Nagaoka University of Technology at TREC 2025

Video Question Answering

Isabel Gonzalez, Shungo Kubosaka, Takashi Yukawa

Abstract

This paper details our approach to two tasks in Video Question Answering (VQA) for TREC 2025 challenge, using VideoLLaMA3-2B as the base model. For both the Answer Generation (AG) task and Multiple Choice (MC) task, the primary training data was the dataset provided by TREC, which was used to finetune the model using LoRA. For the AG task, the approach we used was a methodology to generate diverse answers using sampling and ranked them using the model's average log-probability, which proved effective with a NDCG_BERT score of 0.993. Our generated answers were found to be semantically similar (BERT score: 0.893) but lexically different (METEOR: 0.226). We also identified a potential bias in confidence-based ranking that favors shorter answers. For the MC task, our approach was a two-stage process where the model first generates a "ground truth" answer via greedy decoding, which is then used by a Sentence-Transformer to rank the given options based on cosine similarity. This approach achieved a Top1 Accuracy of 0.499 and a Mean Reciprocal Rank (MRR) of 0.686. The system's effectiveness depended on both the accuracy of the "ground truth" generation and the Sentence-Transformer's similarity measurement. We learned this "generate-then-compare" strategy is viable, but its main limitation is error propagation from the first step. This paper outlines these methods, our experimental findings, and their limitations.

1. Introduction

The Video Question Answering (VQA) track refers to the challenge of a system answering questions about a video, interpreting visual, auditory, and temporal information. The track is divided into two tasks, demands that systems handle open-ended generation for the Answer Generation (AG) task, where answers must be both accurate and concise, as well as discriminative tasks that require ranking pre-defined options in the Multiple Choice (MC) task.

In this paper, we present our submission to the TREC 2025 VQA challenge, focusing on the two tasks: Answer Generation (AG) and Multiple Choice (MC). Our system is built upon the VideoLLaMA3-2B[1] architecture, a state-of-the-art Large Multimodal Model (LMM). We fine-tuned this foundation model using Low-Rank Adaptation (LoRA)[2] on the provided TREC datasets to adapt it specifically to the competition's formatting and domain requirements.

2. Methodology

2.1. Base Model and Finetuning

For both the Answer Generation and Multiple Choice tasks, we employed VideoLLaMA3-2B [1] as our base model, a vision-language model built on the Qwen2 architecture with integrated video understanding capabilities through the SigLIP vision encoder. To adapt the model to the VQA task format

and the TREC dataset domain, we applied Low-Rank Adaptation (LoRA) [2] for parameter-efficient finetuning.

LoRA enables efficient finetuning by introducing trainable low-rank decomposition matrices to selected weight matrices while keeping the original pre-trained weights frozen. We applied LoRA adapters to all attention projections (query, key, value, and output) and feed-forward network layers (gate, up, and down projections) across all transformer layers. The LoRA configuration used a rank of $r = 128$ with a scaling factor of $\frac{\alpha}{r} = 2.0$ and dropout rate of 0.05, resulting in approximately 11% trainable parameters relative to the targeted layers. The model was trained for 10 epochs with a cosine annealing learning rate schedule.

2.2. Task 1: Answer Generation (AG)

Our primary goal for the AG task was to generate as many different and plausible answers as possible for a given video and question. We initially attempted to prompt the model to generate a fixed number of answers (e.g., five or ten). This approach proved ineffective, as the model failed to return concise answers. We then shifted to a sampling-based approach. By running the same prompt for a given video and question multiple times with sampling enabled, we could ensure the model would return different answers. We experimented with different parameters, particularly the temperature, to determine which setting yielded the best and most diverse set of plausible answers.

Given a set of generated answers, the next challenge was to rank them. With only the model's outputs to work with, we decided to utilized the model's confidence scores. The VideoLLaMA3-2B model provided sequence transition scores for its generated tokens. We explored using the average log-probability and the loss function to rank the full answer sequences. Based on tests with our small dataset, the average log-probability of the transition scores gave the best ranking results.

2.3. Task 2: Multiple Choice (MC)

For the MC task, the objective was to rank a given set of answer options. Our first approach was to input the video, question, and all answer options into the model and prompt it to rank them. This method was unreliable. The model could often identify the best answer and possibly rank two of the remaining three, but it frequently failed to correctly place the last remaining option. We devised a more robust, two-step method:

(1) **"Ground Truth" Generation:** We first prompt the model (VideoLLaMA3-2B) to generate its own best answer to the question using greedy decoding. The best results were achieved with no sampling and a beam size of 2. This generated answer served as our "ground truth" for ranking.

(2) **Similarity-Based Ranking:** We then used a Sentence-Transformer[3] model. This model encodes the "ground truth" answer and each of the provided answer options into high-dimensional vectors. We then calculate the cosine similarity between the "ground truth" vector and each answer option vector. The options are then ranked based on this similarity score, with a higher cosine similarity indicating a better rank.

3. Experiments and Results

3.1. Datasets and Evaluation Metrics

The model was trained and evaluated using the dataset provided by TREC. Performance for the AG task was measured using METEOR, BERT score, and Normalized Discounted Cumulative Gain (NDCG) based on both METEOR and BERT (NDCG_METEOR and NDCG_BERT). Performance for the MC task was measured using Top-1 Accuracy and Mean Reciprocal Rank (MRR).

3.2. Answer Generation (AG) Results

The results from the AG task are summarized below:

RunID	METEOR	BERTScore	STSscore	NDCG_METEOR	NDCG_BERTScore
nut-kslab-2025	0.226	0.893	0.284	0.832	0.993

Table 1. Results of the submitted run for the AG task

The low METEOR score alongside a high BERT score indicates that our generated answers were lexically different from the ground truth but were semantically very similar. The high NDCG scores for both metrics, particularly NDCG_BERT, prove that our ranking strategy based on average log-probability was highly effective.

3.3. Multiple Choice (MC) Results

The results from the MC task are summarized below.

RunID	Top-1 Accuracy	MRR
tv25-kslab-mc	0.499	0.686

Table 2. Results of the submitted run for the MC task

These results indicate that our system correctly identified the top answer approximately half the time (0.499). Furthermore, the strong MRR score of 0.686 suggests that even when the correct answer was not ranked first, it was consistently placed near the top of the list, validating the effectiveness of using cosine similarity with a generated ground truth for ranking.

4. Discussion and Limitations

An important and unexpected finding was that the confidence-based ranking proved questionable at times. We identified a potential problem where the answer's length affects the score. This suggests that the model's token-level confidence calculation might intrinsically favor shorter, more concise answers, regardless of whether a longer, more descriptive answer is equally or more accurate in certain contexts. This warrants further investigation into alternative confidence metrics.

The primary limitation of our MC task method is its dependency on the first step. If the model makes a mistake in generating its "ground truth" answer, that mistake is then carried over and propagated to the next step, inevitably leading to an incorrect ranking of the answer options.

Finally, a critical limitation of our current system is its exclusion of auditory information. The model used, as configured for these experiments, processes only the visual frames of the video, completely disregarding the audio track. Consequently, the system is unable to answer questions that rely on spoken dialogue, sound effects, or environmental audio cues. This inevitably lowers performance on a subset of the data where the answer is contained in the audio rather than the visual stream. Future work must address this by integrating an audio encoder to enable true multimodal understanding.

5. Conclusion

This paper described our submission to the TREC 2025 VQA track. We proposed specific methodologies for both open-ended Answer Generation and discriminative Multiple Choice tasks. Our findings demonstrate that Large Multimodal Models, when finetuned with LoRA, can serve as powerful baselines for complex video reasoning tasks. For the Answer Generation task, we showed that a sampling-based generation strategy combined with confidence-based ranking is highly effective. The high NDCG scores (0.9927 for NDCG_BERT) confirm that the model's average log-probability is a reliable indicator of answer quality, despite the potential bias towards shorter sequences. For the Multiple Choice task, our results validated the efficacy of a "generate-then-compare" pipeline. By using the model's own generation as a semantic anchor, we achieved a Top-1 accuracy of 0.499 and an MRR of 0.686. However, the system's performance remains bound by the accuracy of the initial generation; error propagation remains a key challenge to address. The current exclusion of auditory information restricts the system's understanding of videos where sound is a primary information channel. Future iterations of this work will prioritize the integration of audio encoders and the exploration of other methods to mitigate the specific biases and error propagation issues identified in this study.

References

1. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., Jin, P., Zhang, W., Wang, F., Bing, L., & Zhao, D., “VideoLLAMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding”, arXiv:2501.13106, 2025.
2. Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. “LoRA: Low-Rank Adaptation of Large Language Models”, arXiv: 2106.09685, 2021
3. Reimers, N., & Gurevych, I. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", arXiv:1908.10084, 2019.