Seahawk at MedGenVidQA 2026: LLM Segment-Range Selection for Medical Visual Answer Localization

Xiaotian Tian
NKU School of Mathematical Sciences
Tianjin, China
13707579071@163.com

Gulustan Dogan
University of North Carolina Wilmington
Wilmington, USA
dogang@uncw.edu

Abstract

Medical visual answer localization requires identifying the temporal span in a video where a medical question is answered or visually explained. We present a simple retrieval-and-selection pipeline for Task C that treats visual answer localization as segment-level answer paragraph selection over timestamped video transcripts. Given a question and a segmented transcript, our system prompts DeepSeek to select a contiguous range of transcript segments rather than directly generating timestamps. The final start and end times are then computed deterministically from the selected segment boundaries, decreasing the risk of hallucinated or malformed temporal outputs. To support long videos, we apply overlapping sliding-window prompting and rank candidate ranges using lexical question. In a 20-sample sanity check on test dataset, a completeness-biased configuration achieved an mIoU of 0.3217, while a shorter duration-penalized configuration improved performance to 0.4815. These results suggest that constrained LLM-based segment selection, combined with deterministic timestamp extraction, is a practical baseline for medical visual answer localization.

1 Introduction

Instructional medical videos contain procedural and explanatory information that is often difficult to convey using text alone. For consumer health questions, first-aid procedures, and medical education scenarios, the answer may correspond not to an entire video but to a short temporal interval in which a specific action, tool, or explanation is shown. Medical visual answer localization therefore requires a system to identify the relevant start and end timestamps for a natural-language medical query.

The MedGenVidQA shared tasks formalize this problem as medical video question answering with temporal grounding. In visual answer localization, a system is given a medical query and a video and must return the temporal segment where the answer is shown or explained (Gupta and Demner-Fushman, 2022; Gupta et al., 2023, 2026). The task is challenging because medical videos can be long, contain multiple related procedural steps, and include terminology that may be paraphrased or only indirectly expressed in the transcript.

Our system is designed around three principles. First, transcript segments should preserve their original timestamp boundaries. Second, the LLM should make a semantic localization decision but should not invent timestamps. Third, candidate answer spans should be biased toward concise intervals because overly long predictions reduce temporal Intersection-over-Union.

The main contributions of this paper are:

- a constrained LLM prompting strategy that predicts only start and end segment identifiers
- a sliding-window search procedure for long timestamped transcripts
- a deterministic timestamp extraction step that converts selected segment ranges into valid Task-C outputs
- a duration-penalized candidate scoring function that improves mIoU in our sanity-check evaluation

2 Background and Related Work

Medical video question answering extends traditional question answering by requiring systems to reason over visual and temporal evidence. The original MedVidQA shared task introduced medical video understanding tasks including visual answer localization, where the goal is to identify the start and end timestamps of the relevant answer segment (Gupta and Demner-Fushman, 2022). The MedVidQA dataset contains human-annotated

Proceedings of the BioNLP 2026 (Shared Tasks), pages 257–261

medical instructional questions paired with visual answer timestamps (Gupta et al., 2023). Later work expanded the setting with additional medical video resources and monomodal and multimodal approaches for health-related question answering from videos (Gupta et al., 2024).

The current MedGenVidQA Task C setting focuses on visual answer localization. Given a medical query and a video, the system must locate the temporal region where the answer is shown or explained (Gupta et al., 2026).

Our method is transcript-centered rather than fully multimodal. Instead of learning a video-text model, we exploit timestamped transcript segments as the grounding substrate. This makes the method lightweight and submission-oriented: the LLM performs semantic selection over transcript text, while the system preserves temporal reliability through deterministic post-processing.

3 Task Definition

We focus on Task C, Visual Answer Localization. Given a medical question q and a video v, the system must predict a temporal interval $[p_s, p_e]$ indicating where the answer to q is shown or explained. In our setting, each video is represented by a timestamped transcript containing ordered segments $S = \{s_1, \ldots, s_n\}$, where each segment s_i has text, a start time, and an end time.

The output for each test instance consists of an identifier, an answer start time, and an answer end time. The system is evaluated using temporal overlap between the predicted interval and the ground-truth interval. Following the shared-task setup, we report mean Intersection-over-Union (mIoU) and use the same interval-overlap formulation in our internal sanity-check experiments.

4 System Overview

Our system decomposes temporal localization into two steps: semantic segment selection and deterministic timestamp extraction. The LLM is responsible only for identifying which transcript segments contain the answer. It is not asked to generate timestamps directly. Once a segment range is selected, the predicted start time is taken from the first selected segment and the predicted end time is taken from the last selected segment.

This design reduces the output space of the LLM from arbitrary temporal strings to two integer segment identifiers. It also makes the system easier to debug: every predicted timestamp can be traced back to the transcript segment that produced it. For long transcripts, we apply the same segment-selection prompt over overlapping windows and then score all non-null candidate ranges before selecting the final prediction.

5 Experimental Setup

5.1 Data and inputs

We use timestamped video transcripts as the primary input representation. Each transcript is converted into a list of numbered segments, where each segment contains a segment identifier, start time, end time, and transcript text. This format allows the LLM to inspect both the semantic content and the temporal boundaries of each candidate answer span.

5.2 Model configuration

We use DeepSeek as the segment-range selector. The model is prompted with the question and a window of timestamped transcript segments, and it is instructed to output only two fields: the start segment identifier and the end segment identifier. All outputs are parsed with a deterministic post-processing script. Invalid or null outputs are discarded unless no valid candidate is available.

5.3 Candidate selection

For transcripts that exceed the context budget, we divide the segment list into overlapping windows. Each window produces zero or one candidate answer range. We then score candidates using lexical overlap between the question and the selected paragraph, minus a penalty proportional to the predicted duration. This scoring function favors spans that are both semantically relevant and temporally concise.

5.4 Evaluation protocol

We first evaluated the pipeline on the first 20 examples from `MedVidQA/test.json` as a sanity check. We compared a completeness-biased configuration, which allowed expanded ranges, with a shorter duration-penalized configuration. The final Task-C submission was generated using the tuned short-span configuration and formatted as `task_c_submission.json`.

6 Methodology

Our approach is a retrieval-and-selection temporal grounding pipeline using DeepSeek as a segment-range selector, coupled with deterministic timestamp extraction. The design goal is to let the model decide *where* the answer is, but never let the model invent timestamps.

6.1 Segment-based context representation

Given a transcription with segments $S = \{s_i\}$, we build a prompt context of the form:

```
[segment_id] (start_time - end_time)
segment_text
```

This representation provides the model both textual content and precise segment boundaries while keeping the output interface simple.

6.2 LLM as answer paragraph locator (segment-range prediction)

For a question q, we prompt DeepSeek to output only two fields:

```
START_SEGMENT_ID: <int or null>
END_SEGMENT_ID: <int or null>
```

The model is instructed to select a contiguous segment range that corresponds to the answer paragraph. This design avoids brittle JSON generation and mitigates long-form reasoning outputs.

6.3 Long transcript handling via sliding-window search

For long transcripts, we apply a sliding window strategy: we split segments into windows of size W with overlap O, query the model per window, and convert window-local selections to global segment ids. This approximates searching the whole transcript while keeping each LLM call bounded.

6.4 Candidate scoring with duration penalty

When multiple windows yield non-null ranges, we select the best candidate using:

$$\text{Score} = \text{Overlap}(q, \text{paragraph}) - \lambda \cdot \text{Duration} \tag{1}$$

where $Overlap(\cdot)$ counts lexical overlap between question tokens and the concatenated paragraph tokens, $Duration = end\_time - start\_time$, and λ is a tunable length penalty to discourage overly long spans.

6.5 Deterministic timestamp extraction and formatting

Once a segment range $[a, b]$ is selected, timestamps are computed deterministically:

$$\begin{aligned} \text{pred\_start} &= \text{start\_time}(s_a), \\ \text{pred\_end} &= \text{end\_time}(s_b) \end{aligned} \tag{2}$$

For Task-C submission formatting, seconds are converted to `MM:SS`; the start uses floor and the end uses ceil to ensure coverage of the answer.

6.6 Output generation

For Task-C we output a submission file task_c_submission.json with entries id, answer_start, answer_end. For analysis, we retain an intermediate file task_c_answer_paragraphs.json including selected segment ranges, paragraph text, and raw seconds.

7 Results

7.1 Evaluation Metric

We report temporal mean Intersection-over-Union (mIoU) over predicted and ground-truth answer intervals:

$$\text{IoU} = \frac{|[p_s, p_e] \cap [g_s, g_e]|}{|[p_s, p_e] \cup [g_s, g_e]|}, \tag{3a}$$

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \text{IoU}_i \tag{3b}$$

Following the shared-task evaluation protocol, we additionally report the proportion of predictions achieving IoU above fixed thresholds (IoU@0.3, IoU@0.5, IoU@0.7), which measure fine-grained temporal localization accuracy.

7.2 Sanity-Check Results (First 20 Samples)

We first evaluated the pipeline on the first 20 examples from `MedVidQA/test.json` as a sanity check. An initial completeness-biased configuration (with expanded ranges) achieved mIoU = 0.3217. After introducing a duration penalty and disabling expansion (`expand_before` = 0, `expand_after` = 0), performance improved substantially; one setting achieved mIoU = 0.4815 with `length_penalty` = 0.06. A similar short+penalty regime yielded mIoU = 0.4689. These sanity checks confirmed that penalizing overly long spans is beneficial for temporal localization.

7.3 Full Test Set Results (MedGenVidQA Task C)

We ran the complete Task C test set (80 instances, designated C1–C80) using the tuned short+penalty configuration. Table 1 summarizes the final performance.

Table 1: Performance on the full MedGenVidQA 2026 Task C test set.

Metric	Value
mIoU (%)	52.31
IoU@0.3 (%)	71.25
IoU@0.5 (%)	52.50
IoU@0.7 (%)	42.50

The mIoU of 52.31% indicates that, on average, the predicted temporal segment overlaps with the ground truth by more than half of their union. The high IoU@0.3 (71.25%) shows that the majority of predictions have at least a modest overlap with the correct interval, while the gradual drop at stricter thresholds (IoU@0.5 and IoU@0.7) reflects the inherent difficulty of exact boundary alignment in medical procedural videos. These results demonstrate that constrained LLM-based segment selection, combined with deterministic timestamp extraction, provides a practical and competitive baseline for medical visual answer localization, even without any training or fine-tuning of the underlying model.

7.4 Qualitative Behavior

In typical success cases, the model selects a segment range corresponding to the procedural explanation rather than surrounding narration. Even when a tool name in the question is absent verbatim in the transcript, the model can still focus on segments describing the corresponding instrument or action. Failure cases often involve videos with very long transcripts where the answer spans across multiple sliding windows, or where ASR errors introduce misleading or missing text. The duration penalty effectively suppresses excessively long predictions, but occasionally over-penalizes correctly long answer spans, leading to underestimated boundaries.

7.5 Task-C Output Generation

Using the tuned short+penalty configuration, we ran the full Task-C set (C1–C80) and produced a submission file `task_c_submission.json` in the required format. The results reported in Table 1 are computed directly from that submission against the held-out ground truth.

		IoU@0.3	IoU@0.5	IoU@0.7	Mean IoU
1	LAMAR-2	93.75	90.00	77.50	79.55
2	NJUST-KMG	92.50	81.25	67.50	75.48
3	405621	60.00	55.00	47.50	50.78
4	TXT66 (Seahawk)	71.25	52.50	42.50	52.30
5	UNCC	62.50	36.25	22.50	42.57
6	ADAPT	10.00	10.00	8.75	8.62

Table 2: Task C validation results for submitted systems. Scores are shown at IoU thresholds 0.3, 0.5, and 0.7, together with mean IoU. Systems are ordered by IoU@0.7.

8 Conclusion

We presented a practical pipeline for LLM-assisted temporal answer grounding in medical procedural videos using segment-aligned transcriptions. Restricting the LLM to selecting segment ranges rather than generating timestamps enables deterministic and robust timestamp extraction. Sliding-window prompting scales the approach to long transcripts, and duration penalties significantly improve mIoU by discouraging overly long spans. Future improvements include higher-quality ASR, boundary refinement, and embedding-based candidate scoring for robustness to paraphrase.

9 Limitations

Despite the promising results, our approach has several limitations. First, the system relies entirely on timestamped transcripts derived from automatic speech recognition (ASR); errors in transcription or timing can directly degrade localization performance. Second, the current method is text-only and does not use visual information from the video frames, which may be essential for recognizing actions, instruments, or visual cues that are not well described in the transcript. Third, the sliding-window strategy approximates full-video reasoning but may miss answer spans that cross window boundaries or fall into low-overlap regions. Fourth, the duration penalty and other hyperparameters (e.g., window size, overlap, λ) were tuned on a small sanity-check set; their generalizability to the full test set is not guaranteed. Fifth, the segment-range selection prompt assumes that the answer corresponds to a contiguous block of transcript segments, which may not hold for videos where the

answer is interleaved with irrelevant narration. Finally, invalid or null LLM outputs are still possible, and we fill missing predictions with placeholders, which limits the system's robustness in production settings. Future work should address these issues by incorporating multimodal cues, refining boundary detection, and developing more robust prompting or fine-tuning strategies.

References

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2023. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2024. Towards answering health-related questions from medical videos: Datasets and approaches. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16399–16411, Torino, Italia. ELRA and ICCL.

Deepak Gupta, Collin Scott Campbell, Pedram Golnari, and Dina Demner-Fushman. 2026. Overview of the medgenvidqa 2026 shared task on medical generative video question answering. In *Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026)*, San Diego, USA. Association for Computational Linguistics.

Deepak Gupta and Dina Demner-Fushman. 2022. Overview of the MedVidQA 2022 shared task on medical video question-answering. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 264–274, Dublin, Ireland. Association for Computational Linguistics.