Video Question Answering (VQA) 2025 Track

George Awad
National Institute of Standards and Technology (NIST)
USA
george.awad@nist.gov

Sanjay Purushotham
University of Maryland, Baltimore County
USA
psanjay@umbc.edu

Afzal Godil
National Institute of Standards and Technology (NIST)
USA
afzal.godil@nist.gov

Abstract

Recent advancements in large multimodal 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. The Video Question Answering (VQA) Challenge aims to rigorously assess the capabilities of state-of-the-art multimodal models in understanding and reasoning about video content. Participants developed and tested models that answer a diverse set of video segments-based questions covering various levels of complexity, from factual retrieval to complex reasoning. The challenge track serves as a critical evaluation framework to measure progress in video understanding, helping identify strengths and weaknesses in current multimodal AI architectures. By fostering innovation in multimodal learning, this track contributes to advancing AI's ability to process dynamic visual narratives, enabling more reliable and human-like interaction with video-based information. The track completed its first pilot year, which included two subtasks: *Answer Generation Task* and *Multiple Choice Task*. Based on lessons learned and participant feedback, we plan to run the track again in 2026.

Keywords

Video Question Answering (VQA), Large Language models (LLM), Multimodal Language Models

ACM Reference Format:
George Awad, Sanjay Purushotham, and Afzal Godil. 2025. Video Question Answering (VQA) 2025 Track. In *Proceedings of the TREC 2025, December 2025, NIST, Maryland, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The Video Question Answering (VQA) [8] track 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 image-based Visual Question Answering task [1, 4, 6], Video Question Answering systems must capture temporal dependencies, motion cues, and multi-frame context, making it a rich testbed for multimodal reasoning. The track 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.

2 Description of the VQA Track Tasks

2.1 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), the objective is as follows:

For each video, automatically generate up to ten textual answers to the associated question 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.

Evaluation: Systems are evaluated against human-annotated ground truth using established text-generation metrics, including **METEOR** [2], **BERTScore** [7], and **Semantic Textual Similarity (STS)** [3]. To assess the quality of the ranking itself, **Normalized Discounted Cumulative Gain (NDCG)** [5] is employed.

2.2 Multiple Choice (MC) Task

Accurately identifying the most appropriate answer to a question about short video content is a key challenge in multimodal 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 dialogue 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, the objective of this task is to develop a system that can *automatically rank the candidate answers* for every question. Specifically, for each video the system

TREC '25, Gaithersburg, Maryland, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

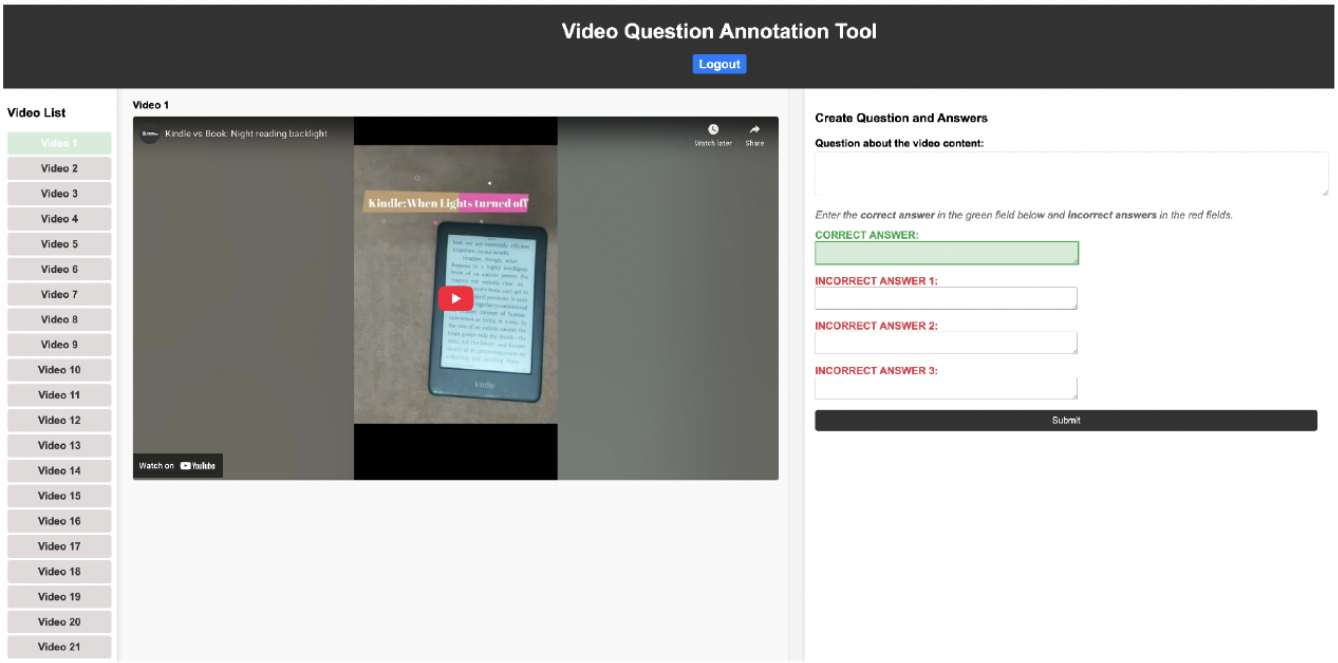

Figure 1: Annotation Tool Interface

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.

Evaluation: System performance is assessed using established ranking metrics, including **Top-1 accuracy**, which measures the proportion of questions for which the top-ranked answer is correct and **Mean Reciprocal Rank (MRR)**, which captures the average inverse rank of the first correct answer across all test instances.

3 Datasets and Annotation Framework

The dataset for this track was constructed this year through a large-scale annotation effort designed to support both the initial evaluation and future reuse. Approximately **80,000 YouTube Shorts links** were collected and randomly distributed among five human annotators hired by NIST, with no overlap in assigned videos. 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.

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 at least one of seven predefined question categories (see section 3.1); 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 (see samples in the appendix).

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, 2025 test, and planned 2026 test subsets. To enable out-of-distribution evaluation, no training data were released for the "Causal Question" category (all question categories are described in section 3.1).

Data Maintenance and Limitations: Because the dataset relies on publicly available YouTube content, some video links may become inaccessible—for example, if an uploader removes a video. We will maintain and distribute an updated 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 retrieval collection that supports long-term research while acknowledging the inherent limitations of external video hosting.

3.1 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, multimodal integration, or real-world knowledge. By organizing annotation guidelines into the seven categories shown in Table 4, 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

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
Train	125	0	125	125	125	500
2025 Test	400	400	400	400	400	2000
2026 Test	121	794	180	164	198	1457

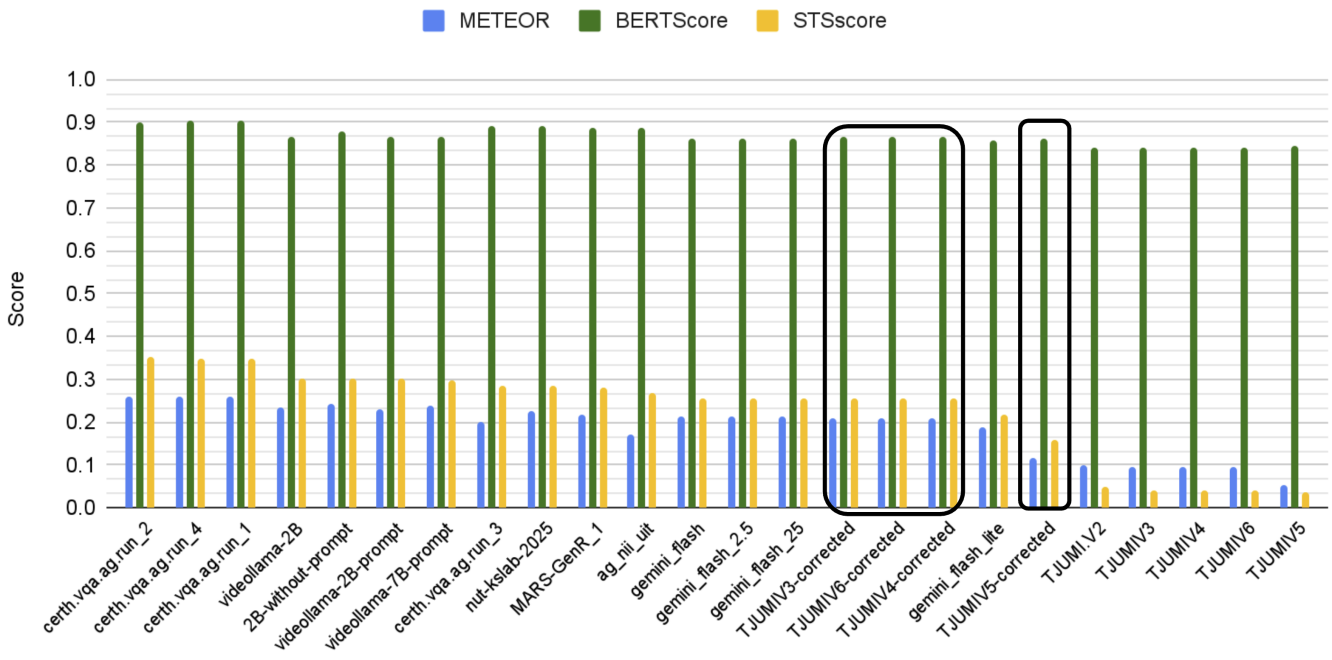

Figure 2: 2025 pilot task results (Answer Generation Task)

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.

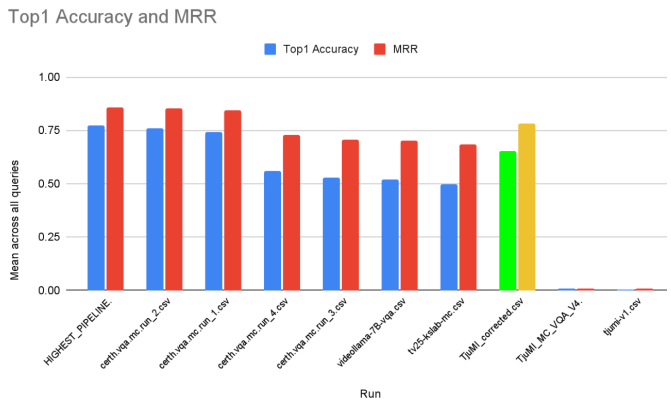

Figure 3: 2025 pilot task results (Multiple Choice Task)

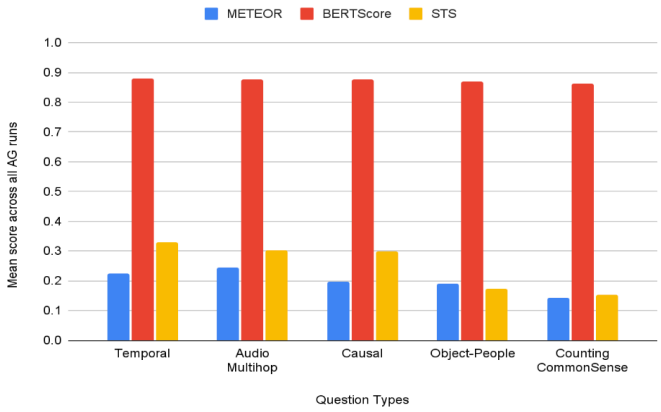

Figure 4: 2025 pilot task results by question types (Answer Generation Task)

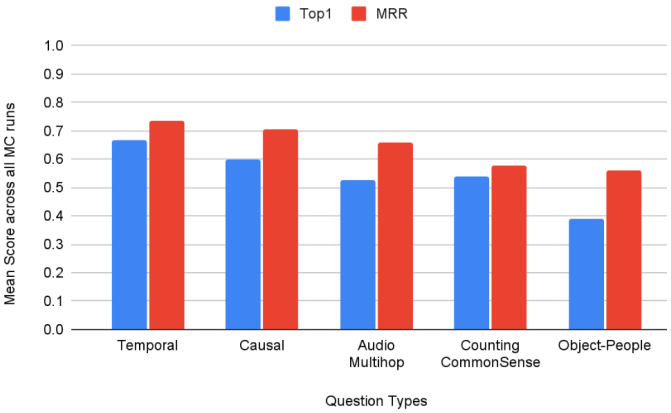

Figure 5: 2025 pilot task results by question types (Multiple Choice Task)

4 2025 Pilot year Participants, Runs, and Results

In 2025, the first year of the track, 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 2 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 snippets. 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 green boxes in figure 2 and with different colors in figure 3.

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).

Figure 3 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 4 and 5 show that the temporal question category performed the highest, while people and object interactions, counting and common sense categories are more harder(or 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.

Table 2 summarizes the answer-generation strategies used by participating teams, showing a mix of fine-tuned and off-the-shelf

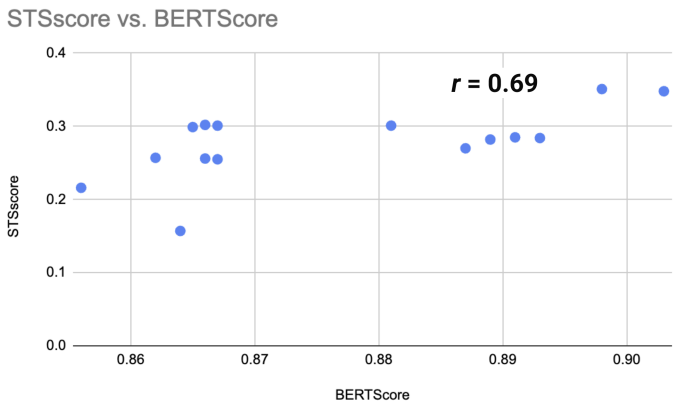

Figure 6: correlation between metrics

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 one month, whereas a few teams used off-the-shelf models without additional training. Overall, the approaches reflect a trade-off between extensive fine-tuning on diverse datasets and faster deployment using pre-trained models.

The table 3 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 TRECVID2025 VQA 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.

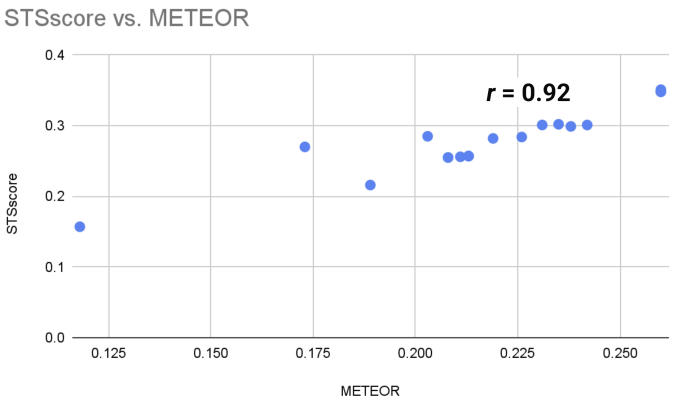

Figure 7: correlation between metrics

5 Conclusion

The TREC 2025 Video Question Answering (VQA) track represents a first, important step toward rigorous and large-scale evaluation of multimodal systems' ability to understand and reason over short-form video content. Through its pilot year, the track 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 multimodal 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. The diversity of submitted approaches—ranging from off-the-shelf multimodal models to fine-tuned systems—also indicates that competitive performance can be achieved under varying computational and development constraints.

Overall, the VQA 2025 pilot 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 track. By expanding coverage, refining question categories, and continuing to challenge models with complex multimodal reasoning scenarios, the VQA track aims to drive sustained progress toward more reliable, interpretable, and human-like video question answering in upcoming years, including the planned 2026 run .

Disclaimer: Certain commercial equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

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.

[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, Michigan, 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

[3] 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. doi: 10.3115/v1/S13-1005. URL https://aclanthology.org/S13-1005.

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

[5] 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. URL https://arxiv.org/abs/1304.6480.

[6] 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.

[7] 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. URL https://arxiv.org/abs/1904.09675.

[8] 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.

Team	Model(s) Used	Technique	Training Dataset	Training Time
TjuMI	VideoLLaMA2	Fine-tuned	ActivityNet-200, TRECVID 2024 VTT/VQA, InternVid-10M, others	Since March
WHU-NERCMS	VideoLLaMA (2B, 7B)	Off-the-shelf	TRECVID 2025 VQA	3 weeks
CERTH-ITI	Qwen2.5-Omni-7B	Fine-tuned	Not specified	2 weeks
HLTCOE	Qwen 2.5 VL + Whisper	Fine-tuned	NIST dev 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 2: Main Approaches (Overview) — Answer Generation Task

Team	Model(s) Used	Training Technique	Training Dataset	Time Taken
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	TRECVID2025 VQA train dataset	Two weeks
CERTH-ITI (run 2)	Qwen2.5-Omni-7B	Fine-tuned	TRECVID2025 VQA train dataset	Two weeks
WHU-NERCMS	VideoLLaMA3	Off-the-shelf	TRECVID2025 VQA train dataset	3 weeks

Table 3: Main Approaches (Overview) — Multiple Choice Task

Appendix A: VideoQA Question Categories

Table 4: Question categories for the Video Question Answering (VideoQA) task, with descriptions and representative examples.

Category	Description	Examples
Cat 1: Temporal / Order / Attribute Tracking	*Temporal*: Questions requiring understanding of the timing or sequence of events. *Time order*: Questions requiring reasoning about the order of actions or events. *Attribute change*: Questions requiring observation of changes over time (state, appearance, location/position, emotion/facial expression, scene/environment).	*Temporal*: (After / before + how/why/what): **When** does the boy enter the room? What happened **after** the man drank the juice? **Before** the woman got into the store, who talked to her? **Why** did the young man fall down after eating breakfast? **How** did the girl prank her dad after leaving the house? **What** did the man give the lady before she paid him the money? *Time order*: (order of events) What happens **after** the man opens the door? Which event occurred **first**: the car crash or the siren sound? What is the **correct order** of actions performed by the chef? *Attribute change*: What **happens** to the balloon ***by the end*** of the video? What does the cup **look** like ***after*** the person pours the drink? **Where** does the cat move ***after*** jumping off the table? How does the room look ***after*** the party **ends**?
Cat 2: Causal / Plot / Goal-Oriented Reasoning	Questions requiring reasoning about causes, motivations, or overall plot/theme.	**Why** did the woman start running? **What caused** the boy to fall down? What is the video mainly about? What lesson does the video try to convey? What is the video trying to show about teamwork? What was the effect of <AAAA> on <BBBB>? What is the main challenge faced by the old man?
Cat 3: Audio / Multimodality	Questions relying on spoken words, music, sound effects, or on-screen text.	What does the man say at the end of the video? What did the old woman tell the boy when he greeted her? What kind of music does the young girl listen to? What is written on the blackboard behind the professor?
Cat 4: Multi-hop Reasoning	Questions that require combining multiple observations or intermediate inferences.	What does the boy do after picking up the ball and seeing the dog? What clues suggest the woman is planning a surprise?
Cat 5: Object–People Interactions	Questions involving interactions between people and objects or animals.	What does the man do with the suitcase? What object does the child hand to the woman?
Cat 6: Counting / Quantitative	Questions involving counting or comparing numbers in the video (possibly with temporal aspects).	How many people entered/left the room? How many cats are there in this video? Are there more cats than dogs?
Cat 7: Common-Sense & World Knowledge Inference	Questions that require everyday reasoning implied but not explicitly shown.	Why is the person wearing a raincoat indoors? (They are about to go outside) Why does the man check the oven multiple times? (He's worried the food might burn or overcook)

Appendix B: Sample annotations

Sample 1

Why does the boy pick up his toys and run off down the street?

1- He thinks that the person sitting next to him broke the parts off his toys.
2- Both of the toys he has lose a part, so he picks them up and runs away.
3- He wants to find a new place to play.
4- He wants to throw away the toys that he thinks are broken.

Sample 2

What happens when the yellow pool ball is hit ?

1- It sinks in the adjacent side pocket
2- other players sink balls
3- The shooter gets a pat on the back
4- Bystanders comment

Sample 3

how many bouquets of flowers are in the video?

1- two
2- three
3- one
4- none