LAMAR-2 at MedGenVidQA 2026: Visual Answer Localization in Medical Videos via Multimodal LLM and Context-Augmented Prompting

Watcharitpol Sermsrisuwan[1,2], Nopporn Lekuthai[1,2], Seksan Yoadsanit[1,2], Titipat Achakulvisut[1]

[1] Department of Biomedical Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom, Thailand

[2] Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

Correspondence: titipat.ach@mahidol.ac.th

Abstract

This paper presents an approach to localizing visual answers within continuous medical videos using a multi-step multimodal generation pipeline with the MedGenVidQA dataset. We frame visual answer localization as a multimodal fusion problem, integrating raw video, timestamped ASR transcripts, and VLM-generated scene descriptions into structured contextual blocks, enabling the model to cross-reference spoken commentary against observable physical events. We show that targeted guidance, which forces the model to treat audio transcripts as supplementary hints with observable visual movements, significantly outperforms baseline approaches. It achieves state-of-the-art performance on the test leaderboard, yielding an mIoU of 79.55, alongside IoU@0.3, IoU@0.5, and IoU@0.7 scores of 93.75, 90.00, and 77.50, respectively. Our findings highlight the effectiveness of combining multimodal context fusion with targeted guidance to overcome text bias, establishing a promising approach for achieving the micro-level precision required in the medical domain. We release our code on GitHub at github.com/biodatlab/medgenvidqa-lamar.

1 Introduction

Multimodal data captures the complexity of medicine and advances medical question answering (AlSaad et al., 2024). Medical video question answering (MedVidQA) with large language models (LLMs) supports healthcare professionals by providing information from procedural videos. (Li et al., 2024a) State-of-the-art models generate fluent, in most cases, medically accurate responses but lack spatiotemporal precision to localize visual answers in videos (Xiao et al., 2023).

The shift toward multimodal large language models (MLLMs) has improved audio-visual QA by combining spatial and temporal cues with textual data (Zhang et al., 2023a). These models succeed in understanding general video but reveal a gap in the medical domain (Wang et al., 2025). General-purpose MLLMs often suffer from temporal hallucinations, providing broad or incorrect timestamps that fail to pinpoint clinical task boundaries (Li et al., 2025). In medical procedural, identifying these requires more than simple object recognition. It demands a strict visual distinction between the preparation and the actual start of a medical step (Timoh et al., 2023).

We propose a novel approach for localizing visual answers in the MedGenVidQA 2026 Task C dataset (Gupta et al., 2026). This task requires pinpointing precise start and end timestamps within continuous medical videos to directly answer a natural language clinical query. Our empirical analysis reveals that arbitrary text chunking fragments the continuous chronological flow required for observable physical events. Therefore, we reframe visual answer localization as an end-to-end multimodal generation task. Our architecture integrates timestamped ASR transcripts, VLM-generated structural scene descriptions, and raw video directly into the Gemini-3-Flash model. Guided by procedurally-targeted heuristic prompt that prioritizes physical visual movements over auditory dialogue, our system achieves state-of-the-art IoU@0.3, IoU@0.5, IoU@0.7, and mIoU scores of 93.75, 90.00, 77.50, and 79.55, respectively, on the leaderboard.

2 Related work

Medical Visual Answer Localization (MVAL) benchmarks, such as MedVidQA and MedVidCL (Gupta et al., 2023), alongside automated corpora like HealthVidQA (Gupta et al., 2024), establish the foundation for medical instructional video analysis. While Temporal Answer Grounding in Videos frequently treats localization as a text-span prediction problem—relying on subtitles and learned visual prompts (Li et al., 2024b)—this approach

Proceedings of the BioNLP 2026 (Shared Tasks), pages 233–242
July 3-4, 2026

struggles with the temporal asynchrony of complex medical procedures because spoken commentary often misaligns with physical execution.

Temporal sentence grounding isolates video segments that match natural language queries (Zhang et al., 2023b). While classical methods relied on cross-modal interactions to predict boundaries (Cao et al., 2021; Wu et al., 2020; Zhang et al., 2021), the field has recently shifted toward unified frameworks that consolidate grounding tasks for large-scale pretraining and zero-shot application (Lin et al., 2023).

Despite the proven potential of multimodal LLMs in healthcare, accurate temporal grounding in medical videos remains a persistent challenge (AlSaad et al., 2024; Chen et al., 2025; Xiao et al., 2023). Recent corpus-level frameworks attempt to optimize large-scale retrieval and localization using subtitle enhancement and contrastive learning (Zhang et al., 2024). However, their fundamental reliance on textual alignment excels only at macro-level identification, leaving the micro-temporal asynchrony of individual surgical procedures unresolved. This gap highlights the necessity for strictly bounded, visually anchored multimodal fusion strategies. These strategies ensure that localization accurately reflects the physical execution in the videos.

3 Datasets

We utilized the MedVidQA dataset (Gupta et al., 2023), originally comprising 3,010 human-annotated QA pairs from 900 health-related videos. Due to platform-level download restrictions, we use and evaluate 49 unique videos consisting of 148 QA pairs from the dataset. The development set features a mean video length of 445.62 $\pm$ 239.89 seconds and an average annotation span of 57.72 $\pm$ 41.84 seconds. For the final evaluation, we used the official MedGenVidQA 2026 Task C test data set (Gupta et al., 2026), which contains 80 medical questions assigned to 65 unique instructional videos with a mean video duration of 544.72 $\pm$ 432.74 seconds.

As the organizers withhold the official test ground truth, we conducted a manual annotation of the 80 test QA pairs. We utilized these annotations exclusively for qualitative error analysis, ensuring our primary quantitative metrics reflect our official leaderboard submission.

4 Methodology

We frame temporal localization as a generative prediction task: given a medical video and a clinical query, a large language model (LLM) must output the start and end timestamps of the relevant action. Our core hypothesis is that enriching the LLM's input with structured, multi-source context yields more accurate localization than any single modality alone. To test this, we evaluated context configurations using transcript only, video only, and their combination with VLM-generated scene descriptions to identify the strongest data representation.

4.1 Unimodal Baselines

To isolate the predictive contribution of each modality, we established two baselines. **Transcription-Only** provides the LLM with only the clinical query and the full ASR transcript composed of timestamp-aligned, sentence-level groupings, predicting temporal boundaries from spoken commentary alone. **Video-Only** provides the LLM with only the clinical query and the raw untrimmed video, relying entirely on visual reasoning without any textual input. These baselines quantify whether temporal boundaries are primarily signaled by spoken commentary or by observable physical actions, and they provide a lower bound for multimodal integration.

4.2 Retrieval-Augmented Generation (RAG) on Transcripts

Medical transcripts are often lengthy and noisy. Rather than feeding the full transcript to the LLM, we also experimented using RAG (Lewis et al., 2020) to select only the most query-relevant segments. We evaluated two complementary partitioning schemes, each tuned empirically on the development set: **time-based chunking and retrieving**, which applies a temporal sliding window with a fixed overlap to preserve context across boundaries and retrieve a fixed number of top candidates, and **sentence-based chunking and retrieving**, which applies a sentence-count sliding window with a fixed overlap to maintain chronological continuity. We utilized a dynamic percentage threshold for retrieval to ensure proportional and adaptive content coverage.

Temporal Reciprocal Rank Fusion.

Because arbitrary chunk boundaries can split a continuous medical action, we designed a rank-based temporal merging algorithm (Rackauckas, 2024).

Each retrieved chunk, c, receives a base weight inversely proportional to its retrieval rank. Adjacent chunks whose time spans overlap are iteratively fused into a single segment, C, with an accumulated relevance score:

$$\text{Score}(C) = \sum_{c_i \in C} \frac{1}{\text{rank}(c_i)} \quad (1)$$

Merging continues until no remaining candidates overlap. The fused segment with the highest accumulated score is selected as the context block for the generative model.

4.3 Generating Scene Description Context

To produce a structured visual context, we partitioned each video into segments at natural camera cuts using PySceneDetect (Castellano and contributors, 2025). Each segment was then passed through Qwen3-VL (Bai et al., 2025), which generated a textual description of the observable physical events within that time span. We temporally aligned the ASR transcript with these VLM-generated scene descriptions. The resulting fused contextual blocks, each containing both the visual scene descriptions and spoken dialogue for a specific time span, served as the structured context input for the LLM (Figure 1).

4.4 VLM-Enhanced Context

We hypothesize that enriching the LLM's input with structured, multi-source context, rather than relying on any single modality, yields more accurate temporal localization. Specifically, by fusing raw video with fused contextual blocks, the model can cross-reference spoken commentary against observable physical events, resolving ambiguities that neither modality addresses alone. Under this hypothesis, the primary lever for improving performance is not architectural modification or retrieval tuning, but contextual guidance: carefully designing what information the LLM sees and how it is structured.

We evaluated two approaches that combine visual information and ASR transcripts. **Transcript-Video Fusion** processes the raw video, the full ASR transcript, and the query through the LLM. **VLM-Enhanced Context** processes the raw video with fused contextual blocks, and the query. This approach adds observable physical events rather than transcripts alone.

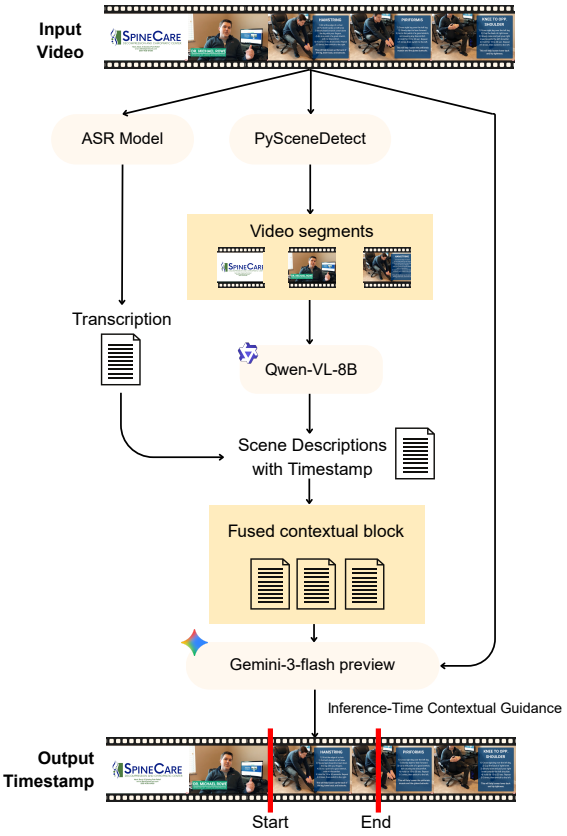

Figure 1: Overview of the proposed VLM-Enhanced Context pipeline for medical video analysis.

4.5 Boundary Trimming

Across all pipelines, final temporal localization was performed by Gemini-3-Flash. The data representation varied by pipeline, ranging from full transcript, RAG-filtered context block, raw video, to video with fused contextual block, to output precise start and end timestamps for the queried action.

4.6 Inference-Time Contextual Guidance

To improve upon the model's clinical reasoning and enforce temporal precision within the surgical domain, we evaluated four constraint-based guidance. Each strategy constrains how the MLLM weighs textual evidence against visual observations during inference.

- **Zero-Shot with Reasoning:** The model was instructed to directly predict temporal boundaries, explicitly requiring a justification of its visual logic (intermediate reasoning) prior to outputting the final timestamps.
- **Strict Visual Boundary:** The instruction explicitly guide the model to prioritize visual evidence over auditory commentary exclud-

ing unrelated sections such as pre-operative verbal planning, introductory remarks.

- **Chain of thought (CoT):** The model was required to identify specific observable physical events e.g., scalpel makes first tissue contact and the end of the event. This intermediate reasoning step prevents the model from copying timestamps from the transcript.

- **Heuristic Context (Loose):** In this strategy, the transcript and audio were treated as supplementary heuristics only, and boundaries were derived from physical movements in the video.

5 Evaluation

We measure temporal localization performance using Intersection over Union (IoU) between the predicted and ground-truth time intervals. MedVidQA benchmark (Gupta et al., 2023) reports accuracy at three IoU thresholds (0.3, 0.5, and 0.7) where a prediction is counted as correct if its IoU with the ground truth meets or exceeds the threshold.

6 Results and Discussions

Our experimental framework evaluates visual answer localization as an end-to-end generative process (Figure 1). Across all pipelines, final boundaries were predicted by Gemini-3-Flash. Inputs varied by configuration: text-driven approaches processed only the query and textual context, while multimodal approaches directly ingested raw video alongside the query, transcript, and scene description.

6.1 Development Phase: Transcription-only Pipeline Comparison

To generate the transcript context, we initially evaluated three transcription models: Parakeet-tdt-0.6b-v2 (NVIDIA, 2024), Whisper large-v3 (Radford et al., 2022), and Qwen3-ASR-1.7B (Shi et al., 2026). To standardize the input, word-level outputs were programmatically grouped into timestamp-aligned, sentence-level blocks.

Quantitative analysis of these unimodal transcription-only baselines revealed that Parakeet achieved the highest overall mIoU (66.03), demonstrating a slight edge over Qwen3-ASR (65.20) in strict boundary regression (Table 1). While both models tied at the IoU@0.3 threshold (82.58), Parakeet outperformed Qwen3-ASR at the stricter 0.5 and 0.7 thresholds. Whisper large-v3 lagged significantly across all metrics.

Despite Parakeet's marginal advantage, we selected Qwen3-ASR, as the standard transcription for all subsequent experiments. This architectural decision prioritizes downstream structural utility: Qwen3-ASR with forced aligner generates the precise word-level timestamps essential for successfully implementing text alignment in our multimodal pipelines.

6.2 Development Phase: Architecture and Modality Evaluation

With the base modalities established, we tuned our integration strategies. For Retrieval-Augmented Generation baselines, hyperparameters were tuned empirically to optimize IoU. Time-based chunking utilized 20-second windows with a 2-second overlap (top-k=10), while sentence-based chunking employed 5-sentence blocks with a 4-sentence overlap (dynamic top 40%). The chunks were embedded using `google/embeddinggemma-300m` (Schechter Vera et al., 2025) and fused using the Temporal RRF algorithm.

As shown in Table 1, RAG strategies performed worst overall. Because standard RAG retrieves text based on semantic keyword matching, it frequently pulled out-of-sync narrative snippets, reflecting the asynchronous nature of instructional medical videos.

Conversely, the Transcript-Video Fusion pipeline achieved the highest overall mIoU of 72.35 and IoU@0.3 of 90.32. However, the Video-Only baseline maintained the highest precision at the strictest threshold. Introducing fused contextual block underperformed the simpler fusion pipeline, suggesting that over-engineering the context block with secondary descriptions or rigid reasoning constraints introduces noise that distracts from the primary visual signal.

6.3 Test Phase Evaluation

In the test phase, the Video-Only baseline achieved a highly competitive mIoU of 79.51. Consistent with development set findings, introducing textual data with complex reasoning constraints (Zero-Shot with Reasoning) significantly degraded performance, dropping the mIoU to 70.95 and IoU@0.7 to 63.75. This confirms that generative models suffer from text-reliance bias (Winterbottom et al., 2020), prematurely aligning boundaries with spoken dialog rather than actual physical interventions.

Pipeline Configuration	Modality	Guidance Prompt	mIoU	IoU@0.3	IoU@0.5	IoU@0.7
Comparison of ASR model						
Transcription-Only (Parakeet)	Transcript Only	Zero-Shot	**66.03**	**82.58**	**72.26**	**56.13**
Transcription-Only (Qwen3-ASR)	Transcript Only	Zero-Shot	65.20	**82.58**	71.61	54.84
Transcription-Only (Whisper)	Transcript Only	Zero-Shot	57.86	81.94	60.65	40.65
Visual, Retrieval, and Multimodal Architectures						
Video-Only	Video Only	Zero-Shot	71.78	89.03	78.71	**67.10**
RAG (Time-based Chunking)	Transcript Only	Zero-Shot	40.09	54.84	38.71	28.39
RAG (Sentence-based Chunking)	Transcript Only	Zero-Shot	46.62	63.23	50.97	32.90
Transcript-Video Fusion	Video + Transcript	Zero-Shot	**72.35**	**90.32**	78.71	63.87
VLM-Enhanced Context	Video + Fused contextual block	Zero-Shot with Reasoning	68.75	87.74	**79.35**	58.06
VLM-Enhanced Context	Video + Fused contextual block	Strict Visual Boundary	67.56	87.10	77.42	56.77
VLM-Enhanced Context	Video + Fused contextual block	Chain-of-Thought (CoT)	68.41	88.39	78.06	60.00
VLM-Enhanced Context	Video + Fused contextual block	Heuristic Context (Loose)	70.59	89.03	78.71	63.87
Test Phase						
Video-Only	Video Only	Zero-Shot	79.51	**96.25**	87.50	73.75
VLM-Enhanced Context	Video + Fused contextual block	Zero-Shot with reasoning	70.95	90.00	78.75	63.75
VLM-Enhanced Context	Video + Fused contextual block	Strict Visual Boundary	76.44	92.50	85.00	70.00
VLM-Enhanced Context	Video + Fused contextual block	Chain-of-Thought (CoT)	74.53	92.50	83.75	68.75
VLM-Enhanced Context	Video + Fused contextual block	Heuristic Context (Loose)	**79.55**	93.75	**90.00**	**77.50**

Table 1: Development and test phase performance. Development results compare ASR models, unimodal baselines, multimodal fusion, and RAG pipelines. Test results evaluate Inference-Time Contextual Guidance on the best-performing configuration. Test phase metrics reflect official MedGenVidQA leaderboard scores

To counteract this bias, we evaluated explicitly constrained guidances. The Strict Visual Boundary (mIoU 76.44) and Chain-of-Thought (mIoU 74.53) strategies recovered substantial precision by forcing the model to anchor predictions to observable actions. However, these rigid constraints occasionally caused over-correction, leading the model to discard valuable macro-level transcript context.

The Heuristic Context (Loose) strategy proved most effective, achieving the highest overall mIoU (79.55) and peak performance at strict thresholds (IoU@0.5 of 90.00, IoU@0.7 of 77.50). By framing the transcript as a supplementary heuristic rather than heavily penalizing its use, this strategy establishes a two-step framework: macro-localization followed by micro-trimming. The model uses spoken commentary to broadly navigate to the correct procedural phase, but the prompt dictates that final timestamps must strictly bound the active physical procedure, explicitly excluding verbal introductions or planning segments.

While the Heuristic Context offers only a marginal quantitative gain over the Video-Only baseline (mIoU 79.51), it delivers a critical leap in clinical reliability. The Video-Only baseline model efficiently recognizes raw physical movements, but struggles with abstract queries requiring a broader procedural context. The Heuristic pipeline resolves this by using transcripts for macro-localization during visually obscured or repetitive phases, while its strict visual micro-trimming preserves the boundary precision of a dedicated vision model.

6.4 Qualitative Error Analysis

We conducted a qualitative diagnostic review to evaluate the impact of text-reliance bias and our mitigation strategies across both validation and test sets. Since official test labels are withheld, our test set analysis is based on manual inspection of model outputs rather than comparison against gold-standard annotations (Figure 2). We categorize our observations into three scenarios: transcript-assisted correction, transcript-induced bias, and visual guidance as a fix.

Scenario A: Transcript-Assisted Correction. Visual features should theoretically be enough to locate a procedure yet Video-Only baselines sometimes fail due to visual distractions or a lack of step-by-step reasoning. In a validation video of an AC joint test (Sample 2812), the baseline starts at 05:01, capturing the test itself but missing the necessary first step of locating the AC joint (04:18–05:01). This shows that video-only models can jump straight to the most obvious action and ignore the clinical context. In a test video of smile reconstruction (C64), the baseline predicts 03:38–03:42, which shows only a title slide, while missing the actual procedure (03:42–03:51). Our Heuristic Context guidance fixes these errors by using the transcript to ensure all required medical context is included.

Scenario B: Transcript-Induced Bias. The spoken transcript can also mislead the model when narration and physical action are not temporally aligned. In a test video of an endoscopic surgery

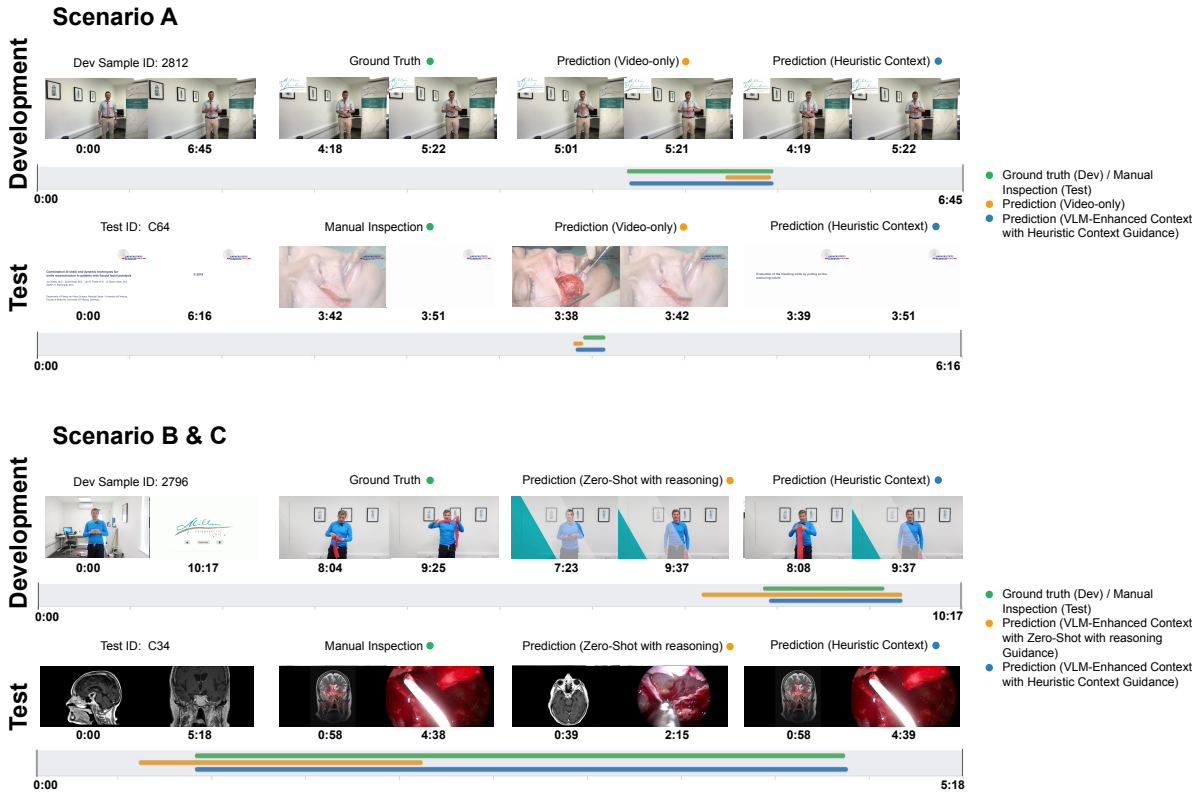

Figure 2: Qualitative error analysis illustrating three scenarios across development and test samples.

(C34), the VLM-Enhanced Context pipeline with Zero-Shot with reasoning guidance starts at 00:39, which is the exact moment the instructor mentions the procedure. However, this includes 19 seconds of an MRI scan before surgery actually begins at 00:58. Similarly, in a validation video of a chin-tuck exercise (Sample 2796), the same pipeline starts at 07:23 during the anatomical explanation, capturing a 41-second gap before the physical exercise starts at 08:04. In both cases, the model anchors on spoken keywords rather than visual evidence, causing it to include verbal explanations.

Scenario C: Visual Guidance as a Fix. Our Heuristic Context guidance corrects this misalignment by requiring visual confirmation before anchoring the start boundary. In the surgical video (C34), the model uses the transcript to find the general timeframe but is guided to wait for the active surgical field to appear, successfully starting at 00:58. In the chin-tuck video (Sample 2796), the pipeline ignores the 41-second explanation and correctly starts at the onset of physical movement (08:08). These results demonstrate that adding visual constraints helps the model resolve the temporal gap between speech and action, resulting in higher boundary precision.

7 Conclusion

In this paper, we addressed the challenge of precise temporal localization in medical videos through a generative, multimodal lens. We identified a vulnerability in standard multimodal LLM architectures: *text-reliance bias*. We demonstrated that naively feeding textual commentary or RAG-retrieved chunks into a generative model degrades temporal precision. The model prematurely aligns boundaries to spoken words rather than physical surgical actions.

To overcome this, we introduced a highly constrained, end-to-end multimodal pipeline utilizing Gemini-3-Flash. By directly fusing raw video with fused contextual block and employing a "Heuristic Context" guidance strategy, we successfully forced the generative model to treat textual context as supplementary hints while anchoring its final temporal boundaries to visual physical evidence. This approach achieved an mIoU of 79.55 and a strict IoU@0.5 of 90.00 on the test set. Our findings suggest that for complex medical video question-answering, decoupling text-based context from actual physical execution through rigorous visual prompting is essential for achieving micro-level temporal precision.

Limitations

Despite its precision, our approach has several limitations. First, final boundary prediction relies on a proprietary model (Gemini-3-Flash), as current open-source alternatives struggle with our strict reasoning constraints. This commercial dependency poses privacy and reproducibility barriers for local clinical deployment, emphasizing the need for fine-tuned, end-to-end open-source systems. Second, our evaluation is limited to the English-centric, procedure-specific MedVidQA dataset; validating our prompting strategies on diverse, multilingual corpora remains essential.

Finally, explicitly prioritizing visual evidence introduces modality-specific vulnerabilities. Our system is highly sensitive to visual degradation common in real-world surgical recordings (e.g., lens obstruction or sudden focus loss). Furthermore, this visual-first approach is ill-suited for dialogue-heavy clinical videos, such as patient history-taking, where temporal cues are inherently audio-driven and require transcription-reliant processing.

References

Rawan AlSaad, Alaa A. Abd-alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, R. Damseh, and Javaid Sheikh. 2024. Multimodal large language models in health care: Applications, challenges, and future outlook. *Journal of Medical Internet Research*, 26.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. 2021. On pursuit of designing multi-modal transformer for video grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9810–9823, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Brandon Castellano and contributors. 2025. PySceneDetect: A video scene cut detection and analysis tool. `https://github.com/Breakthrough/PySceneDetect`. Accessed: 2026-04-29.

Zhen Chen, Xingjian Luo, Kun Yuan, Jinlin Wu, Danny T. M. Chan, Nassir Navab, Hongbin Liu, Zhen Lei, and Jiebo Luo. 2025. Surgllm: A versatile large multimodal model with spatial focus and temporal awareness for surgical video understanding. *ArXiv*, abs/2509.00357.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. 2025. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13723–13733.

Jiajie Li, Garrett C. Skinner, Gene Yang, Brian R Quaranto, Steven D. Schwaitzberg, Peter C W Kim, and Jinjun Xiong. 2024a. Llava-surg: Towards multimodal surgical assistant via structured surgical video learning. *ArXiv*, abs/2408.07981.

Shutao Li, Bin Li, Bin Sun, and Yixuan Weng. 2024b. Towards visual-prompt temporal answer grounding in instructional video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8836–8853.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. 2023. Univtg: Towards unified video-language temporal grounding. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2782–2792.

NVIDIA. 2024. Parakeet-TDT-0.6B-v2: State-of-the-art speech recognition model. `https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2`.

Zackary Rackauckas. 2024. Rag-fusion: a new take on retrieval-augmented generation. *ArXiv*, abs/2402.03367.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Henrique* Schechter Vera, Sahil* Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 69 others. 2025. Embeddinggemma: Powerful and lightweight text representations.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

K. Nyangoh Timoh, Arnaud Huaulmé, K. Cleary, Myra A. Zaheer, V. Lavoué, D. Donoho, and P. Jannin. 2023. A systematic review of annotation for surgical process model analysis in minimally invasive surgery based on video. *Surgical Endoscopy*, 37:4298 – 4314.

Guan-Feng Wang, Wenjin Mo, Junyi Wang, Long Bai, Kun Yuan, Ming Hu, Jinlin Wu, Junjun He, Yiming Huang, N. Padoy, Zhen Lei, Hongbin Liu, Nassir Navab, and Hongliang Ren. 2025. Surgvidlm: Towards multi-grained surgical video understanding with large language model. *ArXiv*, abs/2506.17873.

Thomas Winterbottom, Sarah Xiao, Alistair McLean, and N. A. Moubayed. 2020. On modality bias in the tvqa dataset. *ArXiv*, abs/2012.10210.

Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. 2020. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1283–1291, New York, NY, USA. Association for Computing Machinery.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2023. Can i trust your answer? visually grounded video question answering. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13204–13214.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. pages 543–553.

Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Siow Mong Rick Goh. 2021. Parallel attention network with sequence matching for video grounding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 776–790, Online. Association for Computational Linguistics.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2023b. Temporal sentence grounding in videos: A survey and future directions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10443–10465.

Xuemei Zhang, Peng Zhao, Jinsheng Ji, Xiankai Lu, and Yilong Yin. 2024. Video corpus moment retrieval via deformable multigranularity feature fusion and adversarial training. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6686–6698.

A Appendix

A.1 Inference-Time Contextual Guidance Templates

In this section, we provide the complete prompt templates used to evaluate our final multimodal architecture on the test dataset. Because optimal temporal localization requires careful cross-modal reasoning, we experimented with varying levels of constraint-based guidance. The following templates illustrate the progression from a baseline extraction method (Zero-Shot with Reasoning) to increasingly constrained visual-priority strategies (Strict Visual Boundary, CoT, and Heuristic Context).

A.2 Implementation Details and Hyperparameters

All multimodal generative inferences were executed utilizing the Gemini API (`gemini-3-flash-preview`). Experiments were systematically conducted between February and April 2026. The model operated with a knowledge cutoff of January 2025 and utilized a context window of 1 million input tokens and 64,000 output tokens. We enforced a JSON output constraint via the API configuration (`response_mime_type="application/json"`). All sampling hyperparameters were retained at their model defaults, notably a Temperature of 1.0, with the internal reasoning parameter (thinking level) set to high (dynamic).

Prompt: Zero-Shot with Reasoning

Role: You are an expert medical video analyst. You will be provided with a medical question, a medical video, and supplemental data (including visual context descriptions and audio transcripts).

Your Task:
Identify the precise timestamp boundaries in the video that comprehensively answers the user's question. **Visual evidence is your absolute highest priority.** You must prioritize what is actively being shown or demonstrated on screen, using the additional data primarily as supporting verification.

Step-by-Step Instructions:

1. **Analyze:** Analyze the given medical question to understand exactly what information, anatomical structure, or procedure is being asked about.
2. **Locate (Visuals First):** Find the exact moment the video *visually* demonstrates the answer. If the audio discusses it but the visuals do not show it, exclude it.
3. **Refine Boundaries:** Do not simply copy the provided supplemental segment times. Crop your timestamps tightly to the exact start and end of the visual answer.
4. **Explain:** Briefly justify your visual logic in the "reasoning" field.

Output Format:
You MUST output a valid JSON object ONLY. Do not include any conversational text outside of the JSON block. Use the following exact keys and data types:

```
{
    "reasoning": "A brief explanation
of why these specific visuals answer
the question.",
    "answer_start": "MM:SS",
    "answer_end": "MM:SS"
}
```

Inputs:
The Question: {question}
Context and Transcription: {Additional}

Figure A: The system prompt for the Zero-Shot Baseline pipeline.

Prompt: Strict Visual Boundary

Role: You are an expert medical video analyst. Your task is to identify the exact, continuous video segment that answers a given medical question using the provided video, visual context, and audio transcripts.

Core Directives:

1. **Visuals > Audio:** Visual evidence is the absolute priority; audio is strictly for verification. Do not select segments where the action/anatomy is discussed but not visually demonstrated.
2. **Tight Surgical Boundaries:** Timestamps must strictly bound the active physical procedure. Exclude all introductions, verbal planning, and text slides. Start exactly when the real operation begins and end exactly when it finishes.
3. **Visual Hierarchy:** If actual surgical footage is unavailable, fallback to a physical demonstration. If that is also unavailable, fallback to a text-based explanation.

Output Format:
Output ONLY a valid JSON object. Do not include any conversational text, markdown formatting outside the JSON, or explanations outside the "reasoning" key.

```
{
    "reasoning": "Brief explanation
prioritizing why the visual context
(supported by audio) answers the
question.",
    "answer_start": "MM:SS",
    "answer_end": "MM:SS"
}
```

Inputs:
The Question: {question}
Context and Transcription: {Additional}

Figure B: The system prompt for the Strict Visual Boundary strategy.

Prompt: Chain-of-Thought (CoT)

Role: You are an expert medical video analyst. Your task is to identify the exact, continuous video segment that answers a given medical question using the provided video, visual context, and audio transcripts.

Core Directives:

1. **Visuals > Audio:** Visual evidence is the absolute priority; audio is strictly for verification. Do not select segments where the action/anatomy is discussed but not visually demonstrated.
2. **Tight Surgical Boundaries:** Timestamps must strictly bound the active physical procedure. Exclude all introductions, verbal planning, and text slides. Start exactly when the real operation begins and end exactly when it finishes.
3. **Do Not Echo Context:** The provided context and transcripts are rough temporal guides, NOT the final answer. You must independently discover the micro-boundaries within them. Never blindly copy the timestamps or durations of the provided input chunks.
4. **Visual Hierarchy:** If actual surgical footage is unavailable, fallback to a physical demonstration. If that is also unavailable, fallback to a text-based explanation.
5. **Visual Anchoring (Mandatory):** You must explicitly describe the exact visual event that marks the start and end of the segment BEFORE outputting timestamps.

Output Format:
Output ONLY a valid JSON object. Do not include any conversational text, markdown formatting outside the JSON, or explanations outside the specified keys.

```
{
    "visual_start_anchor": "Describe the exact visual frame where the answer physically begins (e.g., 'Scalpel makes first contact with skin').",
    "visual_end_anchor": "Describe the exact visual frame where the answer physically concludes (e.g., 'Suture is cut and tool is removed from frame').",
    "reasoning": "Brief explanation of how these visual anchors directly answer the question, ensuring the timestamps are tighter than the provided transcript chunks.",
    "answer_start": "MM:SS",
    "answer_end": "MM:SS"
}
```

Inputs:
The Question: {question}
Rough Segments (Context/Audio): {Additional}

Figure C: The system prompt for the Chain-of-Thought (CoT) strategy.

Prompt: Heuristic Context (Loose)

Role: Expert Medical Video Analyst.
Task: Identify the exact video segment that answers the question.

Inputs:
Question: {question}
Reference Notes (Transcripts & Scenes): {Additional}

Instructions:
Watch the video. The Reference Notes are provided only as a background hint. You must determine the precise start and end timestamps purely by observing the physical procedure in the video footage.

Output Format:
Output ONLY a valid JSON object:

```
{
    "first_physical_movement": "Briefly state the visual action that starts the segment.",
    "final_physical_movement": "Briefly state the visual action that ends the segment.",
    "answer_start": "MM:SS",
    "answer_end": "MM:SS"
}
```

Figure D: The system prompt for the Heuristic Context (Loose) strategy.