BodyCam-VQA: Enhanced Body-Worn Camera Video Captioning via Multimodal Reasoning and Probe Question Generation

Karish Gupta[♣] **Matthew Alex**[♣] **Alex Li**[♣]
Yang Wu[♣] **Yun-Wei Chu**[♡] **Kashif Munir**[♡]
Xiaotian Zhou[♡] **Zhengping Ji**[♡] **Xiaozhong Liu**[♣*]
[♣]Worcester Polytechnic Institute, Worcester, MA, USA
[♡]Axon, Scottsdale, AZ, USA
{kagupta, malex, ajli, ywu19, xliu14}@wpi.edu
{ychu, kmunir, xzhou, zji}@axon.com

Abstract

Police body-worn camera (BWC) footage has emerged as a critical aspect of law enforcement that ensures legal transparency, officer accountability, and the protection of civil rights. However, effectively processing this data remains a significant challenge due to its multimodal video format. BWC videos, in many cases, comprise chaotic scenes with low visual quality, rapid movement/interactions, and high-noise audio that make visual understanding a challenge for even SOTA multimodal models. Current Vision-Language Models (VLMs) frequently overlook critical forensic details, such as the presence of valuable evidence or the latent nuances of suspect-officer interactions, which are vital for fair legal outcomes and civilian/officer safety. To address these limitations, we propose an Adaptive Visual Question Answering (VQA) framework engineered for high-stakes law enforcement. Our framework employs a structured reasoning approach to extract fine-grained visual evidence that traditional captioning systems fail to capture. We experiment with multiple question generation models, including foundation models and fine-tuned open-weight models, to observe performance variation among question generation model implementations. Our results demonstrate that this VQA-driven architecture provides a more reliable, objective, and detailed record of enforcement events, ultimately serving as a powerful tool to protect both law enforcement officers and the public through AI-assisted forensic clarity.[1]

1 Introduction

The integration of police body-worn cameras (BWCs) into modern law enforcement has allowed a large amount of transparency into officer activity; however, many overlook the massive intelligence asset that BWC footage has become for the officers themselves (Voigt et al., 2017; Lum et al., 2020; Camp et al., 2024). By providing an objective view of high-stakes law enforcement encounters, these devices can be used to gather a great deal of evidence about law enforcement encounters that may be overlooked without video evidence (White, 2014). A 2022 study looking into the impact of body-worn cameras on police report writing showed that participants who did not review the BWC footage of police encounters were almost four times more likely to make at least one factual error in their written report (Boivin and Gendron, 2022). However, manual report writing is an exceptionally tedious task and can be difficult to implement at large scale when there is a great deal of video footage (Ferguson, 2025; Watts et al., 2026). Thus, with the growing domain of multi-modal large language models (MLLMs) and vision language models (VLMs), we see potential in building captioning pipelines for complex police BWC footage. Similar work has been done with the OpenBWC framework (Srbinovska et al., 2025), which proposes a multi-modal system to analyze interactions and behaviors found in BWC footage. This model, potentially valuable for law enforcement training, does not directly extract details and information from a given scene. We aim to build a system that can understand police interactions, point out valuable evidence (objects, people, etc.), and describe in detail the setting and location of a given video.

BWC videos frequently contain chaotic visual content and noisy audio conditions, making most types of automated understanding much more challenging (Grauman et al., 2022). Our work provides a framework for enhanced captioning of these videos, leading us to believe that a visual question-answering (VQA) system, alongside a strong VLM, could be a viable option for detailed and accurate automated video captioning.

Prior work on VQA has primarily focused on well-structured domains such as medical and sur-

*Corresponding author.
[1]https://github.com/Karish-Gupta/BodyCam-VQA

Proceedings of the Fifth Workshop on NLP for Positive Impact (NLP4PI), pages 103–115
October 28, 2026 ©2026 Association for Computational Linguistics

Pipeline Stage	Descriptions and Generated Outputs
Input Video	60s BWC footage: Video depicts an officer chasing a suspect trying to flee, deploying a Taser, then wrestling and detaining the person against a parked car
Initial VLM Summary	Officer pursues a civilian in a white t-shirt at 01:55:29. Taser deployed at 01:55:45. Physical struggle occurs against a parked vehicle.
Generated VQA	**Q:** What is the duration of the grapple? **A:** Approximately 20 seconds. **Q:** Are additional officers present? **A:** Yes, audio cues and brief visual frames imply multiple officers.
Final QA Caption	At 01:55:15, a BWC reveals a dark street with flashing police cruisers. Following a foot pursuit (01:55:29) and Taser deployment (01:55:45), the officer engages in an **intense 20-second struggle**. Notably, overlapping voices and visual cues confirm the **presence of additional officers**, concluding with the subject pinned against a parked SUV.

Table 1: Qualitative BodyCam-VQA Walkthrough

gical videos. Representative examples include LLaVA-Surg (Liu et al., 2024b) and Surgical-VQA (Salehi et al., 2022), which demonstrate the effectiveness of question-driven supervision for extracting fine-grained visual semantics. The LLaVA-Surg framework developed a two-stage question-answer generation pipeline: the framework first extracts discrete, structured factual information from surgical lecture videos and then uses pairs of structured information to generate intelligent questions. This approach greatly reduces the major risk of generating hallucinations in the questions.

Our goal is to utilize visual question-answering in tandem with powerful multimodal models to enhance the quality of BWC video captioning. A qualitative example of how this change can be observed in the generated captions is laid out in Table 1, where we can observe valuable details incorporated into the original caption as a direct result of intelligent and specific VQA. The addition of the questions pushed the model to look for more individuals in the scene, which changed the narrative from a single-officer scene to one of multi-officer involvement. This directly increases the contextual density of the generated caption.

In summary, our primary contributions to this work are as follows:

- We construct a new Body-Worn Camera (BWC) video dataset comprising 359 annotated 60-second videos with aligned transcripts to support multi-modal reasoning and analysis.
- We propose BodyCam-VQA, an adaptive Visual Question Answering framework tailored for body-worn camera footage. The architecture is designed to handle chaotic scenes, low visual quality, rapid interactions, and noisy audio by employing structured reasoning to extract fine-grained forensic evidence that conventional captioning and existing Vision-Language Models often overlook.
- We empirically validate the effectiveness of BodyCam-VQA through comprehensive experiments and human evaluation, demonstrating improvements in factual accuracy, answer completeness, and visual grounding.

2 Related Work

2.1 Video Understanding

Video understanding in the past few years has shifted towards structured multimodal reasoning (Fu et al., 2026). Vision-Language Models (VLMs) have had the biggest impact, using methods such as visual instruction tuning (Liu et al., 2024a), self-distillation (Naeem et al., 2023), and reinforcement learning (Liu et al., 2026; Guo and Gu, 2025; Wu et al., 2026) to help improve the alignment of these VLMs from recognition to decision making. In more complex settings, Chain-of-Thought(CoT) reasoning was incorporated into the multimodal domain, improving performance (Zhang et al., 2025; Park et al., 2025). Despite many improvements in the field, many blockers are still in place. These mainly focus on temporal hallucination (Fu and et al., 2024), where VLMs fabricate events in longer-form videos. This is because many of these models sample a few frames from thousands, which removes important context required for accurate reasoning. Models also struggle to distinguish between similar actions (e.g., handshakes vs fist-bumps), leading to major problems in complex reasoning, where these mistakes can invalidate a long reasoning chain.

2.2 Visual Question Answering

Visual Question Answering (VQA) has emerged as a powerful paradigm for extracting fine-grained, targeted information that general-purpose captioning often overlooks (Gupta, 2017; Huang et al., 2021), with structured questions shown to elicit more detailed, factually grounded responses than holistic descriptions. This has proven effective in well-structured domains such as radiology (Lau et al., 2018), pathology (He et al., 2020), long-form video (Yu et al., 2019), and legal consultation (Wu et al., 2024; Yao et al., 2025). LLaVA-Surg (Liu et al., 2024b) and Surgical-VQA (Salehi et al., 2022) further show that question-driven supervision extracts fine-grained surgical details missed by direct captioning. However, these works remain confined to controlled, well-lit, single-task environments, leaving open whether this paradigm extends to chaotic, multi-participant settings such as BWC footage.

2.3 VLMs in Public Safety

In the domain of public safety, two papers investigate similar topics. The first paper introduces VIVID (Gonzalez et al., 2025). This model focuses on identifying high-motion keyframes through optical flow, which are most likely to contain violent actions. It then converts the actions into text tokens and compares the generated text to the formal definitions of crimes to determine if they reach the standard. The second paper is an evaluation paper of VLMs on Surveillance Video (Benschop et al., 2025), testing the power of smaller models (8B or fewer parameters) on anomaly detection. In their model, the VLM generates a description, and a separate model scores them based on whether they contain an anomaly label such as "theft". The paper highlighted that the models were good at detecting events when they were "clear" (a fight in the center of the frame), but not as good at detecting events that were noisy or partially obstructed.

3 Methodology

3.1 Problem Definition

The BodyCam-VQA pipeline is designed to transform multimodal inputs into dense, investigative captions. Given a BWC video V and its associated transcript T, our goal is to synthesize a VQA-enhanced caption C. This caption represents a comprehensive analysis of the scene, capturing scene-level context, the individuals involved, and critical environmental details.

3.2 BodyCam-VQA Mechanism

The framework $\mathcal{F}$ consists of four primary stages:

- Multimodal Summarization (S)
- Structured Detail Extraction (D)
- Targeted Question Generation (Q)
- Visual Question Answering and Synthesis (C)

We will introduce each part of the workflow in the following sections.

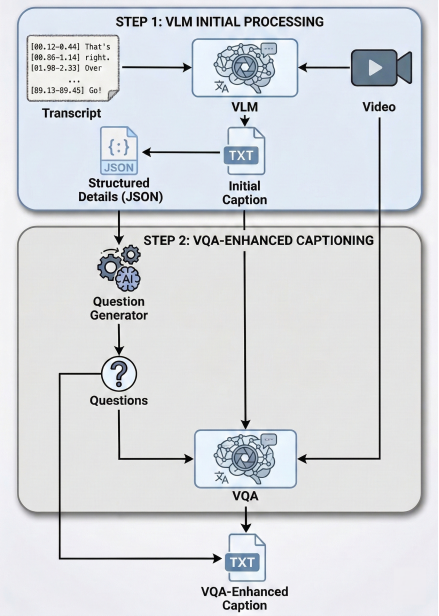


Figure 1: Overview of the BodyCam-VQA Framework and its Two Main Steps. Initial Caption Generation Creates a Baseline VLM Summarization. Final VQA-Enhanced Caption Utilizes VQA Driven by Targeted Question Generation.

3.2.1 VLM Summarization

Preliminary experiments indicated that audio transcripts T often lack the visual granularity necessary for high-quality caption generation. All visual details (e.g., physical interactions, objects in the scene, participant descriptions) are overlooked. To bridge this gap, we utilize a VLM to generate an initial visual summary $S = f_{VLM}(V, T)$. This summary serves a dual purpose: it provides the primary visual context for subsequent stages and acts as a baseline to evaluate the performance gain of our multi-pass pipeline over standard single-pass VLM captioning.

3.2.2 Structured Details

To ensure focused investigative questions, we adapt the LLaVA-Surg (Liu et al., 2024b) framework.

Prior work in PDE-based Bayesian hierarchical modeling (Cen et al., 2025) shows that decomposing complex processes mitigates observational noise. Inspired by this, we extract structured details D in JSON format as observed in Equation 4

$$\mathcal{D} = \{D_{\text{scene}}, D_{\text{entity}}, D_{\text{action}}, D_{\text{semantic}}\} \quad (1)$$

This process compresses raw multimodal information into a discrete set of fine-grained categories (scene-level, entity-level, action-level, and semantic-level), as illustrated in Figure 2. By grounding question generation in these structured details, we enforce factual consistency and mitigate the risk of model hallucination during the question generation phase.

3.2.3 Question Generation Models

The efficacy of the VQA enhanced caption C is inherently based on the quality of the generated questions Q. To optimize this, we formalize question generation as:

$$Q = G(D) \quad (2)$$

Where G is a generative model. We investigated various architectures for G, ranging from large-scale foundation models to specialized, fine-tuned models.

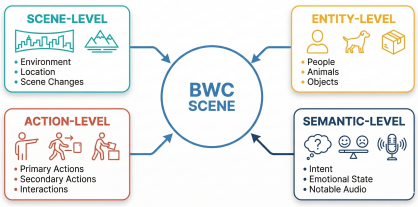


Figure 2: Structured Details Schema. Comprised of Scene, Entity, Action, and Semantic Level Details.

3.2.4 Training Data with Chain of Thought

Preliminary evaluations indicated that smaller language models struggled to consistently generate investigative questions capable of extracting nuanced details from BWC footage. In contrast, larger proprietary models demonstrated significant proficiency in this domain. To bridge this performance gap, we implemented multiple fine-tuning approaches, utilizing a generated dataset of high-quality questions with chain-of-thought reasoning (to encourage smaller models to follow a similar thought process). This training data of curated high quality questions, paired with latent reasoning, was generated with a proprietary model with strong reasoning and CoT capabilities.

3.2.5 Standard Fine-tuned Model

We conducted supervised fine-tuning on open-weight models to evaluate the efficacy of training smaller models on reasoning-dense data. The training set consisted of 288 examples of input structured details, CoT tokens, and high-quality questions. This allows for a direct assessment of how distilled reasoning affects the zero-shot investigative performance of smaller open-weight models.

3.2.6 GRPO Fine-tuned Model

To refine the precision and relevance of our investigative questions, we employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Unlike SFT, GRPO optimizes the model policy by evaluating a group of sampled outputs against their collective mean, facilitating stable reinforcement learning. By utilizing a critic-based reward, this method enables the model to iteratively improve question quality and investigative depth, surpassing the static performance of SFT.

We define the total reward R for generated questions as a multi-objective function. For a given input D, the model generates a set of G outputs $\{o_1, o_2, ..., o_G\}$. The reward for each output o_i is calculated as:

$$\mathcal{R}(o_i) = \alpha \cdot r_{format}(o_i) + \beta \cdot r_{judge}(o_i) \quad (3)$$

Where:

- r_{format} (Rule-based Reward): This component encourages the structural integrity of the generated output. It assigns high values for the correct use of <think> (Chain-of-Thought) and <question> tags. Additionally, it applies a length penalty on short questions and a keyword bonus to prioritize interrogative words (e.g., how, describe, identify).
- r_{judge} (Model-based Reward): To capture semantic nuance, we employ a large foundation model as an automated judge. The judge evaluates each generated question $q \in o_i$ on a binary scale $\{0, 1\}$ based on its relevance to the specific video context S and investigative quality.
- α, β: These represent coefficients used to balance structural compliance and question quality.

3.2.7 VQA and Caption Synthesis

In the final stage, the pipeline performs a second VLM pass on the video. The questions are passed to the VLM for visual question-answering, yielding

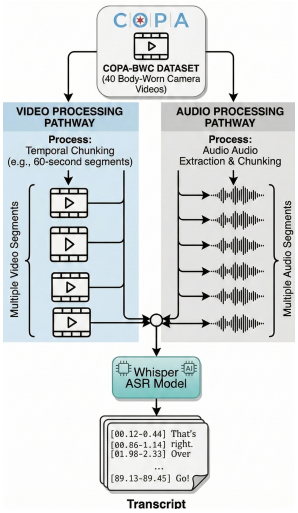


Figure 3: Dataset Preprocessing Pipeline. Ingests Full-Length BWC Videos and Generates 60-Second Video Segment and Transcript Pairs.

a set of answers A. Finally, the original VLM summary S, the questions Q, and the answers A are ingested by a LLM to produce a final synthesized caption C:

$$C = \text{LLM}(S, Q, A) \tag{4}$$

This iterative refinement allows the model to include interactions and objects that may typically be overlooked in a single summarization pass.

The structured details are passed to the VLM once again, but this time for visual question-answering. This acts like the second VLM pass on the video. However, instead of focusing on broad details, the model is encouraged to observe specific interactions, objects in the scene, etc. Finally, the questions, answers, and original VLM summary are all passed to an LLM for VQA-enhanced caption synthesis. This brings us to the end of the BodyCam-VQA pipeline.

4 Source Dataset and Video Preprocessing

We utilized the Chicago Civilian Office of Police Accountability (COPA) dataset, an open-source collection of incidents involving police, which includes body-worn camera footage with audio. From this dataset, we sampled 40 full-length videos covering a diverse range of police incidents and visual contexts.

To reduce computational overhead and keep our VQA pipeline manageable with limited resources, each video is segmented into 60-second clips, giving us a total dataset size of 359 clips. These clips were chosen from high-action videos from the COPA dataset, and frames were sampled at a rate of 1 FPS. Based on previous works (Zhu et al., 2025; Yenduri et al., 2024), current VLMs generally perform best with short videos and relatively simple, visual questions. Since VLMs have to extract a set number of frames from a video, the frame rate can become quite limited. Thus, this clipping strategy helps retain sufficient temporal information for reasoning over actions and events, and allows for traceable inference for large-scale VLMs. In our pipeline, each resulting clip is treated as an independent training instance for downstream processing.

For each 1-minute video segment, we generate high-quality textual transcripts using WhisperX (Bain et al., 2023), a state-of-the-art automatic speech recognition (ASR) model. The quality of these captions was checked through manual human checking of each. WhisperX is selected for its highly accurate transcription capabilities. The transcription pipeline produces labeled and time-aligned transcripts per video clip as observed in Figure 3.

5 Experiment

5.1 Experimental Settings

Gemini 2.5 Flash is used for initial VLM summarization, VQA, and final caption synthesis in our framework. We selected this proprietary model due to its robust performance and significantly lower hallucination rates in the context of chaotic BWC footage compared to open-weight alternatives. To ensure reproducibility across experiments, the decoding temperature is set to 0.1.

To establish an open-source performance VLM summary baseline, we also implement Qwen3-VL, a highly capable open-weight VLM currently available. We specifically utilize Qwen3-VL to generate single-pass summaries, allowing us to quantify the performance gains achieved by our proposed multi-pass VQA framework.

For the question generation task, we evaluated both large-scale foundation models and specialized fine-tuned architectures. The foundation models include Gemini 2.5 Flash and DeepSeek V3.2. To improve question quality, we implemented and compared three fine-tuned variants on Qwen3 [2], a

[2]https://huggingface.co/BodyCam-VQA

specialized reasoning model optimized for complex text generation and chain-of-thought (CoT) logic:

- Qwen3-4B-Thinking-2507 (SFT)
- Qwen3-30B-A3B-Thinking-2507 (SFT)
- Qwen3-4B-Thinking-2507 (GRPO)

All fine-tuned models were trained and deployed using Low-Rank Adaptation (LoRA) to maximize parameter efficiency. For inference, we employed 4-bit quantization to maintain high throughput. The generation parameters were standardized with a temperature of 0.1 and a top-p value of 0.9. All computational experiments were conducted on a single NVIDIA H200 GPU.

5.2 Baselines

To evaluate the performance of the proposed architecture, we compare our results against four distinct baselines categorized into two groups: Direct VLM Summarization and Pipeline-based VQA.

Direct VLM Summaries: We employ Gemini 2.5 Flash and Qwen3-VL-32B-Instruct to generate single-pass captions. These models serve as high-performing proprietary and open-source benchmarks, respectively. By evaluating these initial summaries, we establish a performance floor to quantify the incremental value provided by our multi-stage VQA process.

VQA-Integrated Pipelines: We implement two baseline versions of the BodyCam-VQA architecture. In these configurations, the core architecture remains constant, but the question generation (QG) module is powered by either Gemini 2.5 Flash or DeepSeek V3.2. These baselines demonstrate the upper-bound efficacy of the proposed framework when paired with state-of-the-art foundation models and provide a rigorous benchmark for our fine-tuned QG models.

5.3 Tasks and Metrics

In order to properly evaluate generated captions for these videos, we put together a human-evaluated ground truth dataset that includes important details, visual enrichment details, and extra auxiliary details from the videos. Important details contain all major events that occurred during the video, visual enrichment details include all important visual details from the video, and auxiliary details include any extra details we added. Our evaluation is based on our ground truth dataset.

Generated captions are evaluated using the following three metrics:

- "Factual Accuracy": 0 - 1,
- "Completeness": 0 - 1,
- "Visual Enrichment": 0 - 1,

Factual Accuracy: Measures whether the statements in the caption are based on ground truth. If the general idea is mentioned in the ground truth (any section), it is considered true. If extra details are added that contradict the ground truth, the score is penalized. Each sentence must be checked against the ground truth to verify accuracy, and the final score is calculated by dividing the number of statements grounded in truth by the total number of statements. Since we cannot incorporate all details that the model may potentially extract into the ground truth dataset, we used human evaluation to calculate the factual accuracy for each model.

Completeness: Measures whether the caption captures all relevant and important events. The model first retrieves the list of events from the important details section of the ground truth. The model then checks whether each event from the list is present in the generated caption. The final score is the number of included events divided by the total number of important details events.

Visual Enrichment: Measures the proportion of specific visual details captured in the caption. The model first retrieves the list of items from the visual enrichment details section of the ground truth and checks whether each specific visual detail from that list is described in the generated caption. The final score is the number of included visual details divided by the total number of visual enrichment items.

Since there is potential for the completeness and visual enrichment calculations to differ based on the judge, we utilized two judges (Gemini 2.5 Flash and Deepseek V3.2) and took the averages of their judgment results.

5.4 Main Results

As illustrated in Table 2, the incorporation of a dedicated VQA stage consistently enhances performance across all metrics compared to the Direct VLM Summary baselines. Both QA-enhanced pipelines utilizing closed-source question generation (QG) models—Gemini 2.5 Flash and DeepSeek V3.2—demonstrate superior factual accuracy, completeness, and visual enrichment over their single-pass counterparts.

Metric	Visual Sum. (Qwen3)	Visual Sum. (Gemini)	Gemini QA	Deepseek QA	Qwen3 30B SFT QA (ours)	Qwen3 4B GRPO QA (ours)
Human Factual Acc.	0.86	0.86	0.89	0.90	0.86	**0.91**
Gemini as Judge						
Completeness	0.27	0.51	0.64	0.59	**0.65**	0.56
Visual Enrichment	0.60	0.59	0.65	**0.68**	0.62	0.66
Deepseek as Judge						
Completeness	0.22	0.50	0.59	0.57	**0.62**	0.53
Visual Enrichment	0.49	0.42	**0.57**	**0.57**	0.50	0.49
Judge Averages						
Completeness	0.25	0.51	0.62	0.58	**0.64**	0.55
Visual Enrichment	0.55	0.51	0.61	**0.63**	0.56	0.58

Table 2: Comparative Performance of BWC Video Captioning Models Across Human and LLM-as-a-Judge Metrics

Our fine-tuned Qwen-based models generally exceed the performance of the VLM summary baselines and remain highly competitive with the closed-source QA pipelines. This suggests that structured question-answering effectively bridges the gap between mid-sized open-source models and flagship foundation models for domain-specific tasks.

In particular, the highest Human Factual Accuracy (0.91) was achieved by the Qwen 3 4B GRPO QA pipeline, while the strongest Completeness score (0.65) was attained by the Qwen 3 30B SFT QA variant. The closed-source QA pipelines scored highest in Visual Enrichment, with the DeepSeek-integrated model reaching a peak score of 0.68.

In comparing results using the QA pipeline and those that did not, the QA models showed improvement in performance over the baseline models, scoring between 0.04-0.13 points better in Completeness and 0.03-0.09 points better in Visual Enrichment. All of the QA pipelines also produced significantly denser captions than the baseline.

When comparing the results from the question generator using Foundation model pipelines (Gemini and Deepseek), and the Qwen3 finetuned models, it is clear that foundation models are better than the trained Qwen3 models; however, the performance between all of them did not differ by much, showing that these models are quite competitive and capable of putting up similar performance to models much larger than themselves.

5.5 Ablation Study

To investigate model scaling and the various training methods employed, we conducted an ablation study using the Qwen3 model as the backbone. We examine the impact of SFT versus reinforcement learning (GRPO), as well as the effects of increasing parameter count.

Configuration	HFA ↑	Comp ↑	VE ↑
Impact of Model Scale (SFT)			
Qwen3 4B SFT QA	0.87	0.57	0.56
Qwen3 30B SFT QA	0.86	**0.64**	0.56
Impact of Training Strategy (4B)			
Qwen3 4B QA (Base)	0.78	0.49	0.55
Qwen3 4B SFT QA	0.87	**0.57**	0.56
Qwen3 4B GRPO QA	**0.91**	0.55	**0.58**

Table 3: Ablation Study: Impact of Model Scale and Training Strategy on BodyCam-VQA Performance. (HFA: Human Factual Accuracy, Comp: Completeness, VE: Visual Enrichment).

As observed in Table 3, the transition from the base Qwen model to ones that incorporate specialized QG training yields significant performance gains. The Base Qwen3 4B QG model, while utilizing our VQA architecture, achieved a human factual accuracy of only 0.78, which is lower than that of single-pass VLM Summary models.

The most notable shift occurred with the implementation of the GRPO fine-tuned Qwen model. The GRPO variant was trained directly by a strong judge model, resulting in the highest factual accuracy (0.91) across all tested 4B models. For this test the reward function α and β weights were each set to 0.5, to ensure an equal balance between structure and question quality. Interestingly, while GRPO slightly reduced the completeness score compared to SFT (0.55 vs. 0.57), it effectively minimized hallucinations, which is a critical requirement for forensic body-worn camera (BWC) documentation.

We further evaluated the effect of parameter scaling by comparing the 4B and 30B SFT variants.

Increasing the model size to 30B parameters led to a substantial increase in Completeness (0.64), the highest in our study. This suggests that larger models are more capable of coming up with more specific and effective questions for investigation. However, this increase in information density did not translate to higher factual accuracy, which remained at 0.86. This finding reinforces our hypothesis that reinforcement learning (GRPO) is a more effective tool than raw scaling for ensuring the reliability of generated police reports.

5.6 Case Study

To demonstrate the practical efficacy of the *BodyCam-VQA* framework, we present a qualitative case study of a one-minute BWC video sequence. Generated captions alongside representative key frames are provided in Appendix A Figure 4. This section serves as an illustrated example of the model's ability to capture high-stakes context in a real-world law enforcement scenario. The qualitative comparison of the generated captions reveals a clear progression in quality across the different model tiers, with our Fine-Tuned model performing the best.

The selected clip features a response post-incident with the following ground-truth events:

- **Initial Scene:** The officer observes fellow officers secure a scene at night in a residential driveway.
- **Analyze Surroundings:** The officer turns around to a residential street to gauge the surroundings.
- **Observation of Civilian:** The officer returns to the scene and observes the civilian on the ground.
- **Surveillance:** Fellow officers continue to monitor the scene, keeping a close-eye on the civilian.

Analysis: Comparing the three captions, the Baseline Caption is noticeably shorter than the other two. When comparing the Baseline Caption to the QA Caption, the QA Caption provides substantially more detail, but it does so at the cost of accuracy. For instance, it incorrectly states that the scene takes place in an adjacent yard rather than in the present yard. It also misidentifies the siding material between two houses that were passed through by the bodycam officer. Additionally, it inaccurately captures details about the officers, such as how many are directing flashlights at the scene. The Baseline Caption also contains errors, such as misreporting the number of officers present and the number of officers using flashlights.

When comparing both of these captions to the Fine-Tuned Caption, we observe clear improvement. The Fine-Tuned Caption avoids being overly brief like the Baseline Caption while also avoiding the excessive, speculative detail found in the QA Caption. Instead, it strikes a balance between conciseness and completeness. It captures important and accurate scene details such as the driveway setting, the presence of multiple flashlight beams, the civilian's location and condition, and the fact that a weapon was recovered without introducing major hallucinations.

The only error in the Fine-Tuned Caption is the incorrect attribution of flashlight use to the bodycam officer. This mistake also appears in the QA Caption and is likely due to the complexity of the scene (multiple officers; approximately five, including the bodycam officer) and multiple light sources create visual ambiguity. Given this, the error is understandable and relatively minor compared to the major inaccuracies found in the other captions. Overall, the Fine-Tuned Caption most effectively minimizes hallucinations while maintaining meaningful detail, making it the strongest and most balanced of them.

6 Conclusion and Future Work

BodyCam-VQA improves caption density for BWC footage by asking high-quality questions to the model, verifiably improving the model accuracy and enriching the output. However, some issues remain in final caption generation, particularly the misattribution of physical attributes caused by model hallucination. Object misattribution was observed across all VLMs tested. The unique perspective of the camera and chaotic scenarios of BWC footage can lead to odd angles and clipping of objects in view, which leads to incorrect predictions for object recognition. Furthermore, since these models are given video and transcript data separately, determining the speaker at a given point in the video remains challenging. This problem is especially difficult since the video is taken from an officer's perspective, making it hard to determine whether a piece of dialogue came from the POV

officer or another individual off-camera. Future work will focus on improving the model's ability to distinguish between individual people.

Limitations

While BodyCam-VQA demonstrates consistent improvements in factual accuracy, completeness, and visual enrichment over baselines, we acknowledge several limitations of the current work:

• ***Limited Temporal Scope.*** Our current pipeline and evaluation are constructed around 60-second video segments, a design choice made to keep frame sampling tractable and inference costs manageable within our computational budget. As a result, we do not evaluate the framework's ability to perform long-horizon temporal reasoning or to maintain coherent context across multiple linked segments from the same incident. Real-world BWC footage often spans several minutes to hours and may require linking evidence, entities, and events across segment boundaries to construct a complete investigative narrative. Extending BodyCam-VQA to support cross-segment memory and longer temporal windows is an important direction for future work.

• ***Indirect Alignment in the Reward Signal.*** Our GRPO-based training relies in part on a model-based reward (r_{judge}) in which a large foundation model scores the relevance and investigative quality of generated questions. While this design allows us to scale supervision beyond hand-labeled question quality, it also means the policy is optimized toward what a judge model considers a high-quality question, rather than a signal directly derived from human forensic assessment. Since our human evaluation is applied only to the final generated captions and not to the intermediate question-generation reward, there remains a possibility that the judge's preferences diverge from those of human reviewers in edge cases, or that the policy could learn to exploit regularities in the judge's scoring rather than genuinely improving investigative value. We view incorporating direct human feedback into the reward loop, or periodically auditing judge-policy agreement, as a valuable next step to further ground this training signal.

• ***Proof-of-Concept Evaluation Setting.*** Our evaluation is conducted using a curated ground-truth dataset and automated/human scoring of generated captions, which allows for controlled and reproducible comparison across model configurations. However, we have not yet evaluated BodyCam-VQA within an operational law enforcement workflow, such as direct comparison against officer-written reports, integration into report-drafting tools, or usability studies with investigators and legal reviewers. These deployment-oriented evaluations would help clarify the practical time savings, adoption barriers, and remaining accuracy gaps relevant to real-world use, and we consider this an important next stage beyond the current proof-of-concept study.

Ethical and Societal Implications

The positive impact of our work lies in the enhancement of legal transparency, objective accountability, and the safeguarding of civil rights through AI-assisted forensic clarity. BWCs are intended to be impartial observers; however, the sheer volume and chaotic nature of this footage often lead to critical details being overlooked by human reviewers or hallucinated by standard VLMs. By structuring the captioning process through an adaptive VQA framework, our work directly contributes to a more equitable justice system, ensuring that latent nuances of suspect-officer interactions are accurately documented to protect both the public and law enforcement officers. We view the resulting increase in caption density and factual grounding, alongside a corresponding reduction in manual reporting burden, as a concrete step toward more reliable and equitable documentation of high-stakes encounters.

We use the publicly released COPA dataset for our experiments, which was made available for the purpose of civilian oversight, and we do not collect, process, or attempt to re-identify any footage or individuals beyond what is already disclosed in this source. Our use of this dataset and the overall design of our framework are consistent with the ACL Ethics Policy, and we do not foresee any additional ethical concerns beyond those already inherent to the publicly released source dataset.

Acknowledgment

We gratefully acknowledge the support and collaboration from Worcester Polytechnic Institute and Axon. We also thank the reviewers and colleagues for their helpful comments and feedback.

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-aligned speech transcription with forced alignment and diarization. *arXiv preprint arXiv:2303.00747*.

Pascal Benschop, Cristian Meo, Justin Dauwels, and Jelte P. Mense. 2025. Evaluation of vision-llms in surveillance video. *arXiv preprint arXiv:2510.23190*. Accepted at NeurIPS 2025 Workshop on Space in Vision, Language, and Embodied AI.

Rémi Boivin and Annie Gendron. 2022. An experimental study of the impact of body-worn cameras on police report writing. *Journal of Experimental Criminology*, 18(4):747–764.

Nicholas P Camp, Rob Voigt, MarYam G Hamedani, Dan Jurafsky, and Jennifer L Eberhardt. 2024. Leveraging body-worn camera footage to assess the effects of training on officer communication during traffic stops. *PNAS nexus*, 3(9):pgae359.

Mengqi Cen, Xuejing Meng, X Joan Hu, Juxin Liu, and Jianhong Wu. 2025. Pde-based bayesian hierarchical modeling for event spread, with application to covid-19 infection. *arXiv preprint arXiv:2509.13174*.

Andrew Guthrie Ferguson. 2025. Generative suspicion and the risks of ai-assisted police reports. *Nw. UL Rev.*, 120:299.

Chaoyou Fu and et al. 2024. Video-mme: The first comprehensive multi-modal benchmark for dataset-level video evaluation. *arXiv preprint arXiv:2405.21075*.

Jiahui Fu, Junyu Nan, Lingfeng Sun, Hongyu Li, Jianing Qian, Jennifer L. Barry, Kris Kitani, and George Konidaris. 2026. Novaplan: Zero-shot long-horizon manipulation via closed-loop video language planning.

Jose Gonzalez, Tetsu Matsukawa, and Einoshin Suzuki. 2025. Leveraging vision language models for understanding and detecting violence in videos. In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 2: VISAPP*, pages 99–113. SciTePress.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012.

Yuncheng Guo and Xiaodong Gu. 2025. Mmrl: Multi-modal representation learning for vision-language models. *arXiv preprint arXiv:2503.08497*.

Akshay Kumar Gupta. 2017. Survey of visual question answering: Datasets and techniques. *arXiv preprint arXiv:1705.03865*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Tong-yuan Huang, Yu-ling Yang, and Xue-jiao Yang. 2021. A survey of deep learning-based visual question answering. *Journal of Central South University*, 28(3):728–746.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251.

Haotian Liu et al. 2024a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Shih-Yang Liu et al. 2026. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*.

Shuyue Liu et al. 2024b. Llava-surg: Towards multimodal large language models for surgery. *arXiv preprint arXiv:2408.07981*.

Cynthia Lum, Christopher S Koper, David B Wilson, Megan Stoltz, Michael Goodier, Elizabeth Eggins, Angela Higginson, and Lorraine Mazerolle. 2020. Body-worn cameras' effects on police officers and citizen behavior: A systematic review. *Campbell systematic reviews*, 16(3):e1112.

Muhammad Ferjad Naeem et al. 2023. Silc: Improving vision language pretraining with self-distillation. *arXiv preprint arXiv:2310.13355*.

Chan Young Park, Jillian Fisher, Marius Memmel, Dipika Khullar, Seoho Yun, Abhishek Gupta, and Yejin Choi. 2025. Making VLMs more robot-friendly: Self-critical distillation of low-level procedural reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 32588–32612, Suzhou, China. Association for Computational Linguistics.

Mohammad Salehi, Tal Arbel, and Nicholas Ayache. 2022. Surgical-vqa: Visual question answering in surgical scenes. *arXiv preprint arXiv:2206.11053*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Anita Srbinovska, Angela Srbinovska, Vivek Senthil, Adrian Martin, John McCluskey, Jonathan Bateman, and Ernest Fokoué. 2025. Towards ai-driven policing: Interdisciplinary knowledge discovery from police body-worn camera footage.

Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the national Academy of sciences*, 114(25):6521–6526.

Seth Watts, Michael D White, and Aili Malm. 2026. Ai in policing: Capturing police performance (good and bad) in routine, rare, and novel events. *CrimRxiv*.

Michael Douglas White. 2014. *Police officer body-worn cameras: Assessing the evidence*. Office of Justice Programs, US Department of Justice Washington, DC.

Yang Wu, Chenghao Wang, Ece Gumusel, and Xiaozhong Liu. 2024. Knowledge-infused legal wisdom: Navigating llm consultation through the lens of diagnostics and positive-unlabeled reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15542–15555.

Yang Wu, Jinhong Yu, Jingwei Xiong, Zhimin Tao, and Xiaozhong Liu. 2026. " excuse me, may i say something..." colabscience, a proactive ai assistant for biomedical discovery and llm-expert collaborations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36109–36129.

Rujing Yao, Yang Wu, Chenghao Wang, Jingwei Xiong, Fang Wang, and Xiaozhong Liu. 2025. Elevating legal llm responses: harnessing trainable logical structures and semantic knowledge with legal reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5630–5642.

Gokul Yenduri et al. 2024. Beyond the limits: A survey of techniques to extend the context length in large language models. *arXiv preprint arXiv:2407.03211*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9127–9134.

Juntian Zhang et al. 2025. Weaving context across images: Improving vision-language models through focus-centric visual chains. *arXiv preprint arXiv:2504.20199*.

J. Zhu et al. 2025. Vision-language models: A comprehensive survey. *Information Fusion*, 114.

A Case Study

Key Frames



(a) Officers secure a scene at night in a residential driveway



(b) Emergency vehicles are present



(c) Officers illuminate a subdued civilian on the ground



(d) Officers observe the subdued civilian

Figure 4: Representative Frames From Case Study. Errors in the generated captions below are highlighted in Red. **Notable/key information** is **bolded**.

Baseline Caption: *At night, police officers are in a **residential area** with single-family homes. A **dark sedan** is parked in a driveway **next to a white garage**. A person, identified as the **"offender," is on the ground next to the garage**. Police officers in dark uniforms, one wearing a black knit cap, attend to the person. **One officer** uses a flashlight to illuminate the scene. An officer states that a **weapon has been recovered** and requests **EMS**. The bodycam officer moves through a dark alleyway between houses, then walks on a sidewalk where police vehicles with **flashing blue lights are visible**. Sounds of a police siren are heard briefly, followed by a continuous phone ringing. The bodycam officer returns to the scene by the garage where **three officers** are surrounding the person on the ground. Officers repeatedly ask, "Hey, you okay?" and **instruct the individual to "Keep breathing."** An officer informs the person, "**EMS is on the way**, buddy," and continues to request EMS over the radio."*

QA Caption: *Nighttime envelops a residential property as an officer, wearing a dark uniform and body camera, approaches. This officer first focuses on a dark sedan parked in front of a white garage. In the adjacent yard, multiple officers, also in dark uniforms, attend to a **civilian lying on the ground**. An officer states, "offender is down on the ground in their yard," and another confirms a significant development, "**Weapon recovered**, let's get EMS back here," indicating a weapon was found on or near the civilian, though it remains unseen. The body camera officer rapidly navigates away from the immediate incident, moving down a narrow, dark alleyway **between two houses**, one with dark siding and another lighter, where small, light-colored windows are briefly visible. Using a flashlight to cut through the darkness, the officer quickly rounds a corner onto a street. There, active streetlights illuminate parked cars, including a **white pickup truck**, and the **flashing blue and red lights** of **emergency vehicles** are visible further down the road, indicating a strong police presence. The officer then passes the front of a residential house featuring a white front door with steps, brick siding, and an illuminated porch light, presumably moving to direct arriving medical personnel. Returning to the yard, **the body camera officer directs their flash-***

light to illuminate the scene where other officers assist the civilian.*The civilian lies on their back on the dark, unpaved ground, which consists of dirt, gravel, and some debris next to the white garage. The* ***civilian wears dark long-sleeved clothing and dark pants.*** *While no major injuries or bloodstains are immediately visible, their breathing appears labored. The body camera officer asks, "Hey, you okay?" and instructs the civilian, "Keep breathing, dude," reassuring them, "****EMS is on the way****, buddy." Another officer states, "We're good, we're good," and the body camera user reiterates, "Everybody relax," and firmly instructs, "Squad, get EMS over here," emphasizing the urgent need for medical assistance. All visible doors and windows on the residential property appear closed, and* ***no other individuals are seen****, beyond the responding officers and the civilian, ensuring the scene's immediate security.*

Fine-Tuned GRPO Caption: *At night,* ***emergency vehicles with flashing blue lights line the street*** *of a residential area. A* ***dark-colored sedan*** *is parked* ***in a driveway****, nestled between a* ***white garage*** *and a* ***house****.* ***Multiple flashlight beams****, along with* ***ambient light****, actively illuminate the scene, clearly detailing the sedan's wheels and body lines, the texture of the white garage door and wall, and the ground and driveway, including the tight space where a civilian is located. Officers, wearing dark uniforms, actively use their flashlights to pierce the darkness,* ***directing beams onto the scene****. One* ***officer recovers a weapon;*** *however, the bodycam officer does not handle it. A* ***civilian lies on the ground****, primarily* ***on their stomach or side,*** *in the tight space* ***between the garage and a fence/wall****. The flashlight beams from various officers, along with a blue light, consistently illuminate the civilian's upper body and head. The* ***civilian appears subdued****, showing* ***no visible signs of active struggle, defensive postures, or rapid uncooperative movements,*** *and is breathing regularly, responding verbally with "Yeah" when an officer checks on their well-being. The bodycam officer initially approaches the scene where another* ***officer is already tending to the civilian*** *on the ground. After an officer states, "Weapon recovered." the bodycam officer immediately moves away, walking down a narrow path between houses and towards the street. After a brief pause near the street, the bodycam officer returns to the civilian's location near the garage. The bodycam officer then gets closer to check on the civilian, shining their flashlight directly on them and instructing them to "keep breathing." Officers reassure each other and the civilian, stating, "We're good," and "Everybody relax," as an officer reiterates the request for EMS to respond to their exact location."*