Varja-Dominators at MedGenVidQA 2026: Hybrid Video and Document Retrieval using PubMedBERT, T5 Query Expansion, and Cross-Encoder Re-Ranking

Pratik Dhaktode
Computer Engineering
PICT, Pune
Maharashtra, India
pratik.dhaktode.143@gmail.com

Suhani Bighane
AI and Data Science
PICT, Pune
Maharashtra, India
bighanesuhani@gmail.com

Anupama Phakatkar
Computer Engineering
PICT, Pune
Maharashtra, India
agphakatkar@pict.edu

Abstract

This paper presents our submission to Task A of the MedGenVidQA 2026 shared task (Gupta et al., 2026), which challenges participating systems to simultaneously retrieve relevant PubMed documents and medical videos for 60 consumer health topics (A1–A60). Rather than treating these as independent retrieval problems, we design a unified multi-stage pipeline combining dense and sparse retrieval with cross-encoder re-ranking. For video retrieval, we fine-tune a PubMedBERT bi-encoder on the MedVidQA training set (2,710 samples across 800 unique videos) using BM25-driven hard negative mining. Dense retrieval results are fused with BM25 sparse scores through weighted Reciprocal Rank Fusion (RRF), and queries are expanded using a T5-based doc2query model to bridge the vocabulary gap between consumer-phrased questions and clinical text. A cross-encoder provides final re-ranking. For document retrieval, we query the NCBI PubMed E-utilities API with a progressive keyword fallback strategy. Our system achieves a MAP of 0.4229, Recall@5 of 0.6903, Recall@10 of 0.7226, and NDCG@10 of 0.4971, with complete 60/60 topic coverage for both modalities.

1 Introduction

The proliferation of medical video content on platforms such as YouTube has created both an opportunity and a challenge for health information retrieval. Consumers increasingly seek answers to clinical and procedural questions through video, yet most IR systems index only textual documents such as PubMed abstracts. The MedGenVidQA 2026 shared task (Gupta et al., 2026) directly addresses this gap by requiring systems to perform *dual retrieval*: identifying both relevant PubMed documents and relevant instructional medical videos for consumer health topics.

Task definition. Task A requires each system to submit, for each of 60 test topics (A1–A60), a ranked list of up to 10 PubMed document IDs (PMIDs) and up to 10 video IDs. The test set comprises 155 question-answer samples spanning 50 unique YouTube videos. The input is a natural-language consumer health question; the output is two ranked lists of at most 10 items each.

Challenges. The primary challenges are: (1) *vocabulary mismatch*—consumer health questions use colloquial terminology differing from clinical language in abstracts and transcripts; (2) *transcription noise*—auto-generated subtitles introduce errors degrading text-matching quality; and (3) *corpus heterogeneity*—relevant content spans YouTube videos and MEDLINE abstracts, requiring distinct retrieval strategies per modality.

Approach. We address these challenges through a pipeline combining a domain-specific PubMedBERT bi-encoder fine-tuned with BM25-driven hard negative mining, T5-based query expansion, weighted RRF fusion of dense and sparse retrieval, and a cross-encoder re-ranking stage. PubMed document retrieval uses a multi-pass NCBI API strategy with exponential backoff.

2 Related Work

Dense Retrieval and Bi-Encoders. Dense retrieval using pre-trained language models has substantially advanced open-domain IR. Karpukhin et al. (2020) showed that bi-encoders can match or outperform BM25 for open-domain QA. PubMedBERT (Gu et al., 2021) provides strong initialisation for biomedical dense retrieval. Hard negative mining (Xiong et al., 2021) further improves bi-encoder precision by exposing the model to false-positive candidates.

Hybrid Retrieval and Rank Fusion. Hybrid systems combining BM25 and dense retrievers con-

Proceedings of the BioNLP 2026 (Shared Tasks), pages 243–247

sistently outperform either alone (Ma et al., 2022). Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) is a robust, score-free fusion method widely adopted for its effectiveness across diverse retrieval settings.

Re-ranking with Cross-Encoders. Cross-encoders jointly encode query-passage pairs and provide higher precision than bi-encoders at the cost of inference speed (Nogueira and Cho, 2019). MS-MARCO-trained MiniLM cross-encoders offer a practical speed-accuracy trade-off.

Medical Video Retrieval. Gupta and Demner-Fushman (2023) introduced MedVidQA—consumer health instructional videos with timestamped QA annotations—and established transcript-based retrieval as the dominant paradigm. Subsequent MedVidQA shared task systems (Gupta et al., 2026) explored BM25 baselines, dense bi-encoders, and cross-modal approaches. Our work extends transcript-based retrieval with domain-specific fine-tuning, multi-stage fusion, and a dedicated document retrieval branch.

3 Data

3.1 Dataset

The corpus is drawn from **MedVidQA** (Gupta and Demner-Fushman, 2023), consumer health instructional YouTube videos with timestamped QA annotations. Table 1 summarises the statistics.

Split	Samples	Unique Videos
Train	2,710	800
Validation	145	49
Test	155	50

Table 1: MedVidQA dataset statistics. Each sample contains a question and timestamped answer boundaries used to identify positive transcript segments.

3.2 Transcript Corpus Construction

Video transcripts were downloaded using `yt-dlp`[1] and parsed from VTT format using `webvtt-py`.[2] The corpus covers **833 unique videos** with **217,153 raw transcript lines**. For videos without automatic subtitles, OpenAI Whisper ASR was used as a fallback. After temporal chunking (Section 4.1), the corpus yields **32,489 indexed segments**.

[1] https://github.com/yt-dlp/yt-dlp

[2] https://github.com/glut23/webvtt-py

4 System Description

Our system has four sequential stages: (1) temporal chunking, (2) PubMedBERT bi-encoder fine-tuning, (3) hybrid retrieval with RRF fusion and cross-encoder re-ranking, and (4) PubMed document retrieval. Figure 1 shows the architecture.

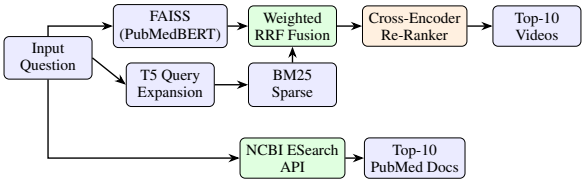

Figure 1: Hybrid retrieval pipeline: dense (FAISS/PubMedBERT) and sparse (BM25/T5-expanded) results fused via weighted RRF, re-ranked by a cross-encoder for videos; NCBI ESearch for documents.

4.1 Temporal Chunking

Given transcript $T = \{(t_i, d_i, s_i)\}_{i=1}^{N}$ (start time, duration, spoken text), we form overlapping chunks with window W=30 s, stride S=10 s:

$$C_k = \left\{ s_i \mid (t_i + d_i) > kS \wedge t_i < kS + W \right\} \quad (1)$$

Each chunk stores its parent `video_id` for deduplication. This yields **32,489 chunks**, each a 768-dimensional dense vector.

4.2 Bi-Encoder Fine-Tuning

Base model. We build on **PubMedBERT** (Gu et al., 2021),[3] pre-trained on full PubMed abstracts for strong biomedical domain coverage.

Training objective. We fine-tune as a bi-encoder using `MultipleNegativesRankingLoss` (Reimers and Gurevych, 2019). Positive passages overlap the ground-truth answer span by $\geq$10 s. Hard negatives are the highest BM25-scoring non-overlapping window in the same video, forcing the model to distinguish passages sharing medical terminology but differing in procedural content.

Hyperparameters. Batch size 8, 4 epochs, 100 warmup steps, max sequence length 256 tokens. A `TripletEvaluator` runs on the validation split twice per epoch.

[3] huggingface.co/microsoft/BiomedNLP-PubMedBERT

Embedding index. All 32,489 chunks are encoded with mean pooling and L2 normalisation (32,489×768 matrix) and indexed in FAISS `IndexFlatIP` (Johnson et al., 2019).

4.3 Hybrid Retrieval, Fusion, and Re-Ranking

4.3.1 Query Expansion via T5

Each query is expanded using a T5 doc2query model[4] generating three passage-style expansions (top-k=10 sampling, max length 64):

$$q_{\text{exp}} = q \oplus [\text{T5-expand}(q)]_{1..3} \tag{2}$$

q_{exp} feeds BM25; the original q feeds FAISS to preserve embedding alignment.

4.3.2 Weighted Reciprocal Rank Fusion

FAISS and BM25Okapi each return top K=200 chunks. Fusion uses weighted RRF (Cormack et al., 2009):

$$\text{RRF}(c) = \frac{w_d}{k + r_d(c)} + \frac{w_s}{k + r_s(c)} \tag{3}$$

with k=60, w_d=0.75, w_s=0.25.

4.3.3 Cross-Encoder Re-Ranking

Top-200 RRF candidates are re-ranked by a MiniLM cross-encoder[5] (Nogueira and Cho, 2019). The final output is the top-10 unique `video_id` values (highest score per video across all chunks).

4.4 PubMed Document Retrieval

We use NCBI ESearch with a sequential fallback chain:

1. Full question text.
2. Top-10 non-stop-word tokens.
3. Top-6 tokens.
4. Top-3 content-bearing tokens.

Failed requests retry with exponential backoff (5 retries, base 0.4 s). PMIDs are scored by rank:

$$\text{score}(r) = 0.5 + 0.49 \cdot e^{-0.3r} \tag{4}$$

mapping rank 0 to ≈0.990 and rank 9 to ≈0.537. Results are cached incrementally.

[4] `doc2query/msmarco-t5-base-v1`

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

5 Experimental Setup

5.1 Datasets and Preprocessing

Training uses the MedVidQA train split (2,710 samples, 800 videos). The validation split (145/49) monitors bi-encoder training only. The test split (155/50) is used exclusively for final evaluation.

5.2 Baselines

We submitted one run (*hybrid_rerank_exp_v3*); no external baseline comparisons are available as no other system runs on this test set are public at time of writing.

5.3 Evaluation Metrics

Standard TREC metrics via `pytrec_eval`: MAP, Recall@5/10, Precision@5/10, and NDCG@5/10.

5.4 Implementation Details

Python 3.10, PyTorch, HuggingFace Transformers, Sentence-Transformers, FAISS. BM25 via `rank_bm25`.[6] Training on a single consumer GPU; inference on CPU or GPU.

6 Results and Discussion

6.1 Quantitative Results

Metric	Score
MAP	0.4229
Recall@5	0.6903
Recall@10	0.7226
Precision@5	0.1381
Precision@10	0.0723
NDCG@5	0.4857
NDCG@10	0.4971
Topic Coverage	60 / 60
Avg. Docs/Topic	7.6
Avg. Videos/Topic	6.6

Table 2: Evaluation results on the MedGenVidQA 2026 Task A test set (*hybrid_rerank_exp_v3* run).

Our system achieves full 60/60 topic coverage for both retrieval modalities. Recall@10 of 0.7226 indicates our top-10 list covers ≈72.3% of all relevant videos per topic. The close agreement between NDCG@5 (0.4857) and NDCG@10 (0.4971) confirms that most relevant videos are ranked in positions 1–5, with only marginal gain from positions 6–10.

The MAP of 0.4229 reflects strong average precision overall. However, 44 of 155 test queries

[6] https://github.com/dorianbrown/rank_bm25

received Recall@10 = 0.0, representing procedure-specific or rare clinical topics for which the transcript corpus contained no closely matching content.

Precision@5 (0.1381) and Precision@10 (0.0723) reveal a recall–precision gap: the system retrieves most relevant videos but does not consistently rank them at the very top positions. We attribute this partly to the cross-encoder being trained on general web passages (MS-MARCO) rather than biomedical text, limiting its ability to make fine-grained relevance distinctions in the medical domain.

6.2 Analysis

Corpus coverage. Our 833-video corpus (32,489 temporal chunks) substantially exceeds the 50 test videos, providing broad candidate coverage and reducing the risk of corpus gaps.

Hard negative mining. BM25-driven hard negatives expose the bi-encoder to passages that share medical terminology with the correct answer but differ in procedural or temporal context—critical for distinguishing adjacent instructional segments within the same recording session.

Query expansion. T5-based expansion bridges colloquial consumer questions (e.g., *"How to fix a crick in my neck?"*) and the clinical register of video transcripts, improving BM25 recall through synonym-level coverage.

PubMed retrieval reliability. The multi-pass fallback strategy was essential: full-question queries frequently failed due to PubMed's query length constraints and low-frequency medical terminology. Incremental caching ensured 60/60 topic coverage without repeated API calls.

Failure mode analysis. Inspection of the 44 zero-recall queries reveals two dominant failure patterns: (1) *topic specificity*—questions about rare clinical procedures for which the corpus contains only peripherally related videos; and (2) *video duplication*—the same popular medical education video dominating retrieval across multiple topics, reducing per-topic diversity.

6.3 Future Scope

Our current system is purely text-based; several directions offer clear paths to improvement:

- **Multi-modal retrieval.** Incorporating visual keyframes, on-screen text (OCR), and audio prosody features alongside transcript embeddings would allow the system to handle visually informative but verbally sparse videos that currently receive low retrieval scores.
- **Domain-adapted cross-encoder.** Fine-tuning the re-ranker on biomedical QA corpora (e.g., MedVidQA training pairs) rather than relying on the general MS-MARCO model would directly address the observed recall–precision gap and improve precision at early ranks.
- **Local PubMed neural indexing.** Replacing the live NCBI API with a local dense index over PubMed abstracts (e.g., using BioLinkBERT or SPECTER) would enable faster retrieval, richer ranking signals, and rate-limit-free operation.
- **Video metadata integration.** Including YouTube video titles, channel descriptions, and view counts as auxiliary signals could improve retrieval for topics where transcript content alone is ambiguous.
- **Ablation study.** A controlled ablation varying individual components (bi-encoder vs. BM25 only, with/without T5 expansion, with/without cross-encoder) would quantify the contribution of each stage and guide future architecture decisions.

7 Conclusion

We presented the **Varja-Dominators** system for MedGenVidQA 2026 Task A: a fine-tuned PubMedBERT bi-encoder with BM25-driven hard negatives, T5 query expansion, weighted RRF fusion over 32,489 temporal chunks, and cross-encoder re-ranking. PubMed document retrieval uses a multi-pass NCBI API strategy, achieving 60/60 topic coverage. Our system achieves MAP 0.4229, NDCG@10 0.4971, and Recall@10 0.7226. Future work includes multi-modal retrieval with video frames and metadata, local neural PubMed indexing, and domain-adapted cross-encoders for improved precision.

Limitations

This system relies entirely on spoken transcript text; no visual or audio features are used. PubMed

retrieval is live via a public API, introducing rate-limiting latency without neural document ranking. We submitted one run without a controlled ablation study, so individual component contributions are not empirically isolated.

Ethics Statement

All data is publicly available: MedVidQA from the competition organisers and PubMed abstracts via the NCBI open API. No personally identifiable information is collected or processed. This system is for research purposes only and must not substitute for professional medical advice.

Acknowledgments

We thank the MedGenVidQA 2026 organisers for providing the dataset and evaluation infrastructure.

References

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. ACM.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

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. 2023. A dataset for medical instructional video retrieval and question answering. In *Proceedings of the 22nd Workshop on Biomedical Language Processing (BioNLP)*, pages 230–240. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, and Xueqi Cheng. 2022. Hybrid retrieval for open-domain question answering over long-form documents. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2740–2745. ACM.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. In *arXiv preprint arXiv:1901.04085*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jiawei Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive estimation for dense text retrieval. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.