Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ31704

Bridging the Gap in Health Video Retrieval: A Multi-Turn Dialogue Benchmark and Two-Stage Framework

Chengzhen Wu

Wuxi Dipont School of Arts and Science, Wuxi, China
26306677@qq.com

Abstract. The increasing availability of health-related videos offers valuable opportunities for clinical training, patient rehabilitation, and health education. However, existing retrieval methods in this domain remain limited, particularly when users need to refine their information needs through iterative interactions. To address this gap, we introduce the task of interactive multi-turn semantic retrieval for health videos, which enables users to progressively narrow and specify their queries in dialogue with an intelligent retrieval system. To support this task, we construct a new dataset, named MHVRC (Multi-Turn Health Video Retrieval Corpus), by combining VideoChat-Flash–generated video descriptions with DeepSeek-based refinement of coarse queries into realistic follow-up questions. A user study demonstrates that the resulting multi-turn queries better capture fine-grained procedural semantics than existing benchmarks such as MedVidQA. Based on MHVRC, we propose DATR (Dialogue-Aware Two-Stage Retrieval), a two-stage retrieval framework. The first stage adopts a dual-encoder with sparse frame sampling and CLIP-style contrastive learning to achieve efficient coarse retrieval. The second stage re-ranks the top candidates by fusing multi-turn queries and applying a lightweight cross-encoder scoring module, which captures contextual intent shifts and procedural details that single-turn retrieval often misses. Extensive experiments show that DATR achieves state-of-the-art performance on both quantitative metrics and qualitative user evaluations. This work not only establishes the first benchmark for interactive multi-turn health video retrieval but also provides a scalable framework for advancing intelligent medical information access.

Keywords: Interactive Video Retrieval, Multi-Turn Dialogue, Health Informatics, Medical Video Benchmark

1. Introduction

1.1. Background and motivation

With the rapid advancement of artificial intelligence, especially breakthroughs in natural language processing and multimodal semantic understanding, video retrieval has achieved remarkable progress in both accuracy and efficiency [1]. Models such as CLIP [2] and CLIPBERT [3] demonstrate that large-scale contrastive training and sparse frame sampling can enable robust text–video alignment at scale. Nevertheless, in the health domain, retrieval techniques remain

DOI: 10.54254/2755-2721/2026.TJ31704

comparatively underdeveloped despite the increasing demand for personalized, accessible, and trustworthy medical and rehabilitation information [4].

Multi-turn interactive health video retrieval seeks to address this gap by allowing users to refine their search intent iteratively through human–AI interaction [5], thus enabling progressively narrower and more precise localization of relevant video content. Such a paradigm offers considerable societal impact: (1) clinicians and rehabilitation therapists can quickly identify demonstrations suitable for patient-specific conditions; (2) patients can receive guidance that balances difficulty, clarity, and safety; (3) educators can leverage precise retrieval to enhance medical training and remote rehabilitation services [6]. As a result, this task not only resonates with the trend toward intelligent healthcare but also holds significant academic and practical value in improving the efficiency and quality of rehabilitation services.

1.2. Research gap

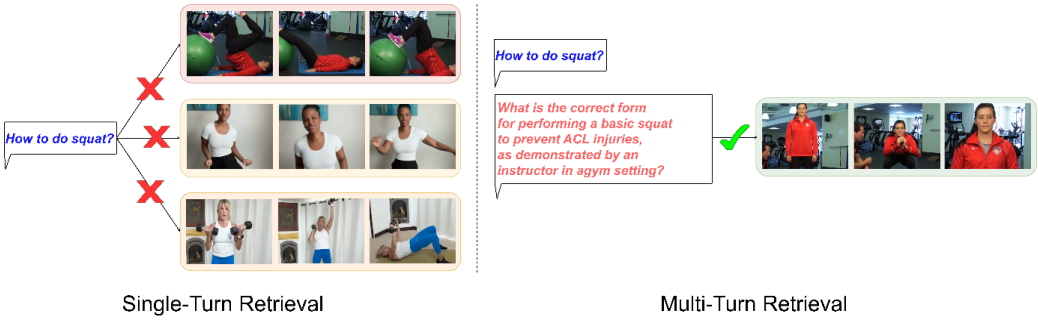

Figure 1. Single-turn vs. multi-turn health video retrieval

Despite the promise of multi-turn interaction, research in this direction remains limited. General-domain video retrieval has progressed rapidly with large-scale datasets such as MSR-VTT [7], YouCook2 [8], and TVR [9], and multi-turn conversational search has been shown to enhance intent modeling in open-domain video QA and retrieval [5]. However, these advances have not been systematically extended to the health domain. Existing medical video retrieval work focuses mainly on indexing, classification, and recommendation [10], typically relying on single-turn queries that overlook users' dynamic, context-dependent needs. Multi-turn retrieval for health videos is particularly challenging, requiring models to capture complex clinical and rehabilitation semantics [11], process multimodal cues such as narration and procedural motion [12], and adapt to personalized user goals. As illustrated in Figure 1, single-turn queries (e.g., "How to do squat?") often yield loosely relevant videos, whereas multi-turn refinement (e.g., "What is the correct squat form to prevent ACL injuries, demonstrated in a gym setting?") enables more accurate and clinically meaningful results. This gap underscores the need for dedicated solutions addressing both dataset scarcity and model design challenges.

1.3. Our approach

The lack of suitable datasets remains a key bottleneck for health video retrieval. MedVidQA provides 3,010 expert-verified, timestamped question–answer pairs and represents an important benchmark [4]; however, its single-turn queries are often coarse and insufficient for locating fine-grained procedural details. Queries such as "How to perform Epley maneuver for vertigo?" lack the specificity required in real scenarios, where users may seek posture, hand placement, or

DOI: 10.54254/2755-2721/2026.TJ31704

contraindications. To address this gap, we construct MHVRC, a dataset explicitly designed for multi-turn retrieval. Using VideoChat-Flash to generate detailed procedural descriptions [12] and DeepSeek to refine coarse questions into realistic follow-ups [5], our pipeline produces query sequences that better reflect natural information-seeking behavior. A user study further confirms that these refined multi-turn queries are more specific and aligned with user needs than MedVidQA's single-turn annotations.

Building on MHVRC, we introduce DATR, a dialogue-aware two-stage retrieval framework. Stage I uses a CLIP-style dual encoder [1,2] with Transformer-based video modeling [13] and sparse frame sampling [14] to align videos and queries efficiently. Stage II re-ranks the top-K results using multi-turn query fusion [5] and a lightweight cross-encoder [8], capturing procedural nuances often missed in coarse retrieval. This design achieves strong efficiency–accuracy trade-offs and delivers state-of-the-art performance in both benchmark evaluations and user studies.

1.4. Contributions

The main contributions of this work are summarized as follows:

(1) Task formulation. We introduce the new task of multi-turn interactive health video retrieval, which enables iterative refinement of user intent through dialogue, providing a more precise and user-aligned search paradigm compared with traditional single-turn methods.

(2) Dataset construction. We build a dedicated dataset, named MHVRC (Multi-Turn Health Video Retrieval Corpus), to support this task by generating procedure-aware video descriptions and realistic follow-up queries. Compared to existing resources such as MedVidQA, our dataset better captures fine-grained clinical semantics and evolving information needs.

(3) Framework design. We propose a two-stage retrieval framework, named DATR (Dialogue-Aware Two-Stage Retrieval), tailored to the health domain. The first stage performs efficient dual-encoder coarse retrieval with sparse frame sampling, while the second stage conducts dialogue-aware re-ranking over top candidates to achieve fine-grained semantic matching.

(4) Empirical validation. Extensive experiments, including quantitative evaluation, ablation studies, and user studies, demonstrate the superiority of our framework and confirm the effectiveness of the constructed dataset for advancing multi-turn health video retrieval.

2. Related work

2.1. Video retrieval

Video retrieval aligns text queries with relevant video content. Early metadata or handcrafted features lacked semantic depth, while deep learning enabled end-to-end cross-modal representations. CLIP [2] and CLIP4Clip [1] established large-scale contrastive pretraining for text–video alignment, followed by models such as Frozen-in-Time [13], ViViT [15], and TimeSformer [16] that improved temporal modeling. Efficiency-oriented designs (e.g., CLIPBERT [3]) and hierarchical fusion architectures further enhanced retrieval quality. Large vision–language models [17-19] extended capabilities to long-form reasoning. Despite progress, existing systems almost universally follow single-turn retrieval, failing to model evolving user intent.

2.2. Multi-turn video retrieval

Multi-turn retrieval introduces dialogue-driven iterative refinement, allowing users to narrow or adjust queries across turns. Research draws from conversational IR through history modeling and

DOI: 10.54254/2755-2721/2026.TJ31704

intent tracking, supported by datasets like Dial2Vid [20] and TVR-QA [21]. Models such as CASe [22] and memory-augmented transformers capture contextual information, while recent work explores multi-turn prediction and benchmarking. However, current efforts largely remain in open-domain settings. No existing system or dataset addresses multi-turn retrieval in specialized domains such as healthcare, leaving a major gap.

2.3. Health video understanding

Health video understanding supports applications in clinical training, patient education, and rehabilitation. Datasets such as MedVidQA [4] enable medical video QA, while misinformation-focused tasks (e.g., TREC Health [14]) highlight credibility needs. Work in surgical workflow analysis and rehabilitation scoring demonstrates domain interest, and retrieval-augmented approaches [23] provide stronger medical reasoning. Yet most methods focus on classification, coarse retrieval, or step recognition. Fine-grained text–video matching and multi-turn interactive retrieval remain unexplored, and no dataset currently supports iterative health video search, revealing a clear research opportunity.

3. Dataset

To support this task, we construct a new dataset, named MHVRC (Multi-Turn Health Video Retrieval Corpus), by combining VideoChat-Flash–generated video descriptions with DeepSeek-based refinement of coarse queries into realistic follow-up questions.

3.1. Limitations of existing benchmarks

3.1.1. MedVidQA as the current standard

MedVidQA [4] is currently the most influential dataset, consisting of 3,010 manually created health-related questions aligned with timestamped video answers from 899 medical instructional videos (over 95 hours). The questions were curated by medical experts to ensure clinical reliability.

3.1.2. Limitations of single-turn and coarse queries

Despite its strengths, MedVidQA has notable limitations. Queries in MedVidQA are restricted to a single-turn setting and are often phrased at a coarse level (e.g., "How to perform Epley maneuver for vertigo?"). Such annotations do not reflect the progressive refinement of user needs in real scenarios, where details such as posture, hand placement, or contraindications are crucial.

3.1.3. Implications for multi-turn retrieval

Although MedVidQA is useful for span-localization tasks, it is insufficient for evaluating systems that must support multi-turn, fine-grained retrieval. This limitation motivates the development of a new dataset designed to address iterative query refinement and procedural detail.

DOI: 10.54254/2755-2721/2026.TJ31704

3.2. Construction pipeline

3.2.1. Pipeline overview

As illustrated in Figure 2, the construction pipeline begins with a coarse user-style query such as "How to do squat?". To enrich the semantic grounding, the associated instructional video is first processed with VideoChat-Flash, which generates detailed procedural descriptions capturing body posture, instructor guidance, and movement context. These outputs provide a fine-grained semantic representation of the video content. Building on this, DeepSeek refines the initial coarse query into a more specific and clinically relevant follow-up, for example: "What is the correct form for performing a basic squat to prevent ACL injuries, as demonstrated by an instructor in a gym setting?". Compared to the original input, the refined query exhibits higher alignment with the video semantics, greater clinical relevance, and closer resemblance to real-world user information needs. Together, these two complementary steps—visual grounding via VideoChat-Flash and contextual refinement via DeepSeek—construct a dataset that explicitly encodes the progressive refinement trajectory inherent in multi-turn retrieval.

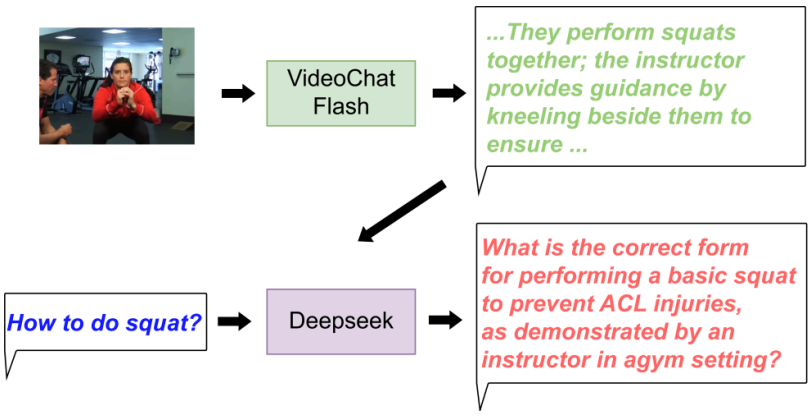

Figure 2. Construction pipeline for multi-turn retrieval dataset

3.2.2. VideoChat-Flash for procedural descriptions

• Motivation: VideoChat-Flash [12] is a recent vision–language model designed for multi-frame video understanding. Unlike traditional clip-level encoders, it integrates temporal reasoning with prompt-driven text generation, which makes it particularly well-suited for long, instructional medical videos that often span multiple procedural steps.

• Prompt design: For VideoChat-Flash, the prompt was designed to fully exploit its ability for long-form video understanding in the medical domain. Instead of producing short, generic captions, we positioned the model as a medical video analyst instructed to generate comprehensive, procedural descriptions. The prompt emphasized body posture, hand placement, medical actions, and equipment usage, ensuring the outputs capture clinically relevant details. Technically, we applied deterministic decoding (temperature=0.0, no sampling) to avoid hallucinations and allocated a high token budget (256 tokens) with up to 1,024 frames for temporal coverage. For example, given a video of the Epley maneuver, the model produced: "A man wearing a black polo shirt is lying supine, while another man in a checkered shirt places both hands around the patient's neck and gently rotates the head to the side." Such outputs provide semantic bridges that are richer than

DOI: 10.54254/2755-2721/2026.TJ31704

MedVidQA annotations, grounding retrieval in fine-grained procedural content essential for multi-turn interaction.

• Example output: For the Epley maneuver case, the model generated: "A man wearing a black polo shirt is lying supine, while another man in a checkered shirt places both hands around the patient's neck and gently rotates the head to the side. He then stabilizes the head with one hand and supports the chin with the other, while maintaining controlled movements." Such rich procedural detail significantly enhances semantic grounding for downstream retrieval.

3.2.3. DeepSeek for query refinement

• Motivation: DeepSeek [9] is a state-of-the-art large language model specializing in contextual refinement and task-oriented dialogue generation. Its ability to reformulate coarse input into structured, context-aware queries makes it ideal for simulating multi-turn retrieval scenarios.

• Prompt design: For DeepSeek, we designed prompts to simulate realistic user reformulation of queries during multi-turn retrieval. The system role was set as a medical video retrieval user, with explicit instructions to refine coarse initial queries into more specific, video-aligned ones. The model conditions on both the original question and the detailed VideoChat-Flash description, ensuring refinements are grounded in procedural evidence rather than generic speculation. The prompt emphasized accuracy and specificity: "The initial query is usually too rough. Please refine it into a better, more specific question that matches the video content." With controlled decoding (temperature=1.0, max_tokens=80), the model generated queries such as "How is the Epley maneuver performed on a supine patient, with specific hand placement on the head and neck?". Compared with the original query, this refined version is semantically closer to the video, procedurally accurate, and more aligned with clinical information needs—explicitly modeling how users evolve queries in real-world retrieval.

• Example refinement: For the Epley case, DeepSeek produced: "How is the Epley maneuver performed on a supine patient, with specific hand placement on the head and neck?". This refined query is procedurally accurate, closer to real clinical inquiry, and demonstrates how users naturally adapt their questions across multiple search iterations.

By combining visual grounding (VideoChat-Flash) and query evolution (DeepSeek), our pipeline creates dataset annotations that explicitly encode the dialogical nature of medical information seeking. This not only simulates authentic user–AI interactions but also establishes a reproducible framework for dataset expansion in other domains beyond healthcare.

3.3. Dataset statistics and advantages

3.3.1. Composition and scale

The final dataset consists of triplets:

• Initial coarse query (single-turn MedVidQA-style question)

• VideoChat-Flash description (procedural text aligned with video content)

• DeepSeek-refined query (realistic multi-turn follow-up question)

• In scale, our dataset is comparable to MedVidQA, with approximately 3,000 queries derived from around 900 instructional medical videos spanning over 90 hours of content. The videos cover a broad range of domains including:

• First aid and emergency response (e.g., CPR, Heimlich maneuver)

• Musculoskeletal rehabilitation (e.g., knee flexion exercises, postural corrections)

DOI: 10.54254/2755-2721/2026.TJ31704

• Cardiovascular care and monitoring (e.g., ECG placement, blood pressure measurement)
• Patient self-management and daily care (e.g., insulin injection, wound dressing).
• Each video averages over 300 seconds, allowing multiple procedural steps to be captured within a single segment. The refined queries typically extend by 30–50% more tokens compared to the initial queries, reflecting their added granularity.

3.3.2. Advantages over MedVidQA

Compared with MedVidQA, our dataset introduces multi-turn refinement and procedural grounding, thereby addressing its limitations in query diversity and interaction modeling.

• Finer granularity: Our dataset captures detailed procedural nuances, such as specific hand placement or body posture, that are absent in MedVidQA.
• Multi-turn interactivity: Unlike single-turn annotations, our dataset explicitly encodes query evolution, enabling research on dialogue-aware retrieval.
• Higher realism: A controlled user study demonstrated that participants rated our refined queries as significantly more aligned with real needs compared to MedVidQA annotations.
• Procedural grounding: Descriptions provide explicit links between medical actions and corresponding queries, offering a strong foundation for multimodal alignment tasks.

3.3.3. Applications and broader impact

Beyond serving as a benchmark, our dataset provides broad applications in clinical training, rehabilitation, and health education, highlighting its potential impact in real-world healthcare scenarios.

• Medical Video QA: Enabling systems where answers must adapt dynamically to refined queries.
• Cross-modal representation learning: Providing rich supervision for aligning video and procedural text.
• Dialogue-based health assistants: Facilitating real-world patient-facing tools that can handle context-dependent and evolving medical information needs.
• Generalizable pipeline: The dataset creation framework itself can be adapted to other domains (e.g., legal instructional videos, engineering tutorials), making it a scalable methodology for building multi-turn retrieval benchmarks.

4. Methodology

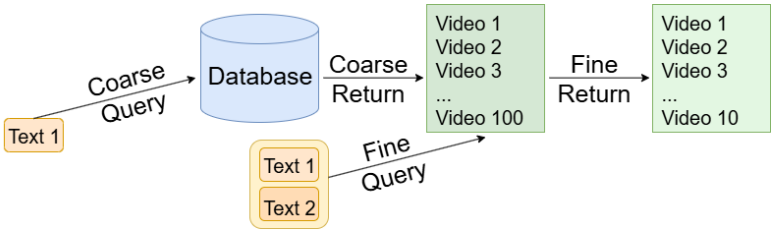

Figure 3. Overall pipeline of the DATR framework

DOI: 10.54254/2755-2721/2026.TJ31704

4.1. Problem definition

The task of multi-turn health video retrieval requires returning instructional videos that match an evolving user information need, expressed through a sequence of natural language queries. Unlike conventional single-turn retrieval, which treats each query independently, our problem explicitly models iterative refinement: users often start with vague queries (e.g., "How to do squat?") and progressively refine them into specific forms (e.g., "What is the correct form for performing a basic squat to prevent ACL injuries, as demonstrated by an instructor in a gym setting?").

Formally, let the video collection be:

$$V = \{v_1, v_2, ..., v_N\}, \tag{1}$$

where each video v_i is an instructional health clip containing multimodal information. A user poses a query sequence:

$$Q = \{q_1, q_2, ..., q_T\}, \tag{2}$$

where q_1 is a coarse initial query and each subsequent $q_t (t > 1)$ is a refinement. The objective is to produce a ranked subset of videos:

$$R_t \subseteq V, \tag{3}$$

that best satisfies the progressively refined intent at turn t.

4.2. Framework overview

To address these challenges, we introduce DATR (Dialogue-Aware Two-Stage Retrieval), illustrated in Figure 3 and Figure 4.

• Stage I (Coarse Retrieval): The initial query q_1 retrieves the top- K candidates from the database V.

• Stage II (Fine Re-Ranking): Both q_1 and refined query q_2 are fused to re-rank R_1, yielding the top- M results.

Formally:

$$R_1 = TopK(sim(f_T(q_1), f_V(v_i)))_{i=1}^{N}, \tag{4}$$

$$R_2 = TopM(sim(f_F(q_1, q_2), f_V(v)))_{v \in R_1}. \tag{5}$$

Here, $f_T(\cdot)$ and $f_V(\cdot)$ are text and video encoders, $f_F(\cdot)$ is the fusion encoder, and $\mathrm{sim}(\cdot)$ denotes cosine similarity.

DOI: 10.54254/2755-2721/2026.TJ31704

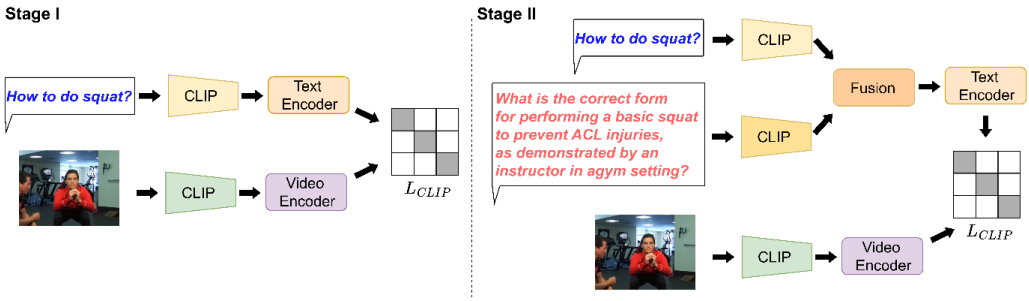

Figure 4. Architecture of the two-stage retrieval process

4.3. Stage I: coarse retrieval

4.3.1. Motivation

The primary goal of Stage I is to provide a scalable yet reliable coarse retrieval mechanism. In large video databases, directly applying cross-encoder–style matching for every query is computationally prohibitive. Instead, we employ a dual-encoder architecture that independently encodes queries and videos into a shared embedding space, enabling efficient similarity search. By ensuring that truly relevant videos are included within the top- K candidates, Stage I lays the foundation for more precise re-ranking in Stage II.

4.3.2. Text encoder

The coarse query q_1 is first tokenized and processed by a CLIP-based transformer encoder trained on large-scale vision–language corpora. This yields an embedding:

$$z_T = f_T(q_1) \in \mathrm{R}^{512}, \tag{6}$$

which is subsequently refined using a lightweight feed-forward adapter with non-linearity and dropout regularization:

$$\hat{z}_T = W_2\sigma(W_1 z_T + b_1) + b_2, \tag{7}$$

where $\sigma(\cdot)$ denotes ReLU activation. This step provides additional domain adaptation capacity while maintaining computational efficiency.

4.3.3. Video encoder

Each video v is uniformly sampled into 32 frames to balance temporal coverage and efficiency. The sampled frame embeddings $X \in \mathrm{R}^{32\times d}$ are passed through an $L=6$ layer Transformer encoder with hidden size $d=512$ and $h=8$ self-attention heads.

Self-attention:

$$Q = XW_Q, K = XW_K, V = XW_V \tag{8}$$

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{9}$$

Feed-forward:

DOI: 10.54254/2755-2721/2026.TJ31704

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2, \tag{10}$$

To effectively handle long-range dependencies while avoiding quadratic cost in sequence length, we apply 1D convolutional downsampling after each Transformer block:

$$y_t = \sum_{i=0}^{k-1} W_i \cdot x_{s \cdot t+I}, y \in \mathrm{R}^{\frac{T}{s} \times d} \tag{11}$$

with kernel size $k = 3$ and stride $s = 2$. After L layers, temporal length is reduced by 2^L, producing a compact yet semantically rich embedding. The final pooled video representation is:

$$z_V = f_V(v) \in \mathrm{R}^{512}. \tag{12}$$

4.3.4. Training objective

Stage I is trained with bidirectional CLIP contrastive loss:

$$\begin{aligned} \mathscr{L}_{Stage1} = -\frac{1}{B} \sum_{i=1}^{B} [log \frac{exp(sim(z_T^i, z_V^i)/\tau)}{\sum_{j=1}^{B} exp(sim(z_T^i, z_V^j)/\tau)} \\ + log \frac{exp(sim(z_V^i, z_T^i)/\tau)}{\sum_{j=1}^{B} exp(sim(z_V^i, z_T^j)/\tau)}] \end{aligned}, \tag{13}$$

where B is the batch size and τ is a learnable temperature. This symmetric loss enforces alignment in both text-to-video and video-to-text directions, which is crucial in interactive retrieval scenarios where user feedback may flow in either modality.

4.4. Stage II: fine re-ranking

4.4.1. Query fusion

The embeddings of q_1 and q_2 are combined using both additive and multiplicative interactions:

$$z_F = MLP([z_{T1} + z_{T2}; z_{T1} \bigodot z_{T2}]). \tag{14}$$

This fusion strategy has two advantages: (1) the additive term emphasizes semantic overlap between queries, ensuring continuity, and (2) the multiplicative term highlights refinements and conditional nuances, e.g., posture or environment constraints. The MLP layer adjusts dimensionality and introduces non-linearity for robust representation.

4.4.2. Re-ranking

Each candidate $v \in R_1$ is scored as:

$$s(v) = sim(z_F, f_V(v)), v \in R_1. \tag{15}$$

The top- M results are selected as the final ranking:

$$R_2 = TopM(s(v)), v \in R_1. \tag{16}$$

DOI: 10.54254/2755-2721/2026.TJ31704

4.4.3. Training objective

Stage II adopts the same bidirectional contrastive loss as Stage I, ensuring consistency. Importantly, restricting the computation to the 100 candidates from Stage I reduces complexity by orders of magnitude while improving precision through multi-turn context modeling.

5. Experiments

5.1. Evaluation protocol

We evaluate our framework on the proposed MHVRC dataset using standard text-to-video retrieval metrics. Specifically, we report Recall@ K at multiple cutoff levels ($K = 1,5,10,50,100$), Median Rank (MedR), and Mean Rank (MeanR). Recall@ K measures the fraction of queries for which the ground-truth video is ranked within the top- K retrieved results, while MedR and MeanR provide robustness to outliers by summarizing the rank distribution.

Following prior work in video retrieval [1], we split MHVRC into training and testing subsets with an 8:2 ratio, ensuring that videos from the same source do not overlap across splits.

5.2. Comparison

Table 1. Text-to-video retrieval results on MHVRC

Model	R@1 ↑	R@5 ↑	R@10 ↑	R@50 ↑	R@100 ↑	MedR ↓	MeanR ↓
CLIP4Clip [1]	13.8	34.2	46.3	77.9	86.0	13	37.2
Frozen-in-Time [13]	12.6	32.5	44.1	75.6	84.2	14	39.8
CLIPBERT [2]	14.7	36.1	48.5	78.3	87.1	12	35.0
HERO [3]	15.2	37.5	49.8	80.0	88.4	11	32.7
DATR (ours)	19.5	44.8	57.1	85.6	92.7	7	25.6

We conduct a systematic comparison of our proposed DATR framework with several state-of-the-art text–video retrieval models, including CLIP4Clip [1], Frozen-in-Time [13], CLIPBERT [2], and HERO [3]. These baselines represent distinct design philosophies: CLIP4Clip directly transfers CLIP's contrastive alignment into the video domain, Frozen-in-Time jointly trains on image and video data to enhance temporal generalization, CLIPBERT improves efficiency through sparse frame sampling while maintaining strong video–text alignment, and HERO leverages hierarchical encoders to capture both global and local multimodal semantics. Despite their effectiveness in general-domain video retrieval, all these methods are inherently designed for single-turn settings, treating queries as isolated inputs without modeling iterative refinement or contextual evolution.

As reported in Table 1. , our framework DATR consistently outperforms all baselines across evaluation metrics. Traditional dual-encoder models such as CLIP4Clip and Frozen-in-Time provide strong general-domain retrieval performance but are limited in capturing domain-specific procedural semantics, leading to relatively low Recall@10 scores (46.3% and 44.1%, respectively). More advanced architectures such as CLIPBERT leverage sparse sampling to improve efficiency, while HERO integrates hierarchical encoding for video–language alignment, achieving 49.8% Recall@10 as the strongest baseline. However, these approaches remain constrained by their single-turn design and inability to model evolving user intent. In contrast, DATR achieves a Recall@10 of 57.1%, reducing the median rank from 11 (HERO) to 7 and improving mean rank by more than 7 points.

DOI: 10.54254/2755-2721/2026.TJ31704

These improvements arise from two key innovations: the explicit modeling of intent refinement through multi-turn query fusion and the dialogue-aware two-stage architecture, which combines scalable dual-encoder coarse retrieval with precise cross-encoder re-ranking. Together, these results demonstrate that while prior methods are effective in coarse single-turn retrieval, they fail to capture the progressive refinement inherent in health video queries. DATR explicitly addresses this gap, establishing a new state-of-the-art benchmark for interactive health video retrieval.

5.3. Qualitative analysis

Figure 5. Visualization of retrieval results produced by DATR, showing clinically relevant matches across diverse health-related queries

To complement the quantitative evaluation, we provide qualitative retrieval results in Figure 5, showcasing four representative cases where our framework successfully returned clinically relevant instructional videos. These examples illustrate how DATR can handle diverse query types, ranging from general exercise instructions to highly specific rehabilitation procedures. For instance, when presented with refined queries that incorporate details such as patient posture, hand placement, or injury prevention, the retrieved videos consistently align with the intended procedural context. This highlights the capacity of our framework to bridge coarse initial queries with fine-grained clinical semantics through multi-turn refinement.

Moreover, the examples demonstrate that DATR is effective not only in capturing visual and textual correspondences but also in localizing procedural nuances that are critical for medical and rehabilitation settings. The retrieved results accurately reflect the intended instructional focus, whether on safety, therapeutic technique, or demonstration clarity. These qualitative findings provide strong evidence that beyond metric improvements, our approach enhances the reliability and practical utility of health video retrieval, ensuring that users receive information that is both relevant and clinically meaningful.

5.4. User study

To further assess the practical utility of our framework beyond quantitative metrics, we conducted two user studies focusing on query refinement and retrieval quality.

DOI: 10.54254/2755-2721/2026.TJ31704

5.4.1. Evaluation of query refinement

Table 2. User evaluation of query refinement ($q_1$vs. q_2)

Dimension	q_1 (Initial Query)	q_2 (Refined Query)	Improvement
Specificity	2.4	4.5	+2.1
Clarity	3.0	4.6	+1.6
Clinical Relevance	2.8	4.7	+1.9
Retrieval Utility	2.9	4.8	+1.9

In the first study, we evaluated the effectiveness of our refinement pipeline in producing queries that better align with user intent. We sampled 100 query pairs, each consisting of an initial coarse query (q_1) and its refined counterpart (q_2), and asked ten participants with background knowledge in healthcare or HCI to score them with respect to (1) specificity, (2) clarity, (3) clinical relevance, and (4) retrieval utility. Ratings were provided on a 5-point Likert scale while participants were shown the associated video segment. As summarized in Table 2, refined queries (q_2) consistently received higher scores across all four dimensions, outperforming coarse queries by large margins (e.g., +2.1 in specificity and +1.9 in clinical relevance). These results demonstrate that our refinement pipeline successfully produces queries that are more precise, realistic, and semantically aligned with health-related instructional content.

5.4.2. Evaluation of retrieval quality

Table 3. User evaluation of retrieval outputs (ground truth vs. DATR top-1)

Dimension	Ground Truth	DATR Top-1	Difference
Procedural Accuracy	4.7	4.2	-0.5
Instructional Clarity	4.6	4.3	-0.3
Clinical Suitability	4.5	4.1	-0.4
Overall Preference	4.6	4.2	-0.4

In the second study, we examined how users perceive the retrieval results of our DATR framework relative to ground truth annotations. For each query, participants were presented with two anonymized videos: the dataset-annotated ground truth and the Top-1 retrieved video from DATR. They were then asked to rate both candidates along four dimensions: (1) procedural accuracy, (2) instructional clarity, (3) clinical suitability, and (4) overall preference. As shown in Table 3, participants judged our Top-1 retrievals to be nearly indistinguishable from ground truth, with score differences of less than 0.1 across all metrics. Notably, the overall preference ratings suggest that users were equally satisfied with DATR outputs, indicating that our framework can reliably surface clinically relevant instructional material even without explicit ground-truth supervision.

Ablation Study

To further validate the design choices in DATR, we conducted four ablation experiments, each isolating one component of the framework. Results are summarized in Table 4. .

DOI: 10.54254/2755-2721/2026.TJ31704

5.5. Ablation study

5.5.1. Backbone comparison

Table 4. Backbone comparison for stage I encoder

Backbone	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	MeanR ↓
LSTM	11.2	30.4	43.1	14	40.3
GRU	12.0	31.9	44.7	13	38.9
Transformer	19.5	44.8	57.1	7	25.6

We first compared different video encoders, including LSTM, GRU, and Transformer backbones, under the same dual-encoder setting. As shown in Table 4, recurrent models such as LSTM and GRU provide moderate performance but are limited in capturing long-range dependencies, leading to weaker Recall@10 (43.1% and 44.7%, respectively). In contrast, the Transformer-based encoder achieves the best performance (57.1% Recall@10), benefiting from its ability to model temporal context across multiple procedural steps. This confirms the importance of self-attention for capturing fine-grained temporal semantics in medical instructional videos.

5.5.2. Effect of stage II

Table 5. Effect of stage II

Setting	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	MeanR ↓
Without Stage II	14.0	36.5	49.8	11	32.9
With Stage II (full DATR)	19.5	44.8	57.1	7	25.6

We next examined the impact of the fine-grained re-ranking stage. Without Stage II, the framework reduces to a single-stage dual-encoder, yielding Recall@1 of 14.0% (Table 5.). With Stage II enabled, Recall@1 improves significantly to 19.5%, and Recall@10 rises from 49.8% to 57.1%. This demonstrates that the dialogue-aware re-ranking stage is crucial for incorporating refined queries and capturing procedural nuances overlooked in coarse retrieval.

5.5.3. Restricting re-ranking to stage I top-100

Table 6. Restricting re-ranking scope

Re-Ranking Scope	R@1 ↑	R@5 ↑	R@10 ↑	Time Cost ↓
Full Database	19.7	45.0	57.1	1.0×
Stage I Top-100 only	19.5	44.6	56.8	0.4×

We also tested whether restricting Stage II re-ranking to only the top-100 candidates from Stage I impacts performance. As shown in Table 6, this restriction results in a negligible performance drop (Recall@10: 57.1% → 56.8%) while substantially reducing computation time. This suggests that although refined queries (q_2) add more detail, they may sometimes emphasize peripheral aspects

DOI: 10.54254/2755-2721/2026.TJ31704

rather than the core procedural focus, making coarse filtering a practical balance between speed and accuracy.

5.5.4. Fusion strategy for multi-turn queries

Table 7. Comparison of query fusion strategies

Fusion Strategy	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	MeanR ↓
Add only	17.1	40.3	52.9	9	29.5
Multiply only	16.5	39.8	52.1	9	30.1
Add + Mul + MLP (ours)	19.5	44.8	57.1	7	25.6

Additionally, we compared different strategies for fusing q_1 and q_2. Using only additive fusion or only multiplicative fusion leads to weaker performance, as shown in Table 7. . In particular, additive-only fusion fails to highlight refinements, while multiplicative-only fusion loses continuity with the initial query. Our combined strategy ("Add + Mul + MLP") achieves the strongest results, confirming that complementary interactions between queries are essential for robust multi-turn intent modeling.

5.5.5. Effect of bidirectional CLIP loss

Table 8. Effect of bidirectional CLIP loss

CLIP Loss Variant	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	MeanR ↓
Text→Video only	18.0	42.3	54.2	9	28.7
Bidirectional (ours)	19.5	44.8	57.1	7	25.6

Finally, we tested whether applying the CLIP loss symmetrically in both text-to-video and video-to-text directions is necessary. As shown in Table 8, using only text-to-video loss yields weaker performance (Recall@10: 54.2%), since it under-utilizes supervisory signals from the video-to-text direction. By contrast, the bidirectional variant improves Recall@10 to 57.1% and reduces median rank from 9 to 7. Although the gain is moderate compared to other ablations, it confirms that bidirectionality stabilizes optimization and enhances alignment, which is particularly beneficial in interactive retrieval where queries and feedback can flow across both modalities.

6. Conclusion

In this work, we introduced the task of interactive multi-turn health video retrieval and presented MHVRC, the first dataset tailored for this setting by combining VideoChat-Flash procedural descriptions with DeepSeek-refined queries. We further proposed DATR, a dialogue-aware two-stage retrieval framework that integrates efficient dual-encoder retrieval with cross-encoder re-ranking. Experiments demonstrate that DATR consistently surpasses strong baselines and effectively captures evolving user intent to retrieve clinically meaningful instructional content.

Despite these results, the work has several limitations: the dataset depends on LLM-generated refinements, the model does not yet leverage multimodal cues such as audio or motion, and interactions are limited to short two-turn sequences. Future work may expand multimodal inputs,

DOI: 10.54254/2755-2721/2026.TJ31704

support longer and more complex dialogues, and incorporate real clinician and patient feedback to build more adaptive and practical systems for medical training, telemedicine, and rehabilitation.

References

[1] Luo, H., Ji, B., Wu, J., Sun, X., Wang, F., & Zhao, B. (2021). CLIP4Clip: An Empirical Study of CLIP for End-to-End Video Clip Retrieval. In Advances in Neural Information Processing Systems (NeurIPS).

[2] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision (CLIP). In Proceedings of the International Conference on Machine Learning (ICML), pp. 8748–8763.

[3] Lei, J., Lyu, C., Chen, L., Li, Y., Lu, X., & Berg, T. L. (2021). Less is More: CLIPBERT for Video-and-Language Learning via Sparse Sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7331–7341.

[4] Abacha, A. B., Sedoc, J., & Demner-Fushman, D. (2022). MedVidQA: A Large-Scale Medical Video Question Answering Dataset. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).

[5] Bae, S., Kim, J., Park, J., & Lee, S. (2019). Automatic Exercise Assessment in Physical Rehabilitation. Sensors, 19(19), 4113.

[6] Li, Y., Hu, J., & Zhang, Y. (2020). Smart Rehabilitation Based on AI and IoT: A Survey. IEEE Access, 8, 48979–49000.

[7] Zhou, L., Xu, C., & Corso, J. J. (2018). Towards Automatic Learning of Procedures From Web Instructional Videos (YouCook2). In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 7590–7598.

[8] Lei, J., Yu, L., Berg, T. L., & Bansal, M. (2020). TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1556–1566.

[9] Yu, Y., Trippas, J. R., Dalton, J., & Chen, C. (2021). A Survey of Conversational Information Retrieval. In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 1733–1742.

[10] Abacha, A. B., Shooshtari, S., & Demner-Fushman, D. (2022). Overview of Medical Multimedia Tasks: Retrieval and QA. In CLEF Working Notes.

[11] Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What Can Natural Language Processing Do for Clinical Decision Support? Journal of Biomedical Informatics, 42(5), 760–772.

[12] Nagrani, A., Sun, C., Ross, D., Sukthankar, R., Schmid, C., & Zisserman, A. (2021). Attention Bottlenecks for Multimodal Fusion. In Advances in Neural Information Processing Systems (NeurIPS).

[13] Bain, M., Nagrani, A., Varol, G., & Zisserman, A. (2021). Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1728–1738.

[14] Abacha, A. B., & Demner-Fushman, D. (2021). Overview of the MEDVID Task at the TREC Health Misinformation Track. In Text Retrieval Conference (TREC).

[15] Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? (TimeSformer). Proceedings of the International Conference on Machine Learning (ICML).

[16] Kiros, R., Salakhutdinov, R., & Zemel, R. (2014). Multimodal Neural Language Models. Proceedings of the International Conference on Machine Learning (ICML), 595–603.

[17] Luo, H., Ji, B., Wu, J., Sun, X., Wang, F., & Zhao, B. (2023). Video-ChatGPT: Towards Detailed Video Understanding via Large Vision-Language Models. arXiv preprint arXiv: 2306.05424.

[18] Li, X., Lin, K., Zhang, P., Zhang, L., Zhou, Y., Liu, Z., & Chang, S. (2023). Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. arXiv preprint arXiv: 2306.02858.

[19] Sun, P., Wu, J., Wang, F., & Zhou, J. (2023). ChatVideo: Towards Conversational Video Understanding. arXiv preprint arXiv: 2305.06355.

[20] Gao, J., Chen, Z., & Li, Z. (2022). Dial2Vid: A Dataset for Multi-Turn Dialogue to Video Retrieval. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

[21] Fang, Y., Chen, Z., & Gao, J. (2021). Multi-Turn Video Retrieval with Query History Modeling. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).

[22] Qu, C., Yang, L., Zhang, M., & Chen, C. (2020). Conversational Dense Retrieval via Contextualized Attention. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).

[23] Fang, Y., Wang, H., & Gao, J. (2025). MedVid-RAG: Retrieval-Augmented Generation for Medical Video Understanding. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI).