Overview of the NLPCC 2026 Shared Task 1: Difficulty-Aware Multilingual and Multimodal Medical Instructional Video Understanding Evaluation

Shenxi Liu[1], Kan Li[1], Mingyang Zhao[2], Yuhang Tian[1], and Bin Li[3⋆]

[1] School of Computer Science and Technology, Beijing Institute of Technology
{liushenxi,likan,tianyuhang}@bit.edu.cn
[2] Department of Computing, The Hong Kong Polytechnic University
25019897r@connect.polyu.hk
[3] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
b.li2@siat.ac.cn

Abstract. Following the CMIVQA, MMI-VQA, and M4IVQA challenges in NLPCC 2023–2025, we introduce the Difficulty-Aware Medical Instructional Video Question Answering (DA-MIVQA) shared task for NLPCC 2026. DA-MIVQA extends previous multilingual and multimodal medical video benchmarks by explicitly distinguishing questions according to the type and complexity of evidence required for answering. Specifically, simple questions can often be answered from subtitle-based textual cues, whereas complex questions require visual grounding, procedural understanding, and cross-modal evidence integration. The challenge contains three tracks: Difficulty-Aware Temporal Answer Grounding in Single Video (DA-TAGSV), Difficulty-Aware Video Corpus Retrieval (DA-VCR), and Difficulty-Aware Temporal Answer Grounding in Video Corpus (DA-TAGVC). The dataset is collected from public medical instructional channels, covers diverse scenarios such as first aid, emergency response, rehabilitation, nursing, and general medical education, and is manually verified with difficulty annotations. This paper presents the task motivation, dataset construction, evaluation protocol, participation overview, competition results, and representative systems of DA-MIVQA. DA-MIVQA provides a practical benchmark for evaluating medical instructional video question answering systems under varying textual, visual, temporal, and procedural reasoning requirements.

Keywords: Multilingual medical instructional video · Multi-hop question answering · Video retrieval · Temporal answer grounding.

1 Introduction

Recent advances in AI-assisted healthcare have shown promise in clinical and educational scenarios such as medical imaging, decision support, and procedural

⋆ Corresponding author

training [1–3]. Medical instructional videos are increasingly used for acquiring practical skills in first aid, emergency response, nursing, rehabilitation, and general medical education [4–6]. Compared with text alone, such videos demonstrate medical actions, tool usage, posture transitions, and temporal procedure flows, providing richer evidence for medical learning and question answering [7–9].

This value has motivated a series of NLPCC shared tasks on medical instructional video question answering, including CMIVQA 2023 [10], MMIVQA 2024 [11], and M4IVQA 2025 [12]. These tasks have expanded from Chinese medical video QA to multilingual, multimodal, and multi-hop reasoning settings [13], with two core directions: temporal answer grounding, which localizes answer-relevant spans in untrimmed videos, and video corpus retrieval, which identifies relevant videos from large collections [14, 15].

However, existing benchmarks do not explicitly distinguish questions by the evidence required for answering in realistic medical scenarios [16, 17]. While some questions can be answered from subtitles or basic semantic matching, others require visual grounding, action understanding, object-state recognition, and procedural context modeling. Thus, difficulty in medical instructional video understanding should not be defined only by reasoning depth or inference hops, but also by whether explicit visual evidence is required beyond textual cues.

To address this gap, NLPCC 2026 introduces the **Difficulty-Aware Medical Instructional Video Question Answering** challenge, or **DA-MIVQA**[1]. Unlike previous tasks emphasizing language scope, modality coverage, or multi-hop reasoning, DA-MIVQA evaluates systems according to the evidence structure required by each question. It divides questions into *simple* and *complex* subsets: simple questions are typically answerable from subtitle-aligned textual cues or explicit single-source evidence, whereas complex questions require visual grounding and procedural understanding. This design better reflects real-world medical use, where users need not only textual hints but also visual evidence about what to do, where to act, and when a procedure step occurs [18–20].

DA-MIVQA retains the three-track formulation of previous NLPCC shared tasks while applying difficulty-aware evaluation to each track. **DA-TAGSV** requires systems to localize the answer span in a single video; **DA-VCR** retrieves the most relevant video from a corpus; and **DA-TAGVC** first retrieves the relevant video and then grounds the answer temporally within it. Evaluating all tracks under simple, complex, and mixed settings enables fine-grained analysis of robustness across procedural and cross-modal difficulty levels.

The DA-MIVQA dataset is collected from public medical instructional channels on YouTube and covers scenarios such as first aid, emergency management, rehabilitation, nursing, and general medical education [21]. Following previous shared tasks, questions and temporal answers are manually verified by annotators with medical backgrounds, and each question is labeled as simple or complex according to its required evidence. This annotation helps reveal whether systems understand visual and procedural content or mainly rely on subtitle matching.

[1] https://med-m3-dataset.github.io/ for more informaton.

Table 1. A conceptual comparison of the NLPCC medical instructional video QA shared tasks.

Task	Languages	Multi-modal	Chain-of-thought	Difficulty-aware
CMIVQA (2023) [10]	Chinese	✓	✗	✗
MMIVQA (2024) [11]	Chinese/English	✓	✗	✗
M4IVQA (2025) [12]	Chinese/English	✓	✓	✗
DA-MIVQA (2026)	Chinese/English	✓	✓	✓

DA-MIVQA supports more reliable medical video QA for education, skill acquisition, emergency guidance, and multilingual knowledge access. It also provides a benchmark for distinguishing systems that mainly match subtitles from those that integrate visual, textual, and procedural evidence in medical scenarios. This paper presents an overview of DA-MIVQA, covering its motivation, task definition, dataset construction, evaluation protocol, participation summary, and representative systems.

2 Task Introduction

The DA-MIVQA challenge aims to promote medical instructional video question answering under a difficulty-aware evaluation setting. Different from previous NLPCC shared tasks that mainly emphasized multilingual understanding, multimodal fusion, or multi-hop reasoning [10–13], DA-MIVQA explicitly evaluates system performance across questions with different evidence requirements [22, 23]. All questions are categorized into *simple* and *complex* subsets. Simple questions are mainly answerable through subtitle-aligned textual matching or basic semantic understanding, whereas complex questions require visual grounding, procedural interpretation, and integration of textual and visual evidence from medical instructional videos. Following the design paradigm of previous NLPCC medical video shared tasks, DA-MIVQA contains three tracks: Difficulty-Aware Temporal Answer Grounding in Single Video, Difficulty-Aware Video Corpus Retrieval, and Difficulty-Aware Temporal Answer Grounding in Video Corpus.

2.1 Definition of Each Track

The challenge contains three complementary tracks covering temporal localization, corpus retrieval, and their joint modeling in medical instructional video understanding.

Track 1: Difficulty-Aware Temporal Answer Grounding in Single Video (DA-TAGSV). Given a medical or health-related question and a single untrimmed medical instructional video, DA-TAGSV requires systems to localize the start and end timestamps of the answer span. It evaluates fine-grained temporal localization ability under both subtitle-dominant simple questions and visually grounded complex questions.

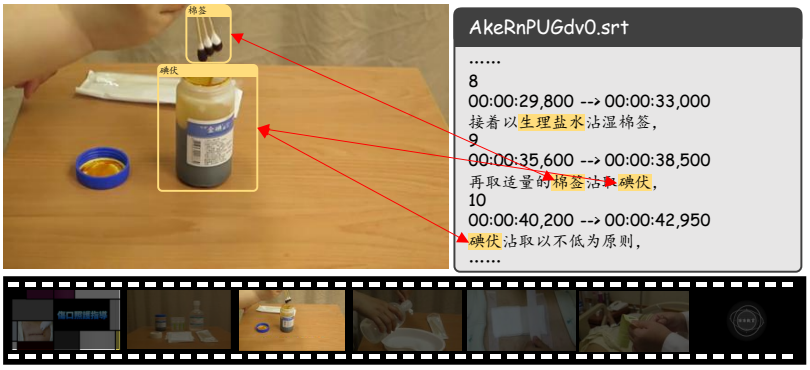

Fig. 1. The multi-modal features of complex questions in DA-MIVQA.

Track 2: Difficulty-Aware Video Corpus Retrieval (DA-VCR). Given a question and a large collection of untrimmed medical instructional videos, DA-VCR requires systems to retrieve the most relevant video from the corpus. It evaluates semantic matching between questions and candidate videos under multilingual and multimodal conditions, while further comparing retrieval robustness across simple and complex questions.

Track 3: Difficulty-Aware Temporal Answer Grounding in Video Corpus (DA-TAGVC). Given a question and a video corpus, DA-TAGVC requires systems to first retrieve the relevant video and then localize the answer span within it. This track combines the challenges of DA-TAGSV and DA-VCR, making it more demanding because retrieval and localization errors may accumulate, especially for complex questions.

2.2 Evaluation Metrics

Following previous NLPCC medical instructional video shared tasks, DA-MIVQA adopts task-specific metrics for the three tracks and reports results on simple-only, complex-only, and mixed subsets [10–12]. This protocol measures not only overall effectiveness but also robustness under different levels of evidence complexity.

For **DA-TAGSV**, evaluation is based on temporal overlap between predicted and gold answer spans. We adopt Intersection over Union (IoU), mean Intersection over Union (mIoU), and $R@1,\ IoU = \mu$, where $\mu \in \{0.3, 0.5, 0.7\}$. The main ranking metric is mIoU.

For **DA-VCR**, evaluation is based on whether the relevant video is ranked highly in the returned list. We use recall-based metrics $R@n$ with $n \in \{1, 10, 100\}$ and Mean Reciprocal Rank (MRR). The main ranking metric is the *Overall* score, computed as the average of R@1, R@10, R@100, and MRR.

For **DA-TAGVC**, evaluation combines retrieval and localization. We report $R@1/mIoU$, $R@10/mIoU$, and $R@100/mIoU$, which reflect localization quality

under different retrieval depths. The main ranking metric is the *Average* score, computed as the mean of these three values.

2.3 Dataset

The DA-MIVQA dataset is collected from public medical instructional channels on YouTube and covers diverse medical and health-related scenarios, including first aid, medical emergency management, rehabilitation guidance, nursing practice, and general medical education [24]. Following previous shared tasks, questions and temporal answers are manually verified by annotators with medical backgrounds to ensure annotation quality and medical validity [25]. Each video may contain multiple question-answer pairs, and each question is associated with a unique temporal answer span marked by start and end timestamps.

Compared with previous datasets, DA-MIVQA further introduces difficulty labels, namely *simple* and *complex*, according to the type of evidence required for answering each question. This re-annotation strategy distinguishes subtitle-dominant matching questions from genuinely multimodal procedural understanding questions [26]. The dataset is divided into training, validation, and test subsets, each containing Chinese and English questions with simple and complex labels. Overall, DA-MIVQA provides a more fine-grained benchmark for multilingual and multimodal medical instructional video understanding.

Table 2. Statistics of sample counts for the three tracks.

Track	Language	Complexity	Training Set	Validation Set	Test Set	Total
Track 1	Chinese	Simple	2101	362	351	2814
		Complex	1842	285	339	2466
	English	Simple	1273	199	202	1674
		Complex	1341	223	219	1783
Track 2	Chinese	Simple	2101	362	394	2857
		Complex	1842	285	280	2407
	English	Simple	1273	199	210	1682
		Complex	1341	223	234	1798
Track 3	Chinese	Simple	2101	362	394	2857
		Complex	1842	285	280	2407
	English	Simple	1273	199	210	1682
		Complex	1341	223	234	1798

3 Baseline

To provide reference points for the proposed DA-MIVQA evaluation, we design two complementary baselines: a text-only oracle baseline and a multimodal encoder–decoder baseline. The former estimates how far a system can go using subtitle information alone, while the latter provides a practical reference for modeling visual evidence, temporal context, and cross-modal interactions. Together,

these baselines allow us to examine the performance gap between subtitle-based reasoning and multimodal video understanding under different evidence requirements.

3.1 Text-only Oracle Baseline

For the text-only oracle baseline, we use a frozen Qwen3.5-9B [27] model as the backbone large language model. Only timestamped SRT subtitles are provided as input, and the model is adapted to DA-MIVQA through task-specific prompts without updating the backbone parameters. Each instance consists of a question, the corresponding subtitle sequence, and the expected task-specific output.

For DA-TAGSV, the model predicts the temporal span in the given video that best supports the answer. For DA-VCR, it identifies the most relevant video from the candidate corpus based on subtitle-level textual evidence. For DA-TAGVC, it first infers the relevant video and then predicts the answer-supporting span within that video. This baseline is not intended as a deployable multimodal solution, but as an estimate of the upper-bound performance achievable from subtitles alone.

3.2 Multimodal Encoder–Decoder Baseline

The second baseline adopts a multimodal encoder–decoder framework that jointly models questions, timestamped subtitles, and visual clips from medical instructional videos. It is conceptually inspired by cross-modal mutual knowledge transfer for visual answer localization [28], but is adapted to the evidence-aware setting of DA-MIVQA rather than directly reproducing the original model.

The framework first extracts subtitle, question, and visual clip representations. It then performs multimodal evidence mining to select question-relevant subtitle spans and video clips, followed by iterative cross-modal reasoning between textual and visual representations. The resulting multimodal representation is fed into task-specific decoders for the three tracks: timestamp prediction for DA-TAGSV, video relevance scoring for DA-VCR, and joint retrieval plus temporal localization for DA-TAGVC. The model is optimized with localization, retrieval, and cross-modal consistency objectives.

Overall, this baseline provides a practical multimodal reference by combining subtitle cues, visual demonstrations, and temporal evidence within a unified architecture. Compared with the text-only oracle, it is expected to better handle complex questions that require visual grounding and procedural understanding.

4 Evaluation Results

In this section, we summarize the participation status and evaluation results of the NLPCC 2026 shared task on Difficulty-Aware Medical Instructional Video Question Answering (DA-MIVQA). Following the reporting style of previous NLPCC shared task overview papers, we present the overall participation overview,

the official rankings of each track, and brief comparative observations across different difficulty settings [10–12].

4.1 Participation Overview

The DA-MIVQA challenge attracted teams from universities, research institutes, and industry participants interested in multilingual and multimodal medical video understanding.

For NLPCC 2025 Shared Task 4, a total of 22, 14, 16 teams registered for Track 1, 2, 3, respectively. During the test phase, 17, 10, and 11 teams submitted valid results, respectively. In the end, **Amazon Inc.**, **Team_WuKong**, and **BIGC** achieved the SOTA for Tracks 1, 2, and 3. In general, the participation distribution across the three tracks can reflect the relative technical difficulty of each task setting, especially because DA-TAGVC requires systems to jointly solve retrieval and temporal grounding under difficulty-aware evaluation.

4.2 Results of Track 1: DA-TAGSV

Track 1, Difficulty-Aware Temporal Answer Grounding in Single Video, evaluates whether a system can accurately localize the temporal answer span in an untrimmed medical instructional video for a given question. The official ranking of this track is based on the *mIoU* score, while *R@1, IoU = 0.3*, *R@1, IoU = 0.5*, and *R@1, IoU = 0.7* are reported as complementary metrics for detailed analysis. Table 3 presents the final ranking results for Track 1.

Table 3. Results of Track 1: Difficulty-Aware Temporal Answer Grounding in Single Video (DA-TAGSV).

Rank	Team ID	R@1,IoU=0.3	R@1,IoU=0.5	R@1,IoU=0.7	mIoU
1	Amazon Inc.	**0.5512**	0.4007	0.2217	**0.3912**
2	Karamay	0.5334	**0.4110**	**0.2269**	0.3904
3	Ouc_AI [29]	0.5101	0.3579	0.2146	0.3608
4	Dvoe protiv vetra	0.4946	0.3688	0.2183	0.3606
5	SETAG[30]	0.4116	0.3363	0.1790	0.3089
6	HIIT	0.4191	0.3287	0.1760	0.3079
7	MM-Baseline [28]	0.4265	0.3211	0.1729	0.3068
8	Text-Only Baseline [27]	0.4206	0.3127	0.1441	0.2925
9	BFSU	0.4147	0.3043	0.1153	0.2782
10	Random Pick Method [31]	0.0774	0.0818	0.0403	0.0665

Under the difficulty-aware setting, Track 1 is expected to reveal a clear performance gap between simple and complex questions, because the latter require more reliable visual grounding and procedural understanding beyond subtitle-based matching. Compared with simple questions, complex questions in this track typically place higher demands on identifying medically relevant actions, body positions, tool usage, and temporal transitions in the video.

4.3 Results of Track 2: DA-VCR

Track 2, Difficulty-Aware Video Corpus Retrieval, evaluates whether a system can retrieve the most relevant medical instructional video from a large corpus for an input question [15]. The official ranking of this track is based on the *Overall* score, which summarizes the retrieval quality measured by *R@1*, *R@10*, *R@100*, and *MRR*. Table 4 shows the final ranking results for Track 2.

Table 4. Results of Track 2: Difficulty-Aware Video Corpus Retrieval (DA-VCR).

Rank	Team ID	R@1	R@10	R@100	MRR	Overall
1	Team_Wukong	**0.3751**	0.4150	0.5189	**0.3931**	**0.4255**
2	Chaldeas	0.3559	**0.4160**	**0.5492**	0.3537	0.4187
3	DIMA[32]	0.3268	0.4355	0.5260	0.3325	0.4052
4	sun [33]	0.3086	0.3721	0.4442	0.3000	0.3562
5	HDU_Team2	0.3139	0.3647	0.4447	0.2982	0.3554
6	DSG-1 [34]	0.3192	0.3572	0.4452	0.2964	0.3545
7	MM-Baseline [28]	0.1570	0.1953	0.2046	0.1546	0.1779
8	AC Automation	0.1518	0.1712	0.1704	0.1461	0.1599
9	Text-Only Baseline [27]	0.1465	0.1470	0.1361	0.1375	0.1418
10	Random Pick Method [31]	0.0361	0.0822	0.0962	0.0636	0.0695

Compared with Track 1, Track 2 places greater emphasis on identifying question-video relevance at the corpus level, and therefore tests the semantic matching ability of systems under multilingual and multimodal conditions. Under the simple setting, retrieval performance may benefit more from lexical overlap and subtitle-level semantic clues, whereas the complex setting is expected to require stronger modeling of visually grounded procedures and action semantics.

4.4 Results of Track 3: DA-TAGVC

Track 3, Difficulty-Aware Temporal Answer Grounding in Video Corpus, is the most comprehensive setting in DA-MIVQA because it requires systems to retrieve the relevant video and localize the answer span within that video at the same time. The official ranking of this track is based on the *Average* score, which is computed from *R@1/mIoU*, *R@10/mIoU*, and *R@100/mIoU*. Table 5 provides the final ranking results for Track 3.

Because retrieval and localization errors may accumulate in this track, its performance is expected to be lower than that of the first two tracks, especially on complex questions requiring both correct video selection and accurate visual-temporal grounding. Therefore, the results of this track are particularly useful for examining the end-to-end capability of systems in realistic medical instructional video question answering scenarios.

4.5 Comparative Observations

Across all three tracks, the difficulty-aware setting makes it possible to compare system behavior on text-dominant and visually grounded questions in a

Table 5. Results of Track 3: Difficulty-Aware Temporal Answer Grounding in Video Corpus (DA-TAGVC).

Rank	Team ID	R@10,mIoU	R@100,mIoU	R@1,mIoU	Average
1	BIGC	**0.1646**	**0.2879**	**0.3660**	**0.2728**
2	UWM	0.1415	0.2727	0.3338	0.2493
3	IIEleven[35]	0.1238	0.2513	0.3566	0.2439
4	UESC	0.1364	0.2551	0.3359	0.2425
5	MedEcho[36]	0.1490	0.2588	0.3152	0.2410
6	Nsddd[37]	0.1436	0.2129	0.3235	0.2267
7	DesiWen	0.1232	0.2351	0.3159	0.2248
8	MM-Baseline [28]	0.1028	0.2572	0.3082	0.2228
9	Text-Only Baseline [27]	0.0885	0.2415	0.2853	0.2051
10	Random Pick Method [31]	0.0454	0.0872	0.0673	0.0666

more fine-grained manner. The expected comparison between *Simple Only* and *Complex Only* results can help reveal whether a model mainly relies on subtitle matching or can truly integrate textual, visual, and procedural evidence. In particular, if a system maintains relatively stable performance across the two difficulty subsets, it may indicate stronger robustness in handling medically grounded multimodal reasoning. By contrast, a large performance drop from simple to complex questions would suggest that current methods still have limitations in visual grounding, action understanding, and temporal procedural modeling for medical videos.

5 Conclusion

In this paper, we presented an overview of the NLPCC 2026 shared task on Difficulty-Aware Medical Instructional Video Question Answering (DA-MIVQA), which extends the CMIVQA, MMIVQA, and M4IVQA research line toward a more fine-grained and practically grounded evaluation setting. Unlike previous benchmarks that mainly focused on language coverage, modality expansion, or multi-hop reasoning, DA-MIVQA distinguishes questions according to the type of evidence required for answering. Simple questions are mainly answerable from subtitle-aligned textual cues, whereas complex questions require visual grounding, procedural understanding, and cross-modal evidence integration.

The challenge consists of three tracks: Difficulty-Aware Temporal Answer Grounding in Single Video, Difficulty-Aware Video Corpus Retrieval, and Difficulty-Aware Temporal Answer Grounding in Video Corpus. Together, they evaluate temporal localization, corpus retrieval, and end-to-end retrieval-grounding ability in multilingual and multimodal medical instructional video scenarios. Built from public medical instructional videos and enriched with difficulty annotations, DA-MIVQA provides a practical benchmark for distinguishing systems that mainly rely on subtitle matching from those that can integrate textual, visual, and procedural evidence. We hope this challenge will encourage future research on difficulty-aware medical video understanding and support more ro-

bust medical question answering systems for education, emergency guidance, rehabilitation training, and cross-lingual knowledge access.

Acknowledgement

This work was supported by National Natural Science Foundation of China (Nos. 4222037 and L181010), and Sanming Project of Medicine in Shenzhen (No. SZZYSM202311002).

References

1. David Escobar-Castillejos, Ari Y Barrera-Animas, Julieta Noguez, Alejandra J Magana, and Bedrich Benes. Transforming surgical training with ai techniques for training, assessment, and evaluation: Scoping review. *Journal of Medical Internet Research*, 27, 2025.
2. Elijah W. Riddle, Divya Kewalramani, Mayur Narayan, and Daniel B. Jones. Surgical simulation: Virtual reality to artificial intelligence. *Current Problems in Surgery*, 61(11):101625, 2024.
3. Aidin Shahrezaei, Maryam Sohani, Soroush Taherkhani, and Seyed Yahya Zarghami. The impact of surgical simulation and training technologies on general surgery education. *BMC Medical Education*, 24(1):1297, Nov 2024.
4. Komal Srinivasa, Fiona Moir, and Felicity Goodyear-Smith. The role of online videos in teaching procedural skills in postgraduate medical education: A scoping review. *Journal of Surgical Education*, 79(5):1295–1307, 2022.
5. Komal Srinivasa, Amanda Charlton, Fiona Moir, and Felicity Goodyear-Smith. How to develop an online video for teaching health procedural skills: Tutorial for health educators new to video production. *JMIR Medical Education*, 10, 2023.
6. Ilana Roberts Krumm, Matthew C. Miles, Alison Clay, W. Graham Carlos II, and Rosemary Adamson. Making effective educational videos for clinical teaching. *Chest*, 161(3):764–772, 2022.
7. Bin Li, Yixuan Weng, Bin Sun, and Shutao Li. Learning to locate visual answer in video corpus using question. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2022.
8. Yixuan Weng and Bin Li. Visual answer localization with cross-modal mutual knowledge transfer, 2022.
9. Shutao Li, Bin Li, Bin Sun, and Yixuan Weng. Towards visual-prompt temporal answer grounding in instructional video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8836–8853, 2024.
10. Bin Li, Yixuan Weng, Hu Guo, Bin Sun, Shutao Li, Yuhao Luo, Mengyao Qi, Xufei Liu, Yuwei Han, Haiwen Liang, Shuting Gao, and Chen Chen. Overview of the nlpcc 2023 shared task: Chinese medical instructional video question answering. In Fei Liu, Nan Duan, Qingting Xu, and Yu Hong, editors, *Natural Language Processing and Chinese Computing*, pages 233–242, Cham, 2023. Springer Nature Switzerland.
11. Bin Li, Yixuan Weng, Qiya Song, Lianhui Liang, Xianwen Min, and Shoujun Zhou. Overview of the nlpcc 2024 shared task 7: Multi-lingual medical instructional video question answering. In Derek F. Wong, Zhongyu Wei, and Muyun Yang, editors, *Natural Language Processing and Chinese Computing*, pages 429–439, Singapore, 2025. Springer Nature Singapore.

12. Bin Li, Shenxi Liu, Yixuan Weng, Yue Du, Yuhang Tian, and Shoujun Zhou. Overview of the nlpcc 2025 shared task 4: Multi-modal, multilingual, and multi-hop medical instructional video question answering challenge. In Xian-Ling Mao, Zhaochun Ren, and Muyun Yang, editors, *Natural Language Processing and Chinese Computing*, pages 367–379, Singapore, 2026. Springer Nature Singapore.
13. Shenxi Liu, Kan Li, Mingyang Zhao, Yuhang Tian, Bin Li, Shoujun Zhou, Hongliang Li, and Fuxia Yang. M^3-med: A benchmark for multi-lingual, multi-modal, and multi-hop reasoning in medical instructional video understanding, 2025.
14. Shuang Cheng, Zineng Zhou, Jun Liu, Jian Ye, Haiyong Luo, and Yang Gu. A unified framework for optimizing video corpus retrieval and temporal answer grounding: Fine-grained modality alignment and local-global optimization. In Fei Liu, Nan Duan, Qingting Xu, and Yu Hong, editors, *Natural Language Processing and Chinese Computing*, pages 199–210, Cham, 2023. Springer Nature Switzerland.
15. Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Video corpus moment retrieval with contrastive learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 685–695. ACM, 2021.
16. Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. Assessing modality bias in video question answering benchmarks with multimodal large language models, 2025.
17. Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges, 2022.
18. Annette Burgess, Christie van Diggele, Chris Roberts, and Craig Mellis. Tips for teaching procedural skills. *BMC Medical Education*, 20(2):458, Dec 2020.
19. Jacqueline Colgan, Sarah Kourouche, Geoffrey Tofler, and Thomas Buckley. Use of videos by health care professionals for procedure support in acute cardiac care: A scoping review. *Heart, Lung and Circulation*, 32(2):143–155, 2023.
20. Huabin Liu, Xiao Ma, Cheng Zhong, Yang Zhang, and Weiyao Lin. Timecraft: navigate weakly-supervised temporal grounded video question answering via bi-directional reasoning. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part V*, page 92–107, Berlin, Heidelberg, 2024. Springer-Verlag.
21. Vernon Curran, Karla Simmons, Lauren Matthews, Lisa Fleet, Diana L. Gustafson, Nicholas A. Fairbridge, and Xiaolin Xu. Youtube as an educational resource in medical education: a scoping review. *Medical Science Educator*, 30(4):1775–1782, Dec 2020.
22. Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. Assessing modality bias in video question answering benchmarks with multimodal large language models. *ArXiv*, abs/2408.12763, 2024.
23. Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online, July 2020. Association for Computational Linguistics.
24. Cynthia J. Brame. Effective educational videos: Principles and guidelines for maximizing student learning from video content. *CBE Life Sciences Education*, 15, 2016.

25. Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158, Mar 2023.
26. Shenxi Liu, Kan Li, Mingyang Zhao, Yuhang Tian, Shoujun Zhou, and Bin Li. Medcraft: Automated construction of interpretable and multi-hop video workloads via knowledge graph traversal, 2025.
27. Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026.
28. Yixuan Weng and Bin Li. Visual answer localization with cross-modal mutual knowledge transfer. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.
29. Huan Zhang, Chen Zheng, Yuanjing He, Yan Zhao, and Yuxuan Lai. Improving multilingual temporal answering grounding in single video via llm-based translation and ocr enhancement. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 145–156. Springer, 2024.
30. Zineng Zhou, Jun Liu, Shuang Cheng, Haiyong Luo, Yang Gu, and Jian Ye. Improving cross-modal visual answer localization in chinese medical instructional video using language prompts. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 221–232. Springer, 2023.
31. Bin Li, Yixuan Weng, Hu Guo, Bin Sun, Shutao Li, Yuhao Luo, Mengyao Qi, Xufei Liu, Yuwei Han, Haiwen Liang, et al. Overview of the nlpcc 2023 shared task: Chinese medical instructional video question answering. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 233–242. Springer, 2023.
32. Yu Wang, Tianhao Tan, and Yifei Wang. Hierarchical indexing with knowledge enrichment for multilingual video corpus retrieval. In Xian-Ling Mao, Zhaochun Ren, and Muyun Yang, editors, *Natural Language Processing and Chinese Computing*, pages 393–404, Singapore, 2026. Springer Nature Singapore.
33. Guyang Yu, Xiaoyang Bi, Jielong Tang, Ming Gu, Tianbai Chen, Zhiqiang Li, and Miankuan Zhu. Mqua: Multi-level query-video augmentation for multilingual video corpus retrieval. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 353–364. Springer, 2024.
34. Ningjie Lei, Jinxiang Cai, Yixin Qian, Zhilong Zheng, Chao Han, Zhiyue Liu, and Qingbao Huang. A two-stage chinese medical video retrieval framework with llm. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 211–220. Springer, 2023.
35. Tianxing Ma, Yueyue Hu, Shuang Jiang, Zhenhao Yin, and Tianning Zang. Multilingual temporal answer grounding in video corpus with enhanced visual-textual integration. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 471–483. Springer, 2024.
36. Yangchengyu Zhou, Jiayuan Wu, and Yunze Li. Multi-hop knowledge-enhanced query reasoning for multi-modal medical video qa. In Xian-Ling Mao, Zhaochun Ren, and Muyun Yang, editors, *Natural Language Processing and Chinese Computing*, pages 380–392, Singapore, 2026. Springer Nature Singapore.
37. Shuang Cheng, Zineng Zhou, Jun Liu, Jian Ye, Haiyong Luo, and Yang Gu. A unified framework for optimizing video corpus retrieval and temporal answer grounding: fine-grained modality alignment and local-global optimization. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 199–210. Springer, 2023.