Evaluating Video Question Answering Multimodal Large Language Models

George Awad
National Institute of Standards and Technology (NIST)
100 Bureau Dr, Gaithersburg, MD 20899
george.awad@nist.gov

Sanjay Purushotham
University of Maryland, Baltimore County (UMBC)
1000 Hilltop Cir, Baltimore, MD 21250
psanjay@umbc.edu

Abstract

Recent advancements in large multi-modal models have significantly improved AI's ability to process and understand complex data across multiple modalities, including text, images, and video. However, true comprehension of video content remains a formidable challenge, which requires AI systems to integrate visual, auditory, and temporal information to answer questions in a meaningful way. In this paper, we present a new Video Question Answering (VQA) evaluation benchmark which aims to rigorously assess the capabilities of state-of-the-art multi-modal models in understanding and reasoning about video content. Participating teams developed and tested models that answer a diverse set of video clips-based questions covering various levels of complexity, from factual retrieval to complex reasoning. The benchmark serves as a critical evaluation framework to measure progress in video understanding, helping identify strengths and weaknesses in current multi-modal AI architectures. The main advantages of this benchmark data include high quality human annotations by dedicated trained in-house human workers, employing real-world data in the wild, and adopting a shared task paradigm under controlled conditions to evaluate multiple systems fairly. The benchmark completed its first pilot year, which included two subtasks: Answer Generation Task and Multiple Choice Task. We plan to continue running the benchmark annually by adding new data sources, refining metrics, and adding new question categories.

1. Introduction

The Video Question Answering (VQA) [21] evaluation task aims to advance research at the intersection of computer vision and natural language understanding by requiring systems to answer natural language questions about dynamic visual content. Unlike the image-based Visual Question Answering task [1, 8, 16], Video Question Answering systems must capture temporal dependencies, motion cues, and multi-frame context, making it a rich testbed for multi-modal reasoning. The task encourages the development of models that jointly learn visual representations and linguistic semantics, driving progress in applications such as video retrieval, human–computer interaction, and automated content analysis. In this paper, we introduce a new benchmark for Video Question Answering that evaluates multi-modal models on reasoning over dynamic video content. The benchmark includes two complementary tasks, an answer generation and multiple-choice ranking along with a large-scale human-annotated dataset spanning diverse reasoning categories. We further define a standardized evaluation protocol and present results from a international competition pilot study, providing insights into current model capabilities and limitations.

2. Related Work

Video Question Answering (VideoQA) [15, 21] has been extensively studied through a variety of benchmark datasets that emphasize different aspects of multi-modal reasoning. Early work such as TGIF-QA [7] introduced tasks requiring spatio-temporal reasoning over short animated clips, highlighting the importance of modeling motion and temporal dependencies in addition to static visual content. Subsequent datasets expanded the scope of reasoning and modality integration. TVQA [11] introduced compositional reasoning over video clips paired with subtitles, requiring systems to jointly process visual and textual streams. Similarly, ActivityNet-QA [17] focused on open-ended question answering in longer web videos, emphasizing high-level semantic understanding and temporal context.

More recent benchmarks have shifted toward evaluating VideoQA in more realistic and challenging settings. EgoSchema targets long-form egocentric videos and evaluates temporal reasoning over extended durations[13], while MVBench provides a comprehensive evaluation suite covering diverse video understanding tasks, including temporal localization and reasoning[12]. In addition, open-vocabulary VideoQA settings have been explored to address the limitations of fixed answer spaces, enabling evaluation of models on rare and previously unseen answers[10]. These efforts reflect a broader trend toward evaluating general-purpose multi-modal systems capable of reasoning across visual, temporal, and linguistic dimensions.

Despite recent advances, most existing benchmarks focus on model performance within static datasets and lack a standardized framework for controlled cross-system comparison. In contrast, this work is conducted within an international shared-task competition (TREC[1]), where multiple teams address identical tasks under the same conditions and are evaluated using unified scoring protocols. Building on this paradigm, the Video Question Answering (VQA) evaluation introduced in this paper supports both answer generation and answer ranking, enabling more comprehensive and systematic assessment of multimodal reasoning systems. Furthermore, unlike prior benchmarks, the proposed VQA benchmark is designed to evolve over time: it will be dynamically updated and re-evaluated in future iterations at an international competition to continually stress-test models in the rapidly changing landscape of multimodal vision-language models (VLMs) and agentic AI workflows.

3. Evaluation Tasks

3.1. Answer Generation (AG) Task

Automatically generating multiple high-quality answers to questions about short video clips is a critical step toward building interactive agents, educational platforms, and assistive technologies capable of nuanced video understanding. Unlike single-response tasks, generating and ranking multiple candidate answers captures the inherent ambiguity of natural language and video events. It also provides richer supervision for evaluating generative models, enabling more fine-grained assessment of both content accuracy and ranking quality.

Task Definition: Given a collection of X short videos (each approximately 30 seconds in duration) and a corresponding set of questions (one question per video), automatically generate up to ten textual answers to the associated question per video and present these answers in a ranked list according to their estimated correctness or relevance. In addition, record and report the generation time (in seconds) for each individual answer.

[1] trec.nist.gov

Metrics: Systems are evaluated against human-annotated ground truth using standard text-generation metrics, including METEOR [2], BERTScore [19], and Semantic Textual Similarity (STS) [6]. These metrics are appropriate for Video Question Answering (VQA) as the task requires generating semantically correct natural language responses grounded in visual content; thus, evaluation must capture both lexical alignment and semantic equivalence beyond exact matching. For ranking evaluation, we use Normalized Discounted Cumulative Gain (NDCG) [14], which measures the quality of ordered predictions by emphasizing highly ranked, relevant answers.

3.2. Multiple Choice (MC) Task

Accurate identification of the most appropriate answer to a question about short video content is a key challenge in multi-modal understanding, with applications in video retrieval, interactive assistants, and educational tools. Ranking candidate answers by their likelihood of correctness enables systems to prioritize the most relevant information and provides a robust foundation for downstream tasks such as dialog generation and knowledge grounding.

Task Definition: Given a collection of X short videos (each approximately 30 seconds in duration) and an associated set of question–answer (QA) pairs for each video, automatically rank the candidate answers for each question. Specifically, for each video, the system must order the provided answer options from the most likely correct response to the least likely, producing a ranked list that reflects its confidence in each option.

Metrics: System performance is assessed using standard ranking metrics, including Top-1 accuracy and Mean Reciprocal Rank (MRR). These metrics are well-suited for the MC task, where systems must prioritize correct answers among multiple candidates. Top-1 accuracy measures whether the highest-ranked prediction is correct, reflecting strict decision quality, while MRR captures the average inverse rank of the first correct answer, providing a graded measure of how effectively correct responses are prioritized.

4. Video Question Categories

Designing diverse question categories is essential for evaluating the full spectrum of video question answering tasks. Short videos often contain complex visual, auditory, and narrative cues that require different forms of reasoning - temporal sequencing, causal inference, multi-modal integration, or real-world knowledge. By organizing annotation guidelines into the seven categories as listed below, we ensure that the dataset captures a broad range of cognitive challenges, from low-level perception (e.g, counting or tracking objects) to high-level reasoning (e.g., inferring intentions or common-sense explanations). This structured question categories not only guides human annotators

in generating balanced and challenging questions, but also enables researchers to analyze system performance across distinct reasoning skills, fostering more robust and interpretable evaluation of Video Question Answering models.

Temporal / Order / Attribute Tracking: This category encompasses questions that demand comprehension of the timing or sequence of events, as well as observations of changes in appearance, state, or location over time. For example, annotators may ask when a specific action occurs, which event happens first, or how an object's attributes evolve during the video.

Causal / Plot / Goal-Oriented Reasoning: Questions in this group require reasoning about causes, motivations, or narrative goals. They probe why an action takes place, what outcome a particular event produces, or the overall message or theme of the video.

Audio / Multimodal Cues: These questions depend on audio signals such as speech, music, or environmental sounds, as well as text appearing on the screen. Correct answers necessitate integrating auditory or textual information with visual content.

Multi-hop Reasoning: Multi-hop questions demand combining multiple observations or intermediate inferences. Annotators create prompts that require linking sequential actions or disparate clues to arrive at the correct answer.

Object - People Interactions: This category focuses on interactions between humans and objects or animals. Annotators formulate questions about what people do with specific objects or how they interact with animals in the scene.

Counting / Quantitative: These questions involve numerical reasoning, such as counting objects or comparing quantities. They often intersect with temporal reasoning when events unfold over time.

Common-Sense & World Knowledge Inference: Finally, these questions call for everyday reasoning or background knowledge not explicitly depicted in the video. Annotators rely on real-world context—for example, inferring intentions or habits from subtle cues.

5. Dataset and Annotation Framework

The dataset for this pilot evaluation benchmark was constructed through a large-scale annotation effort designed to support both the initial evaluation and future reuse. Approximately 80,000 YouTube Shorts links from the LLaVA-Video-178K dataset [20] were collected and randomly distributed among five human annotators hired and trained in-house, with no overlap in assigned videos. We should note here that even though we sub-sampled the existing LLaVA dataset, all our questions and answers are new and should be safe to use even if the data had been available to models. Each annotator initially received 1,000 unique video links and additional videos were allocated to those who were able to contribute beyond their initial quota. In total, all human annotators worked over 400 hrs combined. Relying on hired and in-house trained annotators allowed us to generate high quality data.

To facilitate consistent data entry and quality control, we developed and deployed a custom annotation tool (see Figure 1). The tool enabled annotators to view videos, compose question–answer (QA) pairs, and track progress. Each annotator was assigned only one of seven predefined question categories; two annotators covered two categories to ensure adequate representation. For every video, annotators created one question and four answer options, exactly one of which is correct, while the remaining three were intentionally incorrect but plausible. Table 2 shows two sample annotations.

Annotators were required only to be detail-oriented, capable of watching extended hours of video, and proficient in writing clear, grammatically correct English. No other specific background knowledge was mandated. Table 1 summarizes the number of annotated videos per annotator and question category, the percentage each contributed to the total collection, and the division of data into training, testing dataset, and planned future testing subsets. To enable out-of-distribution evaluation, no training data was released for the category "Causal Reasoning". Figures 2 and 3 show the distribution of questions that start with specific keywords such as why, when, what, and how, as well as the number of words in the questions for the training dataset. On the other hand, figures 4 and 5 show the same analysis for the testing dataset. It can be shown that both are highly overlapping.

Data Maintenance, Strength and Limitations: Because the dataset relies on publicly available YouTube content, some video links can become inaccessible—for example, if an uploader removes a video. We are maintaining an ongoing list of unavailable video IDs so that researchers can exclude these items from their experiments and ground-truth references.

In general, this annotation framework and dataset construction strategy provide a reusable collection that supports long-term research while acknowledging the inherent limitations of external video hosting. In contrast to other static datasets, our approach to dataset building and benchmark evaluation is more dynamic, and we plan to continuously add more data overtime each year using more human annotations. Reserving a subset of the dataset private without exposing the ground truth can allow us to measure progress of models year over year.

6. Evaluation Results

In this open VQA evaluation, we received 20 system runs from 7 participating teams for the answer-generation task, while the multiple-choice task received 9 system runs from 5 participating teams.

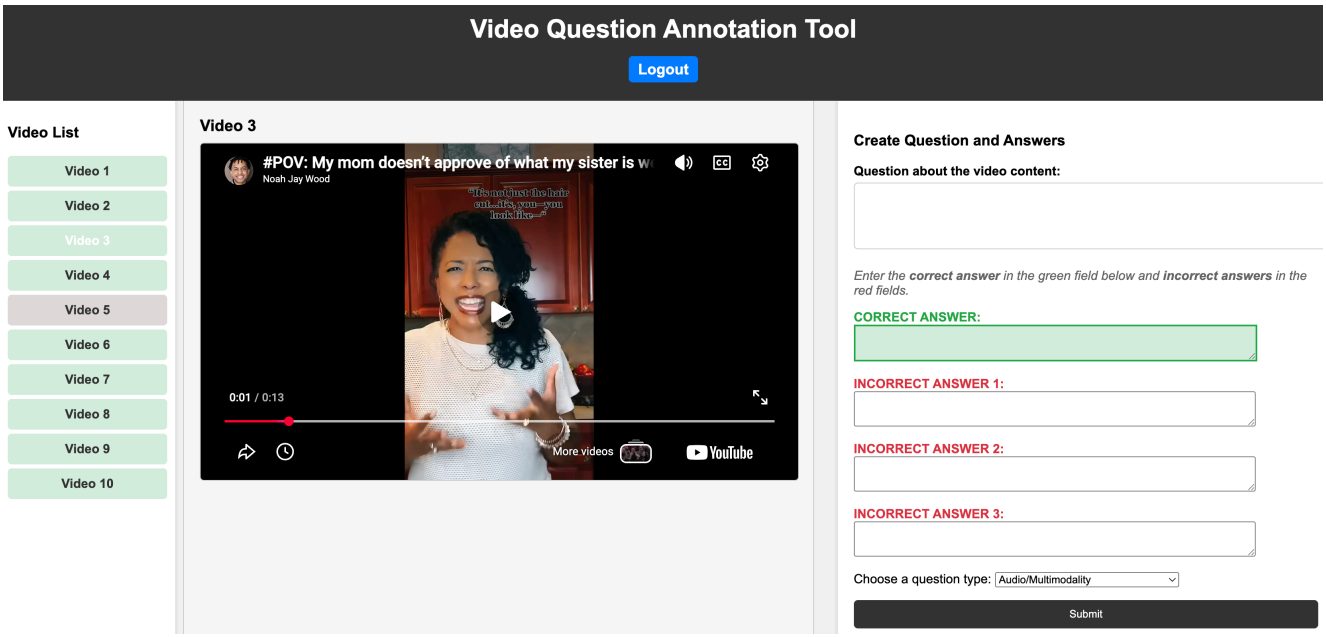

Figure 1. Annotation Tool Interface

Table 1. Dataset Distribution across categories and splits.

	Temporal	Causal	Audio + Multihop	Objects and People	Counting + Common sense	Total
Videos	646	1194	705	689	723	3957
Videos %	16.3	30.1	17.8	17.4	18.2	100
Training	125	0	125	125	125	500
Testing	400	400	400	400	400	2000
Future Testing	121	794	180	164	198	1457

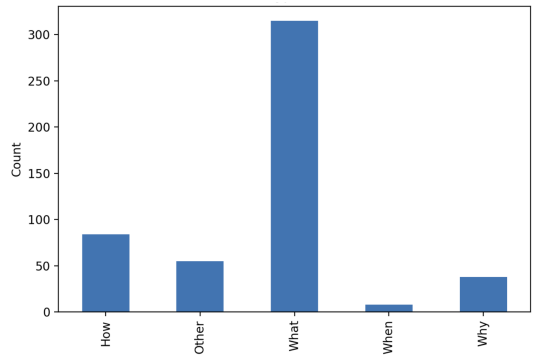

Figure 2. Training Data : Question Types

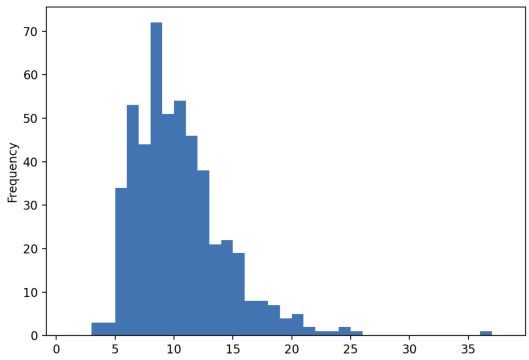

Figure 3. Training Data : words in questions

Figure 8 presents the performance of the 20 submitted runs, reporting the average score of each evaluation metric across the entire test set of videos. We applied three main metrics (METEOR, STS, BERTScore) to measure the similarity between the ground truth answer and the submitted answer. These metrics evaluate the quality of generated text, with each using a different approach to assess the similarity between a machine-generated sentence and a human-written reference sentence. METEOR is a unigram-based metric that calculates a harmonic mean of precision and recall, with a higher weight given to recall. STS measures the degree of semantic equivalence between two text snip-

Table 2. Sample annotations

Sample 1	Sample 2
Why does the boy pick up his toys and run off down the street?	What happens when the yellow pool ball is hit ?
Both of the toys he has lose a part, so he picks them up and runs away	It sinks in the adjacent side pocket
He wants to find a new place to play	other players sink balls
He wants to throw away the toys that he thinks are broken	The shooter gets a pat on the back
He thinks that the person sitting next to him broke the parts off his toys	Bystanders comment

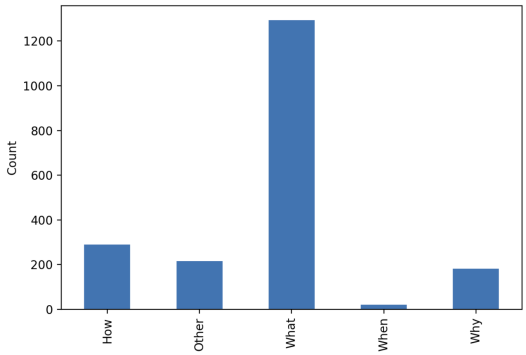

Figure 4. Testing Data : Question Types

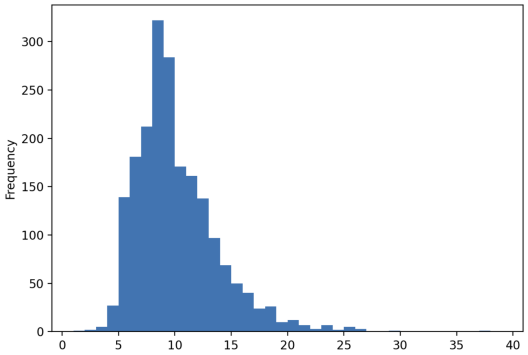

Figure 5. Testing Data : words in questions

pets. BERTScore leverages contextual word embeddings from large pre-trained models like BERT to compute a similarity score.

Scores ranged on average between 0.20 and 0.30 for METEOR and STS, while BERTScore values ranged between 0.80 and 0.90. One team (TJUMI) had a bug in their original runs and later on submitted corrected runs after the official results were released. These new evaluated runs by this team are highlighted by two black boxes in figure 8 and with different colors in figure 9.

Figures 6 and 7 illustrate the pairwise correlations among the evaluation metrics. We observed high correlation between STS and METEOR metrics (r=0.92) and STS and BERTScore (r=0.69). Although the evaluation originally planned to score ranking quality based on NDCG scores, the validation of submitted runs did not force the submission of the required number of answers for each question. Thus, we found that many video-question pairs across different runs had only 1 submitted answer leading to a perfect NDCG score which is not a reliable measure in this case. Eventually, we dropped the NDCG as a metric in this pilot iteration and plan to update the validation of runs so that every video-question pair has a complete set of ranked answers across all runs.

Figure 9 shows the performance of multiple choice task runs. Overall, performance is higher than answer generation task with almost all systems achieving more than 50% mean score across all queries, and top 3 systems reaching 75% performance. Analyzing the performance by question categories, figures 10 and 11 show that the temporal question category performed the highest, while people and object interactions, counting and common sense categories are more more difficult for both tasks. Finally, the multiple choice task showed improved performance in the causal category, despite no training videos being provided to teams on purpose to test systems on out of distribution performance. This may show the importance of providing optional answers in situations where no training data can be provided.

We should acknowledge a limitation in the categories of Audio and Multihop as well as counting and common sense as each pair was annotated by 1 human annotator due to limit of enough annotators per question type (i.e. 5 annotators for 7 question types). This made it hard to separate the analysis within each pair. We plan to avoid this problem in the next iteration by allowing annotators to label the question type they work on and if possible hire enough annotators.

Table 3 summarizes the answer-generation strategies used by participating teams, showing a mix of fine-tuned and off-the-shelf multimodal video–language models. Most teams relied on VideoLLaMA or Qwen-based models, while others used hybrid or proprietary solutions such as Gemini. Fine-tuning was commonly performed on task-specific or large-scale VQA datasets, with training times typically ranging from two weeks to several months, whereas a few teams used off-the-shelf models without additional training. Overall, the approaches reflect a trade-off

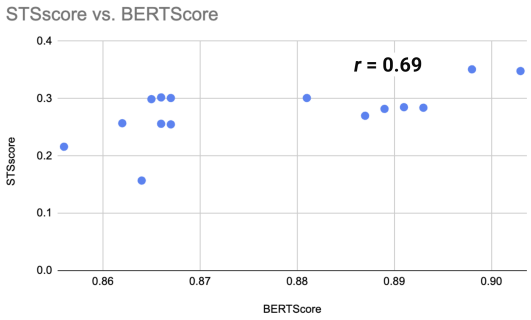

Figure 6. correlation between metrics

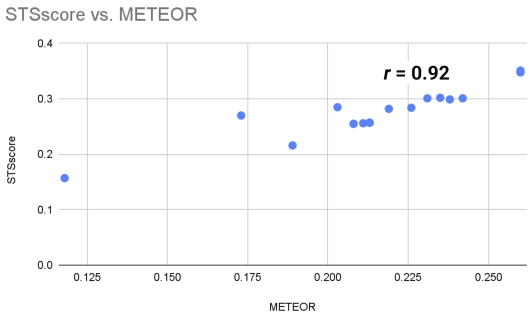

Figure 7. correlation between metrics

between extensive fine-tuning on diverse datasets and faster deployment using pre-trained models.

The table 4 shows that teams used a mix of fine-tuned and off-the-shelf multimodal models for the multiple-choice task, with development times ranging from two weeks to one month. VideoLLaMA3 and Qwen2.5-Omni-7B were the most common choices, often trained or evaluated on the provided training dataset. One team directly compared off-the-shelf and fine-tuned Qwen2.5-Omni-7B within the same two-week timeframe, indicating that fine-tuning does not necessarily increase turnaround time. Overall, the results suggest that both fine-tuned and off-the-shelf approaches are viable, with similar computational time. For more details on system approaches, please refer to specific team papers [4], [18], [3], [5], and [9]

7. Conclusion

Through its pilot year, this VQA benchmark introduced two complementary tasks, Answer Generation and Multiple Choice, that together capture both open-ended generation and ranking-based reasoning over video–language inputs. The results demonstrate that while recent multi-modal models have made measurable progress, video understanding remains a challenging problem, particularly for questions requiring higher-level reasoning, such as object–people interactions, counting, and common-sense inference.

The newly constructed dataset and annotation framework proved effective in supporting diverse reasoning categories, spanning temporal tracking, causal reasoning, audio–visual integration, and multi-hop inference. The analysis of system performance across question types highlights clear disparities in difficulty: temporal questions were consistently easier for current models, whereas categories demanding abstract reasoning or real-world knowledge exposed persistent weaknesses. These findings underscore the value of fine-grained, category-level evaluation for diagnosing model capabilities beyond aggregate scores.

The evaluation results also revealed that multiple-choice formulations remain substantially easier than open-ended answer generation, with higher overall accuracy and robustness between systems. At the same time, strong correlations among METEOR, STS, and BERTScore suggest that existing text-based metrics provide consistent signals for evaluating generated answers, though continued exploration of more video-aware evaluation methods remains an open direction. A plausible direction is hiring human assessors to examine automatic system outputs and rate them compared to ground truth answers. Measuring the correlation between human ratings and automatic metrics can provide high confidence in the quality of automatic metrics. The diversity of submitted approaches—ranging from off-the-shelf multi-modal models to fine-tuned systems, also indicates that competitive performance can be achieved under varying computational and development constraints.

Overall, the VQA pilot benchmark establishes a solid foundation for ongoing evaluation of video understanding systems. The lessons learned from dataset construction, task design, and participant feedback will directly inform future iterations of the evaluation. By expanding coverage, refining question categories, introducing new metrics such as LLM as a judge, adding baseline system, and continuing to challenge models with complex multi-modal reasoning scenarios, this VQA evaluation aims to drive sustained progress toward more reliable, interpretable, and human-like video question answering in upcoming years, including the planned 2026 iteration. Finally, by maintaining a dynamic growing dataset, we can track model progress over time by evaluating them on a fixed subset of data without disclosing the ground truth.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor,

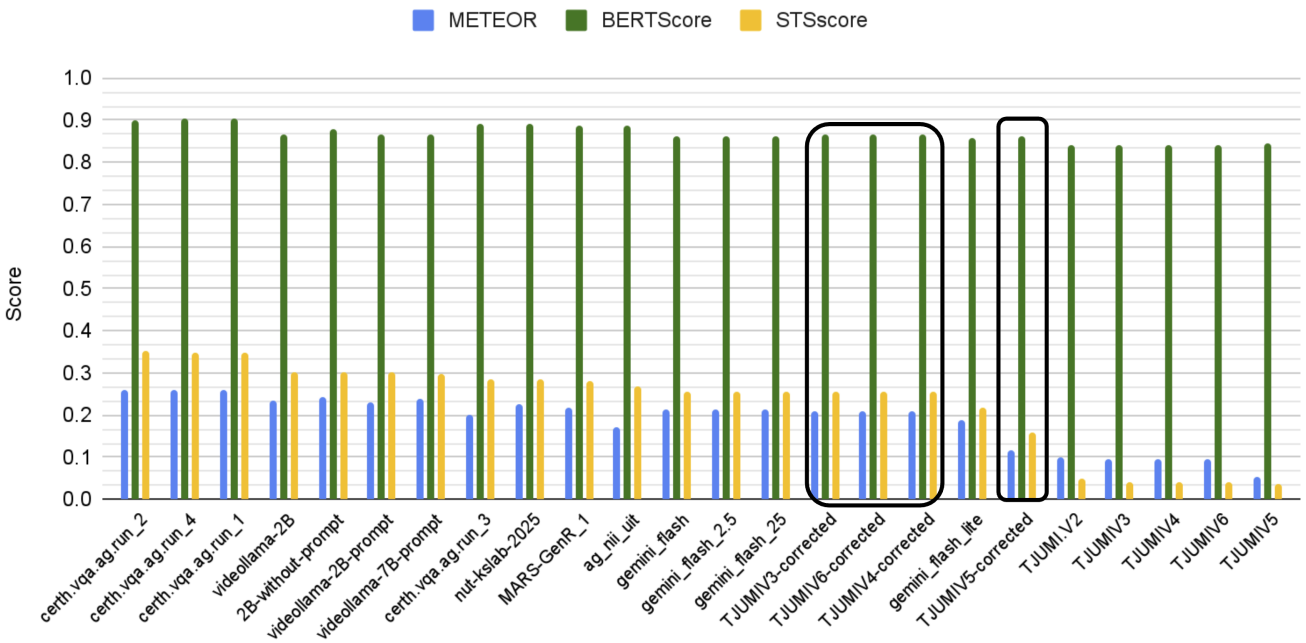

Figure 8. Answer Generation Task Results

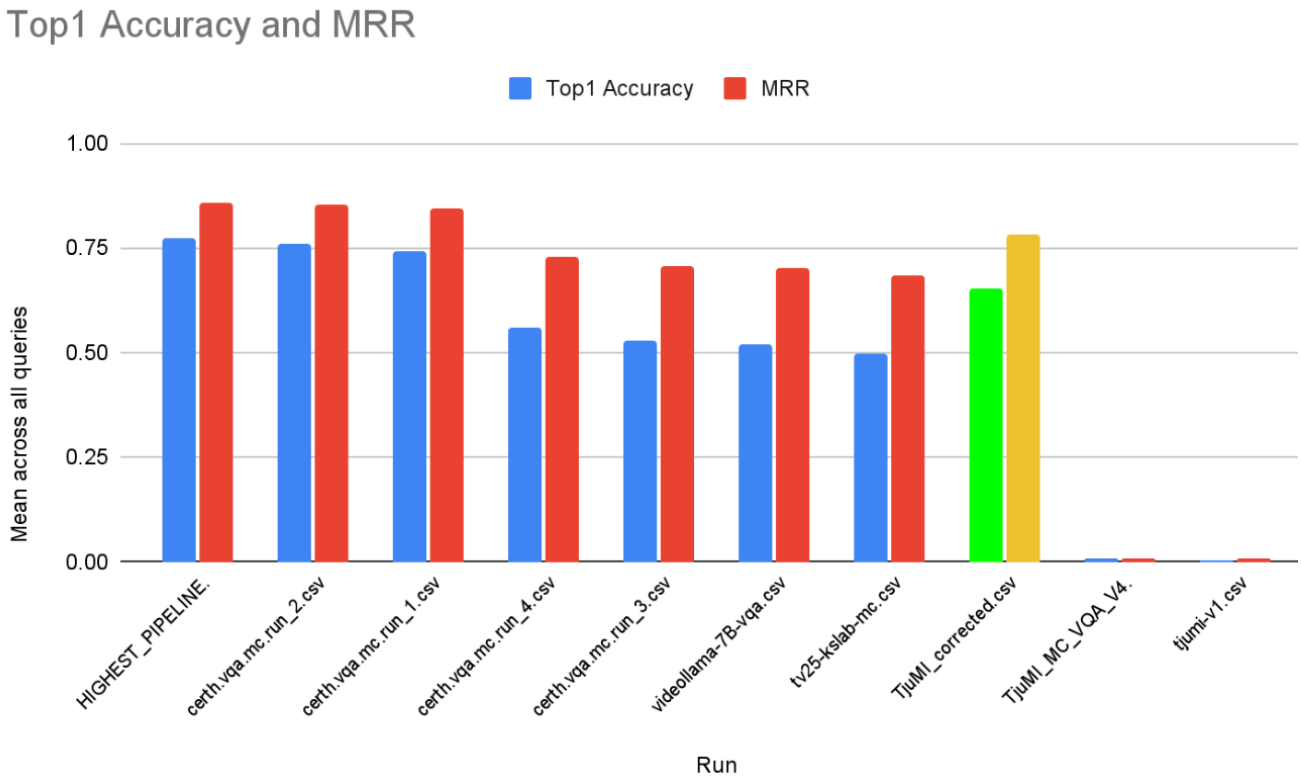

Figure 9. Multiple-choice task results. Run names are shown on the x-axis. The corrected TjuMI run is highlighted in a different color

Michigan, 2005. Association for Computational Linguistics. 2

[3] Fangyun Duan, Haixiang Ni, Xiusong Wang, and Chao Liang. Whu-nercms at trecvid2025: Ad-hoc video search

Team	Model(s) Used	Technique	Training Dataset	Training Time
TjuMI	VideoLLaMA2	Fine-tuned	ActivityNet-200, TRECVID 2024 VTT/VQA, InternVid-10M, others	6 months
WHU-NERCMS	VideoLLaMA (2B, 7B)	Off-the-shelf	VQA Provided Training data	3 weeks
CERTH-ITI	Qwen2.5-Omni-7B	Fine-tuned	Not specified	2 weeks
HLTCOE	Qwen 2.5 VL + Whisper	Fine-tuned	VQA Provided Training data	3 weeks
NII_UIT	Aria 8×3.5B + VideoLLaMA-7B	Off-the-shelf	None	-
kslab	VideoLLaMA3-2B	Fine-tuned	Model-specific dataset	∼1 month
tcna	Gemini Flash 2.5/1.5	Off-the-shelf	None	1 month

Table 3. Main Approaches (Overview) — Answer Generation Task

Team	Model(s) Used	Technique	Training Dataset	Training Time
kslab	Open-source VideoLLaMA3-2B	Fine-tuned	Custom video QA dataset curated for training	1 month
NII_UIT	InternVL-3.5 & Whisper-v3-large	Off-the-shelf	None	1 month
CERTH-ITI (run 1)	Qwen2.5-Omni-7B	Off-the-shelf	VQA Provided Training data	Two weeks
CERTH-ITI (run 2)	Qwen2.5-Omni-7B	Fine-tuned	VQA Provided Training data	Two weeks
WHU-NERCMS	VideoLLaMA3	Off-the-shelf	VQA Provided Training data	3 weeks

Table 4. Main Approaches (Overview) — Multiple Choice Task

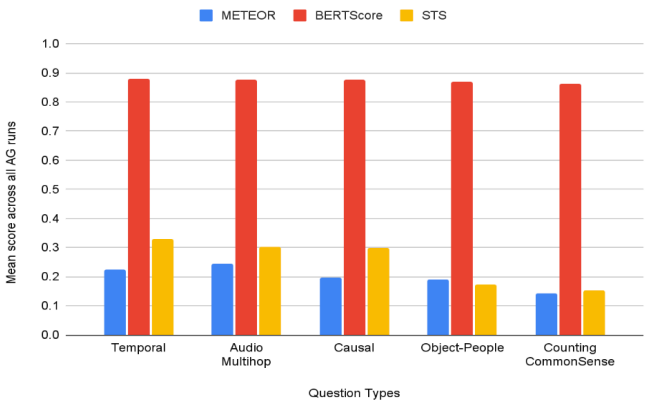

Figure 10. Results by question types (Answer Generation Task)

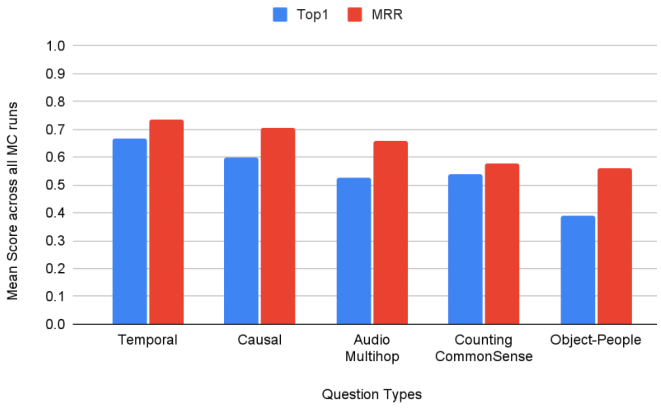

Figure 11. Results by question types (Multiple Choice Task)

(avs) and video question answering (vqa) task. In *Proceedings of The Thirty-Fourth Text REtrieval Conference (TREC 2025)*. National Institute of Standards and Technology (NIST), 2025. 6

[4] Andreas et al. Goulas. Mllm frame subset ensembling for audio-visual video qa... In *Proceedings of the Thirty-Fourth*

Text REtrieval Conference (TREC 2025), 2025. 6

[5] Nguyen Thanh Hai, Nguyen Minh Quan, Tran Minh Khang, and Le Tuan Anh. Nii-uit at trecvid 2025: Video question answering. In *Proceedings of The Thirty-Fourth Text REtrieval Conference (TREC 2025)*. National Institute of Standards and Technology (NIST), 2025. 6

[6] Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. Umbc ebiquity-core: Semantic textual similarity systems. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 44–52, Atlanta, Georgia, 2013. Association for Computational Linguistics. 2

[7] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2758–2766, 2017. 1

[8] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20, 2017. 1

[9] Jaehoon Kim, Seungmin Lee, and Kyunghyun Park. Kslab at trecvid 2025: Video question answering. In *Proceedings of The Thirty-Fourth Text REtrieval Conference (TREC 2025)*. National Institute of Standards and Technology (NIST), 2025. 6

[10] Dohwan Ko, Ji Soo Lee, Miso Choi, Jaewon Chu, Jihwan Park, and Hyunwoo J Kim. Open-vocabulary video question answering: A new benchmark for evaluating the generalizability of video question answering models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3101–3112, 2023. 2

[11] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1369–1379, 2018. 1

[12] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 2

[13] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 2

[14] Yisong Wang, Liwei Wang, Yiming Yu, Di He, and Wei Chen. A theoretical analysis of NDCG-type ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013)*, pages 25–54, 2013. 2

[15] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58. Springer, 2022. 1

[16] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12): nwae403, 2024. 1

[17] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI conference on artificial intelligence*, pages 9127–9134, 2019. 1

[18] Dengjia Zhang, Charles Weng, Katherine Guerrerio, Yi Lu, Kenton Murray, Alexander Martin, Reno Kriz, and Benjamin Van Durme. Hltcoe evaluation team at trec 2025: Vqa track. In *Proceedings of The Thirty-Fourth Text REtrieval Conference (TREC 2025)*. National Institute of Standards and Technology (NIST), 2025. 6

[19] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[20] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 3

[21] Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*, 2022. 1