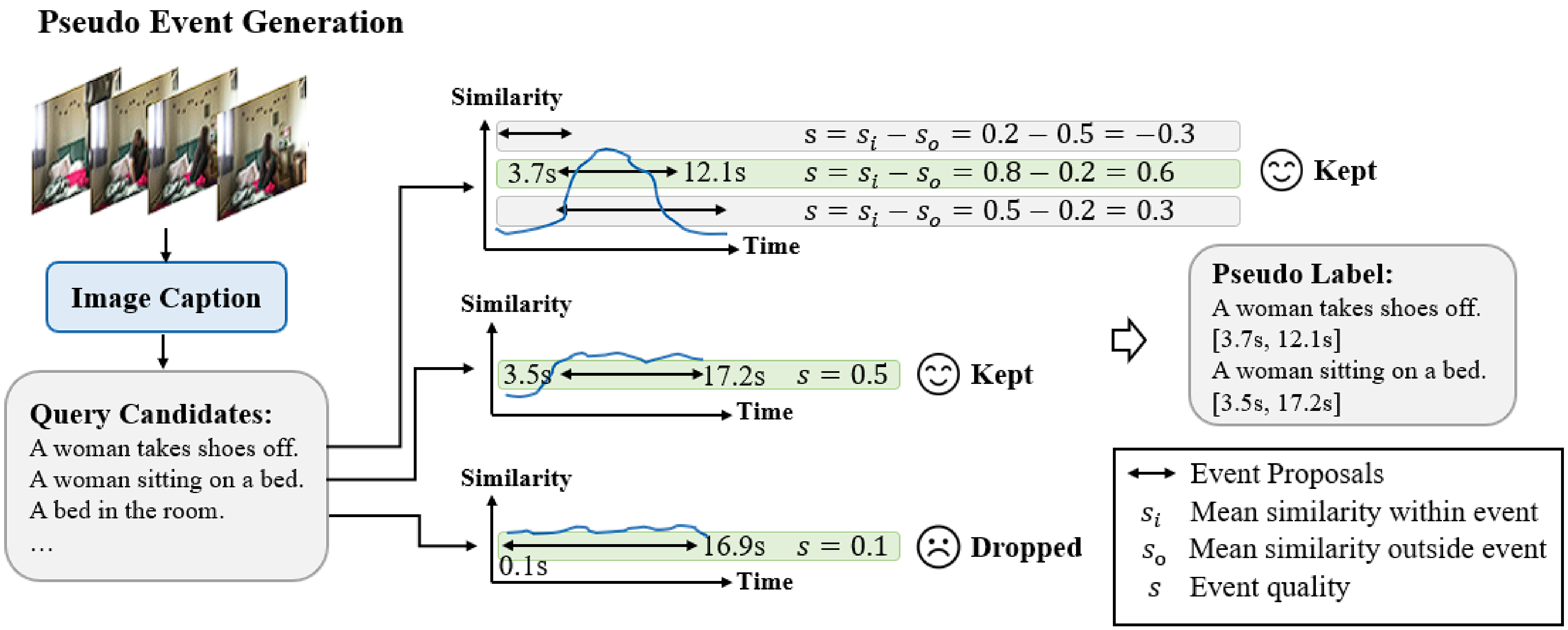
- Calculate similarity S between pseudo-query and video frames
- Event quality: $Q_{ik} = \frac{1}{N_{p_k}} \sum_{j \in p_k} S_{ij} - \frac{1}{N - N_{p_k}} \sum_{j \notin p_k} S_{ij}$
- Choose the event proposal with highest quality

Pseudo Event Generation



2. Pseudo Event Generation

- Filter out **low-quality** pseudo-query event pairs
  - Keep top K pseudo-query-event pairs with high event quality
  - Use non-maximum suppression to eliminate pseudo-query-event pairs with high event overlap.



3. Training with Noisy Pseudo Labels

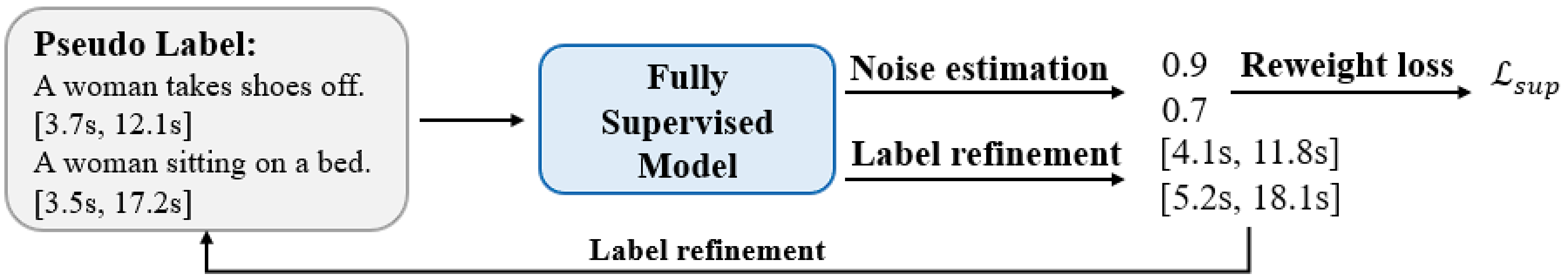
Aims: Train model using pseudo-labels and **reduce noise** in the pseudo-labels

- **Sample re-weight:** Estimate noise based on the confidence score s^{conf} and the IoU s^{iou} of predictions and pseudo-label and weight the sample loss

$$w = \alpha \frac{1}{1 - s^{iou}} + (1 - \alpha) \frac{1}{1 - s^{conf}}$$

- **Label refinement:** If the model's prediction confidence is high, consider the prediction as a new pseudo-label.

Training with Noisy Pseudo Label



Experiments

Datasets

- ActivityNet Captions
- Charades-STA

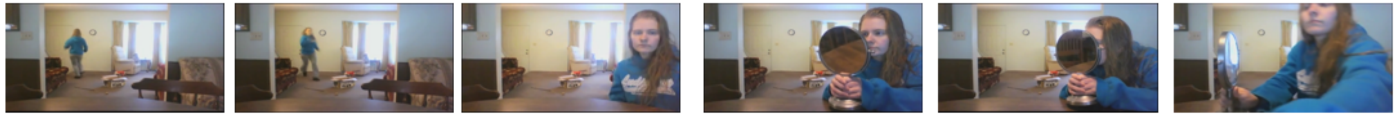
Metrics

- R@m
- mIoU

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

Comparing with SOTA

Results

- **Best zero-shot performance** on most metrics

Method	Sup.	Charades-STA R@0.3	Charades-STA R@0.5	Charades-STA R@0.7	Charades-STA mIoU	ActivityNet Captions R@0.3	ActivityNet Captions R@0.5	ActivityNet Captions R@0.7	ActivityNet Captions mIoU
2D-TAN (Zhang et al., 2020)	fully	-	39.81	23.25	-	58.75	44.05	27.38	-
EMB (Huang et al., 2022)	fully	**72.50**	58.33	39.25	**53.09**	**64.13**	44.81	26.07	**45.59**
MGSL-Net (Liu et al., 2022)	fully	-	**63.98**	**41.03**	-	-	**51.87**	**31.42**	-
CRM (Huang et al., 2021)	weakly	53.66	34.76	16.37	-	55.26	32.19	-	-
CNM* (Zheng et al., 2022a)	weakly	60.39	35.43	15.45	-	55.68	**33.33**	-	-
CPL (Zheng et al., 2022b)	weakly	**66.40**	**49.24**	**22.39**	-	**55.73**	31.37	-	-
Gao et al.* (Gao and Xu, 2021)	no	46.69	20.14	8.27	-	46.15	26.38	11.64	-
PSVL* (Nam et al., 2021)	no	46.47	31.29	14.17	31.24	44.74	30.08	14.74	29.62
PZVMR* (Wang et al., 2022)	no	46.83	33.21	18.51	32.62	45.73	31.26	**17.84**	30.35
Kim et al.* (Kim et al., 2023)	no	52.95	37.24	19.33	36.05	47.61	**32.59**	15.42	31.85
SPL* (ours)	no	**60.73**	**40.70**	**19.62**	**40.47**	**50.24**	27.24	15.03	**35.44**

more experiments and ablation studies can be found in paper

Conclusion

- Propose a zero-shot video sentence localization method based on **structured pseudo-label generation** that is **robust to noise**

- Generate **free-form** pseudo-queries and generate pseudo-events based on **event temporal structure**.

- Reduce the influence of noise in pseudo-labels by **sample reweight** and **label refinement**

- **Best zero-shot performance** on two datasets

Thank you!

Code