NEURAL INFORMATION PROCESSING SYSTEMS

PC-Net: Weakly Supervised Compositional Moment Retrieval via Proposal-Centric Network

Mingyao Zhou[1] **Hao Sun**[1*] **Wei Xie**[1] **Ming Dong**[1] **Chengji Wang**[1] **Mang Ye**[2]

Central China Normal University & Wuhan University

Presenter **Mingyao Zhou**

Date 08/10/2025

Contents

Introduction

- Video Moment Retrieval
- Weakly Supervised Compositional Moment Retrieval

Video Moment Retrieval

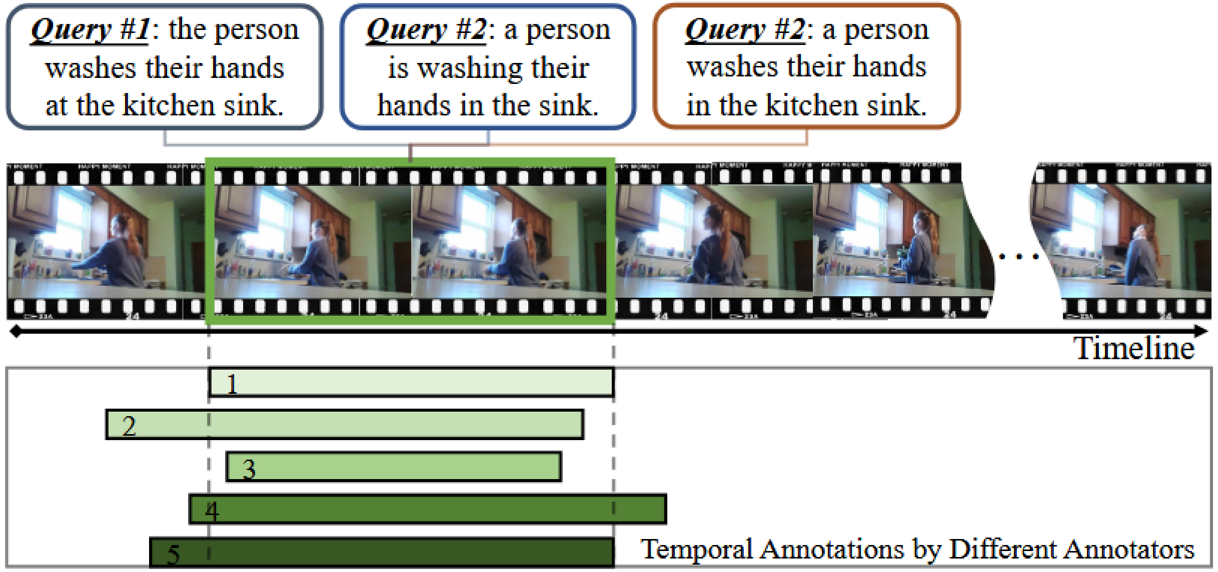
- With the exponential growth of video content, aiming at localizing relevant video moments based on natural language queries, video moment retrieval (VMR) has gained significant attention [1]

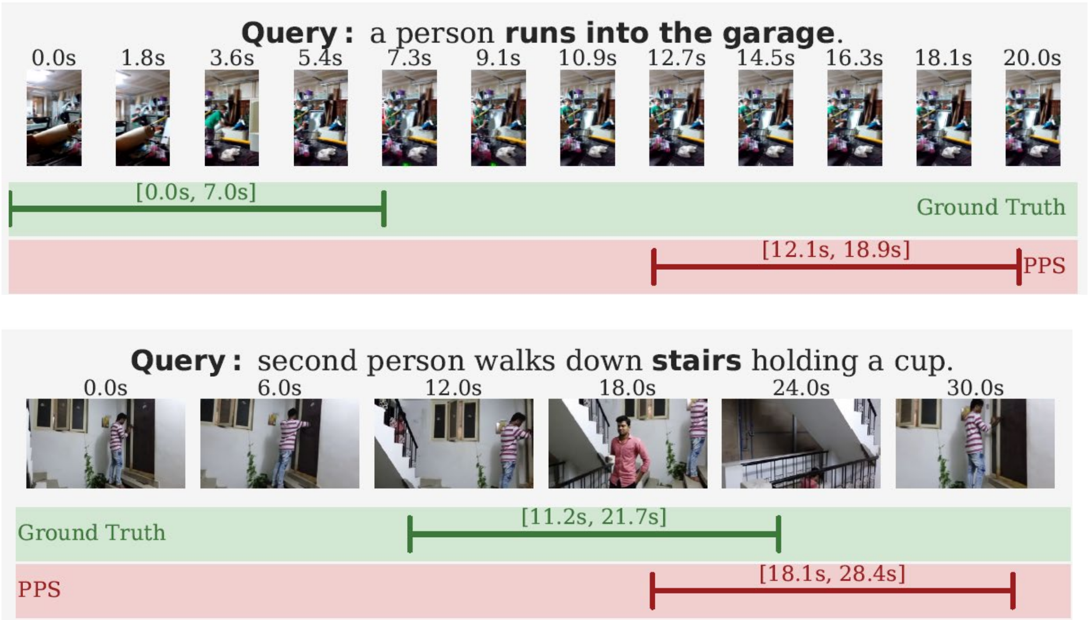
Schematic diagram of video moment retrieval

[1] Dong, Jianfeng, et al. "Temporal sentence grounding with relevance feedback in videos." Advances in Neural Information Processing Systems 37 (2024): 43107-43132.

Weakly Supervised Video Moment Retrieval

- Existing weakly supervised VMR methods focus on designing various feature modeling and modal interaction modules to alleviate the reliance on precise temporal annotations. However, these methods have **poor generalization capabilities on compositional queries** with novel syntactic structures or vocabulary in real-world scenarios [2]

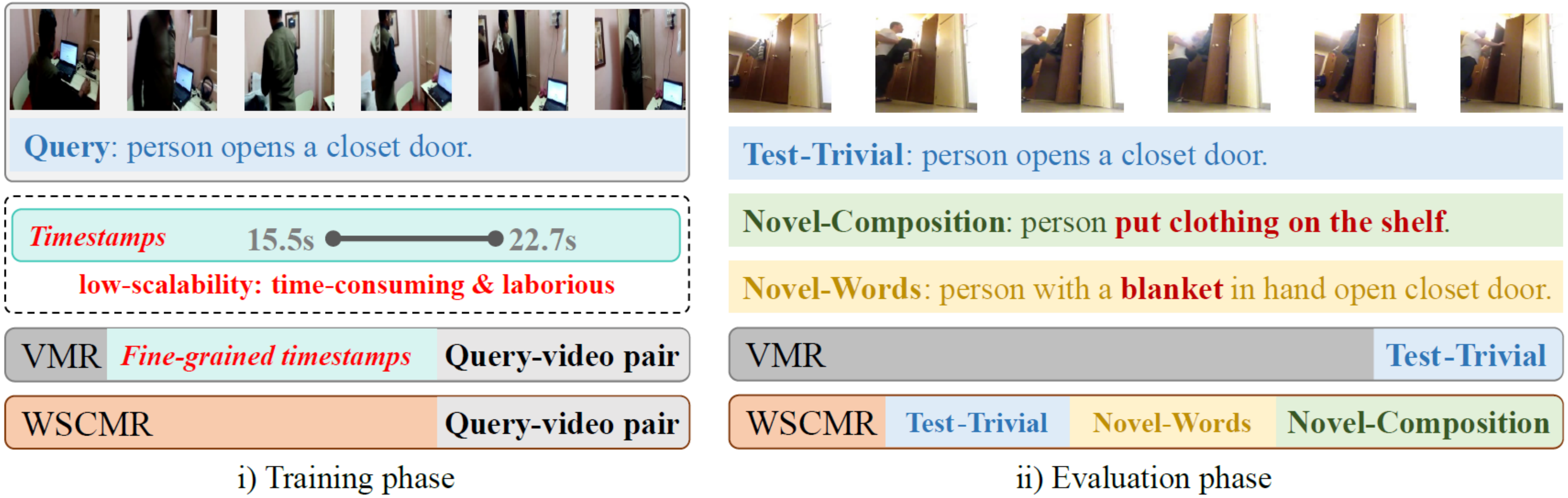




[2] Li, Juncheng, et al. "Compositional temporal grounding with structured variational cross-graph correspondence learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

Weakly Supervised Compositional Moment Retrieval

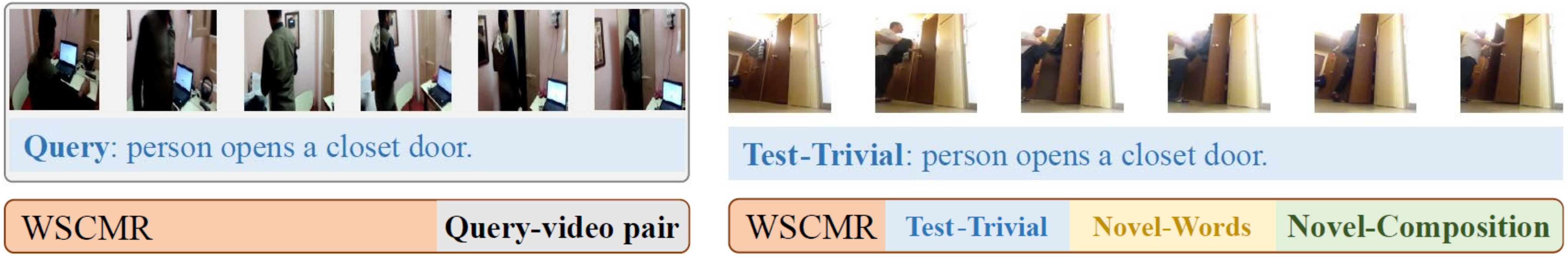
- We propose a new task: weakly supervised compositional moment retrieval (WSCMR). This task trains models using only video-query pairs without precise temporal annotations, while enabling generalization to complex compositional queries.



Weakly Supervised Compositional Moment Retrieval

- Weakly Supervised Compositional Moment Retrieval (WSCMR) is close to practical application
  - ✓ **does not require precise timestamps for training**
  - ✓ **includes generalization evaluation on compositional queries** with unseen grammatical structures or words

- The challenges lie in
  1. modeling fine-grained cross-modal semantic associations **solely based on video-level weak supervision**
  2. **generalizing to queries that contain new grammar, new vocabulary, and complex temporal semantics**



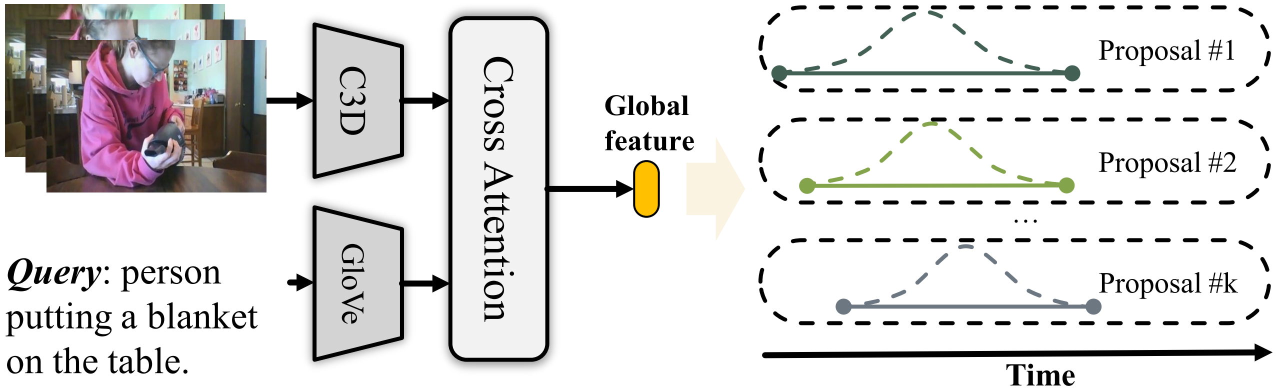
i) Training phase ii) Evaluation phase

Methodology

- The deficiencies of the existing methods
- Implementation details of the proposed Proposal-Centric Network (PC-Net)

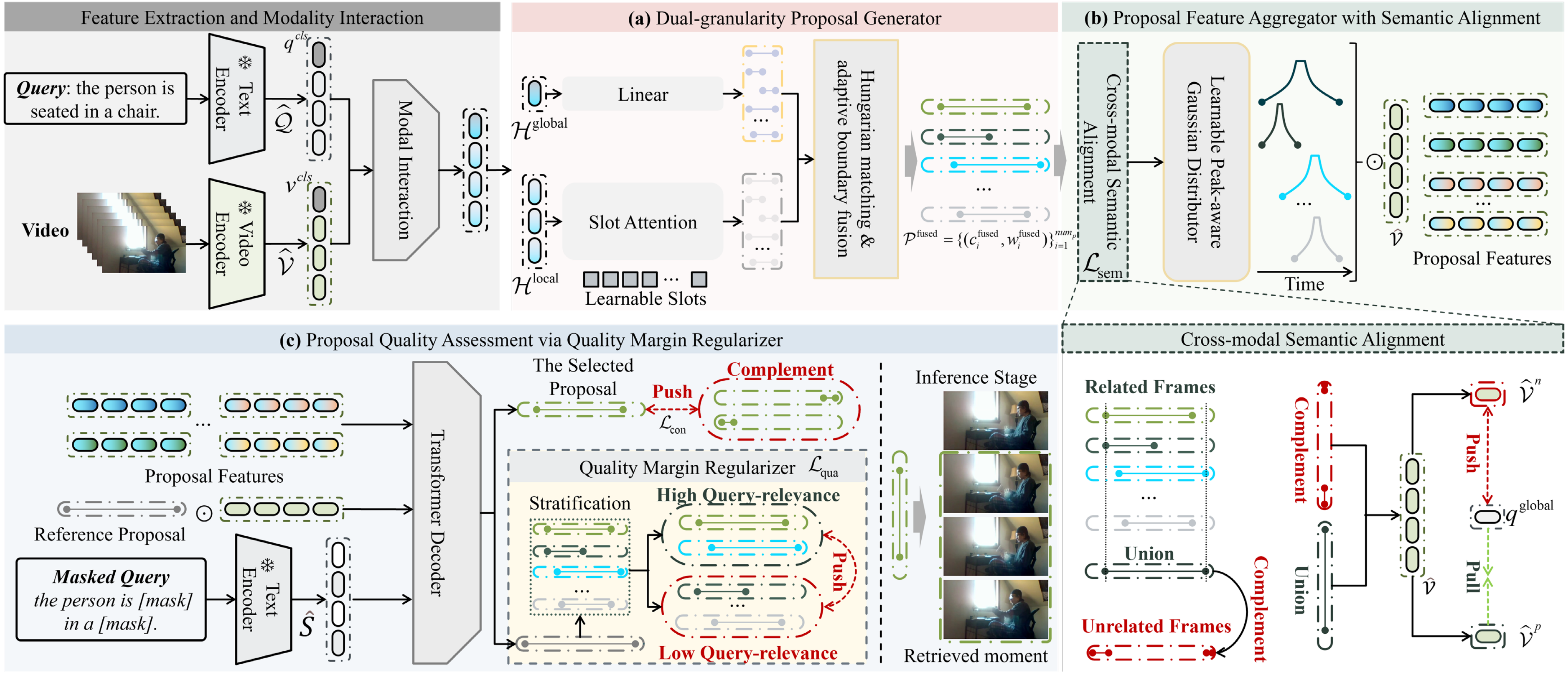
The deficiencies of the existing methods

- An intuitive approach to address the proposed WSCMR is to leverage existing weakly supervised models [3, 4]. However, their inherent limitations hinder them from effectively handling compositional queries.
  - **Coarse Boundary Generation:** Current methods rely on global video-query matching to generate proposal boundaries, which lacks fine-grained temporal perception and fails to handle queries with explicit temporal logic
  - **Inadequate Feature Aggregation:** Using a fixed Gaussian distribution for feature aggregation ignores the semantic gap between frames and queries, as well as varying action durations, resulting in poorly discriminative proposal features
  - **Ineffective Negative Sampling:** Constructing negative samples solely from the proposal with the lowest reconstruction loss discards partially relevant ones, hindering the model's ability to learn fine-grained visual-query associations and undermining compositional generalization



[3] Zheng, Minghang, et al. "Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[4] Zheng, Minghang, et al. "Weakly supervised video moment localization with contrastive negative sample mining." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 3. 2022.

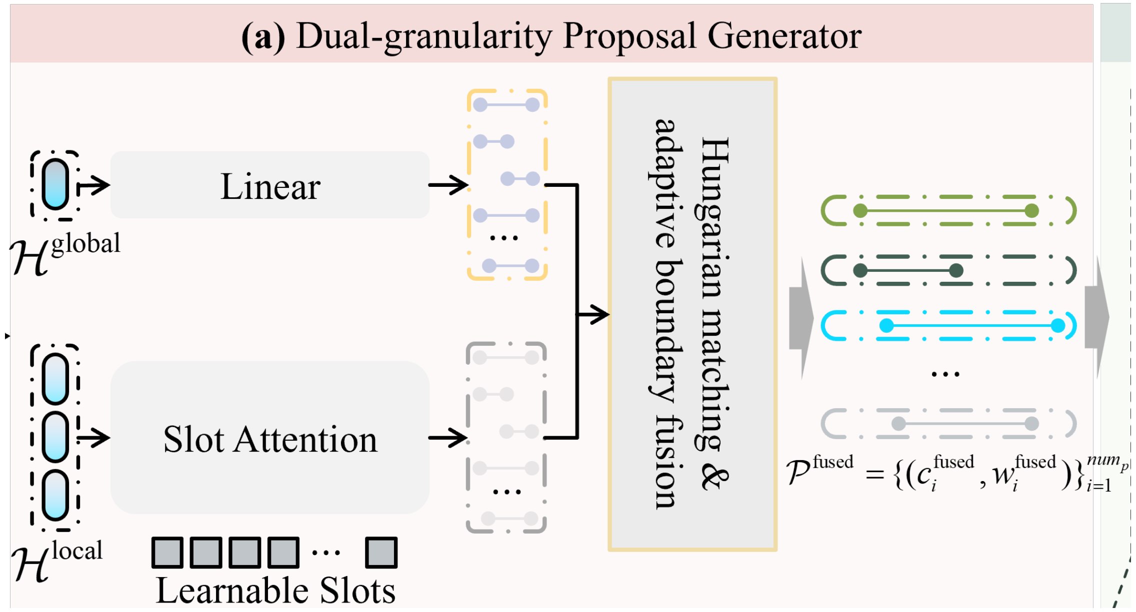
Proposal-Centric Network (PC-Net)



- The PC-Net tackles WSCMR with three key modules: **a dual-granularity generator for precise boundaries**, **a discriminative feature aggregator**, and **a query margin regularizer to suppress spurious correlations**

Dual-granularity Proposal Generator

- To capture global semantic consistency and local temporal precision jointly, for generating boundaries with both holistic scene understanding and fine-grained temporal awareness, the dual-granularity proposal generator is constructed
  - Firstly, global and local proposals are obtained through global-local multimodal features.
  - Then, proposals are matched through the Hungarian algorithm [5] and adaptively fused to obtain the final proposal set



$$\mathcal{P}^{\text{global}} = \text{Linear}(\mathcal{H}^{\text{global}}) \in \mathbb{R}^{num_p \times 2}$$

$$\mathcal{P}_k^{\text{local}} = \text{Softmax}\left(\frac{(\mathcal{H}^{\text{local}})(\mathcal{P}_{k-1}^{\text{local}})^{\mathsf{T}}}{\sqrt{d}}\right) \cdot \mathcal{H}^{\text{local}} + \mathcal{P}_{k-1}^{\text{local}}$$

$$\Pi^* = \arg\min_{\Pi \in \mathcal{A}_N} \sum_{i=1}^{N} \left\| \begin{bmatrix} c_i^{\text{global}} \\ w_i^{\text{global}} \end{bmatrix} - \begin{bmatrix} c_{\Pi(i)}^{\text{local}} \\ w_{\Pi(i)}^{\text{local}} \end{bmatrix} \right\|_2$$

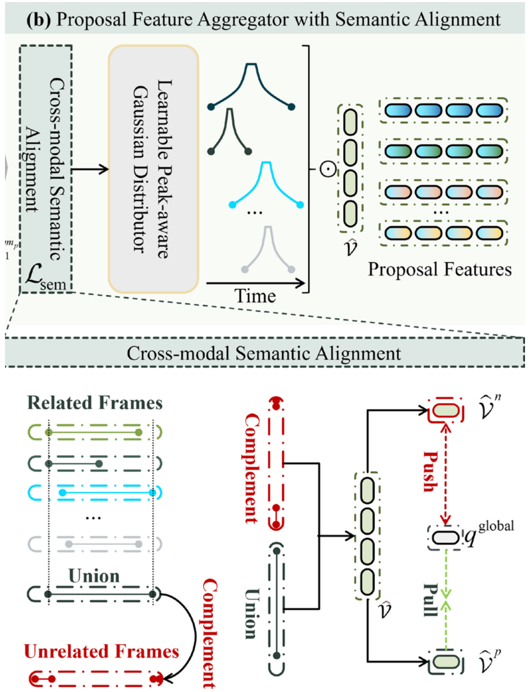
$$c_i^{\text{fused}} = \sigma(\alpha) \cdot c_i^{\text{global}} + [1 - \sigma(\alpha)] \cdot c_{\Pi^*(i)}^{\text{local}}$$

$$w_i^{\text{fused}} = \sigma(\alpha) \cdot w_i^{\text{global}} + [1 - \sigma(\alpha)] \cdot w_{\Pi^*(i)}^{\text{local}}$$

[5] Yu, Tianshu, et al. "Learning deep graph matching with channel-independent embedding and hungarian attention." International conference on learning representations. 2019.

Proposal Feature Aggregator with Semantic Alignment

- To bridge modality gap and fit the diversity of action durations, a proposal feature aggregator with two components is constructed
  - feature triplets of queries, relevant video segments, and irrelevant video segments are constructed to map them into a unified semantic space based on contrastive learning
  - Learnable Peak-aware Gaussian Distributor is used to adaptively adjust the peak area and fit the duration of variable actions



$$\hat{\mathcal{V}}^{p} = \frac{1}{|M_i|}\sum_{t\in M_i}\hat{\mathcal{V}}_t \in \mathbb{R}^{1\times d}, \hat{\mathcal{V}}^{n} = \frac{1}{T-|M_i|}\sum_{t\notin M_i}\hat{\mathcal{V}}_t \cdot \mathbf{1}_{\{|M_i|<T\}} + \frac{1}{T}\sum_{t=1}^{T}\hat{\mathcal{V}}_t \cdot \mathbf{1}_{\{|M_i|=T\}}$$

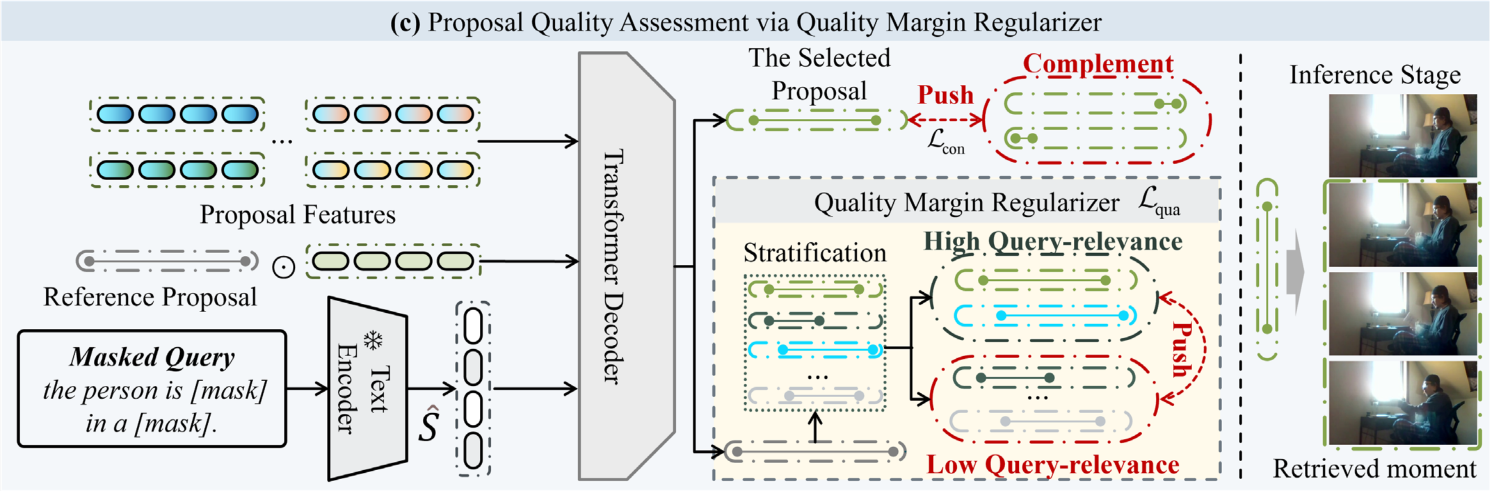
$$\mathcal{L}_{\text{sem}} = \frac{1}{num_p}\sum_{i=1}^{num_p}\max\left(0, \text{sim}(q^{\text{global}}, \hat{\mathcal{V}}^{n}) - \text{sim}(q^{\text{global}}, \hat{\mathcal{V}}^{p}) + \gamma\right)$$

$$M_i(t) = \frac{1}{1+e^{-1000\cdot\eta_i(t)}}, \text{where } \eta_i(t) = \beta\sigma_i - |x_t - c_i^{\text{fused}}|$$

$$W_i(t) = G_i(t)\cdot(1-M_i(t)) + M_i(t)$$

Proposal Quality Assessment via Quality Margin Regularizer

- Single negative sample in existing methods limits learning of subtle semantic associations.
- Quality Margin Regularizer
  - Dynamically groups proposals by reconstruction quality.
  - Enhances semantic correlation via inter-group contrast.



$$\mathcal{L}_i^{re} = -\sum_{j=1}^{N-1} \log P\left(s_{j+1} \,\middle|\, \hat{\mathcal{V}} \odot W_i, \hat{S}_{1:j}\right)$$

$$\mathcal{L}_{\text{high}} = \frac{1}{|\mathcal{X}|} \sum_{i \in \mathcal{X}} \mathcal{L}_i^{re}, \mathcal{X} = \left\{ i \,\middle|\, \mathcal{L}_i^{re} < \mathcal{L}_r^{re} \right\}$$

$$\mathcal{L}_{\text{low}} = \frac{1}{|\mathcal{Y}|} \sum_{i \in \mathcal{Y}} \mathcal{L}_i^{re}, \mathcal{Y} = \left\{ i \,\middle|\, \mathcal{L}_i^{re} \geq \mathcal{L}_r^{re} \right\}$$

$$\mathcal{L}_{\text{qua}} = \max\left(\mathcal{L}_{\text{high}} - \mathcal{L}_{\text{low}} + \theta_3, 0\right)$$

Experiments

- Comparison with SOTAs
- Ablation Study
- Qualitative Results

Comparison with SOTAs

- **Comparison on the Charades-CG (left) and ActivityNet-CG (right) datasets**

Supervision	Method	Params	Test-Trivial			Novel-Composition			Novel-Word		
			R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
Full Supervision	TMN [49]	-	18.75	8.16	19.82	8.68	4.07	10.14	9.43	4.96	11.23
	TSP-PRL [22]	-	39.86	21.07	38.41	16.30	2.04	13.52	14.83	2.61	14.03
	VSLNet [50]	-	45.91	19.80	41.63	24.25	11.54	31.43	25.60	10.07	30.21
	2D-TAN [18]	-	48.06	27.10	43.72	32.74	15.25	31.50	37.12	18.99	35.04
	2D-TAN$_{SSL}$ [51]	-	53.91	31.82	46.84	35.42	17.95	33.07	43.60	25.32	39.32
	LGI [52]	-	49.45	23.80	45.01	29.42	12.73	30.09	26.48	12.47	27.62
	MS-2D-TAN [53]	-	57.85	37.63	50.51	43.17	23.27	38.06	45.76	27.19	40.80
	MS-2D-TAN$_{SSL}$ [51]	-	58.14	37.98	50.58	46.54	25.10	40.00	50.36	28.78	43.15
	VISA [9]	-	53.20	26.52	47.11	45.41	22.71	42.03	42.35	20.88	40.18
	Deco [8]	-	58.75	28.71	49.06	47.39	21.06	40.70	-	-	-
	Moment-DETR [54]	-	49.48	28.04	44.82	39.42	18.62	36.61	46.76	24.75	41.70
	Moment-DETR$_S$ [1]	-	57.14	33.85	49.32	44.65	23.21	39.86	47.05	24.32	41.57
	QD-DETR [55]	7.12M	59.24	33.43	50.92	42.30	21.09	38.55	46.04	26.33	42.89
	QD-DETR$_S$ [1]	7.12M	60.66	38.60	52.53	50.23	27.69	44.14	55.25	35.25	48.10
Weak Supervision	WSSL [31]	-	15.33	5.46	18.31	3.61	1.21	8.26	2.79	0.73	7.92
	CNM [10]	2.52M	36.37	15.25	37.88	25.04	9.12	30.79	31.37	13.24	34.38
	CPL [7]	3.01M	53.04	24.71	45.82	40.79	16.15	37.46	42.45	21.44	39.20
	CCR [32]	9.01M	50.58	24.61	45.62	39.57	16.15	37.03	41.73	21.15	38.19
	QMN [6]	12.51M	51.65	22.64	45.85	40.67	15.72	37.91	46.91	21.58	**41.07**
	PPS [2]	7.31M	51.74	25.87	45.63	40.09	**17.11**	37.07	42.01	21.44	38.23
	PC-Net(Ours)	3.34M	**54.84**	**26.68**	**47.12**	**41.69**	<u>16.73</u>	**38.04**	**46.91**	**23.60**	<u>41.06</u>

Supervision	Method	Params	Test-Trivial			Novel-Composition			Novel-Word		
			R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
Full Supervision	TSP-PRL [22]	-	34.27	18.80	37.05	14.74	1.43	12.61	18.05	3.15	14.34
	TMN [49]	-	16.82	7.01	17.13	8.74	4.39	10.08	9.93	5.12	11.38
	2D-TAN [18]	-	44.50	26.03	42.12	22.80	9.95	28.49	23.86	10.37	28.88
	LGI [52]	-	43.56	23.29	41.37	23.21	9.02	27.86	23.10	9.03	26.95
	VLSNet [50]	-	39.27	23.12	42.51	20.21	9.18	29.07	21.68	9.94	29.58
	VISA [9]	-	47.13	29.64	44.02	31.51	16.73	35.85	30.14	15.90	35.13
	Deco [8]	-	43.98	24.25	43.47	27.35	11.66	31.27	-	-	-
	Moment-DETR [54]	-	42.73	25.31	42.19	29.29	13.71	31.63	26.84	13.34	29.95
	Moment-DETR$_S$ [1]	-	44.19	25.81	43.49	30.60	14.40	33.13	29.59	15.10	32.43
	QD-DETR [55]	7.92M	41.80	20.88	41.15	26.91	10.96	31.01	27.09	11.38	31.21
	QD-DETR$_S$ [1]	7.92M	43.76	25.98	42.86	29.56	14.37	32.44	27.60	13.11	30.98
Weak Supervision	WSSL [31]	-	11.03	4.14	15.07	2.89	0.76	7.65	3.09	1.13	7.10
	CNM [10]	2.38M	28.55	13.44	35.06	18.38	7.22	28.19	21.07	9.59	29.71
	CPL [7]	4.64M	27.62	11.80	32.73	19.31	7.05	26.95	22.50	9.29	28.33
	CCR [32]	268.96M	27.67	12.90	33.56	19.59	7.66	27.50	21.66	9.18	28.42
	QMN [6]	272.38M	24.27	13.19	33.82	15.88	6.09	27.30	19.31	7.76	28.96
	PPS [2]	8.94M	**30.00**	**15.84**	32.98	**20.60**	**9.45**	26.27	**22.98**	**11.25**	27.69
	PC-Net(Ours)	4.97M	<u>29.62</u>	<u>14.35</u>	**36.45**	<u>20.16</u>	<u>8.05</u>	**29.51**	<u>22.88</u>	<u>9.85</u>	**30.76**

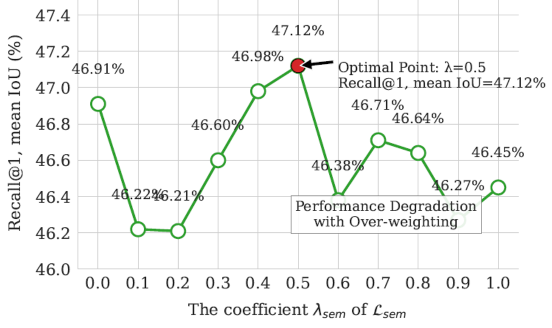
- It can be seen that the proposed PC-Net not only has a high parameter utilization rate, but also has a good generalization ability for queries with new compositions or new words that have not been seen in training

Ablation Study

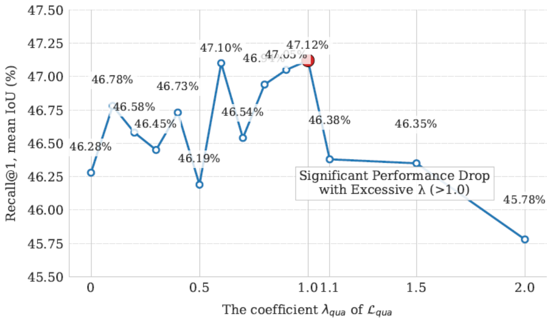
- **A study on the full ablation of the proposed module and losses based on the Charade-CG dataset**

Setting	DPG	PFA		$\mathcal{L}_{\text{qua}}$	Test-Trivial			Novel-Composition			Novel-Word		
		LPG	$\mathcal{L}_{\text{sem}}$		R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
(a)					53.04	24.71	45.82	40.79	16.15	37.46	42.45	21.44	39.20
(b)	✓				54.39	24.87	46.89	40.97	16.40	37.77	46.06	22.32	40.98
(c)		✓			51.87	23.19	45.41	40.38	16.56	37.36	42.16	22.30	38.98
(d)			✓		54.43	24.06	46.42	41.61	16.13	38.02	46.06	21.44	40.98
(e)				✓	51.52	23.55	45.66	39.80	17.00	37.57	43.60	22.16	40.07
(f)	✓	✓			54.07	25.65	47.08	40.74	16.64	37.72	46.20	23.45	41.04
(g)	✓		✓		54.04	24.94	46.59	41.48	16.91	37.49	46.91	23.60	**41.17**
(h)	✓	✓	✓		53.26	24.97	46.28	40.99	16.39	36.94	45.18	22.01	40.28
(i)	✓	✓		✓	54.65	25.16	46.91	40.78	16.55	37.62	45.76	21.73	39.93
(j)	✓		✓	✓	53.88	24.52	46.61	41.04	16.82	38.01	46.19	23.60	40.41
(k)		✓	✓	✓	54.33	25.55	46.75	41.52	**17.81**	37.74	45.32	22.45	40.71
Ours	✓	✓	✓	✓	**54.84**	**26.68**	**47.12**	**41.69**	16.73	**38.04**	**46.91**	**23.60**	41.06

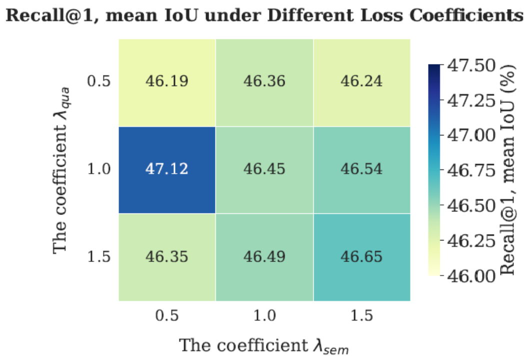
'DPG' denotes the dual-granularity proposal generator, and 'PFA' refers to the proposal feature aggregator, which incorporates both cross-modal semantic contrastive loss ($\mathcal{L}_{\text{sem}}$) and the learnable peak-aware Gaussian distributor ('LPG'). The contrastive loss in quality margin regularizer is denoted as $\mathcal{L}_{\text{qua}}$



(a) Coefficient Ablation of $\mathcal{L}_{\text{sem}}$.



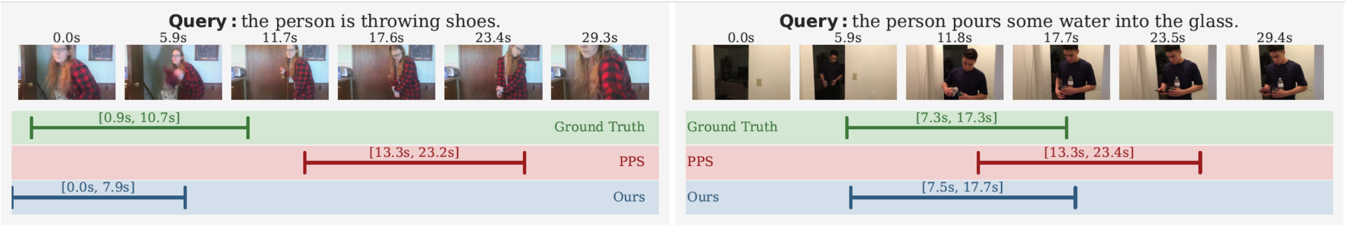
(b) Coefficient Ablation of $\mathcal{L}_{\text{qua}}$.



(c) $\mathcal{L}_{\text{sem}}$ and $\mathcal{L}_{\text{qua}}$ co-ablation.

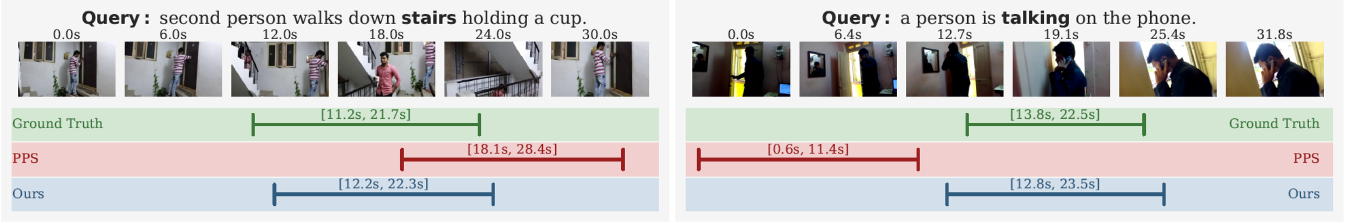
Qualitative Results

- Compared with the weakly supervised PPS (left), PC-Net generalizes better and accurately locates novel queries. Against the fully supervised QD-DETRs (right), it more effectively models multimodal correlations, demonstrating superior architectural efficiency.

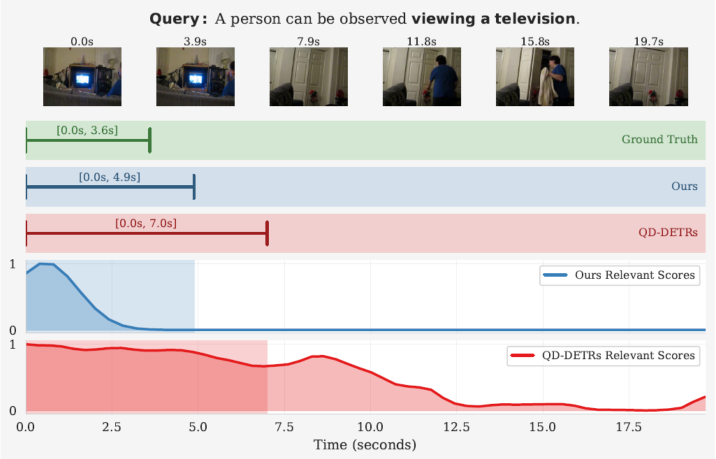


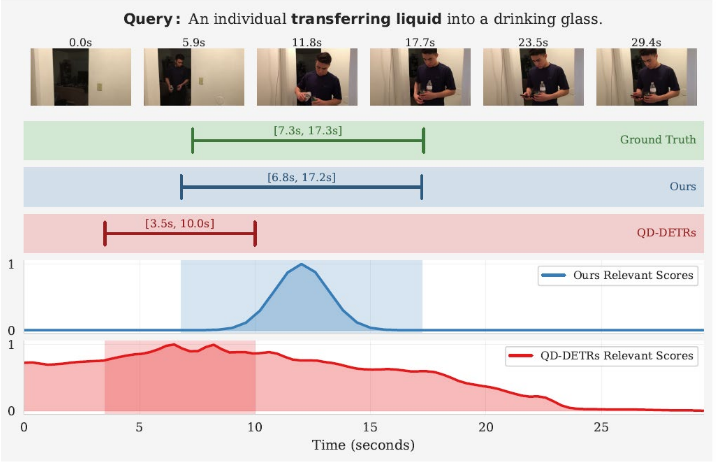
(a) Two Samples from the Test-Trivial split.

(b) Two Samples from the Novel-Composition split.



(c) Two Samples from the Novel-Word split.



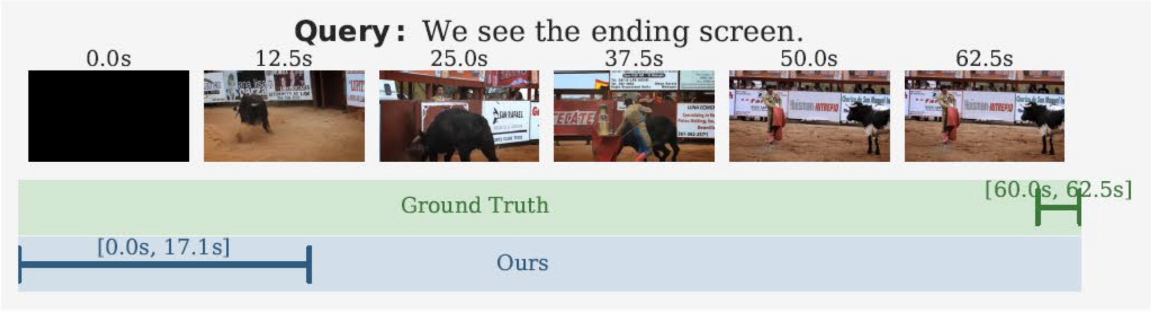


Conclusion

- Discussion
- Future Direction

Summary

- This paper analyzes the shortcomings of existing methods, and proposes a more practical and scalable task, namely WSCMR. PC-Net is constructed to address challenges in WSCMR
  - By fully mining the dual-granularity query semantics and temporal perception to obtain query-relevant and well-bounded proposals,
  - and improving feature discrimination through the semantic alignment and peak optimization,
  - and the quality margin regularizer is used to establish associations between common visual elements in proposals and queries and to suppress spurious associations
- However, **the proposed method has limited modeling ability for actions with a longer duration**. Future work will explore improving the query generalization of weakly supervised moment retrieval in long videos



Thanks for your listening!

Mingyao Zhou

08/10/2025

Paper

GitHub

WeChat