EUROPEAN CONFERENCE ON COMPUTER VISION

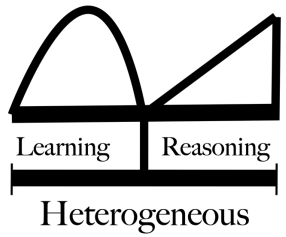
MILANO
2024

SHINE: Saliency-aware Hlerarchical NEgative Ranking for Compositional Temporal Grounding

Zixu Cheng[1]*, Yujiang Pu[2]*, Shaogang Gong[1], Parisa Kordjamshidi[2], Yu Kong[2]

[1] School of Electronic Engineering and Computer Science, Queen Mary University of London

[2] Department of Computer Science and Engineering, Michigan State University (* Equal Contribution)

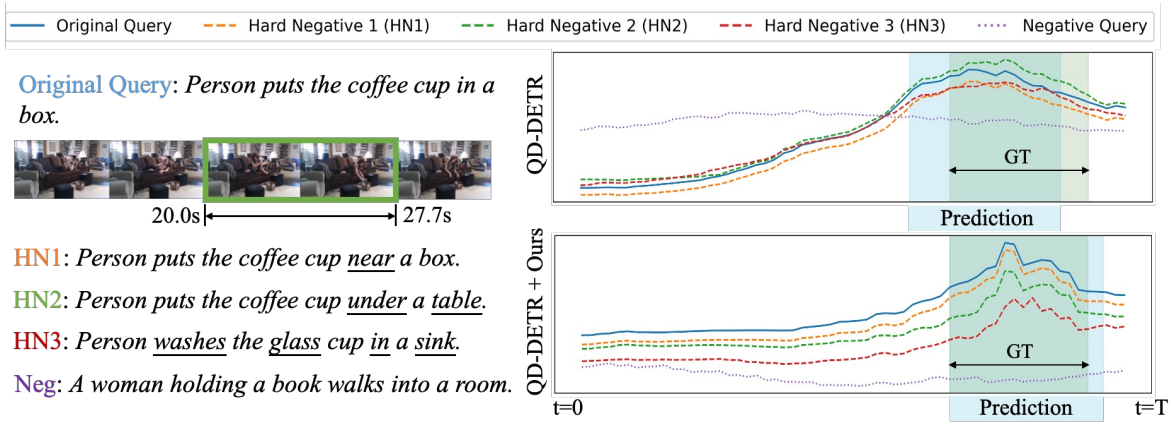


01 Background

Identify the **start and end timestamp** in the video that correspond to the query sentence. The model is expected to:
Generalize to **unseen** video moments by learning **fine-grained semantics** in **compositional concepts** in the training data.

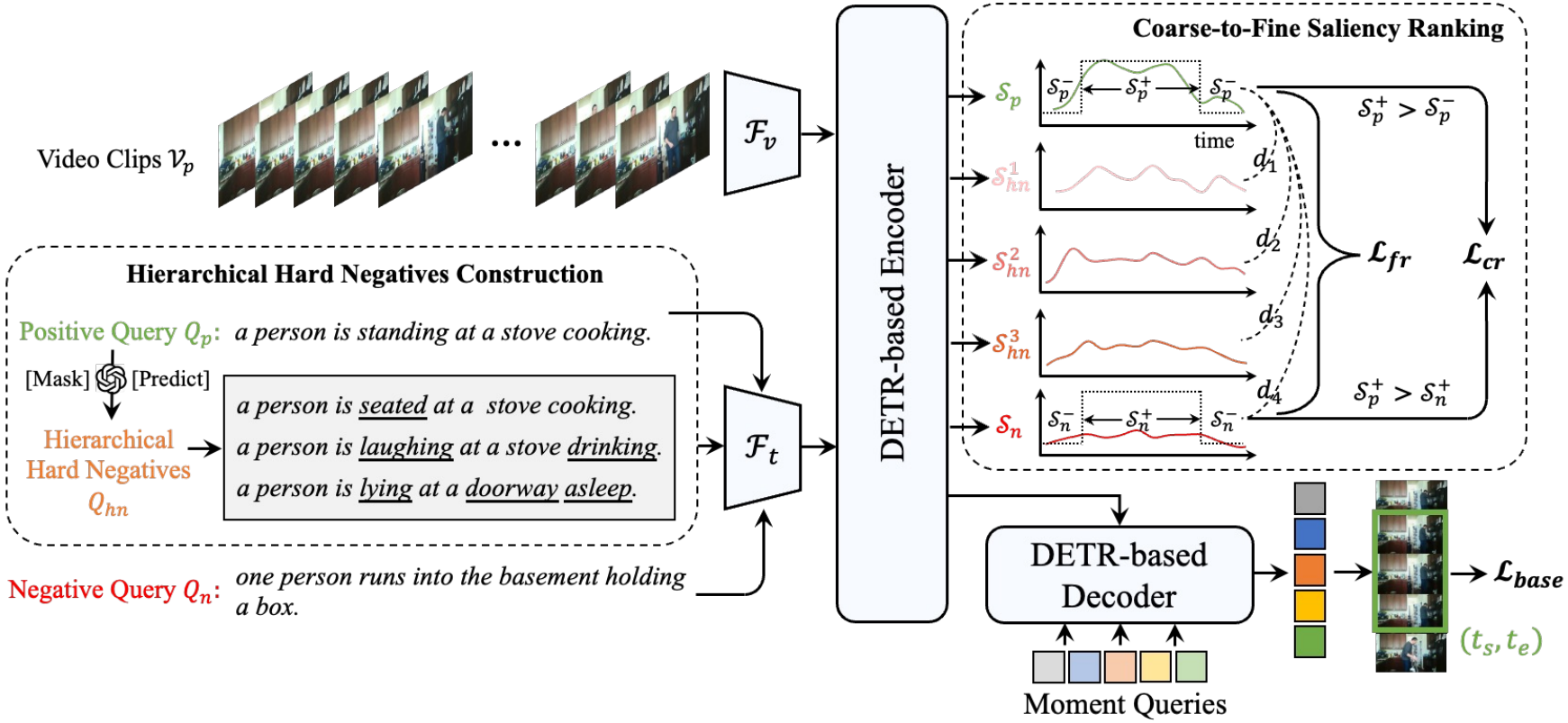
Existing Work

- Only consider dominant verbs and nouns.
- Random sampling leads to implausible compositions.
- Irrational saliency responses to hard negatives due to a lack of compositional generalization.



02

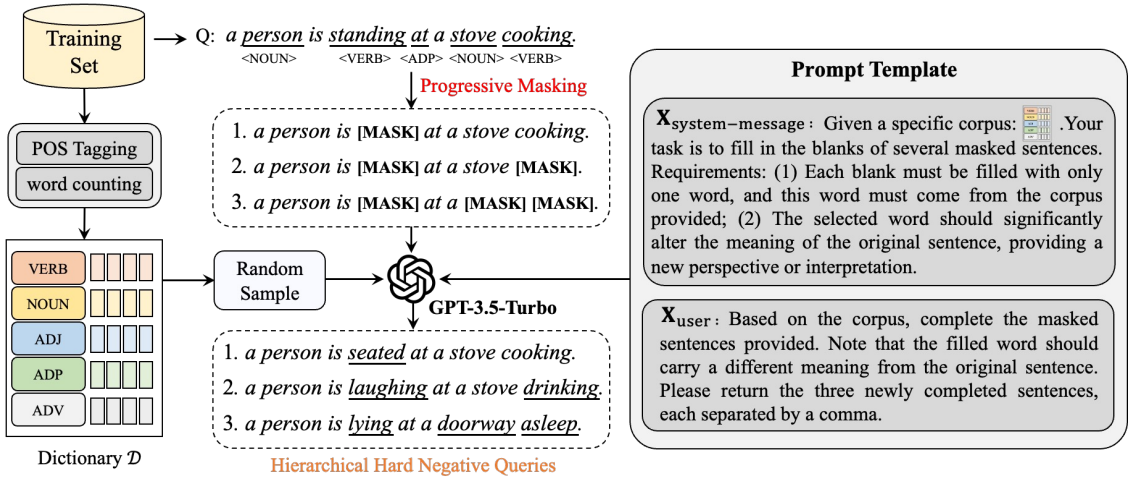
Proposed Method



- We introduce an LLM-driven approach that produces semantically plausible hard negative queries.
- We propose a coarse-to-fine saliency ranking strategy to capture hierarchical semantic differences and boost compositional generalizability.

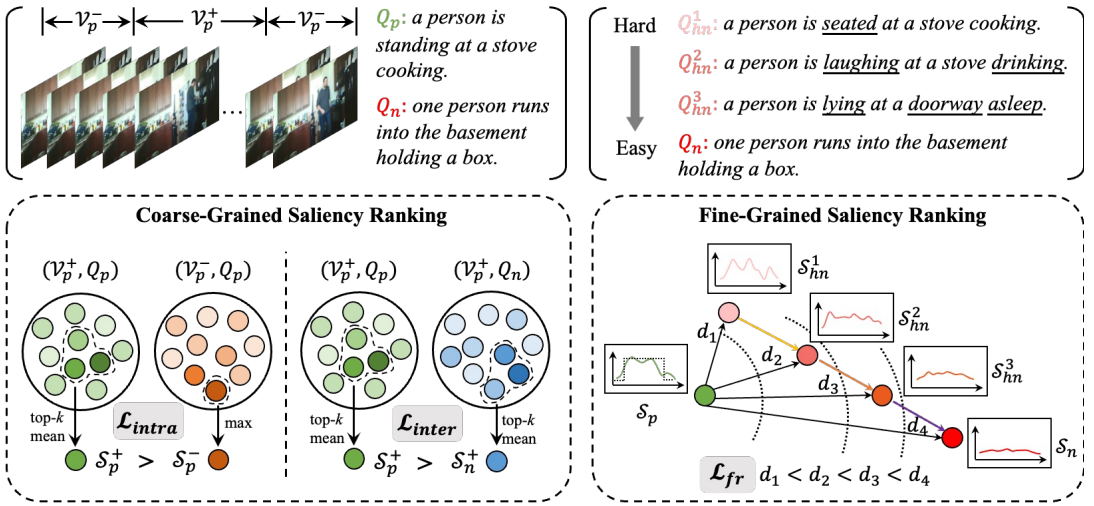
Proposed Method

- **Hierarchical Hard Negatives Construction**



- Part-of-speech tagging on queries to obtain five types of primitives
- Iteratively masking the primitives with different ratios
- GPT-3.5-Turbo fills in the blanks

- **Coarse-to-Fine Saliency Ranking**



- Coarse-grained saliency ranking loss improves the discriminative capability of the video-text representation.
- Fine-grained saliency ranking loss discern the nuances between various primitive words and video moments.

03 Experimental Results

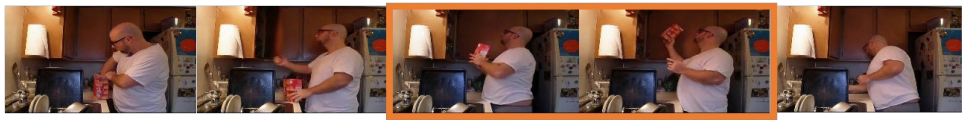
- **Quantitative Comparison**

State-of-the-art performance is achieved on Charades-CG dataset.

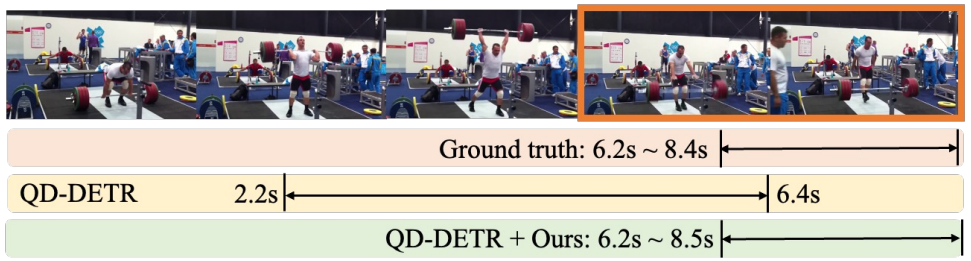
Setting	Method	Test-Trivial			Novel-Composition			Novel-Word		
		R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
WS	WSSL [6]	15.33	5.46	18.31	3.61	1.21	8.26	2.79	0.73	7.92
RL	TSP-PRL [41]	39.86	21.07	38.41	16.3	2.04	13.52	14.83	2.61	14.03
PB	TMN [26]	18.75	8.16	19.82	8.68	4.07	10.14	9.43	4.96	11.23
	2D-TAN [58]	48.06	27.10	43.72	32.74	15.25	31.5	37.12	18.99	35.04
	2D-TAN+SSL [21]	53.91	31.82	46.84	35.42	17.95	33.07	43.60	25.32	39.32
	MS-2D-TAN [57]	57.85	37.63	50.51	43.17	23.27	38.06	45.76	27.19	40.80
	MS-2D-TAN+SSL [21]	58.14	37.98	50.58	46.54	25.10	40.00	50.36	28.78	43.15
PF	LGI [33]	49.45	23.8	45.01	29.42	12.73	30.09	26.48	12.47	27.62
	VLSNet [56]	45.91	19.80	41.63	24.25	11.54	31.43	25.60	10.07	30.21
	VISA* [23]	53.20	26.52	47.11	45.41	22.71	42.03	42.35	20.88	40.18
	Deco [48]	58.75	28.71	49.06	47.39	21.06	40.70	-	-	-
	Moment-DETR† [20]	49.48	28.04	44.82	39.42	18.62	36.61	46.76	24.75	41.70
	Moment-DETR+Ours	57.14	33.85	49.32	44.65	23.21	39.86	47.05	24.32	41.57
	QD-DETR† [32]	59.24	33.43	50.92	42.30	21.09	38.55	46.04	26.33	42.89
	QD-DETR+Ours	**60.66**	**38.60**	**52.53**	**50.23**	**27.69**	**44.14**	**55.25**	**35.25**	**48.10**

- **Qualitative Analysis**

Query: A person **puts** a box **on** the counter.



Query: The man **drops** the **barbell** onto the ground.



Query: Person start **pouring** water **into** a **pot** to begin cooking.