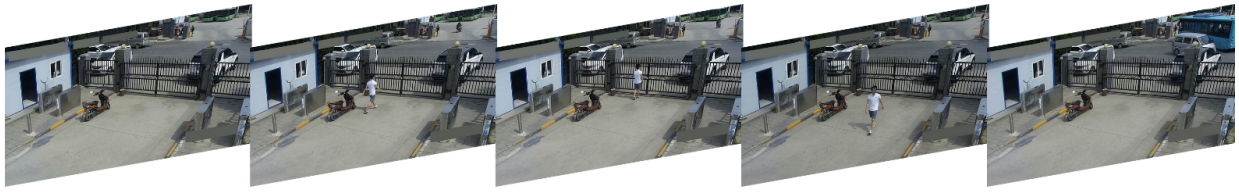
Moraqeb: A Distilled VLM on Surveillance VQA

Anas Aldadi[1], Raghad Alsuhaibani[1], Abdulelah Alturki[1],
Anas Hakami[1], Joori Alsaif[1], Haifa Almakhdhoub[1],
Dr. Tanveer Hussain[2]

[1] KAUST — King Abdullah University of Science and Technology, Saudi Arabia
[2] Edge Hill University, Ormskirk, United Kingdom

https://github.com/AnasKAN/Muraqeb



Summary	Generic	Temporal	Spatial	Reasoning	Short temporal
What happens throughout the entire video, including all key actions and events?	Describe the appearance and attire of the man in the video.	How does the video conclude, and what is the final view shown?	What is the color of the man's jacket and helmet?	Why does the man pass by the same vehicles multiple times in the video?	When does the man in the yellow jacket and red helmet first appear in the video?
Answers					
The video starts with a view of a gated area, likely an entrance to a parking lot or a restricted zone, featuring black metal pillars with yellow and white stripes. A red and black scooter is parked near the gate, and to ...	The man in the video is wearing a yellow jacket and a red helmet. This distinctive combination of clothing makes him easily identifiable as he walks past ...	The video concludes with a static view of the gated area, which is the same as the opening scene. The final view shows the black metal pillars with yellow and white stripes, the red and black scooter ...	The man is wearing a yellow jacket and a red helmet.	The man passes by the same vehicles multiple times because he reappears and walks across the street again, following the same path.	The man in the yellow jacket and red helmet first appears in the video after a moment, walking across the street from right to left and passing by the white car and the green bus.

Figure 1: Test Samples from Surveillance VQA, Containing 6 types of questions and we demonstrate Muraqeeb answers to those questions

Abstract

This paper presents Moraqeb, a distilled vision-language model (VLM) designed for efficient deployment in the domain of video surveillance visual question answering (VQA) scenarios. Although large-scale VLMs have shown promising results in multimodal understanding tasks, their computational overhead makes them unsuitable for time-sensitive or resource-constrained environments. We identify key challenges in the current literature, including the lack of lightweight VLMs tailored for surveillance video data and limited support for varying video resolutions. In response, we distill a state-of-the-art VLM (SmolVLM) [4] on the SurveillanceVQA-589K dataset [3] to improve accuracy of the system in 6 different types of Video Surveillance VQA, inference speed, and reduce memory requirements. Moraqeb achieved 36.3% accuracy in MVBench, a 6.6% gain over the original student model SmolVLM, and enhances the model video scene and action understanding, achieving 41.6% improvement in action antonym tasks. Our approach prioritizes model size, latency, and performance trade-offs, with the aim of making VQA capabilities in the domain of video surveillance more accessible to a wider user base. This work serves as a step toward scalable and practical multimodal systems for video surveillance question answering.

1 Introduction

Recent advances in vision-language models (VLMs) have shown great understanding and reasoning capabilities about visual content using natural language. These models combine computer vision and natural language processing techniques, allowing them to perform complex tasks such as visual question answering (VQA). However, most state-of-the-art VLMs are large and computationally demanding, making them difficult to use for many users where they have limited resources.

The goal of this project is to develop a distilled VLM capable of performing video-based VQA tasks in surveillance settings. We begin by surveying the literature on VLM distillation and surveillance VQA, identifying gaps related to scalability, deployment, and generalization across video quality. Building on these insights, we propose Moraqeb a compact, efficient VLM distilled from larger models and fine-tuned on subset of SurveillanceVQA-589K dataset [3]. By reducing model size and inference latency, Moraqeb aims to bridge the gap between academic research and practical deployment in surveillance VQA applications. Our preliminary experiments lay the groundwork for future development and evaluation of distilled multimodal models in dynamic environments.

2 Methodology

Distillation is any technique that is concerned to transfer knowledge of large, complex model "Teacher Model" to a smaller, more efficient model "Student Model".

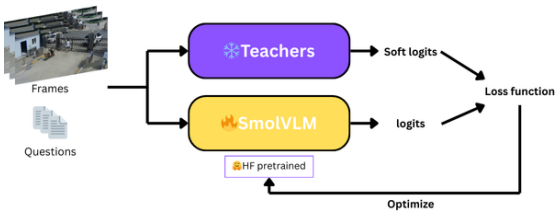


Figure 2: Offline distillation setup.

In 2 demonstrate offline distillation setup which we used in this project. A simple scheme where soft targets from the teacher model are extracted, meaning that the teacher model is frozen, and the student model learns from those soft targets.

2.1 Distillation Strategy:

Domain specific offline knowledge distillation is the framework we adopted, where we extract soft targets from the teacher model for the smaller student to train on. Thus, the student tries its best to mimic the teachers' answers. The student is designed for efficiency, targeting reduced parameter count and lower inference latency while learning on large scale amount of data.

2.2 Dataset Selection

We used the SurveillanceVQA-589K dataset [3], which includes 31,548 video clips with textual annotations and 589,380 QA pairs. The dataset is a combination of four datasets: MSAD, MEVA, NWPU, and UCA [8, 2, 1, 6]. Video clip annotations are created through both human labeling and LLM-based annotation. The QA pairs are defined under six types for normal videos: Summary, Generic, Temporal, Short Temporal, Spatial, and Reasoning QA. For abnormal videos, another six types are used: Detection, Classification, Temporal, Description, Cause, and Result

QA. These different categories help the model generalize to a wider range of tasks within the question answering field.

We specifically used the NWPU dataset within the SurveillanceVQA-589K collection [3]. The original NWPU dataset [1] only contained videos and a classification of whether a given video included abnormal actions or not. However, in the SurveillanceVQA-589K dataset, the NWPU subset has been modified and extended to also include human-labeled and LVLM-generated annotations, along with LLM-generated QA pairs. LLaVA-Video7B-Qwen2 was used to analyze each video and generate descriptions, while Qwen-Max was used to generate question-answer pairs.

2.3 Dataset Structure

The original NWPU dataset only contains videos and normal/abnormal classifications of each video. The SurveillanceVQA-589K dataset [3] improves and extends the original dataset by also including captions and QA pairs for each video.

The captions are created using a hybrid of human-labeled annotations and LVLM-generated annotations [3]. The LVLM captions are produced using LLaVA-Video7B-Qwen2, after which the human-labeled annotations and the LVLM annotations are integrated using Qwen-Turbo to form the final combined annotations. Qwen-Max is then used to generate the QA pairs. for each video according to original paper there are 12 different types [3].

We used the QA pairs in SurveillanceVQA-589K, along with the videos in the NWPU dataset, to train our student model in an offline distillation scheme, ensuring that the data it learns is generated from the teacher models.

In this context, when we refer to the "teacher models," we mean the ensemble of large vision-language models (LLaVA-Video7B-Qwen2, Qwen-Turbo, and Qwen-Max) [7, 5, 3] that collectively generated the captions and QA pairs in SurveillanceVQA-589K. While no single teacher model directly supervises the student, the outputs of this multi-model pipeline serve as the effective teacher signals for knowledge distillation.

Loss Function: The training uses the standard token-level cross-entropy loss for conditional text generation, where the model predicts the next token in the target sequence given the multimodal input (video frames + text prompt), and the loss is averaged over non-padding tokens.

3 Experiments

Setup: We trained and evaluated multiple distilled versions of SmolVLM [4] and then ensembled those finetuned models to create Moraqeb. on standard video VQA tasks using SurveillanceVQA-589K. Models are trained on L4 GPU and hyperparameters tuned for efficiency.

For full finetuned and QLoRA training we used the same settings which are the following, batch size of 2, learning rate 1e-4, optimizer AdamW, weight decay 0.01, and warmup steps 50.

Baselines: The base SmolVLM was considered to be the baseline since we need to improve upon it.

Metrics: We reported model size, inference latency, memory usage, and accuracy across 15 different tasks, reflecting both performance and reliability of the model.

4 Results and Discussion

Performance & Efficiency: Moraqeb achieves a massive increase in benchmark results compared to the base model and reducing parameter count significantly compared to the teacher models. Inference time is improved by a substantial margin, making deployment cheaper.

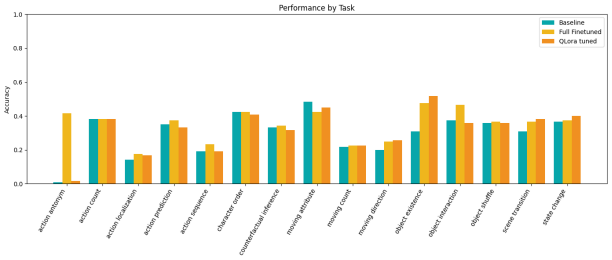


Figure 3: Comparison of accuracies across different tasks.

Analysis: As the figure 3 demonstrate, our full finetuned model have shown a huge jump in accuracy in the following task "action antonym", "object existence", and "object interaction", which indicates that the original base model haven't seen any example of it.

Also our Full finetuned, and QLora models have shown thorough and detailed answers compared to the answers of the base model, the following table shows the answer of our models against the base model.

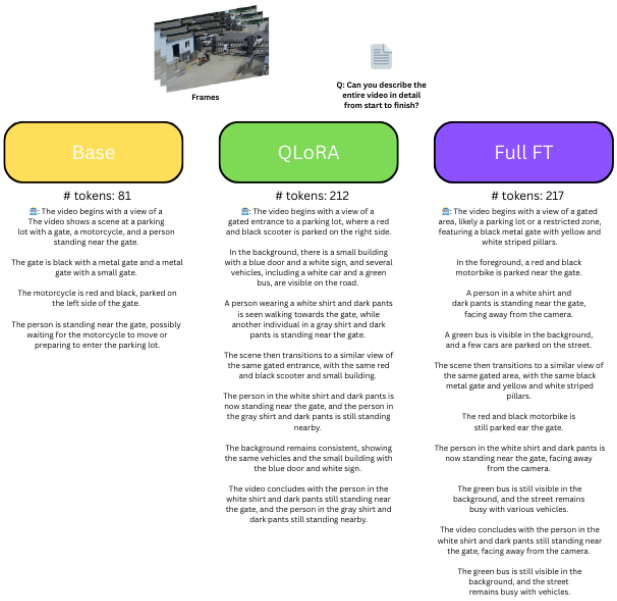


Figure 4: Comparison of responses from our models against the base model for the same prompt.

As the figure 4 illustrates, our models are much more descriptive than the original base model, and also provide more accurate, detailed answers.

Limitations: Challenges remain in generalizing across varying video quality, and night-time surveillance.

5 Conclusion and Future Work

The substantial improvement Muraqeb achieved suggest that exploring different distillation techniques is a promising approach for reducing VLM sizes, GPU memory requirement, and increasing latency. This will make many video VQA tasks be cheaper, and more accessible enabling much more users to utilize these tools.

5.1 Future directions

In this subsection we lay the foundation to build upon our work. include .

Surveillance Video Understanding: After error analysis we concluded that training on much more than a quarter of an epoch will substantially enhance models' performance on various tasks specially on poorly performed tasks such as action antonym and other tasks shown in error analysis subsection. Also we foresee exploring continual learning for dynamic surveillance environments will show intersting results.

Distillation Techniques: We strongly suggest trying different distillation techniques other than offline distillation to yield better student model. Literature proves this by showing when each technique is powerful and in our case online distillation is better, but due to resource constraints we were unable to perform such technique.

Loss function: Since the student model is built on top of huggingface class 'AutoModelForImageTextToText' and since it behaves similar to 'AutoModelForSeq2SeqLM' where:

- If labels are provided, the model computes the token-level cross-entropy loss between the predicted logits and the ground-truth label tokens.
- Padding tokens (usually -100 after preprocessing) are ignored in the loss.

The used vanilla loss function is the standard token-level cross-entropy loss, and we suggest to explore a customized loss function to influence the student model more on teachers' logits.

References

[1] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation, 2023.

[2] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection, 2020.

[3] Bo Liu, Pengfei Qiao, Minhan Ma, Xuange Zhang, Yinan Tang, Peng Xu, Kun Liu, and Tongtong Yuan. Surveillancevqa-589k: A benchmark for comprehensive surveillance video-language understanding with large models, 2025.

[4] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models, 2025.

[5] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[6] Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset, baselines, and challenges, 2023.

[7] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2025.

[8] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

6 Appendix

6.1 Training Details

Additional details about the training will be shown and discussed here. All distillation runs we attempted had different number of steps which are shown in the following learning curve diagram, the reason behind this is different Parameter Efficient Fine Tuning (PEFT) techniques lead to different training speeds. Given the previous information it justify why QLoRA requires more number of steps than full FT runs to reach convergence.

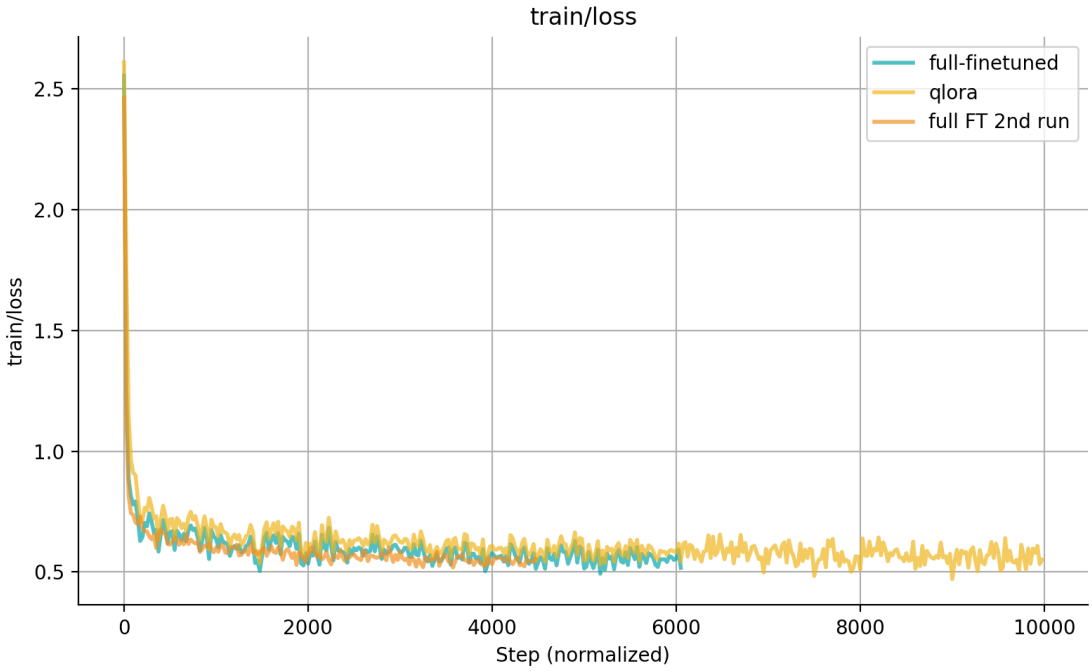


Figure 5: Training Loss Comparison between Full Fine-Tuning and QLoRA.

As figure 5 illustrate, almost all 3 runs converged on the same value, but figure 3 have shown better performance for the full FT run while the QLoRA performed worse. The reason behind this behavior will be

revealed in the next figure.

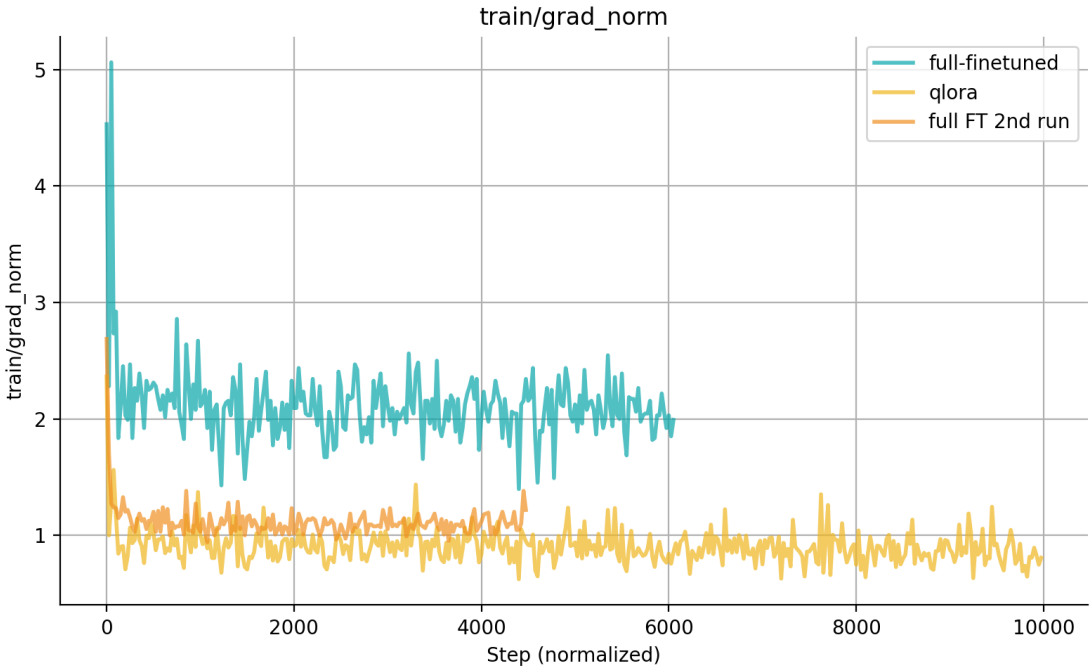


Figure 6: Gradient Norm Comparison between Full Fine-Tuning and QLoRA.

Gradient norm measures the overall strength of activation signals a neural network during backpropagation, providing crucial information whether the neural network is taking small gradient steps or large gradient steps.Higher values indicate larger update magnitudes, which can accelerate learning but also risk instability, and vice versa. The figure 6 shows the first full FT run to have large gradient norm values meaning it learned alot in the span of 6k steps.

In contrast, QLoRA displayed lower and more stable gradient norms, consistent with the fact that it only optimizes a small set of adapter parameters rather than the entire network.

6.2 Additional Comparisons

Model	**Size (B)**	**MVBench**	**GPU Mem (GB)**
Baseline smolVLM-Video-Instruct (Student)	0.5	29.7	0.9
full fine-tuned SmolVLM	0.5	35.3	0.9
QLora-tuned SmolVLM	0.5	31.8	0.9
Ensembled SmolVLM (Muraqeb)	1	**36.3**	1.8

Table 1: Benchmarking results comparing the teacher, distilled student, and an external baseline.