DR.SIMON : Domain-wise Rewrite for Segment Informed Medical Oversight Network

Seohyun Lee[1], Suhyun Choe[2*], Jaeha Choi[3*], Jin Won Lee[4*]

[1] Korea University [2] Yonsei University [3] Incheon National University [4] McGill University

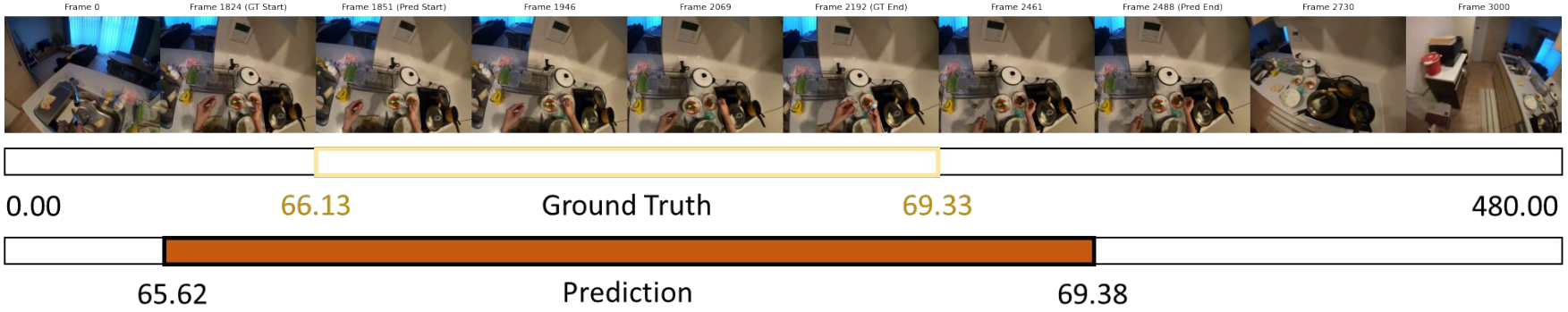
(*Equal Contribution)

Background

> Video Temporal Grounding (VTG)

Video Temporal Grounding : locate the start–end segment for a text query

Query: Where did I wash the white dishes?

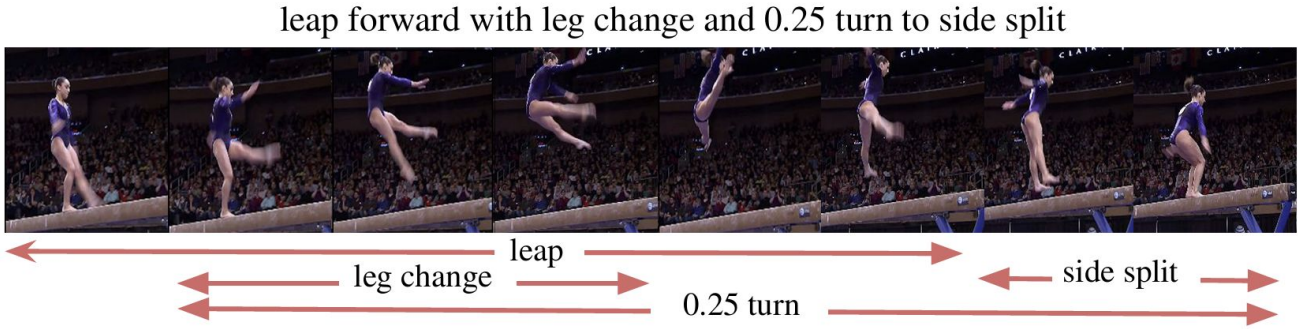


Background

> Video Temporal Grounding (VTG)

Early Approaches : video/clip-level cues→ limited boundary precision

Temporal Query Networks : fine-grained action-level localization



Temporal Query Networks : Fine grained action recognition

Temporal Query Networks for Fine-grained Video Understanding, Zhang et al., 2021.

Background

> Video Temporal Grounding (VTG)

Video LLMs : better grounding → improved boundary cues and query alignment

- VTimeLLM : boundary-aware training, but **limited for long videos**
- ReVisionLLM : extends context **hierarchically** but requires **heavy training**

VTimeLLM: Empower LLM to Grasp Video Moments, Huang et al., 2023.
ReVisionLLM: Recursive Vision-Language Model for Temporal Grounding in Hour-Long Videos, Hannan et al., 2024.

Problem Statement

> Challenges in Medical Videos VTG

Limitations in Long-Form Video Handling & Query–Video Alignment

- **General training queries:** explicit visual cues → **easy grounding**

  (e.g., “man in a blue shirt is sitting”)

- **Medical videos:** long videos, medical terms → queries **not directly visual**

  (e.g., “how to remove a fishhook via the string-jerk method”)

Problem Statement

> Challenges in Medical Videos VTG

Limitations in Fine-Tuning the Model with Medical-Domain Data

- **Data scarcity:** limited labeled medical datasets
- **Resource cost:** extensive compute/time
- **Limited generalization:** struggles to adapt to newly coined or rare terms

Problem Statement

> Motivation

(a) Original query

Query : During which frames can we see how to treat **patellofemoral pain** by performing straight side leg raises?

(b) Masked query

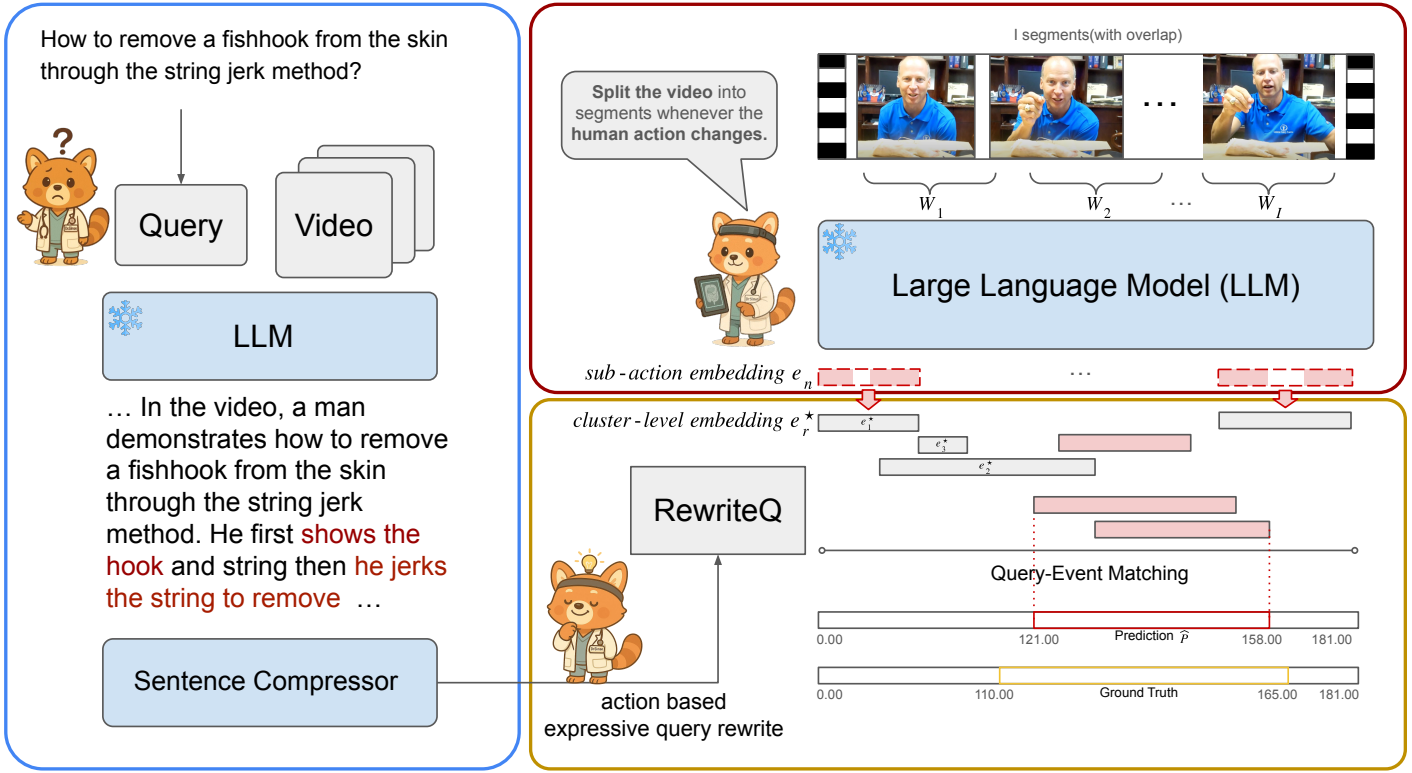
Query : During which frames can we see how to treat **something** by performing straight side leg raises?

Method	mIoU ↑	R@0.3 ↑	R@0.5 ↑	R@0.7 ↑
VTimeLLM	6.32	9.86	4.93	2.11
+ Masked Query	7.92	11.97	6.34	3.52

"Rewrite the query into visually explicit form before grounding"

Method

> Overall Pipeline



Method

> Query Rewriting Module (QRM)

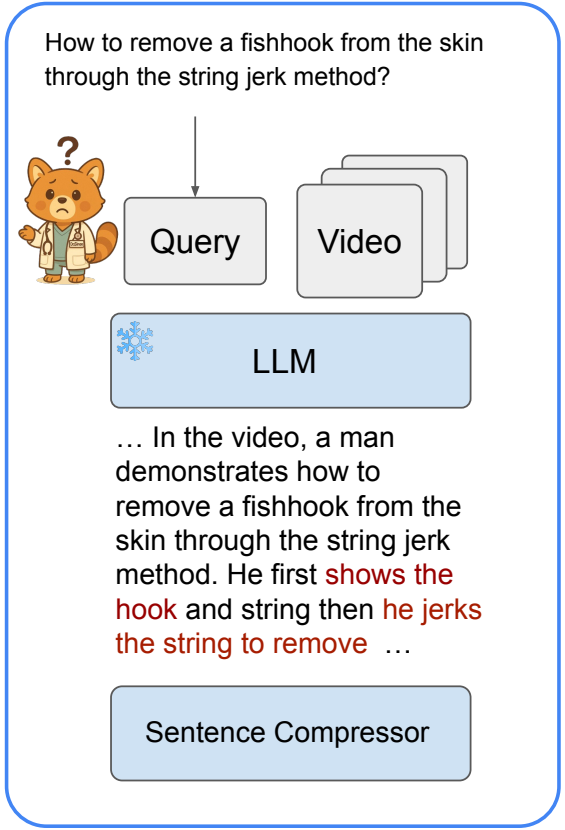
Goal

create a **more action-focused version of the query**.

Prompt

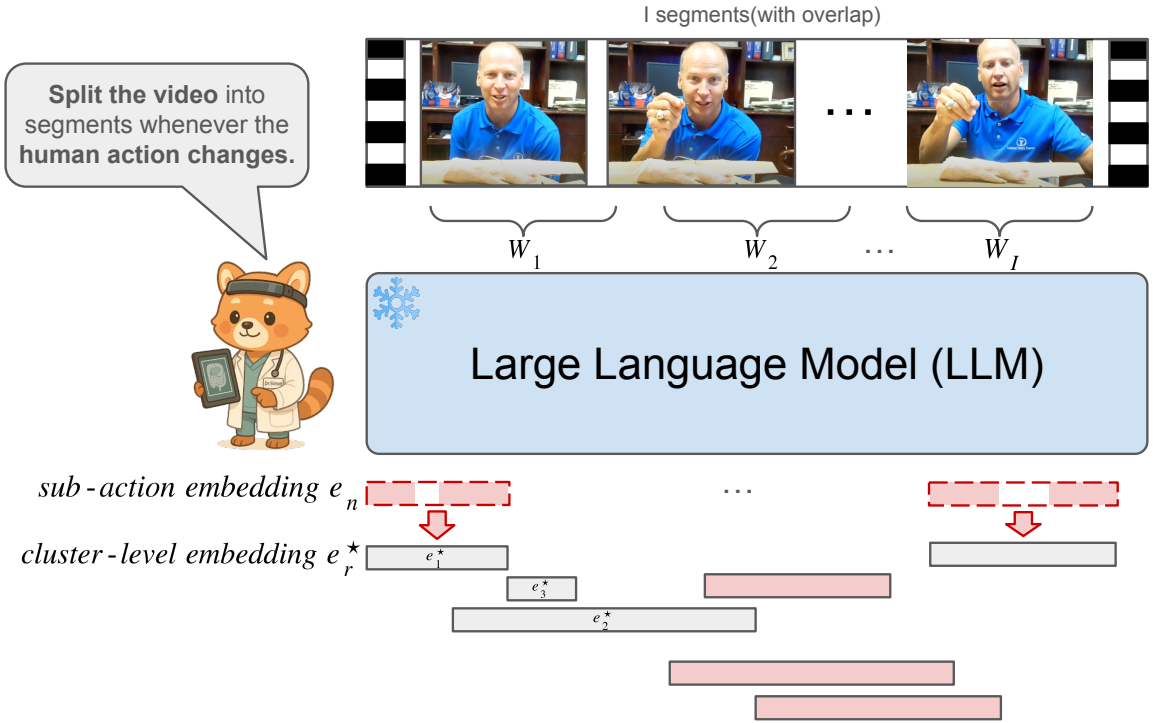
"Explain '{original query}' by **describing the actions of people** in the video."

=> output concise action-focused reformulated query



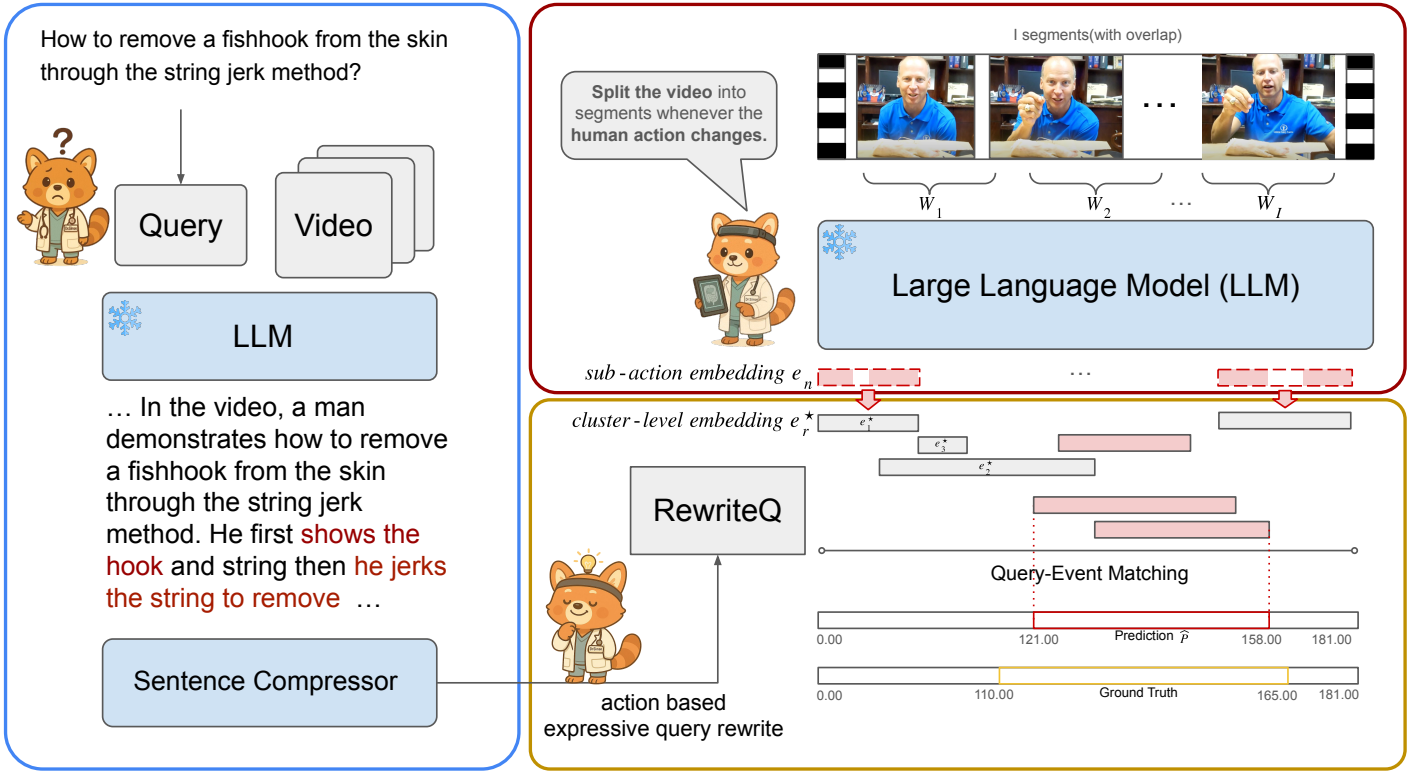
Method

> Boundary Event Segmentation Module (BESM)



Method

> Query Event Matching Module (QEM)

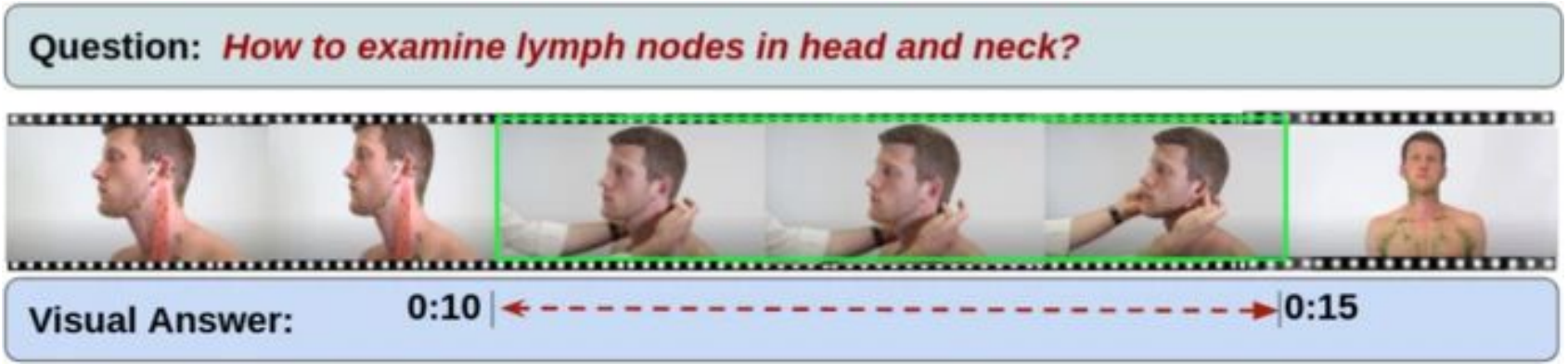


QEM

Experiment

> Evaluation Dataset

MedVidCL



A Dataset for Medical Instructional Video Classification and Question Answering, Gupta et al.

Experiment

> Result

Backbone : Vicuna-7B + CLIP ViT/L-14, videos @ 2fps

Method	mIoU↑	R@0.3↑	R@0.5↑	R@0.7↑
VTimeLLM	6.32	9.86	4.93	2.11
+ Masked Query	7.92	11.97	6.34	3.52
+ Rewrite Query	7.15	10.56	5.63	3.52
+ Rewrite Query (Summ.)	9.18	15.49	7.04	1.41
VSL-QGH	20.12	25.81	14.20	6.45
RevisionLLM	21.18	28.50	**26.10**	**14.28**
DR.SIMON (ours)	**28.08**	**40.14**	20.42	10.56

VTimeLLM: Empower LLM to Grasp Video Moments, Huang et al., 2023.
ReVisionLLM: Recursive Vision-Language Model for Temporal Grounding in Hour-Long Videos, Hannan et al., 2024.
A Dataset for Medical Instructional Video Classification and Question Answering, Gupta et al.

Experiment

> Result

Domain-specific terms hurt performance

Method	mIoU↑	R@0.3↑	R@0.5↑	R@0.7↑
VTimeLLM	6.32	9.86	4.93	2.11
+ Masked Query	7.92	11.97	6.34	3.52
+ Rewrite Query	7.15	10.56	5.63	3.52
+ Rewrite Query (Summ.)	9.18	15.49	7.04	1.41
VSL-QGH	20.12	25.81	14.20	6.45
RevisionLLM	21.18	28.50	**26.10**	**14.28**
DR.SIMON (ours)	**28.08**	**40.14**	20.42	10.56

VTimeLLM: Empower LLM to Grasp Video Moments, Huang et al., 2023.
ReVisionLLM: Recursive Vision-Language Model for Temporal Grounding in Hour-Long Videos, Hannan et al., 2024.
A Dataset for Medical Instructional Video Classification and Question Answering, Gupta et al.

Experiment

> Result

Query rewriting helps coarse localization

Method	mIoU↑	R@0.3↑	R@0.5↑	R@0.7↑
VTimeLLM	6.32	9.86	4.93	2.11
+ Masked Query	7.92	11.97	6.34	3.52
+ Rewrite Query	7.15	10.56	5.63	3.52
+ Rewrite Query (Summ.)	9.18	15.49	7.04	1.41
VSL-QGH	20.12	25.81	14.20	6.45
RevisionLLM	21.18	28.50	**26.10**	**14.28**
DR.SIMON (ours)	**28.08**	**40.14**	20.42	10.56

VTimeLLM: Empower LLM to Grasp Video Moments, Huang et al., 2023.
ReVisionLLM: Recursive Vision-Language Model for Temporal Grounding in Hour-Long Videos, Hannan et al., 2024.
A Dataset for Medical Instructional Video Classification and Question Answering, Gupta et al.

Experiment

> Result

Query rewriting helps coarse localization

Method	mIoU↑	R@0.3↑	R@0.5↑	R@0.7↑
VTimeLLM	6.32	9.86	4.93	2.11
+ Masked Query	7.92	11.97	6.34	3.52
+ Rewrite Query	7.15	10.56	5.63	3.52
+ Rewrite Query (Summ.)	9.18	15.49	7.04	1.41
VSL-QGH	20.12	25.81	14.20	6.45
RevisionLLM	21.18	28.50	**26.10**	**14.28**
DR.SIMON (ours)	**28.08**	**40.14**	20.42	10.56

VTimeLLM: Empower LLM to Grasp Video Moments, Huang et al., 2023.
ReVisionLLM: Recursive Vision-Language Model for Temporal Grounding in Hour-Long Videos, Hannan et al., 2024.
A Dataset for Medical Instructional Video Classification and Question Answering, Gupta et al.

Experiment

> Result

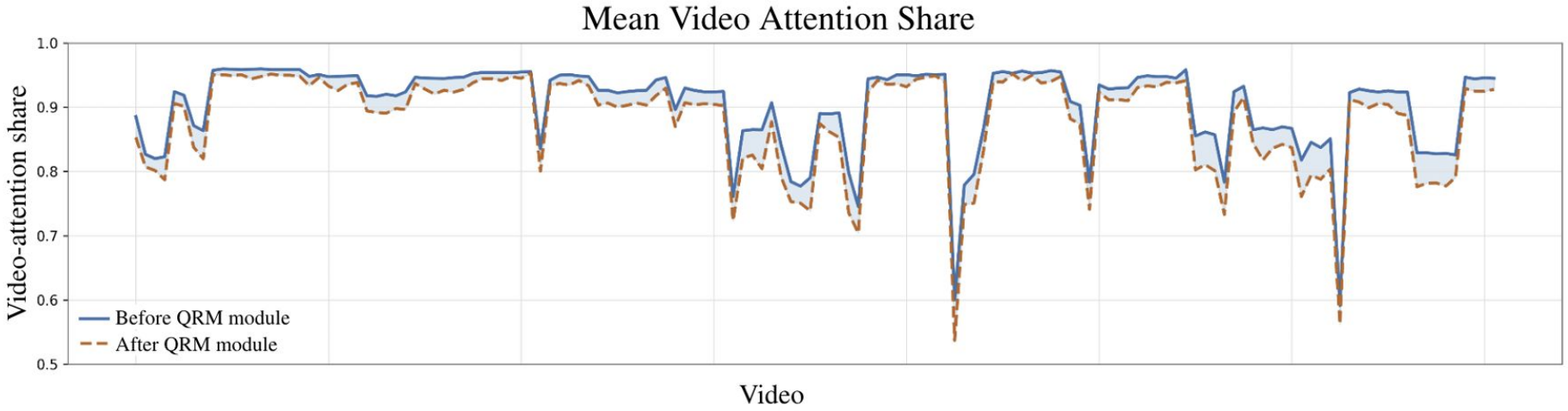
DR.SIMON produces the highest mIoU and R@0.3

Method	mIoU↑	R@0.3↑	R@0.5↑	R@0.7↑
VTimeLLM	6.32	9.86	4.93	2.11
+ Masked Query	7.92	11.97	6.34	3.52
+ Rewrite Query	7.15	10.56	5.63	3.52
+ Rewrite Query (Summ.)	9.18	15.49	7.04	1.41
VSL-QGH	20.12	25.81	14.20	6.45
RevisionLLM	21.18	28.50	**26.10**	**14.28**
DR.SIMON (ours)	**28.08**	**40.14**	20.42	10.56

VTimeLLM: Empower LLM to Grasp Video Moments, Huang et al., 2023.
ReVisionLLM: Recursive Vision-Language Model for Temporal Grounding in Hour-Long Videos, Hannan et al., 2024.
A Dataset for Medical Instructional Video Classification and Question Answering, Gupta et al.

Experiment

> Further Analysis on Query Rewriting Module



More balanced use of **both video and language**

Experiment

> Hyperparameter Sensitivity

Varying τ at $k=7$

τ	mIoU↑	R@0.3↑	R@0.5↑	R@0.7↑
1.0	22.41	32.39	12.68	7.04
0.99	**28.08**	**40.14**	**20.42**	**10.56**
0.90	25.31	30.99	15.49	9.15
0.80	22.73	24.65	10.56	4.93
0.70	21.53	21.13	9.15	4.23
0.60	19.65	16.20	9.86	4.93
0.50	18.17	14.08	7.75	3.52

Varying k at $\tau \in \{0.99, 0.90\}$

τ	k	mIoU↑	R@0.3↑	R@0.5↑	R@0.7↑
0.99	10	25.99	30.28	15.49	7.75
	7	**28.08**	**40.14**	**20.42**	**10.56**
	5	23.20	30.28	16.20	10.56
	3	17.42	23.24	12.00	5.63
	1	7.65	8.45	3.52	0.70
0.90	10	23.94	25.35	13.38	7.75
	7	25.31	30.99	15.49	9.15
	5	24.86	**33.80**	**17.61**	**9.86**
	3	20.51	26.06	13.38	7.75
	1	8.66	9.86	4.23	0.70

Ablation Studies

> Effect of QRM

- Disabled BESM and QEM
- LoRA fine-tuned on the MedVidCL training set

Setting	mIoU ↑	R@0.3 ↑	R@0.5 ↑	R@0.7 ↑
Base FT + Orig. Q (Row 1)	14.58	21.13	9.86	4.23
Base FT + Rewrite Q (Row 2)	14.79	21.13	10.56	2.82
RewriteQ FT + Orig. Q (Row 3)	14.99	24.65	9.86	3.52
RewriteQ FT + Rewrite Q (Row 4)	**15.18**	**24.94**	**10.15**	**4.52**

Ablation Studies

> Effect of QRM

Rewriting the query is useful **even at train time**

Setting	mIoU ↑	R@0.3 ↑	R@0.5 ↑	R@0.7 ↑
Base FT + Orig. Q (Row 1)	14.58	21.13	9.86	4.23
Base FT + Rewrite Q (Row 2)	14.79	21.13	10.56	2.82
RewriteQ FT + Orig. Q (Row 3)	14.99	24.65	9.86	3.52
RewriteQ FT + Rewrite Q (Row 4)	**15.18**	**24.94**	**10.15**	**4.52**

Ablation Studies

> Effect of QRM

Rewriting is **less critical** once the model is **domain-tuned**

Setting	mIoU ↑	R@0.3 ↑	R@0.5 ↑	R@0.7 ↑
Base FT + Orig. Q (Row 1)	14.58	21.13	9.86	4.23
Base FT + Rewrite Q (Row 2)	14.79	21.13	10.56	2.82
RewriteQ FT + Orig. Q (Row 3)	14.99	24.65	9.86	3.52
RewriteQ FT + Rewrite Q (Row 4)	**15.18**	**24.94**	**10.15**	**4.52**

⇒ Query rewriting helps **most** when backbone is **not domain-tuned**

Conclusion

> Limitations

- Relies **heavily** on the **segmentation quality** of the Video-LLM
- Performance drops at stricter IoU
- Sensitive to **hyperparameter choices**

Conclusion

> DR.SIMON

- Efficient framework **without any additional training**
- **Practical for medical domains** with emerging terminology or scarce annotated data
- By redefining temporal grounding as **event selection**, it narrows the search space and reduces computation
- Injecting visually grounded actions into queries **bridges the lexical gap**, reducing the need for domain-specific fine-tuning

Thanks