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

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

[1] Korea University, Seoul, Republic of Korea
happy8825@korea.ac.kr
[2] Yonsei University, Seoul, Republic of Korea
gydbs0925@yonsei.ac.kr
[3] Incheon National University, Incheon, Republic of Korea
chlgocks2000@inu.ac.kr
[4] McGill University, Montreal, Canada
jinwon.lee@mail.mcgill.ca

Abstract. Humans are capable of understanding language, even when encountering unfamiliar words. Rather than requiring precise definitions, we often infer meaning from the surrounding linguistic context or visual cues. Inspired by this capability, we address the long-standing challenge of aligning medical terminology in queries with visual content for temporal grounding in medical videos. While bridging this gap typically relies on costly, domain-specific fine-tuning, such methods frequently lack generalization and struggle to adapt to newly coined or rarely encountered terms. To deal with this limitation, we present **DR.SIMON** (**D**omain-wise **R**ewrite for **S**egment-**I**nformed **M**edical **O**versight **N**etwork), a simple yet efficient query-rewriting framework that runs on a frozen backbone. DR.SIMON first segments the video into coarse events, then rewrites the user query into visually explicit paraphrases under global visual context, and finally localizes the most relevant segment. Evaluated on MedVidCL, DR.SIMON achieves remarkable gains over recent video-LLMs—without any additional training. Our results show that mitigating lexical misalignment alone can unlock substantial performance improvements and provide a scalable route to keep pace with continually emerging medical vocabulary. Code will be released at https://drsimon-rewrite.github.io/.

Keywords: Video temporal grounding · Domain specific query rewriting · Video event segmentation

1 Introduction

The rapid advancement of vision-language model (VLM) has significantly accelerated progress in video understanding tasks [13, 15, 17, 22]. In particular, video temporal grounding (VTG) localizes segments in a video and underpins downstream tasks such as retrieval and summarization [6, 9, 12]. While VTG has been

⋆ Equal contributions, ordered alphabetically

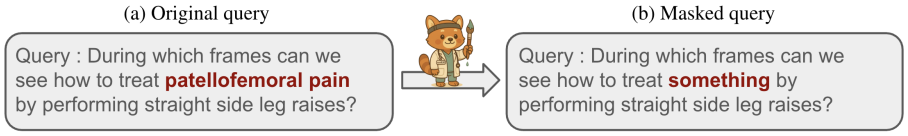

Fig. 1. Diagnostic Masking Procedure. (a) Original query containing a domain-specific medical term (e.g., *polypectomy*). (b) The same query after replacing that term with the neutral placeholder token "something." This isolates the influence of the medical term and forms the basis of the ablation in Table 1.

widely investigated in general-domain videos [4, 6, 8, 11, 19, 25], it remains challenging in domain-specific videos, particularly medical videos, due to their long durations and the need for expert-level semantic and terminological understanding. Recent approaches address this by integrating external medical knowledge graphs or structured report data [5, 18, 24]. However, these methods typically require extensive training and large-scale domain-specific datasets.

To determine how much the lexical gap hurts grounding, we perform a simple ablation(Fig. 1) in which every domain-specific medical term in queries is masked. Counter-intuitively, overall performance improves (Table 1). The improvement is small yet revealing: the limiting factor is not visual perception but the mismatch between domain-specific terminology and the generic visual cues on which the model was trained. This motivates our key idea: rewrite the query into visually explicit form before grounding.

In light of this, we propose DR.SIMON, a simple yet efficient framework that rewrite the query itself into visually explicit language, achieving comparable gains with significantly lower cost. Without additional training, DR.SIMON improves VTG in long-form medical videos in three stage process. (1) Query rewriting. It employs a frozen Video-LLM to generate a visually explicit paraphrase of the medical question, augmenting the medical query with observable actions. (2) Segmentation. It divides the input video into semantically coherent clips and extracts representative actions of each segment. (3) Selection and localization. It computes similarity between each rewritten query and actions extracted from the video and selects the most relevant segment.

Our experimental results demonstrate that DR.SIMON outperforms existing methods without any training overhead, showing that query rewriting alone effectively mitigates linguistic–visual misalignment in medical video understanding.

2 Related Works

2.1 Temporal Grounding in Video-Language Models.

Temporal grounding locates the start and end times of an event in a video given a natural-language query. Early studies focused on generating high-level summaries of entire videos but struggled to perform fine-grained temporal ground-

ing. Zhang *et al.* [26] addresses this challenge by introducing a model, Temporal Query Networks, which allows fine-grained action recognition and temporal localization in untrimmed videos.

With the advent of large language models (LLMs), recent studies have significantly improved grounding performance. VTimeLLM [11] employs a boundary-aware training pipeline for precise event localization; however, it relies on fixed-frame sampling and short context windows, making it challenging to handle long videos. ReVisionLLM [8] addresses this limitation by introducing a hierarchical, recursive strategy, requiring extensive hierarchical training on large-scale video datasets. While effective for general-domain videos, this approach necessitates additional domain-specific fine-tuning to accurately interpret specialized terminology, such as medical queries. To address these limitations we employ a sliding-window segmentation approach to efficiently handle long videos. Also we reformulate temporal grounding as an event-matching problem, significantly reducing computational overhead and effectively handling domain-specific terminology without additional training.

2.2 Query Reformulation and Handling Domain-Specific Terminology.

Rewriting queries (e.g., acronym expansion, simplification) improves alignment between language and video modalities. Sun *et al.* [23] employ an LLM (MiniGPT-4 [27]) to rephrase a user's query into an enriched version and to generate detailed textual descriptions of video frames. The system improves moment retrieval performance by matching the rewritten queries with frame-level descriptions. However, this text-only rewriting method is blind to the actual video content and, without domain-specific fine-tuning, often fails to handle specialized medical terms. In the medical QA context, handling domain-specific terminology is a key challenge. Recent approaches explore ways to make medical queries more interpretable to models. For example, Cho and Lee [3] propose a method that detects medical terms in a question and injects definitions or explanations of those terms, thereby enabling a general LLM to better grasp the meaning of the question. Although effective, these strategies overlook visual context which limits its cross-domain scalability and overall performance. By contrast, our framework explicitly rewrites the query under global visual context. The Video-LLM jointly attends to the video and the original query, injects visually observed actions into the sentence, and produces an action-centric query rewrite. This video-guided rewriting eliminates reliance on external knowledge bases, bridges the lexical gap even for unseen terms, and scales to diverse domains without additional training.

3 Method

Overview. We perform temporal grounding of a natural language query in an untrimmed video as illustrated in Figure 2. Section 3.1 rewrites the original

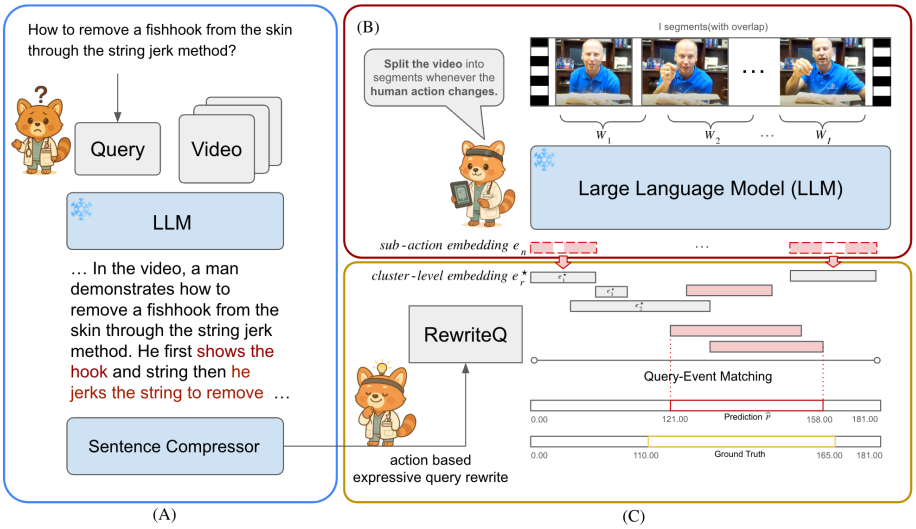

Fig. 2. Overall pipeline of DR.SIMON. (A) Rewrite the query into visually explicit language. (B) Slice the video into overlapping windows and extract representative events. (C) Match the rewritten query to those events and output the most relevant time span $\widehat{P}$. Visual encoder and VtimeLLM adapter is omitted for brevity.

query q_0 using global visual context, yielding an action-centric query $\hat{q}$ that better aligns with the video's semantics. Section 3.2 sweeps an overlapping window across the video, clusters the resulting action proposals, and refines their start–end boundaries to yield a concise set of representative events' embedding $e_i^\star$. Section 3.3 ranks the representative events by cosine similarity and selects the most coherent high-score cluster; the outermost boundaries of this cluster constitute the predicted interval $\widehat{P}$.

3.1 Query Rewriting Module (QRM)

Given an untrimmed video V of duration T seconds, we extract a sequence of frame-level CLIP [20] embeddings, denoted by $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_N] \in \mathbb{R}^{N \times d}$, where N is the total number of frames. In parallel, the original query q_0 is processed by the CLIP [20] text encoder to obtain the initial textual embedding $\mathbf{q}_0$. To obtain an action-centric rewritten query, we first pass the frame embeddings $\mathbf{F}$ through the frozen VTimeLLM [11] visual adapter to convert them into LLM-readable visual tokens. These tokens are then given as input to a frozen LLM, which produces the action-focused description $\tilde{q}$. Since $\tilde{q}$ can contain repetitive or semantically redundant phrases, we decompose it into individual sentences. We embed each with a Sentence-BERT [21] encoder and merge those whose pairwise cosine similarity exceeds a predefined threshold τ. This processing produces a concise reformulation $\hat{q}$, a clear action-driven query, that better aligns with the

video. Notably, enabling QRM by itself already boosts performance (see "Rewrite Query (Summ.)" in Table 1).

3.2 Boundary Event Segmentation Module (BESM)

To effectively handle action-centric queries, we require precise, action-oriented video captions. However, directly generating these captions faces practical challenges due to overly fragmented outputs. For instance, feeding entire videos into a Video-LLM often produces excessively detailed micro-events accompanied by repetitive captions [14]. Additionally, large language models suffer from positional biases, notably under-utilizing information located in the middle of lengthy inputs [16]. Consequently, we adopt a sliding-window strategy followed by density-based clustering, which robustly merges redundant sub-actions and yields one representative caption per semantic event.

The video is first divided into I overlapping windows $W_1, \ldots, W_I$ using a fixed stride and overlap ratio. For every window W_i, VLM model returns a set of sub-action captions $\mathcal{A}_i = \{a_{ij}\}_{j=1}^{m_i}$, where $m_i = |\mathcal{A}_i|$, each with start/end percentages that are converted into absolute seconds inside W_i. Encoding every caption once with Sentence-BERT yields the unit-norm vectors $\mathcal{E}_i = \{\mathbf{e}_{ij}\}_{j=1}^{m_i} \subset \mathbb{R}^d$.

All sub-actions from all windows are pooled into $\left\{(t_n^{\text{start}}, t_n^{\text{end}}, \mathbf{e}_n)\right\}_{n=1}^{M}$, where $M = \sum_i m_i$ and t_n^{start}, t_n^{end} are absolute start/end times. We construct a pairwise cosine-distance matrix $D_{nm} = 1 - \mathbf{e}_n^{\mathsf{T}}\mathbf{e}_m$ and run average-linkage agglomerative clustering with threshold $1 - \tau$. The influence of the density threshold τ is analyzed in Section 4.3.

Each resulting cluster C_r yields the earliest start, $\min_{n \in C_r} t_n^{\text{start}}$, and the latest end $\max_{n \in C_r} t_n^{\text{end}}$. The caption whose embedding is closest to the cluster centroid[5]. We denote this representative caption by $a_r^\star$ and its embedding by $\mathbf{e}_r^\star$. The sequence $\left\{(a_r^\star, \mathbf{e}_r^\star)\right\}_{r=1}^{R}$, where $R = |\mathcal{C}_r|$, which contains one event for each semantic cluster, is forwarded to Section 3.3 for query–event matching.

3.3 Query–Event Matching Module (QEM)

We reformulate the VTG task as an event-matching problem where an action-centric query is aligned with event-level captions. This allows the model to focus on video segments most likely to contain the answer. Given the rewritten query $\hat{q}$, we embed it as a vector $\mathbf{q}$ and compute cosine similarities $\text{score}_i = \mathbf{q}^{\mathsf{T}}\mathbf{e}_r^\star$ for every event embedding $\mathbf{e}_r^\star$. Instead of scanning the full video, we retain the seven highest–scoring events, ordered on the timeline $t_{a_1^\star}^{\text{start}} \prec \ldots \prec t_{a_7^\star}^{\text{start}}$. Focusing on top–7 strip drastically reduces the search space.

When the three best events already form a compact chunk—each pair either overlaps (IoU > 0.2) or their temporal gap is at most θ_{gap}=1.5 s—we treat them as a single action and return the union of their boundaries. This covers the majority of queries in which the relevant frames are naturally adjacent.

[5] $\hat{\mathbf{e}}_r = \frac{1}{|C_r|} \sum_{n \in C_r} \mathbf{e}_n, \quad \mathbf{e}_r^\star = \arg \max_{\mathbf{e}_n :\, n \in C_r} \hat{\mathbf{e}}_r^{\mathsf{T}}\mathbf{e}_n$

If the top–3 events are scattered, a stronger discriminator is needed. We merge adjacent segments whose gaps do not exceed θ_{gap}, yielding clusters $\{\mathcal{C}_r\}_{r=1}^{R}$. For each cluster, we define

$$S_r = (1 + \lambda\, \rho_r)\, \frac{1}{|\mathcal{C}_r|} \sum_{g \in \mathcal{C}_r} score_g, \qquad \rho_r = \max\Big(0, 1 - \tfrac{\text{span}(\mathcal{C}_r)}{|\mathcal{C}_r|\,\theta_{\text{gap}}}\Big), \tag{1}$$

with $\lambda = 0.1$ and $\text{span}(\mathcal{C}_r) = \max_{g \in \mathcal{C}_r} t_g^{\text{end}} - \min_{g \in \mathcal{C}_r} t_g^{\text{start}}$. The term $\rho_r \in [0, 1]$ rewards tightly packed clusters: it approaches 1 when segments abut and falls to 0 for diffuse groups, thus modulating the average similarity by the density of the cluster. The cluster with the highest score, $r^\star = \arg\max_r S_r$, defines the prediction

$$\widehat{P} = \Big[\min_{g \in \mathcal{C}_{r^\star}} t_g^{\text{start}},\ \max_{g \in \mathcal{C}_{r^\star}} t_g^{\text{end}}\Big]. \tag{2}$$

4 Experiments

Experiments are carried out on the MedVidCL [7] test split, which contains 50 medical instructional clips. The average video length is 345.4s, ranging from 46 seconds to 20 minutes. Across these clips there are QA pairs, which requires the model to precisely localize answer segments within long instructional videos. We follow the conventional metrics in VTG: mean Intersection-over-Union (mIoU) and recall at IoU thresholds 0.3, 0.5, and 0.7. A prediction is correct if its IoU with the ground-truth span exceeds the threshold.

4.1 Experiment Configuration

All methods except VSL-QGH [7] share a Vicuna-7B [2] backbone with a CLIP-ViT/L-14 [20], and process videos uniformly subsampled at 2.0fps. We compare seven methods: (1) VTimeLLM [11] (2) VTimeLLM with Masked Query, obtained by replacing every medical term in the query with `"something"`; (3) VTimeLLM with Rewrite Query, which inserts a query-rewriting module; (4) VTimeLLM with Summarized Query, which further applies an abstractive summary to the rewritten query; (5) VSL-QGH, a span-prediction baseline enhanced with query-guided highlighting and fine-tuned on the MedVidQA train split; (6) ReVisionLLM [8] and (7) DR.SIMON (ours).

4.2 Results and Analysis

Table 1 presents the quantitative results, highlighting three key observations. (1) Domain specific term harms performance, if the model isn't fine-tuned to the domain. Comparing Row 1 with Row 2, masking medical terms lifts all metrics over the backbone, showing that VTimeLLM [11] rarely grounds such vocabulary visually. (2) Query rewriting helps coarse localization. Feeding a visually explicit rewrite (Row 3) boosts R@0.3, confirming that linguistic alignment is the primary bottleneck. Compressing the rewrite query (Row 4) further

Table 1. Temporal localization results on the MedVidQA test set. Best score in each column is shown in **bold**, and the second-best is underlined.

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

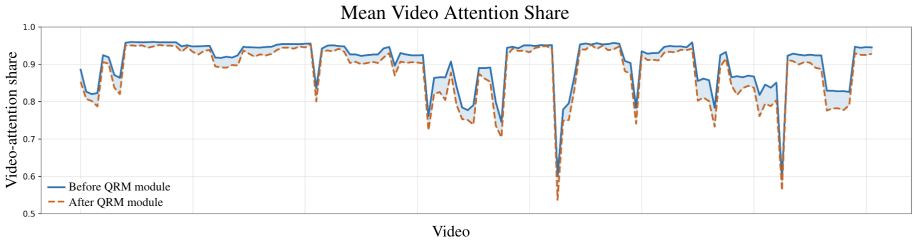

Fig. 3. Cross-modal attention ratio before and after query rewriting. The y-axis is normalized so that 1.0 corresponds to attention devoted exclusively to video tokens, whereas 0.5 indicates a perfectly balanced split between video and text. After QEM module, attention shifts toward the language modality, alleviating the original video bias.

aids coarse windows but erodes strict IoU, suggesting a loss of fine cues. As further evidence, we plot the cross-modal attention ratio in Fig. 3. Before the QEM module is applied, the model exhibits a clear bias toward video information. However, after query rewriting, attention shifts more toward the language modality, producing a more balanced output and thus reducing modal bias. By rewriting the query to match latent video features, the model can rely on textual cues that point precisely to the relevant segment. Fig. 3 mirrors the findings of Chen *et al.* [1]: rephrasing the question while preserving its meaning weakens single-modality shortcuts. (3) DR.SIMON (Row 7) produces the highest mIoU and R@0.3 without any fine-tuning. By enforcing explicit action boundaries, rewriting queries with visual context, and ranking candidate windows by density-weighted relevance, DR.SIMON achieves +6.9%p mIoU and +11.6%p R@0.3 over the strongest baseline in one pass.

4.3 Hyper-parameter Sensitivity

To quantify hyper-parameter sensitivity in QEM and BESM, we analyze two key parameters: the cluster-density threshold τ, which determines how similar segments must be merged, and the fallback strip size k, the number of addi-

Table 2. Impact of cluster density τ and fallback pool k. τ is the similarity threshold for merging segments, and k is the number of extra top-scoring segments checked when the initial top-3 are not contiguous.

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

tional high-scoring segments reviewed when the initial top-3 test fails to form a compact cluster. Since the MedVidCL [7] dataset contains mostly static, visually similar shots, tightening τ to around 0.99 maintains clusters that are pure yet comprehensive, optimizing both mIoU and recall. Lower thresholds merge less related segments, reducing performance, while excessively high thresholds fragment clusters, diminishing recall (Figure 2). Moderate fallback sizes (around $k = 7$) notably enhance recall without sacrificing overall precision, whereas larger fallback sizes produce marginal improvements but degrade cluster cohesion.

5 Ablation Studies

5.1 Effect of QRM

To measure how QRM alone aids a backbone not yet domain-tuned, and how that gain shifts after a single-epoch domain fine-tune, we disable BESM and QEM and test both original and rewritten queries. We fine-tune the frozen VTimeLLM [11] with a single-epoch LoRA [10] training on the MedVidQA train split[6]—once using the original queries Base FT and once using our rewritten queries as a form of data augmentation RewriteQ FT.

As shown in Table 3, rewriting remains useful even at train-time. Augmenting the training set with rewritten queries lifts test-time recall by +3.5%p (R@0.3) compared with Base FT, regardless of whether the test query is rewritten. Moreover, we can observe that rewriting at test-time is less critical once the model is domain-tuned. Within each checkpoint the gap between Orig. Q and Rewrite Q is only 0.2–0.3%p mIoU, suggesting the model now understands medical terms. Hence, query rewriting yields its largest benefit when the backbone has not yet been adapted to the domain specific vocabulary.

[6] LoRA settings: rank r=64, α=128; applied to all FFN, self-attention, and cross-attention blocks; batch size 24; 4 × A100-40 GB.

Table 3. Verification of QRM. LoRA fine-tuning results on MedVidQA.

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**

6 Conclusion

We propose **DR.SIMON**, a simple yet efficient framework that achieves superior performance on the MedVidQA dataset without any additional training. We expect rapid application to medical domains with emerging terminology or scarce annotated data, highlighting the power of query reformulation for robust video understanding on medical domains. It also effectively narrows the search space and significantly reduces computational resources by redefining the video temporal grounding task as an event-selection problem. Finally, by injecting visually grounded actions into the rewritten query, DR.SIMON bridges the lexical gap between medical terms and generic video features, reducing the need for domain specific finetuning.

7 Limitations and Future Work

Although our method efficiently achieves remarkable improvements in mIoU and R@0.3 without additional training, it also has certain limitations. Our method relies heavily on the initial segmentation produced by the Video-LLM. If this segmentation results in overly long or overly fragmented segments, or if the boundaries are incorrectly placed, our QEM module may select inaccurate spans. This issue is evident from the performance drop observed at higher IoU thresholds (R@0.5 and R@0.7), despite strong results at lower thresholds (R@0.3). Another limitation is the sensitivity to hyperparameter choices, including window size, overlap ratio, and clustering thresholds. Optimal settings for these parameters vary significantly based on video characteristics.

To address these limitations, we plan to explore finer-grained approaches that can precisely refine segment boundaries after coarse event selection. Additionally, adaptive strategies for automatically tuning hyperparameters according to video content will be investigated to enhance robustness across diverse videos. Lastly, we aim to extend our framework beyond medical videos, establishing a generalizable solution to handle linguistic-visual alignment challenges in other specialized domains.

References

1. M. Chen, Y. Cao, Y. Zhang, and C. Lu. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective, 2024. 7

2. Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. 6
3. J. Cho and G. Lee. K-comp: Retrieval-augmented medical domain question answering with knowledge-injected compressor. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6878–6901, 2025. 3
4. J. Gao, C. Sun, Z. Yang, and R. Nevatia. Tall: Temporal activity localization via language query. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5277–5285, 2017. 2
5. T. Gu, K. Yang, D. Liu, and W. Cai. Lapa: Latent prompt assist model for medical visual question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4971–4980, 2024. 2
6. Y. Guo, J. Liu, M. Li, Q. Liu, X. Chen, and X. Tang. Trace: Temporal grounding video llm via causal event modeling, 2025. 1, 2
7. D. Gupta, K. Attal, and D. Demner-Fushman. A dataset for medical instructional video classification and question answering, 2022. abs/2201.12888. 6, 8
8. T. Hannan, M. Islam, J. Gu, T. Seidl, and G. Bertasius. Revisionllm: Recursive vision-language model for temporal grounding in hour-long videos, 2024. 2, 3, 6
9. B. He, J. Wang, J. Qiu, T. Bui, A. Shrivastava, and Z. Wang. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1
10. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021. 8
11. B. Huang, X. Wang, H. Chen, Z. Song, and W. Zhu. Vtimellm: Empower llm to grasp video moments. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14271–14280, 2024. 2, 3, 4, 6, 8
12. M. Islam, M. Hasan, K. Athrey, T. Braskich, and G. Bertasius. Efficient movie scene detection using state-space transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18749–18758, 2023. 1
13. P. Jin, R. Takanobu, W. Zhang, X. Cao, and L. Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13700–13710, 2024. 1
14. J. Lee, J. Kim, J. Na, J. Park, and H. Kim. Vidchain: Chain-of-tasks with metric-based direct preference optimization for dense video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4499–4507, 2025. 5
15. B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan. Video-llava: Learning united visual representation by alignment before projection, 2024. 1
16. N. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. 5
17. S. Liu, C. Zhang, C. Zhao, and B. Ghanem. End-to-end temporal action detection with 1b parameters across 1000 frames. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18591–18601, 2024. 1

18. V. Nath, W. Li, D. Yang, A. Myronenko, M. Zheng, Y. Lu, Z. Liu, H. Yin, Y. Tang, P. Guo, C. Zhao, Z. Xu, Y. He, G. Heinrich, Y. Law, B. Simon, S. Harmon, S. Aylward, M. Edgar, M. Zephyr, S. Han, P. Molchanov, B. Turkbey, H. Roth, and D. Xu. Vila-m3: Enhancing vision-language models with medical expert knowledge, 2025. 2
19. L. Qian, J. Li, Y. Wu, Y. Ye, H. Fei, T. Chua, Y. Zhuang, and S. Tang. Momentor: advancing video large language model with fine-grained temporal reasoning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. 2
20. A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021. 4, 6
21. N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019. 4
22. S. Ren, L. Yao, S. Li, X. Sun, and L. Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323, 2024. 1
23. Y. Sun, Y. Xu, Z. Xie, Y. Shu, and S. Du. Gptsee: Enhancing moment retrieval and highlight detection via description-based similarity features. *IEEE Signal Processing Letters*, 31:521–525, 2024. 3
24. C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21315–21326, 2023. 2
25. Y. Wu, X. Hu, Y. Sun, Y. Zhou, W. Zhu, F. Rao, B. Schiele, and X. Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 13754–13765, 2025. 2
26. C. Zhang, A. Gupta, and A. Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3
27. Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3