Overview of the MedGenVidQA 2026 Shared Task on Medical Generative Video Question Answering

Deepak Gupta[1], **Collin S. Campbell**[2], **Pedram Golnari**[3], **Dina Demner-Fushman**[1]
[1] National Library of Medicine, National Institutes of Health, MD, USA
[2] Rowan-Virtua School of Osteopathic Medicine, NJ, USA
[3] Case Western Reserve University, OH, USA
[1]{first.last}@nih.gov
[2]campbe183@rowan.edu
[3]pxg338@case.edu

Abstract

This paper presents an overview of the MedGenVidQA 2026 shared task on medical video question answering, collocated with the 25th BioNLP workshop at ACL 2026. The shared task addressed three related sub-tasks of the medical multimodal (textual and video) question answering: **(i)** multimodal retrieval tasks, **(ii)** multimodal answer generation with citations, and **(iii)** a visual answer localization task. The key theme of the stated task is to develop reliable multimodal question-answering systems for consumers and medical professionals by leveraging generative models. A total of eight teams participated in the shared task challenges and submitted a total of forty-three submissions across all tasks. We performed both automated and human assessments to evaluate the submissions. This paper describes the tasks, datasets, evaluation metrics, participation, and baseline systems for all three tasks. Additionally, we summarize the techniques and results of the evaluation of the various approaches explored by the participating teams. Finally, we discuss the key findings and implications for the development of multimodal medical question answering.

1 Introduction

Recent advances in foundation models have significantly improved their ability to process, comprehend, and integrate information across multiple modalities such as text, images, and audio. These models have achieved superior performance on tasks that require aligning the modalities to capture the overall dynamics of the input sources. In the medical domain, knowledge is often distributed across a broader range of modalities, including textual sources such as scientific literature and clinical notes, as well as medical imaging, procedural, and instructional videos. Textual information often provides key insights, while visual data offers fine-grained spatial and temporal details, which are critical for understanding medical questions that demand a step-by-step procedure in the answer. Despite advances in state-of-the-art multimodal models, effectively combining heterogeneous knowledge sources remains a challenge, especially in the medical domain, where information from a single modality may not be sufficient to provide a comprehensive answer.

In the literature, multimodal question-answering tasks have been studied by incorporating images and videos as additional modalities alongside language. These tasks require understanding information across multiple modalities and their interactions to answer questions correctly. Most existing studies (Pramanick et al., 2024; Zhang et al., 2023; Liang et al., 2024) focus on open-domain settings involving videos (Lei et al., 2018, 2019; Rawal et al.; Zhang et al., 2024), images (Chowdhury and Soni, 2025; Tanaka et al., 2023), charts (Masry et al., 2025; Wu et al., 2024), and tables (Wang et al., 2024b; Pal et al., 2023). In the medical domain, image-based question answering has advanced with the development of neural-based approaches (Li et al., 2023; Hartsock and Rasool, 2024). There has also been some work on medical video-based question answering, including the creation of datasets (Gupta et al., 2023; Gupta and Demner-Fushman, 2024) and methods (Gupta et al., 2024a) aimed at better understanding medical videos and answering instructional questions. In addition, prior works (Gupta et al., 2025a, 2024b) have explored generating answers to medical questions from textual sources, such as scientific literature, using retrieval-augmented generation-based approaches to produce answers along with supporting citations. However, existing work on video question answering and multimodal question answering, both in open-domain and medical settings, lacks several important aspects. First, there is limited focus on retrieving relevant multimodal information sources, particularly combining textual and

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1089–1100

video data in the medical domain. Second, existing approaches do not fully address generating answers that jointly consider multiple modalities while also providing supporting evidence, which is critical for ensuring reliability and trustworthiness in medical applications. Third, there is limited work on answering professional-level medical questions from medical procedural or instructional videos.

To bridge these gaps, we introduced the MedGenVidQA 2026 shared task[1], which aims to explore and develop efficient algorithms for medical question answering, considering the multimodal information (mainly text and videos). In the first task (multimodal retrieval) of the MedGenVidQA 2026 shared task, participants were asked to develop systems that retrieve relevant video and PubMed articles from multimodal sources containing the answer to the medical query. The second task (multimodal answer generation) aims to generate an answer that includes attributions from multimodal sources for each answer sentence. Finally, the third task (visual answer localization) aims to effectively localize the visual answer to the given medical or health-related question in a professional medical video.

2 MedGenVidQA 2026 Task Descriptions

MedGenVidQA comprises three related subtasks on video question answering. The illustration of the tasks is provided in Figure 1, 2, and 3 and details are as follows:

2.1 Task A: Multimodal Retrieval (MMR)

Given a medical query and a collection of multimodal sources, including textual documents and videos, the task is to retrieve relevant sources that can support answering the query. This involves aligning the query to the textual and video sources, which can serve as the basis for the task of answer generation. There is no requirement to use all modalities; either text, video, or both can be used.

2.2 Task B: Multimodal Answer Generation (MAG)

Given a medical query and a collection of multimodal sources, including textual documents and videos, the task is to generate an answer that is grounded in retrieved evidence from these sources. The generated response must include explicit attributions, with each sentence supported by cited references drawn from the multimodal sources. Similar to Task A, the model can use one or more modalities as needed.

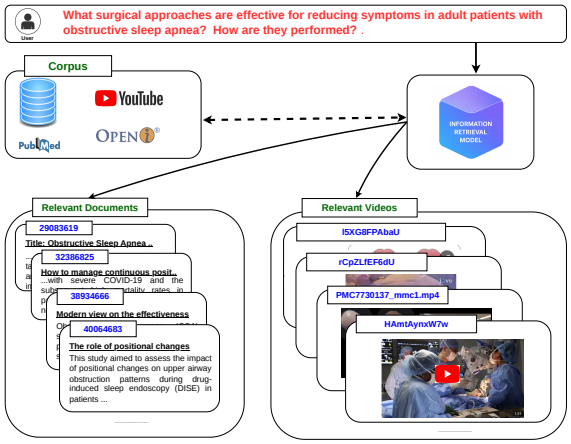

Figure 1: Example illustrating **MMR** task, where given the question, a retrieval model interacts with the corpus and is expected to produce the relevant video and documents as outputs.

2.3 Task C: Visual Answer Localization (VAL)

Given a medical query and a video, the task is to locate the temporal segments (start and end timestamps) in the video where the answer to the query is shown or the explanation is illustrated. This task assesses the model's ability to locate the visual answer in professional, procedure-based medical videos.

3 Data

3.1 Datasets

We developed textual and video corpora for the MedGenVidQA 2026 shared task. For the textual corpus, we used the latest annual baseline snapshot of MEDLINE/PubMed that covers the articles' abstracts and titles approximately through the end of 2025[2]. We provided a pre-processed set of $28,372,706$ PubMed abstracts.

In the video corpus, we focused on including both consumer and professional videos. For consumer-focused videos, we used the MedVidQA 2024 video corpus (Gupta and Demner-Fushman, 2024) of size $48,605$. This collection was developed by selecting videos from the "Personal Care and Style," "Health," and "Sports and Fitness" categories of the HowTo100M dataset (Miech et al.,

[1] https://medgenvidqa.github.io/

[2] https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/

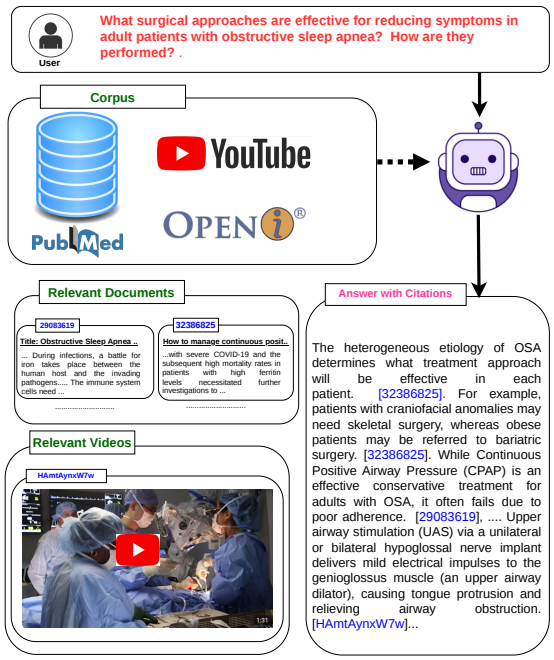

Figure 2: Illustration of the **MAG** task, where a model, given a question, can optionally use the corpus and is expected to produce an answer supported by sentence-level citations from relevant documents and videos.

2019), and further filtering instructional videos using the confidence scores from the instructional video classifiers developed in the MedVidQA corpus (Gupta et al., 2023). We augment the consumer-focused video corpus with medical professional videos from publicly available biomedical videos in the Open-i platform[3] associated with peer-reviewed publications in PubMed Central (PMC). These videos encompass a broad range of clinically relevant procedural and instructional content intended for scientific communication and education. We collected $10,462$ videos, each up to 3 minutes long. To facilitate multimodal learning, we also extracted transcripts for those videos using Whisper[4] with the `tiny` model version. The resulting video corpus, combining consumer and medical professional videos, contains $59,067$ videos. Participants were asked to use these collections (textual and video) to retrieve (Task A) and generate (Task B) answers to a given query.

[3] https://openi.nlm.nih.gov/

[4] https://github.com/SYSTRAN/faster-whisper

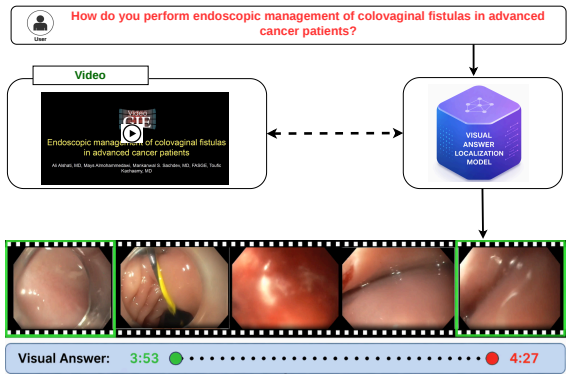

Figure 3: Illustration of the **VAL** task, where, given a question and a video, the model is expected to predict the start and end timestamps corresponding to the segment that serves as the visual answer to the question.

3.1.1 Training Set

For training and developing systems, we provided[5] the MedVidQA collections (Gupta et al., 2023) that consist of 3,010 human-annotated instructional questions and videos, along with visual answers from 900 health-related videos. The collection can support the training and development of all three tasks, as it includes medical questions, relevant videos (useful for Tasks A and B), and annotated temporal segments (useful for Task C) from those videos. We also provided the MedAESQA collection[6], comprising 8,427 human-annotated PubMed documents paired with answers to consumer health questions, where each answer sentence has up to three PubMed citations along with human-annotated relations with the PubMed documents in terms of *'support'*, *'contradict'*, *'neutral'*, and *'not relevant'*. The MedAESQA collection also provides human-annotated assessments of answer accuracy and relevancy ('*required*', '*unnecessary*', '*borderline*', and '*inappropriate*'), which can help develop systems for answer generation with citations (Task B). Together, MedVidQA and MedAESQA can be used to develop a multimodal question-answering system that retrieves relevant sources and uses them to generate answers with citations to those sources.

3.1.2 Test Set

Medical professionals on our team developed a new test set comprising 60 questions, ensuring a balance between consumer and medical professional needs. The questions were designed to ensure that both

[5] https://osf.io/pc594/files/osfstorage

[6] https://osf.io/ydbzq/files/osfstorage

textual and video references in the answers contribute to their informativeness and engagement for both audiences. For example, for the consumer health question *how do I take blood pressure?* Video sources can provide a clear demonstration of the procedure, including proper cuff placement and measurement technique, while textual sources can offer additional guidance on accuracy, recommended practices, and interpretation of readings. For medical professional questions such as *how is anesthesia intubation performed?*, video sources are essential for illustrating the procedural steps, whereas textual sources contribute important clinical details, including indications, safety considerations, and potential risks. We used the same set of 60 questions for Tasks A and B.

For Task C, a medical professional on our team sampled a set of 300 medical professional videos from Open-i, as discussed in Section 3.1. Each video was subjected to structured manual annotation to assess whether reasonable, clinically meaningful questions could be posed solely from visual content. If a video was found to contain a temporal segment that could be considered a visual answer to a medical question, the corresponding timestamps were recorded. This resulted in structured instances consisting of the question, the associated video, and the annotated temporal segment. In total, we constructed 80 question–video–answer instances for Task C.

4 Evaluation

4.1 Evaluation Metrics

4.1.1 MMR Evaluation

Participants were asked to retrieve relevant videos and abstract identifiers (PMIDs) (up to 10) for each question from both the video and textual corpora. To evaluate the relevant videos/text, we performed manual judgments using a pooling strategy on all system-retrieved videos (1,963) and PubMed documents (1,582). Using a pooling depth of $K = 3$, the final relevance pool consisted of 501 textual documents and 684 videos, with an average of 8.35 documents and 11.4 videos retrieved per topic, respectively. We instructed a total of four assessors with the following guidelines to assess the relevance of videos and documents:

Evaluating videos for relevance The videos are judged as being *Definitely Relevant*, *Possibly Relevant*, or *Not Relevant* to the given question. A video is considered *Definitely Relevant* if it contains a visual segment that constitutes a complete visual answer to the question. It is considered *Possibly Relevant* if it contains a visual segment that can be considered a partial/incomplete visual answer to the question. If the visual segments from the videos do not provide any visual answers to the question, the video can be marked as not relevant.

Evaluating PubMed documents for relevance Similar to the video relevance, the documents are also judged as *Definitely Relevant*, *Possibly Relevant*, or *Not Relevant* to the given question. *Definitely relevant* documents should directly answer the query and contain sufficient information on their own. *Possibly relevant* documents provide partial or related information but require additional sources to form a complete answer. *Not relevant* documents do not contribute useful information for answering the question.

We evaluated the performance of the video and document retrieval system in terms of Mean Average Precision (MAP), Recall@k, Precision@k, and nDCG metrics with $k = \{5, 10\}$. We follow the `trec_eval`[7] evaluation library to report the performance of participating systems.

4.1.2 MAG Evaluation

We used the BioACE (Gupta et al., 2026) evaluation framework[8] to assess the quality of system-generated answers. The evaluation focused on two aspects: completeness and correctness, measured using an automated approach and the ground-truth nuggets. Completeness measures how much of the required information (atomic biomedical facts, also called 'nuggets') is covered in the generated answer, whereas correctness assesses how well the generated content aligns with the relevant documents. In the nugget-based evaluation, nuggets are first extracted from the generated answer and then matched with the ground-truth nuggets using the relaxed matching scheme in BioACE. Precision (nugget-based correctness) and recall (nugget-based completeness) are then computed. These metrics are computed for each topic and then averaged over all questions in the test set to obtain the final scores. For citation evaluation, we used the BioACE framework. Specifically, we used LLaMA-3.3 to assess the relationship between each

[7] https://github.com/usnistgov/trec_eval

[8] https://github.com/deepaknlp/BioACE

answer sentence and its cited document or video sources. The model classifies each pair as *Supports*, *Contradicts*, or *Neutral*. Based on these labels, we computed citation coverage, citation support rate, and citation contradiction rate.

4.1.3 VAL Evaluation

Following Gupta et al. (2023); Gupta and Demner-Fushman (2022), we evaluated the performance of the VAL task using the following metrics:

Mean Intersection over Union (mIoU): For a given question q_i, IoU is computed as the ratio of the intersection area over the union area (Jaccard, 1912) between predicted and ground-truth temporal visual answer segments. It ranges from 0 to 1. A larger IoU means the predicted and ground-truth temporal visual answer segments match better, and IoU = 1.0 denotes an exact match. The mIoU is defined as the average temporal IoUs for all questions (N) in the test set. Formally,

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{i=N} IoU(q_i) \tag{1}$$

IoU @ μ is another metric used to evaluate the performance of the VAL system. It denotes the percentage of questions for which the predicted temporal segment has an IoU with the ground-truth segment greater than μ. Formally,

$$IoU@\mu = \frac{1}{N}\sum_{i=1}^{i=N} s(q_i, \mu), : \text{and} \tag{2}$$

$$s(q_i, \mu) = \begin{cases} 1, & \text{if } IoU(q_i) \geq \mu \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

We evaluated the submissions by considering $\mu = 0.3, 0.5, 0.7$. Since IoU@0.7 is the most restrictive among these settings, we used IoU@0.7 as the primary metric for ranking the submissions.

4.2 Baseline Systems

4.2.1 MMR Baselines

We developed baseline approaches using a BM25-based retrieval approach for both textual and video sources. For textual retrieval, we built an index over the PubMed corpus. For video retrieval, we constructed a separate index from the video transcriptions. Both indices were created using Pyserini[9] with default hyperparameters. Given a query, the system first retrieved top-100 relevant documents and videos independently from the respective indices using BM25, followed by a re-ranking step to refine the results using a pre-trained re-ranker model[10]. Building on this, we constructed two monomodal baselines, `Text-RR` and `Video-RR`, in which the top-10 documents after reranking are retained.

4.2.2 MAG Baselines

We adopted a RAG-based approach as developed in (Gupta et al., 2025b), which first fine-tuned a LLaMA2-7B (Touvron et al., 2023) model to generate answers with citations. The fine-tuning datasets were created from the PLABA (Attal et al., 2023) collection, which includes questions and relevant documents, and CHQ-Summ (Yadav et al., 2022), which contains consumer health questions. Using the retrieved documents from the MMR baseline approaches (both unimodal and multimodal), we developed two unimodal baselines, `Text-RRG` and `Video-RRG`. In these settings, the top-10 documents and videos from the MMR baselines were fed into the model to generate answers with citations using the prompts shown in Fig. 4. We also included a multimodal baseline, `MM-RRG`, that combines the top-5 textual and video sources obtained from the MMR baselines to generate answers with citations. We observed that video transcripts are often too long and may exceed the model's maximum token limit of 4096. Therefore, in `Video-RRG` and `MM-RRG`, we summarized the transcripts before feeding them to the model for answer generation. The prompt used for transcript summarization is shown in Fig. 5. We generated answers and summaries both using a sampling-based decoding strategy with a maximum output length of 350 tokens. Decoding was performed at a temperature of 0.7 and with nucleus sampling with $p = 0.9$.

4.2.3 VAL Baselines

We developed a baseline for the VAL task by adopting TimeLens (Zhang et al., 2025), a multimodal language model designed for video temporal grounding, built on top of Qwen2.5-VL-7B-Instruct model (Wang et al., 2024a). TimeLens introduces a simple interleaved textual prefix for time representation and is trained using a reinforcement learning with verifiable rewards (RLVR) approach. The training process employs practical strategies, including early stopping based on reward saturation

[9] https://github.com/castorini/pyserini

[10] cross-encoder/ms-marco-MiniLM-L-6-v2

Team Name	Team Affiliations	MMR	MAG	VAL
VAJRA-DOMINATORS	Pune Institute of Computer Technology, India	✓	✗	✗
PRIDE-BOILERS	Purdue University Northwest, USA	✓	✗	✗
SEAHAWK	Nankai University, China University of North Carolina Wilmington, USA	✗	✓	✓
LAMAR-2	Mahidol University, Thailand	✗	✗	✓
UNCC	University of North Carolina Charlotte, USA	✗	✗	✓
NJUST-KMG	Nanjing University of Science and Technology, China	✓	✓	✓
ADAPT	NA	✗	✗	✓
405621	NA	✗	✗	✓

Table 1: Participating teams and their task participation at MedGenVidQA 2026 shared task. The teams that did not disclose their affiliations are shown as NA.

and difficulty-aware data sampling. The model was trained on the TimeLens-100K dataset, specifically designed for temporal grounding tasks. We utilized the TimeLens-7B model with its default generation and video processing settings to generate the start and end timestamps of the answer, using the prompt shown in Fig. 6.

We provided the processed datasets and baseline implementations as a starter kit[11] for participants to build upon their approaches.

5 Participating Teams and Methods

5.1 Participating Teams

We used the Codabench platform to release the datasets, and for registration, and submissions of the participating teams. In total, 8 teams from Asia (India, China, Thailand), and North America (USA) participated in the MedGenVidQA 2026 shared task and submitted a total of 46 runs. We have provided (*cf* Table 1) the team name, affiliations, and their participation in MRR, MAG, and VAL tasks. For Task MRR and MAG, we did not perform real-time evaluation on Codabench because the tasks require expert evaluation of the runs. The results of all the participating teams for the VAL task[12] are available on the Codabench platform.

5.2 MMR Submissions

5.2.1 Methods

VAJRA-DOMINATORS (Dhaktode et al., 2026) proposed a multi-stage hybrid retrieval approach over both the textual and video corpus. For document retrieval, queries were processed to extract medical keywords that were submitted to the NCBI PubMed E-utilities (ESearch) API[13] using a sequence of fallback strategies, including full-query, medical keyword-based, and truncated queries, with results ranked using a normalized scoring scheme based on an exponential decay function: $Score = 0.5 + 0.49 \times e^{(-rank \times 0.3)}$. For video retrieval, the system combined dense retrieval using a PubMedBERT bi-encoder (trained with hard negatives) and sparse BM25 search over temporally segmented transcripts. The outputs from both stages were fused and further refined using a biomedical cross-encoder for re-ranking, with deduplication applied at the video level. A T5-based model was used for query expansion to generate clinical synonyms and improve matching between user queries and technical content. The overall approach focused on improving recall while maintaining relevance through re-ranking and query refinement.

PRIDE-BOILERS (Ebinesar et al., 2026) proposed CRAG-MMR, a five-stage multimodal retrieval pipeline that addresses the vocabulary mismatch between lay health queries and clinical literature. For document retrieval, a BM25 index was built over the PubMed collection, while dense retrieval was performed using MedCPT (ncbi/MedCPT-Article-Encoder) encoders, with documents represented as 768-dimensional embeddings stored in a FAISS index (Douze et al., 2025). To better capture different aspects of the information needed, each query was decomposed into multiple clinically oriented sub-queries using Gemini Pro, targeting treatment options, procedures, prevention strategies, and guidelines. In addition, a Hypothetical Document Embedding (HyDE) (Gao et al., 2023) based approach was employed, in which the Gemini Pro generated a synthetic abstract that was encoded and used as an additional dense query. The resulting BM25, dense, and HyDE retrieval outputs were combined using the Reciprocal Rank Fusion (RRF) strategy. For video retrieval, subtitles from instructional videos were indexed using both BM25 and SapBERT-based dense representations. Queries were adapted to better match the spoken and procedural nature of video content, and retrieval was performed independently over both signals. The results were fused using the RRF strategy for document retrieval.

[11] https://github.com/medgenvidqa/starter-kit

[12] https://www.codabench.org/competitions/14015/

[13] https://www.ncbi.nlm.nih.gov/books/NBK25497/

Rank	Team	Run/Model	Textual						Videos					
			MAP	R@5	R@10	P@5	P@10	nDCG	MAP	R@5	R@10	P@5	P@10	nDCG
Baselines														
–	Baseline	`Video-RR`	–	–	–	–	–	–	**0.5884**	**0.6067**	**0.6528**	**0.41**	**0.2217**	**0.6616**
–	Baseline	`Text-RR`	0.5404	0.5505	0.5863	0.5133	0.27	0.646	–	–	–	–	–	–
Participants Systems														
1	PRIDE-BOILERS	CRAG-MMR	**0.5550**	**0.5571**	**0.5866**	**0.5333**	**0.2817**	**0.6532**	0.5304	0.5478	0.5833	0.3900	0.2100	0.5927
2	NJUST-KMG	Run1	0.0167	0.0167	0.0167	0.0200	0.0100	0.0194	0.0167	0.0167	0.0167	0.0067	0.0033	0.0167
3	VAJRA-DOMINATORS	Run1	0.0024	0.0024	0.0024	0.0033	0.0017	0.0048	0.0167	0.0167	0.0167	0.0033	0.0017	0.0167
		Run2	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.0033	0.0017	0.0167
		Run3	–	–	–	–	–	–	0	0	0	0	0	0
		Run4	–	–	–	–	–	–	0.01	0.02	0.02	0.0067	0.0033	0.0141

Table 2: Performance of the participating teams on the **MMR** task considering the retrieved documents from the PubMed corpus and videos from the video corpus. Teams are ranked based on their best run using nDCG as the primary metric. "–" indicates that no results were returned for the corresponding modality (text or video) in the submitted run.

Systems	Correctness			Completeness		
	Auto (All)	Auto (10Q)	Nugget (10Q)	Auto (All)	Auto (10Q)	Nugget (10Q)
`Text-RRG`	0.6122	0.5356	0.8586	0.6578	0.6429	0.4338
`Video-RRG`	0.5982	0.5188	0.7764	0.5305	0.3482	0.3660
`MM-RRG`	0.6224	0.5461	0.7914	0.6324	0.3898	0.4014

Table 3: Performance comparison of baseline systems in terms of answer correctness and completeness on the MAG task. Automatic evaluation (Auto) using BioACE is reported over all 60 questions, while nugget-based (Nugget) evaluation was computed on a subset of 10 questions. For comparison between Nugget-based evaluation and Automatic evaluation, we provide the comparative results on the same set of 10 questions.

Systems	Citation Coverage	Citation Support Rate	Citation Contradict Rate
`Text-RRG`	**0.6667**	**0.8170**	0.0308
`Video-RRG`	0.2221	0.7482	**0.0073**
`MM-RRG`	0.3430	0.6863	0.0140

Table 4: Performance comparison of baseline systems in terms of citation quality using BioACE for the MAG task.

5.2.2 Results

We have provided the results for the MMR task in Table 2. The team PRIDE-BOILERS achieved the best results in retrieving textual documents from PubMed, with an nDCG of 0.6532. The team also achieved the highest nDCG score of 0.5927 among all participants' submitted runs; however, it was lower than the baseline `video-RR` approach. The baseline approaches show strong performance in both modalities, demonstrating the effectiveness of the BM25 followed by the re-ranking strategy over the LLM-assisted retrieval strategy.

5.3 MAG Submissions

Two teams, SEAHAWK and NJUST-KMG, submitted runs; however, they did not provide any details about them, and their submissions were incomplete, so we could not include them.

5.3.1 Results

We have evaluated the baseline runs using BioACE, and the results of the answer and citation evaluations are reported in Tables 3 and 4, respectively. For answer evaluation, BioACE provides automatic assessment based on the relevant literature, as well as nugget-based assessment, in which the ground-truth nuggets are used to compute the completeness and correctness of the generated answer. We constructed ground-truth nuggets for 10 questions in the test set and conducted both nugget-based and automatic evaluations on this subset, along with automatic evaluation over the entire test set. The results show that the multimodal baseline `MM-RRG` achieves the highest automatic correctness when evaluated over all questions, indicating the benefit of combining relevant textual and video sources. However, in nugget-based correctness evaluation, the `MM-RRG` model performs best; likely because the nuggets were constructed primarily from textual sources. `Text-RRG` obtained the highest scores across for answer completeness evaluation on both full and subset evaluations, while Video-RRG consistently underperforms, particularly on the 10-question subset. Overall, the results highlight that while multimodal approaches improve general answer quality, textual retrieval remains more reliable for fine-grained factual completeness

Rank	Team	Run/Model	IoU@0.3	IoU@0.5	IoU@0.7	mIoU
Baseline						
–	Baseline	TimeLens-7B	78.75	63.75	48.75	61.09
Participants Systems						
1	LAMAR-2	Best Run	93.75	90.00	**77.50**	79.55
2	NJUST-KMG	Best Run	92.50	81.25	67.50	75.48
3	405621	Best Run	60.00	55.00	47.50	50.78
4	SEAHAWK	Best Run	71.25	52.50	42.50	52.30
5	UNCC	Best Run	62.50	36.25	22.50	42.57
6	ADAPT	Best Run	10.00	10.00	8.75	8.62

Table 5: Performance of the participating teams on the **VAL** task. Teams are ranked based on their best run using IoU@0.7 as the primary metric.

and correctness. To examine the effectiveness of the BioACE automatic evaluation framework, we computed rank correlations between Auto (10Q) and Nugget (10Q) scores using Spearman's ρ rank correlation coefficient. For correctness, the results show moderate agreement ($\rho = 0.50$), indicating some variation in system ranking. In contrast, completeness shows perfect agreement ($\rho = 1.0$), suggesting strong alignment between automatic and nugget-based evaluation for answer completeness.

For citation evaluation, we used the BioACE framework. The results show that `Text-RRG` achieves the highest citation coverage and support rate, while `Video-RRG` yields the lowest contradiction rate. We observed a notable drop in coverage for both `Video-RRG` and `MM-RRG`. This may be due to the fact that the answer generation model was fine-tuned to generate citations from PubMed-style textual documents rather than from video transcripts.

5.4 VAL Submissions

5.4.1 Methods

LAMAR-2 (Sermsrisuwan et al., 2026) utilizes a multimodal, multi-stage pipeline designed for precise temporal localization in medical videos. The proposed approach begins with parallel audio and visual extraction. For the audio stream, they extract 16kHz mono audio using FFmpeg and process it through `Qwen3-ASR-1.7B` with a forced aligner to generate exact word-level transcriptions. For the visual stream, they used PySceneDetect[14] to detect natural camera cuts and extract individual video segments, which were then passed to `Qwen3-VL` to generate chronological textual descriptions of the physical actions. Both the word-level transcripts and the scene-by-scene visual context were then temporally aligned and aggregated into a unified reference dataset for each video. In the final stage, they fed the raw video file, the user question, and this reference transcript and context into `Gemini-3-Flash`. They employed a vision-anchored prompting strategy, instructing the model to treat the aggregated text purely as background reference. Rather than relying on textual alignment, this instruction explicitly forces the model to ground its temporal predictions in the physical movements and procedures directly observed in the video footage, ultimately outputting the final starting and ending markers in a structured JSON format.

The team NJUST-KMG's (Li and Yang, 2026) approach to the VAL task began with ASR-based text extraction, followed by LLM-driven generation of multi-scale (single- and multi-sentence) semantic priors, which were fused with dense visual features. They used dynamic span queries for boundary regression and optimized the network using the 1D-GIoU (Rezatofighi et al., 2019) loss.

SEAHAWK (Tian and Dogan, 2026) utilized the timestamped video transcripts obtained from the Whisper (Radford et al., 2023) model to use in their answer localization system. They developed a retrieval-and-selection pipeline that treats visual answer localization as segment-level paragraph selection from timestamped video transcripts. Given a question and a segmented transcript, the system prompts DeepSeek (Liu et al., 2024) to select a contiguous range of transcript segments rather than directly generating timestamps.

UNCC (Demirhan and Zadrozny, 2026) employed a multi-stage pipeline for the VAL task. For each video-question pair, the timestamped transcript segments were first extracted using `GPT-4o Transcribe Diarize`, forming a structured

[14] https://www.scenedetect.com/

representation of the video content. Based on the query, candidate answer intervals were proposed from these segments. These candidates were then evaluated and refined using `GPT-5.4`, which considered both the transcript context and sampled key video frames. The pipeline further applied contextual augmentation and consistency-aware span selection, followed by a final refinement step to determine the start and end timestamps for the output.

5.4.2 Results

The VAL results in Table 5 show that the team LAMAR-2 achieved the highest IoU performance across multiple thresholds and the mIOU metric. It outperformed the baseline systems in terms of IoU@0.7, increasing from 48.75 to 77.50. Their strong performance is due to fine-grained modeling of the modalities (audio, visual), explicitly leveraging visual cues rather than relying on textual descriptions of the videos. The team NJUST-KMG achieved the second-best performance by using LLM-driven generation of multi-scale (single- and multi-sentence) semantic priors. The team UNCC achieved a competitive IoU@0.3. However, their performance declined for strict metrics such as IoU@0.5 and IoU@0.7. This may be because their approaches rely on textual candidate generation and on varying visual cues and temporal alignment. The baseline system achieved competitive performance, with an IoU@0.7 score of 48.75, outperforming the best approaches from four participants. We find that approaches that use multimodal grounding and temporal alignment achieve better results than those that rely solely on textual descriptions.

6 Discussion

For the MMR task, we found that the BM25-based approach is a strong method for extracting an initial set of relevant videos from the PubMed Central video transcripts. Hybrid approaches that combine lexical and dense retrieval show improved performance. We observed that video collections sourced from YouTube contain very few relevant videos for professional queries, particularly those involving medical procedures. To address this limitation, we plan to enrich the collection with more specialized videos and, in future work, explore open retrieval from a broader range of video sources. We were also unable to evaluate generative retrieval (Li et al., 2025) and the joint retrieval setup (video and document projection in the unified space) in the current challenge, and we leave these to future work.

We note that the MAG task received limited participation. This is attributed to the complexity of the MAG task, which requires retrieving and synthesizing textual and video evidence to generate coherent and informative answers to medical questions. However, we provide baseline approaches and evaluation metrics to support further research on this important and timely task for the research community. It is to be noted that for the MMR and MAG baseline approaches, we use video information in the form of transcripts, which do not capture temporal dynamics. Despite this limitation, the models still achieve competitive performance and serve as reference points to guide further research on incorporating and enhancing visual information in retrieval and answer generation approaches.

The VAL task results show that the majority of approaches struggle to identify precise video segments that can be considered as answers. However, methods that incorporate fine-grained integration of multimodal information and alignment tend to achieve better performance, as observed in some of the submitted runs. Further analysis of the VAL ground-truth dataset reveals that many annotated segments are relatively long and correspond to many procedural or instructional steps. The developed test questions tend to be more general than focusing on a specific step. In future work, we plan to design more fine-grained video-centric questions and annotate shorter, more precise segments to better evaluate the models' ability to localize exact answer spans.

The results of this shared task highlight several practical challenges in developing medical multimodal question answering. The textual sources play a critical role in providing reliable answers. The video information is difficult to use effectively, even though it provides key information for medical professionals, including procedures and medical education. Specifically for the answer generation task, synthesizing answers from multimodal sources and ensuring that they remain attributable to the corpora is challenging. Moreover, evaluating multimodal medical question answering systems remains challenging because it relies on costly and time-intensive expert annotation. BioACE offers a useful alternative for efficient evaluation and ranking of system outputs and aligns well with nugget-based evaluation for completeness, but less so for correctness.

7 Conclusion

This paper provides an overview of MedGenVidQA 2026 shared task, organized as part of the BioNLP 2026 workshop. We discussed the tasks, datasets, evaluation metrics, and baseline systems. We also provided a summary of the approaches developed by the participating systems. We observed that efficient use of multimodal information sources remains challenging for the related tasks. Although multimodal approaches show promise, they do not consistently outperform strong textual baselines, suggesting that current fusion strategies remain limited. We are optimistic that introducing these tasks, baseline systems, and datasets will foster research toward reliable and trustworthy medical question answering systems for consumers and medical professionals.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH) and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). The contributions of the NIH author(s) are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Souvik Chowdhury and Badal Soni. 2025. R-vqa: A robust visual question answering model. *Knowledge-Based Systems*, 309:112827.

Hilmi Demirhan and Wlodek Zadrozny. 2026. UNCC at MedGenVidQA 2026: Structured Temporal Grounding for Multimodal Medical Video Question Answering. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, USA. Association for Computational Linguistics.

Pratik Vijay Dhaktode, Suhani Bighane, and Anupama Phakatkar. 2026. Varja-Dominators at MedGenVidQA 2026: Hybrid Video and Document Retrieval using PubMedBERT, T5 Query Expansion, and Cross-Encoder Re-Ranking. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, USA. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.

Basil Ebinesar, Keyuan Jiang, Charansai Maddineni, and Ashok Raja. 2026. Pride-Boiler at MedGenVidQA 2026: LLM-Augmented BM25 Retrieval with Corrective Self-Verification for Biomedical Evidence Retrieval. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, USA. Association for Computational Linguistics.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

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. 2024a. 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.

Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2025a. a dataset of medical questions paired with automatically generated answers and evidence-supported references. *Scientific Data*, 12(1):1035.

Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2026. Bioace: An automated framework for biomedical answer and citation evaluations.

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.

Deepak Gupta and Dina Demner-Fushman. 2024. Overview of trec 2024 medical video question answering (medvidqa) track.

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2024b. Overview of TREC 2024 Biomedical Generative Retrieval (BioGen) Track. In *The Thirty-Third Text REtrieval Conference Proceedings (TREC 2024)*, NIST Special Publication. National Institute of Standards and Technology (NIST).

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2025b. Overview of TREC 2025 Biomedical Generative Retrieval (BioGen) Track. In *The Thirty-Fourth Text REtrieval*

Conference Proceedings (TREC 2025), NIST Special Publication. National Institute of Standards and Technology (NIST).

Iryna Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence*, 7:1430984.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2019. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*.

Jinglong Li and Yang Yang. 2026. NJUST-KMG at MedGenVidQA 2026: Cascade Multi-modal Alignment with Gaussian Soft Priors for Medical Visual Answer Localization. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, USA. Association for Computational Linguistics.

Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. 2023. Self-supervised vision-language pretraining for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025. From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*, 43(3):1–62.

Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. 2024. Scemqa: A scientific college entrance level multimodal question answering benchmark. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 109–119.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, and 1 others. 2025. Chartqapro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19123–19151.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

Vaishali Pal, Andrew Yates, Evangelos Kanoulas, and Maarten de Rijke. 2023. Multitabqa: Generating tabular answers for multi-table question answering. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 6322–6334.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, 37:118807–118833.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666.

Watcharitpol Sermsrisuwan, Nopporn Lekuthai, Seksan Yoadsanit, and Titipat Achakulvisut. 2026. LAMAR-2 at MedGenVidQA 2026: Visual Answer Localization in Medical Videos via Multimodal LLM and Context-Augmented Prompting. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, USA. Association for Computational Linguistics.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

Xiaotian Tian and Gulustan Dogan. 2026. Seahawk at MedGenVidQA 2026: LLM Segment-Range Selection for Medical Visual Answer Localization. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, USA. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and 1 others. 2024b. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *arXiv preprint arXiv:2401.04398*.

Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. 2024. Chartinsights: Evaluating multimodal large language models for low-level chart question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12174–12200.

Shweta Yadav, Deepak Gupta, and Dina Demner-Fushman. 2022. Chq-summ: A dataset for consumer healthcare question summarization. *arXiv preprint arXiv:2206.06581*.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024. A simple llm framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Ying Shan, and Limin Wang. 2025. Timelens: Rethinking video temporal grounding with multimodal llms. *arXiv preprint arXiv:2512.14698*.

Liang Zhang, Anwen Hu, Jing Zhang, Shuo Hu, and Qin Jin. 2023. Mpmqa: multimodal question answering on product manuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13958–13966.

Instruction: Write an accurate, engaging, and concise answer for the given question using only the provided search results (some of which might be irrelevant) and cite them properly. Use an unbiased and journalistic tone. Always cite for any factual claim. When citing several search results, use [1][2][3]. Cite at least one document and at most three documents in each sentence. If multiple documents support the sentence, only cite a minimum sufficient subset of the documents.
Documents:
{Document 1}
{Document 2 }
...
{Document 10}
Question : {question}

Figure 4: Prompt used for answer generation with citations.

Instruction: Write an accurate and concise summary of the provided document in relation to the given question. Focus on the information that directly helps answer the query and ignore irrelevant details. Use an unbiased and journalistic tone. Base the summary strictly on the document and do not add external knowledge.
Document: {document}
Question : {question}

Figure 5: Prompt used to summarize the video transcript.

Instruction: You are given a full video and the question: '{query}'. Locate the single continuous video segment in the video that best answers the question. Return exactly one start time and one end time for the most relevant segment. Choose the shortest segment that contains enough evidence to answer the question. If multiple candidate segments exist, return only the best one. If the video does not contain an answer, return 'none'. Output exactly in this format: 'The answer segment is <start time> - <end time> seconds' or 'none'.

Figure 6: Prompt used to locate the visual answer from the medical professional video.