

EgoPolice: A Benchmark for Egocentric Video Understanding in High-Stakes Police Body-Worn Camera Footage

Max Gonzalez Saez-Diez^{*1}  Jihoon Chung^{*1} 
Adam D. Wolsky¹  Gregory Lanzaletto²  Dean Knox² 
Jonathan Mummolo¹  Brandon M. Stewart¹  Olga Russakovsky¹ 

¹ Princeton University

{saezdiez, jc5933, awolsky, jummolo, bms4, olgarus}@princeton.edu

² University of Pennsylvania

glanza@wharton.upenn.edu, dcknox@upenn.edu

Abstract. We introduce EgoPolice, a carefully curated dataset of real, egocentric police-civilian interactions, sourced from publicly available body-worn camera videos. We select police-civilian action labels that are critical for police behavioral research and annotate them at a second-by-second granularity. The videos feature rapid and irregular camera motion, dense human interactions, and rare high-stakes events, making the dataset a challenging benchmark for motion-robust and context-aware egocentric perception. We provide two different tasks, classification and multiple-choice question-answering, and benchmark both open-source and closed-source models. We find that even the best video models like Gemini 2.5 Pro still struggle to accurately predict high-risk actions such as “Weapon Out”. Beyond serving as a benchmark, EgoPolice provides a foundation for developing models capable of identifying events of interest in large-scale body-worn camera video repositories, enabling more efficient downstream human review.

Content Warning: This paper includes real police body-worn camera footage, including potentially distressing scenes.

Keywords: Egocentric Dataset, Video Understanding, Police Body-Worn Camera Analysis

1 Introduction

Recent advances in Vision-Language Models (VLMs) have generated significant interest in their application to a wide range of tasks and environments, including complex, high-stakes settings such as medical diagnosis, autonomous navigation, and disaster response [16, 41, 68]. Despite their potential to support real-world decision-making, failures of these models have exposed important ethical, legal,

^{*} Equal Contribution

Project Website: <https://github.com/princetonvisualai/egopolice>



Fig. 1: Sample frames of EgoPolice. We show representative frames for *Easy* (Top, Cyan) versus *Hard* (Bottom, Red) samples. *Easy* samples are correctly classified by VideoMAE V2 and CLIP [52,66], while *Hard* samples are misclassified by both. Models perform well on clear viewpoints but struggle under low-light, occlusion, and distance.

and safety concerns [5, 25, 45]. These are especially pronounced in the context of law enforcement, where errors can carry civil, legal, and safety consequences. The proliferation of Body-Worn Cameras (BWCs) has generated massive archives of footage, driving demand for automated analysis. However, the data captured by these cameras differs from standard training datasets as the videos are characterized by rapid camera motion, severe occlusions, and actions that are usually filtered out in academic datasets. While the failure to distinguish between an innocent gesture and a real threat can carry serious social consequences, commercial vendors [1, 6, 21, 50, 62] have entered the space with tools designed to interpret these records, even though the research community lacks the instruments to characterize their limitations and stress-test them against footage that is often difficult to parse even for a trained human observer (Figure 1).

To address these challenges, we introduce a carefully curated dataset of real-world, egocentric police-civilian interactions. Our work is motivated by the urgent need to benchmark and improve model reliability in first-person environments. Specifically, our primary contributions are threefold:

- *Bridging the domain gap for in-the-wild video datasets:* Existing datasets focus on benign, everyday activities, yet fail to capture more nuanced and rapidly-evolving events. We construct EgoPolice to provide the first benchmark for evaluating video models in egocentric law enforcement scenes.
- *Systematically benchmarking video understanding models in high-stakes environments:* We conduct an evaluation of standard classification techniques and state-of-the-art VLMs. Our analysis reveals that despite recent advancements, models still struggle with the complexities of BWC footage.
- *Demonstrating real-world transferability:* In ongoing work, we deploy a model trained on EgoPolice in a human-in-the-loop setting to analyze a large private repository of uncured body-worn camera videos, enabling systematic review at a scale that would be otherwise infeasible manually. This provides preliminary evidence that EgoPolice can serve as a foundation for scalable police oversight tools capable of operating on real-world footage.

2 Related Works

Police body-worn camera research. Police body-worn cameras (BWCs) are recording devices mounted on an officer’s chest that capture continuous video while the officer is on duty. They provide first-person documentation of police activity and were introduced with the goal of supporting police accountability and criminal investigations. Existing BWC-related research is mainly conducted by social scientists and focuses on evaluating their effectiveness, for example by measuring changes in civilian complaints, policing quality, and the civility of police–civilian encounters [11, 12, 34, 42, 49]. Recent advancements in machine learning have rendered the analysis of raw BWC footage computationally viable, overcoming previous constraints imposed by the sheer volume of video data. Nevertheless, current work remains focused on transcribing audible conversations and then using natural language processing techniques to study and potentially improve law enforcement behavior [13, 14, 51, 60, 64]. Apart from [30], which introduces a video annotation tool, and [3], which provides a face dataset based on body-worn camera footage, the analysis of the video itself remains largely unexplored, even though it offers rich context for understanding events. Nonetheless, several companies claim to analyze such footage without supporting research on their models’ performance [1, 6, 21, 50, 62]. To the best of our knowledge, EgoPolice is the first police BWC dataset designed to evaluate the video understanding capabilities of models, providing a first step toward enabling independent verification and rigorous auditing of AI tools deployed in this domain.

Egocentric data/methods. Egocentric videos [8, 18, 23, 24, 43, 48] capture videos from a first-person viewpoint, often by a head-mounted or a chest-mounted camera. Datasets used in controlled research settings typically capture actions in scripted or semi-scripted lab environments [8, 48]. However, recent works have extended this to unscripted, everyday activities [18, 19, 23], aligning much closer with our contribution. Nonetheless, the majority of these datasets are collected in settings where camera wearers and/or people included in the footage are aware that they are being recorded for research [23, 24, 43]. In addition, most clips record only a few individuals and primarily capture routine activities such as cooking, playing sports, or visiting parks [9, 18, 23]. EgoPolice targets these existing gaps by proposing a dataset that uses recordings of diverse and unscripted, but also *real and high-stakes* encounters between law enforcement officers and civilians.

Video annotation. The development of scalable video annotation, which demands both robust infrastructure and significant human labor, has been a recurring challenge in computer vision [29, 40, 54, 65], particularly for unscripted, real-world, in-the-wild videos that cover a huge variety of scenarios [18, 26, 61]. Previous work finds that minimizing cognitive load and decomposing annotation into binary subtasks are central design principles for maximizing throughput in such settings [58, 65]. To address the cost of manual labeling, recent works utilize vision models and LLMs [15, 27, 28] in model-in-the-loop frameworks, combining automated predictions with human annotators to label new and more challenging

data [31, 46, 53, 56]. However, LLM-assisted annotation can distort label distributions, introduce systematic errors, and propagate biases into ground truth during annotation [4, 57]. In high-stakes domains, annotator subjectivity adds additional complexity as disagreement over temporal boundaries even for well-defined classes is common [10, 20, 33, 63]. This difficulty is further compounded when footage requires multiple labels to be assigned to every frame [67]. We address these challenges by employing a multi-stage pipeline with objective, intent-free definitions, ensuring accurate annotations of the high-stakes dataset.

3 Data Collection Pipeline

3.1 Data Sources

EgoPolice consists of police BWC footage collected from multiple sources across the United States. The dataset includes recordings released by police departments (PD), independent news networks, and a civilian oversight office. Our main source is the Chicago PD, published by the City’s Civilian Office of Police Accountability (COPA). We collected all BWC videos released between July 10, 2016 and June 23, 2024. The footage includes incidents involving civilians who were critically injured or died in police custody, as well as all instances in which an officer discharged a firearm. The videos are grouped into 83 *cases*. A case corresponds to an incident involving between one and 33 officers. Similarly, we collected BWC footage from five other departments (Pasadena PD, Dallas PD, Metropolitan PD of the District of Columbia, Los Angeles PD, and San Antonio PD) and two news networks, which provide recordings from additional departments. Only COPA’s footage is organized into cases.

Videos from COPA and Pasadena are subject to only minor edits, involving occasional audio omissions, particularly at the beginning of videos, and the censoring of nudity. Videos from other departments are heavily edited (*e.g.* subtitles, press-conference-style statements, slowed-down segments, or zoom-ins). We remove non-BWC segments during pre-processing, and refer the reader to the Supplementary Material for sampled frames illustrating the difference.

EgoPolice does not reflect the breadth of police activity, most of which goes unrecorded or unreleased, but rather *overrepresents* high-stakes police-civilian interactions, as footage containing these incidents is more likely to be released.

3.2 Annotation

Compiling a list of police-civilian interactions. Our dataset captures a broad spectrum of interactions between law enforcement officers and civilians, ranging from routine conversations to high-stakes incidents such as shootings or medical emergencies. Many event types are rare due to the vast space of possible encounters. To address this challenge, we first compile a list of actions that are critical for policing research, explicitly distinguishing between the officer wearing

Table 1: Action labels and definitions in EgoPolice.

Role	Action	Definition
BWC Wearer	Physical Interaction	• The Body-Camera Wearer is touching the civilian.
	Medical Treatment	• BWC Wearer is touching the civilian with hands. • Both civilian and BWC Wearer’s actions are visible. • BWC Wearer is directly involved in treatment.
	Weapon Out	• The clip displays a firearm held by the BWC Wearer.
	Running	• The camera is moving/shaking substantially. • The officer is moving forwards or backwards quickly.
Other Officer	Physical Interaction	• An officer (not the BWC wearer) is touching the civilian.
	Medical Treatment	• An officer (not the BWC wearer) is touching the civilian with hands. • Both civilian and police are visible. • The officer is directly involved in treatment.
Any Police Officer Handcuffing		• Any officer is attempting to use or using handcuffs. • Handcuffs are clearly identifiable.
Civilian	Visibly Injured	• The civilian is visibly injured. • The clip shows bleeding or a wound.
	On Ground	• The civilian is sitting, kneeling, or lying on the ground. • Ground position is clear from the video.

the camera (BWC wearer), other officers visible in the recording who are not the recording officer (Other Officer), and civilians involved (Civilian).

We then group these actions into classes that occur frequently enough to support computer vision research while retaining social science significance. To avoid subjective interpretation and reduce disagreement between annotators, we defined each action using objective criteria, ensuring that no prior knowledge of policing is required for annotation. Table 1 summarizes the action taxonomy. For handcuffing, we combine officers into “Any Police Officer” since multiple officers are often in close proximity, which makes it difficult to reliably distinguish individual hands. Figure 1 presents example frames for four different actions, illustrating both easy and challenging instances for each class.

Labeling actions. We designed the annotation process as a two-stage pipeline. In Stage 1, which serves as an initial filtering step, three independent annotators reviewed the candidate videos and flagged any timestamps containing visible interactions, including verbal exchanges between police officers and civilians. To ensure comprehensive coverage, we aggregated these annotations by taking the union of all identified time windows and adding a 10-second buffer before and after the start and end of each interaction. In Stage 2, the resulting interaction-rich clips were assigned to annotators for labeling across the compiled action categories at a per-second granularity. Given the subjective and high-stakes nature of the data, annotators were provided with objective action definitions that explicitly avoid inferring intent or making normative judgments. For example,

Table 2: Data sources.

Dept.	Source	# Vids	Time (h:mm)
Chicago PD	COPA	600	109:15
Los Angeles PD	Dept.	82	2:03
D.C. PD	Dept.	418	12:20
Pasadena PD	Dept.	76	10:44
San Antonio PD	Dept.	135	3:03
Dallas PD	Dept.	38	3:00
Misc. PDs	News	1,335	44:29
Total		2,684	184:57

Table 3: Annotated actions per data source.

Class	COPA Pasadena Others			EgoPolice	
	# sec.	# sec.	# sec.	# sec.	%
BWC-Physical Interaction	13,772	1,749	17,036	32,557	4.89%
BWC-Medical Treatment	6,387	851	644	7,882	1.18%
BWC-Weapon Out	23,961	1,142	15,063	40,166	6.03%
BWC-Running	2,893	434	3,297	6,624	0.99%
Other-Physical Interaction	29,167	2,553	13,223	44,943	6.75%
Other-Medical Treatment	15,810	1,376	1,042	18,228	2.74%
Any Officer-Handcuffing	1,138	152	2,310	3,600	0.54%
Civilian-Injured	25,274	1,041	1,793	28,108	4.22%
Civilian-On Ground	43,083	2,865	18,025	63,973	9.61%
Any Class	75,542	6,122	49,075	130,739	19.63%

the action “BWC Wearer-Physical Interaction” is defined as cases in which *the officer wearing the camera is touching the civilian*, rather than relying on subjective interpretation. This approach ensures that action labels are based on observable behavior, reducing ambiguity for both annotators and researchers. As “BWC Wearer-Running” and “BWC Wearer-Weapon Out” can occur outside Stage 1 segments, these actions are annotated over the entire dataset.

We obtain second-by-second annotations by determining whether the corresponding action class is visible at any given second. Videos from COPA and Pasadena, which are of higher quality and serve as the primary data source for training and evaluation, are annotated independently by at least two annotators to ensure high-quality labels, while videos from the remaining departments are used as auxiliary training data and annotated by a single reviewer. Table 2 summarizes the data sources and Tab. 3 shows the prevalence of each annotated class stratified by origin. In total, EgoPolice contains over 180 hours of BWC footage. We refer the reader to the Supplementary Material for additional statistics.

3.3 Annotator Management

Our Institutional Review Board determined our project to be non-human-subjects research. Given the nature of the dataset, we implemented additional safeguards, incorporating strategies rarely used in prior data collection, to protect annotators’ mental health and ensure data quality.

Hiring procedure. To ensure the best quality, we opted to directly hire annotators from a local university, rather than using crowdsourcing platforms. We paid the annotators \$16/hour when the project began, which increased to \$18/hour for the final tasks. We hired 33 annotators who spent approximately 5,400 hours with the annotation cost totaling around \$92,000. The job listing clearly stated that “annotators that work on this project will be exposed to distressing or graphic images and sounds [...] we ask potential applicants to consider their mental and emotional ability to work with such footage before applying.” Next, prior to hiring annotators, we interviewed each prospective annotator, again emphasizing the potential challenges to mental health due to the nature of the body camera footage. Once the annotators were hired, we asked them to sign an

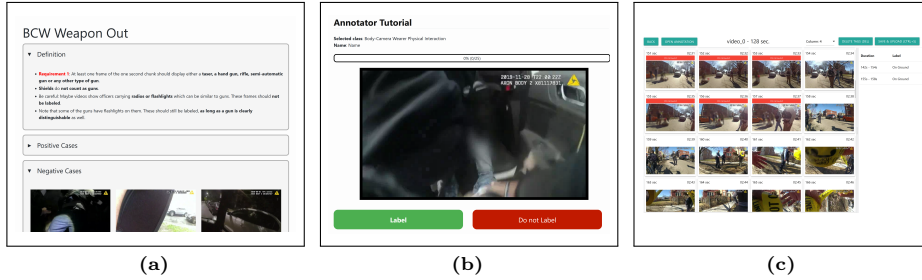


Fig. 2: Annotation Interface. (a) Excerpt from the document providing action definitions with paired positive and negative examples. (b) Training module with practice clips. (c) Custom web-based annotation tool to support efficient labeling.

acknowledgment of risk form, which provided even more detail about the types of situations that may be visible in the body camera footage.

Mitigating vicarious traumatization. Research has shown that repeated exposure to violent footage can cause vicarious traumatization [59]. We aimed to clarify this risk as early as possible, beginning with the job listing itself and reinforced again during the interview, so that annotators could make an informed decision before ever committing to the role. Prior to beginning annotation work, we provided mental health training, outlining vicarious traumatization and secondary traumatic stress and provided tools to cope with challenging video clips including taking breaks, working during the day, and self-care practices.

Initial training. We provided structured onboarding and conducted practice annotations with feedback to ensure consistency across labels. We compiled a document (fig. 2a) that included definitions along with positive and negative examples for every class. After reviewing the definitions, annotators completed a tutorial (fig. 2b) consisting of 25 practice annotations. The tutorial served as an assessment with detailed feedback explaining inclusion and exclusion criteria.

Annotation tool. We reviewed available off-the-shelf annotation tools, but ultimately decided to develop our own to ensure quality and speed of annotation. Our tool allows both managers to assign video clips to annotators, and annotators to have password-protected access to their unique assignments. Annotators can easily click-and-drag across multiple one-second frames of a video, which allows for a faster annotation process than existing tools that require multiple clicks to annotate a single time window. Unlike existing video annotation tools where annotators often have a single large video player in the center of the screen, we found that showing a grid of 1-second clips that each loop led to much faster annotation without increase in annotation errors. Additionally, we found less discomfort from observing disturbing scenes when the video is presented as a small grid versus being played in a full-sized video player. Although the tool

supports multi-label annotation, we assigned the annotators one action item at a time when labeling. Throughout our process, we found that annotators often labeled incorrectly or omitted labels due to complexity and annotation fatigue when given multiple actions to label simultaneously. Figure 2c shows a screenshot of our annotation tool which is developed with PHP and JavaScript, using MySQL as the main database. The tool’s source code is available on the project website, allowing anyone to download and host it on their own web server.

Feedback and monitoring. We implemented ongoing mental health check-ins and close monitoring to support annotators’ long-term well-being. In parallel, we continuously monitored annotation quality by reviewing annotations from each annotator to ensure adherence to the action definitions. When we found deviations, we provided targeted feedback, retraining, and retroactive correction of affected annotations. Furthermore, we manually verified approximately 25% of all annotations and viewed every video in the dataset. We find that our mean inter-annotator agreement (Krippendorff’s α [37]) is 79.4%.

4 Data Challenges

The characteristics of BWC footage can make action recognition more challenging than in conventional video benchmarks. We highlight two properties of EgoPolice that contribute to this difficulty: severe camera motion during annotated actions and limited discriminative signal from global appearance.

Egocentric Camera Motion. Most egocentric datasets capture routine, low-stakes activities [18, 23]. These datasets are designed for machine learning research, so actions are typically recorded clearly, with minimal camera movement and the camera generally pointed at the target. In real-world settings, however, actions can occur amid significant motion from either the subject or the camera, and in high-stakes environments such as policing, events like foot pursuits or physical altercations are inherently associated with extreme camera motion.

To quantitatively compare the amount of motion across datasets, we compute optical flow magnitude using FastFlowNet-v2 [36]. Figure 3 (a,b) shows a density plot of optical flow magnitudes for EgoPolice and other video datasets [18, 23, 29, 35]. When computed over the entire dataset (fig. 3a), EgoPolice exhibits only slightly more motion than Ego4D [23], another standard in-the-wild egocentric dataset. Differences emerge when motion statistics are computed specifically on annotated action segments (fig. 3b). In existing datasets, optical flow magnitudes drop off quickly during action segments, reflecting that the camera is typically stationary while interacting with objects or completing tasks. In contrast, EgoPolice shows a heavy-tailed distribution, with substantially more probability mass at high flow values (> 20). This highlights a central challenge in our domain: the moments that are most semantically important contain severe visual degradation, including blur, large displacements, and rapid viewpoint shifts.

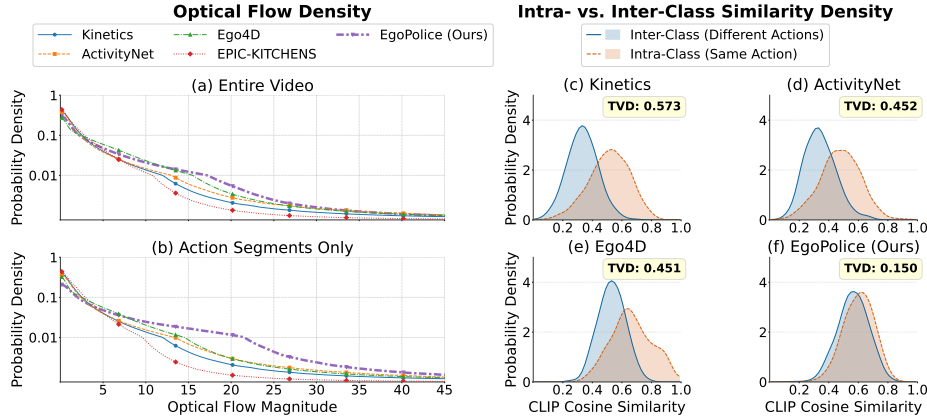


Fig. 3: Comparisons with existing datasets. **(a-b)** EgoPolice shows higher motion overall, which is more noticeable around segments where actions are labeled. **(c-f)** Distribution of CLIP cosine similarity between frames from the same class and frames from different classes. EgoPolice shows similar distributions for both, while other datasets show higher similarity when frames are sampled from the same class.

Visual Cue Diversity. In many standard video benchmarks, action classes are strongly tied to backgrounds or overall scenes (*e.g.*, “playing soccer” and “playing cards”). As a result, models can often rely on the global scene context to classify an action. In contrast, the high-stakes scenarios in EgoPolice unfold in far more visually uniform settings such as streets, hallways, and vehicle interiors.

We illustrate this challenge in fig. 3 (c-f) by plotting the cosine-similarity distributions of CLIP embeddings [52] for pairs of random frames sampled either within the same action class or across different classes. In datasets like Kinetics [35] and ActivityNet [29], these distributions differ substantially, reflected in high Total Variation Distance (TVD) scores (0.573 and 0.452). This indicates that frames from the same class tend to be noticeably more similar to each other than to frames from other classes. This property allows models to rely on scene context and exploit shortcuts to classify actions. In contrast, for EgoPolice, the two distributions almost overlap perfectly, with a TVD of only 0.150. Moreover, both distributions are shifted toward higher similarity values, around 0.6, which is substantially higher than Kinetics and ActivityNet in particular. This combination of low inter-class separability and high overall visual similarity indicates that global appearance cues can offer little discriminative signal for the model to work with. As a result, models must rely on genuine contextual understanding and fine-grained visual cues to correctly determine the visible action.

5 Experiments

To better understand how models transfer to the BWC domain, we evaluate EgoPolice under two paradigms reflecting dominant approaches in video under-

standing. These are supervised classification, which probes representation quality under task-specific training, and multiple-choice question answering (MCQ) with video-language models, which assesses zero-shot reasoning. Together, these experiments quantify how well existing models handle this challenging setting.

5.1 Linear probing for action classification

Model families and training protocols. Training strong video models from scratch typically requires orders of magnitude more data than we can feasibly collect for the BWC domain. Instead we evaluate how well representations transfer to this setting. To this end, we use standard embedding models [44, 47, 52, 55, 66] as frozen feature extractors, training lightweight classification heads on top of the resulting representations. For models with a CLS token, we use the token as the global frame-level representation. For models that do not use a CLS token, we compute the 1-second-level representation by mean-pooling the frame-level features, sampling frames at 16 fps. Training follows the linear-probing protocol of DINOv2 [47]. We evaluate models under 1-second, 10-second and 1-minute window settings, where a window is labeled positive if the action happens at least once. For evaluation, predictions are made independently for each second and aggregated over each window by taking the maximum score. We report both class-level F1 and mean F1 score averaged across classes in Tab. 4.

Data splits and generalization. We evaluate classification performance using 6-fold cross-validation. As described in Sec. 3, videos from COPA and Pasadena are minimally edited and thus best reflect real-world BWC footage. Videos from the remaining departments are often heavily edited (*e.g.*, with overlaid subtitles), making them unsuitable as clean evaluation data. For this reason, we restrict our cross-validation splits to COPA videos. All COPA *cases* recorded *before* 2024 are randomly split into six folds, ensuring each fold contains a similar number of police cases. For each fold, one split is used as the test set (called the **in-distribution (ID)** test set), while the remaining cases are split into training and validation sets with a 4:1 ratio. Splitting is performed at the case level to prevent leakage across splits (as cases are constituted by a varying number of videos of the same event recorded by different officers). Videos from the remaining departments are treated as auxiliary training data and are added to each fold’s training set as augmentation, but are not included in validation or test splits. In addition, we evaluate generalization under two out-of-distribution settings. First, **Out-of-distribution Time (OT)** includes COPA videos recorded in 2024, measuring temporal generalization since these videos are recorded in the same location but later in time than the data in the training set. Second, **Out-of-distribution Location (OL)** includes videos from the Pasadena Police Department, measuring geographic and departmental transfer.

Table 4: Linear probing classification results.

Model	1sec			10sec			1min		
	ID	OT	OL	ID	OT	OL	ID	OT	OL
<i>Random Baseline</i>	7.9	13.9	6.7	11.3	20.2	8.4	18.1	32.3	13.3
Hiera [55]	34.2	40.9	26.1	44.8	48.9	33.4	48.8	56.4	35.6
DINOv2 [47]	40.2	<u>43.9</u>	25.7	<u>49.4</u>	<u>50.5</u>	35.3	54.2	<u>56.9</u>	<u>42.0</u>
X-CLIP [44]	36.2	39.9	<u>30.8</u>	49.3	48.3	<u>45.8</u>	<u>55.4</u>	54.7	51.8
CLIP [52]	<u>39.2</u>	42.1	28.5	48.5	49.8	40.6	53.5	54.9	46.1
VideoMAE V2 [66]	37.9	48.5	30.9	51.8	59.3	46.2	58.4	63.7	<u>49.8</u>
Per-Class F1 Breakdown (ID, 1s)									
Model	<i>BWC</i>			<i>Oth. Off. Any</i>			<i>Civ.</i>		
	MT	PI	Run Wpn	MT	PI	Cuf	Inj	Gnd	
Random Baseline	2.9	5.9	0.9	12.2	6.9	12.3	0.6	11.2	18.2
CLIP [52]	40.0	<u>63.4</u>	<u>2.9</u>	<u>30.7</u>	36.4	52.6	<u>4.1</u>	56.3	66.0
VideoMAE V2 [66]	<u>31.2</u>	63.7	15.0	40.0	<u>25.1</u>	<u>49.7</u>	15.8	<u>39.7</u>	<u>61.2</u>

Table 5: Zero-shot VLM results.

Model	param	1sec	10sec	1min
<i>Random Baseline</i>	-	20.0	20.0	20.0
Proprietary Models				
GPT-4.1-Nano [2]	-	18.6	18.7	10.8
Gemini 2.5 Flash [22]	-	65.8	69.7	66.1
GPT-4.1 [2]	-	77.7	<u>76.2</u>	<u>74.4</u>
Gemini 2.5 Pro [22]	-	<u>76.9</u>	77.5	76.9
Open-source Models				
LLaVA-Mini [70]	8B	15.1	15.7	15.5
LLaMA-VID [38]	7B	38.1	37.7	30.2
VideoLLaVA [39]	7B	39.4	35.6	31.8
MERV [17]	7B	43.0	41.2	39.5
CogVLM2-Video [32]	8B	46.4	48.7	43.8
Qwen2.5-VL [7]	7B	52.5	55.8	55.5
VideoLLaMA3 [69]	7B	<u>58.0</u>	<u>61.1</u>	<u>58.5</u>
InternVL3 [71]	8B	64.7	64.5	60.6

5.2 Zero-shot evaluation of Video-LLMs

In addition to trained classifiers, we evaluate video-language models on EgoPolice in a zero-shot multiple-choice question answering setting. This probes semantic understanding and temporal reasoning without task-specific fine-tuning. We construct a benchmark of 12,000 questions, with 500 questions per action (including *None of the above*) on 1-second and 10-second-long clips and 200 questions per action on 1-minute-long clips. Each question consists of one correct answer and four incorrect choices. Wrong action choices are uniformly randomly selected among actions that never happened in the clip, with the last option always being *None of the above*. For image-based VLMs [2, 22], we sample frames at 1 fps with a minimum of 8 frames and provide them as image inputs. For video-based VLMs [7, 17, 32, 38, 39, 69–71], we follow the author-recommended video sampling strategies. All models are evaluated in a zero-shot setting using a fixed prompt template shared across models. We only use videos from COPA and Pasadena.

5.3 Results

Quantitative analysis. Table 4 and Tab. 5 display the performance of the trained classification models and the zero-shot VLM models, respectively.

In the classification setting, we find VideoMAE V2 [66] and X-CLIP [44], both trained on video data, to be the best-performing models outperforming CLIP [52] and DINOv2 [47] which are trained on images. These results suggest that video pretraining is useful and allows for a more robust transfer to unseen situations. Table 4 (bottom) shows that models perform well on classes such as “Civilian-Injured” and “Civilian-On Ground”, due to strong visual cues such as visible injuries. However, performance is substantially lower for “Any Officer-Handcuffing”. This is partly due to the rarity of this class relative to others in the dataset, and partly because the task requires detecting handcuffs that are often small and partially occluded. Additionally, we observe that CLIP beats VideoMAE V2 on some appearance-grounded classes while VideoMAE V2 leads on

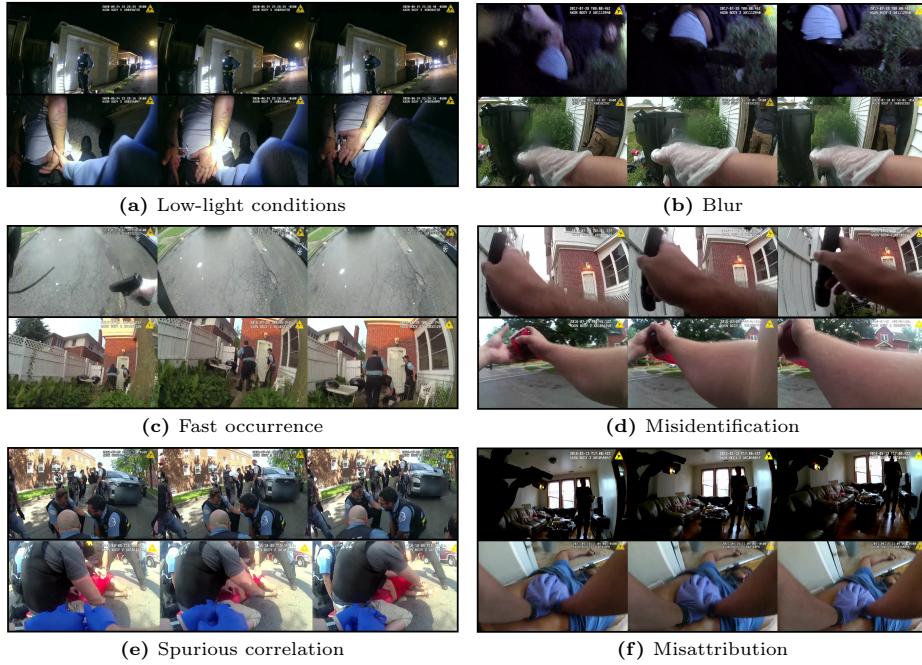


Fig. 4: Examples of common failure modes. EgoPolice combines in-the-wild egocentric video challenges with law-enforcement activity recognition. These include low-light occlusion, motion blur, rapid event occurrence, object misidentification, and action misattribution, which frequently co-occur within the same clip. Together, they highlight the gap between benchmark performance and the reliability required for real-world deployment.

motion-dependent ones, suggesting that robust BWC action recognition requires combining appearance and motion-based representations.

In the zero-shot task, we see LLaVA-Mini [70] and GPT-4.1-Nano [2] performing worse than random guessing due to outputting captions instead of answering the corresponding letter of the correct choice or strong tendencies to choose *None of the above* option. There is a tendency for some models to perform better on 10-second clips than on 1-second clips. We believe that 1-second clips may be too short for VLMs to correctly analyze the action happening in the video, while 10-second clips provide some temporal context that the model can use to infer the correct label. Meanwhile, with 1-minute clips, the task turns into a *needle-in-the-haystack* problem, leading to a reduction in accuracy. Among the models we have tested, we find Gemini 2.5 Pro [22] to perform the best. However, given its performance is only 76.9% on 1-minute video clips, we argue that this model is not yet suitable for autonomous, real-world deployment.

Qualitative analysis. We observe consistent failure modes across all classification models. Figure 4 groups representative examples by type, with each sequence showing three uniformly sampled frames from a one-second clip. Performance drops in low-light conditions, where flashlights create over- and underexposed regions that obscure actions such as “Civilian-On-Ground” or “Any Officer-Handcuffing” leading to false negatives (fig. 4a). Blur from motion or lens obstructions also degrades performance, as in fig. 4b (bottom), where a water droplet hides a gun. Short event durations cause additional errors (fig. 4c), since actions (civilian on ground or weapon out) only appear within part of a one-second window. We also observe false positives from misidentification (fig. 4d), where objects like flashlights or red police tape are confused with firearms. In addition, models rely on spurious correlations, such as using a red T-shirt as a cue for injury (fig. 4e), and misattribute actions in complex scenes, such as assigning a taser held by another officer to the BWC wearer (fig. 4f).

6 Real-World Police Department Application

EgoPolice was designed by both social and computer scientists with the goal of supporting responsible analysis of real-world BWC footage. Many law enforcement agencies maintain large repositories of videos which capture police-civilian interactions of significant evidentiary value, yet the volume of this footage makes consistent human review practically infeasible. EgoPolice provides a foundation for developing models that can efficiently identify events of interest.

In ongoing work, we have found that models trained on EgoPolice can successfully support large-scale analysis of body-worn camera videos in a human-in-the-loop setting, making systematic review of large BWC repositories tractable. In this setting, we focus on the action class “BWC Wearer-Physical Interaction.” Rather than treating model predictions as final determinations, outputs are used to surface candidate video segments that are subsequently reviewed and verified by trained annotators. This workflow has proven effective at surfacing relevant events for review and illustrates the potential of models trained on EgoPolice to support scalable analysis of body-worn camera footage.

This deployment suggests two preliminary insights. First, models trained on EgoPolice learn representations that transfer to real-world BWC data. Second, the immediate value lies in assisting human reviewers by narrowing attention to potentially critical moments. This way, EgoPolice facilitates scalable auditing and retrospective analysis of large volumes of body-worn camera footage. A full analysis of this deployment will be presented in a forthcoming companion paper. We include this discussion to emphasize that EgoPolice is not only a challenging benchmark, but also a concrete example of how to design pipelines that can be deployed in real-world settings where human judgment remains central.

7 Ethical Concerns

In this work, we study police body-worn camera (BWC) footage, a sensitive domain in which individuals may be recorded without consent. Although all videos in EgoPolice are publicly available, privacy concerns remain. We restrict annotations to coarse action-level labels and avoid personal attribute annotations. We also recognize that BWC footage can be distressing for annotators and viewers. For annotators, we mitigate this through assigning narrowly scoped tasks for short clips (instead of full length videos), providing intensive training, and developing a custom tool designed for this type of video. For viewers, we provide a trigger warning urging caution before downloading and examining the dataset.

We acknowledge that releasing BWC footage raises ethical questions, while recognizing that withholding it is not a normatively neutral choice. BWC technology exists to promote accountability, and the public has a role in that process. We believe it is appropriate to release such recordings when there is a clear public interest (*e.g.* promoting accountability, enabling independent oversight, informing public understanding of policing and the lawful use of force, or supporting research) and appropriate safeguards protect the dignity of those depicted.

Beyond privacy, automated analysis of police footage carries broader risks, including reinforcing institutional biases and enabling unaccountable decision-making. High performance on EgoPolice should not be interpreted as a basis for deployment. Instead we aim to highlight the potential and limitations of current systems. It is also important to note that the data may reflect demographic and institutional biases tied to the jurisdictions from which it was collected.

8 Conclusion

In this work, we present EgoPolice, a dataset of egocentric police-civilian interactions designed to challenge vision models with complex, real-world scenarios. Our analysis shows that police-recorded footage is a particularly difficult domain. Hence, EgoPolice serves as a challenging benchmark for computer vision researchers focused on advancing research on vision models. Simultaneously, we believe that EgoPolice can provide a resource for developing video understanding systems that are not only more capable, but also more responsible and accountable. As automated video analysis becomes more prevalent in law enforcement, the lack of transparent, rigorous benchmarks increases the risk of unexamined bias and system failures. We hope EgoPolice encourages the community to confront the ethical, technical, and social challenges that arise as models are increasingly deployed in sensitive, real-world environments.

Acknowledgements

This work is partially supported by the Princeton University Data-Driven Social Science, the Princeton Language and Intelligence (PLI) Initiative, the Microsoft Corporate, External, and Legal Affairs (CELA) legal diversity gift, and Princeton School of Science and Engineering awards from the Helen Shipley Hunt fund and from the Project X fund. We also thank all the annotators for the help curating the dataset, Felix Yu and Yu Wu for help with initiating the project, and Sanghyuk Chun for valuable feedback on the paper.

References

1. Abel Police: Abel police: Automating police paperwork (2026), <https://abelpolice.com/>, accessed: 2026-02-09
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
3. Almadan, A., Krishnan, A., Rattani, A.: Bwcfacenet: Open-set face recognition using body-worn camera. In: International Conference on Machine Learning and Applications (2020)
4. Ashwin, J., Chhabra, A., Rao, V.: Using large language models for qualitative analysis can introduce serious bias. arXiv:2309.17147 (2023)
5. Attia, A., Webb, J., Connor, K., Johnston, C.J., Williams, M., Gordon-Walker, T., Rowe, I.A., Harrison, E.M., Stutchfield, B.M.: Effect of recipient age on prioritisation for liver transplantation in the uk: a population-based modelling study. *The Lancet Healthy Longevity* (2024)
6. Axon Enterprise, Inc.: Axon draft one (2026), <https://www.axon.com/products/draft-one>, accessed: 2026-02-09
7. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv:2502.13923 (2025)
8. Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Zhang, F., Fountain, J., Miller, E., Basol, S., Newcombe, R., Wang, R., et al.: Introducing hot3d: An egocentric dataset for 3d hand and object tracking. arXiv:2406.09598 (2024)
9. Bansal, S., Arora, C., Jawahar, C.V.: My view is the best view: Procedure learning from egocentric videos. In: ECCV (2022)
10. Barrett, D.P., Xu, R., Yu, H., Siskind, J.M.: Collecting and annotating the large continuous action dataset. *Mach. Vis. Appl.* (2016)
11. Braga, A.A., MacDonald, J.M., McCabe, J.: Body-worn cameras, lawful police stops, and nypd officer compliance: A cluster randomized controlled trial. *Criminology* (2022)
12. Braga, A.A., Sousa, W.H., Coldren Jr, J.R., Rodriguez, D.: The effects of body-worn cameras on police activity and police-citizen encounters: A randomized controlled trial. *Journal of Criminal Law and Criminology* (2018)
13. Camp, N.P., Voigt, R.: Body camera footage as data: Using natural language processing to monitor policing at scale & in depth. *Behavioral Science & Policy* (2024). <https://doi.org/10.1177/23794607241308636>
14. Camp, N.P., Voigt, R., Hamedani, M.G., Jurafsky, D., Eberhardt, J.L.: Leveraging body-worn camera footage to assess the effects of training on officer communication during traffic stops. *PNAS Nexus* (2024). <https://doi.org/10.1093/pnasnexus/pgae359>

15. Chen, T., Siarohin, A., Menapace, W., Deyneka, E., Chao, H., Jeon, B.E., Fang, Y., Lee, H., Ren, J., Yang, M., Tulyakov, S.: Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In: CVPR (2024)
16. Chen, Z., Asadi Shamsabadi, E., Jiang, S., Shen, L., Dias-da Costa, D.: Integration of large vision language models for efficient post-disaster damage assessment and reporting. *Nature Communications* (2026)
17. Chung, J., Zhu, T., Saez-Diez, M.G., Niebles, J.C., Zhou, H., Russakovsky, O.: Unifying specialized visual encoders for video language models. In: ICML (2025)
18. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV (2018)
19. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *IJCV* (2022)
20. Davani, A.M., Díaz, M., Prabhakaran, V.: Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Trans. Assoc. Comput. Linguistics* (2022)
21. Dyno Public Safety: Dyno public safety (2026), <https://www.dynopublicsafety.com/>, accessed: 2026-02-09
22. Gemini Team, Google: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv:2507.06261* (2025)
23. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR (2022)
24. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024)
25. Grother, P., Ngan, M., Hanaoka, K.: Face Recognition Vendor Test Part 3: Demographic Effects. National Institute of Standards and Technology (2019). <https://doi.org/https://doi.org/10.6028/NIST.IR.8280>
26. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: AVA: A video dataset of spatio-temporally localized atomic visual actions. In: CVPR (2018)
27. Gutiérrez, J., Gutiérrez, V., Ángel Mora, Rodríguez, S., Blanco, J.L.: An evaluation of hybrid annotation workflows on high-ambiguity spatiotemporal video footage. *arXiv:2510.21798* (2026)
28. He, J., Wang, X., Wong, K.K., Huang, X., Chen, C., Chen, Z., Wang, F., Zhu, M., Qu, H.: Videopro: A visual analytics approach for interactive video programming. *IEEE Transactions on Visualization and Computer Graphics* **30**(1), 87–97 (2023)
29. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR (2015)
30. Hejabi, P., Padte, A.K., Golazizian, P., Hebbar, R., Trager, J., Chochlakis, G., Kommineni, A., Graeden, E., Narayanan, S., Graham, B.A., Dehghani, M.: Cvat-bwv: a web-based video annotation platform for police body-worn video. In: *IJCAI* (2024)
31. Hong, W., Cheng, Y., Yang, Z., Wang, W., Wang, L., Gu, X., Huang, S., Dong, Y., Tang, J.: Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In: CVPR (2025)

32. Hong, W., Wang, W., Ding, M., Yu, W., Lv, Q., Wang, Y., Cheng, Y., Huang, S., Ji, J., Xue, Z., et al.: Cogvlm2: Visual language models for image and video understanding. arXiv:2408.16500 (2024)
33. Idrees, H., Zamir, A.R., Jiang, Y., Gorban, A., Laptev, I., Sukthankar, R., Shah, M.: The THUMOS challenge on action recognition for videos "in the wild". *Comput. Vis. Image Underst.* (2017)
34. Jennings, W.G., Lynch, M.D., Fridell, L.A.: Evaluating the impact of police officer body-worn cameras (bwcs) on response-to-resistance and serious external complaints: Evidence from the orlando police department (opd) experience utilizing a randomized controlled experiment. *Journal of Criminal Justice* (2015). <https://doi.org/https://doi.org/10.1016/j.jcrimjus.2015.10.003>
35. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. arXiv:1705.06950 (2017)
36. Kong, L., Shen, C., Yang, J.: Fastflownet: A lightweight network for fast optical flow estimation. In: ICRA. IEEE (2021)
37. Krippendorff, K.: Content analysis: An introduction to its methodology. Sage publications (2018)
38. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: ECCV (2024)
39. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: EMNLP (2024)
40. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
41. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Zhao, M., Chow, A.K., Ikemura, K., Kim, A., Pouli, D., Patel, A., et al.: A multimodal generative ai copilot for human pathology. *Nature* (2024)
42. Lum, C., Koper, C.S., Wilson, D.B., Stoltz, M., Goodier, M., Eggins, E., Higginson, A., Mazerolle, L.: Body-worn cameras' effects on police officers and citizen behavior: A systematic review. *Campbell systematic reviews* (2020)
43. Ma, L., Ye, Y., Hong, F., Guzov, V., Jiang, Y., Postyeni, R., Pesqueira, L., Gamino, A., Baiyya, V., Kim, H.J., Bailey, K., Fosas, D.S., Liu, C.K., Liu, Z., Engel, J.J., De Nardi, R., Newcombe, R.A.: Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In: ECCV (2024)
44. Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R.: X-CLIP: end-to-end multi-grained contrastive learning for video-text retrieval. In: ACM (2022)
45. MacMillan, D., Ovalle, D., Schaffer, A.: Arrested by ai: Police ignore standards after facial recognition matches. *The Washington Post* (2025), <https://www.washingtonpost.com/business/interactive/2025/police-artificial-intelligence-facial-recognition>, accessed: 2025-10-20
46. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS* (2023)
47. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. *TMLR* (2024)
48. Pan, X., Charron, N., Yang, Y., Peters, S., Whelan, T., Kong, C., Parkhi, O.M., Newcombe, R.A., Ren, C.Y.: Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In: ICCV (2023)

49. Patterson, Q., White, M.D.: Is there a civilizing effect on citizens? testing the pre-conditions for body worn camera-induced behavior change. *Police Quarterly* (2021)
50. Polis Solutions, Inc.: Polis solutions: Public safety technology and training (2026), <https://www.polis-solutions.ai/>, accessed: 2026-02-09
51. Prabhakaran, V., Griffiths, C., Su, H., Verma, P., Morgan, N., Eberhardt, J.L., Jurafsky, D.: Detecting institutional dialog acts in police traffic stops. *Transactions of the Association for Computational Linguistics* (2018). https://doi.org/10.1162/tac1_a_00031
52. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: *ICML* (2021)
53. Ravi, N., Gabeur, V., Hu, Y., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C., Girshick, R.B., Dollár, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. In: *ICLR* (2025)
54. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. *IJCV* (2015)
55. Ryali, C., Hu, Y., Bolya, D., Wei, C., Fan, H., Huang, P., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., Malik, J., Li, Y., Feichtenhofer, C.: Hiera: A hierarchical vision transformer without the bells-and-whistles. In: *ICML* (2023)
56. Salehi, M.R., Park, J.S., Kusupati, A., Krishna, R., Choi, Y., Hajishirzi, H., Farhadi, A.: Actionatlas: A videoqa benchmark for domain-specialized action recognition. *NeurIPS* (2024)
57. Schroeder, H., Roy, D., Kabbara, J.: Just put a human in the loop? investigating LLM-assisted annotation for subjective tasks. In: *Findings of the Association for Computational Linguistics: ACL 2025* (2025)
58. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: *ECCV* (2016)
59. Spence, R., Bifulco, A., Bradbury, P., Martellozzo, E., DeMarco, J.: The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* (2023)
60. Srbinovska, A., Srbinovska, A., Senthil, V., Martin, A., McCluskey, J., Fokoué, E.: Towards ai-driven policing: Interdisciplinary knowledge discovery from police body-worn camera footage. *arxiv:2504.20007* (2025)
61. Tezcan, O., Duan, Z., Cokbas, M., Ishwar, P., Konrad, J.: Wepdtof: A dataset and benchmark algorithms for in-the-wild people detection and tracking from overhead fisheye cameras. In: *CVPR* (2022)
62. Truleo, Inc.: Body camera analytics (2026), <https://truleo.co/>, accessed: 2026-02-09
63. Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M.: Learning from disagreement: A survey. *J. Artif. Intell. Res.* (2021)
64. Voigt, R., Camp, N.P., Prabhakaran, V., Hamilton, W.L., Hetey, R.C., Griffiths, C.M., Jurgens, D., Jurafsky, D., Eberhardt, J.L.: Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences* (2017). <https://doi.org/10.1073/pnas.1702413114>
65. Vondrick, C., Patterson, D., Ramanan, D.: Efficiently scaling up crowdsourced video annotation. *IJCV* (2013)

- 66. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae V2: scaling video masked autoencoders with dual masking. In: CVPR (2023)
- 67. Yeung, S., Russakovsky, O., Jin, N., Andriluka, M., Mori, G., Fei-Fei, L.: Every moment counts: Dense detailed labeling of actions in complex videos. IJCV (2018)
- 68. Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., Xu, M., Wei, X., Guo, N.: Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. arXiv:2505.17685 (2025)
- 69. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv:2501.13106 (2025)
- 70. Zhang, S., Fang, Q., Yang, Z., Feng, Y.: Llava-mini: Efficient image and video large multimodal models with one vision token. arXiv:2501.03895 (2025)
- 71. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv:2504.10479 (2025)